

The NCBI Handbook

2nd edition

Last Updated: January 26, 2018



National Center for Biotechnology Information (US)
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

NLM Citation: The NCBI Handbook [Internet]. 2nd edition. Bethesda (MD): National Center for Biotechnology Information (US); 2013-.

The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health, is a leader in the field of bioinformatics; it studies computational approaches to fundamental questions in biology and provides online delivery of biomedical information and bioinformatics tools. NCBI hosts approximately 40 online literature and molecular biology databases—including PubMed, PubMed Central, and GenBank—that serve millions of users around the world. The second edition of the NCBI Handbook, released in November 2013 in conjunction with the 25th anniversary of NCBI, aims to provide a comprehensive overview of the breadth of informatics resources at NCBI, and an in-depth account of the scope, data, infrastructure, processing, and access for each major database or resource. The databases and resources are organized here into seven concept areas: literature, genomes, variation, health, genes and gene expression, nucleotide, proteins, and small molecules and biological assays. Three additional categories encompass tools, infrastructure, and metadata. Each concept area begins with an overview chapter that provides a contextual framework for the resources discussed under that concept; the overview is followed by separate chapters that cover individual databases or resources.

As with the first edition, The NCBI Handbook 2nd Edition is geared towards advanced users of NCBI resources to provide an understanding of how bioinformatics resources at NCBI work. It is not a step-by-step user manual but complements NCBI user guides, tutorials, help information, and other existing documentation. It is our intent that the handbook will reflect, to the extent possible, the current state of databases, resources, and tools at NCBI, with information updated periodically.

Table of Contents

Editors and Reviewers	vii
A Brief History of NCBI's Formation and Growth	ix
Literature	1
NCBI Literature Resources	3
NLM Catalog	7
PubMed: The Bibliographic Database	13
PubMed Central	25
Bookshelf	61
NLM DTD to NISO JATS Z39.96-2012	71
The NIH Manuscript Submission System	77
GENOMES	85
What's in a Genome at NCBI?	87
GenBank	379
Protein Clusters	385
Eukaryotes	95
Clone	97
Genome Reference Consortium	115
Eukaryotic Genome Annotation Pipeline	133
Prokaryotes	157
About Prokaryotic Genome Processing and Tools	159
Prokaryotic Genome Annotation Pipeline	173
Viruses	187
About Viral and Phage Genome Processing and Tools	189
Virus Variation	205
Variation	213

Some chapters are represented in multiple sections in this book, so pagination in the Table of Contents may be out of sequence.

Variation Overview.....	215
The Database of Genotypes and Phenotypes (dbGaP) and PheGenI	223
The Database of Short Genetic Variation (dbSNP)	259
dbVar	299
ClinVar.....	313
Health.....	321
MedGen	323
ClinVar.....	313
Genes and Gene Expression	341
Genes and Gene Expression	343
Gene Expression Omnibus (GEO)	347
Gene.....	357
UniGene	363
Nucleotide	377
GenBank.....	379
Protein	381
NCBI Protein Resources	383
Protein Clusters	385
Small Molecules and Biological Assays.....	401
Small Molecules and Biological Activities.....	403
NCBI PubChem BioAssay Database	411
Tools.....	423
The BLAST Sequence Analysis Tool	425
The Entrez Search and Retrieval System	437
C++ Toolkit.....	443
LinkOut: Linking to External Resources from NCBI Databases	447
Metadata	457
BioSample	459

BioProject	467
Glossary	477

Editors and Reviewers

Editors

Jeffrey Beck

Dennis Benson

Janet Coleman

Marilu Hoeppner

Mark Johnson

Donna Maglott

Ilene Mizrachi

Rana Morris

Jim Ostell

Kim Pruitt

Wendy Rubinstein

Eric Sayers

Karl Sirotkin

Tatiana Tatusova

Reviewers

Kathi Canese

Renata Geer

Sharmin Hussain

Evgeny Kireev

Adriana Malheiro

Karanjit Siyan

Bart Trawick

Copy-editor

Stacy Lathrop

Assignments

Section	Editor(s) / Reviewer(s)
Front Matter	Dennis Benson Janet Coleman Bart Trawick
Literature	Jeffrey Beck
Genomes	Kim Pruitt Tatiana Tatusova
Variation	Donna Maglott Adriana Malheiro
Health	Wendy Rubinstein
Genes and Gene Expression	Donna Maglott
Nucleotide	Ilene Mizrachi
Protein	Eric Sayers
Small Molecules and Biological Assays	Rana Morris Renata Geer
Tools	Kathi Canese Sharmin Hussain Mark Johnson Evgeny Kireev Karl Sirotkin Karanjit Siyan
Infrastructure	Karl Sirotkin
Metadata	Ilene Mizrachi
Book	Marilu Hoeppner Jim Ostell

A Brief History of NCBI's Formation and Growth

Kent Smith^{✉1}

The establishment of the National Center for Biotechnology Information (NCBI) in November of 1988 occurred primarily through the convergence of three independent but related actions. They were:

- 1984-86—Advocacy groups convened meetings on Capitol Hill to educate legislators and their staffs on the value of supporting genomic research.
- 1986—NLM's Long Range Plan was completed; it contained a recommendation that a new NLM Division be created to manage and process molecular biology information.
- 1987—The House Select Committee on Aging, Chaired by Senator Claude Pepper, introduced a Bill to establish the NCBI.

In 1984, the Delegation for Basic Biomedical Research began briefing sessions on the Hill, using Nobel winners like Dr. James Watson and Dr. David Baltimore to inform legislators about the importance of genomic research as a new and integral part of the advancement of scientific research. These briefing sessions were thought to be critical in creating an atmosphere in Congress that was receptive to the creation of a biotechnology information center like that of the NCBI.

Also in 1984, Dr. Donald A. B. Lindberg became the director of the NLM, and soon thereafter led the National Library in a major long-range planning effort. Over 100 leaders in the biomedical community participated in this rigorous process, forming 5 panels covering the principal domains of NLM. Of particular note was panel 3—titled “Obtaining Factual Information from Data Bases”—which would prove to be the source from which the idea for NCBI was initially conceived.

The germination of the idea for a center emanated in great part from Dr. Allan Maxam, a professor of biological chemistry at Harvard who was a pioneer of molecular genetics and served as a key member of the 1986 NLM Long Range Planning Panel. He instructed his fellow panel 3 members on the importance they should assign to the field of biotechnology and informed them of the country's need to harness the large volume of data that would be generated by the oncoming genetic revolution in science. The Planning Panel, and the NLM Board of Regents, embraced the idea of the need for an organization that could serve as both a repository and distribution center for the growing body of genomic and genetic knowledge and also serve as a unique resource for developing new computer analysis and communication tools for managing molecular biology information.

¹ NCBI; Email: kents@ncbi.nlm.nih.gov.

[✉] Corresponding author.

The newly formed Friends of the NLM saw this as an opportune time to approach the Congress on the need for a biotechnology information center and sought out Senator Claude Pepper, a major champion for medical research. Realizing his congressional colleagues would need to be educated about the benefits of biotechnology research, Senator Pepper asked that NLM develop a document that could be used to explain the need for a center. The resulting document, known as “Talking One Genetic Language: The Need for a National Biotechnology Information Center,” formed the background for the introduction of the initial bill (H.R.5271) to create the NCBI as part of the NLM. The bill was introduced late in the congressional session, and no action was taken on it, but it was reintroduced in the following session by a determined Senator Pepper.

On March 6, 1987, Senator Pepper, as Chairman of the House Committee on Aging, introduced his new bill (H.R.393) to establish NCBI, stating that the center would deal “with nothing less than the mystery of human life and the unfolding scroll of knowledge, seeking to penetrate that mystery, which is life itself.” The hearing had a compelling slate of 15 witnesses, including senior federal and academic health officials as well as five patients who had benefitted from biotechnology.

Although the bill encountered a number of legislative obstacles, Senator Pepper kept the effort alive by securing an appropriation of \$3.85 million to begin the biotechnology information program at NLM. During this timeframe, Dr. Daniel Masys, director of the Lister Hill Center for Biomedical Communications, and his branch chief, Dr. Dennis Benson, initiated NLM’s early biotechnology information activities.

The following year Senator Pepper, with the help of Congressman Henry Waxman and Senators Edward Kennedy and Lawton Chiles, incorporated H.R.393 into the NIH reauthorization bill known as the Health Omnibus Extension Act P.L.100-607. It passed the Congress and was signed into law by President Reagan on November 4, 1988.

Following enactment, Senator Pepper, in ceremonies conducted in the Capitol’s Mike Mansfield Room, said about biotechnology and the new center: “I hope and pray it’s going to realize the dreams that many of us have cherished for a long, long time, by being able to prolong the lives and promote the health and happiness of human beings.”

The act stipulated the following functions for the new National Center for Biotechnology Information:

- 1 *design, develop, implement, and manage automated systems for the collection, storage, retrieval, analysis, and dissemination of knowledge concerning human molecular biology, biochemistry, and genetics;*
- 2 *perform research into advanced methods of computer-based information processing capable of representing and analyzing the vast number of biologically important molecules and compounds;*
- 3 *enable persons engaged in biotechnology research and medical care to use systems developed under paragraph (1) and methods described in paragraph (2); and*

- 4 *coordinate, as much as is practicable, efforts to gather biotechnology information on an international basis.*

Beginning with a modest budget of \$8 million and a dozen staff members, NCBI began its journey to become a national resource for molecular biology information. Dr. David Lipman, a key developer of the FASTA algorithm, was recruited from NIDDK and was appointed as NCBI Director. Along with the support of his three key appointees—Dr. Dennis Benson, Chief, Information Resources Branch; Dr. David Landsman, Chief, Computational Biology Branch; and Dr. James Ostell, Chief, Information Engineering Branch—Dr. Lipman rapidly grew the center into a major information hub in the molecular biology revolution.

NCBI is now a leading source for public biomedical databases, software tools for analyzing molecular and genomic data, and research in computational biology. Today NCBI creates and maintains over 40 integrated databases for the medical and scientific communities as well as the general public. There are over 3 million visitors daily to its website, approximately 27 terabytes of data downloaded per day, and the number of users as well as downloads increases dramatically each year.

Listed below are some of the major milestones from the many that have occurred over the past quarter of a century:

- **1990—BLAST**—the Basic Local Alignment Search Tool (BLAST) is introduced; optimized for speed, the sequence comparison algorithm quickly finds similar sequences to one's query.
- **1991—Entrez**—The search and retrieval system for NCBI's linked databases is introduced in CD form, allowing users to easily find related information from different databases.
- **1992—GenBank at NCBI**—NCBI assumes responsibility for GenBank, a database of nucleotide sequences, and collaborates in its development with international partners at the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).
- **1993—Network Entrez**—Network Entrez, a client-server version of the CD-ROM, is introduced, bringing Entrez to the Internet.
- **1994—NCBI Website**—NCBI establishes its own website, mounting initially BLAST, Entrez, **dbEST** (Expressed Sequence Tags), and **dbSTS** (Sequence Tagged Sites).
- **1995—Genomes**—This new resource organizes information on genomes, including sequences, maps, chromosomes, assemblies, and annotations.
- **1995—Bankit**—The online tool is introduced to facilitate submissions to GenBank.
- **1996—OMIM**—NCBI mounts the Online Mendelian Inheritance in Man (OMIM), a directory of human genes and genetic disorders.
- **1997—PubMed**—NCBI introduces PubMed, a freely accessible bibliographic retrieval system to the entire MEDLINE database. The new service is launched at a Capitol Hill event by Vice President Al Gore and the ranking Labor/HHS

Appropriation Subcommittee members, Senator Tom Harkin (D-IA) and Senator Arlen Specter (R-PA), highlighting its significance.

- **1998—New NIH Disease-Based Services**—Collaborations with NIH Institutes for Disease-Based Services are established such as CGAP, the Cancer Genome Anatomy Project, to identify the human genes expressed in different cancerous states.
- **1999—Human Genome**—Human Genome Project researchers completely sequence the first human chromosome (#22) and deposit the sequence data at NCBI. A working draft of the entire human genome is completed the following year and made freely accessible from NCBI.
- **1999—Suite of Genomic Resources**—NCBI releases a number of resources to support comprehensive analysis of the human genome, including: **LocusLink**—key descriptors of genetic loci; **RefSeq**—a non-redundant set of human reference sequences; and **dbSNP**—a collection of data on human genetic variation.
- **2000—PubMed Central**—NCBI debuts its free full-text digital archive of biomedical and life sciences journal literature. PubMed Central (PMC) serves as an online counterpart to NLM's extensive print journal collection and is in keeping with the National Library's legislative mandate to collect and preserve the world's biomedical literature.
- **2000—GEO**—The Gene Expression Omnibus database is launched in response to community interest in a public repository for data generated from high-throughput microarray experiments.
- **2001—Bookshelf**—The new Entrez database is introduced to provide free access to books and documents in the life sciences and healthcare fields.
- **2002—WGS**—GenBank begins including Whole Genome Shotgun sequences, which are generated by a semi-automatic technique.
- **2003—DTDs**—NCBI Introduces Document Type Definitions (DTDs) for archiving and exchanging journal content.
- **2003—Entrez Gene**—The Entrez Gene database (formerly known as LocusLink) is developed to supply key connections between maps, sequences, expression profiles, structure, function, homology data, and the scientific literature.
- **2004—PubChem**—The PubChem database is released, providing information on the chemical structure and biological activities of small molecules.
- **2005—NIH Public Access**—NIH develops a Public Access Policy to provide scientists, researchers, and the general public with access to the published results of NIH-funded research through NCBI's PubMed Central. NCBI develops a NIH Manuscript Submission System that allows researchers to submit their published papers to PubMed Central.
- **2005—My NCBI**—NCBI introduces the My NCBI tool, which retains user information and database preferences to provide customized services for many NCBI databases.
- **2006—dbGaP**—NCBI launches the database of Genotypes and Phenotypes (dbGaP) to archive and distribute the results of studies that investigate the interaction of genotypes and phenotypes. Studies include Genome-Wide

Association Studies (**GWAS**), medical sequencing, and molecular diagnostic assays, among others.

- **2007—Genome Reference Consortium**—The Consortium of NCBI, EBI, Sanger Institute, and the Genome Institute is created to improve the sequence quality and accuracy of the human reference genome. It takes on the task of improving the reference sequences for other model organisms, including the mouse and zebra fish, often used as models for human disease.
- **2008—Discovery Initiative**—NCBI embarks on a program to help users better explore the myriad of data contained in NCBI's resources. Automated methods are employed to surface related data that may not be apparent to the user in their original search query but which could lead to serendipitous discoveries.
- **2008—Public Access Becomes Mandatory**—Congress enacts legislation mandating that scientists submit final peer-reviewed journal manuscripts that arise from NIH funding to PubMed Central. The policy requires that the papers be made public on PubMed Central no later than 12 months after publication.
- **2008—1000 Genomes Project**—NCBI archives and distributes data from this international public-private consortium, which aims to build the most detailed map available of human genetic variations. In **2012**, NCBI improved the accessibility of these data by collaborating on an effort to make them available on the cloud through Amazon Web Services.
- **2010—dbVar**—NCBI establishes the dbVar archive of large scale genomic variation data and associated defined variants with phenotypic information.
- **2010—My Bibliography**—NCBI introduces the My Bibliography tool to simplify the process for gathering one's published articles and other materials. The tool, which is connected to NIH's grants management system, also assists researchers in complying with the NIH Public Access Policy.
- **2011—PubMed Health**—The service is introduced to provide information for consumers and clinicians on prevention and treatment of diseases and conditions, with an emphasis on reviews of clinical effectiveness research.
- **2012—Genetic Testing Registry (GTR)**—NCBI, in collaboration with NIH, launches the GTR to address the need for information about genetic tests for healthcare providers, researchers, patients, and others. The database provides information about the availability, validity, and usefulness of genetic tests.
- **2013—ClinVar**—NCBI creates the ClinVar database to aggregate information about sequence variation and its relationship to human health. The database includes submissions from outside research and testing groups as well as internal data drawn from such sources as dbSNP, dbVar, dbGap, and Gene Reviews.
- **2013—PubReader**—NCBI develops a new presentation style that optimizes reading of PMC articles through a browser on a desktop, laptop, or tablet computer.

An in depth account of the history of NCBI and NLM can be found in a published version of the Joseph Lieter NLM/MLA lecture presented at the annual meeting of the Medical Library Association in 2007 (1).

References

1. Smith KA. Laws, leaders, and legends of the modern National Library of Medicine. *J Med Libr Assoc.* 2008 Apr;96(2):121–33. PubMed PMID: 18379667.

Literature

NCBI Literature Resources

Ed Sequeira¹

Created: November 14, 2013.

Introduction

NCBI offers the following electronic services in the realm of literature resources. Figure 1, Overview of NCBI Literature Resources, depicts how they relate to one another.

The **NLM Catalog** database contains bibliographic records for journals, books, and other monographs and audiovisual materials. Many, but not all, of these items are in NLM's collection of traditional (print, film, etc.) and digital information resources. Note that the catalog contains just one record for a journal or a book, not individual records for every journal article or book chapter.

PubMed is a database of literature citations, primarily for articles from journals in the life sciences, but also for books and technical reports that are included in the NCBI Bookshelf. A large majority of the journal article records are curated by NLM subject experts who add appropriate MeSH terms (described below) to the PubMed records for better indexing and retrieval. This curated subset is also known as Medline. PubMed records provide links to full text in PMC and the Bookshelf as well as to thousands of external journal sites. PubMed records also have links to related data in dozens of biological databases offered by NCBI.

PubMed Central (PMC) is a repository of journal articles and a digital extension of NLM's permanent print collection. Everything in PMC is free to read, much of it from the time of publication, the rest after a delay that generally is 12 months or less. Almost all the articles in PMC have a corresponding citation in PubMed. The exceptions are a few types of material, such as book reviews, that PubMed does not cover.

Bookshelf is a companion to PMC for books (reference texts in the life sciences) and technical reports (clinical practice guidelines, health technology assessment reports, public health policy reports and similar material). User guides and technical documentation for NCBI's online services, such as the current document, the NCBI Handbook, are also part of the Bookshelf.

Functionally, all the material in the Bookshelf is part of the **LitArch** repository. LitArch also includes some document collections that are not accessible from the Bookshelf but are available through other NCBI services such as PubMed Health.

This section also discusses some complementary resources:

¹ NCBI.

MeSH is the Medical Subject Headings thesaurus. It is a controlled set of terms, organized in a subject hierarchy, that is used by NLM staff to categorize (index) the subject matter of journal articles and other material that NLM manages.

The **PubMed DTD** is a simple XML Document Type Definition (DTD) that publishers and other content providers use to provide NLM with journal article citations and abstracts for inclusion in PubMed. The PubMed DTD essentially is a DTD for article metadata, whereas the DTDs derived from JATS support the full-text of articles and other material that is deposited in PMC.

JATS is the Journal Article Tag Suite (JATS), officially identified as ANSI/NISO standard Z39.96-2012. JATS is a library of XML element and structure definitions from which one can create schemas for journal articles. It has its origins in a DTD developed at NCBI in 2001 to provide a common archival format for all content taken into PMC. Over the next few years, that DTD evolved into a tag suite and three journal article DTDs intended to be used variously for original content markup and as a standard, universal interchange format for exchanging data between parties who each use their own article modes, e.g., two publishers who have their own specific DTDs can convert to and from NLM DTD to exchange their content. As their use spread within the scientific publishing community, these DTDs came to be known as the NLM DTDs.

NIHMS is the NIH Manuscript Submission system. The NIH Public Access Policy (<http://publicaccess.nih.gov/>) requires all researchers who are supported by NIH to deposit in PMC the accepted manuscript of any peer-reviewed journal article that arises from their NIH-funded work. Some journals deposit the final published versions of such articles directly in PMC on behalf of their authors. In the remaining cases, the manuscript must be deposited in the NIHMS, where it is converted to a standard XML format before being transferred to PMC.

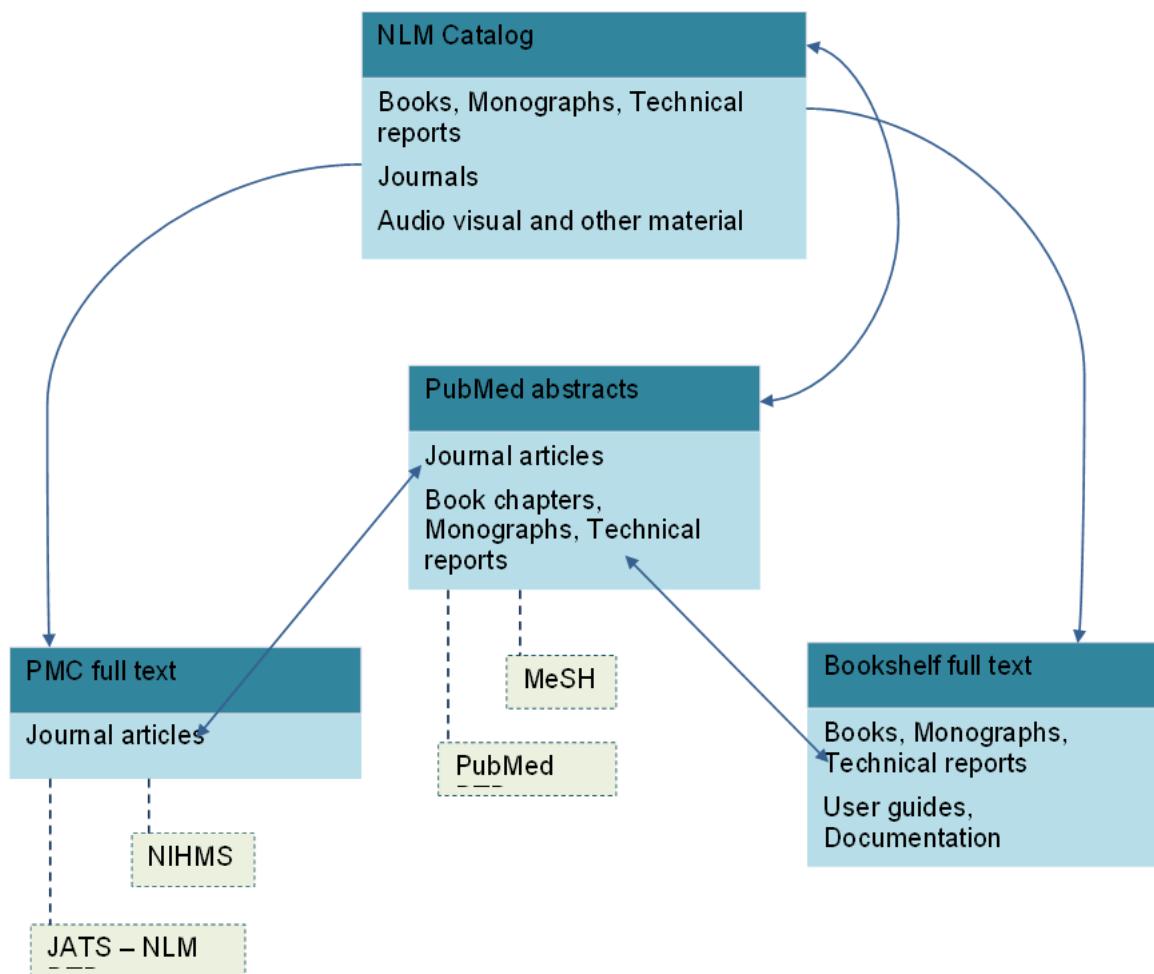


Figure 1. Overview of NCBI Literature Resources

NLM Catalog

Sarah Weis

Created: February 21, 2013; Updated: July 19, 2013.

History and Scope

The NLM Catalog provides access to National Library of Medicine (NLM) bibliographic data for over 1.4 million journals, books, audiovisuals, computer software, electronic resources, and other materials. All journals and books in the National Center for Biotechnology Information (NCBI) databases, including PubMed, PMC, and Books, have NLM Catalog records.

NCBI introduced the NLM Catalog in 2004 as an alternate search interface to the bibliographic records in LocatorPlus, the web-based public access catalog component of the Voyager Integrated Library System (ILS) used at the NLM. Detailed MEDLINE indexing information for journals in PubMed and other NCBI databases was housed in a separate Journals Database.

In 2010, the NLM Catalog and the Journals Database were merged, and the MEDLINE indexing information about the journals in PubMed and other NCBI databases was integrated into the NLM Catalog. At that time, the NLM Catalog was also redesigned to provide users with a streamlined interface and enhanced search and display options.

Data Sources

The NLM Catalog's primary data source is LocatorPlus, a component of the Voyager ILS used at the NLM. LocatorPlus contains over 1.4 million catalog records, holdings information for journals and other materials, links from catalog records to online resources, and circulation status information for materials at NLM. All NLM Catalog records contain a link to the item's LocatorPlus record.

Detailed MEDLINE indexing information for the nearly 29,000 journals in the NCBI databases is contained in the Serials Extract File (SEF), an internal product that serves as a bridge between various NLM applications.

NCBI integrates the data from LocatorPlus and the SEF to create a single, comprehensive NLM Catalog record for each item. Users may access an item's Serials XML and/or NLM Catalog XML using the "XML display" setting after running an NLM Catalog search. The E-utilities [ESummary](#) and [EFetch](#) retrieve only NLM Catalog XML.

Using the NLM Catalog

Basic Searching

Users can run a simple search from the NLM Catalog home page by entering one or more terms in the search box and clicking “Search”. Search terms are automatically ANDed together.

The NLM Catalog search features are similar to those available in PubMed, particularly when searching by journal title abbreviations and author names. Additional search indexes commonly used by searchers of bibliographic records are available for full author names, corporate authors, and other identifiers such as ISSN, ISBN, and the NLM ID.

Search Type	Example
Journal Title	n engl j med [ta] clinical scenarios in thoracic surgery [ti]
Author Name	remington js remington [au]
Full Author Name	david m oshinsky
Corporate Name	american medical association [cn]
ISSN	0028-4793
ISBN	0944235530 [other num]
NLM ID	0255562 [nlmid] 0255562

Phrase Searching

When a phrase is entered as the search term, it is checked against the MeSH Translation Table used in Automatic Term Mapping. If a match is found, the term is searched as a MeSH term and in all fields.

Example:
Search: breast cancer
Query Translation: "breast neoplasms"[MeSH Terms] OR ("breast"[All Fields] AND "neoplasms"[All Fields]) OR "breast neoplasms"[All Fields] OR ("breast"[All Fields] AND "cancer"[All Fields]) OR "breast cancer"[All Fields] OR breast cancer[All Fields]

If a phrase is not recognized, users can bypass Automatic Term Mapping by entering the phrase in double quotes or qualifying the phrase with a search tag.

Example
Search: heart beat

Table continues on next page...

Table continued from previous page.

Example
Query Translation: ("heart"[MeSH Terms] OR "heart"[All Fields] OR heart[All Fields]) AND beat[All Fields]
Search: "heart beat"
Query Translation: "heart beat"[All Fields]

Complex Searching

There are a variety of ways to search the NLM Catalog in a more sophisticated manner. It is possible to construct complex search strategies using the functions listed below.

- Combine search terms with the [Boolean operators](#) AND, OR, and NOT. The NLM Catalog processes searches in a left-to-right sequence. Use parentheses to “nest” concepts that should be processed as a unit and then incorporated into the overall search.
- Search terms may be qualified using [Search Field Descriptions and Tags](#). Each search term should be followed by the appropriate search field tag, which indicates the field to be searched, e.g., Spanish [la].
- Search for all terms that begin with a word by entering the word followed by an asterisk (*), the wildcard character, e.g., “flavor*.”
- Search by [MeSH term](#). To search a term only as a MeSH term qualify it using the search field tags, e.g., “[mh]” for MeSH Terms or “[majr]” for MeSH Major Topic. A qualified term is checked against the [MeSH Translation Table](#) and mapped to the appropriate MeSH term. To turn off mapping to multiple MeSH terms, enter the tagged MeSH term in double quotes.
- [Filters](#) narrow search results by journals referenced in the NCBI databases, journal subsets, language, publication type, material type, text availability, publication year, NLM collection, and search fields.
- Browse and select terms from the [search field indexes](#) using the [Advanced Search Builder](#). Select a search field from the menu. Enter a term in the search box, click “Show index list”, and select a term. Repeat as necessary and click “Search” when finished.
- [History](#) holds previous search strategies and results. The results can be combined to make new searches. Selecting “Add to history” in the Advanced Search Builder also enables users to preview the number of search results before displaying the records.
- Consult the sidebar discovery tool “Search details” to see how NLM Catalog translated a search.

Searching for journals in the NLM Catalog

To search for a journal in the NLM Catalog, click on [Journals in NCBI Databases](#) on the home page of the NLM Catalog. Enter a topic, journal title or abbreviation, or ISSN in the search box. Automatic suggestions will display as you type. When finished, click “Search”.

On the “Summary” display of search results, click the journal title for a specific record or select “Full” from the “Display Settings” menu to view additional information. The “Full” display contains all available fields, including indexing information.

Narrow a search to various journal subsets by using [Filters](#) or entering terms in the search box. See the table below for a selection of journal subsets and corresponding search terms.

Journal Subset	Enter in Search Box
Journals referenced in the NCBI databases	ncbijournals
Current Indexing Status	currentlyindexed notcurrentlyindexed
PubMed Central journals	journalspmc
Journals in electronic-only format	electronic only [current format status]
Version currently indexed	currentlyindexedprint currentlyindexedelectronic
Indexing subset	jsubsetaim – Currently indexed Core Clinical journals jsubsetd – Currently indexed Dental journals jsubsetim – Currently indexed Index Medicus journals jsubsetk – Currently indexed Consumer Health journals jsubsetn – Currently indexed Nursing journals
NLM Collection Only	nlm collection[call number]
Links to Full Text	all[sb] NOT none[URL]

Building a PubMed search for journals

To build a [PubMed search for journals](#) from the NLM Catalog, run a search and use the check boxes to select journals. Click the "Add to search builder" button in the PubMed Search Builder portlet, and the journal title abbreviation(s) will be sent to the search builder box. If a book or a non-PubMed journal is sent to the PubMed search builder, an error message warns the user that the PubMed search builder only retrieves citations for PubMed journals. The search builder will apply an OR Boolean operator if multiple journals are added to the search box. When finished adding journals, click “Search PubMed” to view the citations from the selected journal(s) in PubMed.

Results

NLM Catalog retrieves and displays search results in Summary format in publication date order. Records can be viewed in several other [formats](#) and [sorted](#) differently.

Saving and e-mailing search results

Search results can be e-mailed or [saved](#) in the Clipboard, a [My NCBI Collection](#), or as a text file.

Related Tools

The following resources are available to facilitate effective searches:

- [MeSH Database](#) allows searching of MeSH, NLM's controlled vocabulary. Users can find MeSH terms appropriate to a search strategy, obtain information about each term, and view the terms within their hierarchical structure.
- NLM assigns [Broad Subject Terms](#) to MEDLINE journals to describe the journal's overall scope. All of these subject terms are valid MeSH headings.
- A list of [all journals](#) included in PubMed is available via FTP.
- The [List of Serials Indexed for Online Users](#) provides bibliographic information for all journals whose articles were ever indexed with MeSH and cited in MEDLINE. It includes titles that ceased publication, changed titles, or are no longer indexed.
- Query the NCBI databases programmatically using [E-utilities](#).

PubMed: The Bibliographic Database

Kathi Canese and Sarah Weis

Created: October 9, 2002; Updated: March 20, 2013.

Summary

PubMed is a free resource developed and maintained by the National Center for Biotechnology Information (NCBI), a division of the U.S. National Library of Medicine (NLM), at the National Institutes of Health (NIH).

PubMed comprises over 22 million citations and abstracts for biomedical literature indexed in NLM's MEDLINE database, as well as from other life science journals and online books. PubMed citations and abstracts include the fields of biomedicine and health, and cover portions of the life sciences, behavioral sciences, chemical sciences, and bioengineering. PubMed also provides access to additional relevant websites and links to other NCBI resources, including its various molecular biology databases.

PubMed uses NCBI's Entrez search and retrieval system. PubMed does not include the full text of the journal article; however, the abstract display of PubMed citations may provide links to the full text from other sources, such as directly from a publisher's website or PubMed Central (PMC).

Data Sources

MEDLINE

The primary component of PubMed is MEDLINE, NLM's premier bibliographic database, which contains over 19 million references to journal articles in life sciences, with a concentration on biomedicine.

The majority of journals selected for MEDLINE are based on the recommendation of the Literature Selection Technical Review Committee (LSTRC), an NIH-chartered advisory committee of external experts analogous to the committees that review NIH grant applications. Some additional journals and newsletters are selected based on NLM-initiated reviews in areas that are special priorities for NLM or other NIH components (e.g., history of medicine, health services research, AIDS, toxicology and environmental health, molecular biology, and complementary medicine). These reviews generally also involve consultation with an array of NIH and outside experts or, in some cases, external organizations with which NLM has special collaborative arrangements.

Non-MEDLINE

In addition to MEDLINE citations, PubMed also contains:

- In-process citations, which provide a record for an article before it is indexed with NLM Medical Subject Headings (MeSH) and added to MEDLINE or converted to out-of-scope status.
- Citations that precede the date that a journal was selected for MEDLINE indexing.
- Some OLDMEDLINE citations that have not yet been updated with current vocabulary and converted to MEDLINE status.
- Citations to articles that are out-of-scope (e.g., covering plate tectonics or astrophysics) from certain MEDLINE journals, primarily general science and general chemistry journals, for which the life sciences articles are indexed with MeSH for MEDLINE.
- Citations to some additional life science journals that submit full-text articles to PubMed Central and receive a qualitative review by NLM.

Journal Selection Criteria

Journals that are included in MEDLINE are subject to a selection process. The [Fact Sheet on Journal Selection for Index Medicus®/MEDLINE®](#) describes the journal selection policy, criteria, and procedures for data submission.

History

PubMed was first released in January 1996 as an experimental database under the Entrez retrieval system with full access to MEDLINE. The word "experimental" was dropped from the website in April 1997, and on June 26, 1997, free MEDLINE access via PubMed was announced at a Capitol Hill press conference. Use of PubMed has grown exponentially since its introduction: PubMed searches numbered approximately 2 million for the month of June 1997, while current usage typically exceeds 3.5 million searches per day.

PubMed was significantly redesigned in 2000 to integrate new features such as LinkOut, Limits, History, and Clipboard. PubMed began linking to PubMed Central full-text articles and the Bookshelf's initial book, *Molecular Biology of the Cell*. The Entrez Programming Utilities, E-Utilities, and the Cubby (My NCBI subsequently replaced the Cubby) also were released.

In 2002, the PubMed database programming was completely redesigned to work directly from XML files, and two new NCBI databases, Journals (now the NLM Catalog) and MeSH, were created to provide additional search capabilities for PubMed.

Electronic Data Submission

Publishers of journals indexed for MEDLINE are encouraged to submit citation and abstract data electronically for inclusion in PubMed. Electronic submissions ensure that citations and abstracts are available to the public within 48 hours of uploading a properly formatted XML file. See Figure 1 for information about the PubMed data flow.

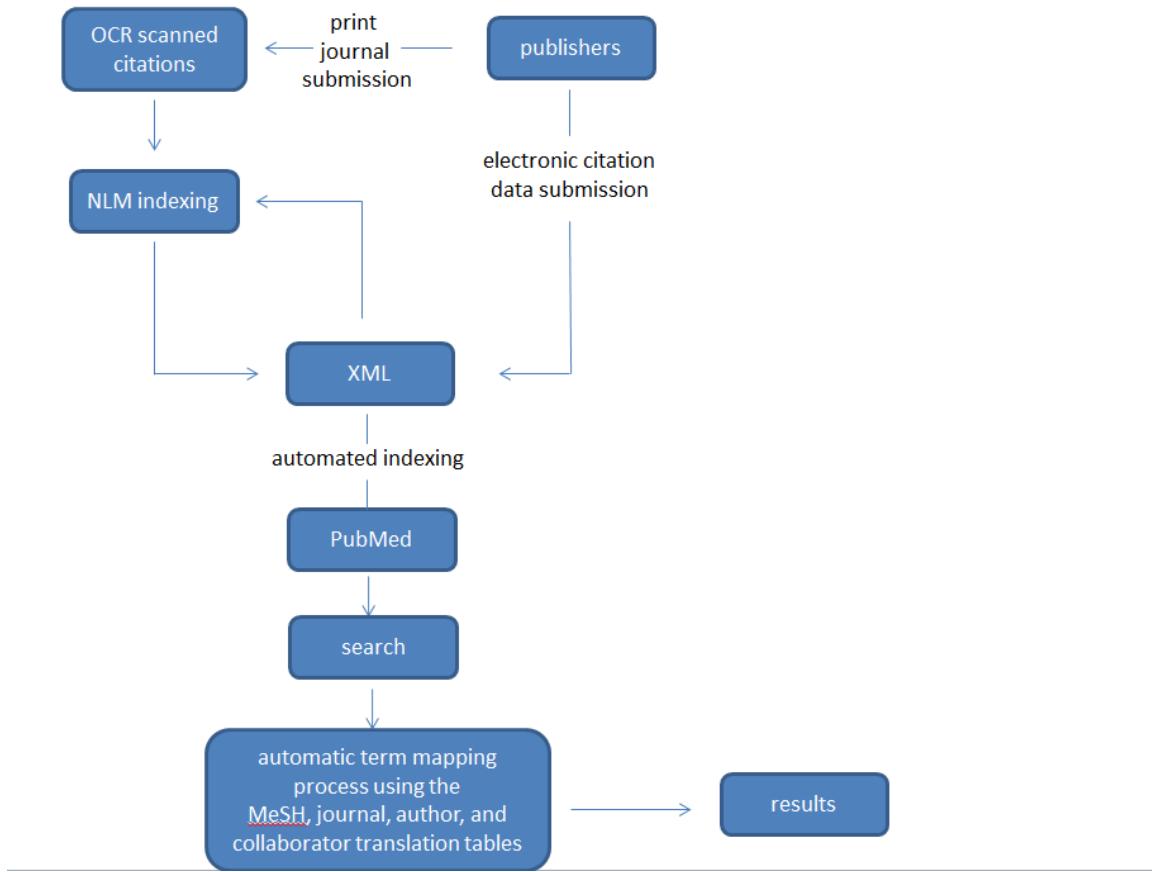


Figure 1. A schematic representation of PubMed data flow

NCBI works with many publishers and commercial data providers to prepare and submit electronic citation and abstract data. In 2012, over 90% of the citations added to PubMed were submitted electronically.

Electronic citation data submission also fulfills one of the requirements to add an icon to PubMed citations that links to full text of the articles available on the journal website. Linking can be achieved using LinkOut, which is the feature that generates a direct link from a PubMed citation to the journal website.

NLM manually creates citations for journals uninterested or unable to submit electronic citation data by scanning the print copy of the journal and using optical character recognition (OCR). This process can take significantly longer than electronic submissions due to the manual nature of the process.

Electronic Data Submission Process

Electronic citation and abstract data are submitted via File Transfer Protocol (FTP) in XML according to the PubMed Document Type Definition (DTD). Details about the

PubMed XML tagged format, [XML tag descriptions](#), [sample XML files](#), and how to handle special characters are available in the online [documentation](#).

NCBI staff guide new data providers through the approval process for file submission. New data providers are asked to submit a [sample XML file](#), which is reviewed for XML formatting and syntax and for bibliographic accuracy and completeness. The file is revised and resubmitted until all criteria are met. Once approved, a private account is set up on the NCBI FTP site to receive XML citation data for future journal issues.

NCBI loads publisher-supplied data daily, Monday through Friday, at approximately 8:30 a.m. Eastern Time. New citations are assigned a PubMed ID (PMID), a confirmation report is sent to the data provider, and the citations become available in PubMed after 6:00 a.m. the next day, Tuesday through Saturday. Citation data submitted Friday after 8:30 a.m. through Sunday will be available in PubMed after 6:00 a.m. the following Tuesday.

After posting in PubMed, the citations are sent to NLM's Index Section for bibliographic data verification and the addition of MeSH terms from NLM's controlled vocabulary. The indexing process can take up to a few months, after which time the completed citations flow back into PubMed, replacing the original data.

Publishers or others interested in submitting electronic data may view the [XML Data Provider Help](#) or write to publisher@ncbi.nlm.nih.gov for more information.

Database Management and Hardware

PubMed uses its own proprietary Web-based search engine. The Web server software is the open-source Apache HTTP server.

PubMed runs on approximately 62 standard Linux servers, each with two quad core 2.6-3.6 GHz Intel Nehalem CPUs, 48-64 GB of memory, 1TB of local storage, and a Gigabit Ethernet connection. Fourteen NCBI Portal servers render the PubMed Web interface and eight servers search the PubMed index. PubMed records are retrieved from four proprietary XML article servers, while the "Related Articles" service relies on 20 servers running a specialized non-relational database system. The remaining servers support services such as links to related information in other NCBI databases (e.g., PubMed Central, Nucleotide, and the Sequence Read Archive), LinkOut, History, and My NCBI.

To accommodate the volume of data output by PubMed and other Web-based services, the NLM has a 3-Gbps connection to the commercial Internet as well as a 20-Gbps connection to Internet2, the non-commercial network used by many leading research universities at the NIH campus, and similar connectivity at a second redundant data center.

Indexing

The Automatic Indexing Process

The indexing process automatically generates access points for each field of a PubMed citation.

- Terms within a citation are initially extracted (stopwords are ignored) and matched against a list of useful phrases.
- Individual terms are added to the corresponding field, e.g., title words are added to the ***title field*** index and the ***title/abstract field*** index.
- All terms are also added to the ***all fields*** index (except for the terms found in the ***place of publication*** (Country) and ***transliterated title*** fields).
- Some fields use a special set of rules for extracting data:
 - Several index points are created for authors, e.g., indexes for the author Harold Varmus will include Varmus H; Varmus, Harold; and Varmus.
 - Indexes for MeSH terms include the ***MeSH term*** and ***MeSH major term*** fields, subheading fields, etc.

Field indexes may be browsed using the PubMed [Advanced Search Builder](#) “Show index list” feature.

How PubMed Queries Are Processed

Automatic Term Mapping

Untagged terms that are entered in the search box are matched against the following translation tables and indexes in this order:

1. a MeSH (Medical Subject Headings) translation table,
2. a journals translation table,
3. the full author translation table,
4. author index,
5. the full investigator (collaborator) translation table, and
6. an investigator (collaborator) index.

On the right side of a search results page there are a number of tools designed to enhance user discovery, including one called “**Search details**” that shows the search term translations.

When a match is found for a term or phrase in a translation table the mapping process is complete and does not continue on to the next translation table.

1. MeSH Translation Table

The MeSH Translation Table contains:

- MeSH Terms

- See-reference mappings (also known as entry terms) for MeSH terms
- MeSH Subheadings
- Publication Types
- Pharmacologic Actions
- Terms derived from the Unified Medical Language System (UMLS) that have equivalent synonyms or lexical variants in English
- Supplementary Concepts (chemical, protocol or disease terms) and their synonyms

If a match is found in the MeSH translation table, the term will be searched as MeSH (that includes the MeSH term and any specific terms indented under that term in the MeSH hierarchy), and in all fields.

For example, if you enter “multiple sclerosis” in the search box, PubMed will translate this search to:

“multiple sclerosis”[MeSH Terms] OR (“multiple”[All Fields] AND “sclerosis”[All Fields]) OR “multiple sclerosis”[All Fields]

If you enter a MeSH Term that is also a Pharmacologic Action, PubMed will search the term as [MeSH Terms], [Pharmacologic Action], and [All Fields].

If you enter a synonym for a MeSH term, the translation will also include an all fields search for the MeSH term associated with the synonym.

Search term: ear infection

“ear infection” is an synonym for the MeSH term “otitis” in the MeSH translation table.

Search translated to: “otitis”[MeSH Terms] OR “otitis”[All Fields] OR (“ear”[All Fields] AND “infection”[All Fields]) OR “ear infection”[All Fields]

When a term is searched as a MeSH term, PubMed automatically searches that term plus the more specific terms underneath in the [MeSH hierarchy](#):

Search term: breast cancer

“Breast cancer” is an entry term for the MeSH term “breast neoplasms” in the MeSH translation table.

“Breast neoplasms” includes the specific headings below, all of which are also searched:

Breast Neoplasms, Male

Carcinoma, Ductal, Breast

Heredity Breast and Ovarian Cancer Syndrome

Inflammatory Breast Neoplasms

2. Journals Translation Table

If the search term(s) is not found in the MeSH translation table, the process continues on to look for a match in the journals translation table, which contains full journal title, journal title abbreviation, and International Standard Serial Numbers (ISSNs). Search term(s) automatically map to the journal abbreviation:

Search term: New England Journal of Medicine

“New England Journal of Medicine” maps to N Engl J Med.

Search translated to: “N Engl J Med” [Journal Name]

Journal titles are included in the all fields index; therefore, a search for a MeSH term that is also a journal title will retrieve citations for the journal as well:

Search term: nature

Search translated as: "nature"[MeSH Terms] OR "nature"[All Fields]

The search will include the journal Nature

3. Full Author Index

The full author translation table includes full author names for articles published from 2002 forward, if available.

4. Full Investigator (Collaborator) index

If the term is not found in the above tables, except for full author, and is not a single term, the full investigator index is consulted for a match. The full investigator (collaborator) translation table includes full names, if available.

5. Author Index

If the term is not found in the above tables, except for full author or full investigator, and is not a single term, PubMed checks the author index for a match.

An author name search should be entered in the form: last name (space) initials, e.g., o'malley f, smith jp, or gomez-sanchez m.

If only one initial is used, PubMed retrieves all names with that first initial, and if only an author's last name is entered, PubMed will search that name in All Fields. It will not default to the author index because an initial does not follow the last name:

Search term: o'malley f

Search retrieves authors: o'malley fa, o'malley fb, o'malley fc, o'malley fd, o'malley jr, etc.

Search term: o'malley

Search translated as: “o'malley” [All Fields]

A history of the NLM's author indexing policy regarding the number of authors to include in a citation is outlined in Table 1.

Table 1. History of NLM author-indexing policy

Dates	Policy
1966-1984	MEDLINE did not limit the number of authors.
1984-1995	NLM limited the number of authors to 10, with “et al.” as the eleventh occurrence.
1996-1999	NLM increased the limit from 10 to 25. If there were more than 25 authors, the first 24 were listed, the last author was used as the 25th, and the twenty-sixth and beyond became “et al.”
2000-present	MEDLINE does not limit the number of authors.

6. Investigator (Collaborator) index

If the term is not found in the above tables, except for full author, author, or full investigator, and is not a single term, PubMed checks the investigator index for a match.

7. If no match is found?

PubMed breaks apart the phrase and repeats the above automatic term mapping process until a match is found. PubMed ignores [stopwords](#) in searches.

If there is no match, the individual terms will be combined (AND-ed) together and searched in all fields (see Simple Searching section below).

One exception: PubMed interprets a sequence of numbers, e.g., 23436005 23193264, as PubMed IDs, and the IDs will be OR-ed individually rather than combined (AND-ed).

Search Rules and Field Abbreviations

It is possible to override PubMed's automatic term mapping by using search rules, syntax, and specific search field tags.

The Boolean operators AND, OR, and NOT should be entered in uppercase letters and are processed left to right. Nesting of search terms is possible by enclosing concepts in parentheses. The terms inside the set of parentheses will be processed as a unit and then incorporated into the overall strategy, e.g., therapy AND (hay fever OR asthma).

Terms may be qualified using PubMed's [Search Field Descriptions and Tags](#). Each search term should be followed by the appropriate search field tag, which indicates which field will be searched. For example, the search term cell [jour] will only search the *journal field*. Specifying the field precludes the automatic term mapping process that would result in using the translation tables, e.g., MeSH.

Using PubMed

Searching

Simple Searching

A simple search can be conducted from the [PubMed](#) homepage by entering terms in the search box and clicking the **Search** button or pressing the Enter key.

Term suggestions will display for search terms entered in the search box.

If more than one term is entered in the search box, PubMed will go through the automatic term mapping process described in the previous section, looking for exact matches for each term. If the exact phrase is not found, PubMed clips a term off the end and repeats the automatic term mapping, again looking for an exact match, but this time to the abbreviated query. This continues until none of the words are found in any one of the translation tables. In this case, PubMed combines terms (with the AND Boolean operator) and applies the automatic term mapping process to each individual word. PubMed ignores [stopwords](#), such as "about," "of," or "what." Users may also apply their own Boolean operators (AND, OR, NOT) to multiple search terms; the Boolean operators must be in uppercase.

If a phrase of more than two terms is not found in any translation table, then the last word of the phrase is dropped, and the remainder of the phrase is sent through the entire process again. This continues, removing one word at a time, until a match is found.

If there is no match found during the automatic term mapping process, the individual terms will be combined with AND and searched in all fields:

Search term: heart attack bad diet

Automatic term mapping process:

heart attack bad diet => no matches

heart attack bad => no matches

heart attack => MeSH translation table match, remove "heart attack" from the query

bad diet => no matches

bad => no matches, search in all fields, remove "bad" from the query

diet => MeSH translation table match, remove "diet" from the query

Processing stops because the query string is empty

Translated as: ("myocardial infarction"[MeSH Terms] OR ("myocardial"[All Fields] AND "infarction"[All Fields]) OR "myocardial infarction"[All Fields] OR ("heart"[All Fields] AND "attack"[All Fields]) OR "heart attack"[All Fields]) AND bad[All Fields] AND ("diet"[MeSH Terms] OR "diet"[All Fields])

Consult the sidebar discovery tool “**Search details**” to see how PubMed translated a search.

Complex Searching

There are a variety of ways that PubMed can be searched in a more sophisticated manner than simply typing search terms into the search box and clicking Search. It is possible to construct complex search strategies using Boolean operators and the features listed below:

- Use the [Advanced](#) search page to:
 - Search by a [specific field](#)
 - Browse the [index](#) of terms
 - [Combine searches](#) using history
 - [Preview](#) the number of search results
- [Filters](#) narrow search results by article types, text availability, publication dates, species, languages, sex, subjects, journal categories, ages, and search fields.

Additional PubMed Features

The following resources are available to facilitate effective searches:

- Use the [MeSH Database](#) to find MeSH terms, including Subheadings, Publication Types, Supplementary Concepts, and Pharmacological Actions, and then build a PubMed search
- The [Clinical Queries](#) page provides searching by clinical study categories that use built-in search filters to limit retrieval to citations to articles reporting research conducted with specific methodologies, including those that report applied clinical research.
- The [NLM Catalog](#) includes information about the journals in PubMed and the other NCBI databases.
- Use the [Batch Citation Matcher](#) to retrieve PMIDs (PubMed IDs) for multiple citations in batch mode.
- [My NCBI](#) saves searches, results, bibliographies, and features an option to automatically update and email search results. Preferences include storing and changing an email address, highlighting search terms, opening the abstract display

supplemental data by default, and turning off the auto suggest feature. Additional features include filtering search results, managing recent activity, and setting a LinkOut icon, document delivery services, and outside tool preferences.

- [PubMed Mobile](#) provides a simplified mobile-friendly web interface to access PubMed.

Results

PubMed search results are displayed in a summary format; citations are initially displayed 20 items per page with the most recently entered citations displayed first. (Note that this date can differ significantly from the publication date.)

A spell-checking feature suggests alternative spellings for search terms that may include misspellings.

To provide users with targeted results, query sensors may display for searches that match specific characteristics. For example, a citation sensor will display for searches that include author names, journal titles, publication dates, or article titles, e.g., zong science 2012. A gene sensor checks for queries that include gene symbols, e.g., brca1

Depending on the search, additional sensors and discovery tools, e.g., **Results by year**, **Recent activity**, **Related searches**, **PMC Images**, may also display.

Citations can be viewed in other [formats](#) and can be [sorted](#), [saved](#) or [e-mailed](#), and [printed](#). The [full text](#) may also be available online or ordered from a library.

How to Create Hyperlinks to PubMed

To create Web URL links that search and retrieve PubMed data the following tools are useful:

- The [Entrez Programming Utilities](#) (E-utilities), which are a set of eight server-side programs that provide a stable interface into the Entrez query and database system at NCBI. E-Utilities provide a fast, efficient way to search and download citation data without using the front-end query engine.
- Generate the URL manually using the [Creating a Web Link to PubMed](#) online documentation.

Customer Support

If you need more assistance:

- Contact the Help Desk by selecting the [Support Center](#) link displayed on all PubMed pages
- Call the NLM Customer service desk: 1-888-FIND-NLM (1-888-346-3656)

PubMed Central

Chris Maloney,¹ Ed Sequeira,¹ Christopher Kelly,¹ Rebecca Orris,¹ and Jeffrey Beck¹

Created: November 14, 2013; Updated: December 5, 2013.

Overview of PMC

PubMed Central (PMC) is NLM's digital archive of medical and life sciences journal articles and an extension of NLM's permanent print collection. It was launched in early 2000 with a single issue each of two journals, and has grown steadily since. As of June 2013, it contained over 2.7 million articles; more than 1200 journals were depositing all their published content in PMC, and a few thousand other journals were depositing selected articles. (See <http://www.ncbi.nlm.nih.gov/pmc/> for current counts and titles.) Almost all the articles in PMC have a corresponding citation in PubMed. The exceptions are a few types of material, such as book reviews, that PubMed does not cover.

In its early years, PMC received all its content from journals that deposited complete issues. In 2006, NLM began offering publishers the additional option to deposit just selected articles from an issue. In both cases, the publisher provides PMC with the final published versions of articles; deposits are covered by formal participation agreements that address copyright and other rights and responsibilities. Participating publishers must deposit full-text XML and PDFs, along with high resolution image files and any supplementary data that is published with an article. Details about these participation agreements are available on the Publisher Information webpage (1).

Although publishers began providing material to PMC just a few months before its launch in 2000, the archive contains a substantial number of articles that were published long before that. Publishers often include several years of back files when they begin participating in PMC. In addition, in 2002, NLM undertook a project to scan and digitize the complete print archives of journals that were depositing current content in PMC. Two years later, the Wellcome Trust and the Joint Information Systems Committee in the UK joined in supporting the effort. The project ran for about 6 years and resulted in the addition of more than 1.2 million articles to PMC, dating back to the early 1800s.

Article Data Formats

Newly published articles, and much of the material in PMC that has been published since the late 1990s, are archived as full-text XML. For the articles from the print issue digitization project, most of which were published before 2000, PMC has PDFs of scanned page images, along with automatically extracted OCR text that is used to support full-text

¹ NCBI.

searching, and abstracts in XML form (e.g., <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC361653/>). For some journals, there may also be a relatively short period when they were making the transition from print to electronic publication. Before they moved to creating full-text XML (or SGML) they had electronically formatted PDFs. For this period (generally the mid-1990s, but extending to the mid-2000s for some journals) PMC has a PDF and an XML abstract for each article (e.g., <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC148429/>).

Full-text XML is at the heart of PMC's design philosophy and therefore is a requirement for all current content. XML files are machine- and human-readable and are not technology dependent. This makes XML easy to migrate as technology changes and, therefore, an excellent archival format. Regardless of the source DTD to which incoming content is structured, all full-text XML in PMC is converted to a common archival format, the "NLM DTD," which is now a NISO standard (see "NLM DTD to NISO JATS Z39.96-2012"). Full-text HTML pages (e.g., <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3382486/?report=classic>) are created dynamically from the XML at retrieval time. This allows article presentation styles to be changed relatively quickly and easily. Even a completely new format like the PubReader display (e.g., <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3382486/?report=reader>) can be introduced without any changes to, or preprocessing of, the source article record in the PMC database.

Access and Copyright

Everything in PMC is free to read, much of it from the time of publication, the rest after a delay that generally is 12 months or less. Participating journals are expected to deposit their articles at time of publication even if they are not made available publicly in PMC immediately. A growing number of articles in PMC are available under a Creative Commons (2) or similar license that generally allows more liberal redistribution and reuse than a traditional copyrighted work. However, the majority of the material in PMC is still protected by standard copyright, held by the respective publishers. NLM does not hold copyright on any of the material.

Author manuscripts

Since 2005, PMC also has been the designated repository for NIH's Public Access Policy (3) and similar policies of other research funders in the US and abroad. Researchers supported by these agencies are required to deposit in PMC the accepted manuscript of any peer-reviewed journal article that arises from the funding they have received. Some journals deposit the final published versions of such articles directly in PMC on behalf of their authors under one of the PMC participation agreements mentioned above. In the remaining cases, the author or publisher deposits manuscript files (e.g., a Word document or a minimally formatted PDF that has not yet gone through final copy editing by the journal) in a Manuscript Submission system—the NIHMS in the US or similar, derivative

systems in the UK and Canada. The manuscript is converted to JATS XML format and reviewed and approved by the author before being transferred to PMC.

PMC International

NLM supports the operation of journal archives similar to PMC in the UK and Canada. These two archives have a copy of the majority of the articles in PMC, based on agreements with the respective publishers. Information about the PMC international collaboration is available on the PMC International webpage (4).

Architecture Overview

The PMC processing model is diagrammed in Figure 1. For each article, we receive a set of files that includes the text in SGML or XML, the highest resolution figures available, a PDF file if one has been created for the article, and any supplementary material or supporting data. The text is converted to the current version of the NISO Z39.96-2012 Journal Article Tag Suite (JATS) Archiving and Interchange article model, and the images are converted to a web-friendly format. The source SGML or XML, original images, supplementary data files, PDFs, and NLM XML files are stored in the archive. Articles are rendered online using the NLM XML, PDFs, supplementary data files, and the web-friendly images.

Ingest

Participating PMC publishers submit the full text of each article in some "reasonable" SGML or XML format along with the highest-resolution images available, PDF files (if available), and all supplementary material. Complete details on the PMC's file requirements are available (5).

A reasonable SGML or XML format is one where there is sufficient granularity in the source model to map those elements critical to the understanding of the article (and/or its functioning in the PMC system) from the original article to the appropriate place in the PMC XML model. Currently we convert all incoming articles into the NISO Z39.96-2012 Journal Article Tag Suite (JATS) version 1.0 Archiving and Interchange article model, but we have articles in nearly every version of the NLM DTD/NISO JATS article model in the PMC database.

The Journal Evaluation Process

Journals joining PMC must pass two tests. First, the content must be approved for the NLM collection (6).

Next the journal must go through a technical evaluation to "be sure that the journal can routinely supply files of sufficient quality to generate complete and accurate articles online without the need for human action to correct errors or omissions in the data." (1)

PMC Archive Model

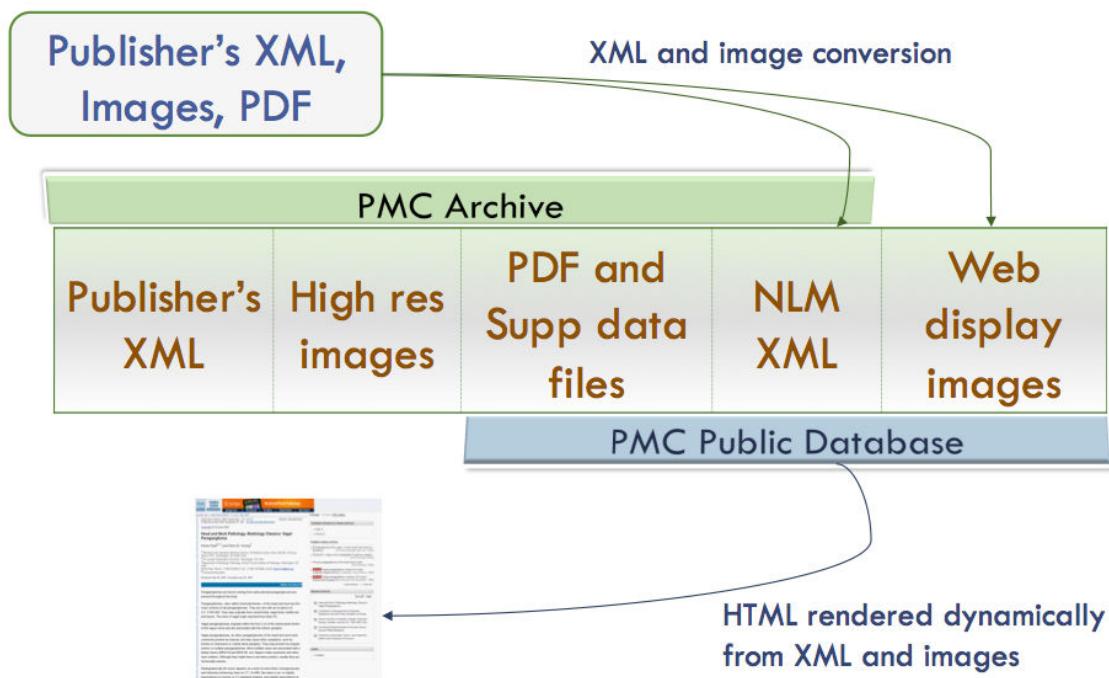


Figure 1. PMC Processing Model.

For the technical evaluation, a journal supplies a sample set of articles. These articles are put through a series of automated and human checks to ensure that the XML is valid and that it accurately represents the article content. There is a set of "Minimum Data Requirements" that must be met before the evaluation proceeds to the more human-intense content accuracy checking (7). These minimum criteria are listed below briefly:

- Each sample package must be complete: all required data files (XML/SGML, PDF if available, image files, supplementary data files) for every article in the package must be present and named correctly.
- All XML files must conform to an acceptable journal article schema.
- All XML/SGML files must be valid according to their schema.
- Regardless of the XML/SGML schema used, the following metadata information must be present and tagged with correct values in every sample file:
 1. Journal ISSN or other unique Journal ID
 2. Journal Publisher
 3. Copyright statement (if applicable)
 4. License statement (if applicable)
 5. Volume number
 6. Issue number (if applicable)

7. Pagination/article sequence number
 8. Issue-based or Article-based publication dates. Articles submitted to PMC must contain publication dates that accurately reflect the journal's publication model.
- All image files for figures must be legible, and submitted in high-resolution TIFF or EPS format, according to the PMC Image File Requirements.

These seem like simple and obvious things—XML files must be valid—but the minimum data requirements have greatly reduced the amount of rework that the PMC Data Evaluation group has to do. It helps to be explicit about even the most obvious of things.

PMC's XML Philosophy

PMC's XML philosophy is a balance between strictness and flexibility that enables control of the quality of data being loaded into the system without being too restrictive on submitters.

PMC does a complete review of any new schema in which content is being submitted, as described above. We do not take articles in HTML. We also do a complete review of sample articles for each new journal to be sure that the content provider is able to provide content that is structurally and semantically correct.

Another thing we are strict about is that all content must be valid according to the schema in which it was submitted—not just during data evaluation but in the ongoing production process as well. This seems obvious, but there was a surprising amount of controversy about this in the early days of PMC, and we still receive invalid files. Problems usually arise now because the submitter has made a schema change (as simple as adding a new character entity to the DTD or a new required element) without telling us or sending an updated schema.

Also, we do not fix text; all content changes must be made by the submitter, and the content must be resubmitted.

Some things we are more flexible about, which reduces some of the burden on our submitters. First, we don't require all content to be in our format or to follow our tagging rules. We don't force updates of content to the latest DTD version, and we can generally follow journal style where it does not interfere with processing.

PMC Internal DTD

We use the JATS Archiving and Interchange DTD ("out of the box") as the format for all articles loaded to the PMC database. This model was created specifically for archiving article content. It was designed to be an "easy target to hit" when transforming content from the over 40 different input models that we receive content in. Currently we are writing content into version NISO JATS version 1.0.

We do not migrate all content to each new version of the JATS DTD when one is released. The system is robust enough to handle content from versions 1.0 through 3.0 of the NLM DTD and version 1.0 of the NISO JATS DTD, so we are not constantly churning the data.

All of the versions of the DTD are managed with an XML Catalog (8), which we also use to manage all of the input DTDs (SGML and XML). We maintain all mappings of PUBLIC and SYSTEM IDs for any DTD that we use in the XML catalog on our Linux machines and then create other catalogs from it each time it is updated. We create an SGML Catalog for the SGML tools that we use; a single "Oxygen" catalog that everyone on the team can use over the network with the XML editor; and a copy of the catalog that refers to http-based copies of the DTDs for PMC International sites. The XML Catalog is an essential piece of the PMC system.

PMC Tagging Style

Next, we've defined a set of rules for objects within articles that is more restrictive than the DTD. This allows us to have normalized structures (figures, tables, contributors) in articles for ease of processing and rendering. We call these rules the PMC Tagging Style, and all articles must "pass style" before being loaded to the database. They are documented in the PMC Tagging Guidelines (9).

(Re)Usability of XML

Finally, our XML must be useable by others. The NLM XML that we create from whatever was submitted to us is always available to the submitting publisher (the content owner), and a subset of the articles that are Open Access are available to anyone for download through the PMC Open Archives Service (10). This keeps us honest. We can't allow ourselves to take shortcuts with the data. All articles must be valid according to the JATS schema version that they reference, and we only use Processing Instructions for instructions about processing.

Text Processing

There are four main principles to PMC text processing:

First, we expect to receive marked-up versions of an article that are well-formed, valid, and accurately represent the article as it was published, (i.e., that it represents the version of record). The question of what is the "Version of Record" is left up to the publisher. It may be a printed copy, a PDF version, or the journal's website.

We do not correct articles or files. That is, we will not fix something that is wrong in the version of record nor will we make a correction to an XML file. All problems found in either processing or QA of files are reported to the publisher to be corrected and resubmitted.

The goal of PMC is to represent the content of the article and not the formatting of the printed page, the PDF, or the journal's website.

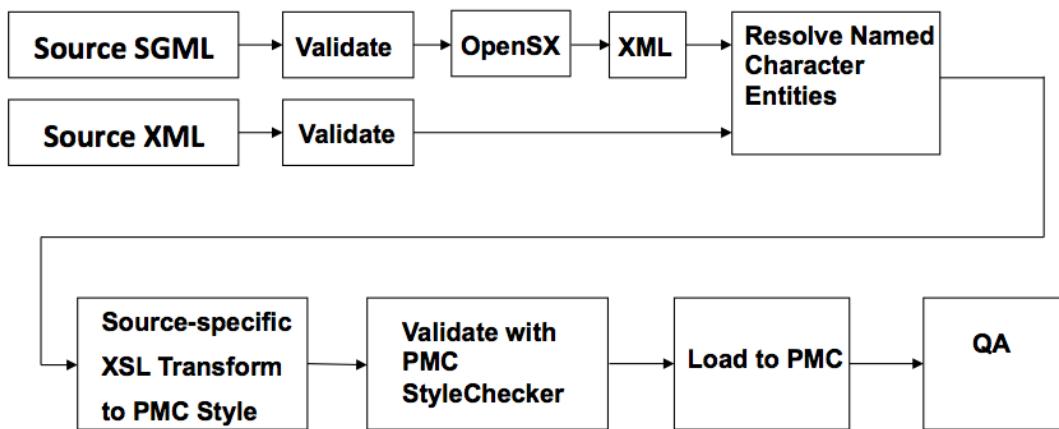


Figure 2. PMC Text Processing.

Finally, we run a Quality Assessment on content coming into PMC to ensure that the content rendered in PMC accurately reflects the article as it was published. Our QA is a combination of automated checks and manual checking of articles. To help ensure that the content we are spending time ingesting to PMC is likely to be worthwhile, journals must pass through an evaluation process before they can send content to PMC in a regular production workflow.

Figure 2 shows the workflow for text coming into PMC.

Images

Generally, authors submit raw image data files to a publishing house in various formats (PPT, PDF, TIF, JPG, XML, etc.). The files are then normalized to produce print or electronic output. PMC requires the normalized output, which is high-resolution, and of sufficient width and quality to be considered archival. Images generated at low resolution for display purposes are not acceptable.

During ingest, the submitted images are converted to web-friendly thumbnail (Figure 3) and full-sized (Figure 4) versions for display within the article.

The thumbnail from "Plate 1" links to a full view of the figure including the caption.

A very large version of the image is also created so that users can zoom in and inspect the image up close. Linking from the full image view takes you to the Tileshop view (Figure 5). The image index in the lower right corner shows which part of the whole image is available on the screen.

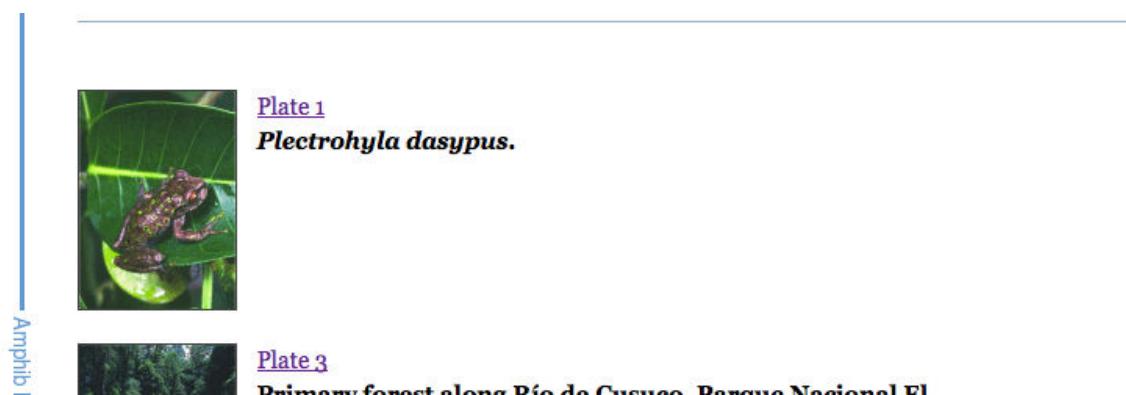


Figure 3. An image thumbnail.

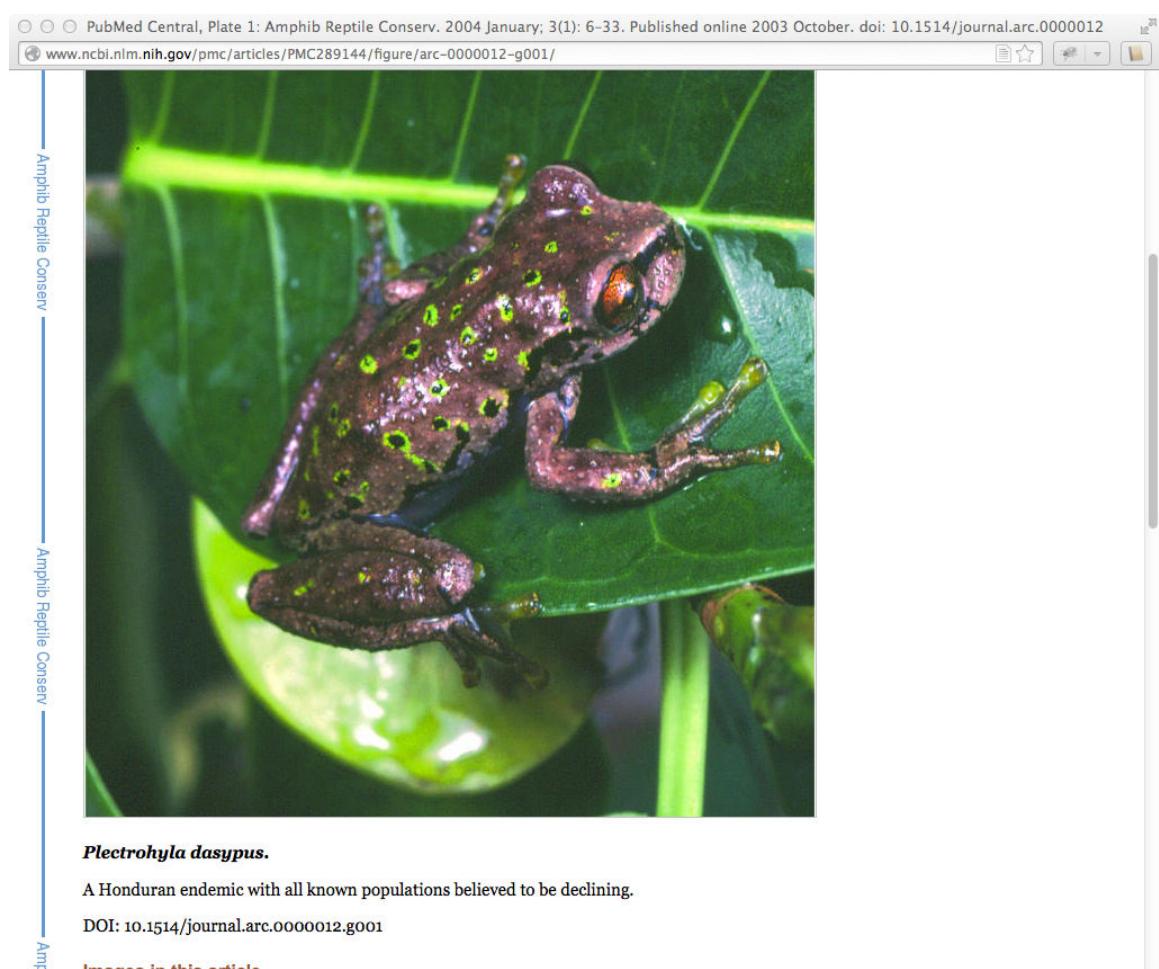


Figure 4. Full sized image displayed in a new window.

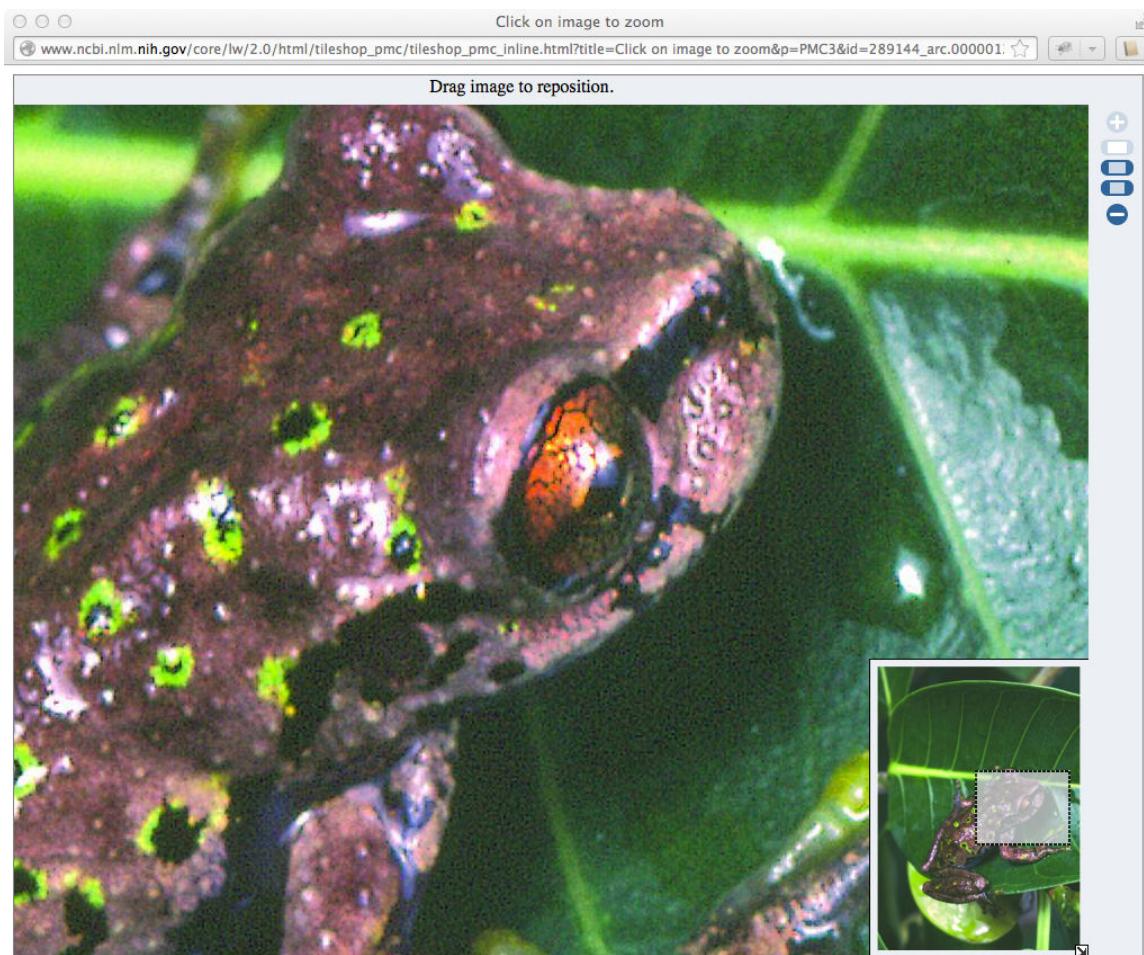


Figure 5. Very large representation of the image that allows zooming.

With the original high-resolution images stored in the archive, when these display technologies become out of date, the images can be generated in whatever the latest image display technology is available.

PDFs and Supplemental Data

PDF versions of the article may be supplied to accompany the XML version in PMC, but they are not required.

PMC requires all available supplementary material to be submitted in a portable format, such as PDF, DOC, CSV, etc. Supplementary material should not be externally linked to a www location from the article text as a substitute for submission. Supplementary material includes the following:

- Voluminous material that was used to support the conclusions of the narrative, such as a genomic database or the multiple data sets for an article that presents the highlights, which can never accompany a paper based on sheer mass.

- "Extra" tables that do not display with the work, but that record the measurements on which the article is based, for example, that need to be available so the peer reviewers can check the article.
- Material added to the work for enhancement purposes, such as a quiz, an instructional video, the 3-minute version of the reaction that was described in the work with narrative and a few still images, a form that can be filled out, etc.

Quality Assessment and Workflows

Quality Assessment (QA) is done for all content coming into PMC to ensure that the content in PMC accurately reflects the article as it was published. Our QA is a combination of automated and manual checks and is managed by a team of Journal Managers (JM's) who are each assigned responsibility for a large number of journals. JM's are also responsible for ensuring that content is deposited on schedule, passes successfully through the automated workflow, and is released to the live site in a timely manner. JM's interact regularly with the publishers and content providers to resolve problems and answer questions.

For journals in our regular production workflow, an automated workflow is set up so that new content uploaded to our FTP site for the journal is picked up and processed automatically, usually within several hours of the upload. A notification email is sent to the responsible JM indicating whether the session succeeded or failed. If the session succeeds, the content has been successfully ingested and processed and an entry will be added into our QA system. If the session resulted in an "error," the logs will be reviewed by the responsible JM and resolved often after updated files are sent from the publisher or content provider. For this automated system to work, it is important the publishers and content providers adhere to the File Submission Specifications (5) and follow a consistent naming scheme. Submissions that don't adhere to a consistent naming scheme will remain on the FTP site and must be reviewed by the JM before they can proceed through the automated workflow.

Automated QA checks that occur as part of this workflow include checking that the XML/SGML is valid according to its schema, that all images and supplementary files referred to in the files are present and properly named, and that volume and issue information in the filename of the submission package (typically a ZIP file) correspond correctly to the volume and issue tagging in the XML/SGML files. All content that is not well-formed (if XML) or valid is returned to the provider to be corrected and resubmitted. Furthermore, the PMC Style Checker (9) is used during the automated workflow processing to ensure that all content flowing into PMC is in the PMC common XML format for loading to the database. The errors reported by the Style Checker provide us with a level of automated checking on the content itself that can highlight problems, but it only goes so far. For example, the Style Checker can tell if an electronic publication date is tagged completely to PMC Style (contains values in year, month, and day elements) in a file, but it can't tell if the values themselves are correct and actually represent the electronic publication date of the article.

Manual QA is done by the JM's after the automated workflow has successfully completed. PMC's QA system shows each JM the journals that are assigned to her, and which articles need to be checked. The QA System marks a percentage of articles from each "batch" of new content deposited by the journal for manual QA. By default, new journals that come out of data evaluation and move into production are set with a higher percentage of articles selected for manual QA. Once the JM is confident in the journal's ability to provide good, clean data, the percentages are lowered. If the JM begins to see problems on an ongoing basis, the percentage of articles checked may be increased. QA errors are grouped into eight major categories: Article Information, Article Body, Back Matter, Figures and Tables, Special Characters and Math, Generic Errors, Image Quality, and PDF Quality. Within each of these major categories, there may be one or more sub-categories. For example, in the "Article Body" section there is a subcategory for "Sections and Subsections," containing errors for missing sections, or sections that have been nested incorrectly in the flow of the body text. The JM looks at each article selected for QA and goes through all the categories and subcategories in this checklist that apply, and records any problems found. Error reports are then sent to the publisher or content provider and revisions are requested.

PMC also has a series of automated data integrity checks that run nightly for articles that have successfully passed through the automated workflow and have been loaded to the database. The integrity checks can identify, among other things, problems like duplicate articles submitted to the system, and potential discrepancies in issue publication dates for a group of articles in the same issue.

Article Identifiers and Version Numbers

There are several different types of identifiers that are used within the PMC system. This section describes these various IDs, and how they are related. Conversion among some of these IDs is available through the PMCID - PMID - Manuscript ID - DOI Converter tool (12), which is described in more detail below and summarized in Table 1.

PMCID, Article IDs, and UIDs

The most basic type of identifier used in the PMC system is the PMCID, which uniquely identifies an article. The PMCID is composed of the letters "PMC" followed by a string of decimal digits, for example, "PMCID1868567." This is also sometimes referred to as the "PMC accession number." Once assigned, the PMCID is permanent, and can be used to unambiguously refer to a particular article within PMC, from that point on.

The numeric portion of the PMCID (without the "PMC" prefix) is referred to as the Article ID, or AID. For a given article, this numeric identifier is the same across all PMCI sites (see PMCI, below).

The NCBI Entrez system refers to items in any of its databases by a numeric identifier that is known, in that system, as the UID. Every Entrez database defines a numeric UID that identifies a record in that system. In the case of PMC, the UID is the same as the AID.

Versions

A recently added enhancement to the PMC system is the ability to handle multiple versions of the same article. Each version is a distinct instance of an article, which is archived separately and made permanently available for retrieval. Versions of an article can be accessed through a URI that uses the same form as that for a canonical article URI, but with a PMCID + version number. For example, here are the three versions of a PLoS Currents article that were available at the time of this writing:

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037.1/>
- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037.2/>
- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037.3/>

Note that there are actually two URIs that can be used to access the latest version of an article. Each of these URIs refers to the same resource, but with different semantics. For example,

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037/> - This URI will always point to the latest version of this article
- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037.3/> - This URI will always point to version number 3 of this article

Every article in PMC has a version number, whether or not it actually has multiple versions. In other words, an article that only has a single version has the version number "1".

PubMed IDs

PMC articles are often identified by their PubMed IDs, or PMIDs. This is the numeric identifier in the PubMed database (see “PubMed: The Bibliographic Database”) corresponding to this article, and is independent of the PMCID. Note that not every PMC article has a PMID (although most do). Articles can be accessed by URIs using their PMID, and these cause a redirect to the canonical URI for that article. For example,

- <http://www.ncbi.nlm.nih.gov/pmc/articles/pmid/17401604/> redirects to
→ <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1868567/>

Manuscript IDs

PMC also maintains a Manuscript ID, or MID, for those articles that arrive as manuscripts, most often through the NIHMS system. In general, these manuscripts continue to be available even after the final published version of the article arrives. As with article versions (described above) these are unique article instances that are archived separately.

For example, the following article does not have a "final published version", and is available through two URIs that refer to the same document:

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3159421/>
- <http://www.ncbi.nlm.nih.gov/pmc/articles/mid/NIHMS311352/>

Whereas the following article has both a manuscript and a final published version, so these two URIs refer to different documents (different article instances):

- <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1434700/>
- <http://www.ncbi.nlm.nih.gov/pmc/articles/mid/NIHMS5786/>

DOIs

A well-known external identification system that is used to specify articles is the Digital Object Identifier (DOI). PMC does not assign DOIs, but records them when they are supplied to us, and makes articles available using these identifiers. Articles in PMC can be accessed with URIs using the DOI, which then causes a redirect to the canonical URI for that article. For example,

<http://www.ncbi.nlm.nih.gov/pmc/articles/doi/10.1172/JCI62818> redirects to → <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3484440/>

At the time of this writing, PMC does not support DOIs that refer to specific article versions, but support is planned in the near future.

ISSN, Volume, Issue, and Page

Finally, articles can also be identified by their citation information: ISSN of the journal, volume, issue, and page. For example, the Journal of Clinical Investigation has ISSN 0021-9738. To access an article from that journal's volume 117, issue #9, page 2380, you could construct a URI using the "ivip" path segment. For example:

- <http://www.ncbi.nlm.nih.gov/pmc/ivip/0021-9738/117/9/2380/> redirects to → <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1952647/>

For electronic journals that don't have pagination, the e-ID replaces that page number. For example,

- <http://www.ncbi.nlm.nih.gov/pmc/ivip/1932-6203/8/5/e52147/> redirects to → <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3653908/>

Summary

Table 1. Summary of PMC Identifiers and URIs.

Identifier	Example	Description	URI
PMCID	PMC1868567	PMC accession number	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1868567/
AID	1868567	Numeric part of PMCID	http://www.ncbi.nlm.nih.gov/pmc/articles/1868567/ (redirects)
UID	1868567	Entrez ID for PMC articles	http://www.ncbi.nlm.nih.gov/pmc/?term=1868567%5Buid%5D (Entrez result)
PMCID +version	PMC3283037.2	Specific version of an article	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283037.2/

Table 1. continues on next page...

Table 1. continued from previous page.

Identifier	Example	Description	URI
PMID	17401604	PubMed ID	http://www.ncbi.nlm.nih.gov/pmc/articles/17401604/ (redirects)
MID	NIHMS5786	Manuscript ID	http://www.ncbi.nlm.nih.gov/pmc/articles/mid/NIHMS5786/
DOI	10.1172/JCI33375	Digital Object Identifier	http://www.ncbi.nlm.nih.gov/pmc/articles/doi/10.1172/JCI33375 (redirects)
IVIP	0021-9738/117/9/2380	ISSN + volume, issue, page	http://www.ncbi.nlm.nih.gov/pmc/ivip/0021-9738/117/9/2380/ (redirects)

Retrieval / Data Processing

Indexing

Every day, the contents of the PMC database are indexed so that they are available via the NCBI Entrez interface. The Entrez interface is used by the PMC home page search facility, as well as by the Entrez Programming Utilities (EUtils; (13)). EUtils allows for third-party tools to provide discovery and search capabilities that mirror those provided by the NCBI website. In this respect, PMC is just one of the approximately 50 NCBI databases (at the time of this writing) that provide data to the public via this interface.

Fields, Filters, and Links

Every NCBI database has its own indexing criteria, including a unique set of fields, filters, and links to other databases. The PMC Help book (14) describes these for the PMC database, and gives information about how to use them to perform effective searches using Entrez.

Entrez searches allow you to enter search criteria with complex boolean expressions, using text phrases that are (optionally) qualified with either fields or filters. For example:

```
wilson eo[author] OR (eusociality AND author manuscript[filter])
```

Search *fields* are entered into a query with a text string followed by the field name in square brackets. For example,

```
wilson eo[author]
```

Consult the PMC Help book to get the list of available search fields. Search fields are added or changed occasionally, and the most up-to-date list of fields can be retrieved from one of two places:

1. The PMC Advanced Search Builder (15). Click the "All Fields" drop-down option box, for a list of all the available search field names.

2. The EInfo utility, at <http://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=pmc> (the result will be in XML). The <FieldList> element includes content describing each of the fields.

Filters are actually a special kind of field (the field named "filter") and are similar to the "tags" or "categories" that are used in many social media sites. (For example, just as a blog post might have several tags, a given record in the PMC database might correspond with several filters). There are two types of filters: standard (built-in) and custom.

Examples of built-in filters are "author manuscript," "reply," "retraction," "open access," and "CC BY NC license." To find records corresponding to a built-in filter value, enter the value in quotes, followed by the word "filter" in square brackets. For example, to find all the author manuscripts in PMC that are in the open access subset, enter the search phrase

```
"author manuscript"[filter] AND "open access"[filter]
```

You can get the complete list of filter values available by going to the PMC Advanced Search Builder, selecting "filter" in the first drop-down option box, and then clicking "Show index list".

Filters can be used by setting up your MyNCBI account to specify specific filter values that will appear as links on every Entrez search results page. This is done through the MyNCBI Filters webpage (16), selecting the "PMC" database. In the panel on the right, you can select any of the built-in filters. Clicking the checkbox enables that filter, so that it appears on every Entrez search results page, for easy access. Managing filters with MyNCBI is described in more detail in the MyNCBI Help Book under *Working with Filters* (17).

From the MyNCBI page, you can also manage custom filters, which are simply named Entrez queries. That is, any arbitrary Entrez query can be set up as a custom filter. A given record matches a custom filter value if it would be found by the corresponding Entrez query.

Links allow you to discover records in other NCBI databases that correspond to PMC documents. You can find the list of PMC-related links at the master Entrez Link Description webpage (18).

The data that correlates fields, filters, and links with particular objects in the PMC database are produced by our internal indexing tasks.

Indexing Tasks

There are two types of indexing tasks: full and merge. Merge indexing occurs daily. In the PMC database, an IndexStatus field is maintained, which keeps track of which articles have been indexed, and when. Merge indexing only operates on those articles that are new or that have been updated since the last time they were indexed.

Full indexing is scheduled to occur once per week, but might also occur if there is a specific need to re-index all of the PMC content; for example, if there is a change to the database structure, or to a search field or filter.

Note that, currently, indexing is done on the basis of UID (as described above) not individual versions of an article. Therefore, Entrez search results will always display links to the most up-to-date version of an article.

The PMC indexing tasks are integrated with the new NCBI CIDX indexing system, which is an automated workflow driven by Ergatis (19), a Web-based tool that allows workflows to be defined as pipelines, composed of reusable steps.

In addition to full text indexing, which generates the data required for searching by fields from within Entrez, the indexing tasks also generate data for filters, links, a spell-checker, and for auto-complete dictionaries. The indexing script accesses the article full text and the metadata from the PMC database, and produces XML files that are fed into the Entrez system. The XML gives full-text search strings and keywords, for each article, that are broken down by the Entrez fields that are defined for PMC.

For built-in filters, there are three possible sources of data, depending on how the filter is defined in the system:

- Explicit Filters—The indexing script produces these as explicit lists of article identifiers that match the given filter.
- Derived Filters—Like user-defined custom filters, these are based on Entrez queries.
- Links Filters—Any article that is the subject of at least one link of a given type automatically matches the filter with the same name. For example, any PMC article that cites another article in PMC matches the filter "pmc pmc cites."

For a base set of Entrez links, the indexing task generates the link from PMC records to other NCBI Entrez databases, and writes these into the Entrez links database. This database then produces some derived links automatically. For example, the link data for pmc_pmc_cites is generated by querying the PMC database to find, for a given article, all of the other articles in PMC that this article cites. The links database utilities then automatically generate the reciprocal link pmc_pmc_citedby, and stores that data.

The results of the linking processes are available from the discovery column of a PMC article display, as described in the [Entrez Help book](#), or through the Entrez utility ELink. For example, to find all of the articles in PMC that are cited by a given PMC article, you could retrieve the URI

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?  
dbfrom=pmc&id=14909&linkname=pmc_pmc_cites.
```

Text Mining

PMC mines the full text of articles in its collection for references to specific types of entities, as agreed to with PMC participating publishers. These references are to entities that are stored in other NCBI databases. The results of this text mining are stored in the TagServer database, described below. When an article is requested from a Web browser, the TagServer software then retrieves this data, and uses it to enrich the presentation.

The text mining script is technically part of the TagServer software. It is implemented as a Perl script written in modern Perl style, and based on PMC-specific Perl modules that are, in turn, based on Perl Moose (20).

The text mining process is scheduled to run every day, and it continuously re-indexes all of the articles in PMC. During each daily iteration, it processes articles in this order:

- new articles never-before mined,
- articles updated since the last time they were mined,
- all others

Thus, all of the PMC articles are re-mined on a periodic basis. (Currently the time between successive mining of a given article is about two months.) It is necessary to continuously re-index all of the PMC articles, even if they have not changed, because the databases that the mining software uses to determine referents are not static, they are constantly being updated. Therefore, re-mining the same article some period of time later can find results that were not found before.

The text-mining software currently mines the articles for references to other journal articles (in PMC and PubMed), as well as these types of data that are stored in other Entrez databases:

- taxonomy
- nucleotide
- unists
- protein
- snp
- geo
- structure

In addition to merely recognizing these terms in the articles, the software validates the results, by verifying that each of the terms actually exists in the target database.

The text mining software is used in NCBI Books and PubMed Health, as well as in PMC. The software is very configurable. Among the things that can be configured are:

- The types of terms to search for, and where (which logical sections) in the articles to search.
- What parts of articles to ignore.

The text mining software is organized as a set of plugins, with each plugin implemented as a Perl module, and mining the article for a particular kind of data.

Web Services and APIs

PMC provides several Web services and APIs to facilitate programmatic access to our resources. Among these are the OAI-PMH service, an FTP server, and the Open Access (OA) Web service.

External users wishing to reuse the content in the PMC Open Access subset should retrieve the articles from our FTP servers, rather than attempting to download them by other means.

If you have questions or comments about these, or any of the other services provided by PMC, please write to the PMC help desk. To stay informed about new or updated tools or services provided by PMC, subscribe to the [PMC-Utils-Announce mailing list](#).

You can read about each of these services on their respective description pages: the [OAI-PMH Service](#), the [FTP Server](#), and the [OA Web Service](#).

The OAI-PMH service is implemented as a CGI program, written in C++, which uses the NCBI C++ Toolkit (21).

The FTP site is populated by a "dumper" script, which encapsulates knowledge about which articles within the Open Access subset have been updated, and how to propagate those updates to the various resources on the FTP site.

The OA web service is implemented as a Perl script deployed as a fast CGI under the Apache Web server. The OA Web service uses database tables that are maintained by the same dumper script as generates the artifacts that are available on the PMC FTP site. Those tables store the information needed by this service, including, for each article in the OA subset, the most recent dates and times that a file (of either format TAR.GZ or PDF) were updated.

Usage Statistics

Each PMC participant has password-controlled access to a website that presents usage reports for that participant's journals at both the journal and article level. The reports, updated daily, include counts of available articles, total retrieval by format (such as full-text HTML and PDFs), total number of unique IP addresses that have accessed the content, and the most frequently retrieved articles. At the individual article level, usage statistics are available for every article beginning with a journal's first submission to PMC.

The reports may be downloaded as a CSV file, for analysis with a spreadsheet package such as Microsoft Excel. PMC also provides a CSV file each month, with usage for the month at the article level. Article-level usage data can also be retrieved directly via a Web service call.

PMC's usage reports generally contain all information called for in the COUNTER specification, except that PMC does not report use by specific institutions. NLM's privacy policy prevents the reporting of use at an individual or organizational level.

Rendering

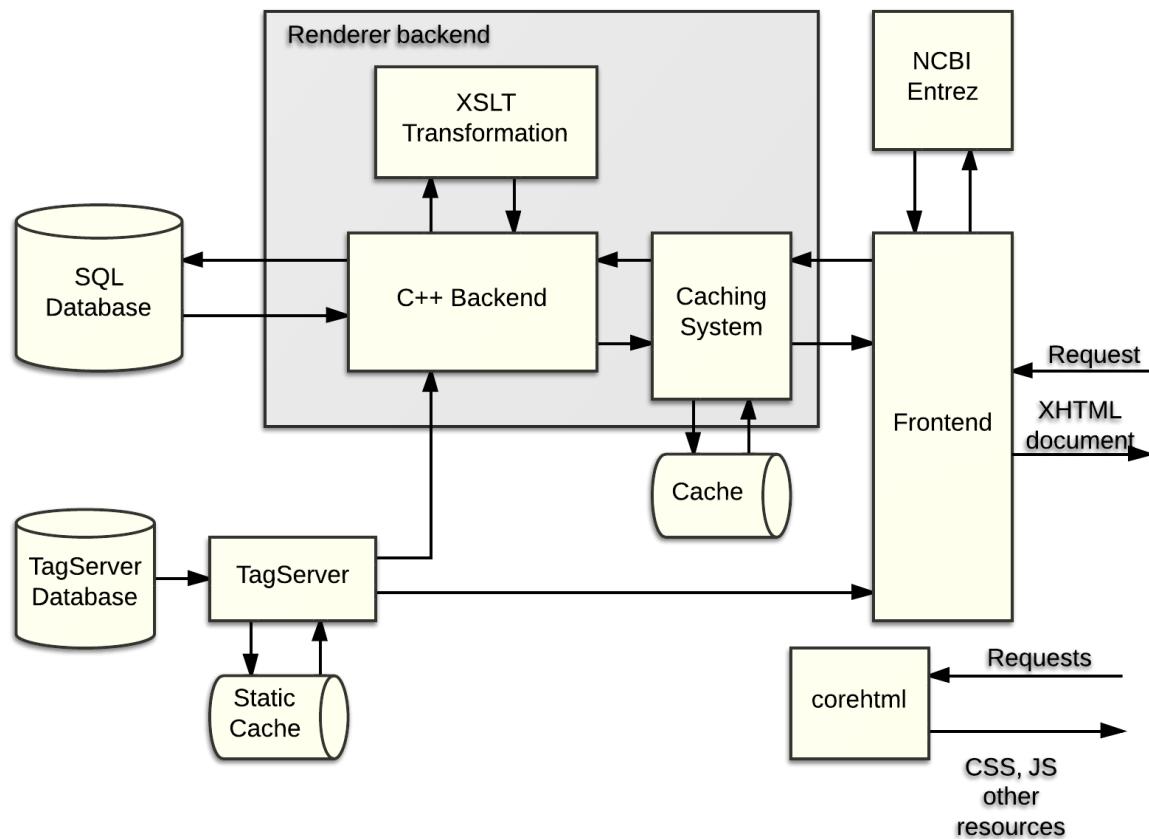


Figure 6. Components of the PMC Renderer.

Rendering Architecture Overview

PMC implements dynamic rendering of the articles at request time, passing source XML through an XSLT transformation, and integrating external data. Journal archive pages, table-of-contents (issue) pages, and articles, as well as the various discovery portlets that are displayed alongside articles, are all generated dynamically from data in the database and external sources.

Figure 6 depicts the main components of the renderer, which handle most of the Web pages and other resources available through PMC.

The SQL database is the core of the PMC archive, storing all the necessary information related to journals, issues, and individual articles that we receive from a variety of sources.

When a request arrives from a client browser, the frontend system (NCBI Portal) analyzes it, and determines, at a high level, how to handle it. Most requests are for dynamic PMC resources such journal archive pages, issue table-of-contents, or articles themselves, and these requests are routed to the renderer backend.

The renderer backend has a custom short-lived, filesystem-based caching system that is designed to improve performance under certain conditions. Currently, this caching system is disabled for the PMC system, but is enabled in the NCBI Bookshelf and PubMed Health. When enabled, the caching system first checks the request to determine if it matches a previous request that has been stored in the filesystem cache. If so, then the cached value is returned. If not, then the request is processed further.

Requests that are not cache hits are parsed and transformed into a set of SQL database queries. The results from those queries are then patched with TagServer data, and then passed through a set of XSLT transformations. Those results are then stored in the filesystem cache (in case there are later requests that match this one) and delivered back to the frontend. The frontend also accesses other NCBI resources, such as Entrez, for data that is added to the display to enhance the usefulness, and the finished results are delivered back to the client.

Many resources, such as images, javascript and css files, are served directly from a static library called corehtml.

Other pages, such as the [Home page](#), [about pages](#), and the [Entrez search results page](#) are served through the frontend, but without accessing the renderer backend or the TagServer.

Pages, Views, and URI Structure

The PMC site provides access to many different types of Web pages and other resources. Table 2 below lists those, with hyperlinks to examples.

PMC URIs are designed to be clear, concise, and resource-oriented, as well as to be consistent with the URIs of other resources across the NCBI site. The PMC home page, at <http://www.ncbi.nlm.nih.gov/pmc>, is the base URI of all of the PMC resources.

The URI space is designed hierarchically, using path segments to identify resource collections, and identifiers to specify items within those collections. For example, the URI </pmc/journals/> specifies the collection of academic journals that have deposited material into the PMC archives. The URI </pmc/journals/2/> specifies one particular journal, the Proceedings of the National Academy of Sciences.

With the PMC URI scheme, there is some flexibility in accessing some types of material because, in a number of cases, more than one URI can be used for the same resource. In these instances, one URI is considered to be canonical (primary) and using any of the ancillary (secondary) URIs for that resource will cause a redirect to the canonical form.

For example, a given article has the canonical URI </pmc/articles/PMC2150930/>, in which the identifier PMC2150930 is the PMCID for this article. However, the article could also be accessed through any of several other URIs, which each use a different scheme to identify the unique article, such as the PubMed ID (pmid), the DOI, or the issn-volume-

issue-page (ivip). When accessed via those other URIs, the client receives an HTTP redirect to the canonical URI. See Table 2 for a list of URIs supported by the PMC site.

Table 2. A list of URIs supported by the PMC site. Canonical URIs are given in bold.

Resource	URI(s)
PMC home page	/pmc/
Entrez search results	/pmc/?term=protein
Static "about" pages	/pmc/about/intro/
List of journals	/pmc/journals/
List of journals matching search	/pmc/journals/?term=respiratory
A specific journal archive	/pmc/journals/2/ /pmc/journals/domain/pnas/ /pmc/journals/issn/1091-6490/ /pmc/journals/ivip/1091-6490/
Latest issue	/pmc/journals/2/latest/
Issue	/pmc/issues/157490/ /pmc/ivip/0021-9738/117/8/
Article full text	/pmc/articles/PMC2150930/ /pmc/articles/2150930/ /pmc/PMC2150930/ /pmc/2150930/ /pmc/articles/pmid/16511247/ /pmc/articles/doi/10.1107/S1744309105040984 /pmc/ivip/0021-9738/117/9/2380/
Article alternative views: PubReader, classic, printable	/pmc/articles/PMC2150930/?report=reader /pmc/articles/PMC2150930/?report=classic /pmc/articles/PMC2150930/?report=printable
Scanned article browse page	/pmc/articles/PMC2483210/
Scanned article page	/pmc/articles/PMC2483210/?page=3
Article manuscript or version	/pmc/articles/mid/NIHMS20955/ /pmc/articles/PMC3283037.2/
Article PDF and EPub	/pmc/articles/PMC2150930/pdf/f-62-00001.pdf /pmc/articles/PMC2150930/epub/
Article abstract	/pmc/articles/PMC2150930/?report=abstract
Figure	/pmc/articles/PMC2278217/figure/F5/
Table	/pmc/articles/PMC2278217/table/T1/
Cited-by list	/pmc/articles/PMC369838/citedby/

Furthermore, we use a couple of NCBI-standard query string parameters to specify various views (report) and formats of these resources. So, for example, "?report=reader" accesses the PubReader view of a full text article.

SQL Databases

PMC uses Microsoft SQL Server to store all of the articles, supplementary material, and metadata of the archive.

Within the database, we define the term "domain," which corresponds roughly to an individual journal. In a simplified view, each publisher can have many domains, each domain can have many issues, each issue has many articles, and each article has one-to-many versions.

There is also a separate table to store information about the citations within an article, including the identifier of the item that is referenced, and another table to store a set of relationships between various articles, including, for example, links to commentary, corrections, and updates.

The actual source content for the articles are stored as "blobs" in the database, including the source XML, images, thumbnails, PDF files, media files, supplementary material, etc. These blobs are stored within dedicated database instances. Once a blob database is full, then it is closed, and never written to or modified again.

The ArticleBlobInfo table cross-references the articles to their associated blobs in the corresponding blob database. This table allows for the dereferencing of request URIs to their requested resources, at render-time. If a particular blob must be changed or deleted for whatever reason, and the blob database in which it is stored is already closed, then we simply write the new version of the blob to a new blob database, and update the pointer in the ArticleBlobInfo table.

Rendering Backend

C++ Backend

The C++ backend is based on the NCBI C++ toolkit, and is written in C++ so that its performance is as fast as possible, and so that it can take advantage of built-in features of that library for logging, database access, and XML processing.

The software runs as a Fast CGI, and provides an HTTP API to the frontend. The frontend passes several parameters to the backend in an XML document via HTTP POST. Among those parameters are the path portion of the original URI (which identifies the requested resource), a session ID, and a user ID (which identifies the MyNCBI account, if the user is logged in). The frontend also passes the Apache environment, which contains important information like the client's user-agent string.

The C++ backend always returns an XML document that encapsulates the response, that is divided into a response header (including status code, response type, error messages if appropriate), and a response body, which includes the document payload.

When the C++ backend gets a request, it first parses the URI, to determine whether or not it is of canonical form. If the URI is canonical, then the requested resource will be

returned directly; if not, then a redirect will be performed. The results of parsing the URI also identify the resource that is being requested. Based on that information, the backend queries the PMC database.

If the query results in an error, or if there is no resource matching the identifiers provided, then the backend will return a response document to the frontend that indicates the error, with appropriate status code and error message.

When the request is for a full-text article, and there is no error, then the renderer backend will also make a request to the TagServer application to retrieve data related to the tags for this article, and that data is patched into the document (see the TagServer section for more information). The backend will pass an MD5 signature of the article document, and the TagServer will compare that to the signature that it has stored. If they do not match, then the request is rejected, and no tag data is returned. This ensures that the patching mechanism, which relies on precise byte offsets into the document, is robust.

The TagServer response actually contains two parts: ids and markers that are used to mark where in the document the tags occur, and the tag attributes. The tag attributes are bundled together and stored in memory, where they are made available to the XSLT converters (described in the next section) through an extension to the XSLT document() function.

The C++ backend also takes some of the metadata related to the resource, and writes that into the document as processing instructions (PIs).

The document is then passed into the XSLT transformer, which uses one of several possible "entry points" within the XSLT stylesheet library to determine how to process the document. The XSLTs are described in more detail in the next section.

XSLTs

The XSLT files that are used by the PMC Renderer are written in XSLT 1.0, since they are processed by libxslt within the NCBI C++ toolkit. They use a few extensions:

- EXSLT (22)
- A few custom XSLT extension functions written in C, to support internationalization
- A custom document() function, to allow the fast retrieval of the TagServer tag attributes.

As mentioned above, the C++ backend invokes the XSLTs at an "entry point," which is a main module that imports all of the others, defines top-level XSLT variables and parameters, and the top-level matching template (the template that matches the document root node). There are different entry points for, for example, full text articles in classic view versus PubReader view.

The XSLTs are designed in a modular fashion, such that there are a core set of templates that are included by all (or most) of the applications, and then each application imports those, but then overrides them selectively, as needed, in order to customize the result.

The XSLT processor gets its input from the following:

- The main XML document—this is passed to the processor as its input document, and comprises:
 - NXML—the main article as stored in the PMC database.
 - Patches inserted by the TagServer patching mechanism. These are XML elements and attributes that mark the location of tags in the NXML.
 - Processing instructions (PIs) inserted by the C++ backend, with metadata about the article from the PMC database.
- XSLT parameters, which come from:
 - Parameters passed in from the frontend,
 - Renderer backend configuration files (ini files),
 - Metadata from the PMC database.
- In-memory document—this is the tag attribute data from the TagServer. This is accessed from within the XSLTs using the `document()` function, which invokes a custom extension integrated with the C++ backend.

NCBI Portal

The frontend system that is used to render PMC content is an internally-developed XML-based web application server known as NCBI-Portal (referred to simply as "Portal" for the rest of this section). Portal is written in C++, based on the NCBI C++ toolkit, and uses the libxml and libxslt libraries for XSLT and the Zorba XQuery processor (23) for XQuery.

The Portal system is used throughout NCBI, and individual applications are implemented with their own applications, which are bundles of components called snapshots. Each snapshot is versioned independently. The PMC site is implemented with three snapshots: PMCVIEWER, PMCStatic, and the PMC Entrez snapshots.

Request Router

Since the Portal system handles requests for all of NCBI, the first order of business for the Portal system, when it receives a request, is to dispatch that request to the correct snapshot. This is done in two steps:

1. The combination of the top-level domain (i.e., "www") and the topmost path segment of the URI (in our case, "pmc") is used to select a request router. (Within NCBI we use other top-level domains, for example, "test", for development and testing, and these might resolve to different request routers.)
2. The request router contains a list of regular expression rules that are matched against the URI and the request environment. The first rule that matches selects the snapshot that will handle this request.

This design allows each application within NCBI to be independent, which is important in such a large and diverse organization. It also provides great flexibility in handling requests, because it permits fine-grained control over how specific URIs are handled within a site.

As mentioned above, the PMC site is handled by three separate snapshots, which are described in more detail below.

PMCViewer Snapshot

The PMCViewer snapshot is the main snapshot used by the PMC system. It handles the rendering tasks for the journal list, journal, issue, and article pages, as well as many others. This is the snapshot that interacts with the renderer backend, as shown in Figure 6. (The other snapshots produce different types of pages that do not originate with the renderer backend).

The snapshot contains a set of rules which further examine the requested URI and other HTTP headers in order to correctly dispatch the request. As the result of this, the snapshot might:

- Immediately respond with a redirect to another URI (for example, if a page has moved, or if the user has used a non-canonical URI).
- Reverse-proxy a binary resource, such as a PDF, figure or table thumbnail, etc.
- Invoke the rendering backend to retrieve page data from the database (as described above).

When invoking the rendering backend, the snapshot passes the URI and other request parameters to the rendering backend, which responds with information about how to render this particular page. The snapshot then does some further processing of the XML response from the backend, and produces the final integrated HTML page that is sent to the browser.

PMCStatic Snapshot

The PMCStatic snapshot handles the rendering of information and news pages, for example, the [PMC Overview](#). These are generated from a set of XML, XHTML and other types of files that are stored in a specific directory within the renderer backend. The directory includes these types of resources:

- nav.xml—specifies the navigation bar menu items
- redirects.xml—this specifies any redirects that should occur (in those cases where a page has moved, the old URI will redirect to the new)
- XHTML pages—these are the content of the individual pages, and are divided into the groups about, pub (for publishers), and tools (related resources)
- Images, PDFs, and other downloadable resources

When a page is requested, the URI is translated into the pathname of an XHTML file within this directory, that file is retrieved, and then it is processed with XSLT to combine

it with the navigation bar, and the other NCBI-standard components such as the header and the footer.

PMC Entrez Snapshot

The PMC Entrez snapshot is derived from the NCBI standard Entrez package of components, which allows PMC to share the same look-and-feel and many functional aspects with other applications at NCBI. In particular, the [home page](#) is delivered by this snapshot, as well as many of the search-related pages such as [limits](#), the [Advanced Search Builder](#), the [clipboard](#), and [search details](#).

Customizations to the default Entrez packages exist to provide PMC-specific behavior. For example, on the home page, a panel in the center displays up-to-date highlights on the number of articles currently archived in PMC. This display is produced from an XML file that is generated daily and pushed to the renderer backend location.

Customizations for the Entrez search results provide for, for example, displaying [imagedocsum results](#).

PubReader

PubReader is a set of JavaScript and CSS files that provide for rendering journal articles in a way conducive to reading, especially on a tablet or a small-screen device.

The article document for the PubReader view is generated by the same mechanism as the classic view, utilizing the renderer Backend, the TagServer, and the Frontend, but the format of the XHTML document is somewhat different, in terms of the specific elements and attributes used, CSS class names, and the document structure. This difference is achieved within the renderer by invoking the XSLTs with a PubReader-specific entry point.

PubReader uses some of the latest features from the HTML5 and CSS3 standards, in order to achieve the dynamic display. Chief among these is the [CSS3 multi-column layout module](#), in conjunction with a fixed page height.

Each JavaScript component is written as a jQuery extension, based on the "Nested Namespacing Plugin Pattern," by Doug Neiner, and described in *Learning JavaScript Design Patterns* (24).

Among the components of PubReader are:

- PageManager—controls and performs page turning in the PubReader.
- HistoryKeeper—monitors and controls the fragment identifier (the part after the "#" symbol) of the URI, and instructs the PageManager to turn pages to the destination defined in that fragment identifier.
- ObjectBox—this component handles the modal box that opens to display a figure, a table, or a message box.

- PageProgressBar—gives a visual indication of the current location within the document, and provides a control to allow the user to move to a new location. This component uses a modified version of the rangeinput widget from jquerytools (25).
- Links—provides the ability to hijack clicks on links.

The PubReader code is managed in the master NCBI Subversion repository, and is mirrored to the GitHub repository [NCBITools/PubReader](#).

Caching System

The caching system in PMC is used primarily to boost performance in those cases when there is a surge of requests, over a short period of time, for one or a few articles or resources. It uses the filesystem to store copies of each resource as it is requested, using a pathname that is generated by a unique key. The key value is an MD5 signature of a string that is composed of query parameters and environment variables that uniquely specify the requested resource. When another request arrives that results in the same key and before the cache entity has expired (a cache hit), then the resource is retrieved from the filesystem instead of being regenerated dynamically.

Currently, the caching system is disabled for the PMC renderer backend, but is being used with NCBI Bookshelf and PubMed Health. Each of these is each set up with an independent cache.

The types of requests that are cached are configurable. Also configurable is a setting for the minimum amount of free space on the filesystem. This prevents the caching system from filling up the disc, and is typically set to four gigabytes. Once that limit is reached, no new requests will be cached until old ones are purged.

There is a purging mechanism that runs continuously and "garbage collects" cache entities on the filesystem that have expired.

The full path for each entity in the cache is derived from the MD5 signature by using the first four hex digits and directory names, and the rest of the MD5 as a filename. The disc files that make up the cache include, in a header, metadata about the resource and about this particular cache entity, including the request parameters that resulted the cache hit, the date and time the entity was stored, and metadata from the PMC database such as article id, blobname, etc.

The caching system is configured so that an entity will expire after one hour. The reason for this is that the final rendered output for a request typically changes every day.

The system is implemented in libraries in both C++ and in Perl. They are maintained independently, but both use the same format for the filesystem and the headers.

TagServer

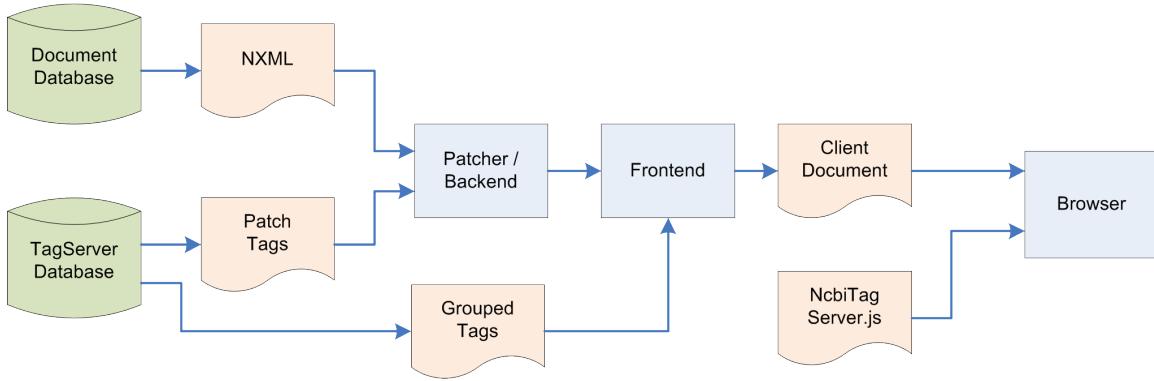


Figure 7. TagServer data flow.

Overview

The TagServer is a database system that contains metadata which has been mined from PMC articles. Tags are (typically) strings of text within a source document that are associated with a set of metadata attributes. For example, the gene name "D17Jcs48" might occur within an article in PMC. The text mining process would recognize this as a gene name, determine its numeric ID number in the Entrez gene database, and would store information about this instance of this gene name into the TagServer database, as a single tag, associated with the article instance.

The TagServer system was designed to be generic, and the TagServer database highly configurable, such that it can store metadata about a wide variety of tags that can occur in a variety of source documents.

Figure 7 illustrates the data flow involved in mashing up TagServer data with an article rendered in PMC.

As can be seen in this illustration, during rendering, the TagServer delivers two types of tags associated with an article. The first, "patch tags," are integrated with the XML document immediately as it comes out of the document database, before it is processed as XML. These are tags that are associated with very specific parts of the document, for example, specific strings of text like gene names. The second type of tags is grouped tags, and these are associated with areas within the article that are identified by XML id attributes, for example, paragraphs. The data for these tags is integrated into the XML document by the frontend XML processing engine.

Finally, the generated XHTML document is delivered to the client browser, along with a JavaScript module that performs the final steps required for rendering, such as positioning discovery blocks, implementing tooltips, etc.

Tags

Tags in the TagServer database are classified in a number of ways.

The most basic classification is the nature of the entity within the PMC article (i.e., the subject of the tag). This is associated with whether or not, and how the tag marks are patched into the document. For example, some tags refer to short snippets of text, and these are patched in as new XML elements. Others are patched in "raw," meaning that the content of the tag is simply inserted into the document. Others are not patched in at all; for example, tags that refer to the document as a whole, rather than to a specific part.

Tags are also classified according to the semantics of the reference (i.e., the object of the tag). This classification is "tag type," and is fully configurable. Examples are entrez-gene, entrez-nucleotide, glossary (for glossary terms), and reference. New tag types can be defined as needed.

There are also special types of tags that are used to store XML id values that get patched into the instance documents, to enhance their addressability.

Every tag has a set of attributes associated with it, which are key-value pairs. These attributes are stored in the database by the text mining process, and are not constrained by the TagServer software itself. However, some tag attributes have a set meaning defined by the application layer. For example, the tag attribute term_id is used to specify the numeric ID of the object of a tag. In the case of an item in an Entrez database, term_id stores the UID of the item. Other attribute names with fixed meanings are reference_id, pubmed_id, etc.

Database Design

The TagServer data is organized around the concept of a presentation, which is like a stored query. When the database is accessed with a given presentation and an object identifier, then it will respond with a specific set of tags, in a defined format. A given presentation is specified by three parameters: object type, site, and request type.

Each tag refers to a specific object in the PMC database, which is an XML instance document. Because tags are patched into this document before they are parsed as XML, it is essential that the instance document is byte-for-byte identical with the copy that was used for text mining. To ensure this, an MD5 signature of each object is stored in the TagServer database. When the article is rendered, the MD5 is again computed, and if it doesn't match, then the tag data is discarded, and not patched into the document.

API

The TagServer is accessed through a RESTful Web service API, with resources identified in the path portion of the URIs, and query parameters used to define the desired set of tags and their format. An example of a TagServer request is

```
/tags/prod/srv/pmc/2464733/tags?site=prod&rt=backend
```

This specifies the PMC article instance with ID 2464733, and the site and rt (request type) parameters specify the desired presentation. The presentation, as described above, specifies exactly what tag types are desired, and the format, sorting, and grouping.

Implementation

The TagServer is implemented as a Perl Fast CGI script, and has a separate database from the main PMC database. It has a Perl Catalyst based Web interface that allows it to be configured for each application in which the TagServer is used. That configuration consists of definitions of tag types, sites, request types, and the various presentations, or collections and groupings of tags that are returned for each type of request.

TagServer Static Cache

The TagServer static cache system is used to boost the performance of the TagServer. The static cache itself is a large binary file that holds all of the possible "realistic" TagServer responses for all of the PMC articles in our archives. "Realistic" responses mean those that are actually generated by requests from production servers, given the way they are currently configured.

Currently, for every article, there are three different realistic responses, corresponding to the three request types: backend, frontend, and indexer. Multiplying those by about three million articles means that there are on the order of ten million responses that are stored in the static cache file.

The static cache file stores the responses as a large hash table, using a form of open addressing (26) collision resolution algorithm known as Cuckoo hashing (27). This file is regenerated every day from all the PMC articles, including any that have been newly mined.

This design means that the static cache system is very efficient: since it contains (nearly) all of the possible responses, the hit rate is very high (about 99.9%). The only requests that do not hit the cache are those for articles that were updated since the last time the static cache was regenerated. Also, the response time when there is a hit is on the order of 10 milliseconds, versus 150 milliseconds when there is no hit.

PMCI

PMC International (PMCI) is a collaborative effort between the NLM, publishers, and organizations in other countries, to create a network of digital archives that share content. Currently there are two active PMCI sites: Europe PubMed Central ((28); originally UKPMC), which went online in January, 2007, and PMC Canada (29), which went online in October, 2009. See the PMC International page on the PMC website (4) for more general information about this collaboration effort.

The PMCI sites use the same database architecture, the same backend software, and the same business rules as NCBI PMC. They deploy the following software developed by NCBI:

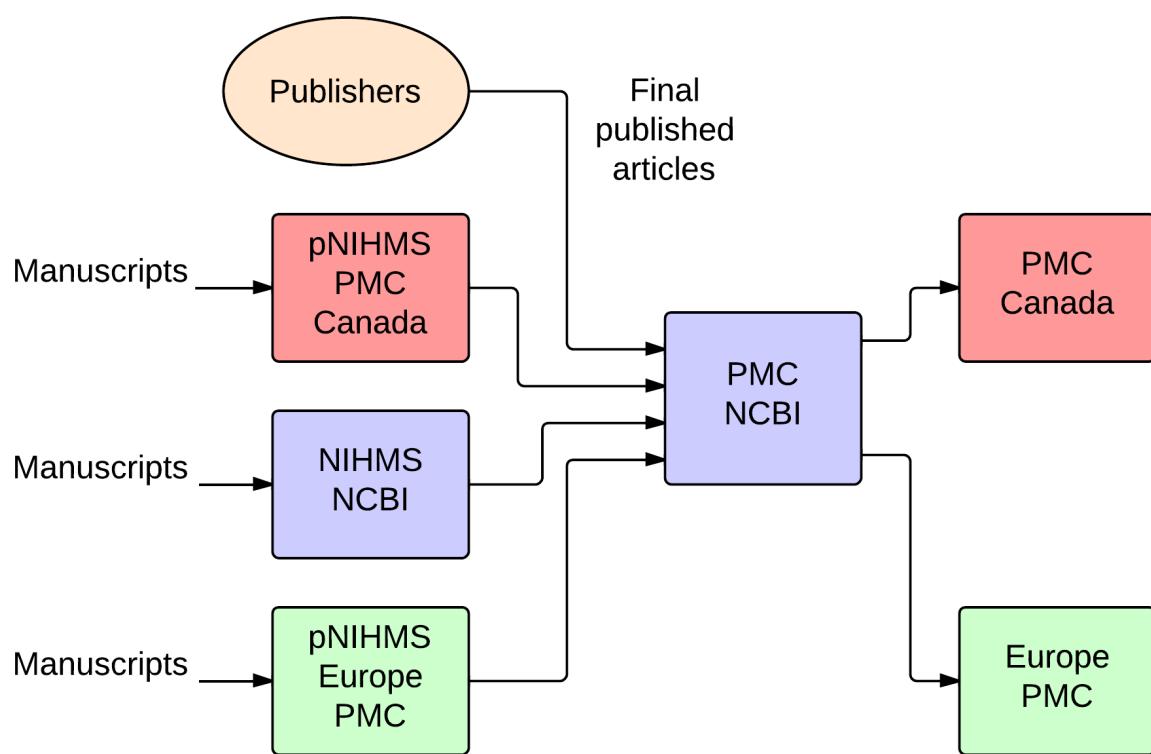


Figure 8. Data exchange between PMCI sites and NCBI PMC.

- Portable NIHMS (pNIHMS)—a portable version of NIH manuscript submission system ([NIHMS](#)) that allows authors and publisher to submit manuscripts directly to one of the PMCI sites.
- Portable PMC (pPMC)—an archiving and rendering system used to store and deliver the content to the end users.

The pPMC software has a collector component that communicates with NCBI PMC to update content in the pPMC database. The collector enables the exchange of content and metadata between the PMCI site and NCBI PMC, as illustrated in Figure 8. Whereas publishers submit final published versions of all articles directly to NCBI PMC, PMCI sites receive author manuscripts through the pNIHMS system. For example, Europe PMC accepts and processes author manuscripts of journal articles funded by the Europe PMC sponsoring agencies (30).

All content is initially sent to NCBI PMC, and then redistributed to the PMCI sites. NCBI controls which new and updated content is distributed to which sites, and makes that content available through the PMCI collector system. The PMCI sites retrieve that content from the PMCI collector, as frequently as necessary.

The pPMC sites use the same database structure as NCBI PMC, and the software includes most of the same rendering components, including the rendering backend . The frontend

is different for each, which allows each PMC site to customize and enrich the article display in its own way.

Note that in order to support PMCI, the PMC renderer software is internationalized, to allow for multilingual translations of various components.

Other Tools and Utilities

PMC provides a number of tools and utilities, mostly to aid publishers who deposit content in our archive, but a few others that are of general usefulness. The publisher tools and utilities are described on the File Validation Tools page [31].

Here is a list of the tools and utilities provided by PMC:

- Citation Search
- PMCID - PMID - Manuscript ID - DOI Converter
- XML validator and SGML validator
- PMC Style Checker
- Article Previewer

Most of these tools are served through the PMCStatic snapshot, described above.

Citation Search

This is a very simple form interface to the Entrez search system.

PMCID - PMID - Manuscript ID - DOI Converter

This is served by the PMCStatic snapshot of the frontend system, at the URL <http://www.ncbi.nlm.nih.gov/pmc/pmctopmid/>. This tool is just a wrapper for the ID converter API service. See the [ID Converter API](#) documentation page for more information.

XML and SGML Validators

These are served by the PMCStatic snapshot, which accesses a CGI backend written in Perl to handle the validation. The uploaded file is sent the CGI backend via an HTTP POST request, and a separate Perl module is engaged to validate the document, and report errors.

PMC Style Checker

The Style Checker verifies that an uploaded document conforms to the PMC XML Tagging Guidelines (9). This utility is served by the PMCStatic snapshot, via a CGI backend written in Perl. That CGI script sends the uploaded document through a set of XSLT transformations that check the document conforms to the set of rules defined by the tagging guidelines. A downloadable version of these XSLT transformations is available.

Article Previewer

The article previewer allows users to view uploaded articles the way they would appear in PMC. The tool requires users to have a MyNCBI account, and it associates any uploaded articles with that account.

The tool runs the same processes that are used during PMC production. This allows users to view an article as it would appear in PMC that is tagged in NLM XML according to PMC Style, or any XML or SGML DTD that PMC currently accepts for data submissions. See the Article Previewer Instructions and FAQ for more information.

The article previewer is currently implemented as a stand-alone CGI program (not served through any NCBI Portal frontend snapshot). It accesses a separate and independent database that is created with the same schema and tables as the main PMC production database. To render the converted articles, it uses a stand-alone version of the renderer backend.

References

1. Add a Journal to PMC [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pub/pubinfo/>.
2. License your work [Internet]. Mountain View, CA: Creative Commons; 2013 [cited 2013 Nov 1]. Available from <http://creativecommons.org/about/license/>.
3. Public Access [Internet]. Bethesda, MD: National Institutes of Health; 2013 [cited 2013 Nov 1]. Available from <http://publicaccess.nih.gov/>.
4. PMC International [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/about/pmc/>.
5. PMC File Submission Specifications [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pub/filespec/>
6. Fact Sheet: Technical Services Division [Internet]. Bethesda, MD: National Library of Medicine; 2013 [cited 2013 Nov 1]. Available from <http://www.nlm.nih.gov/pubs/factsheets/tsd.html>.
7. Minimum Criteria for PMC Data Evaluations [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2009 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pmcdoc/mindatareq.pdf>
8. XML Catalogs. OASIS Standard, Version 1.1. 7 October 2005. (Available at: <http://www.oasis-open.org/committees/download.php/14810/xml-catalogs.pdf>)
9. PubMed Central Tagging Guidelines [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/style.html>
10. OAI-PMH Service [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/about/oai.html>.

11. W3C. “Processing Instructions.” Extensible Markup Language (XML) 1.0 (Fifth Edition). 2008 [cited 2013 Nov 1]. Available from (<http://www.w3.org/TR/REC-xml/#sec-pi>).
12. PMCID/PMID/NIHMSID Converter [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pmctopmid/>.
13. Entrez Programming Utilities Help [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2010 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/books/NBK25501/>.
14. PMC Help [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2005 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/books/NBK3825/>.
15. PMC Advanced Search Builder [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/advanced/>.
16. My NCBI – Filters [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/sites/myncbi/filters/>.
17. “Working with Filters” My NCBI Help [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2010 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/books/NBK53591/>.
18. Entrez Link Descriptions [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://eutils.ncbi.nlm.nih.gov/entrez/query/static/entrezlinks.html#pmc>.
19. Ergatis: Workflow creation and monitoring interface. [Internet]. [cited 2013 Nov 1]. Available from <http://ergatis.sourceforge.net/>.
20. Moose: A postmodern object system for Perl [Internet]. Infinity Interactive; 2006 [cited 2013 Nov 1]. Available from <http://moose.iinteractive.com/en/>.
21. The NCBI C++ Toolkit Book [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2004 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/books/NBK7160/>.
22. EXSLT [Internet]. [cited 2013 Nov 1]. Available from <http://www.exslt.org/>.
23. Zorba NoSQL Query Processor [Internet]. 2008 [cited 2013 Nov 1]. Available from <http://www.zorba.io/>.
24. Osmani, Addy. Learning JavaScript Design Patterns [Internet]. 2012 [cited 2013 Nov 1]. Available from <http://addyosmani.com/resources/essentialjsdesignpatterns/book/>.
25. RANGEINPUT [Internet]. jQuery Tools [cited 2013 Nov 1]. Available from <http://jquerytools.org/documentation/rangeinput/index.html>.
26. Open Addressing [Internet]. Wikipedia [cited 2013 Nov 1]. Available from http://en.wikipedia.org/wiki/Open_addressing.
27. Cuckoo hashing [Internet]. Wikipedia [cited 2013 Nov 1]. Available from http://en.wikipedia.org/wiki/Cuckoo_hashing.
28. Europe PubMed Central. [cited 2013 Nov 1]. Available from <http://europepmc.org/>.
29. PMC Canada. [cited 2013 Nov 1]. Available from <http://pubmedcentralcanada.ca/pmcc/>.

30. Europe PubMed Central Funders. [cited 2013 Nov 1]. Available from <http://europepmc.org/funders/>.
31. File Validation Tools [Internet]. Bethesda, MD: National Center for Biotechnology Information; 2013 [cited 2013 Nov 1]. Available from <http://www.ncbi.nlm.nih.gov/pmc/pub/validation/>.

Bookshelf

MariLu Hoeppner^{✉1} Martin Latterner,¹ and Karanjit Siyan¹

Created: March 18, 2013; Updated: November 4, 2013.

Scope

Bookshelf is a biomedical literature trove, whether you are preparing for a college biology test, studying health trends, or investigating the molecular basis of a gene mutation.

Bookshelf (<http://www.ncbi.nlm.nih.gov/books/>) is an online resource providing free access to the full text of books and documents in life sciences and health care, built and maintained by the National Center for Biotechnology Information (NCBI) within the National Library of Medicine (NLM) (1). Bookshelf includes books, reports, documentation, and databases in life sciences and health care.

Bookshelf data is tagged in XML in the NCBI Book DTD (Document Type Definition), which is modeled after the NLM Journal Article DTDs. Book content follows a processing route similar to journal articles; tagging book data in a format similar to journal articles in PMC (PubMed Central) has enabled Bookshelf to use existing PMC infrastructure and workflows for processing book content.

Bookshelf aims to further advance science and improve health care through the collection, exchange, and dissemination of books and related documents in life sciences and health care. As a literature resource at NCBI, Bookshelf serves to provide annotations for the factual information residing in the genomic and molecular databases such as Gene and PubChem, and facilitate the discovery of this information.

History

Bookshelf started in 1999, with a single book, the third edition of *Molecular Biology of the Cell*, Alberts et al. (2). The first few books in Bookshelf were college text books. In the early days of Bookshelf, terms in PubMed abstracts were linked to the books which served as encyclopedic references for these terms. With the introduction of the [Health Services/Technology Assessment Texts](#) (HSTAT) collection to Bookshelf in 2004, a large number of health reports were added to Bookshelf. Today, there are over 1700 titles in Bookshelf (see Figure 1).

¹ NCBI; Email: hoeppner@ncbi.nlm.nih.gov; Email: latternm@ncbi.nlm.nih.gov; Email: siyan@ncbi.nlm.nih.gov.

[✉] Corresponding author.

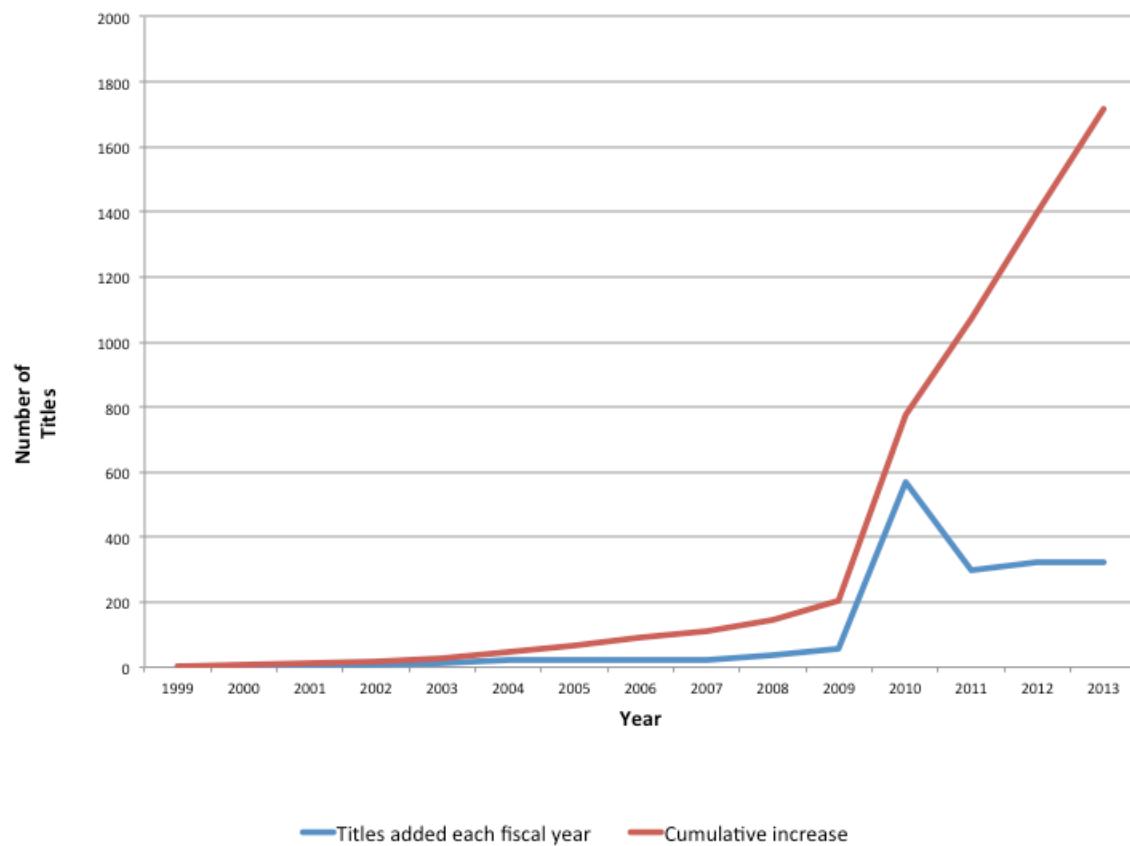


Figure 1. Number of titles added in Bookshelf each fiscal year (October to September) and cumulative growth. The spike in 2010 represents a restructuring of the HSTAT collection.

The Collection and Content

The collection is broadly biomedical in scope, comprising a diversity of works. They include books, reports, literature databases, and documentation ranging from basic undergraduate text books to specialized publications in life sciences and healthcare. Titles are selected for the collection based on three criteria: (1) scope, as defined by NLM's [Collection Development Manual](#); (2) scientific and editorial quality of the content; and (3) technical considerations, such as the quality of submitted files. Some works are in the public domain, whereas others are copyrighted works for which the copyright holders have granted NCBI distribution rights. Once content is selected for the collection, participants sign an agreement. See [Information for Authors and Publishers](#) for details on the selection process, how to apply, and to view the agreement.

Bookshelf serves users seeking biomedical information; they include college and graduate students, scientists, healthcare professionals, and patients. The free availability of the content ensures that the information is accessible by users who might otherwise not have access to this data. The content providers agree to make the content freely available; they

include authors, editors, publishers, and administrators from universities, publishing houses, US and international government agencies, as well as organizations in the health sector. Publishers and content providers also benefit when their content is widely distributed to the general public, to health care professionals, and to a population of students who will become the next generation of biomedical researchers, clinicians, and teachers.

Some content providers also agree to participate in the [Open Access subset](#). For content in the Open Access subset, XML, image and supplementary files are shared, allowing for redistribution and reuse of the content.

Data Model

Format and Structure

Early in the project, Bookshelf used a DTD based on the ISO 12083 article DTD for tagging data in XML format. As the project grew with more data being added, the tag set had to be modified, complicating data management and rendering. This led to the development of the [NCBI Book DTD](#), which is modeled along the same design principles as the DTDs of the Journal Article Tag Suite (JATS), and utilizes many of same modules. Bookshelf XML data are currently tagged in the NCBI Book DTD, v2.3. The similarities between book chapters and journal articles, and between their shared tag sets, have permitted Bookshelf to leverage the robust PMC architectural framework as well as existing PMC workflows and tools for handling the data. The NCBI Book DTD in the context of JATS has been discussed in detail (3).

Submission, XML Conversion, and Storage

Tagging content semantically in XML is one of the most complex and costly operations for Bookshelf. To enable continued maintenance of the corpus of book data, and continued growth of Bookshelf, it has been necessary to balance the needs of the publisher with the resources of the Bookshelf by streamlining the number of submission formats. To this end, Bookshelf recently moved toward a requirement for data submission in semantically tagged XML, which permits partial or complete automation of data processing. XML data are submitted either in the NCBI Book DTD or in an alternate DTD (e.g., DocBook). When submission utilizes an alternate DTD, Bookshelf employs XSLT converters to transform the XML to the NCBI Book DTD format. For submission of data in NCBI Book DTD XML, [tagging guidelines](#) have been developed and are based on similar tagging guidelines for PMC. These guidelines are intended to guide proper tagging practice through tagged samples, to reduce the variability in tagging data elements, and to facilitate data exchange.

A subset of Bookshelf projects that require frequent updates are authored in a specialized Microsoft Word template that utilizes styles to semantically tag the document elements, such as the title, author list, etc. The documents are converted to XML using the in-house NCBI Word Converter tool that utilizes the eXtyles product (Inera, Inc.) for reference

processing. Documents are updated in Microsoft Word and reprocessed using the Word Converter. Legacy projects involving print publications are submitted in PDF format and are converted by third-party vendors to NCBI Book DTD XML. FTP is the main portal for data submission.

For the majority of books (>99.5%), XML, image, source files (example, publisher-supplied PDFs, Word), and supplementary files are stored in a content management system (CMS), built in-house for the Bookshelf project. The CMS is the destination hub for NCBI Book DTD XML data that is received through a number of workflows and the staging area for ingest and subsequent processing of book data. All XML content stored in the CMS is in the form of a master XML document that describes the book's metadata and individual book part elements such as chapters and appendices. For convenience in editing the book, the individual book chapters and appendices are in separate XML files. Support data files for the book such as figure images, PDFs, supplementary files, as well as original source files are also stored in CMS. In the CMS, book data are checked for validation against the DTD, conformance to an in-house stylechecker (which runs additional checks beyond XML validation to ensure data quality), and additional integrity checks to ensure that all files associated with the book are available (see below, Performing Quality Assurance).

The different operations such as validation, style check, integrity check and loading to PMC can be selected and run separately by the user. However, these operations can also be defined as a workflow and the workflow can be run as an interactive or batch process that ensures that the operations are executed in the intended order specified in the workflow. The workflow is described as an XML document. The elements of the workflow are described using W3C schema and include the CMS operations and conditional and branching logic to execute the next step dependent on the success of previous steps. Defining workflows using XML gives users the flexibility of creating custom workflows and modifying them as future needs change.

The CMS is set up so that most operations dealing with the content processing can be done or initiated from the CMS. For example, an XML file can be edited by selecting it from CMS and running the Oxygen XML Editor (SyncRO Soft SRL) and saving the results back to CMS. There is no need to copy the file outside CMS and edit it separately and then upload the edited file to CMS. Another example is authoring content using the Microsoft Word template (above). There is a separate area in the CMS for storing book chapters authored in MS Word. These Word documents can be converted to XML by initiating the conversion action from CMS. The Word documents are converted to XML and the results stored in CMS.

Books contents in CMS can be searched using XQuery. The XQuery scripts can be stored and edited directly in CMS and run against any set of books. The XQuery and workflows can be set to run immediately or at a future time using a built-in scheduler. This enables workflows and queries that require heavy processing to be performed at times when the system is not so heavily used.

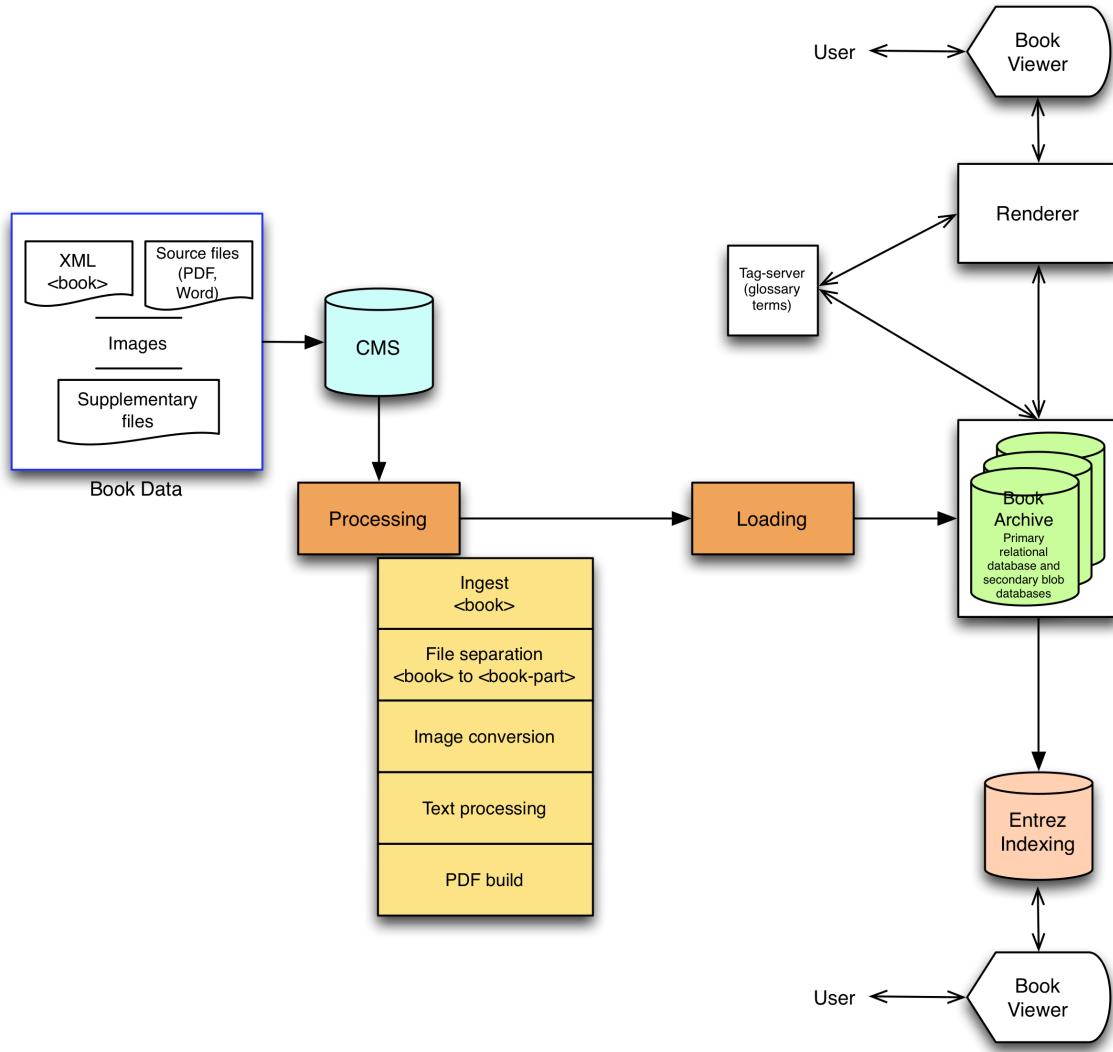


Figure 2. Books data workflow.

Dataflow

From the CMS, content is then processed for storage in the books archive to enable fast delivery to the Web, and for the automated creation of alternative formats (example, PDF). The main steps of data processing are: (a) ingest, (b) “chop-it-up” process, (c) text and image processing, and (d) PDF build (see Figure 2). Ingest begins with downloading XML, image, and supplementary files from the CMS onto the file system then bundling them to create a tar file; in cases where the CMS is bypassed (<0.5% of books), data is directly ingested following deposit to the FTP site.

Chop-it-up and text processing involve XSLT transformation on XML data, creating XML output. During the chop-it-up process, the single independently validating NCBI Book

DTD XML document with root element <book> is separated into independently validating XML documents with root element <book-part>; i.e., the book is divided into standalone book units such as front-matter sections, chapters, appendices, or reference lists. Book-metadata is carried into every book part. The creation of article-like <book-part> XML files from the <book> XML has provided the basis for using the PMC workflows and tools for Bookshelf data processing.

Text processing and image conversion occur in parallel. For text conversion, the software resolves named entities, handles special or custom characters and custom math, validates XML, and runs the stylechecker. For image conversion, the software which runs on open-source ImageMagick (ImageMagick Studio) determines image dimensions and properties, such as size, type, and resolution, resizes images per Bookshelf specifications, and creates for each image a thumbnail, a Web-resolution JPEG file, and a high-resolution JPEG file (if the source files were of high resolution).

PDFs are created for book chapters if not provided by the content provider and if their creation and display in Bookshelf is permitted. The PDF build software uses the XML output of text conversion and creates a formatting object (FO) file, gathers image heuristics, and resizes images so they are compatible with print layout. The Antenna House formatter (Antenna House, Inc.) creates the PDF from the formatting object file.

Loading to the Database

The loading software identifies the XML files for addition or replacement and loads them to the database. Each book in the database is referred to as a domain. The loader validates the data, and performs checks for file types and associated files; resolves loading of files associated with each XML file, such as images, equations, multimedia, and supplementary files. It parses the XML for key metadata information, such as book-part identifiers for storage in the main database tables. Citations that have PubMed identifiers are stored in the database. The loader creates a unique accession ID, with the “NBK” prefix for each book part.

The book database is very similar in design to the PMC article database (see SQL Databases). It is actually a database cluster with a primary database for the main relational tables holding book and book part information, as well as their properties and attributes; and several secondary blob databases for holding the XML and associated file blobs.

Rendering

Bookshelf dynamically renders the book-part XML to HTML Web pages at request time. The architecture closely corresponds to the PubMed Central (PMC) rendering model: The NCBI frontend system analyzes a request from a client browser and routes it to the renderer, a FastCGI program written in C++. The program retrieves the book-part XML as well as additional information about the book-part, for example PubMed IDs of references cited in the content. It runs the data through an XSLT transformation and then passes it back to the frontend, which returns an HTML page to the client. Bookshelf uses

the PMC Caching system in order to deliver its pages faster. It also exploits the PMC TagServer as a tool to enrich the content, for example by mining and storing glossary terms mentioned across a book-part.

Performing Quality Assurance

Quality assurance checks aim to protect the fidelity of data through all stages of processing and ensure accurate rendering and retrieval by the user. Bookshelf uses both manual and automated procedures for performing quality assurance checks. Metadata checks against the source documents, as well as integrity checks (to ensure that all book files are included) are performed in the CMS. Following ingest, processing and loading to the SQL databases, checks are also performed in the Book Viewer application to ensure that all data is accurately rendered.

Indexing

Bookshelf records are indexed in Entrez, NCBI's global indexing, retrieval, and discovery system. Entrez records are created for a complete book, for its individual chapters as well as for lower-level units, such as sections or tables. A Bookshelf Entrez record mainly contains:

- Main search text which comprises the body of the content unit;
- Search fields based on bibliographic and subject metadata, for example, authors or title; and
- Specially computed keywords and phrases.

The indexing process runs each night. A Perl program retrieves the book part XML files from the database. It passes it through an XSLT transform to produce simplified “indexing documents,” extracting the bibliographic search fields and the search text. It also interfaces with a program maintained by NCBI's Computational Biology Branch to compute important keywords from the book XML and merges those into the indexing document. The latter is then fed into the global Entrez indexing pipeline.

In addition to the main indexing records, the process also produces Entrez filters and links: it collects, for example, all records belonging to a particular bibliographic series or set into a filter, which enables the user to limit her or his search to a particular collection of interest. It creates link pairings to other NCBI databases, for example, to PubMed Records cited in a chapter or to a Gene records tagged in the book XML.

Access

Search

Users can search Bookshelf for a term or phrase across all books or in a single book. An advanced search builder and the ability to apply limits to the search query are available. Standard search features familiar to PubMed users, such as Save search, Send to

Clipboard, and Search details are also available. See [Searching Bookshelf](#) for details on performing a Bookshelf search.

Example

Search for term: heart attack

Bookshelf uses some of the query processing facilities available in the Entrez system. Search terms, for example, are expanded via a Medical Subject Headings (MeSH) translation table used also in PubMed. Similarly, the system employs a spell-checker or uses phrase tokenization if an original user query yields no results.

Browse

Books can be browsed using an application that allows users to filter the list of books by entering a term into a text box or by selection of one or more of the following categories: subject, type of publication, and publisher. A URL request is sent from the client to the browse application backend and the backend response is handled using AJAX (Asynchronous JavaScript and XML), allowing fast loading of the page without a reload. This tool is available at: <http://www.ncbi.nlm.nih.gov/books/browse/>. See [Browsing Bookshelf](#) for details on using the browse tool.

Read

The book viewer application presents book content to the reader, as in the page you are currently reading. It facilitates navigation within the book, as well as within the page. Through this application, users can access all features of the book such as tables, figures, glossaries, bibliographic reference lists, download alternate formats, view bibliographic information, copyright and permissions, and cite the content.

Related Resources

A subset of books and book chapters in Bookshelf are indexed in PubMed. They are searchable using the filters “pmcbook” and “pmcbookchapter” respectively in PubMed. They are identifiable in the PubMed result summary by the label “Books and Documents.” MARC records are available for Bookshelf titles and can be downloaded from the following FTP site: <ftp://ftp.ncbi.nlm.nih.gov/pub/bookshelf/>. Bookshelf catalog records can also be found in the NLM Catalog.

References

1. Hoeppner MA. NCBI Bookshelf: books and documents in life sciences and health care. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D1251–60. [10.1093/nar/gks1279](https://doi.org/10.1093/nar/gks1279). Epub 2012 Nov 29PubMed Central PMID: PMC3531209; doi. PubMed PMID: 23203889.

2. Alberts B, Bray D, Lewis J, et al. Molecular Biology of the Cell. 3rd edition. New York: Garland Science; 1994. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK20684/>
3. Latterner M, Hoeppner M. Bookshelf: Leafing through XML. 2010 Oct 12. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2010 [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK47113/>

NLM DTD to NISO JATS Z39.96-2012

Jeffrey Beck¹ and Laura Randall¹

Created: November 14, 2013.

Scope

The Journal Article Tag Suite (JATS) is a description of a set of elements and attributes that is used to build XML models of journal articles for archiving, publishing, and authoring. JATS became an American National Standard (ANSI/NISO Z39.96-2012) in August 2012, but it was already a well-established specification (known by the colloquial name “NLM DTD”) by the time work began on standardization in late 2009.

Normalizing the structure of journal articles enables interchange of articles among publishers, authors, data conversion vendors, and aggregators such as archives and indexing services. An existing, well used, and freely available article model also allows new, small journal publishers to start creating articles in XML significantly faster, more easily, and at less cost than if they had to create a model and persuade their vendors and publishing partners to use it.

History

PMC DTD

PubMed Central (PMC), developed and maintained by the National Center for Biotechnology Information ([NCBI](#)), is the NLM’s digital library of full-text life sciences journal literature. The project’s mission is to make full-text article content (submitted by participating publishers) available through a public database. The only technical requirement when PMC started in 1999 was that publishers supply the articles in either SGML or XML format and include all images.

It quickly became obvious that article content needed to be normalized into a single article model on ingest to reduce the stress on the database and the software that rendered the articles on the Web. The PMC Document Type Definition (DTD), known as `pmc-1.dtd`, was written based on the two article models that were being submitted to PMC at the time, and its main focus was the online representation of the articles.

The original article model was built based on a small sample set, and as publishers submitted new formats for inclusion in PMC, the `pmc-1.dtd` grew to handle new article structures. This approach did not scale, so NCBI contacted Mulberry Technologies, Inc., in Rockville, Maryland, to perform an independent review of the `pmc-1.dtd` and to work on a replacement model.

¹ NCBI; Email: beck@ncbi.nlm.nih.gov; Email: lrandall@ncbi.nlm.nih.gov.

Universal DTD for Electronic Journals

In 2001, the Harvard University Library E-Journal Archiving Project (using funds from the Mellon Foundation) commissioned a study into the feasibility of having one DTD that could be used to archive all electronic journals (6). The report prepared by Inera, Inc., Belmont, Massachusetts, was a survey of the journal article DTDs from PubMed Central and the following publishers:

- American Institute of Physics
- BioOne
- Blackwell Science
- Elsevier Science
- Highwire Press
- Institute of Electrical and Electronics Engineers
- Nature Publishing Group
- University of Chicago Press
- John Wiley & Sons

The report concluded that there could be a single DTD that could accommodate any electronic journal article, but none of the existing DTDs in the study met all of the requirements.

At this point, the modification of the pmc-1.dtd was well under way. Many of the suggestions from the study were incorporated into the modified PMC article model. When the modified model was shared with Bruce Rosenblum from Inera, he determined that the pmc-2.dtd was almost the one model that they had been looking for during the feasibility study.

A meeting was held in the spring of 2002 at the NLM that included representatives of NCBI/NLM, the Harvard Library, the Mellon Foundation, Mulberry Technologies, and Inera to try to work out the details of adopting the new pmc-2.dtd to general use for archiving any electronic journal article.

At this meeting it was decided that:

1. The project would be a set of “standard” XML elements and attributes that could be used to build article models.
2. Work should continue on the new models to expand them to handle any journal article content, including a survey of articles across many disciplines, to ensure that all article objects could be accommodated in the new model.
3. There should be two initial article models: one for existing content, providing a broad target for conversion of any article content, and one for creating new content, having a more prescriptive model to provide explicit rules for tagging content.

4. The new models should be easily extensible. For example, it should be easy to swap the OASIS CALS (Continuous Acquisition and Life-cycle Support) table model for the default XHTML table model.

The NLM DTDs

The National Library of Medicine (NLM) DTDs were created based on that initial meeting. Version 1 of the NLM Archiving and Interchange Tag Suite was released in early 2003 and included two article tag sets: the Archiving and Interchange model and the Journal Publishing model. The Archiving model was intended for tagging existing content, and the more prescriptive Publishing model was intended for authoring and tagging new article content (or article content that would be marked up in XML for the first time).

The intention of the NLM DTD project was to enable what publishers are already doing with their content rather than to define what they should be doing. In order to keep the suite relevant, because publishing practices are not static, the NLM assembled the Archiving and Interchange Tag Suite Working Group—a group of individuals who advised the NLM or recommended changes to the suite. The Working Group, responding to public feedback and, drawing from their experience, released several updated versions of the Tag Suite and the individual models over the next several years.

In 2005, with the release of the Tag Suite version 2.1, a new article model was introduced: the Article Authoring model. Between versions 1.0 and 2.0, the modifications to the Tag Sets had made models far more permissive, and the Working Group realized the Journal Publishing set was no longer suited for authoring new content. The Article Authoring model is the most prescriptive of the sets and is targeted toward new content creation.

Backward-compatibility is a significant factor in adoption of a new version of any tag set, so to facilitate the adoption of updated versions, the Working Group tabled all non-backward-compatible changes through the version 2.3 release. Concurrent with the release of version 2.3, the Working Group made the decision that the next major version release, version 3.0, would incorporate all of the non-backward-compatible changes that had been accumulating.

Involvement of NISO

The decision to make version 3.0 non-backward-compatible was part of the discussion about formalizing the Tag Suite with the National Information Standards Organization (NISO). The original plan had been to submit the latest version of the suite and models for registration, but because standardization would bring a lot of attention and new users to the suite, the Working Group chose to make the non-backward-compatible changes prior to registration.

Once version 3.0 was released in November 2008, the work of the NLM Archiving and Interchange Tag Suite Working Group concluded and the NISO Standardized Markup for

Journal Articles Working Group was created. Like the NLM Working Group before it, the NISO Working Group saw its role as normalizing and documented existing practices rather than dictating what should be done.

On March 30, 2011, after approval by the NISO Standardized Markup for Journal Articles Working Group and the NISO Content and Collection Management Topic Committee that oversaw the Working Group, NISO released NISO Z39.96, JATS: Journal Article Tag Suite, as a Draft Standard for Trial Use. Officially, this was NISO JATS version 0.4, but in essence it was a minor update to the NLM version 3.0 Tag Suite and article models. The draft standard was available for public comment until September 30, 2011.

The Working Group responded to each of the comments received and created JATS version 1.0, which was approved by NISO voting members and the American National Standards Institute as ANSI/NISO Z39.96-2012 in August 2012.

The Standard and the Supporting Information

ANSI/NISO Z39.96-2012 defines elements and attributes that describe metadata and full content of scholarly journal articles. It is not designed to describe magazines, books, or other publishing formats that may have some similar structures to journal articles but could also have significantly different structures.

The Tag Suite is the complete set of elements and attributes described in the standard. Along with these descriptions the standard includes three article models, or Tag Sets:

- The Journal Archive and Interchange Tag Set
- The Journal Publishing Tag Set
- The Article Authoring Tag Set

The Tag Suite has been designed to be extensible. Any of the tag sets may be extended or restricted to meet the needs of a given project. Also, new tag sets can be built from the elements and attributes in the Tag Suite and should be considered conforming to the standard.

Non-normative Information

The standard includes neither schemas nor much usage information. Non-normative supporting information, available from the NLM site, includes:

1. Schemas for each of the Tag Sets described above in three schema languages: DTD, W3C Schema (XSD), and RELAX NG.
2. Detailed “Tag Libraries” for each Tag Set that include the element and attribute definitions from the standard, remarks on usage, tagged examples, and detailed discussions of topics ranging from customizing a tag set to tagging names and dates.

3. A basic set of style sheets for rendering articles in HTML or in PDF through XSL-FO. These style sheets are intended as “starters” to be modified and personalized by each user.

Additional Schemas

The article models for NLM version 1.0 in 2003 were released only as DTDs. Beginning with version 1.1, WC3 Schema expressions of the Tag Sets were released along with the DTDs, and RelaxNG schema versions were added beginning with version 2.1. The additional schema languages were created from the DTD versions. Because the three languages have different features and limitations, the DTD version was declared as the version intended for maintenance and the other two as derivatives. This ensured that data tagged in one of the tag sets would be valid according to all of that Set’s schemas.

Tools

The NLM (NCBI) released tools for use with the NLM DTDs to the public. These tools include an XSL conversion to HTML for previewing NLM DTD content and an NLM DTD-to-XSL-FO conversion for creating PDFs. These basic tools are intended to be launching points for groups and it is expected that groups will customize these basic stylesheets for their own uses.

The public tools also include an XSL stylesheet that will transform data from any version prior to 3.0 into 3.0. This was released to help ease the transition to the non-backward-compatible version.

The Future of JATS

The plan with NISO is to maintain JATS continuously. Continuous maintenance is an option for American National Standards that allows comments and requests for enhancements to be submitted at any time, with a published regular schedule of when a Standing Committee will meet to evaluate such requests. When a sufficient number of substantive changes have been approved, a revision is balloted for approval and publication. (The alternative default option of periodic maintenance provides for a five-year review of the standard and, if a revision is deemed to be needed after such a review, a revision working group is initiated.) Continuous maintenance will allow revisions to be issued on a more timely basis and ensure ongoing interaction with the community that is using the standard. We look forward to working with users as the JATS grows to accommodate the needs of its growing user community.

References

- JATS standard (ANSI/NISO Z39.96-2012) jats.niso.org
- JATS supporting documentation <http://jats.nlm.nih.gov>
- JATS-Con Available at: <http://jats.nlm.nih.gov/jats-con/>

JATS-Con Proceedings Available at: <http://www.ncbi.nlm.nih.gov/books/NBK65129/>

JATS E-mail List Available at: <http://www.mulberrytech.com/JATS/JATS-List>

Inera, Inc. E-Journal Archive DTD Feasibility Study. December 5, 2001. <http://www.diglib.org/preserve/hadtdfs.pdf>

NLM Archiving and Interchange DTD, version 1.0 <http://dtd.nlm.nih.gov/archiving/1.0/>

NLM Journal Publishing DTD, version 1.0 Available at: <http://dtd.nlm.nih.gov/publishing/1.0/>

OASIS CALS table model Available at: <https://www.oasis-open.org/specs/tablemodels.php>

The NIH Manuscript Submission System

Abigail Acland^{✉1}

Created: March 15, 2013.

Summary

The NIH Manuscript Submission system (NIHMS) handles the submission of full-text manuscript material to PubMed Central (PMC) in support of the NIH Public Access and other related funding agency policies for material published in non-PMC participating journals. The NIHMS system allows users, such as authors, principal investigators (PIs), and publishers to supply material for conversion to XML documents in a format that can be ingested by PMC. Documents go through multiple stages in the conversion process; initial submission, grant reporting, initial conversion approval, staff QA, conversion, QA of converted material, Web version approval, and citation matching.

History

NIHMS was created in 2005 as part of the [NIH Public Access Policy](#). The policy calls for deposit of NIH funded research in PubMed Central. For authors whose articles were published in journals that are not submitting content to PMC directly, a method for submission was needed. NIHMS was created to allow users to deposit manuscript source files for conversion to PMC articles.

¹ NCBI; Email: aclanda@ncbi.nlm.nih.gov.

[✉] Corresponding author.

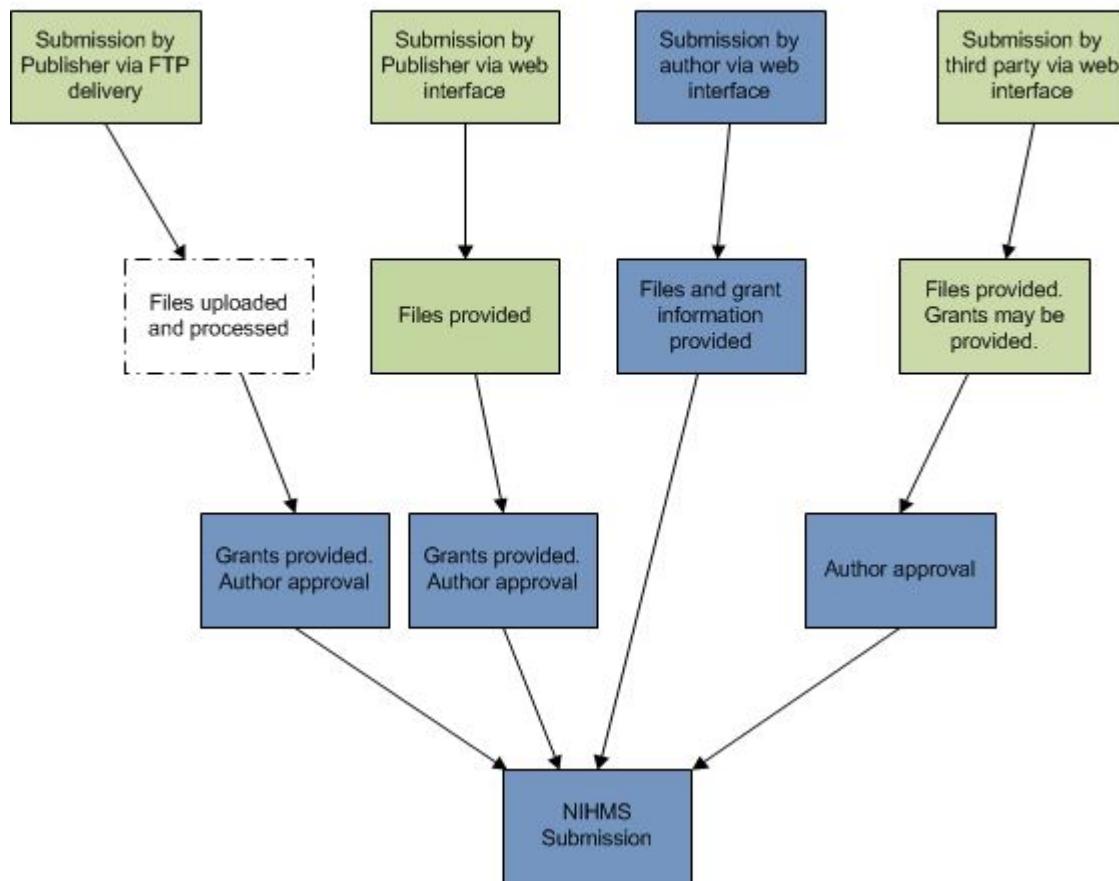
Overview of Submission Methods

	Method A Journal deposits final published articles in PubMed Central without author involvement	Method B Author asks publisher to deposit specific final published article in PMC	Method C Author deposits final peer-reviewed manuscript in PMC via the NIHMS	Method D Author completes submission of final peer-reviewed manuscript deposited by publisher in the NIHMS
Version of Paper Submitted	Final Published Article	Final Published Article	Final Peer-Reviewed Manuscript	Final Peer-Reviewed Manuscript
Task 1: Who starts the deposit process?	Publisher	Publisher	Author or designee, via NIHMS	Publisher
Task 2: Who approves paper for processing?	Publisher	Publisher	Author, via NIHMS	Author, via NIHMS
Task 3: Who approves paper for Pub Med Central display?	Publisher	Publisher	Author, via NIHMS	Author, via NIHMS
Participating journal/publisher	Method A Journals	Make arrangements with these publishers	Check publishing agreement	Make arrangements with these publishers
Who is Responsible?	NIH Awardee	NIH Awardee	NIH Awardee	NIH Awardee
To cite papers, from acceptance for publication to 3 months post publication	PMCID or "PMC Journal- In Process"	PMCID or "PMC Journal- In Process"	PMCID or NIHMSID	PMCID or NIHMSID
To cite papers, 3 months post publication and beyond	PMCID	PMCID	PMCID	PMCID

The policy was initially voluntary and took affect in May 2005. Compliance remained low in the months following. The policy then became mandatory in April 2008.

Ingest

NIHMS Ingest



NIHMS receives source files from a variety of submitters. The majority of submissions are entered as single manuscripts through a Web interface. Submissions may be started by anyone with an account in the system, but they must be approved by an author on the paper, and appropriate funding must be reported for the record.

A high percentage of manuscripts in NIHMS are started by publishers as a service to their authors.

Ingest QA

Once the material has been submitted to NIHMS, it must undergo a QA review by staff to ensure that the material is complete, suitable for conversion, within scope of the NIH Public Access Policy, and does not represent a duplicate submission. Staff also checks funding to ensure reported funding is applicable and that the approving party on the submission is an author on the paper.

Submissions that pass the staff QA evaluation will be sent on in processing for conversion. Items that do not pass QA may be blocked from further processing, merged with an existing record, or returned to the submitter or author for correction.

Additional automated checks are performed on the submission at this stage to ensure that the material is matched to a journal recognized by the NLM and does not fall within any PMC-participating journal's period of PMC participation. PIs holding associated funding who were not involved in the submission process are notified of the deposit of material associated with their grants at this time.

Conversion

Once these requirements have been met, the manuscript is sent to document conversion vendors for conversion to XML.

The conversion vendors return to NIHMS the full-text XML file for the manuscript, figure files in standard formats for thumbnail, Web, and PDF presentation, and any extracted supplemental files that may have originally been provided as embedded in the original source files.

XML in NIHMS

Regardless of source document format, all submitted manuscripts are converted to full-text XML using the NLM JATS (Journal Article Tagging Suite) Journal Publishing model.

The XML document is considered the converted document. The Web and PDF displays rendered from the XML are display versions of the document, but alterations and corrections are performed on the base XML or rendering software, rather than directly on the Web or PDF version.

Complex mathematical content that cannot be captured in plain JATS markup and display formulas are captured using MathML.

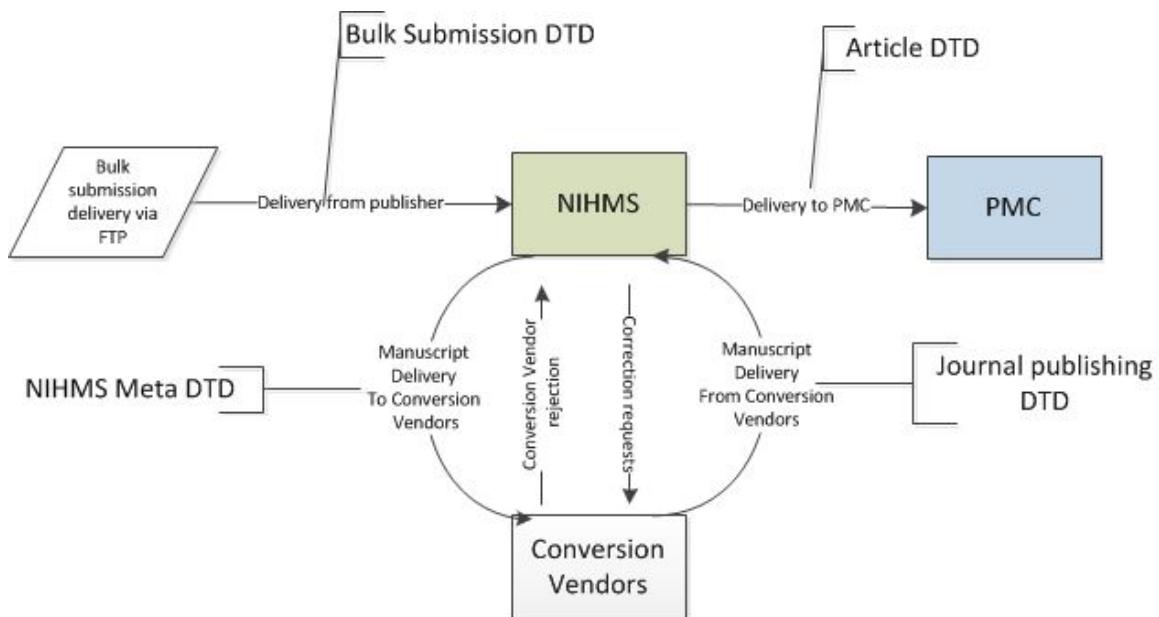
Source File	$p_{\lambda}(t) = \frac{1}{2}[\lambda^2 - (\lambda - t)_+^2]$
MathML	<pre><mml:math id="M19" display='block'> <mml:msub> <mml:mi>p</mml:mi> <mml:mi>&#x003BB;</mml:mi> </mml:msub> <mml:mo stretchy='false'>(</mml:mo> <mml:mi>t</mml:mi> <mml:mo stretchy='false'>)</mml:mo> <mml:mo>=</mml:mo></pre>

Table continues on next page...

Table continued from previous page.

<pre> <mml:mfrac> <mml:mn>1</mml:mn> <mml:mn>2</mml:mn> </mml:mfrac> <mml:mo stretchy='false'>[</mml:mo> <mml:msup> <mml:mi>&#x003BB;</mml:mi> <mml:mn>2</mml:mn> </mml:msup> <mml:mo>&#x02212;</mml:mo> <mml:msup> <mml:mrow> <mml:mo stretchy='false'>(</mml:mo> <mml:mi>&#x003BB;</mml:mi> <mml:mo>&#x02212;</mml:mo> <mml:mi>t</mml:mi> <mml:mo stretchy='false'>)</mml:mo> </mml:mrow> <mml:mo>+</mml:mo> <mml:mn>2</mml:mn> </mml:msup> <mml:mo stretchy='false'>]</mml:mo> </mml:math> </pre>	
Rendered Equation	$p_\lambda(t) = \frac{1}{2} [\lambda^2 - (\lambda - t)_+^2]$

XML in Communication

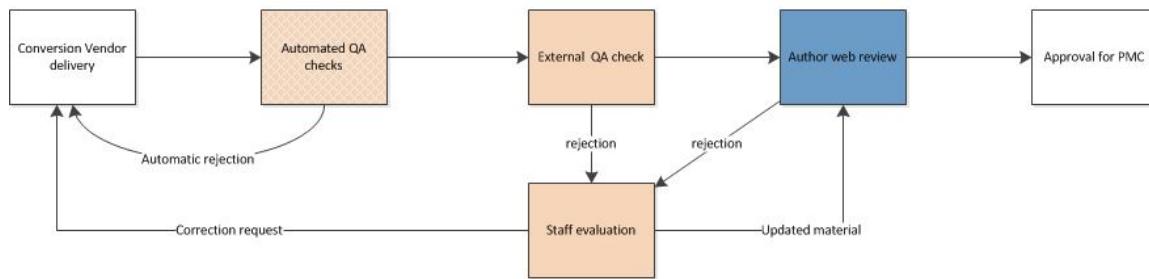


XML is also used for much of the communication and data handoffs between systems. Incoming deliveries from publishers submitting via FTP delivery include metadata documents in XML which contain information about the article needed to start a manuscript record in the system.

Packages sent to the conversion vendors include metadata documents that contain system-specific information that must be included in the tagged manuscript.

QA of Converted Materials

QA In NIHMS



Converted material in NIHMS undergoes several QA steps by various groups before it may be sent to PMC.

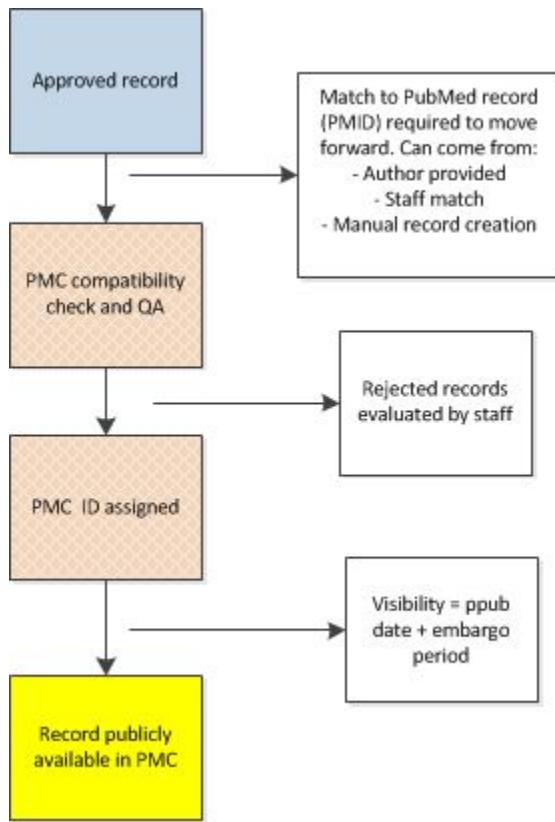
The first QA step consists of automated checks for validity and conformity to NIHMS standards of files. Failures at this level are returned directly to the conversion vendors. When the delivery passes these checks, the system generates HTML and PDF versions of the manuscript.

Once these display versions have been generated, the manuscript is sent to an external QA team that confirms that no errors were introduced in either the conversion or rendering processes. Manuscripts that fail this outside evaluation are directed to NIHMS staff for further investigation. Manuscripts that pass this evaluation are sent to the author for review and approval of the converted material.

The assigned reviewing author for the manuscript reviews the display versions (HTML and PDF) and may either accept the manuscript as is or reject it with comments. Authors may report not only errors in conversion or rendering, but request any necessary updates to the material to ensure scientific accuracy.

Material rejected by either external QA or authors is evaluated by NIHMS staff. Staff may return material to the conversion vendors for updates, request additional files from the submitter or author, make manual updates, or request rendering updates in the NIHMS system.

Additional Processing



Delivery to PubMed Central and subsequent assignment of a PubMed Central ID number (PMCID) is not automatic after author approval of the NIHMS Web version. The record must also be matched to a PubMed citation to confirm publication and calculate release dates in PMC. This is fairly straight forward for records started from a PMID match by the authors, or deposited with trusted publication information from the publisher that may be used to perform an automated match. For records not matched by these methods, staff must either make a match based on suggested matches by the system, manual PubMed searches, or matches provided by the author separately from the submission of the manuscript.

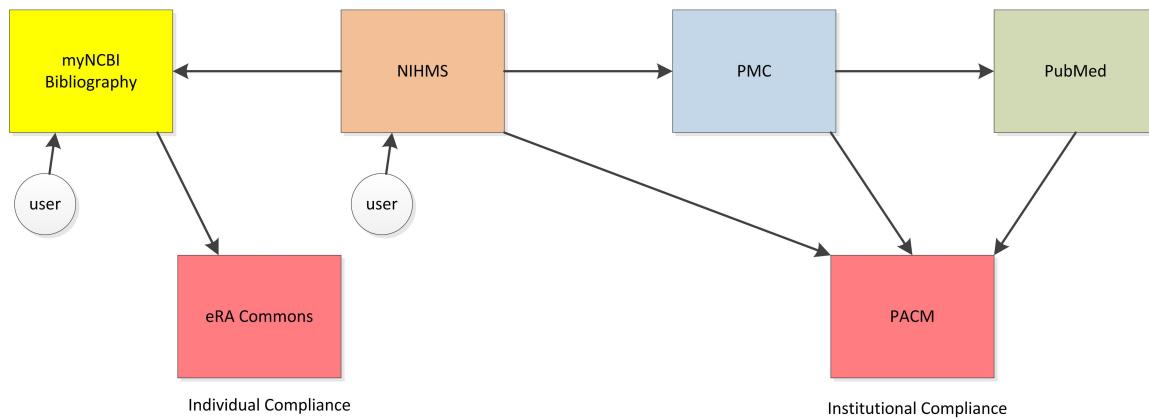
For records from journals that are not indexed in MEDLINE, a manual citation must be created for the NIHMS record. This involves National Library of Medicine team members reviewing the manuscript record to match the material to information in a variety of external citation indexing sites or publishers' websites to confirm publication information. In rare cases, for journals not indexed in MEDLINE that do not have a publicly-accessible online presence, we may request archive documents from the author in order to verify the citation.

Once the Web version has been approved by the author and matched to a citation record, the record may be sent to PubMed Central. PMC then performs additional automated QA

checks on the submission. If the manuscript passes these checks, a PMCID is generated and assigned to the record. If the record fails these checks, it is returned to NIHMS staff for evaluation. Records may also be returned to NIHMS staff following regular PMC integrity checks.

Records in PMC will remain “hidden” until at least the final publication date of the record, plus any additional embargo period specified by the submitter or author.

Reporting Systems And NIHMS



In most cases, records created in NIHMS will generate a PMCID, which demonstrates full compliance with the NIH Public Access Policy, although some records in NIHMS are created only to report the association of a grant with a particular paper. NIHMS IDs are an important aspect of both documenting the relationship between grants and articles and the compliance status of those articles.

The NIH Public Access Policy calls for submission of records for non-PMC journals to NIHMS at the time of manuscript acceptance; there is no grace period in the requirements. Lack of an NIHMS ID for applicable papers will cause authors to be reported as non-compliant. NIHMS IDs may be used by NIH-funded parties as evidence of compliance with the Public Access Policy for up to 90 days after the print publication date of the article, after which a PMCID must be used. An NIHMS ID is not permanent evidence of compliance.

However, NIHMS does not report compliance directly. The system supplies grant and paper associations to a variety of resources and provides status information of NIHMS IDs and PMCID when assigned. From these, resources compliance monitoring systems—such as PACM, which provides compliance reports to institutions, and eRA Commons, which provides compliance reports for an individual—calculate article compliance status with respect to the NIH Public Access Policy.

Genomes

What's in a Genome at NCBI?

James Ostell, PhD¹

Created: November 8, 2013.

Scope

It seems like a simple request to a bioinformatics center like NCBI—"Download the human genome", "Display the *HIV-1* genome"—and yet this is a complicated question in terms of biology, experimental data, the current state of knowledge, and the use to which a particular scientist may wish to put the data. This chapter provides an introduction to those questions, a brief history of genome representations at NCBI as the state of the science evolved over the last few decades, and a summary of some of the many resources and tools that are relevant to Genomes at NCBI.

What is a Genome?

Biologically speaking, of course, a "genome" of an organism is the complement of genes that make up the organism. However, this is already an abstraction. Traditionally "genes" are heritable units that manifest some observable trait over related generations of an organism. It was only later we discovered they are pieces of DNA contained in a larger DNA strand comprising a chromosome, leading to the current assumption that a genome is the set of DNA strands of the chromosomes. However, we already know this isn't quite right, because any particular individual from whom we measured the DNA may not be typical in every gene, in fact, may be completely missing some genes that are found in other individuals of the same species. In addition, there may be DNA sequences that are found in NO real organism (because they are chimeras from many individuals or because they contain errors), which are nonetheless the standard reference for a particular community, and thus the sequence they expect to find when they ask for e.g., "the human mitochondrial genome" (also known as "the revised Cambridge Reference Sequence") (1, 2).

In addition to these biological and historical issues, science is a moving target. At various points in time, the genome data for a particular organism may be incomplete to varying degrees, in which case the best "genome" available may still have unique difficulties. Over time as technology changes, a different set of data may become a better approximation for the biological reality, then the community may suddenly find itself getting a different answer for "download the genome" than they got last time.

Finally, there are many contexts for use of "the genome." For many types of research, "the genome" doesn't mean DNA at all but instead means the set of transcripts or protein

¹ NCBI.

sequences coded for by that organism's genome. In other cases, for example medical genetics, it means not only DNA, but a very specific piece of reference standard DNA that is used by the medical community to record single base changes at specific locations (3). This important piece of DNA may only cover one gene, and for that length, it may not be identical to the commonly used complete "human genome" sequence in that region. Still other contexts involve investigations of large and small sequence variation across a population of individuals, by browsing or downloading data for the [human 1000 Genomes](#) project (4), monitoring sequence variation during an [influenza outbreak](#) (5), or comparative analysis datasets as provided in [HomoloGene](#) (6), [Protein Clusters](#) (7), or the viral [Pairwise Sequence Comparison \(PASC\)](#) tool (8).

So, NCBI's efforts to provide a response for "give me the genome of my research organism" has necessarily had to change over time, may be different for different organisms, and may bring complications and nuances that surprise users who may only be familiar with their particular view of what "the genome" is.

History

The history of genomes at NCBI is intimately tied to the history of [GenBank](#) and [RefSeq](#). GenBank is part of the [International Nucleotide Sequence Database Collaboration \(INSDC\)](#) (9) of which NCBI is the US member, which collects annotated nucleotide sequences from contributing scientists, typically when they publish the sequence in a journal article. As such, GenBank is like the primary research literature. Each "article" or sequence entry represents the view of the contributing author at the time. The database staff does not curate the sequence records for correctness other than conformance to standards and internal consistency. The biological validity of the record is reflected in the peer review process of the published article. Just as with the primary literature, articles can go out of date or later articles may contradict, or even invalidate, a previous article. There may be legitimate disagreement about which view is more correct. But this evolution of scientific knowledge over time is normal for any experienced scientist reading the literature, and it is normal and healthy for GenBank.

NCBI's initial expectation was that we would simply identify the GenBank record which was currently the most widely accepted genome sequence for a particular organism when someone requested to view or download "the genome." However, as early as the 1990's it was clear there were problems with this. At that time the *HIV-1* "genome" had been published (in fact several times) in GenBank. The problem was that these were the first sequences, and they were partial in different ways, in particular missing the long terminal repeat (LTR). So, scientists working on *HIV-1* had developed and were sharing a "standard" *HIV-1* genome by editing and combining the partial GenBank sequences into one complete genome, but it had never been deposited into GenBank. So, it was not possible to provide any single GenBank record in response to the request "download the *HIV-1* genome." In an attempt to remedy this problem, NCBI staff worked with the authors of an authoritative book on retroviruses, to both deposit the retrovirus genome sequences into GenBank (e.g., AF033819.3), and to cite those sequences when referring to

the genomes in the [RETROVIRUSES](#) book, which is available on the NCBI [Bookshelf](#). While this solution continued to use GenBank as the vehicle for authoritative genomes, it was no longer a pure model of the data coming voluntarily from the scientific community. NCBI had a guiding and active role to make it happen.

Shortly after that the yeast genome was being completed chromosome by chromosome. Each chromosome was sequenced by a different US or European group, and published as a submission of a chromosome, a chromosome arm, or a collection of contigs and scaffolds. It was only at the time of publication that the chromosome sequence was deposited and shared through GenBank. This was a long slow process, especially when the work on one chromosome might be the result of the work by many separate labs working at different rates. During this time, it was only possible for NCBI to provide a partial answer from GenBank to “download the yeast genome.” Once the whole genome was completed, however, much work, especially in annotating the genes, but also in correcting sequence problems, continued. The scientists in the US who did the initial work on half of the yeast chromosomes “ceded” their GenBank records to a single database group, the [Saccharomyces Genome Database \(SGD\)](#) (10). This gave SGD the right to update “their” GenBank records as annotation improvements were made and sequence problems corrected. SGD worked with NCBI to make this happen, and that half of the yeast genome was kept current and up to date in GenBank. Unfortunately, the European half of the genome was not held under this model, so even though annotation improvements and sequence updates were being made to those chromosomes, they were not being deposited into GenBank. The result, again, was the NCBI was unable to respond with a complete up-to-date set of records when someone wanted to “download the yeast genome.” After much negotiation to make this approach work through GenBank, it became clear that half of the yeast genome in GenBank would always be out of date. In this same time-frame, the human genome project was actively underway and again sequencing was completed and submitted chromosome by chromosome but by this time the “Bermuda Principles” (11) had been established and data was now starting to be submitted in advance of peer-reviewed publication. The international community was actively discussing the human gene count and planning for genome annotation; during this time several groups attempted to define the gene content before the sequencing was completed (12, 13).

Thus, the climate at that time with regards to both data submission and scientific discussion were instrumental to the decision that NCBI needed a new database, called RefSeq (14). Just as you may think of GenBank as equivalent to the primary research literature written by many authors, RefSeq could be considered the “NCBI review article” on genomes. Unlike GenBank, it reflects the judgment of a single group, NCBI, about what the most useful sequences are to represent a particular genome and its products, typically in collaboration or consultation with experts wherever possible. Like a review article it is drawn from the primary literature and archival sequence databases, but it may be aggregated, edited, or reorganized by NCBI to represent a better summary overview in light of current knowledge. RefSeq included the *HIV-1* genome from GenBank with no annotation or sequence changes compared to the GenBank record. But it could also now include human transcripts and proteins (organized by genes in LocusLink (14), the

precursor to NCBI's Gene resource), in preparation for annotating the human genome. As for the *Saccharomyces cerevisiae* genome, for many years the RefSeq annotated genome—which was taken entirely from SGD so that it would be complete, consistent, and up to date—was the only way that NCBI could answer the request to view or download all chromosomes of the yeast genome. Much more recently, the complete yeast genome was also made available, by SGD, as a Third Party Annotation submission so that the current up-to-date version is also available at all member databases of the INSDC.

The RefSeq initiative allowed NCBI to start building representations of genomes that might vary considerably in how they were built or the sources they came from. For example, when the first bacterial genomes were sequenced, NCBI took a chromosome-centric view of the organism because we had the whole DNA sequence, but not much information on transcripts or proteins. In contrast, for human, at the time there were many cDNA sequences of transcripts from individual genes, but only relatively short stretches of chromosomal DNA for a few regions. So “the human genome” in RefSeq was actually the most comprehensive collection of cDNAs we could aggregate and curate at the time. During the early and middle phases of sequencing and assembling the human genome, the cDNA sequences were still the highest quality data for the genes, but the assembled chromosome fragments started filling in the intragenic regions and providing long range order. In this version, “download the human genome” would get chromosomal sequence and aligned cDNAs, where the sequence of the cDNA may not match exactly the sequence of chromosomal DNA. Because the cDNA is more reliable, the protein from the coding region is derived from the cDNA, not the chromosome. Today, the chromosomal sequence of human is very good quality (although not without flaws), and NCBI has gone to considerable lengths to ensure that the sequence of chromosome and the cDNA do match, sometimes correcting the chromosome, sometimes correcting the cDNA, in collaboration with sequencing laboratories and other scientists.

While “the human genome” is now very consistent across chromosomal DNA and cDNA, there remain cases where there are exceptions, such as when the human genome represents a rare allele or a base that is suspected of being an error. NCBI is one of the collaborators in the [Genome Reference Consortium](#) (15) which is actively involved in ongoing maintenance and improvement of the reference human genome sequence. NCBI's [RefSeqGene](#) project (16) is another case where there may be differing definitions of the reference genomic sequence. RefSeqGene provides a collection of chromosomal DNA regions, in chunks covering single genes, which are intended for use in clinical genetics. Since the traditional sequence for some genes predates the complete human genome sequence, or where the gene in the complete human genome may not be a common allele or traditional allele, the RefSeqGene record may not be identical to “the human genome.” RefSeqGene genomic records are aligned, as with the cDNAs in the middle period of human genome sequencing, to the human genome to support (using the [NCBI Genome Remapping Service](#)) mapping coordinates and sequence back and forth between the two. NCBI has made every effort to make the RefSeqGene identical to the human genome when it can support the needs of the medical reporting community, but it

is not totally consistent. So the “complete human genome” in the research world may not be exactly the same thing as “the complete human genome” in the clinical world.

The evolution continues as we start to accumulate many genomes for a single organism. If one asks for the “*Salmonella* Genome” today, the question is which strain of *Salmonella* do you really want? In some cases you want the well annotated typical genome, but in other cases you may specifically want the one from “the Montevideo outbreak.” Even with humans, we know that some humans contain genes not contained in other humans and vice versa. So by “the human genome” do you want a single example of a real human genome? Or do you want that single example, plus the additional genes that are found in other humans, to get the full complement of possible human genes?

Finally, in some aspects, RefSeq is coming full circle back to GenBank. NCBI has gone to considerable efforts to persuade the scientific community to keep GenBank up to date directly themselves, so that RefSeq can again be simply a selection of particular records. One case in point is *Drosophila*. The *Drosophila melanogaster* genome records were “ceded” to the [FlyBase](#) database by the original sequencing team. Unlike yeast, in this case FlyBase “owns” all the chromosomes. NCBI has worked closely with FlyBase both to provide computational support and evidence for building and validating the gene models, and for facilitating the update of the GenBank records from the FlyBase database, and maintaining the RefSeq records from those GenBank records. FlyBase provides valuable communication within the *Drosophila* community and manual curation of gene models by experts in that organism.

RefSeq necessarily continues to contain “genomes” for different organisms done in a variety of possible ways, with a variety of possible consequences. But for the goal of comprehensiveness, one necessarily pays the price of complexity. NCBI makes every effort to provide simple, intuitive views of genomes where possible, while still hinting at the additional layers and nuances to the genome concept so users who may need that more sophisticated view are aware it exists.

Resources, Tools, and Access

NCBI’s early commitment to reliably and robustly support the simple request to download, view, or analyze a genome resulted in a large suite of resources, tools, and shareable code-base (via NCBI [toolkit](#) libraries) that range from broadly scoped multi-kingdom resources such as [RefSeq](#), [Gene](#), and [Genome](#), and eukaryotic and prokaryotic genome annotation pipelines—to niche resources such as [Viral Variation](#) or [CloneDB](#). Some resources reflect natural organizing principles of genomic data and support access from a gene- or organism-centric perspective, whereas others were developed in response to a particular disease outbreak ([NCBI FLU resource](#)) or based on collaboration, community feedback or requests (e.g., [International Standards for Cytogenomics Array](#), [HIV-1:human protein interactions](#), [Conserved CDS database](#)). Along the way we also developed a suite of viewing platforms including [Map Viewer](#), the Graphical sequence viewer (for example, the RefSeq *Escherichia coli* genome record [NC_000913.3](#)), and a

standalone, multiplatform, downloadable graphical user interface (GUI) [Genome Workbench](#). NCBI continues to be fully committed to supporting access to genome data and several of our newer resources—[BioProject](#), [BioSample](#), and [Assembly](#)—describe aspects of the research project, the biological sample, and how the genome assembly is organized.

References

1. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, et al. Sequence and organization of the human mitochondrial genome. *Nature*. 1981;290(5806):457–65. PubMed PMID: 7219534.
2. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics*. 1999;23(2):147. PubMed PMID: 10508508.
3. Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, et al. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med*. 2010;2(4):24. PubMed PMID: 20398331.
4. Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–73. PubMed PMID: 20981092.
5. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the National Center for Biotechnology Information. *Journal of virology*. 2008;82(2):596–601. PubMed PMID: 17942553.
6. Database resources of the National Center for Biotechnology Information. *Nucleic acids research*. 2013;41(Database issue):D8–D20. PubMed PMID: 23193264.
7. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, et al. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic acids research*. 2009;37(Database issue):D216–23. PubMed PMID: 18940865.
8. Bao Y, Chetvernin V, Tatusova T. PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. *Viruses*. 2012;4(8):1318–27. PubMed PMID: 23012628.
9. Nakamura Y, Cochrane G, Karsch-Mizrachi I.; The International Nucleotide Sequence Database Collaboration. *Nucleic acids research*. 2013;41(Database issue):D21–4. PubMed PMID: 23180798.
10. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. *Saccharomyces Genome Database*: the genomics resource of budding yeast. *Nucleic acids research*. 2012;40(Database issue):D700–5. PubMed PMID: 22110037.
11. Marshall E. Bermuda rules: community spirit, with teeth. *Science* (New York, NY. 2001;291(5507):1192.
12. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, et al. A physical map of 30,000 human genes. *Science* (New York, NY. 1998;282(5389):744–6.
13. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature genetics*. 2000;25(2):239–40. PubMed PMID: 10835646.

14. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* 2000;16(1):44–7. PubMed PMID: 10637631.
15. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. *PLoS Biol.* 2011;9(7):e1001091. PubMed PMID: 21750661.
16. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research.* 2012;40(Database issue):D130–5. PubMed PMID: 22121212.

Eukaryotes

Clone

Valerie Schneider, Ph.D.¹

Created: November 14, 2013.

Scope

The NCBI [Clone DB](#) is a database that integrates information about eukaryotic genomic and cell-based clones and libraries, including sequence data, genomic location, and distribution sources (1). At Clone DB, users can find library metadata, search for clones containing genes or other sequences of interest, or find contact information for distributors of libraries and clones. In addition, Clone DB provides mapping data that can be used to help researchers assess and improve genome assemblies. Although Clone DB is a resource whose aim is to help users connect data with physical clone reagents, NCBI is not itself a distributor of libraries or clones. The database contains library and clone records for over 150 taxa, is indexed in Entrez, and can be searched by many terms, including clone, library or gene name, organism or sequence identifier. Clone DB maps genomic clones to reference assemblies when such data is available. These placements can be viewed as graphical displays in the clone records themselves, as well as in the NCBI [Clone Finder](#), where clone placements can be searched by location, genome features, or transcript names.

Clone DB maintains records for genomic and cell-based libraries and clones that are available from commercial or academic distributors, along with a limited collection of clone libraries of sufficient scientific significance to warrant their representation even in the absence of distribution. At this time, Clone DB contains over five hundred genomic library records that represent more than 150 different eukaryotic taxa, which include both animal and plant species. The current Clone DB collection of records for cell-based clones includes gene trap and gene target libraries produced by the International Knock-out Mouse Consortium (IKMC) (2, 3) and International Gene Trap Consortium (IGTC) (4), as well as the Lexicon Genetics gene trap collection (5). These libraries and their associated metadata are provided to Clone DB by [Mouse Genome Informatics](#) (MGI). Genomic library records in Clone DB include the original set of libraries imported from the former NCBI Clone Registry database, as well as additional library records generated by database curators. Curators continue to update the database with new library records, emphasizing representation for genomic libraries that contribute to the generation of reference assemblies, are extensively end or insert sequenced or fingerprinted, as well as libraries whose representation is specifically requested by users contacting the Clone DB (clonereg-admin@ncbi.nlm.nih.gov).

¹ NCBI; Email: schneiva@ncbi.nlm.nih.gov.

History

Clone DB replaces and extends upon the former NCBI Clone Registry. The Clone Registry was developed during the Human Genome Project (HGP) as a resource to assist the many large-scale sequencing centers involved in this effort track the sequencing of clones in the tiling paths for the human and mouse reference assemblies. Importantly, the Clone Registry developed the notion of a standardized clone naming system that could be used to unambiguously identify clones from different libraries. This naming system was adopted by many of the sequencing centers and facilitated the consolidation of clone data from various and disparate databases. Clone Registry records contained information about sequencing status, links to end and insert sequence records, mapping locations, and clone distributors. Although the Clone Registry later grew to include records for genomic clones from other eukaryotic taxa, these generally lacked the depth of the human and mouse records, because of the relative paucity of genomic data. As the HGP drew to a close, it was clear that the Clone Registry would need to evolve in order to remain a relevant resource.

Clones continue to play an important role in biological research in the current era of next generation sequencing technologies, though their specific uses have changed. Although whole genome sequencing (WGS) has largely obviated the use of clone tiling paths in the generation of genome assemblies, genomic clones are still among the best means for resolving sequences in complex regions. Clone end alignments are used to assess assembly quality, and the technique of end-sequence profiling has proven to be a valuable means for discovering genomic variation (6, 7). In organisms in which large amounts of repetitive content or variation confound WGS assemblies, genomic clones remain the sequencing reagent of choice (8). Cell-based clones, such as gene trap and gene targeting clones, are used in the study of many model organisms to define gene function and study genotype-phenotype relationships (2, 9-11). In all these instances, clones may be associated with a variety of data types. Thus, there remains a need for a clone-focused database that can consolidate this information and assist users in obtaining these important biological reagents. Clone DB now serves this function, providing users with a central location to access information about library construction details and clone sequence, gene content, map location, and distribution.

Data Model

The data objects in Clone DB represent physical objects. Thus, the physical attributes of clones and libraries inform the Clone DB data model (Figure 1).

Libraries

Functionally speaking, a library is a collection of clones, and Clone DB utilizes this hierarchical relationship in its data model. In Clone DB, all clone records must be associated with a library record. Within a species, each library is uniquely identified in Clone DB by its library name. Metadata collected for genomic clone libraries includes

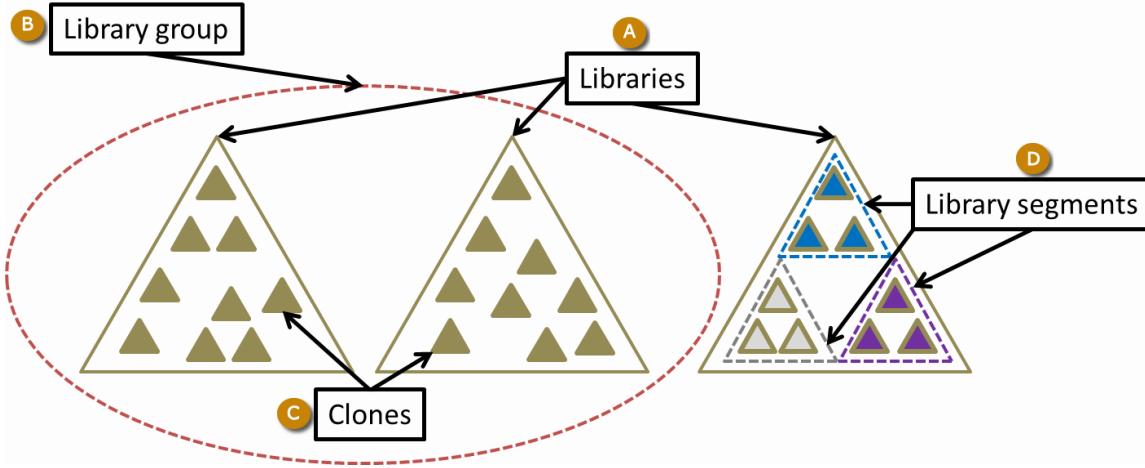


Figure 1. Clone DB data model. A: Three different libraries are shown (large triangles). B: Libraries sharing user-defined attributes may be assigned a library group. C: Clones (small triangles) are always associated with a library. D: Library segments distinguish subsets of clones within a single library that share different sets of common attributes. Figure taken from (1).

details about source DNA, library construction, library statistics, alternate names and abbreviations by which the library is known, and library distributors, as well as publications describing the construction or end sequencing of the library. Metadata captured for cell-based libraries includes the library creator, cloning vector, parental cell line, parental cell line strain, and allele type.

Library Groups

Within the database, library groups may be defined for collections of libraries that share one or more common features. Any common feature may be used as the basis for the creation of a library group. For example, murine cell-based libraries generated as part of the International Knock-Out Mouse Consortium (IKMC) belong to the same library group.

Library Segments

Within a single library, there may be subsets of clones that are distinguished by different sets of common attributes. Such attributes may include, but are not restricted to, cloning vector, vector type, source DNA, or average insert size. Clone DB uses the notion of library segments to capture any such subsets within a library record. Displays for library records report both the features common to the library as a whole, as well as segmental differences. For example, the record for the [Caltech Human BAC Library D](#) in Clone DB has 5 segments (Figure 2). All of the segments share a common DNA source and cloning vector, but the DNA for clones in one segment were digested by and cloned into the site for a different restriction enzyme. The 5 segments of this library, which were defined by the library creators, are distinguished by different average insert sizes. Within the physical

Annotation A: Library segment ALL. Sex: male. Cell type: sperm.

Annotation B: Library segment 1-2. Vector Name: pBeloBACII. Vector Cloning Site(s): HindIII, EcoRI.

Annotation C: Library segment 2-5. Vector Name: pBeloBACII. Vector Cloning Site(s): HindIII, EcoRI.

Annotation D: Library segment 1-5. Avg insert(kb): 129, 202, 182, 142, 166. Plate range(s): 2001 to 2423, 2501 to 2565, 2566 to 2671, 3000 to 3253, 3254 to 4869.

Library segment	Vector Name	Vector Cloning Site(s)
1	pBeloBACII	HindIII
2-5	pBeloBACII	EcoRI

Library segment	Avg insert(kb)	Plate range(s)
1	129	2001 to 2423
2	202	2501 to 2565
3	182	2566 to 2671
4	142	3000 to 3253
5	166	3254 to 4869

Figure 2. Details from genomic library record for Caltech Human BAC library D, which has 5 segments. Clones in all segments are derived from the same DNA source (A) and were cloned into the same vector (B). Clones in the 5th segment were cloned into a different restriction enzyme site (C) and the average insert sizes for clones in the 5 segments are all different (D).

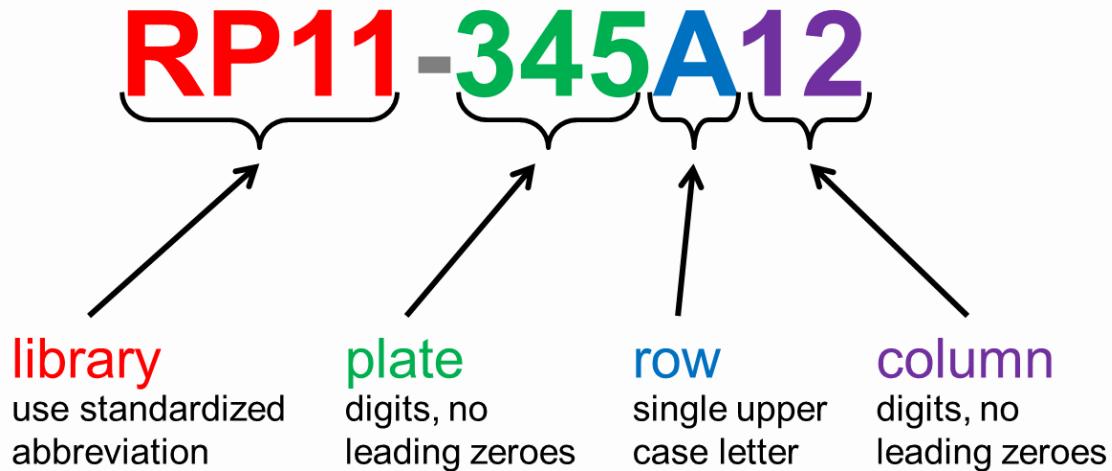


Figure 3. Clone DB standard nomenclature for genomic clones. Standardized library abbreviations are unique for each species in Clone DB.

library, these segments correspond to different sets of numbered microtiter dishes (plate ranges).

Clones

Clone DB also includes records for individual clones. As noted above, all clone records must be associated with a library record. Although library records are representations of physical libraries, Clone DB does not necessarily maintain records for all of the physical

clones that are associated with a particular physical library. Instead, records are only created for those clones for which there are sequence or mapping data to be represented in Clone DB.

A major feature of each clone's record is the clone name assigned by Clone DB. The assignment of names to genomic clones often presents a challenge, as it is common to find that different submitters have provided different permutations of a clone name on different data submissions representing the same clone object. Whenever possible, Clone DB attempts to parse submitter-provided names and assign a standardized name comprised of the clone's microtiter plate address (plate number, row and column), prefixed by a Clone DB library abbreviation to each record (Figure 3). In such cases, the submitter-provided name will be stored as a searchable alias of the standard name. If a standard name cannot be parsed from the submitter-provided name, the submitter-provided name will be assigned as the clone name. All names and aliases associated with genomic clone records are indexed and can be used as search queries. In the case of murine cell-based clones, Clone DB simply adopts the clone name provided by MGI.

Dataflow

Record Data

Metadata associated with records in Clone DB are supplied by different sources. On a weekly basis, MGI provides the clone library, gene, allele, and sequence identifier information, as well as creator and distributor details, for all murine cell-based clone records in Clone DB. In contrast, data affiliated with genomic clone records are derived from a variety of NCBI databases and external providers. Clone DB queries the [Nucleotide](#) database daily to retrieve high throughput genomic (HTG) insert sequences and their associated metadata for all organisms that have at least one library represented in the clone database. End sequences and their metadata are retrieved from both the [GSS](#) and [Trace Archive](#) databases via library-specific queries on an ad-hoc basis. Fingerprint data for genomic clones has been obtained from the FPC database maintained by the Michael Smith Genome Sciences Centre in Vancouver, Canada (<http://www.bcgsc.ca/data/data>) and The Genome Institute at Washington University, St. Louis. Likewise, data for cytogenetic map positions and STS markers mapped to human genomic clones are taken from the work of the [BAC resource consortium](#) (12) and [National Cancer Institute's Cancer Chromosome Aberration Project](#) (13).

Record Curation

All new genomic library records are generated and loaded to Clone DB by database curators. Weekly queries produce reports that identify insert and end sequences without corresponding libraries for organisms already represented in Clone DB. Curators also perform literature reviews to identify additional organisms for which record creation may be needed. Furthermore, new library records may be added at the behest of library creators, distributors, and database users by contacting Clone DB (clonereg-

admin@ncbi.nlm.nih.gov). All queries to retrieve end sequences from the Trace Archives and GSS are also defined by curators. In addition, curation is performed to address issues with library and clone record metadata that are identified by automated processes that check for data integrity. For example, retrieved insert and end sequence records that contain clone names that cannot be associated with existing clone records, or from which new standardized clone names cannot be parsed and created, are flagged for curatorial review. Likewise, externally provided data receives curatorial review to ensure consistency with the Clone DB data model. End sequence and clone placements are also reviewed by curators with respect to size, number, and concordance to ensure that the NCBI clone placement algorithm is producing results consistent with published library and genome characteristics.

Genomic Clone Placements

Clone DB maps genomic clones to assemblies that are annotated by the NCBI eukaryotic genome annotation pipeline. As a genomic clone is a physical object containing a specific fragment of DNA, this mapping provides context for the clone with respect to its genomic origin and to other clones in the same and other libraries. At the time of this handbook's writing, the NCBI clone placement algorithm only uses end sequences to create clone placements. In the future, these placements may also be informed by insert sequences. End sequences associated with clone records are screened to remove vector contamination and low quality bases. The set of processed ends is aligned to the genome assembly of interest with NG Aligner, an NCBI BLAST-derived aligner (see Genome Workbench chapter). In most cases, the ends are aligned to a genome representing the same species as the clone library. However, in instances in which no such genome is available, ends may be mapped to the genome of a closely related species.

The data flow for the generation of clone placements by Clone DB is illustrated in Figure 4. Clone placements are created by pairing end placements representing opposite ends of the same clone. The NCBI clone placement algorithm uses two mechanisms to minimize self-overlapping clone placements and present users with the most likely placement(s) for each clone. First, the algorithm clusters any overlapping end placements for a given clone end and selects only a single prototype for use in clone placements. The prototype is the end placement that holds the 5'-outermost position on the scaffold to which it aligns. Second, if the end placement prototypes contribute to a set of self-overlapping clone placements, the algorithm uses a set of defined heuristics to select a single clone placement from the set as an archetypal placement. A clone may therefore have more than one archetypal placement, but these may not overlap. Only archetypal placements are reported and displayed in clone records.

Clone DB defines an average insert size and standard deviation for each library based on its clone placements. It should be noted that these values, which are provided in reports on the Clone DB [FTP](#) site, may differ slightly from the library creator-provided values reported in the library record displays. As the latter values are commonly defined using

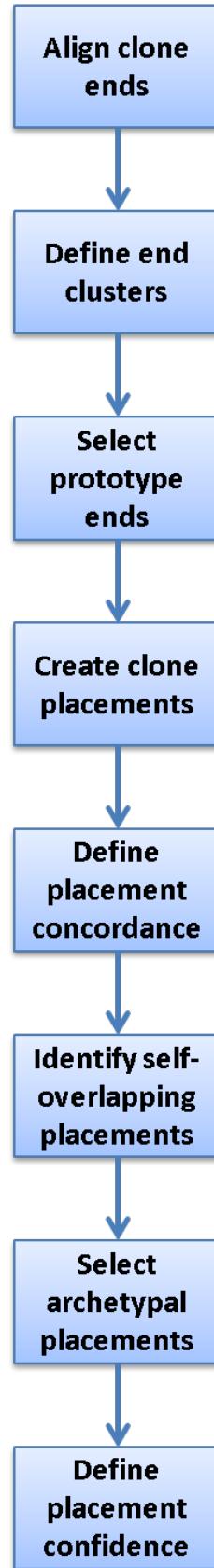


Figure 4. Diagram illustrating data flow for genomic clone placements in Clone DB.

techniques other than end mapping (e.g. gel sizing), some discrepancies are to be expected.

Clone DB defines the average insert size and standard deviation for each library using only the following subset of clone placements on an assembly:

- the placement is comprised of 1 forward and 1 reverse end
- both ends are uniquely placed
- both ends are placed on the same assembly scaffold
- end placements face each other
- placement length is between 50-500 kb (BAC/PAC) or 10-100 kb (fosmids).

Clone DB defines concordant placements as those in which:

- placement length is within 3 standard deviations of the library average
- contributing end placements are facing one another on opposite strands
- Any clone placement not meeting either of these criteria is defined as discordant.
No placement-related data is provided for clones for which no placements are found.

Clone placements are also assigned a [category](#) reflecting the level of confidence in the placement. Confidence assignments are reported in the “Clone Placement” tab of individual genomic clone records.

- **Unique:** There is a unique placement for the clone within an assembly unit. All end placements associated with the clone support this placement.
- **Unique-dissent:** There is a unique placement for the clone within an assembly unit, but there are end placements that do not support the clone placement.
- **Multiple-conflict:** There are multiple placements for the clone in an assembly unit; every clone placement is comprised of two uniquely placed ends and the multiple placements are due to non-overlapping end clusters.
- **Multiple-best:** There are multiple placements for the clone in an assembly unit; this clone placement is comprised of two top-ranked end placements and the multiple placements are due to the existence of end sequences with multiple end placements.
- **Multiple-other:** There are multiple placements for the clone in an assembly unit; this clone placement is comprised of lower-ranked end placements and the multiple placements are due to the existence of end sequences with multiple end placements.

Additional details describing the clone placement process are provided in [documentation](#) on the Clone DB website.

Access

Data from Clone DB can be accessed via FTP, direct Entrez query, or through use of tools such as the Clone DB library browsers or Clone Finder.

FTP site

At the Clone DB [FTP](#) site, users will find a number of reports that provide details about clone-associated sequences and clone placements. These reports, which are organized by species, include:

- **clone_acstate:** details of genomic clone insert sequences (updated weekly)
- **clone_placement_report:** summary information for clone placements generated by Clone DB (updated whenever new placements are generated)
- **endinfo:** details of genome clone end sequences (updated weekly)
- **end_placement_report:** summary information for end sequence placements generated by Clone DB (updated whenever new placements are generated)
- **library:** summary details for all clone libraries (updated weekly)

The FTP site also contains text files with the clone placements themselves. There may also be additional text files available for some organisms, depending on data availability.

Entrez search

As an Entrez-indexed database, Clone DB can be directly queried by entering text into the search box found at the top of all Clone DB web pages. For assistance with the construction of complex text queries, users may click on the “Advanced” link located beneath the search box. [Documentation](#) describing the complete set of indexed terms is accessed by clicking on the “Help” link found on each Clone DB web page. Query results are presented in tabular format where each row provides a result summary and a link to the corresponding library or clone record page (Figure 5).

Library browsers

Clone DB also provides a pair of library browsers to facilitate user access to cell-based and genomic library records. These browsers, which are accessed via links on the Clone DB homepage, are sortable tables that provide summary information for each of the libraries represented in Clone DB (Figure 6). A set of filters can be used to restrict the displays to subsets of libraries meeting certain characteristics, such as organism, library or vector type, sequence count or distributor. Each row in the browser table provides a link to the corresponding library record. FAQ pages for the [genomic](#) and [cell-based](#) library browsers are provided to assist users with their navigation.

Clone DB Records

Libraries

Users can view cell-based and genomic library details on Clone DB’s individual library record pages. Library records can be accessed via links in the “Library Name” and “Library Abbreviation” columns in the genomic and cell-based library browsers, as well as from the “Library Name” column of the table in which results from an Entrez search of Clone DB are displayed. At the top of each library record page, a summary provides easy

Clone Name A	Clone Name Aliases	Library Name B	Library Abbreviation	Library Type	Organism	Vector Type	Placed
HEPD0835_4_H04		Helmholtz Zentrum Muenchen GmbH Targeted (Reporter) JM8.N4 C57BL/6N L1L2_Bact_P		gene_targeting	Mus musculus	plasmid	N
HEPD0835_4_G01		Helmholtz Zentrum Muenchen GmbH Targeted (Reporter) JM8.N4 C57BL/6N L1L2_Bact_P		gene_targeting	Mus musculus	plasmid	N
HEPD0835_4_D03		Helmholtz Zentrum Muenchen GmbH Targeted (Reporter) JM8.N4 C57BL/6N L1L2_Bact_P		gene_targeting	Mus musculus	plasmid	N
HEPD0835_4_D01		Helmholtz Zentrum Muenchen GmbH Targeted (Reporter) JM8.N4 C57BL/6N L1L2_Bact_P		gene_targeting	Mus musculus	plasmid	N

Figure 5. Screenshot of results returned from an Entrez query of Clone DB ("gene targeting"[Library Type]) AND mouse[Organism]). Only clones are returned in this tabular display. Users can access individual clone (A) or library (B) records by clicking in the data in the Clone Name and Library Name columns.

Library Name	Library Abbreviation	Vector types	Distributors	Items 1 - 10 of 32	<< First	< Prev	Page <input type="text" value="1"/> of 4	Next >	Last >>
				Total clones	Total end sequences	Total insert sequences			
Zea mays fosmid library	Z_AI	fosmid	AGI	433,875	776,583	63			
Zea mays High-CoT library	ZMMBTa	plasmid		271,076	445,631	0			
CHORI-201 Maize B73 BAC Library	CH201	BAC	CHORI	196,982	726,825	12,788			
Zea mays BAC HindIII library ZMMBBb	ZMMBBb	BAC	CUGI	183,586	616,039	4,447			
Zea mays methyl-sensitive linking library ZMMBLd	ZMMBLd	BAC	AGI	5,614	10,916	0			
Zea mays HMPR library ZMMBHf	ZMMBHf	plasmid	AGI	5,288	10,428	0			
Zea mays methyl-sensitive linking library ZMMBLc	ZMMBLc	BAC	AGI	4,858	9,639	0			
Zea mays methyl-sensitive linking library ZMMBLi	ZMMBLi	BAC	AGI	4,608	8,495	0			
Zea mays methyl-sensitive linking library ZMMBLb	ZMMBLb	BAC	AGI	4,582	8,733	0			
Zea mays methyl-sensitive linking library ZMMBLe	ZMMBLE	BAC	AGI	4,545	8,920	0			

Figure 6. Screenshot of Clone DB genomic library browser. In this image, a filter has been applied that restricts the browser to the display of Zea mays genomic libraries only. Clicking on the data in either of the first two columns will take the users to the corresponding individual library record display.

access to key library attributes, including library name, library group, organism and counts of the number of clones and associated sequences in the database. For genomic library records, this summary also includes distributor information. For cell-based library records, the summary provides the allele type. Below the summary, a tabbed table provides details about the library's DNA source, construction, statistics, and aliases by which it is known (Figure 2). In the case of cell-based libraries, the table also provides

information about the library host. On these pages, users will also find a link to the results of an Entrez query that returns the records for all clones in the library, as well as links that direct them to publications describing library construction and/or sequencing. The data presented in these records are intended to help researchers determine whether clones from the library will be suitable for their research needs and to facilitate their use in a research setting. For more information about these pages, please see the Clone DB [Help](#) page.

Clones

Details for individual clones can be viewed in Clone DB's individual clone record pages, which can be accessed by performing an Entrez query of Clone DB and clicking on links in the "Clone Name" column of the table in which search results are displayed. Similar to the library record pages, the clone record pages also contain a summary section that presents a digest of key attributes. These include the standardized clone name, along with any aliases by which the clone is known, the library to which it belongs, and the library type. A tabbed table beneath the summary provides additional information about the clone. A tab specific to the murine cell-based clone record pages presents allele information, including type, name, and links to corresponding gene records at MGI and the NCBI [Gene](#) database (Figure 7). Cell-based and genomic clone records both have tabs that provide distribution details and information about corresponding sequences. In genomic clone records, users will find additional tabs containing data about genetic markers mapped to the clone, fingerprint contigs to which the clone belongs, and clone placement details (see below for more information about accessing clone placements). Detailed descriptions of the clone record pages can be found on the Clone DB [Help](#) page.

Clone Placements

Clone Records

Graphical displays of NCBI clone placements can be accessed in individual clone records. Within Entrez, users can search for clones having placements using the query "placed"[Properties], and can also query to identify clones whose placements belong to any of the above-noted confidence categories by using the search field [Placement Confidence]. An ideogram at the top of each record page provides genomic context for any placements the clone may have (Figure 8). The "Genome View" tab displays an instance of the NCBI Sequence Viewer showing the placement of the clone in the context of other clones in the same library (Figure 9). If a clone has multiple placements, users can select the specific placement to be displayed. The selected placement for the clone of record is highlighted and shown at top. This display also shows assembly components and NCBI annotated genes. A [FAQ page for clone placements](#) provides a legend that explains the rendering scheme used in this graphical display. Holding the mouse over any placement will bring up a tool tip that includes additional placement details, including the concordance and uniqueness, as well as the sequence identifiers of the prototype ends that contributed to the placement. Additional placement information for the clone, including

A

Allele Information

Allele Name: gene trap OST109778, Lexicon Genetics

Allele Type: Approved

MGI Allele Symbol: [Glmn<Gt\(OST109778\)Lex>](#)

MGI Gene Name: [Glmn](#)

Entrez Gene: [Glmn](#)

IMSR Distributors **B**

Resource Type	This allele: Glmn<Gt(OST109778)Lex>	Any mutation in Glmn	IMSR Distributors
ES cells	0	3	TIGM, WTSI
Strains	1	2	JAX, TAC

Figure 7. Screenshots showing details from individual murine cell-based clone record page. A: “Allele Information” section with links to allele and gene records at MGI and the NCBI Gene database. B: The table in the “Distributors” tab provides users with information about how to obtain clones from the International Mouse Strain Resource (IMSR) in various formats.

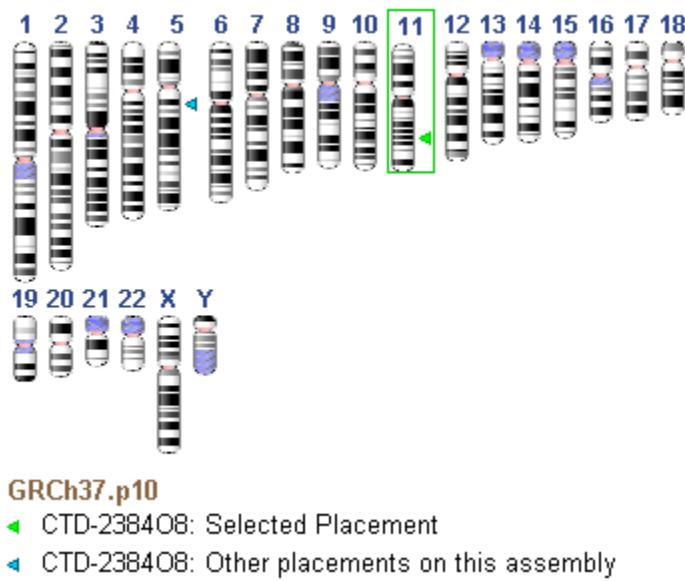


Figure 8: Ideogram from individual genomic clone record displays showing clone placements (arrowheads).

any non-sequence based placements (i.e., cytological), is provided in tabular format in the “Clone Placements” tab (Figure 10).

Clone Finder

Clone placements may also be viewed with the NCBI Clone Finder resource, which can be accessed from the Clone DB home page. While the individual clone record pages enable users to review the placement details for a specific clone in the context of other clones from the same library, Clone Finder allows users to search for and visualize placements of

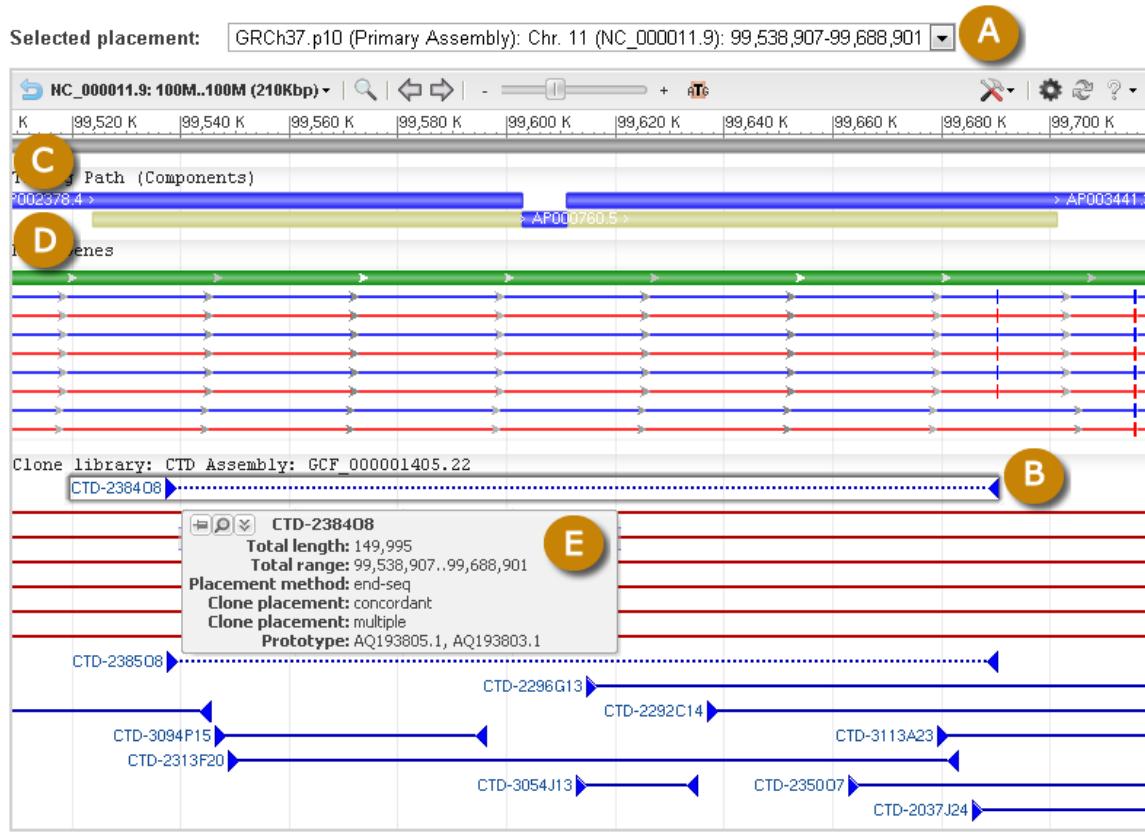


Figure 9. Screenshot showing graphical display of clone placement in an individual genomic clone record. A: Menu for selecting clone placement. B: Selected placement; note that it is highlighted and displayed above all other clone placements. C: Assembly components. D: NCBI annotated genes. E: Hovering over any of the placements with the mouse will bring up a tool-tip with additional placement details.

Sequence-Based Placements A											
Assembly	Asm. unit	Chr	Seq. ID	Start	End	Length	Method	Conc.	Confidence	Placed by	
GRCh37.p10	Primary Assembly	11	NC_000011.9	99,538,907	99,688,901	149,995	end-seq	Y	Multiple-best	NCBI	
GRCh37.p10	Primary Assembly	5	NC_000005.9	66,135,910	66,259,424	123,515	end-seq	Y	Multiple-best	NCBI	
CHM1_1.0	Primary Assembly	5	NC_018916.1	66,054,944	66,178,542	123,599	end-seq	Y	Unique-dissent	NCBI	
HuRef	Primary Assembly	11	AC_000143.1	95,472,235	95,622,249	150,015	end-seq	Y	Multiple-best	NCBI	
HuRef	Primary Assembly	5	AC_000137.1	63,092,865	63,216,338	123,474	end-seq	Y	Multiple-best	NCBI	

Non-Sequence Based Placements B							
Chr	Location	Method	Placed by	NCBI Remapped To	Seq. ID	Start	End
5	5q12-5q13	FLpter	LBNL	GRCh37.p10	NC_000005.9	58,900,001	76,900,000

Figure 10. Tabular placement displays from individual genomic library record page. A: Details for sequence-based clone placements. B: Details for non-sequence based clone placements.

clones from different libraries that have been mapped to specific genomic regions. Users may perform Clone Finder searches based on genomic location or feature, such gene or

Specify Region

A Search by Position **Search by Feature**

Region	Feature type	Feature name
Chromosome: - All -	From: Any	is myod
Assembly: GRCm38	To: Any	is

Select Region

Assembly	Chromosome	Begin	End	Length
GRCm38	7	46,376,474	46,379,092	2,619

B Set Data display filters

Dataset selection: Clones, Check all, Clear all

DNA Source: normal, cancer, unknown, Check all, Clear all

Strain Selection: 129S1/SvImJ, 129S6/SvEvTac, 129S7/SvEvBrd-Hprt-b-m2, AKR/J, BALB/cByJ
 C3H/HeJ, C3HeB/FeJ, C57BL/6J, C57BL/6NCrlCrlj, CAST/Ei
 DBA/2J, MSM/MS, NOD/MrkJac, NOD/ShiLtJ, RP21

Library Selection: Check all, Clear all

Clone vectors: B6Ng01, C3H, CH25, CH26, CH28
 CH29, CH33, CH34, CH36, CT7
 DN, MM_DBa, MSMg01, RP21, RP22
 RP23, RP24, WI1, bMQ

Figure 11. Clone Finder search interface. A: Users may search for clone placements by chromosome coordinate or genomic feature. B: A number of filters allow users to restrict the display of placed clones.

transcript name, marker or SNP. Filters are available to restrict searches by DNA source, library, or vector type (Figure 11). In contrast to the placement displays provided in individual clone records, Clone Finder can simultaneously display the placements for clones from different libraries. The Clone Finder graphical display distinguishes concordant and discordant placements and includes assembly components and annotated genes in the selected region (Figure 12). The placement data is also displayed in a tabular format (Figure 13) and can be downloaded in Excel.

Related Tools

Several related tools at NCBI are available that may be of interest to users of Clone DB.

- The Map Viewer “Clone” track presents clone placements generated by Clone DB.
 - Only clones with concordant placements are displayed in this track.
- Utilities for accessing end sequence records
 - `endseq_dp.pl`
 - This is a perl script provided by Clone DB that dumps FASTA files for end sequences with records in dbGSS or Trace Archives

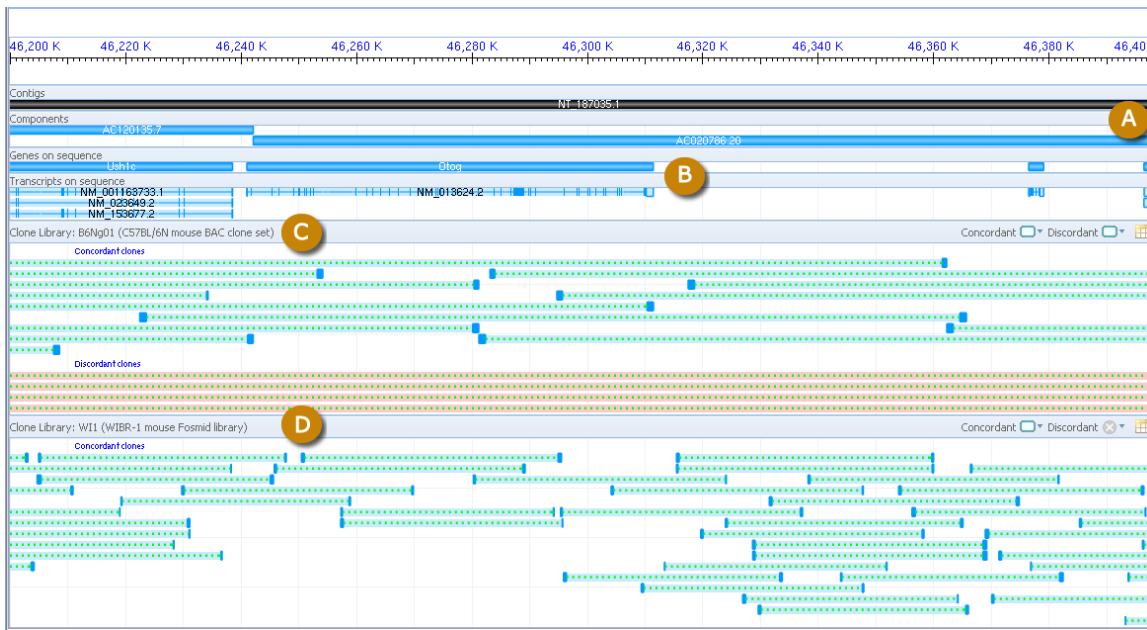


Figure 12. Screenshot of Clone Finder graphical display. A: Assembly scaffolds and components. B: Gene and transcript annotation. C, D: Clone placements from two different libraries are shown. Concordant placements are in green, discordant placements are red.

	Clone	End 1	End 2	Start	Stop	Span	Conco...	Unique
⊕	B6Ng01-166K14	GI_292786049	GI_292786048	12,112,274	139,834,232	127,721,959	X	✓
	Feature: B6Ng01-166K14							
	Type: Clone							
	Description(s): Library: B6Ng01							
	Chrom: 7							
	Chr Pos: 12,112,274 - 139,834,232							
	Span: 127,721,959							
	Clone Ends:							
	GI_292786049 offset: 127,720,898 span: 1,062 strand: plus							
	GI_292786048 offset: 1 span: 496 strand: minus							
⊕	B6Ng01-302C18	GI_322255590	GI_322255591	11,603,495	113,581,196	101,977,702	X	X
⊕	B6Ng01-106C19	GI_292797419	GI_292797420	13,172,447	52,201,461	39,029,015	X	X
⊕	B6Ng01-264F21	GI_322320038	GI_322320059	27,915,565	62,513,425	34,597,881	X	✓
⊕	B6Ng01-81B18	GI_292777838	GI_292777839	46,141,814	46,362,284	220,471	✓	✓
⊕	B6Ng01-310M3	GI_322288653	GI_322288654	46,034,686	46,254,166	219,481	✓	✓

Figure 13. Screenshot of Clone Finder tabular display. A: One of the rows has been expanded to show additional placement details.

- It takes a list of NCBI GI numbers (max 1000) or Trace Archive identifiers (max 4000) as input and returns the corresponding FASTA sequences.
- The script and usage directions are located in the [utility](#) directory of the [Clone DB](#) FTP site.
 - [query_tracedb.pl](#)
 - This is a perl script provided by the Trace Archive that can be used to download large datasets from the Trace Archive.
 - The script and usage directions are located in the “[Obtaining Data](#)” tab on the Trace Archive home page.
 - End sequence BLAST databases

- NCBI BLAST databases comprised of end sequences from individual or collections of genomic clone libraries are available for several organisms, including human and mouse.
- These databases are listed in the “Database” drop-down menu on organism-specific BLAST pages, the complete list of which can be accessed via the [Map Viewer home page](#).
- Unless otherwise noted, BLAST databases named “Clone end sequences” only contain end sequences whose records are in dbGSS, not the Trace Archive.
- Genome Reference Consortium (GRC) annotated clone assembly problems files
 - Available for human, mouse and zebrafish clones that are components of the respective reference assemblies for each of these organisms; these files map individual clone assembly problems, such as unsure sequence, single clone coverage, or low sequence quality annotated on the insert sequence records in GenBank, to the corresponding location in the current reference assembly.
 - Available in GFF3 or ASN.1 format
 - The files are found in organism-specific directories on the GRC’s [public FTP site](#).
 - [Human GRCh37 assembly](#)
 - [Mouse GRCm38 assembly](#)
 - [Zebrafish Zv9 assembly](#)

References

1. Schneider VA, Chen HC, Clausen C, Meric PA, Zhou Z, Bouk N, et al. Clone DB: an integrated NCBI resource for clone-associated data. *Nucleic acids research*. 2013;41(Database issue):D1070–8. PubMed PMID: 23193260.
2. Skarnes WC, Rosen B, West AP, Koutsourakis M, Bushell W, Iyer V, et al. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*. 2011;474(7351):337–42. PubMed PMID: 21677750.
3. Pettitt SJ, Liang Q, Raideran XY, Moran JL, Prosser HM, Beier DR, et al. Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nature methods*. 2009;6(7):493–5. PubMed PMID: 19525957.
4. Skarnes WC, von Melchner H, Wurst W, Hicks G, Nord AS, Cox T, et al. A public gene trap resource for mouse functional genomics. *Nature genetics*. 2004;36(6):543–4. PubMed PMID: 15167922.
5. Hansen GM, Markesich DC, Burnett MB, Zhu Q, Dionne KM, Richter LJ, et al. Large-scale gene trapping in C57BL/6N mouse embryonic stem cells. *Genome research*. 2008;18(10):1670–9. PubMed PMID: 18799693.
6. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008;453(7191):56–64. PubMed PMID: 18451855.
7. Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajadian S, Graves TA, et al. Gorilla genome structural variation reveals evolutionary parallelisms with chimpanzee. *Genome research*. 2011;21(10):1640–9. PubMed PMID: 21685127.

8. Safár J, Bartos J, Janda J, Bellec A, Kubaláková M, Valárik M, et al. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *The Plant journal*. 2004;Sep39(6):960–8. PubMed PMID: 15341637.
9. Babiychuk E, Fuangthong M, Van Montagu M, Inze D, Kushnir S. Efficient gene tagging in *Arabidopsis thaliana* using a gene trap approach. *Proceedings of the National Academy of Sciences of the United States of America*. 1997;94(23):12722–7. PubMed PMID: 9356517.
10. Hsing YI, Chern CG, Fan MJ, Lu PC, Chen KT, Lo SF, et al. A rice gene activation/knockout mutant resource for high throughput functional genomics. *Plant molecular biology*. 2007;63(3):351–64. PubMed PMID: 17120135.
11. Lukacsovich T, Yamamoto D. Trap a gene and find out its function: toward functional genomics in *Drosophila*. *Journal of neurogenetics*. 2001;15(3-4):147–68. PubMed PMID: 12092900.
12. Cheung VG, Nowak N, Jang W, Kirsch IR, Zhao S, Chen XN, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*. 2001;409(6822):953–8. PubMed PMID: 11237021.
13. Jang W, Yonescu R, Knutsen T, Brown T, Reppert T, Sirotnik K, et al. Linking the human cytogenetic map with nucleotide sequence: the CCAP clone set. *Cancer genetics and cytogenetics*. 2006;168(2):89–97. PubMed PMID: 16843097.

Genome Reference Consortium

Valerie Schneider, PhD¹ and Deanna Church, PhD¹

Created: November 14, 2013.

Scope

NCBI is a member of the [Genome Reference Consortium](#) (GRC), an international collaboration that oversees updates and improvements to the human, mouse, and zebrafish reference genome assemblies. These reference assemblies include linear chromosome representations, unlocalized and unplaced scaffold sequences, and alternate loci scaffolds providing alternate sequence representations for genome regions too complex to be adequately represented by the linear chromosome path. The GRC produces two types of assembly updates: (1) major releases, in which chromosome coordinates are changed, and (2) minor releases, in which chromosome coordinates do not change and updates are provided as standalone patch scaffold sequences. All GRC assemblies are submitted to the International Nucleotide Sequence Database Collaboration (INSDC) databases and made publicly available. The GRC is not responsible for annotation of the reference assemblies. For information about the National Center for Biotechnology Information's (NCBI) annotation of the GRC assemblies, please see the handbook chapter titled, "About Eukaryotic Genome Processing and Tools".

History

In 2004, the Human Genome Project (HGP) published a finished version (Build35) of the human genome assembly (1). This was a major accomplishment that represented over a decade of effort by more than a dozen institutions and resulted in the highest quality vertebrate genome ever produced and a new tool for understanding human biology. Despite this achievement, a limited number of gaps, sequence and tiling path errors remained in the reference assembly. Thus, at the conclusion of the HGP and the release of their final assembly version (Build36 (UCSC name: hg18)), the GRC was conceived as a mechanism for continued stewardship and improvement of the human reference assembly. The GRC was subsequently tasked with updating the mouse reference genome upon conclusion of its major sequencing effort and assembly release (MGSCv37) (2), and in 2010 the GRC also assumed responsibility of the zebrafish reference genome after the release of the Zv9 assembly.

The GRC is comprised of four institutions. NCBI supplies the database and provides bioinformatics support for the consortium, and also develops public-facing GRC assembly resources. Sequencing and other wet lab work associated with updating the assembly is performed by The Genome Institute at Washington University, St. Louis and at the

¹ NCBI; Email: schneiva@ncbi.nlm.nih.gov; Email: church@ncbi.nlm.nih.gov.

Wellcome Trust Sanger Institute. The latter, along with the European Bioinformatics Institute (EBI) provide additional bioinformatics support and tool development for the GRC.

Although the GRC's primary role was initially envisioned to be one of gap-filling and sequence correction, advances in genomic and population biology made possible by the availability of the human reference genome soon defined new assembly management tasks for the consortium. Notably, many studies of the human genome revealed previously unrecognized degrees and forms of genetic variation (3-10). The original assembly model, comprised of linear chromosome sequences, proved insufficient in its ability to represent this variation. Thus, the GRC, in addition to correcting assembly errors, also makes updates to the assembly model used to represent these organisms' genomes and works to provide additional representations of diversity in the reference assemblies (11). In 2009, it produced an updated human assembly (GRCh37 (UCSC name: hg19)) and, in 2012, released a revised mouse assembly (GRCm38 (UCSC name: mm10)), the first two assemblies to be represented by the new model. Today, the GRC remains dedicated to producing improved reference assemblies that serve as valuable substrates for a variety of analyses.

Data Model

Assembly Model

It is important to recognize that a genome assembly and a genome are not the same thing. A genome is the physical genetic entity that defines an organism. An assembly is not a physical object; it is the collection of all sequences used to represent the genome of an organism. The GRC utilizes a specific assembly model for the reference genomes under its auspices (Figure 1). However, this assembly model can be adopted for use with almost any eukaryotic genome. Within this model, sequences belong to different hierarchies and are assigned to various assembly units, depending upon their role in assembly.

Sequence Hierarchies

Because current sequencing technologies do not allow for chromosomes to be sequenced from end-to-end in a continuous fashion, they must be fragmented, sequenced, and reassembled for purposes of representation. The minimal collection of sequences needed to reconstruct a molecule of interest is referred to as its tiling path. The reference assembly model includes three tiers of accessioned sequences. Figure 2 uses human chromosome 6 ([CM000668.1](#)) to illustrate this hierarchy. At the bottom of this hierarchy are the tiling path components, which in the case of the GRC reference assemblies are primarily genomic clones or Whole Genome Shotgun (WGS) contigs. In the middle are scaffolds, which are sets of ordered and oriented components. At the top of this hierarchy lie the chromosome sequences. These are assembled from scaffolds that have been localized and oriented with respect to one another and that are separated from one another by gaps representing unresolved sequence. A genome assembly may also contain scaffold

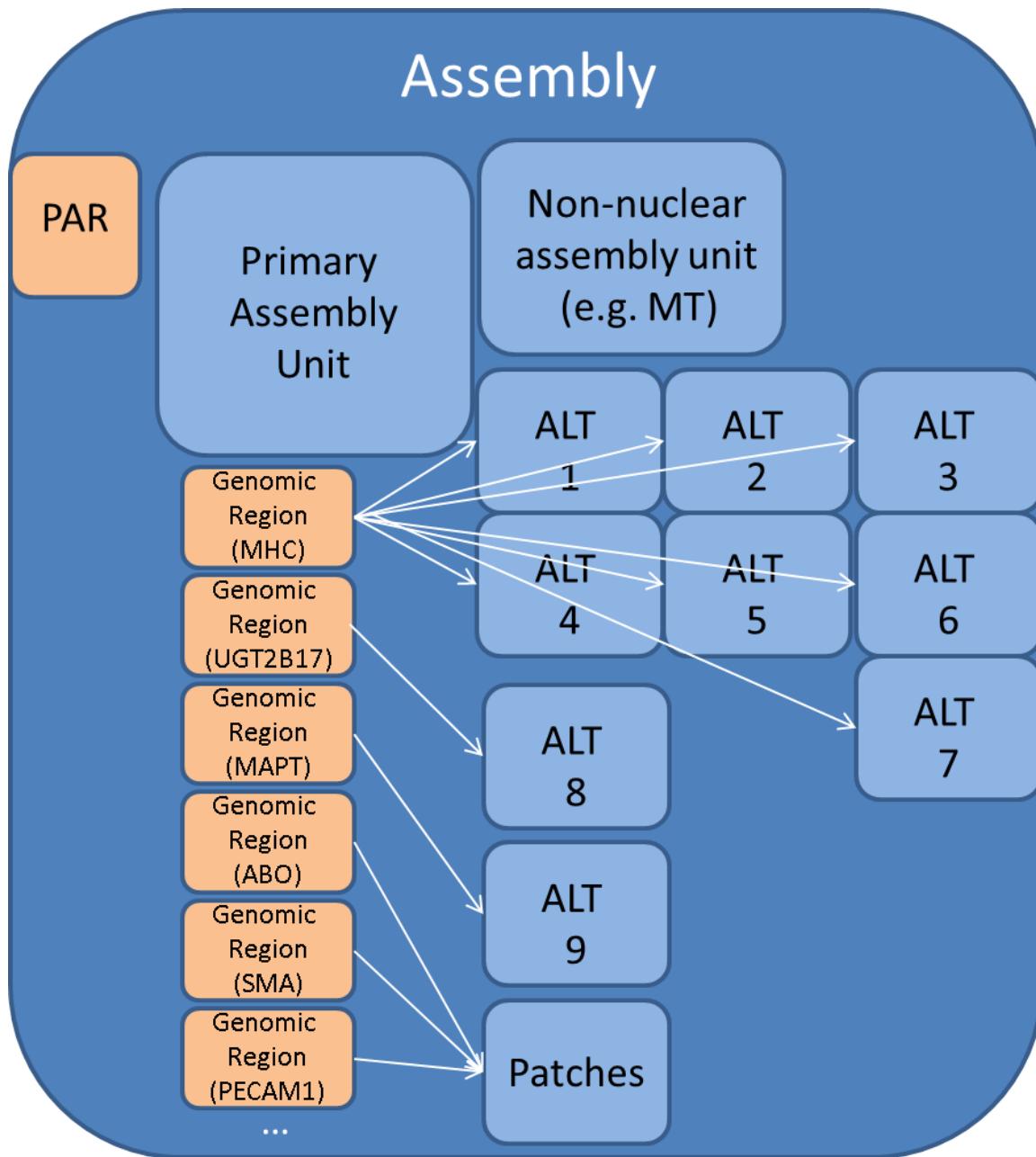


Figure 1. Schematic representation of the assembly model, showing assembly units and regions. The primary assembly unit is the collection of sequences that provides a haploid representation of the genome. This includes chromosome sequences, as well as unlocalized and unplaced scaffolds. Alternate loci assembly units consist of scaffold sequences that represent variants of sequence present in the primary assembly unit. The Patches assembly unit includes scaffolds that represent updates made to the reference assembly since its last major release. Genomic regions define chromosome extents for which there are alternate loci or patch scaffold representations. The PAR (pseudoautosomal region) defines the extents of homology between the sex chromosomes.



Figure 2. Sequence hierarchy in human chromosome 6 (CM000668.1). A: component sequences. In this chromosome, the components are either clone sequences or WGS contigs. The ordered set of components shown here comprises the tiling path for this chromosome. B: localized scaffold sequences. C: chromosome sequence. The large gap between the first and second scaffolds occurs at the location of the centromere, which appears as Ns in the chromosome sequence record.

sequences whose chromosomal context is either poorly defined or not known. The former category describes unlocalized scaffolds. These are genomic sequences that have been assigned to a particular chromosome, but whose location within that chromosome cannot be unambiguously defined at this time. Scaffolds entirely without chromosomal context are known as unplaced scaffolds.

Primary Assembly Unit

The primary assembly unit is the collection of sequences that, all together, provide a haploid representation of an organism's genome. Prior to the development of this assembly model, the human reference assembly only consisted of the sequences in the primary assembly unit. As a result, researchers sometimes mistakenly continue to refer to the collection of sequences in the primary assembly unit as the reference assembly. However, this is only one of several assembly units that together comprise GRC assemblies.

The primary assembly unit includes the chromosome sequences and the collection of unlocalized and unplaced scaffolds. These scaffold sequences make important contributions to the primary assembly unit. For example, in the GRCh37 primary assembly unit, an unlocalized scaffold associated with chromosome 1 provided the only representation for the HYDIN2 locus ([GL000192.1](#)). Although this locus is known to reside on chromosome 1, a complex repeat structure confounded the chromosome assembly and made the assignment of this scaffold to any one of three gaps equally likely. Consequently, the scaffold was designated unlocalized.

Alternate Loci Assembly Units

Alternate loci assembly units contain sequences that represent variants of sequence present in the primary assembly unit. As such, they permit an assembly to provide more

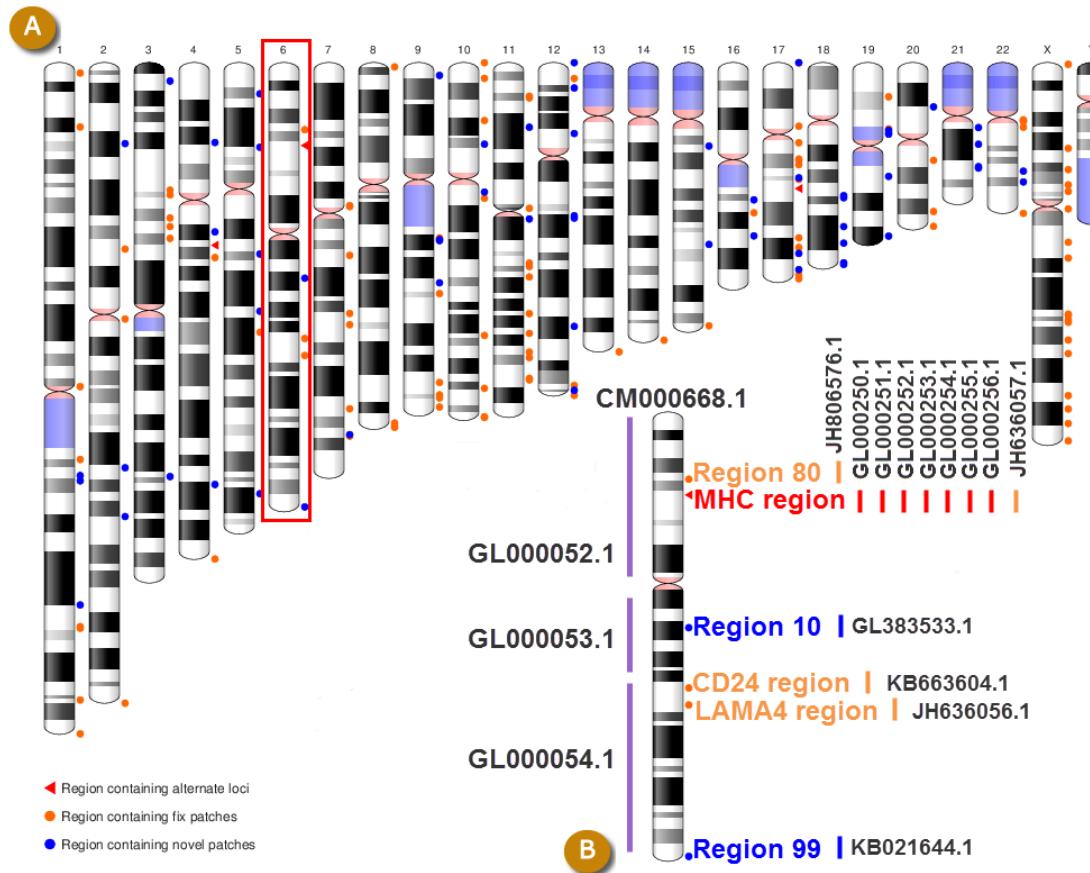


Figure 3. A: Ideogram representation of the human genome, with the locations of regions represented by alternate loci and patch scaffolds in the GRCh37.p12 assembly. B: An enlarged view of chromosome 6 (CM000668.1) shows the locations of 3 localized scaffolds (GL000052.1-GL000054.1) belonging to the primary assembly unit, along with 6 regions: MHC (associated with 7 alternate loci unit scaffolds (GL000250.1-GL000256.1) and one fix patch scaffold (JH636057.1)), REGION80 (associated with FIX patch JH806576.1), REGION10 (associated with the novel patch scaffold GL383533.1), CD24 (associated with the fix patch KB663604.1), LAMA4 (associated with the fix patch JH636056.1), and REGION99 (associated with NOVEL patch KB021644.1).

than a haploid representation of a genome. While there are no size limits for sequences in alternate loci assembly units, these are generally scaffold sequences less than 5 Mb in length. In the human reference assembly, which does not represent an individual genome, alternate assembly units are not organized by haplotype. In contrast, alternate assembly units in the mouse reference assembly are organized by strain; they only include sequences from strains other than C57BL/6J, which is represented in the primary assembly unit. No alternate assemblies have yet been defined for the zebrafish reference assembly. For GRCh37, the GRC instantiated 7 alternate loci assembly units so that the reference assembly might better represent the diversity that exists in the major histocompatibility complex (MHC) region on human chromosome 6, one of the most variable regions of the human genome (Figure 3). There are therefore 8 sequence

representations for the MHC in GRCh37: one on the chromosome sequence from the primary assembly unit ([CM000668.1](#)), and 7 from scaffolds belonging to 7 alternate loci assembly units ([GL000250.1](#)-[GL000256.1](#)).

Patches Assembly Unit

All patches belong to the patches assembly unit. Patches are scaffold sequences that represent updates made to the reference assembly since its last major release. Thus, the patches assembly unit is empty at the time of an assembly's major release. The GRC releases patches on a quarterly basis; the patches assembly unit always contains the complete collection of patches associated with the reference assembly. Patches do not change the coordinates of any sequences in the primary assembly or alternate loci units. The assembly model includes the concept of patches because they provide a mechanism for providing users with timely access to assembly improvements without the need for frequent major assembly releases involving chromosome coordinates updates that many researchers find disruptive. The GRC does not integrate the patch scaffolds into the chromosomes; they exist only as scaffold sequences.

There are two types of patch scaffolds in this assembly unit. Fix patches correct errors in the primary and alternate loci assembly units, while novel patches add new sequence variants to the assembly. As illustrated in Figure 4, the fix patch [GL339450.1](#) provides a single haplotype representation for the [ABO](#) locus, correcting the mixed, non-existent haplotype found in GRCh37 where the locus spanned two components with different haplotypes. In Figure 5, the novel patch [GL383583.1](#) is shown to represent a deletion variant involving the [APOBEC3A](#) and [APOBEC3B](#) genes, which are involved in innate immunity and retroviral infections. The deletion variant, which is common in Asia but rare in Europe and Africa, creates a gene fusion, [APOBEC3A_B](#) (12). At the time of an assembly's next major release, all fix patch scaffold sequences will be deprecated, as the changes they represent will be reflected in sequences in the primary assembly and alternate loci assembly units. In contrast, novel patch scaffold sequences will be retained, though they will be moved from the patches assembly unit to the appropriate alternate loci assembly unit.

Non-Nuclear Assembly Unit

Although the GRC is not responsible for the maintenance of the mitochondrial reference sequences of the human, mouse, or zebrafish genomes, the assembly model includes a unit for non-nuclear assemblies. The human mitochondrial reference sequence is maintained by the [Mitomap](#) group and is distributed by the GRC with the reference genome assembly for user convenience.

Alignments

Although scaffolds in the patches and alternate assembly units do not have chromosome coordinates, they may be placed in chromosome context by virtue of their alignment to primary assembly sequences. All patch scaffolds and scaffolds in the human alternate assembly units contain at least one anchor sequence as either the first and/or last

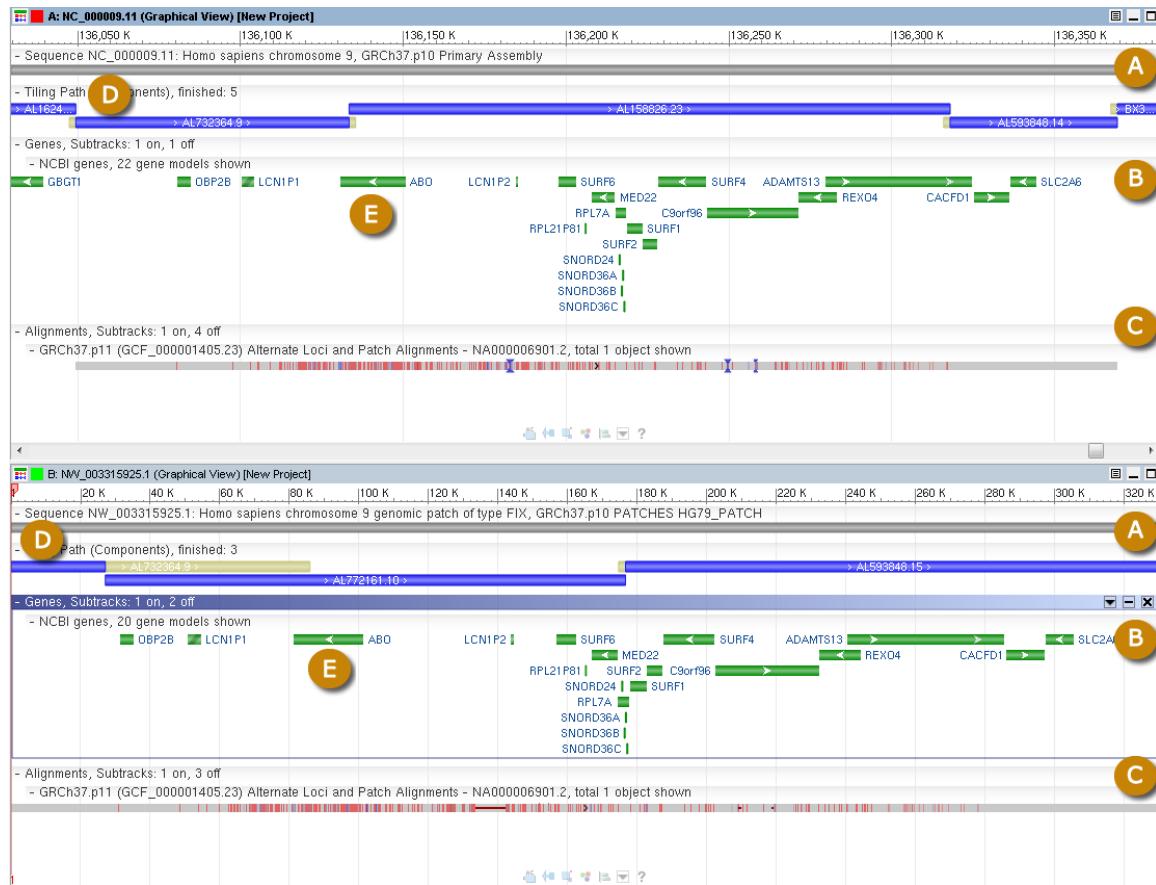


Figure 4. Top panel: RefSeq copy of GRCh37 chromosome 9 (NC_00009.11). The annotated RefSeq chromosome NC_00009.11 is a copy of the GRC chromosome CM000671.1. Bottom panel: Annotated RefSeq copy (NW_003315925.1) of the GRC fix patch GL339450.1. A: The blue bars represent each of the components that make up the tiling paths of the patch scaffold and chromosome. B: NCBI annotated genes. C: Top panel: alignment of chromosome to patch; Bottom panel: alignment of patch to chromosome. Red ticks in the alignment highlight mismatches, blue triangles represent deletions, and thin lines indicate insertions. The anchor component (AL732364.9) of the patch is marked (D). Note how the ABO locus (E) in the fix patch is derived from a single component, as opposed to the two components on the GRCh37 chromosome.

component (Figures 4 and 5). These anchor sequences are components that are also found in the primary assembly unit and are included to ensure a good alignment of the alternate locus scaffold to the primary assembly. Because the alternate loci assembly units in the mouse assembly are strain specific, their scaffolds do not contain anchor sequences from the primary assembly unit. As a result, mouse alternate loci scaffolds may not always have an alignment to the primary assembly unit.

The GRC generates alignments of the alternate loci and patch scaffolds to the primary assembly unit and submits these alignments to the NCBI Assembly <http://www.ncbi.nlm.nih.gov/assembly/>database with every assembly release. As a result, these alignments are part of the assembly definition and are distributed on the GenBank FTP

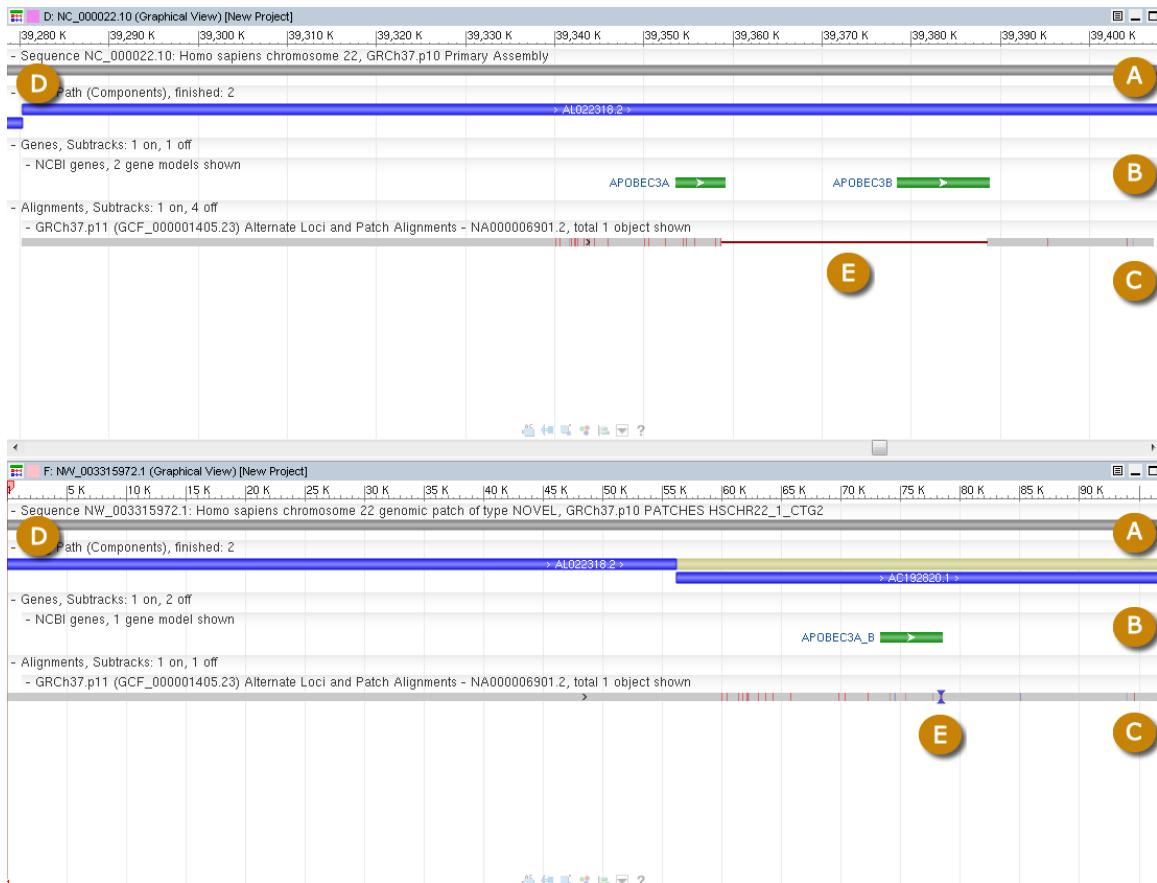


Figure 5. Top panel: RefSeq copy of GRCh37 chromosome 22 ([NC_000022.10](#)). The annotated RefSeq chromosome NC_000022.11 is a copy of the GRC chromosome CM000684.1. Bottom panel: Annotated RefSeq copy ([NW_003315972.1](#)) of the GRC novel patch GL383583.1. A: The blue bars represent each of the components that make up the tiling paths of the patch scaffold and chromosome. B: NCBI annotated genes. C: Top panel: alignment of chromosome to patch; Bottom panel: alignment of patch to chromosome. Red vertical lines in the alignment highlight mismatches, blue triangles represent deletions, and thin red horizontal lines indicate insertions. The anchor component ([AL022318.2](#)) of the patch is marked (D). Note that the APOBEC3A_B locus in the patch overlaps its deletion (E) relative to the chromosome sequence.

site with the assembly sequences. The alignments distinguish how scaffold sequences from the patches or alternate loci assembly units differ from the primary assembly unit sequence. Figures 4 and 5 also show the alignments between the annotated RefSeq copies of the aforementioned fix and novel patches, and the corresponding GRCh37 chromosome sequences.

Assembly Regions

The GRC defines discrete regions on sequences in the primary assembly unit where alternate loci and patch scaffolds are aligned. A region may contain more than one patch or alternate loci scaffold and the extent of a region is defined by the outermost edges of the corresponding alignments. The GRC also defines regions on the X and Y

chromosomes corresponding to the extents of the pseudo-autosomal regions (PAR), as defined by their alignments to one another. The ideogram in Figure 3 shows the location of regions associated with the GRCh37 assembly.

Assembly Accessions

All GRC assembly sequences are submitted to [GenBank](#) and the assembly itself is submitted to the NCBI [Assembly](#) database. Every scaffold and chromosome in the assembly receives an accession.version, which is a unique identifier of the sequence. Likewise, the assembly units and full assembly also receive accession.versions. These identifiers enable users to track the collections of sequences within each assembly. The GRC strongly recommends that authors include the accession.versions of all assembly sequences referenced in their publications. Because sequence coordinates may change with each accession.version update, use of these identifiers provides an unambiguous definition of the coordinate-sequence relationship. Such usage eliminates any possible reader confusion with respect to the particular sequence on which coordinates may be reported for genes, regulatory regions or other assembly features.

Dataflow

Figure 6 provides a schematic of the GRC dataflow for assembly updates. GRC assemblies start with a set of text files known as [TPFs](#) (tiling path files). TPFs provide an ordered list of the components and gaps that make up a scaffold or chromosome. However, they specify neither the orientation of the components, nor the specific sub-regions of the components that will contribute to the final sequence. GRC curators download TPF files from an NCBI database and update them with changes to the tiling path by adding, removing, or reordering components as indicated by their analyses. All updates are made in accordance with a series of GRC-developed standard operating procedures for assembly curation and the GRC uses a centralized system to track the regions of the assembly under review. The TPF files are then reloaded to the database, where they are validated for format and content. A versioning system ensures that all TPF updates are recorded, and a check-in/check-out system for the files prevents simultaneous modification of a TPF by more than one curator.

A modified version of the NCBI NGAligner software identifies and evaluates alignments between adjacent components with respect to criteria such as length and percent identity. Adjacent assembly components are generally expected to have dovetail overlaps (Figure 7), though other alignment types are sometimes observed. Pairs without alignments or those whose alignments do not meet established GRC evaluation criteria are prioritized for review. There are three possible outcomes of review: (1) the TPF may be further updated to solve the problem, (2) a new alignment meeting evaluation criteria may be curated and stored, or (3) the GRC may provide external evidence supporting the pairing of the sequences despite the low quality alignment (join certification). If a pair exhibits more than one alignment, a curator will designate the preferred alignment. The pairwise

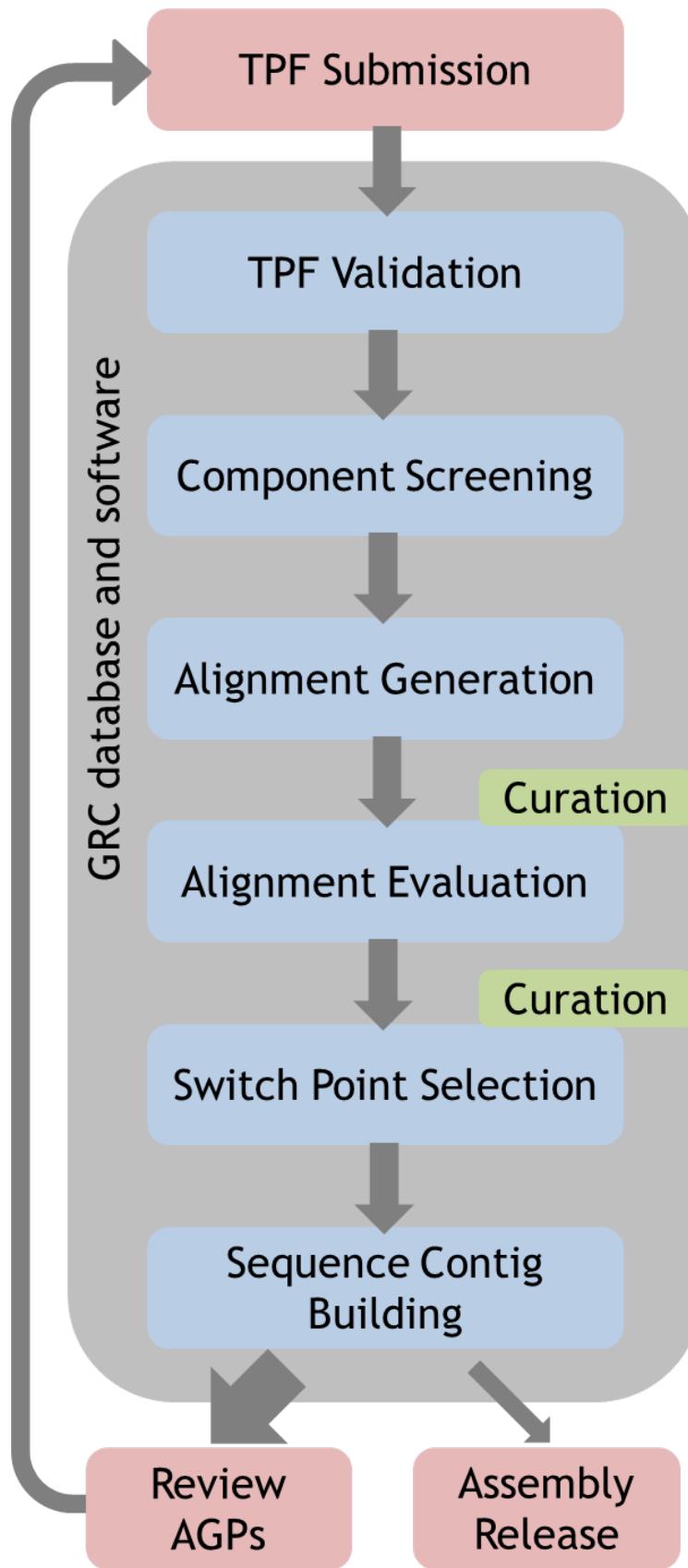


Figure 6. Dataflow for GRC assembly updates.

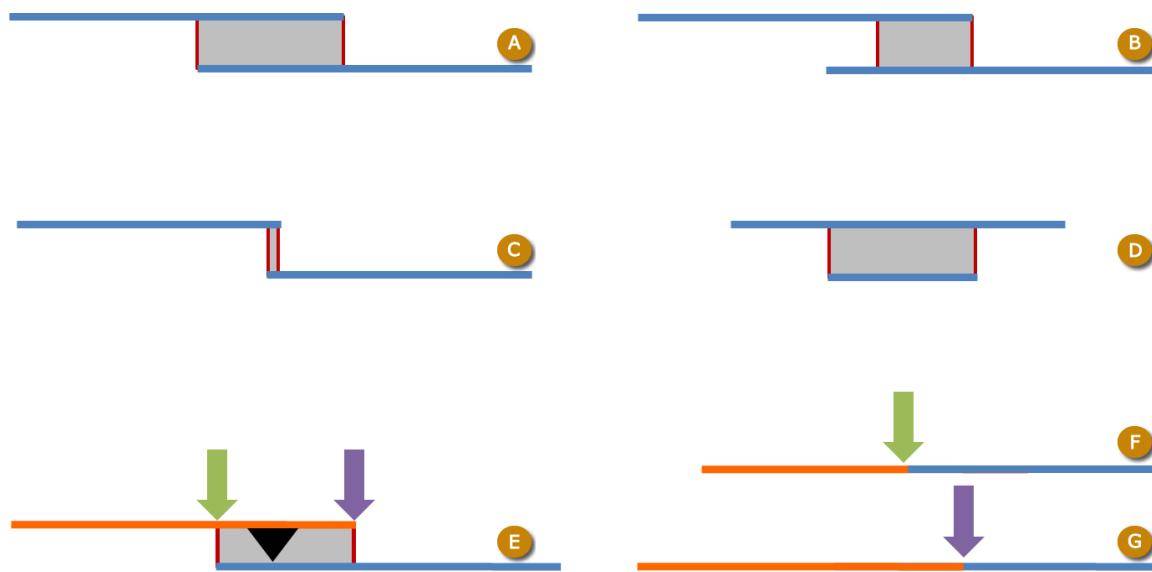


Figure 7. Schematic of component overlaps and switch points. Blue and orange bars represent components, gray boxes indicate aligned regions, and switch points located at either extent of each alignment are indicated by thin red lines. A: Full dovetail alignment. This is the type of alignment that is expected for adjacent TPF components. B: Half-dovetail alignment, in which the end of one of the components does not align. While such alignments may be indicative of two components that do not belong together, this situation can also occur if the components contain untrimmed vector sequence or overlap in a repetitive sequence of variable length. C: Short/blunt overlap (< 50 bp). These alignments always require external evidence in the form of a join certificate. D: Contained alignment, in which one component's sequence is contained in the other. This situation is generally observed when the shorter component is being used to correct an error in the longer component. E: Default switch point position (purple arrow) between two assembly components. Because of an indel in the alignment (black triangle), moving the switch point (green arrow) may change the resulting sequence. F: Sequence constructed from alternate switch point (green arrow) in panel E. G: Sequence constructed from default switch point (purple arrow) in panel E.

alignments and evaluation results are stored to the database. As a result, alignments need only be generated and evaluated for new sequence pairs on new or updated TPFs.

NCBI-developed software is also used to select switch points for each aligned pair (Figure 7). The switch points define the start and stop positions of the individual components in the scaffolds. By default, this occurs at the last base of the first component in the aligned pair. If an alignment does not exhibit 100% identity, which may occur when components represent different haplotypes or other forms of variation, the GRC may curate the switch points in order to include or exclude sequence unique to one of the components. Like the alignments, switch points are stored in the database and are only generated for new sequence pairs on new or updated TPFs. All switch points are validated to ensure they occur at aligned bases.

NCBI sequence contig building software known as `tpf_builder` uses the component order specified on the TPFs and the stored alignments and switch points to build sequence contigs and generate [AGP](#) (A Golden Path) files that describe the assembly scaffolds and chromosomes (Figure 6). During the inter-release period for an assembly, this software

runs every time there is a sequence-changing TPF update. Any errors encountered in the process are reported to curators for their review, and the entire assembly curation process is repeated as necessary. At the time of a public assembly release, tpf_builder is triggered to produce a final set of AGP files. The alignments of the patch and alternate loci scaffold alignments to the primary assembly are also produced at this time, as are the genomic region definitions. These files are submitted to the NCBI GenColl database and subsequently loaded to GenBank, culminating in an assembly release.

There are two types of assembly releases. Minor releases are used by the GRC for updates to the patches assembly unit. In a minor release, the accession.version of the patches assembly unit and the full assembly will increment. However, the accession.version of the primary assembly unit and the alternate loci subunits will not change. As a result, there are no changes to the sequences or of any of the assembly chromosomes. In a major assembly release, all assembly unit accession.versions will increment. Major assembly releases are associated with coordinate changing chromosome updates. Users can distinguish whether a new GRC assembly represents a major or minor release by comparing the accession.version of the primary assembly unit in the latest assembly version to that of the previous assembly version: if the version is unchanged, it is a minor release; if it has incremented, it is a major release. Users can find accession.version information for all GRC assemblies in the [NCBI Assembly resource](#).

Access

Users can download GRC assembly data from the [GenBank FTP](#) site. This data includes the sequences, alignments, assembly region definitions, and join certifications. The genome browsers at [UCSC](#), [Ensembl](#) and [NCBI](#), which obtain the assembly data from GenBank, provide displays for the GRC assemblies. The GRC generates a file that provides the genomic locations for all issues under review, which Ensembl and UCSC display as a track in their browsers. All three browsers have tracks showing the regions in the primary assembly for which there are patch and alternate loci scaffold sequences.

The GRC provides users with access to the inter-assembly TPF and AGP files on the [GRC FTP](#) site. While these files are not recommended for publication-level analyses, due to their instability and lack of corresponding accessioned sequences, they provide users with a preview of genome changes. At this FTP site, the GRC provides a file with the genomic locations of annotated clone assembly problems in the component sequences, which can also be loaded as a browser track.

The GRC strives to make its efforts to update the human, mouse, and zebrafish reference assemblies as transparent as possible. It maintains a [public website](#) (Figure 8) where users can find assembly statistics for current and past assembly releases, plans for future updates, and a link to the [GRC blog](#). At the GRC website, users will find pages describing the current status and genomic locations of individual issues under GRC review (Figure 9). Users can search the GRC website for specific issues by features such as genome location, gene name, accession, or clone name, and links are provided to view the

The Genome Reference Consortium

Putting sequences into a chromosome context.

The original model for representing the genome assemblies was to use a single, preferred tiling path to produce a single consensus representation of the genome. Subsequent analysis has shown that for most mammalian genomes a single tiling path is insufficient to represent a genome in regions with complex allelic diversity. The GRC is now working to create assemblies that better represent this diversity and provide more robust substrates for genome analysis.

Slides from the GRC's presentation at ASHG 2012 are available on the new [Workshops](#) page.

We are planning to update the human reference assembly to GRCh38 in the summer of 2013. If you have questions or concerns about this let us know.

See our [blog](#) for more information on why we think this is important.

We are planning to update the zebrafish reference assembly to GRCz10 in late 2013. If you have questions or concerns about this let us know.

The Genome Reference Consortium consists of:



The Wellcome Trust Sanger Institute



The Genome Institute at Washington University



The European Bioinformatics Institute



The National Center for Biotechnology Information

A

GRC Blog

Genome Update: Highly variant immune regions retiled as single haplotype paths 09 Jan 2013

The GRC and the 10th International Zebrafish Genetics and Development Meeting (June 20-24, 2012 - Madison, Wisconsin) 26 Jul 2012

[see all](#)

Resolved Issues

Human (HG-1033) Mar 29, 2013
AC233266 will be removed from the TPF b/c it represents a different haplotype. Orientation does not need to be set at this time.

Human (HG-1577) Mar 29, 2013
AC236040.3 is a finished component it represents a sequence insertion of 12.1kb relative to the reference assembly and contains a duplication of CYP2D6 (NM_001025161.2). The component has been added to the ALT_REF_LOCI_2 TPF.

[see all](#)

B

C

Figure 8. GRC website. A: GRC announcements. B: Link to and highlights from GRC blog. C: Link to and highlights of recently resolved assembly issues.

corresponding regions in the major browsers. Additionally, the GRC website includes region-centric pages that provide links to the issue reports and sequence records for all patches, alternate loci, and issue reports associated with a specified region, along with a graphical view of the region (Figure 10). The website also provides forms for users to [report assembly issues](#) directly to the GRC, which are entered into the GRC tracking system, as well as to [contact the GRC](#) with general assembly questions.

The GRC also provides users with access to the evaluated alignments, switch points, and join certificates for all sequence pairs on the assembly TPFs (Figure 11). Users can search for specific TPFs by component accession or clone name. The TPF Overview pages present an enhanced view of the TPF files that includes information such as the evaluation status, length, and percent identity for all component alignments. The OverlapView pages, accessed by clicking on the evaluation status markers in the TPF Overview pages, provide alignment and switch point details for each sequence pair in graphical and text formats. There is a link on each OverlapView page that can be used to

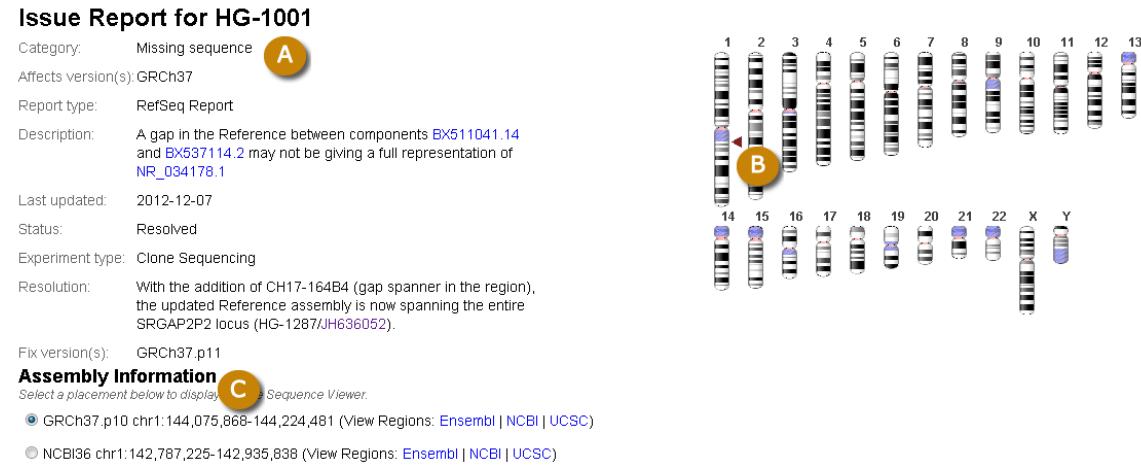


Figure 9. Detail from issue-specific page at GRC website. A: Summary of issue status, as stored in GRC issue tracking system. B: Ideogram showing issue location (triangle). C: Links to display the associated region in an NCBI Sviewer instance found on the page (not shown in figure) or the Ensembl, NCBI, or UCSC browser sites.

view the alignment in Genome Workbench. The OverlapView pages provide information about the database history for the sequence pair, genomic clones whose ends map to either of the components, as well as the coordinates of RepeatMasked regions within the alignment. Links to pages showing join certificates submitted by GRC curators are found in the OverlapView pages for sequence pairs with sub-optimal alignments.

Related Tools

MapViewer and Sviewer

Users can view GRC assemblies and sequences in the NCBI MapViewer and Sviewer resources. These resources can be configured to show different tracks containing assembly data.

Clone DB

The NCBI [Clone DB](#) maintains records for the genomic clones that are components of the GRC assemblies, as well as for other, non-component clones. These records include sequence, distributor, and mapping information.

Assembly database

All GRC assemblies are submitted to the NCBI [Assembly](#) database.

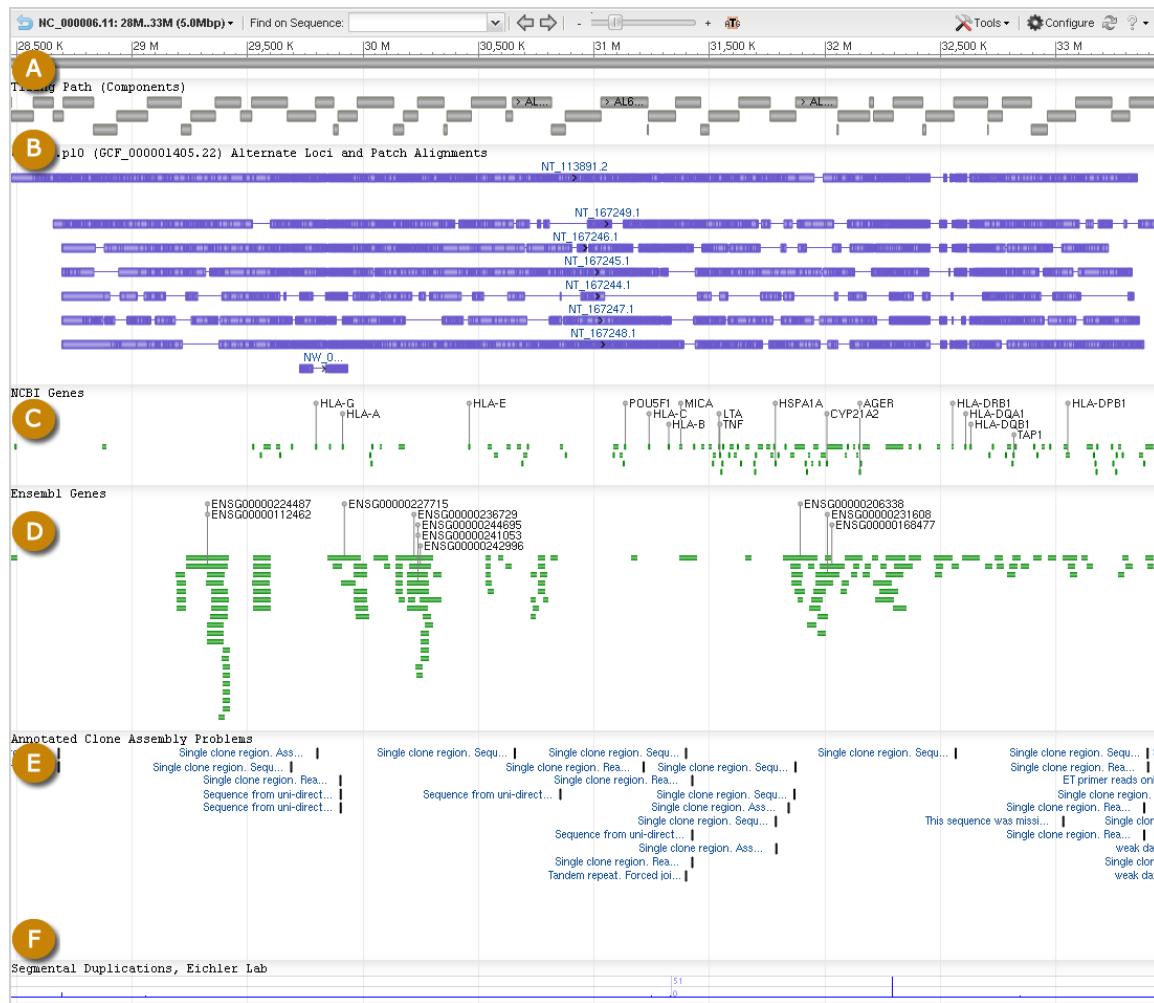


Figure 10. Screenshot of NCBI Sviewer display from the GRC region-specific page for the human major histocompatibility complex (MHC) region on chromosome 6. This Sviewer instance includes several default tracks useful for evaluation of the chromosome or scaffolds in the region. A: Tiling path of chromosome components. Note: The Sviewer display can be toggled to the reciprocal perspective so that it shows the tiling path for any of the scaffolds in this region. B: Alignments of all alternate loci and patches in this region to the chromosome. C: NCBI genes annotated in the region. D: Ensembl genes annotated in the region. D: Annotated clone assembly problems. F: Segmental duplications.

Genome Remapping Service

The NCBI genome remapping service can be used to remap features between different assembly versions.

Eukaryotic Genome Annotation Pipeline

All GRC assemblies are annotated as part of NCBI's eukaryotic genome annotation pipeline.

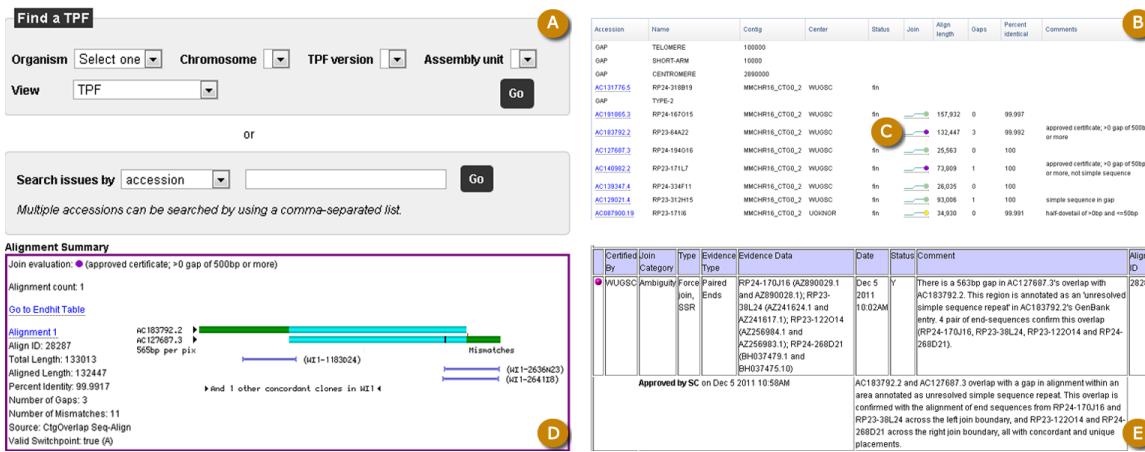


Figure 11. A: Search interface for TPF pages. B: Detail from TPFOverview page for the mouse chromosome 16 TPF showing enhanced TPF file display table. Clicking on any join evaluation icon (C) will take a user to the OverlapView page for the specified pair. D: Detail from OverlapView page for the highlighted join in B showing graphical rendering of the alignment and alignment summary details. E: Join certificate for the alignment shown in D. The certificate provides external evidence supporting the alignment and an explanation of the major alignment issues. All certificates are reviewed prior to approval.

References

1. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931–45. PubMed PMID: 15496913.
2. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*. 2009;7(5):e1000112. PubMed PMID: 19468303.
3. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. *Nature genetics*. 2004;36(9):949–51. PubMed PMID: 15286789.
4. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nature genetics*. 2006;38(1):75–81. PubMed PMID: 16327808.
5. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature genetics*. 2006;38(1):82–5. PubMed PMID: 16327809.
6. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. *Nature genetics*. 2005;37(7):727–32. PubMed PMID: 15895083.
7. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*. 2006;16(9):1182–90. PubMed PMID: 16902084.
8. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007;318(5849):420–6. PubMed PMID: 17901297.

9. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008;453(7191):56–64. PubMed PMID: 18451855.
10. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental duplications and copy-number variation in the human genome. *American journal of human genetics*. 2005;77(1):78–88. PubMed PMID: 15918152.
11. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. *PLoS biology*. 2011;9(7):e1001091. PubMed PMID: 21750661.
12. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS genetics*. 2007;3(4):e63. PubMed PMID: 17447845.

Eukaryotic Genome Annotation Pipeline

François Thibaud-Nissen, PhD,¹ Alexander Souvorov, PhD,¹ Terence Murphy, PhD,¹ Michael DiCuccio, MD,¹ and Paul Kitts, PhD¹

Created: November 14, 2013.

Scope

The NCBI Eukaryotic Genome Annotation Pipeline is an automated pipeline producing annotation of coding and non-coding genes, transcripts, and proteins on finished and unfinished public genome assemblies. It provides content for various NCBI resources including Nucleotide, Protein, BLAST, Gene, and the Map Viewer genome browser. The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of raw and curated data from public repositories (sequence and [Assembly](#) databases) through the alignment of sequences and the prediction of genes, to the submission of the accessioned and named annotation products to public databases.

Core components of the pipeline are the alignment programs Splign (1) and ProSplign, and Gnomon, a gene prediction program combining information from alignments of experimental evidence and from models produced *ab initio* with an HMM-based algorithm.

The annotation pipeline produces comprehensive sets of genes, transcripts, and proteins derived from multiple sources, depending on the data available. In order of preference, the following sources are used:

1. RefSeq curated annotated genomic sequences (2), such as the human beta globin gene cluster located on chromosome 11 (NG_000007.3)
2. Known RefSeq transcripts (2)
3. Gnomon-predicted models

Both the set of genes and the placements of the genes in the annotation on the genomic sequences comprise the output of the annotation pipeline.

Organisms in scope

Those eukaryotic organisms annotated by NCBI span a wide range of taxa among invertebrates, vertebrates, and plants. Annotation priorities are based on several considerations, including:

¹ NCBI; Email: thibaudf@ncbi.nlm.nih.gov; Email: souvorov@ncbi.nlm.nih.gov; Email: murphyte@ncbi.nlm.nih.gov; Email: dicuccio@ncbi.nlm.nih.gov; Email: kitts@ncbi.nlm.nih.gov.

- National Institutes of Health (NIH) priorities: Mammals are important to the NIH, so high-quality genome assemblies for new mammalian species are given a higher priority for annotation
- Biological or economic importance: highly-studied organisms or organisms with agricultural (e.g., crops) or industrial use
- Community interest/requests: requests from research communities, communicated in person or in writing through the NCBI Support Center. To write to the NCBI Support Center, click on the “Support Center” link in the bottom right corner of any NCBI web page.

The annotation process depends heavily on the availability of transcript or protein evidence for the species. Some annotation plans for high-priority organisms may be put on hold pending submission and public availability of transcriptome data.

Assemblies in scope

Only genomes with assemblies that are public in the International Nucleotide Sequence Database Collaboration (INSDC) ([DNA Data Bank of Japan](#), [European Nucleotide Archive](#) or [GenBank](#)) are considered for annotation. These assemblies are available in the [Assembly](#) resource. Assemblies with assembled chromosomes are preferred, but assemblies made of unplaced scaffolds only may also be annotated. Assemblies for which only contigs are available are not annotated.

Assemblies with high contig and scaffold N50 are prioritized. No single quality metric is used as a strict threshold, but assemblies that have a contig N50 above 50,000 bases and/or a scaffold N50 above 2,000,000 bases are preferred, as more complete gene sets are generally produced for assemblies with higher N50 statistics. NCBI may decide not to annotate assemblies that are extremely fragmented, even if they meet other criteria.

If multiple assemblies are available for the same organism, NCBI will annotate the higher quality assembly as the reference. Alternate assemblies of lower quality may also be included. This decision depends on the quality of the alternate assemblies, their importance to the community, as well as the estimated gain from annotating extra assemblies (number of extra genes identified, compensation of low-quality regions in the reference by higher-quality regions in the alternate assembly, value to variation studies).

Some assemblies are submitted to INSDC with annotation. NCBI may elect to propagate this annotation onto RefSeq sequences. This is typically the case for model organism assemblies with well-curated annotation, such as *Drosophila melanogaster* (maintained by FlyBase), *Saccharomyces cerevisiae* (maintained by the Saccharomyces Genome Database) or *Caenorhabditis elegans* (maintained by WormBase) but annotation propagation from GenBank to RefSeq records may also be done for other organisms (e.g., *Sorghum bicolor*). For some organisms with annotation submitted to INSDC (e.g., *Ailuropoda melanoleuca*), NCBI may opt to annotate RefSeq copies of the assemblies, primarily to provide a more consistent RefSeq dataset across organisms of interest to the NIH.

History

NCBI's original Eukaryotic Genome Annotation Pipeline began development in the year 2000 to annotate draft versions of the human genome assembly produced by the Human Genome Project. NCBI's annotation process has grown over the last 13 years to accommodate non-human organisms. It has also become an automated pipeline that annotates more feature types using a wider range of input data and new or improved algorithms.

In its infancy, NCBI's Eukaryotic Genome Annotation Pipeline was a semi-manual process to annotate known genes by aligning mRNAs from GenBank and RefSeq to the genome using BLAST (3), and to generate *ab initio* gene model predictions in the spaces between the known genes with GenomeScan (4) guided by protein alignments. One early advance was to use EST alignments to produce model transcripts that represented EST and mRNA chains that shared introns. Another major improvement came in 2003 when Gnomon, a gene prediction program developed at NCBI based on GenScan (5), replaced GenomeScan. Gnomon allowed us to generate gene models using a combination of mRNA, EST and protein alignments as evidence, supplemented by *ab initio* prediction where evidence was lacking. The next major advance was the development and incorporation of splicing-aware alignment algorithms capable of placing transcripts and proteins independently while following established rules of eukaryotic splicing. NCBI's first splicing-aware transcript alignment program, Spidey (6), was developed as a research project but this program did not scale to very large data sets and it was not sufficiently robust for routine use in our annotation pipeline. Splign (1) was developed as a replacement for Spidey and was incorporated into the annotation pipeline in 2004. Splign allowed accurate placement of transcripts and aided efforts to identify problematic areas of both the genome and the transcript set. ProSplign, NCBI's splicing-aware protein alignment program, was incorporated into the annotation pipeline in 2006 to improve the accuracy of the protein-to-genomic sequence alignments used as evidence in the Gnomon gene model prediction process. In 2013, NCBI made another major enhancement to the annotation process that allowed effective use of RNA-Seq data as evidence for making transcript models. This greatly improved the quality of the annotation for many organisms that have little or no mRNA or EST data in GenBank.

As the rate that new genome assemblies deposited in GenBank increased, deficiencies in the annotation pipeline that limited our ability to scale the process beyond a small number of organisms became more apparent. In parallel to the improvements to the annotation algorithms described above, we twice re-engineered the existing process to create a new framework for parallel execution that also provides extensibility, robustness, tracking, and reproducibility. By 2009, development of the re-engineered pipeline was sufficiently advanced to switch production annotation runs from the old pipeline to the new framework. Further refinements to the process and more automation continue to improve throughput. In 2011, we annotated twice as many eukaryotic genomes as in any

previous year and as of the second half of 2013 are releasing an average of 8 eukaryotic genome annotations per month.

Dataflow

Methods

Alignments

Both Splign (1) and ProSplign are global alignment tools that enable alignment of transcripts and proteins with high resolution of splice sites. The computational cost of these algorithms requires that approximate placements of the query sequences (transcripts or proteins) on the target (genome) be first identified with a local alignment tool, such as BLAST. Since a query often aligns at multiple locations, the BLAST hits are analyzed by the Compart algorithm to identify compartments prior to running Splign or ProSplign.

BLAST

See the BLAST chapter.

Compart algorithm

A compartment is defined as a sequence of compatible hits. Two BLAST hits are said to be compatible if they follow the natural flow of the target sequence. On a given strand, the relative position of the hits should be the same on both the query sequence and the genome. Compatible hits may overlap but may not be contained within one another. This definition of compatibility is transitive.

The Compart algorithm finds all non-overlapping compact compartments on the genome for a given query using a maximal coverage algorithm. Each compartment is assigned coverage, Φ^c , which is a measure of how well it represents the target sequence:

$$\Phi^c = \sum_h w^h L_{\text{eff}}^h$$

In this equation L_{eff}^h is the effective length of the hit h . It is usually the hit length, but if the hit overlaps with a neighbor hit, its effective length is decreased by a half of the overlap.

For cDNA alignments, where most useful hits are of very high identity, the weight w^h equals the identity of the hit and the coverage Φ^c is the number of matches. For protein alignments, the weight is a constant equal 1. In this case the coverage Φ^c is simply the target sequence length covered by the hits.

When there is more than one compartment, the query sequence is covered multiple times, and to a certain extent finding all compartments is equivalent to maximization of the total coverage. In the case of exon duplication events, the additional hits should be ignored rather than turned into additional compartments. Since typically only a relatively small

portion of the gene is duplicated we introduce a penalty P_{new} for an additional compartment. This penalty ensures that a new compartment is created only if there is enough gene material for it. The value of this parameter is usually 25%–40% of the target sequence length. So our maximal coverage algorithm finds the compartments configuration that maximizes the following total coverage:

$$\Phi = \sum_c (\Phi^c - P_{\text{new}})$$

The process of optimization is performed very effectively using the dynamic programming algorithm.

Splign – Transcript alignment

Splign (1) is a tool for aligning spliced cDNA sequences against their genomic counterparts using pre-computed compartments. The program produces accurate spliced alignments via solving a score S optimization problem formulated specifically to account for splice signals and introns.

$$S = B_m N_m - P_{\text{mis}} N_{\text{mis}} - \sum_{\text{gaps}} (P_{\text{gopen}} + P_{\text{gextend}} l) - \sum_{\text{introns}} (P_{\text{iopen}} + P_{\text{iextend}} l)$$

In this formula B_m and N_m are the bonus for a match and the number of matches, P_{mis} and N_{mis} , are the penalty for a mismatch and the number of mismatches, P_{gopen} and P_{gextend} , are the penalties for opening and extending a gap. These parameters are similar to the ones used in Blastn. The introns are accounted for by introduction of a special type of gap with P_{iopen} and P_{iextend} as the penalties for an opening and extending an intron. The formulation discriminates between the most frequent consensus (GT/AG), less frequent consensus (GC/AG, AT/AC), and not consensus donor/acceptor sites by giving different values to P_{iopen} .

Since the complexity of solving the global sequence alignment problem is proportional to the product of lengths of the sequences, the hits are arranged into compartments as described above and the dynamic programming matrix split into smaller blocks by seeding the global alignment with the high identity portion of the hits (Figure 1).

For each compartment, its genomic search space is expanded by the length of query cDNA ends not covered by local alignments. This allows detecting the end exons if they are missed by the local alignment tool for reasons such as the alignment length being shorter than the word size or the exon residing in a masked region. Each hit may correspond to an exon, a part of an exon, or even a number of exons. Therefore, it is important to be conservative when using local alignments for alignment seeding. Within each compartment, parts of alignments that overlap on the query are dropped. From the remaining alignments, the longest perfectly matching diagonals are extracted, and the cores are used to seed the global alignment.

Hits comprising compartments determine whether the query and the subject sequence align on the same strand. Most mRNA sequences have natural biological order and

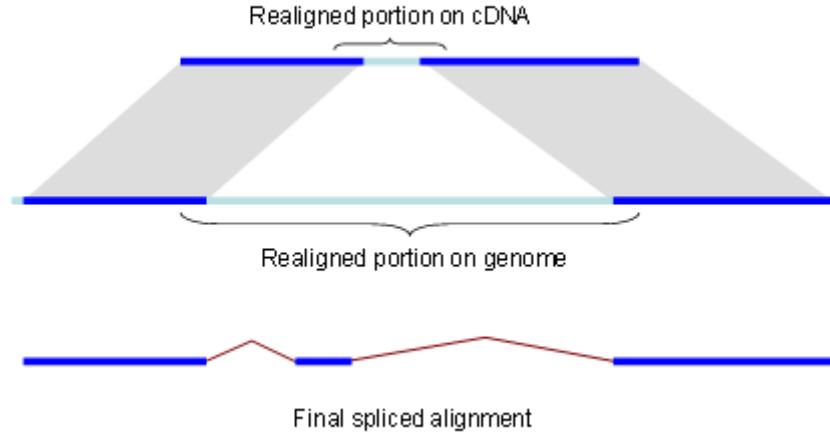


Figure 1. Salign reduces the computational complexity by using the high identity portions of the hits (dark blue) for the bulk of the alignment and realigning only small portions of the transcript (light blue).

positive strand can be assumed when aligning them. On the contrary, EST and frequently RNA-Seq sequences are not oriented, so both the original sequence and its reverse complimentary have to be aligned and the strand is determined by comparing the resulting alignments.

ProSalign – Protein alignment

Protein alignments are produced by ProSalign. Similarly to Salign, ProSalign is a global protein-to-genome alignment tool that produces accurate spliced alignments from pre-computed compartments. ProSalign uses a modified Needleman Wunsch type (7) global alignment algorithm for aligning. ProSalign scores the target protein sequence against translation of the genomic sequence using the following score:

$$S = \sum_{\text{diag}} S_{\text{diag}} - \sum_{\text{gaps}} (P_{\text{gopen}} + P_{\text{gextend}}l) - \sum_{\text{introns}} (P_{\text{iopen}} + P_{\text{iextend}}l)$$

where S_{diag} is the score for an ungapped part of the alignment calculated using a BLOSUM62 matrix (8). Insertions and deletions for which the length is a multiple of three are scored with the default Blastp gap penalties P_{gopen} and P_{gextend} . Gaps for which the length is not a multiple of three are frameshifts and have a much higher opening penalty P_{gopen} . The introns are scored as a special type of gap with a very small extension cost and an opening cost which is different between the most frequent consensus splices (GT/AG), less frequent consensus splices (GC/AG, AT/AC), and non-consensus splice sites.

Unlike Salign, ProSalign doesn't use seeds because Blast hits for cross-species proteins do not give reliable information about seeds. Instead, ProSalign aligns the protein against a slightly extended genomic region identified by Compart as the compartment.

Not all parts of a protein are conserved well enough to provide a reliable alignment. In fact, some parts may not correspond to anything on the genome. The global alignment algorithm will align the whole protein, rendering a very low-identity alignment for the non-conserved portions of the protein. These unreliable and often misleading pieces of the alignment are filtered out by ProSplign during a post-processing step.

Gene prediction

Gnomon is a two-step gene prediction program maintained by NCBI. The Chainer algorithm assembles overlapping alignments into “chains” and is followed by the *ab initio* prediction step which extends these chains into complete models and creates full *ab initio* models, using a Hidden Markov Model (HMM).

Chainer

Spliced alignments obtained using Splign and ProSplign are likely partial, either because the aligned sequences are partial or in the case of protein alignments because only conserved portions of the protein could be aligned. Chainer analyzes and assembles these partial alignments to provide longer gene models and additional information about alternative variants.

Because of their short length and high redundancy, RNA-Seq alignments with identical introns are first combined into single alignments with larger weights (Figure 2). Boundaries of these “micro-chains” don’t cross splices known from other alignments and their extension is limited to 20 bp.

These “micro-chains” are then combined by Chainer with cDNA and protein alignments based on their exon structure compatibility using a modified version of the Maximal Transcript Alignment algorithm (9) based on frame compatibility of the coding regions. For protein and annotated full-length cDNA alignments, the coding regions can be inferred. For other cDNA alignments, possible coding regions are predicted and scored using a 3-periodic fifth-order Markov model for coding propensity and Weight Matrix Method (WMM) models for splice signals and translation initiation and termination signals (10). All cDNAs with coding sequence (CDS) scores above a given threshold are marked as coding, and the CDS information is used when assembling chains. In many cases, this process determines the orientation of an EST if it was unknown before. RNA-Seq and some EST alignments are too short to score above the threshold and, if they are not spliced, their orientation is often also unknown. For these alignments, Chainer will consider that these sequences can be part of the 5' end and harbor a start codon, or be part of the 3' end and harbor a stop codon, or be internal to the CDS or to an untranslated region (UTR), and select the scenario that contributes to the longest CDS.

Afterward, UTRs are added if the necessary translation initiation or termination signals are present. There are no restrictions on the extension of a 5'-UTR other than the exon-intron structure compatibility.



Figure 2. Combining the alignments with the same introns into one alignment (micro chaining) reduces the computational complexity.

The assembled full-length chains that share splices or CDS are combined into genes with alternative isoforms. Among the partial chains for a gene, the variant with the longest CDS is selected for extension by *ab initio* prediction.

HMM-based prediction

The core algorithm of the *ab initio* prediction capability of Gnomon is based on Genscan (5), which uses a 3-periodic fifth-order HMM for the coding propensity score and incorporates descriptions of the basic transcriptional, translational and splicing signals, as well as length distributions and compositional features of exons, introns and intergenic regions. The most important distinction of Gnomon from Genscan and other *ab initio* prediction programs is its ability to conform to the supplied alignments and extend and complement them when necessary.

Mathematically, an HMM-based *ab initio* prediction is a search in the gene configuration space for the gene that provides the maximal score. If all configurations that are not compatible with the available alignments are excluded from the search space, then the optimization process in the resulting collapsed space will yield a gene configuration that is possibly suboptimal from the *ab initio* point of view but exactly follows the experimental information available. This approach allows extension or connections of partial alignments (Figure 3). Untranslated regions, if present in the alignments, are also included in the gene model.

Gnomon recognizes as HMM states coding exons and introns on both strands and intergenic sequences. Translational and splice signals are described using WMM (10) and WAM (11) models. A 12-bp WMM model, beginning 6 bp prior to the initiation codon, is used for the translation initiation signal (12). A 6-bp first order WAM model starting at the stop codon is used for the translation termination signal. The donor splice signal is described by a 9-bp second order WAM model, and the acceptor splice signal is described by a 43-bp second order WAM model. Both donor and acceptor models include 3-bp of the coding exon. Coding portions of exons are modeled using an inhomogeneous 3-

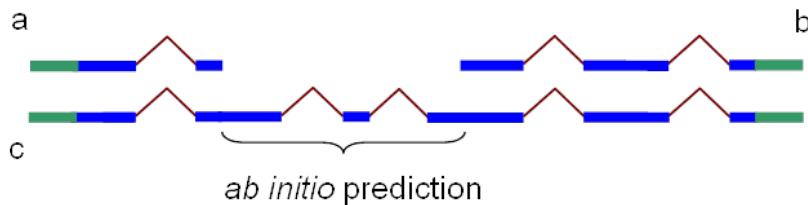


Figure 3. Partial chains a and b produced by Chainer may be combined into one chain, c, by addition of the HMM prediction of missing coding sequence. In blue: coding sequence. In green: untranslated region

periodic fifth-order Markov model (13). The noncoding states are modeled using a homogeneous fifth-order Markov model.

Input data

Assemblies

The Eukaryotic Annotation Pipeline can annotate one or multiple assemblies at once (see below). All assemblies must be publicly available in the Assembly database. Since the INSDC sequence records constituting the submitted assemblies are owned by submitters and may not be modified by NCBI, all annotation is done on RefSeq copies of the INSDC assemblies. Prior to the annotation process, RefSeq accessions are assigned to the assembly's scaffolds and chromosomes. These RefSeq sequences are based on the sequences in the INSDC records, but their records will bear the NCBI annotation. Note also that a new assembly accession, with the prefix GCF_, is given to the assembly which contains the RefSeq sequences.

Source of evidence

The evidence used to predict gene models is selected from available public data. Same-species transcripts, proteins, and short reads, and if not sufficient, transcripts and proteins from closely related species are included.

More specifically the following sets of transcripts are included:

- Known [RefSeq](#) transcripts: coding and non-coding [RefSeq](#) transcripts, with NM_ or NR_ prefixes respectively. These are generated by NCBI staff based on automatic processes, manual curation, or data from collaborating groups (see more details in the RefSeq chapter and 2)
- Other long transcripts
 - [GenBank](#) transcripts from the taxonomically relevant GenBank divisions, and the Third-Party Annotation ([TPA](#)), High-throughput cDNA (HTC) and Transcriptome Shotgun Assembly ([TSA](#)) divisions
 - ESTs from [dbEST](#)

- Long RNA-Seq sequences (e.g., from the GS FLX TITANIUM 454 platform) from the Sequence Read Archive [SRA](#)
- Short read RNA-Seq data available in [SRA](#)

And the following proteins:

- Known RefSeq proteins, with NP_ prefixes
- INSDC proteins derived from transcripts (as much as possible, conceptual translations are excluded)

In addition, if available for the annotated organism, curated RefSeq genomic sequences are used. These sequences have accessions with NG_ prefixes and represent non-transcribed pseudogenes, manually annotated gene clusters that are difficult to annotate via automated methods, or human [RefSeqGene](#) records (2).

Process flow

Figure 4 provides an overview of the annotation pipeline. Transcripts from RefSeq, GenBank, and the Sequence Read Archive, proteins, and, if available, RefSeq curated genomic sequences are aligned to the masked genome. Gene models are predicted by Gnomon based on these alignments, and searched against the curated database UniProtKB/SwissProt. The final set of models is then chosen among the Gnomon predictions (model RefSeq) and the known and curated RefSeq. Names and type of loci and GeneIDs are assigned to model RefSeq and retrieved from the Gene database for known RefSeq. In the final steps, the annotation is formatted, submitted to the sequences databases, and published.

Fetching of inputs

All evidence identifiers are retrieved from Entrez at the very beginning of the annotation run and the date of sequence retrieval is tracked and reported as the annotation run “freeze” date. Any sequence added to archival databases after that day will not be used.

Genome sequence masking

The assemblies are retrieved from the Assembly resource and masked using either WindowMasker (14) or RepeatMasker (15). RepeatMasker is generally used for organisms for which a comprehensive repeat library is available.

Alignment of curated RefSeq genomic sequences

If available for the organism of interest, curated RefSeq genomic sequences are aligned to the masked genome using BLAST. The alignments are ranked and filtered based on identity, coverage, and placement information kept in a RefSeq tracking in-house database. The features annotated on the alignments passing the filter are then projected onto the genomic sequences and evaluated with the other aligning evidence when choosing the best model.

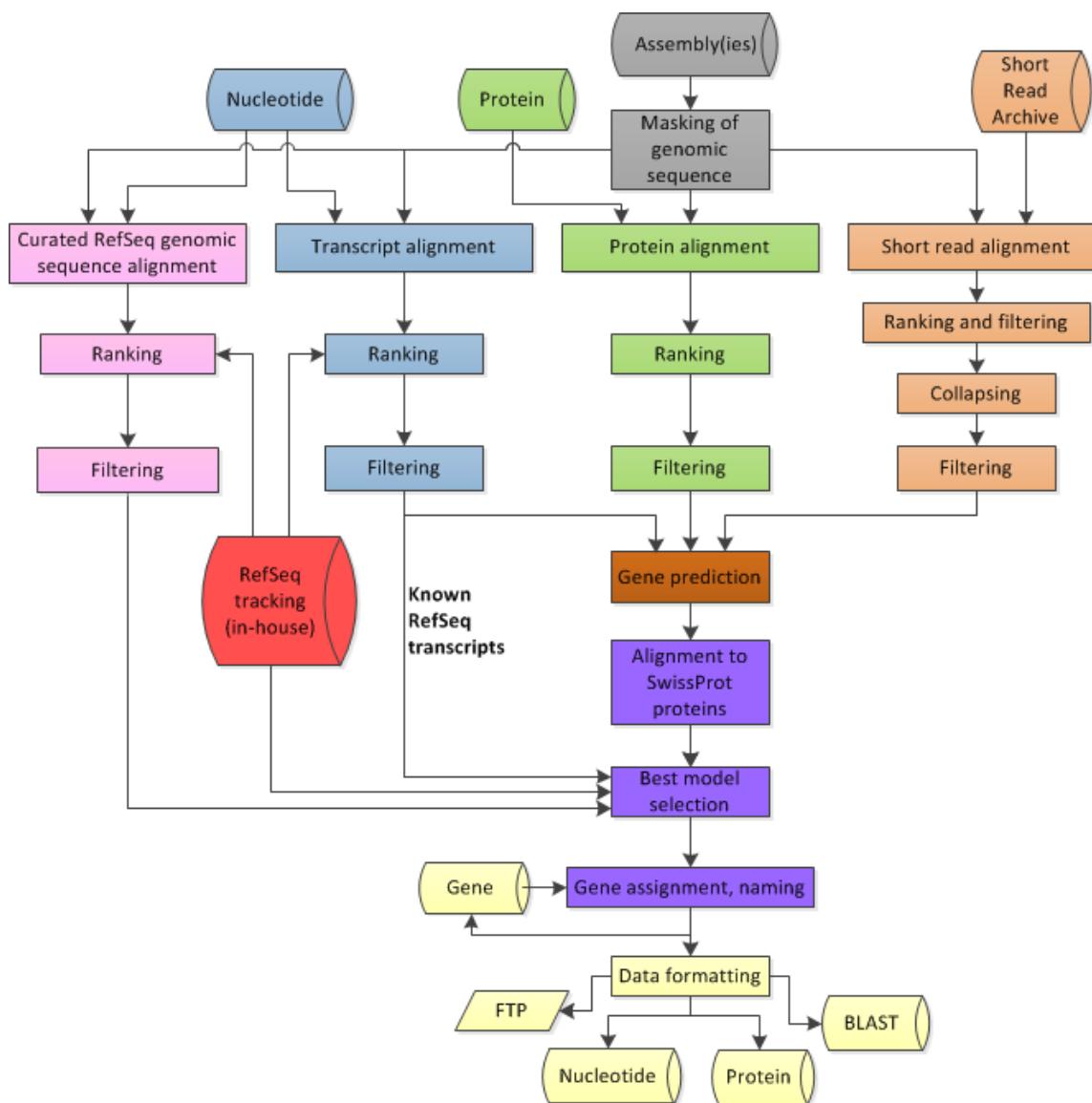


Figure 4. Overview of the process flow in the Eukaryotic Genome Annotation Pipeline. In grey: genomic sequence preparation; in blue: alignments of transcripts; in green: alignment of proteins; in orange: alignment of short reads; in pink: alignment of curated genomic sequences (if available); in brown: gene prediction based on all available alignments; in red: internal tracking database of RefSeq sequences; in purple: selection of the best models and protein naming; in yellow: formatting of annotation sets for deployment to public resources.

Alignment of protein and transcript evidence

After retrieval, sequences are aligned to the masked genome following this general strategy: sequences are aligned locally to the genome using BLAST. Based on the BLAST hits, Compart identifies genomic compartments to which query sequences are re-aligned globally. This second round of alignments is necessary for accurate determination of splice sites and for the identification of small terminal exons that may be missed by BLAST. The

global alignments are performed by Splign for transcripts and ProSplign for proteins. Resulting alignments are then ranked based on coverage and identity and filtered before hand-off to downstream tasks. Adjustments to the alignments and filtering parameters, and variation to this general dataflow are made based on the source and characteristics of the evidence and are described below.

Alignment of known RefSeq transcripts

Since many of the known RefSeq sequences are curated (most notably for Vertebrates) and, as such, are high-value targets when annotating a genome, special attention is given to their proper placement. Masking may interfere with the alignment process, so RefSeq transcripts for which all alignments on the masked genome are under a coverage threshold may be re-aligned to the unmasked genome.

The alignments are ranked and filtered based on adjustable criteria (such as coverage, identity, rank) as well as location information contained in the RefSeq tracking database. Typically, only the best-placed alignment for a given query is selected for use in downstream steps.

Alignment of non-RefSeq transcripts

INSDC mRNAs, ESTs and 454 sequences are first screened against a database of mitochondrial sequences, cloning vectors, adaptors, bacterial IS-elements and repetitive sequences, and excluded from further processing if a large portion of their sequence hits a contaminant. In addition, transcripts identified as low-quality by curation staff are screened out.

Following this initial screen, the sequences are aligned with BLAST and Splign, as explained above, and ranked and filtered. For a given transcript, typically only the best-placed alignment (rank 1) is selected. For sequences that cannot be oriented (e.g., unspliced ESTs), alignments to both strands are passed downstream. If used, cross-species transcripts are aligned with more stringent criteria than same-species transcripts to insure that only the most-likely ortholog transcript is passed downstream.

Alignment of proteins

Similarly to transcripts, proteins are first screened against a database of repeats and the curated list of low-quality transcripts. Proteins are then aligned to the masked genome with BLAST and ProSplign. The alignments are further ranked and filtered and passed to the gene prediction step.

Alignment of short reads

Short reads (RNA-Seq) available in the SRA can be used for gene prediction. A specific dataflow was engineered to handle the large volume and short length of sequences produced by new generation sequencing technologies.

RNA-Seq data from so-called next-generation sequencing platforms present several challenges for use in gene prediction. First, the reads are substantially shorter than

conventional transcript data such as ESTs and mRNAs, so an individual read contains relatively little information. For example, typically only 5-25% of reads from the Illumina platform span an intron, which is the most useful data for building gene models. Second, the reads are extremely numerous and redundant, with highly-expressed genes being represented by tens of millions of reads. This presents a challenge for throughput. And third, the depth of coverage results in apparent background expression in most of the genome that isn't desirable to represent in the final gene models.

The annotation pipeline addresses these issues in several ways to reduce the complexity of the RNA-Seq data and convert it to a form useful for gene predictions:

1. Datasets and associated metadata are obtained from the SRA and [BioSample](#) databases, enabling robust tracking of evidence.
2. The reads are "uniquified" so that 100% identical sequences are aligned only once.
3. Unique reads are aligned, ranked, and filtered for high identity and coverage alignments.
4. Alignments with the same splice structure and the same or similar start and end points are collapsed into a single representative alignment. The number of reads from each SRA run is tracked for each collapsed alignment.
5. Alignments containing rare introns or that represent apparent noise or background are filtered from the dataset.

Taken together, these steps reduce the size and complexity of a typical RNA-Seq dataset by 100-1000x. The resulting collapsed alignments can be used by themselves or combined with transcript and/or protein alignments for the gene prediction step.

Gene prediction by Gnomon

Protein transcript and short read alignments are passed to Gnomon for gene prediction. Chainer assembles alignments with the same exon structure and with coding regions in compatible frames into putative models. Gnomon then extends the models missing a start or a stop codon or internal exon(s) using an HMM-based algorithm. Gnomon additionally creates pure *ab initio* predictions where open reading frames of sufficient length but with no supporting alignment are detected (see Methods).

This first set of predictions is further refined by alignment against a subset of the nr (non-redundant) database of protein sequences. The additional alignments are added to the initial alignments and the chaining and *ab initio* extension steps are repeated. The results constitute the set of Gnomon predictions.

Alternate variants, complete or partial, may be produced for each gene.

Frameshifts, indels, and stop codons may occur in the resulting Gnomon predictions. They reflect sequence differences between the input transcript and protein alignments and the genome assembly.

Annotation of small RNA

tRNAs are annotated using tRNAScan-SE (16). Other small RNAs are annotated by placement of same-species curated RefSeq transcripts. Hence, these are only part of the annotation if they were incorporated in the RefSeq set for the organism being annotated. Currently the RefSeq set may include small RNAs identified by curation, collaboration, or external sources, which is currently limited to microRNAs obtained from miRBase (17).

Choosing the best model(s)

The final set of annotated features comprises, in order of preference, pre-existing known RefSeq sequences and a subset of well-supported Gnomon-predicted models. It is built by evaluating together at each locus the known RefSeq transcripts, the features projected from the curated RefSeq genomic alignments, and the models predicted by Gnomon.

Models based on known and curated RefSeq

RefSeq transcripts are given precedence over overlapping Gnomon models with the same splice pattern. Alignments of known same-species RefSeq transcripts or curated genomic sequences are used directly to annotate the gene, RNA, and CDS features on the genome. Since the RefSeq sequence may not align perfectly or completely to the genomic sequence, a consequence of this rule is that the annotated product may differ from the conceptual translation of the genome.

Models based on Gnomon predictions

Gnomon predictions are included in the final set of annotations if they do not share all splice sites with a RefSeq transcript and if they meet certain quality thresholds including:

- Only fully- or partially-supported Gnomon predictions, or pure *ab initio* Gnomon predictions with high coverage hits to UniProtKB/SwissProt proteins are selected.
- When multiple fully-supported transcript variants are predicted for a gene, only the Gnomon predictions supported in their entirety by a single long alignment (e.g., a full-length mRNA) or by RNA-Seq reads from a single BioSample are selected.
- Poorly-supported Gnomon predictions conflicting with better-supported models annotated on the opposite strand are excluded from the final set of models.
- Gnomon predictions with high homology to transposable or retro-transposable elements are excluded from the final set of models.

Integrating RefSeq and Gnomon annotations

As a result of the model selection process, a gene may be represented by multiple splice variants, with some of them known RefSeq and others model RefSeq (originating from Gnomon predictions).

Gnomon predictions selected for the final annotation set are assigned model RefSeq accessions with XM_ or XR_ prefixes for protein-coding and non-coding transcripts, respectively, and XP_ prefixes for proteins to distinguish them from known RefSeq with

NM_/_NR_ and NP_ prefixes. Model RefSeq can be searched in Entrez with the query “srcdb_refseq_model[properties]” while known RefSeq sequences can be obtained with the query “srcdb_refseq_known[properties]”.

Locus typing and protein naming

Genes are categorized into different locus types according to the type and quality of the model and based on orthology information.

- Known RefSeq features are annotated according to their locus type (e.g., protein-coding vs. pseudogene) established before the annotation run.
- Most Gnomon models with insertions, deletions, or frameshifts are labeled as pseudogenes and annotated without a CDS feature or protein product.
- Gnomon models that appear to be single-exon retrocopies of protein-coding genes may also be annotated as pseudogenes.
- Gnomon models with insertions, deletions, or frameshifts may be considered coding if they have a strong unique hit to the SwissProt database or appear to be orthologs of known protein-coding genes. Titles for these models are prefixed with “PREDICTED: LOW QUALITY PROTEIN.” There may be defects in the assembly and/or the model in these cases.
- Gnomon models that have no predicted CDS or a short CDS with no supporting alignments may be annotated as non-coding models or removed from the annotation.
- When multiple assemblies are annotated, a partial or imperfect model may be called coding because a complete model exists at the corresponding locus on one of the other annotated assemblies.

Gene and protein names are assigned based on the locus type, protein homology, and orthology information, and data from the [Gene](#) database, which may in turn be based on nomenclature from an external group such as the HUGO Gene Nomenclature Committee (HGNC). Predicted genes are evaluated for orthology to genes in a reference species using a pairwise comparison process based on protein alignments and local synteny information.

If a likely ortholog can be determined, the gene symbol and name is transferred from the reference species, if applicable.

If an ortholog cannot be determined, predicted genes are named based on the name of the most similar SwissProt protein, adding the suffix ‘-like’ to indicate the putative nature of the assignment.

Predicted genes for which no name can be determined are assigned a generic gene and protein name of the form “uncharacterized LOC” plus the GeneID.

Assignment of GeneIDs

Genes in the final set of models are assigned GeneIDs in the [Gene](#) database.

- A gene represented by at least one known RefSeq transcript receives the GeneID of the RefSeq transcript(s).
- Genes mapped from a previous annotation (see Re-annotation below) are assigned the same GeneIDs as in the previous annotation.
- Genes that are not mapped from a previous annotation and genes that are represented by Gnomon models only are assigned new GeneIDs.
- Genes mapped to equivalent locations on co-annotated assemblies are assigned the same GeneIDs (see Annotation of multiple assemblies).

Packaging of the annotation

The output of the annotation pipeline is labelled with an Annotation Release number. For a given annotation, the combination of organism and Annotation Release number (e.g., NCBI Homo sapiens Annotation Release 105) is used throughout NCBI as a way to uniquely identify annotation products originating from the same annotation run.

The annotation pipeline output is composed of the scaffolds and the chromosomes of the assembled genome(s) annotated with the genes, RNAs and proteins as features, and also the RNAs and proteins themselves. The RefSeq scaffolds and chromosomes are assigned accessions with NW_ or NT_ and NC_ prefixes and submitted to the Nucleotide database with the features annotated. Sequences submitted to the sequence databases are labelled with the Annotation Release (Figure 5).

The annotated products may include known RefSeq transcripts and proteins, Gnomon-predicted models and tRNA genes that were predicted by tRNAscan-SE. The Gnomon models that were retained by the best model selection process are submitted to the Nucleotide, Protein, and Gene database and the tRNAs genes are submitted to Gene. The known RefSeq features are updated independently from the annotation process and are not re-submitted to the sequence or Gene databases (see Access section below). The origin of the annotation can be deduced from the \note on the feature annotated on the genomic sequences (Table 1).

For transcripts and proteins produced by Gnomon, the sequence records provide the level of support for predicted models. For low-quality proteins, the records also detail the difference between the model and the genomic sequences that was introduced to compensate for a possible error in the assembly (Figure 6).

As explained above, a known RefSeq transcript may not align perfectly to the genome but may be selected as a gene representative in the set of annotation products. These discrepancies are noted on the genomic sequence records (Figure 7).

LOCUS NW_004457742 5497 bp DNA linear CON 06-MAY-2013
DEFINITION *Dasypus novemcinctus* isolate 3-136 unplaced genomic scaffold,
Dasnov3.0 Scaffold2, whole genome shotgun sequence.
ACCESSION NW_004457742 GPS_001484838
VERSION NW_004457742.1 GI:477502311
DBLINK BioProject: PRJNA196486
Assembly: GCF 000208655.1 **A**
KEYWORDS WGS.
SOURCE *Dasypus novemcinctus* (nine-banded armadillo)
ORGANISM *Dasypus novemcinctus*
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Xenarthra; Cingulata; Dasypodidae; Dasypus.
COMMENT **B** REFSEQ INFORMATION: The reference sequence is identical to
JH561178.1.
Assembly name: Dasnov3.0
The genomic sequence for this RefSeq record is from the
whole-genome assembly released by the Baylor College of Medicine on
2012/01/06 (see
<http://www.hgsc.bcm.tmc.edu/content/armadillo-genome-project>). The
original whole-genome shotgun project has the accession
AAGV00000000.3.
C

```

##Genome-Annotation-Data-START##
Annotation Provider :: NCBI
Annotation Status   :: Full annotation
Annotation Version :: Dasypus novemcinctus Annotation Release 100
Annotation Pipeline :: NCBI eukaryotic genome annotation pipeline
Annotation Method   :: Best-placed RefSeq; Gnomon
Features Annotated :: Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##

```

Figure 5. Typical RefSeq record for a scaffold annotated by the Eukaryotic Genome Annotation Pipeline. (A) Links to the RefSeq BioProject and RefSeq assembly. (B) The comment field is prefixed with REFSEQ INFORMATION, and provides a link to the GenBank sequence on which the record is based. (C) The Genome Annotation structured comment provides the Annotation Release number and other information relating to the annotation process. ‘Annotation Status :: Full annotation’ and ‘Annotation Method :: Best-placed RefSeq; Gnomon’ indicate that the annotation used the placement of RefSeq sequences and Gnomon prediction as the source for the annotation.

LOCUS XM_004484857 1991 bp mRNA linear MAM 06-MAY-2013
 DEFINITION PREDICTED: *Dasyurus novemcinctus HHIP-like 1 (HHIPL1)*, mRNA.
 ACCESSION XM_004484857
 VERSION XM_004484857.1 GI:488509918
 DBLINK BioProject: PRJNA196486
 KEYWORDS .
 SOURCE *Dasyurus novemcinctus* (nine-banded armadillo)
 ORGANISM *Dasyurus novemcinctus*
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
 Mammalia; Eutheria; Xenarthra; Cingulata; Dasypodidae; Dasypus.
 COMMENT MODEL REFSEQ: This record is predicted by automated computational analysis. This record is derived from a genomic sequence (NW 004458804.1) annotated using gene prediction method: Gnomon.
 Also see: Documentation of NCBI's Annotation Process

```
##Genome-Annotation-Data-START##
Annotation Provider :: NCBI
Annotation Status   :: Full annotation
Annotation Version  :: Dasyurus novemcinctus Annotation Release 100
Annotation Pipeline :: NCBI eukaryotic genome annotation pipeline
Annotation Method   :: Best-placed RefSeq; Gnomon
Features Annotated :: Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##
```

FEATURES Location/Qualifiers
 source 1..1991
/organism="Dasyurus novemcinctus"
/mol_type="mRNA"
/isolate="3-136"
/db_xref="taxon:9361"
/chromosome="Unknown"
/sex="female"
/country="USA: National Hansen's Disease Programs at Louisiana State University, School of Veterinary Medicine"
 gene 1..1991
/gene="HHIPL1"
/note="Derived by automated computational analysis using gene prediction method: Gnomon. Supporting evidence includes similarity to: 10 Proteins, and 33% coverage by RNAseq alignments"
/db_xref="GeneID:101426433"
 CDS 94..1500
/gene="HHIPL1"
/note="The sequence of the model RefSeq protein was modified relative to its source genomic sequence to represent the inferred complete CDS: substituted 1 base at 1 genomic stop codon"
/codon_start=1
/transl_except=(pos:1195..1197,aa:OTHER)
/product="LOW QUALITY PROTEIN: HHIP-like 1"
/protein_id="XP_004484914.1"
/db_xref="GI:488509919"
/db_xref="GeneID:101426433"
/translation="MWQECRALFRHSPDRELWALEGNRAKFCRYLALDDVDYCFPRL"

Figure 6. Example of a RefSeq record for a transcript model predicted by Gnomon. (A) The title in the DEFINITION line is prefixed with PREDICTED (B) The comment field is prefixed with MODEL REFSEQ and indicates the gene prediction method and refers to the genomic sequence on which the model is annotated. (C) The note on the gene indicates the type and number of supporting evidence for the model. (D) The note on the CDS describes the modification that was done relative to the genomic sequence to produce the model. (E) The product name is prefixed with LOW QUALITY PROTEIN.

```

gene      <12978..15024
/gene="GAD3"
/gene_synonym="LeGAD3"
/note="glutamate decarboxylase isoform3; Derived by
automated computational analysis using gene prediction
method: BestRefSeq."
/db_xref="GeneID:100147723"
mRNA      join(<12978..13103,13177..13381,13461..13661,13739..13953,
14033..14286,14372..14424,14502..15024)
/gene="GAD3"
/gene_synonym="LeGAD3"
/product="glutamate decarboxylase isoform3"
/inference="similar to RNA sequence, mRNA (same
species) :RefSeq:NM_001246898.1"
/exception="annotated by transcript or proteomic data"
/note="The RefSeq transcript has 9 substitutions and 2
indels and aligns at 98% coverage compared to this genomic
sequence; Derived by automated computational analysis
using gene prediction method: BestRefSeq."
/transcript_id="NM_001246898.1"
/db_xref="GI:350538350"
/db_xref="GeneID:100147723"
CDS       join(13024..13103,13177..13381,13461..13661,13739..13953,
14033..14286,14372..14424,14502..14948)
/gene="GAD3"
/gene_synonym="LeGAD3"
/note="Derived by automated computational analysis using
gene prediction method: BestRefSeq."
/codon_start=1
/product="glutamate decarboxylase isoform3"
/protein_id="NP_001233827.1"
/db_xref="GI:350538351"
/db_xref="GeneID:100147723"

```

Figure 7. Example of a known RefSeq transcript annotated on a genomic scaffold. (A) The note on the gene indicates that the gene was annotated by projection of a best-placed RefSeq transcript on the genome. (B) The inference identifies the RefSeq transcript from which the annotation is inferred. (C) The note describes the alignment of the known transcript to the genomic sequence.

Table 1. Guide to the features annotated on scaffolds and chromosomes. The note provides information on the origin of the feature. *For predicted models, the note is also on the records of individual annotation products.

Annotated Product	Accession prefix	Origin of the product	Note provided for the feature annotated on scaffolds and chromosomes records*
Known transcripts/proteins	NM_, NR_, NP_	Curated RefSeq genomic alignment	Derived by automated computational analysis using gene prediction method: Curated Genomic
Known transcripts/proteins	NM_, NR_, NP_	Known RefSeq transcript alignment	Derived by automated computational analysis using gene prediction method: BestRefseq

Table 1. continues on next page...

Table 1. continued from previous page.

Annotated Product	Accession prefix	Origin of the product	Note provided for the feature annotated on scaffolds and chromosomes records*
Model transcripts/proteins	XM_, XR_, XP_	Gnomon	Derived by automated computational analysis using gene prediction method: Gnomon
tRNAs	no accession	tRNAscan-SE	tRNA features were annotated by tRNAscan-SE
RefSeq non-transcribed pseudogenes	no accession	Curated RefSeq genomic alignment	Derived by automated computational analysis using gene prediction method: Curated Genomic
Gnomon non-transcribed pseudogenes	no accession	Gnomon	Derived by automated computational analysis using gene prediction method: Gnomon
Full set of Gnomon predictions	no accession	Gnomon	Not in the sequence database. Available on the FTP site and as BLAST databases.

Special considerations

Annotation of multiple assemblies

When multiple assemblies of good quality are available for a given organism, the annotation of all is done in coordination. To ensure that matching regions in multiple assemblies are annotated consistently, assemblies are mapped to each other using a BLAST-based process prior to the annotation. The reciprocal best hits are used to pair corresponding regions on two assemblies.

As explained on Figure 8, these paired regions allow the coordinate ranking of the alignment of a given transcript on both assemblies.

This strategy ensures that mapped regions are annotated the same way and that the same genes are assigned the same GeneID and locus type on both assemblies. It reduces the redundancy in the Gene set for a given organism and helps navigation between multiple assemblies. Note that for Gnomon models, although a single GeneID represents the locus in multiple assemblies, a different transcript and protein accession is instantiated for each individual assembly.

For more on the assembly-assembly alignment process, see the Remapping Service chapter.

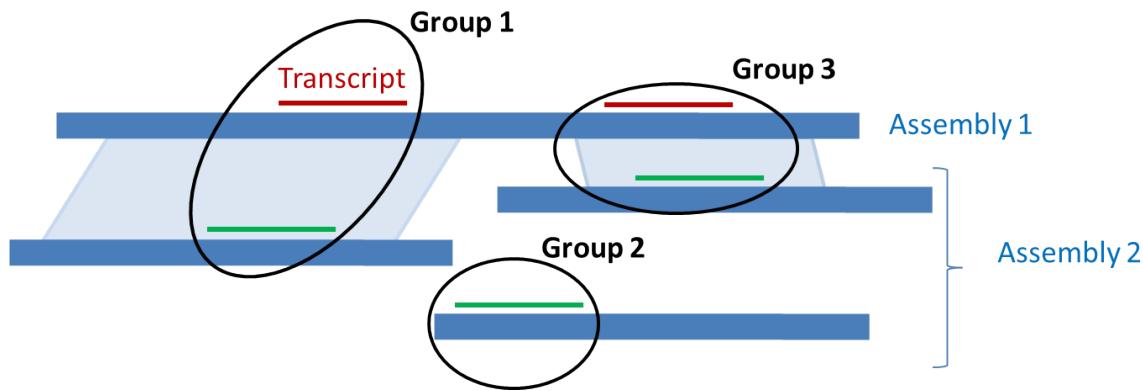


Figure 8. Ranking of alignments across multiple assemblies. Alignments of a given transcript are represented in red to Assembly 1 and in green to Assembly 2. If a genomic alignment exists between two regions harboring a transcript alignment (light blue parallelograms), the alignments in the paired regions are placed in the same group (Group 1 and Group 3). All alignments in a given group are given the same rank, different from the rank of other groups, based on the quality of the alignments.

Re-annotation

Special attention is given to tracking of models and genes from one release of the annotation to the next. Previous and current models annotated at overlapping genomic locations are identified and locus type and GeneID of the previous models are taken into consideration when assigning GeneIDs to the new models. If the assembly was updated between the two rounds of annotation, the assemblies are aligned to each other and the alignments used to match previous and current models in mapped regions.

Access

The status of annotation runs in progress or completed recently is updated nightly on the Eukaryotic Genome Annotation Pipeline public page:

http://www.ncbi.nlm.nih.gov/genome/annotation_euk/status/

This page provides links to the resources where data for a specific Annotation Release is available (Figure 9).

Products of NCBI's eukaryotic annotation pipeline are available in several resources (Table 2) including:

- In the [Nucleotide](#) and [Protein](#) databases
- In the [Gene](#) database
- On the [FTP](#) site in GFF, FASTA, GenBank flat file and ASN formats
- As [Map Viewer](#) tracks
- In BLAST databases available from organism-specific [BLAST](#) pages
- In the Consensus CDS project ([CCDS](#))

A					
Annotation runs in progress					
Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Status	
Bubalus bubalis (water buffalo)	UMD_CASPUR_WB_2.0	100	2013-10-25	Automated processing in progress	
Bos grunniens mutus (NA)	BosGru_v2.0	100	2013-10-22	Automated processing in progress	
Myotis brandtii (Brandt's bat)	ASM41265v1	100	2013-10-21	Automated processing in progress	
Haplochromis burtoni (Burton's mouthbrooder)	AstBur1.0	100	2013-09-26	Automated processing in progress	

B					
Recently completed annotation runs					
Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links
Xiphophorus maculatus (southern platyfish)	Xiphophorus_maculatus-4.4.2 (GCF_000241075.1)	100	2013-10-18	2013-10-23	FTP
Pundamilia nyererei (NA)	PunNye1.0 (GCF_000239375.1)	100	2013-09-26	2013-09-30	FTP

Figure 9. Public report of annotation runs (A) in progress and recently completed annotation runs (B). Information in the tables are linked to the Taxonomy database (Species), the Assembly database (RefSeq Assemblies), and resources where the data is available (Links). For each annotation run, the name of the Annotation Release, the Freeze date when the input data used for the annotation was fetched, and the Release date when the annotation was first made public are also provided.

Table 2. Availability of annotation products in NCBI resources.

Annotation products	In sequence databases	In Gene	In BLAST database	On the FTP site	In a Map Viewer track
Chromosomes	Yes	Yes	Yes	Yes	Yes
Scaffolds	Yes	No	Yes	Yes	Yes
Curated RefSeq transcripts and proteins	Yes	Yes	Yes	Yes	Yes
Predicted transcripts and proteins	Yes	Yes	Yes	Yes	Yes
tRNA	No	Yes	No	Yes	Yes
<i>Ab initio</i> Gnomon models	No	No	Yes	Yes	Yes

Future development: annotation reports

The quality of the end-products produced by the Eukaryotic Genome Annotation Pipeline is highly dependent on the quality of the assembly and on the amount and quality of same-species or close cross-species evidence.

To facilitate the users' understanding of the annotation process and provide context for the annotation results, NCBI will start publishing reports for each annotation run by the end of 2013. These reports will include a description of the assemblies that were annotated and summary counts of the products of the annotation. Additionally, intermediate statistics summarizing which transcripts and protein sets were used and how well the evidence aligned to the genomes will be provided.

References

1. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Salign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct*. 2008 May 21;3:20. PubMed PMID: 18495041.
2. Pruitt KD, Tatusova T, Brown GR, Maglott DR. Nucleic Acids Res. 2012 Jan; 40(Database issue):D130–5. PubMed PMID: 22121212.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990 Oct 5;215(3):403–10. PubMed PMID: 2231712.
4. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Res*. 2001 May;11(5):803–816. PubMed PMID: 11337476.
5. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997 Apr 25;268(1):78–94. PubMed PMID: 9149143.
6. Wheelan SJ, Church DM, Ostell JM. Spidey: A Tool for mRNA-to-Genomic Alignments. *Genome Res*. 2001 Nov;11(11):1952–1957. PubMed PMID: 11691860.
7. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970 Mar;48(3): 443–53. PubMed PMID: 5420325.
8. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992 Nov 15;89(22):10915–9. PubMed PMID: 1438297.
9. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 2003 Oct 1;31(19):5654–66. PubMed PMID: 14500829.
10. Staden R. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*. 1984 Jan 11;12(1 Pt 2):505–19. PubMed PMID: 6364039.
11. Zhang MQ, Marr TG. A weight array method for splicing signal analysis. Computer applications in the biosciences. *Comput Appl Biosci*. 1993 Oct;9(5):499–509. PubMed PMID: 8293321.

12. Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 1984 Jan 25;12(2):857–72. PubMed PMID: 6694911.
13. Borodovsky M, McIninch J. GenMark: Parallel gene recognition for both DNA strands. *Computers & Chemistry.* 1993;17(2):123–33.
14. Morgulis A, Gertz EM, Schäffer AA, Agarwala R.. [WindowMasker: window-based masker for sequenced genomes](#). *Bioinformatics.* 2006 Jan 15;22(2):134-41
15. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2004. Available at: <http://www.repeatmasker.org>
16. Lowe TM and Eddy SR. Nucleic Acids Res. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. 1997 Mar 1;25(5):955-64.
17. Griffiths-Jones S. The microRNA Registry. *Nucleic Acids Res.* 2004 Jan 1;32(Database issue):D109–11. PubMed PMID: 14681370.

Prokaryotes

About Prokaryotic Genome Processing and Tools

Tatiana Tatusova, PhD,¹ Stacy Ciufo, PhD,¹ Boris Fedorov, PhD,¹ Kathleen O'Neill, PhD,¹ Igor Tolstoy,¹ and Leonid Zaslavsky, PhD¹

Created: January 23, 2014.

Scope

RefSeq Prokaryotic Genome Project

As of October 2013, the prokaryotic genome dataset contains more than 15,000 genomes from almost 4,500 species representing a wide range of organisms. They include many important human pathogens, but also organisms that are of interest for non-medical reasons, biodiversity, epidemiology, and ecology. There are obligate intracellular parasites, symbionts, free-living microbes, hyperthermophiles and psychrophiles, and aquatic and terrestrial microbes, all of which have provided a rich insight into evolution and microbial biology and ecology. There is almost a 20-fold range of genome sizes, spanning from ultra-small 45 kb archaeal genome of *Candidatus Parvarchaeum acidiphilum* recently obtained from mine drainage metagenome project (1) to the largest (14,7 Mb) strain of *Sorangium cellulosum*, an alkaline-adaptive epothilone producer (2).

The NCBI Reference Sequence (RefSeq) prokaryotic genome collection represents assembled genomes with different levels of quality and sampling density. Largely because of interest in human pathogens and advances in sequencing technologies (3), there are rapidly growing sets of very closely related genomes representing variations within the species. Some bacteria are often indistinguishable by means of current typing techniques. Whole-genome sequencing may provide improved resolution to define transmission pathways and characterize outbreaks. In order to support genome pathogen detection projects, RefSeq is changing the scope of the prokaryotic genome project to include all genomes submitted to public archives. Next generation technologies are changing the conventional use of microbial genome sequencing. Not so long ago genome sequencing projects were focused on a single bacterium, isolated and cultured from a single initial sample. Most of the sequencing technologies require DNA library preparation (DNA extraction and purification) followed by amplification and random shotgun sequencing. More recently new approaches have been developed that skip some of these steps. Metagenome sequencing shifted the focus from a single bacterium to multi-isolate and multi-species bacterial populations found in a single environmental sample. Individual organisms are not isolated and cultured but can be assembled computationally. RefSeq is taking a conservative approach of representing the genomic sequence of a single organism. Metagenomic assembly usually represents not a single organism but rather a

¹ NCBI.

composition of bacterial population. Metagenomic assemblies are not taken into RefSeq, however, that policy may change as the technologies and methods evolve. Single-cell sequencing technology (4) is another new technology that is being used to expand the catalog of uncultivated microorganisms. Genome assemblies that are generated from single-cell sequencing are taken into RefSeq when they meet basic validation criteria (see Quality Control).

RefSeq Prokaryotic Re-annotation Project

Historically, RefSeq prokaryotic genomes relied on annotation submitted to one of the archival sequence databases maintained by the International Nucleotide Sequence Database (INSD) Collaboration . RefSeq curation focused primarily on the correction of protein names using protein clusters (first COG (5), later PRK (6)). Some attempts to correct the start sites were made but were not comprehensive and were based on manual review that didn't scale when the number of genomes grew to many thousands. The problem of missing genes has not been addressed at all. The result was inconsistent annotation even in closely related genomes with a good reference genome such as *Escherichia coli*. (7). To address these problems, NCBI developed its own prokaryotic genome annotation and analysis pipeline (PGAAP) that has been successfully used for many genomes submitted to GenBank in the last 5 years. This pipeline produces more consistent and high quality automatic annotation that in many cases surpasses the original author-provided annotation.

More recently, we have re-designed the PGAAP pipeline using a more structured framework that enables faster processing of batches of bacterial genomes and integrates additional automated quality checks. All RefSeq genomes, newly or previously submitted, will be re-annotated using the updated pipeline to further improve consistency and quality in the RefSeq prokaryotic genomes dataset. This process will include both existing RefSeq genomes and new submissions of complete and draft (WGS) bacterial genomes that meet basic quality thresholds.

RefSeq Targeted Loci Project

The small subunit ribosomal RNAs (16S in prokaryotes and 18S in eukaryotes) are useful phylogenetic markers that have been used extensively for evolutionary analyses. The large subunit ribosomal RNAs (23S and 5S in prokaryotes and 28S in eukaryotes) have also been used for evolutionary analyses although to a lesser extent than the 16S or 18S. The 16S bacterial/archaeal ribosomal RNA project is the result of an international collaboration with the Ribosomal Database Project (8)), [GreenGenes](#), and [Silva](#) ribosomal RNA databases that curate and maintain sequence datasets for these markers. The fungal 18S and 28S ribosomal RNA projects are the result of an international collaboration with the [Fungal Tree of Life](#) project.

History

RefSeq Prokaryotic Genome Project History

The scale of genome sequencing and the production of data have reached astounding proportions since the completion of the first microbial genome of *Haemophilus influenzae* Rd KW20 (9) was released in 1995 and both the number of genomes and the number of unique genera for which a completely sequenced genome is available are increasing rapidly. We can see a shift in paradigm around 2010 when genome sequencing shifted from a single sample of a bacterial organism to hundreds of samples of bacterial populations.

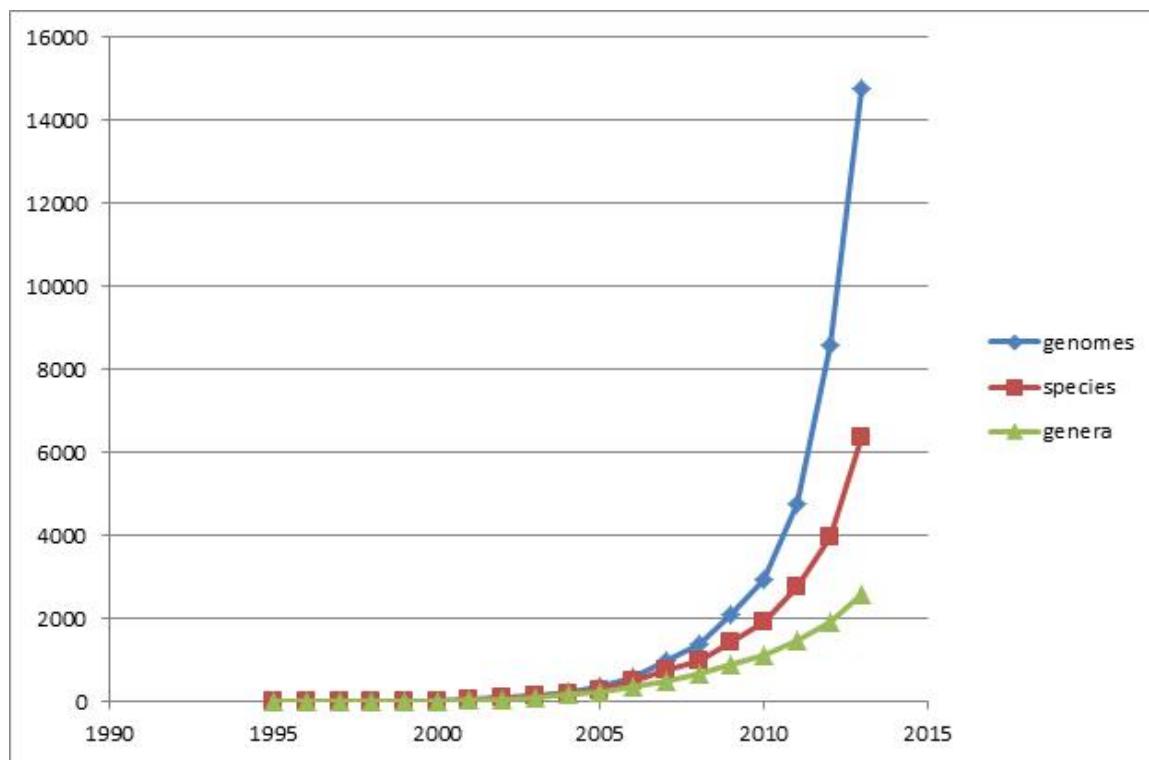


Figure 1. Growth of genomes, species, and genera: rapid growth of the number of isolates with relatively slow growth of new genera. Note that the data does not include assemblies from environmental studies where the number of novel species is growing much faster.

RefSeq Targeted Loci Project History

The Targeted Loci Project initiated in 2009 as a database of molecular markers used for phylogenetic analyses and identification of bacteria, archaea, and fungi. The initial project consisted of 16S ribosomal RNA from bacterial and archaeal type strains and has expanded to include 23S and 5S ribosomal RNA from bacterial and archaeal genomes as well as 18S and 28S ribosomal RNA from fungi.

Data Model

Genome and Assembly

RefSeq prokaryotic genomes are organized in several new categories based on curated attributes and assembly and annotation quality measures.

Reference genomes—manually selected “gold standard” complete genomes with high quality annotation and the highest level of experimental support for structural and functional annotation. Available at: <http://www.ncbi.nlm.nih.gov/genome/browse/reference/>

Representative genomes—representative genome for an organism (species); for some diverse species there can be more than one. For example, pathogenic and non-pathogenic *E. coli* will each be assigned a reference. Available at: www.ncbi.nlm.nih.gov/genome/browse/representative/

Variant genomes—all other genomes from individual samples representing genome variations within the species. Corresponds to Sequence Ontology- [SO:0001506].

How to define a genome?

BioProject ID can no longer define a genome for many multi-isolate and multi-species projects.

Taxonomy ID (taxid) can no longer define a genome since a unique taxid will not be assigned for individual strains and isolates. The collection of DNA sequences of an individual sample (isolate) will be represented by unique BioSample ID and if raw sequence reads are assembled and submitted to GenBank they will get a unique Assembly accession. The Assembly accession is specific for a particular genome submission and provides a unique ID for the set of sequence accessions representing the genome. Therefore, sequence data associated with a BioSample ID could be assembled with two different algorithms which may be submitted resulting in two sets of GenBank accessions, each with its own Assembly accession.

For example, BioProject [PRJNA203445](#) is a multi-species project with multiple strains and isolates of different food pathogens. Each isolate has its own BioSample ID and each assembled genome has its own Assembly accession. An isolate of *Listeria monocytogenes* strain R2-502 was registered as BioSample [SAMN02203126](#), and its genome is comprised of GenBank accessions [CP006595-CP006596](#), which are in the Assembly database as accession [GCA_000438585](#).

Autonomous Proteins

In order to manage the flood of identical proteins arising from annotation of variant genomes and decrease existing redundancy from bacterial genomes, NCBI is introducing a new protein data type in the RefSeq collection signified by a “WP” accession prefix. [WP](#)

accessions provide non-redundant identifiers for protein sequences. This new data type is provided through NCBI's genome annotation pipeline but is managed independently of the genome sequence data to ensure the dataset remains non-redundant.

We are doing this for two major reasons: 1) WP protein records represent a non-redundant protein collection that provides information about the protein sequence and name with linked information to genomic context and taxonomic sample; 2) use of WP accessions allows us to avoid creating millions of redundant protein records in the RefSeq collection. See more details at <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/announcements/WP-proteins-06.10.2013.pdf>

Targeted Loci

16S ribosomal RNA project: Archaea and Bacteria

Initially, the 16S ribosomal RNA project compared curated, near-full-length 16S sequences that corresponded to bacterial and archaeal type strains and from all contributing databases. RefSeq records corresponding to the original INSD submission were created from sequences and taxonomic assignments that were in agreement in all databases. Curation provided additional information, such as culture collection information or type strain designations and corrections to the sequence or taxonomy as compared to the original INSD submission. The 16S ribosomal RNA project has been expanded to include full length 16S ribosomal sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

23S ribosomal RNA project: Archaea and Bacteria

The 23S ribosomal RNA project includes full length 23S sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

5S ribosomal RNA project: Archaea and Bacteria

The 5S ribosomal RNA project includes full length 5S sequences from complete and incomplete genomes to provide representatives at the species level for the entire taxonomic range of bacteria and archaea.

18S and 28S ribosomal RNA projects

18S and 28S markers that correspond to type specimens and near full length sequences from all contributing databases were compared. RefSeq records corresponding to the original INSD submission were created from sequences and taxonomic assignments that were in agreement in all databases. The RefSeqs may contain corrections to the sequence or taxonomy as compared to the original INSD submission, and may have additional information added that is not found in the original.

Dataflow

The source of the genomic sequence in the RefSeq collection is a primary sequence record in the INSD public archives. Genomic sequences (nucleotide) in prokaryotic RefSeqs are identical copies of the underlying primary INSD records.

Quality Control

Genome representation: only assemblies with full representation of the genome of the organism are taken into RefSeq. Many genomes assemblies coming from single cell sequencing technology give only partial representation of DNA in a cell, ranging from 10% to 90%. Genome representation can be validated by comparative analysis if other genomes are available in closely related groups (species or genera). For novel phyla or kingdoms, some indirect criteria are applied (presence of universally conserved genes and total genome size)

Genomes and Genome Groups (Clades)

Historically, prokaryotic organisms were organized by classical taxonomic ranking system (species, genus, family, order, and phylum). Unlike eukaryotes, prokaryotes do not have clear definition of a species. Delineation of prokaryotic species was originally based on phenotypic information, pathogenicity, and environmental observations. More recently, several complementary approaches have been developed including molecular techniques such as DNA-DNA hybridization, phylogenetic markers (16S or universally conserved genes), and whole genome comparison (ANI – average nucleotide identity). Sequence-based methods using single-copy universally conserved genes are used for delineation of prokaryotic species. We have implemented a similar approach to define bacterial clades based on comparison of universally conserved ribosomal proteins (markers).

Table 1. 23 universally conserved markers, shown here by Cluster ID.

1	30S ribosomal protein S12
2	30S ribosomal protein S7
3	30S ribosomal protein S2
4	50S ribosomal protein L11
5	50S ribosomal protein L1
6	50S ribosomal protein L3
7	50S ribosomal protein L22
8	30S ribosomal protein S3
9	50S ribosomal protein L14
10	50S ribosomal protein L5
11	30S ribosomal protein S8

Table 1. continues on next page...

Table 1. continued from previous page.

12	50S ribosomal protein L6
13	30S ribosomal protein S5
14	30S ribosomal protein S13
15	30S ribosomal protein S11
16	50S ribosomal protein L13
17	30S ribosomal protein S9
18	30S ribosomal protein S15
19	30S ribosomal protein S17
20	50S ribosomal protein L16
21	50S ribosomal protein L15
22	50S ribosomal protein L18
23	30S ribosomal protein S4

The pipeline for calculating genome clades consists of three major components. The first step is collecting the input data from NCBI main sequence repositories. The genomic data are dynamic: hundreds of new genomes and assembly updates are submitted to NCBI each day. We create a snapshot of all live genome assemblies and their nucleotide sequence components (chromosomes, scaffolds, and contigs) and store them in an internal database with a date stamp. The genome dataset is organized into large groups, phyla and super-phyla as defined by NCBI Taxonomy, see ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/CLADES/Phyla.txt.

Assemblies are then filtered by quality and passed to the processing script. Ribosomal protein markers are predicted in every genome to overcome problems with the submitted genome annotations (missing and/or incorrect annotations) and to normalize the predicted markers data set. Marker predictions are performed by aligning reference protein markers against full genome assemblies. Assemblies with at least 17 markers are passed to the next step. Genome distance is calculated as an average of pairwise protein distances of markers shared in a pair of genomes. Finally, agglomerative hierarchical clustering trees are built within phylum-level groups. Clades at the species level are calculated using a species-aware algorithm. Sub-clade trees are provided by cutting out the trees at the distance of 0.25.

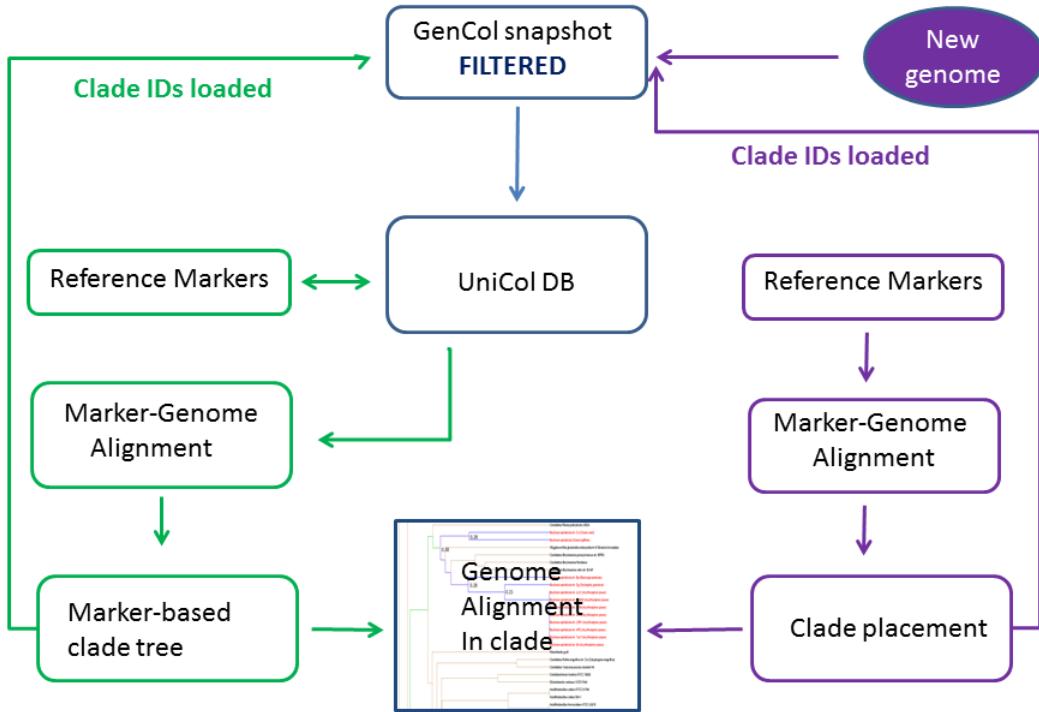


Figure 2. Calculating genome groups (clades) using universally conserved clusters. Snapshots are made every 6 months. The “FILTERED” step removes: partial assemblies, chimera, hybrid, mixed-cultured, metagenome assemblies.

Re-Annotation

The goal of the RefSeq re-annotation project is to improve the quality, normalize annotation, and reduce redundancy by creating reference sets of genomes, genes, and proteins. Information about the NCBI Prokaryotic Genome Annotation Pipeline is available [here](#).

Improved consistency in RefSeq annotation across prokaryote genomes will provide a common ground for experimental and computational analysis. However, automatic pipelines cannot replace the manual curation and experimental validation of unusual features and biological artifacts such as ribosomal slippage, pseudogenes, mobile elements, Insertion Elements (IS), and rare non-standard start codons. Connecting the experimental studies of individual genes or gene families to genome annotation—despite some attempts to make it automatic—continues to be a laborious manual process. We do not want to overwrite manual curation with automatic prediction so continue to emphasize an integrated approach of computation supplemented by curation. We also encourage the research community to submit experimental data and manually curated

data to RefSeq. There are two ways of making a contribution to the prokaryotic RefSeq collection:

1. **Organism/genome experts**—submit regular updates of your community-curated or experimentally validated genome annotation to GenBank and RefSeq. For example: *Escherichia coli* K-12, *Mycobacterium tuberculosis*, *Bacillus subtilis*, *Pseudomonas aeruginosa* PAO1, and *Salmonella enterica* LT2. These genomes are all in the the RefSeq “Reference genome” dataset
2. **Gene or gene family experts/experimental data providers**—partial annotation updates to RefSeq records. For example: proteomics, ColleCT, REBASE. The update can be implemented as an automatic pipeline, or experimental validation studies published in journals indexed in PubMed can submit an annotated citation (a GeneRif) through NCBI’s Gene resource.

There are ongoing efforts to establish relationships with the research community to provide accurate and up-to-date annotation for specific organisms or metabolic pathways. The “gold standard” Reference genome annotation is a result of the comparison of community annotation and NCBI automatic annotation reviewed by RefSeq curators.

Access

New genomes processed for RefSeq, are made public in NCBI resources and added to FTP directories daily.

Genome Groups

Reference Genomes: <http://www.ncbi.nlm.nih.gov/genome/browse/reference/>

Representative Genomes: www.ncbi.nlm.nih.gov/genome/browse/representative/

Complete list of prokaryotic genomes is available in Entrez Genome browser: <http://www.ncbi.nlm.nih.gov/genome/browse/>

Text version of the table can be downloaded from the FTP site: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

Species-level clades : ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/CLADES/

Reference set of universally conserved markers: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/MARKERS

Sequence Data

Complete genomes : <ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria>

Draft genome assemblies: ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT

Plasmids: <ftp://ftp.ncbi.nlm.nih.gov/genomes/Plasmids>—this directory contains a complete list of all plasmids that are submitted as part of whole genome or individual

complete plasmids that are sequenced and submitted separate from the chromosomes in plasmid targeted studies.

The genomes FTP area supports users who are interested in downloading data for one or a specific subset of organisms and/or in downloading the data that corresponds to an annotated genome. Users who are interested in comprehensive downloads can do so via the existing RefSeq release: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Genomes are automatically linked to many other databases and resources in Entrez. These include Bioproject, Biosample, Assembly, PubMed, Taxonomy, and many others. See the Genome chapter for details.

Related Tools and Resources

Entrez links allow navigation through different databases. Metadata and sequence data are stored separately but are easily linked.

Resources

Taxonomy

The Taxonomy database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet. The NCBI taxonomy group began assigning strain-level taxids for prokaryotes with complete genome sequences as a convenience for those at INSDC institutes and their users when that sequencing was a major achievement. That practice was extended to prokaryotes with draft genome sequences and to some eukaryotic microbial organisms, e.g., yeasts. With high-throughput sequencing, manual curation of strain-level taxids is no longer possible. Therefore, the practice will be discontinued in January 2014. However, the thousands of existing strain-level taxids will remain, and we will continue to add informal strain-specific names for genomes from specimens that have not been identified to the species level, e.g., “*Rhizobium sp.* CCGE 510” and “*Salpingoeca sp.* ATCC 50818”. The strain information will continue to be collected and displayed. Submitters of genome sequences will be required to register a BioSample ID for each organism that they are sequencing. Available at: <http://www.ncbi.nlm.nih.gov/taxonomy/>

BioSample

The BioSample database contains descriptions of biological source materials used in experimental assays. The BioSample record includes strain information and other metadata, such as culture collection and isolation information, as appropriate. The BioSample accession will be included as a “DBLINK” on GenBank records, and the GenBank records themselves will continue to display the strain in the source information. Available at: <http://www.ncbi.nlm.nih.gov/biosample/>

BioProject

A BioProject record is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project. Available at: <http://www.ncbi.nlm.nih.gov/bioproject/>

Assembly

Each genome assembly is loaded to the Assembly database and assigned an Assembly accession. The Assembly accession is specific for a particular genome submission. Genome assemblies are hierarchical. The shortest assembly components are contigs. Contigs are assembled into longer scaffolds, and scaffolds are assembled into chromosomes if there is sufficient mapping information. The Assembly resource provides the information on genome assembly structure and statistics. Available at: <http://www.ncbi.nlm.nih.gov/assembly/>

Protein Clusters

This collection of related protein sequences (clusters) consists of proteins derived from the annotations of whole genomes, organelles and plasmids. It currently limited to Archaea, Bacteria, Plants, Fungi, Protozoans, and Viruses. Available at: <http://www.ncbi.nlm.nih.gov/proteinclusters/>

Microbial Genome Resources:

The Microbial Genome resource page provides a central hub for many of NCBI's tools and resources. These include the gene prediction tools GeneMark (10) and Glimmer (11), and a statement of availability for the NCBI Genome Annotation Pipeline, now available as part of the Genbank submission process.

An expandable menu lists all of NCBI's prokaryotic genome tools and resources, guides for genome submission, information about NCBI's Annotation Workshop, and dynamically updated statistics on the growth of microbial genome data.

The page also contains an expandable taxonomic tree with all bacterial and archaeal genomes with submitted genome data, and includes both complete and draft genomes. Available at: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html

Tools

Microbial Genomes BLAST has new database options including "Representative genomes," now the default database, and "All genomes." Representative genomes provide a smaller, less redundant set of records for a given bacterial species. These representatives are selected by the research community and NCBI computational processes and are especially helpful for microbial species that are highly represented by genomes for numerous strains in NCBI databases, such as *Escherichia coli*. The "All genomes" option offers the choice of Complete genomes, Draft genomes, or Complete plasmids. You can search these sets individually or in any combination. The microbial BLAST report also has

a new “Genome” link to the species page in Entrez Genome in the alignments section of the BLAST report.

Concise BLAST includes a representative protein from each cluster. This allows a more comprehensive taxonomic BLAST search while eliminating much of the noise from similar species: <http://www.ncbi.nlm.nih.gov/genomes/static/conciseblasthelp.html>

The Submission Check Tool is available for users to check the validity of their genome data prior to submission to Genbank. Checks include gene overlaps, RNA overlaps, partial overlaps, frameshifts, truncated proteins, missing RNAs, and RNA strand mismatches.

gMap is a graphical representation of pre-computed genomic comparisons of closely related strains. Syntenic blocks are detected through analysis of BLAST hits between every pair of the input sequences. Hits are split or combined to keep the number and lengths of syntenic blocks in accordance with the length of selected genomic intervals, as well as to ensure consistency of the blocks across multiple sequences. The results are displayed in a simple graphic that shows color-coded and numbered segments indicating similarity between two or more genomes. This tool can be used to visually detect chromosomal similarities, rearrangements, and the above-mentioned genomic islands, as well as smaller insertions or deletions. Care must be taken when interpreting results from incomplete genomes as hit coverage may be affected by the number of contigs and the gaps between them. Other tools that are useful for examining pairwise genomic comparisons are GenePlot and HitPlot which are linked on this page.

ProtMap is a graphical gene neighborhood tool that displays clickable, linked genes upstream and downstream of the target. This resource is useful in identifying paralogs.

GenePlot combines protein-sequence similarity searches with sequence location, unlike gMap, which is solely based on nucleotide sequence similarity. This tool can be used to detect syntenic regions or chromosomal rearrangements in closely related species, as well as contiguous regions in distantly related organisms. Small insertions or deletions and major genomic islands in closely related species are also identifiable using this tool and a table of the best hits between both organisms is available. Geneplot provides a more detailed view of the pairwise comparison of two genomes.

TaxPlot compares two reference proteomes to a query proteome and thus provides a three-way comparison of proteome similarities. A single protein can be searched for and highlighted, and entire COG functional categories can be examined, allowing a three-way comparison of a single category of proteins. This tool can be used to detect potentially horizontally transferred genes by using a distantly related organism in comparison to two closely related strains.

tRNAscan-SE (12) is a program for detection of tRNA genes in genomic sequence.

References

1. Fujishima K, Sugahara J, Miller CS, Baker B, Di Giulio M, Tomita M, Banfield JF, Kanai A. A novel three-unit tRNA splicing endonuclease found in ultra-small

- Archaea possesses broad substrate specificity. *Nucleic Acids Res.* 2011 Dec;39(22):9695–704. PubMed PMID: 21880595.
- 2. Han K, Li ZF, Peng R, Zhu LP, Zhou T, Wang LG, Li SG, Zhang XB, Hu W, Wu ZH, Qin N, Li YZ. Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieus. *Sci Rep.* 2013;3:2101. PubMed PMID: 23812535.
 - 3. Loman NJ, Constantinidou C, Chan JZ, Halachev M, Sergeant M, Penn CW, Robinson ER, Pallen MJ. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nat Rev Microbiol.* 2012 Sep;10(9):599–606. PubMed PMID: 22864262.
 - 4. Blainey PC. The future is now: single-cell genomics of bacteria and Archaea. *FEMS Microbiol Rev.* 2013 May;37(3):407–27. PubMed PMID: 23298390.
 - 5. Koonin EV. The Clusters of Orthologous Groups (COGS) Database: Phylogenetic Classification of Proteins from Complete Genomes. The NCBI Handbook (Internet). 2002.
 - 6. O'Neill K, Klimke W, Tatusova T. Protein Clusters: A collection of proteins grouped by sequence similarity and function. NCBI Help Manual. 2007.
 - 7. Poptsova MS, Gogarten JP. Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiol.* 2010;156:190917. PubMed PMID: 20430813.
 - 8. Maidak BL, Olsen GJ, Larsen N, Overbeek R, MacCaughey MJ, Woese CR. The Ribosomal Database Project (RDP). *Nucleic Acids Res.* 1995;24:82–85. PubMed PMID: 8594608.
 - 9. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* 1995 Jul 28;269(5223):496–512. PubMed PMID: 7542800.
 - 10. Borodovsky M, McIninch J. GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry.* 1993;17(2):123–133.
 - 11. Salzberg S, Delcher A, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research.* 1998;26(2):544–548. PubMed PMID: 9421513.
 - 12. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* (1997) Mar 1;25(5):955–64.

Prokaryotic Genome Annotation Pipeline

Tatiana Tatusova, PhD,¹ Mike DiCuccio, MD,¹ Azat Badretdin, PhD,¹ Vyacheslav Chetvernin,¹ Stacy Ciufo, PhD,¹ and Wenjun Li, PhD¹

Created: December 10, 2013.

Scope

The process of annotating prokaryotic genomes includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons, and other mobile elements. Bacterial and archaeal genomes have the considerable advantage of usually lacking introns, which substantially facilitates the process of gene boundary identification. A protein coding gene in a prokaryotic genome can be defined as a single interval open reading frame (ORF)—a region starting with a valid start codon and ending with a stop codon bounding a region of in-frame translation covering three nucleotides per codon. In the absence of introns, it might seem that ORFs can be designated as any substring of DNA that begins with a start codon and ends with a stop codon. However, applying this straightforward and simple rule to any bacterial or archaeal genome will result in many overlapping and short ORFs. Determining which one of the overlapping ORFs represents a true gene is a particularly difficult task. In addition, designating the cutoff for filtering short ORFs that might encode small polypeptides presents a special challenge.

Additional complications arise from the fact that Bacteria and Archaea often use alternative start codons—codons other than the traditional ATG. Gene prediction tools must distinguish between six potential candidate start codons (ATG, GTG, TTG, and sometimes ATT, CTG and ATC). Several approaches have been developed for accurate prediction of translation initiation site in prokaryotes (1, 2). Stop codons can also have a dual function, as stop codons TGA or TAG may encode selenocysteine and pyrrolysine. With the rapid and continuous growth of prokaryotic genome sequencing, automated annotation techniques will remain the main approach in the future. Given advances in population studies and analysis of outbreaks, automated annotation processes will shift toward comparative analysis and away from individual genome annotation. On the other hand, diversity studies generate genome sequences from extreme environments and deep taxonomic lineages. These genomes may encode novel genes with no similarity to those available in public archive databases, and must be handled differently. NCBI has developed an automated pipeline that takes advantage of both statistical and similarity-based methods, using similarity when sufficient quantities of comparative data are available and relying more on statistical predictions in the absence of supporting material.

¹ NCBI.

NCBI provides a prokaryotic genome annotation service to GenBank submitters using this pipeline, which is also used to annotate RefSeq prokaryotic genomes.

History

Gene prediction or gene finding is one of the fundamental challenges in computational biology.

The main goal of gene prediction is to identify the regions of DNA that are biologically functional. The history of gene prediction dates to the works of Fickett, Gribskov, and Staden (3-5) that started in the early 1980s. The first generation of gene prediction algorithms used a local Bayesian approach analyzing one ORF at a time.

Generation of the first complete bacterial genome sequence of *Haemophilus influenzae* in 1995 heralded a new era in genome sciences. The second-generation of gene prediction algorithms analyzed the global properties of the genomic sequence of a given organism and gave rise to several successful programs such as GeneMark (6) and Glimmer (7). These programs employ an inhomogeneous Markov model for short DNA segments (i.e., k-tuples), from which an estimate of the likelihood for the segment belonging to a protein coding sequence can be derived after training against existing validated gene data.

Similarity searches gave rise to another broad category of gene prediction methods (8). Experimentally derived or known protein sequences were used to determine putative placements and base gene models on those placements, using programs such as BLASTx and FASTA. The third generation of automated annotation programs combined execution of multiple gene-calling programs with similarity-based methods. These third generation approaches attempt to balance evidence-based gene model selection with computationally derived predictions. In 2013, NCBI released the current incarnation of its third-generation pipeline, based on an infrastructure that permits efficient parallel computation and high-throughput annotation.

The first version of NCBI's Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) using second-generation gene prediction approaches was developed in 2001-2002. The approach combined hidden Markov model (HMM)-based gene prediction algorithms with protein homology methods. Gene predictions were done using a combination of GeneMark (6) and Glimmer (7). Conserved proteins from curated clusters, Clusters of Orthologous Groups (9) and NCBI Prokaryotic Clusters (10), were used to search for genes that may have been missed by pure *ab initio* annotations. Ribosomal RNAs were predicted by sequence similarity searching using BLAST against an RNA sequence database and/or using Infernal and Rfam models. Transfer RNAs were predicted using tRNAscan-SE (11). This standard operating procedure was previously published (12).

With recent advances in genome sequencing technology, the paradigm has shifted from individual genomes to population studies represented by a so-called pan-genome. Newly submitted genomes can be annotated using data already available for closely related

genomes. In order to incorporate information from closely related isolates, NCBI's annotation pipeline was extensively redesigned. Our new approach is based on the assumption that proteins conserved in a genome clade (core proteins) should be found in a new genome of the same clade. The major difference compared to the previous pipeline is that the alignment-base information is calculated upfront and passed to a customized version of GeneMarkS (13), termed GeneMarkS+, which can incorporate external data in the analysis of the statistical evidence of coding potential and transcriptional start site.

Annotation Standards

Certain metrics can be used to assess the quality of the annotation of the prokaryotic genomes. NCBI has established a relationship with other major archive databases and major sequencing centers in an effort to develop standards for prokaryotic genome annotation. This collaboration has resulted in a set of annotation standards approved and accepted by all major annotation pipelines (14). Many groups still use a simplified set of rules for annotation, and as such may miss critical annotations. Some simplifications include eliminating alternative starts and applying hard-coded length cutoffs for acceptance of short proteins. In addition to these standards, many groups also apply "soft" validation checks.

Minimum standards for annotating complete genomes

1. ANNOTATION SHOULD FOLLOW INSDC SUBMISSION GUIDELINES
(GenBank/ENA/DDBJ)
 - a. Prior to genome submission a submitted Bioproject record with a registered locus_tag prefix is required according to accepted guidelines
<http://www.ncbi.nlm.nih.gov/genomes/locustag/Proposal.pdf>
 - b. The genome submission should be valid according to feature table documentation
http://insdc.org/documents/feature_table.html
2. MINIMAL GENOME ANNOTATION SHOULD HAVE
 - a. At least one copy of rRNAs (5S, 16S, 23S) of appropriate length and corresponding genes with locus_tags
 - b. At least one copy of tRNAs for each amino acid and corresponding genes with locus_tags
 - c. Protein-coding genes with locus_tags (see below) and corresponding CDS
3. VALIDATION CHECKS AND ANNOTATION MEASURES
Validation checks should be done prior to the submission. NCBI has already provided numerous tools to validate and ensure correctness of annotation. Additional checks will be put in place to ensure the minimal standards are met. Statistical measures that are used for annotation quality assessment include:
 - a. Feature counts by feature type
 - b. Protein coding gene count vs genome size ratio
 - c. Percent of short (<30 aa) proteins
 - d. Percent of coding regions with a standard start codon

- e. Count of protein coding regions with “hypothetical protein” product
- 4. EXCEPTIONS

Exceptions (unusual annotations, annotations not within expected ranges) should be documented and strong supporting (experimental) evidence should be provided.

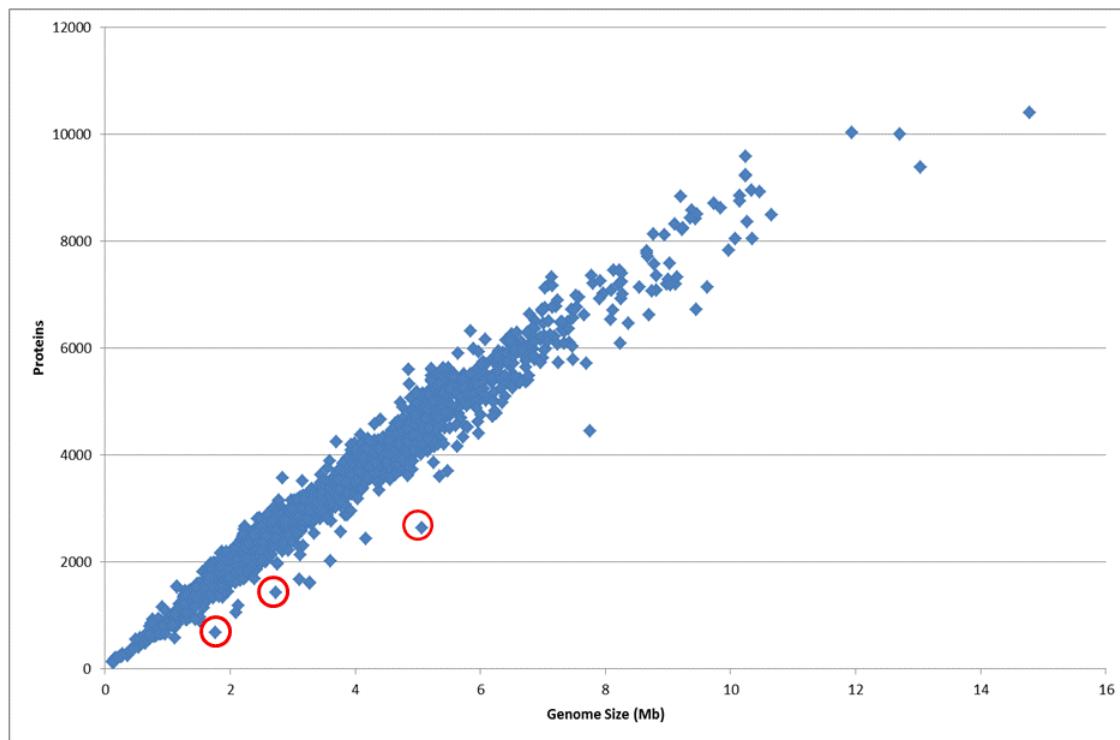


Figure 1. The ratio of genome size to the number of protein coding genes for all genomes. The number of protein coding genes is directly proportional to the genome size; on average the density is one gene per 1000 kb. Small obligate parasites and symbionts that undergo rapid gene reduction have less protein coding genes than average. Some examples are indicated in red: *Serratia symbiotica* str. 'Cinara cedri', *Synergistetes bacterium* SGPI, and *Yersinia pestis* CO92. Methods

Non-coding RNA (structural RNA, small ncRNA, tRNA)

Structural ribosomal RNAs in prokaryotes (5S, 16S, and 23S) are highly conserved in closely related species. The NCBI Refseq collection contains a curated set of rRNA reference sequences for each of these three types of rRNA. The pipeline uses a nucleotide (BLASTn) search against the reference set. We further pass 5S rRNA hits through cmsearch for refinement against known structural motifs (15, 16). Partial alignments that fall below 50% of the average length are dropped. Prediction of small ncRNAs involves a two-step process similar to the identification of 5S rRNAs: first, we use a BLASTn search

against sequences of selected Rfam families; second, we use cmsearch with default parameters to produce the final annotation.

TRNAscan-SE

The NCBI annotation pipeline uses tRNAscan-SE to identify tRNA placements. The tRNAscan-SE program identifies 99–100% of transfer RNA genes in DNA sequence with less than one false positive per 15 gigabases and is currently one of the most powerful and widely used tRNA identification tools. To identify tRNA genes, the input genome sequence is split into 200 nucleotide (nt) windows with overlap of 100 nt and run through tRNAscan-SE program (11). We automatically provide separate parameterization for Archaea and Bacteria. All tRNA calls with a score below 20 are discarded.

Protein Alignments—ProSplign

The current incarnation of NCBI’s automated annotation pipeline uses data derived from prokaryotic population studies. Our approach uses a pan-genome approach to identify the core set of proteins that we expect to find in all genomes belonging to a given group. We define several groups of proteins that can be used: a “target set,” comprising proteins that are expected to be found in all members of a group, such as universally conserved ribosomal proteins, clade-specific core proteins, and curated bacteriophage protein clusters; and a separate “search set,” comprising all automatic clusters, including curated and non-curated protein clusters, curated bacteriophage protein clusters, and all bacterial UniProtKB/Swiss-Prot proteins.

Proteins from the target set are aligned to genomic sequence using ProSplign, an application developed at NCBI for handling partial-frame and spliced protein alignments. ProSplign offers the advantage of being frameshift-aware and can align proteins correctly in the face of genome sequence errors.

Complete gapless alignments with 100% identity to a target protein are accepted for final annotation. Frameshifted alignments and partial alignments of good quality are passed to GeneMarkS+ for further refinement.

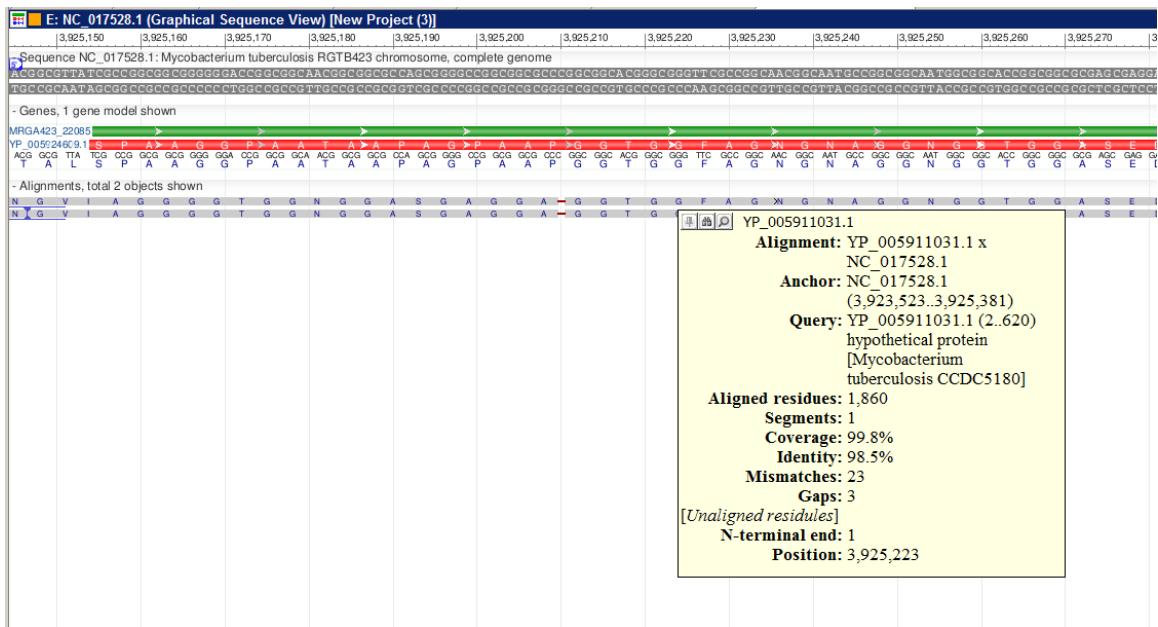


Figure 2. A fragment of ProSplign alignment against an annotated peptide. The similarity is too low for BLAST to find a significant hit, but ProSplign was able to locate the corresponding protein and identify a frameshift as well.

Frameshift detection

Detecting frameshifted genes is a critical component of resolving ambiguities in automated annotation and provides important feedback in assessing the quality of an assembly. A shift of a reading frame in a coding region is caused by indels (insertions or deletions) of a number of nucleotides in a genomic sequence that is not divisible by three. These events may represent artifacts resulting from technical errors, or they may have biological causes. Sequencing errors are common with the Next Generation Sequencing (NGS) technologies leading to the potential for a high rate of frameshifted genes in assemblies generated using NGS techniques. In addition, gene inactivation during evolution allows selective mutation across ancient ORFs, representing a true biological event that can be marked as a pseudogene with a disrupted ORF. Further, programmed frameshift mutations that are tolerated during translation are known to play an important role in the evolution of novel gene function.

Two-pass improvement

The introduction of a two-pass method improved the original gene-calling procedure. An initial gene call is made by alignment or *ab initio* prediction, followed by extraction of the protein and BLAST comparison to a set of known conserved proteins. We scan these BLAST hits to identify candidates that are partial or incomplete matches to single or adjacent models. Candidate proteins are realigned to an expanded region using ProSplign. The candidate frameshift alignments are then combined with the original evidence and fed back to GeneMarkS+ in a two-step iterative process. This method is similar to a one-

pass annotation with an extended protein reference set but is both more accurate and efficient.

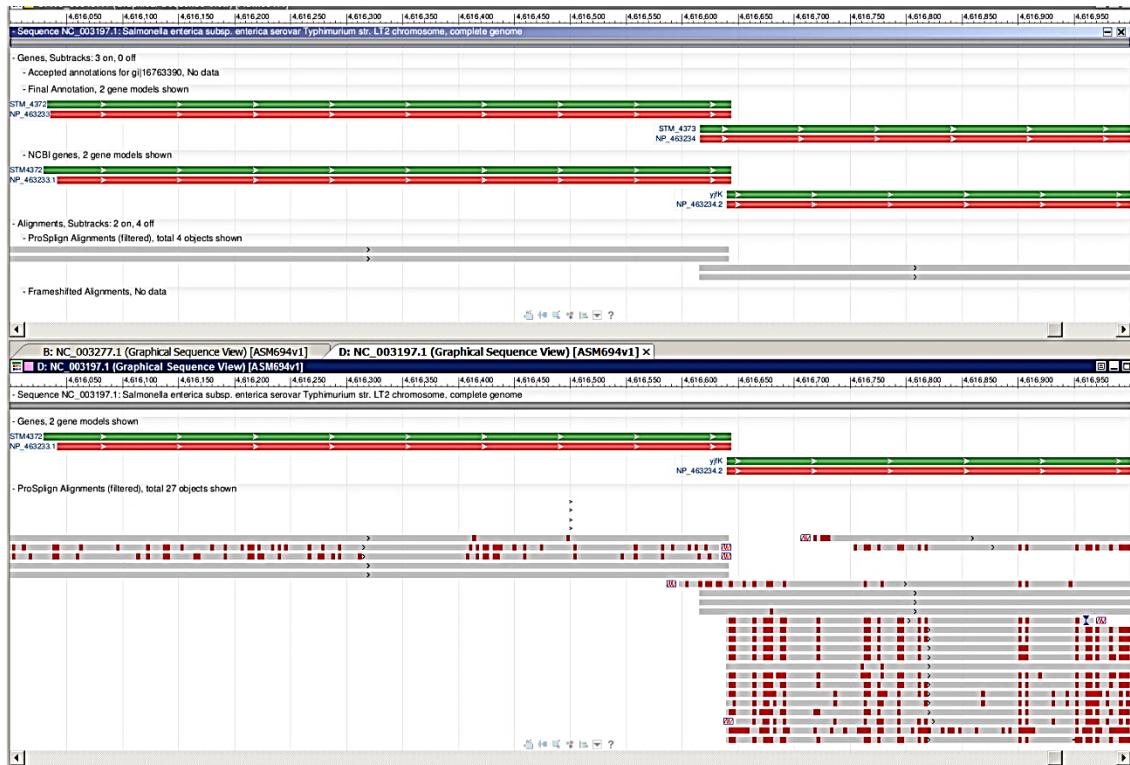


Figure 3. Two-pass protein alignment process produces improved gene model. The first pass, shown in the upper panel, does not get enough alignments; the second iteration, seen in the bottom panel, produced expanded evidence supporting a more conserved start.

GeneMarkS+

In collaboration with NCBI, The GeneMark team has developed GeneMarkS+, a special version of GeneMarkS that can integrate information about protein alignments and non-coding RNA features. NCBI's pipeline first collects evidence based on placement of known proteins and structural elements as described above. These placements are then passed to GeneMarkS+, which combines information about statistical ribosomal binding sites and likelihood estimations for the start of transcription with provided evidence about high-quality placements to determine a final set of predictions.

Mobile or fast evolving genes (phage, CRISPR)

The annotation of phage related proteins is based on homology to a reference set of curated phage proteins. The bacteriophage protein reference data set comes from an independent effort to calculate and curate protein clusters from all complete bacteriophage genomes.

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are a family of DNA direct repeats of 20 to 40 nucleotides separated by unique sequences of similar length and are commonly found in prokaryotic genomes.

The CRISPR database (17) allows users to search and identify repeats of interest. These defense systems are encoded by operons that have an extraordinarily diverse architecture and a high rate of evolution for both the cas genes and the unique spacer content. For classification and nomenclature of CRISPR-associated genes see (18). For CRISPR prediction the pipeline uses a wrapper around CRISPR Recognition Tool (CRT) (19) and PILER-CR (20).

Protein naming

The final component of the pipeline is identifying protein function and naming the protein product of the coding region. Assignment of a predicted model to a cluster for purposes of naming is based on protein homology to members of the cluster: we require high coverage, high-scoring alignments to at least three members of the same cluster in order to assign a protein to a cluster.

Dataflow

NCBI's Prokaryotic Genome Annotation Pipeline combines a computational gene prediction algorithm with a similarity-based gene detection approach. The pipeline currently predicts protein-coding genes, structural RNAs (5S, 16S, 23S), tRNAs, and small non-coding RNAs. The flowchart below describes the major components of the pipeline.

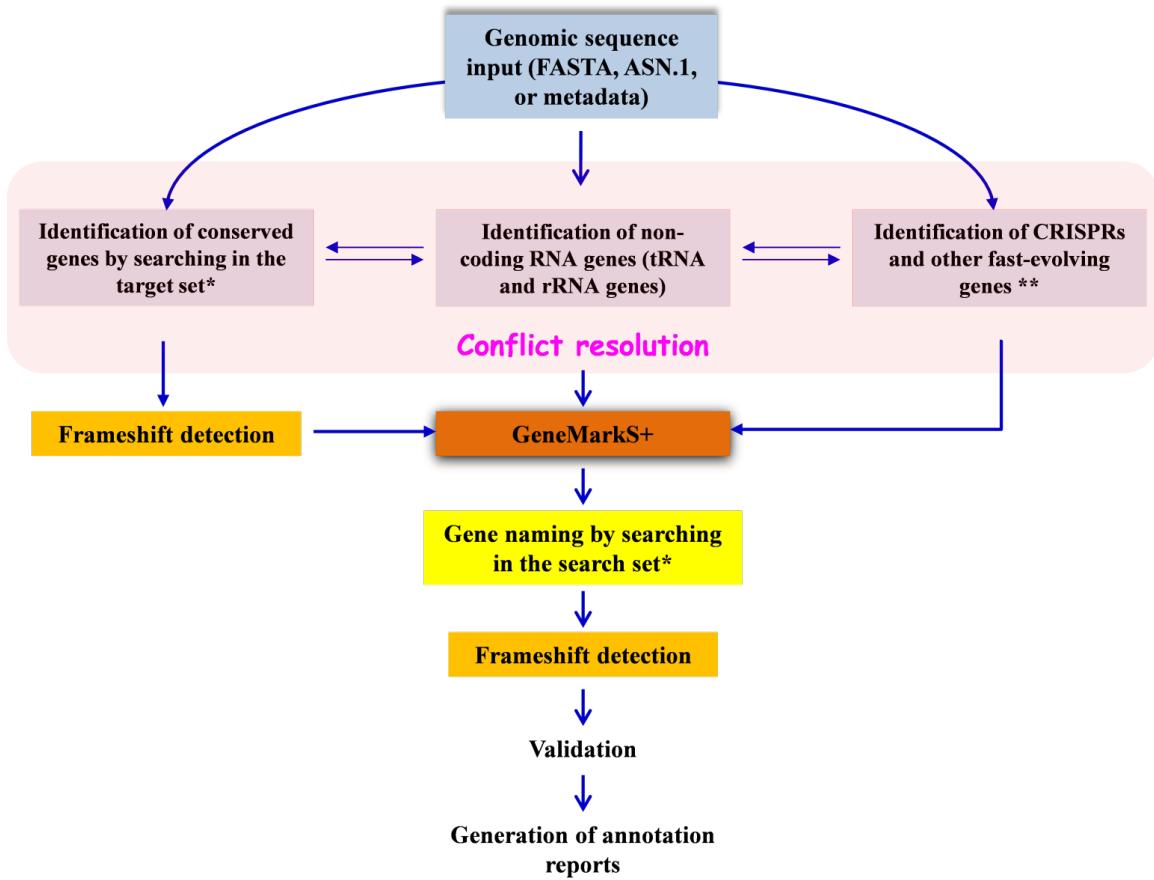


Figure 4. NCBI Prokaryotic Genome Annotation Pipeline diagram. * Target and search sets are described in the Protein Alignments section. ** Described in the protein naming section.

GenBank Submission Service

The NCBI prokaryotic annotation pipeline is a genome annotation service that is intended to help GenBank submitters with prokaryotic genome annotation. The pipeline can be used with complete genomes as well as whole genome sequences (WGS) consisting of multiple contigs. NCBI's submission standards require that genomic sequences deposited in GenBank meet a minimum level of quality, including passing contamination screening to eliminate foreign sequence elements and having all sequence contigs be at least 200 bases in length. The annotation pipeline is integrated into the submission system for those who choose this option: sequence data must pass initial validation within GenBank to ensure proper formatting and the presence of required information needed for annotation (organism information, genetic code, and locus-tag prefix). More details on the requirements for submission are available here: http://www.ncbi.nlm.nih.gov/genome/annotation_prok/

RefSeq Genome Annotations

In addition, the prokaryotic genome annotation pipeline is used to annotate NCBI reference sequence (RefSeq) genomes, with the exception of a small number that are manually curated by collaborating groups (for example, *Escherichia coli K12* which is provided by [EcoCyc](#)). The RefSeq collection provides a comprehensive, integrated, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. RefSeq sequences form the foundation of medical, functional, and diversity studies. They provide a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.

Autonomous Protein Records

In order to manage the flood of identical proteins and decrease representational redundancy, particularly from bacterial genomes, NCBI has introduced a new protein data type in the RefSeq collection signified by a ‘WP’ accession prefix. WP accessions provide non-redundant identifiers for protein sequences. This new data type is provided through NCBI’s prokaryotic genome annotation pipeline and is managed independently of the genome sequence data to ensure that the dataset remains non-redundant. There are two main reasons for this paradigm shift:

1. Autonomous WP protein records represent a non-redundant protein collection that provides independent information about the protein sequence and name with linked information to genomic context and taxonomic sample.
2. A WP accession may be annotated on numerous genomes (when the genome encoded proteins are identical) thereby providing a mechanism to avoid creating millions of redundant protein records in the RefSeq collection.

When the NCBI genome annotation pipeline annotates a bacterial protein that is 100% identical and the same length as an existing WP accessioned protein, NCBI is no longer creating a new protein record, with one exception (noted below). NCBI is instead annotating such proteins on the genome by referencing the existing WP accession in the annotated coding sequence (CDS) feature, indicating that the genome represents an exact example of that known protein sequence. Any annotation of protein function on the genome record (such as the product name and functional characteristics) reflects the independent WP record. WP records, therefore, always represent one exact sequence that may be observed one or many times in different strains or species. Also, WP records will always have a version of “1” and the sequence not be updated like taxon-specific RefSeq records.

Data Access

Genomes annotated by NCBI’s annotation pipeline include a relevant comment on the nucleotide record, and each feature specifies which gene prediction method was used. Within nucleotide records, users will find a generated comment and a structured

comment block indicating the version of the annotation software used and the date on which a given genome was annotated:

```

LOCUS      CP005492          1692823 bp    DNA     circular BCT 30-JUL-2013
DEFINITION Helicobacter pylori UM037, complete genome.
...
COMMENT    Annotation was added by the NCBI Prokaryotic Genome Annotation
           Pipeline (released 2013). Information about the Pipeline can be
           found here: http://www.ncbi.nlm.nih.gov/genome/annotation\_prok/

##Genome-Annotation-Data-START##
Annotation Provider      :: NCBI
Annotation Date         :: 07/22/2013 11:47:17
Annotation Pipeline     :: NCBI Prokaryotic Genome Annotation Pipeline
Annotation Method       :: Best-placed reference protein set;GeneMarkS+
Annotation Software revision :: 2.1 (rev. 406590)
Features Annotated      :: Gene; CDS; rRNA; tRNA; repeat_region
Genes                  :: 1,692
CDS                    :: 1,615
Pseudo Genes           :: 35
rRNAs                  :: 6 ( 5S, 16S, 23S )
tRNAs                  :: 36
Frameshifted Genes     :: 31
##Genome-Annotation-Data- END##

```

Within the Flat File report, users will find CDS regions marked up with information about evidence used and methods applied to generate such annotations:

```

CDS 17236..18585
      /locus_tag="K750_07885"
      /inference="EXISTENCE: similar to AA
sequence:RefSeq:YP_005789361.1"
      /note="Derived by automated computational analysis using
gene prediction method: Protein Homology."
      /codon_start=1
      /transl_table=11
      /product="phospho-2-dehydro-3-deoxyheptonate aldolase"
/protein_id="AGL70499.1"
      /db_xref="GI:499061585"

CDS complement(808991..809746)
      /locus_tag="K750_04005"
      /inference="COORDINATES: ab initio prediction:GeneMarkS+"
      /note="Derived by automated computational analysis using
gene prediction method: GeneMarkS+."
      /codon_start=1
      /transl_table=11
      /product="restriction endonuclease R.HpyAXII"
/protein_id="AGL69752.1"
      /db_xref="GI:499060838"

```

Re-annotation Consortium

Historically, RefSeq prokaryotic genomes relied on author-supplied annotation. Curation focused primarily on the correction of protein names using protein clusters. Attempts to correct start sites based on manual review were not comprehensive. Because of the rapid increase in the number of genomes (many thousands), manual review became impractical. Moreover, the issue of genome submissions with unannotated genes (missing genes) was not resolved. As a result, the original RefSeq prokaryotic annotation dataset contained inconsistent annotation even in closely related genomes that had high-quality annotated references, such as *E. coli*. NCBI's updated annotation pipeline can produce a consistent, high quality, automatic annotation that in many cases surpasses the original author-provided annotation. Therefore, NCBI is re-annotating prokaryotic RefSeq genomes to improve the overall consistency and quality of this dataset.

Related Tools and Resources

Protein Clusters—A collection of related protein sequences (clusters) consists of proteins derived from the annotations of whole genomes, organelles, and plasmids. It is currently limited to Archaea, Bacteria, Plants, Fungi, Protozoans, and Viruses. Protein clusters can be searched and viewed at <http://www.ncbi.nlm.nih.gov/proteinclusters/>

ProSplign—This tool produces accurate spliced alignments and locates alignments of distantly related proteins with low similarity and is an integral component of the NCBI's Genome Annotation Pipeline (Gnomon). The integration of ProSplign with the genome annotation pipeline significantly improved the quality of genome annotation over existing available methods. ProPSalign is available as an online tool at <http://www.ncbi.nlm.nih.gov/utils/static/prosplign/prosplign.html>

Frameshift detection—This tool is available as a standalone tool or Web application. Adjacent genes on the same strand are analyzed for hits against the same subject (common BLAST hit) by comparing BLAST results. Since gene fusions and splits occur in prokaryotic genes, the BLAST hits are analyzed for any subject (not the common BLAST hit) that covers 90% of the query protein, in which case the frameshift is not reported under the assumption that this gene is “real.” Any pair of genes failing to meet these criteria is reported as potential frameshifted genes and should be manually inspected. The Web version is available at: <http://www.ncbi.nlm.nih.gov/genomes/frameshifts/frameshifts.cgi>

NCBI is hosting two *ab initio* prokaryotic gene prediction programs: GeneMark and Glimmer.

The programs can be used for a rapid draft annotation of prokaryotic genomes.

GeneMark—The GeneMark family of gene finding programs has been used for prokaryotic genome annotation since 1995 when GeneMark contributed to launching the genomic era by providing automatic gene annotation of complete genomes of

Haemophilus influenza, *Methanoccus jannaschii*, as well as *Escherichia coli* and *Bacillus subtilis*. <http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi>

Glimmer—GLIMMER (20) is a system for finding genes in microbial DNA, especially the genomes of bacteria and archaea. GLIMMER (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models to identify coding regions. Glimmer version 3.02b is the current version of the system. http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi

References

1. Yada T, Totoki Y, Takagi T, Nakai K. A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res.* 2001 Jun 30;8(3):97–106. PubMed PMID: 11475327.
2. Hu GQ, Zheng X, Zhu HQ, She ZS. Prediction of translation initiation site for microbial genomes with TriTISA. *Bioinformatics*. 2009 Jul 15;25(14):184–5. PubMed PMID: 19015130.
3. Staden R, McLachlan AD. Codon preference and its use in identifying protein coding regions in long DNA sequences. *Nucleic Acids Res.* 1982 Jan 11;10(1):141–56. PubMed PMID: 7063399.
4. Gribskov M, Devereux J, Burgess RR. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* 1984 Jan 11;12(1 Pt 2):539–49. PubMed PMID: 6694906.
5. Fickett JW. Finding genes by computer: the state of the art. *Trends Genet.* 1996 Aug; 12(8):316–20. PubMed PMID: 8783942.
6. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998 Feb 15;26(4):1107–15. PubMed PMID: 9461475.
7. Salzberg SL, Delcher AL, Kasif S, White O. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 1998 Jan 15;26(2):544–8. PubMed PMID: 9421513.
8. Guigó R, Burset M, Agarwal P, Abril JF, Smith RF, Fickett JW. Sequence similarity based gene prediction. In: Suhai S, editor. *Genomics and proteomics: Functional and computational aspects*. New York, NY: Kluwer Academic / Plenum Publishing; 2000. pp. 95–105.
9. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 2001 Jan 1;29(1):22–8. PubMed PMID: 11125040.
10. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S, Schafer S, Tolstoy I, Tatusova T. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D216–23. PubMed PMID: 18940865.
11. Lowe T.M., Eddy S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.* 1997;25:955–964. PubMed PMID: 9023104.

12. Angiuoli S, et al. Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS. 2008; 2008;12:137–41. PubMed PMID: 18416670.
13. Besemer J., Lomsadze A., Borodovsky M. 2001; GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res.26(No. 4)pp1107–1115. PubMed PMID: 11410670.
14. Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, Mizrahi I, Pruitt KD, Tatusova T. Solving the Problem: Genome Annotation Standards before the Data Deluge. Stand Genomic Sci. 2011 Oct 15;5(1):168–93. PubMed PMID: 22180819.
15. Eddy S.R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics. 2002;3:18. PubMed PMID: 12095421.
16. Nawrocki EP, Eddy SR. Query-dependent banding (QDB) for faster RNA similarity searches. PLoS Comput Biol. 2007;3(3) PubMed PMID: 17397253.
17. Grissa I., Vergnaud G., Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007;8:172. PubMed PMID: 17521438.
18. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, Moineau S, Mojica FJ, Wolf YI, Yakunin AF, van der Oost J, Koonin EV. Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol. 2011 Jun;9(6):467–7. PubMed PMID: 21552286.
19. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007 Jun 18;8:209. PubMed PMID: 17577412.
20. Biswas A., Gagnon J.N., Brouns S.J., Fineran P.C., Brown C.M. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. RNA Biol. 2013;10(5):817–827. PubMed PMID: 23492433.
21. Delcher A.L., Harmon D., Kasif S., White O., Salzberg S.L. Improved microbial gene identification with GLIMMER. Nucleic Acids Research. 1999;27(23):4636–4641. PubMed PMID: 10556321.

Viruses

About Viral and Phage Genome Processing and Tools

Yiming Bao, PhD,^{✉1} J. Rodney Brister, PhD,¹ Olga Blinkova, PhD,¹ Danso Ako-adjei, PhD,¹ and Chetvernin Vyacheslav, PhD¹

Created: March 30, 2013; Updated: May 10, 2013.

Scope

The National Center for Biotechnology Information (NCBI) Viral Genome Resource hosts all virus-related data and tools. All complete viral genome sequences deposited in the International Nucleotide Sequence Database Collaboration (INSDC) databases are collected by the NCBI Viral Genome Project (1). A RefSeq record is created from one of the complete genome sequences for each virus species, and the others are tagged as neighbors to the RefSeq. RefSeq records are subjected to curation procedures, which include automated gene locus_tag assignment, validation of molecule information and protein names, and annotation of novel proteins. Proteins encoded by RefSeqs are used to generate Protein Clusters, and the curated Protein Clusters are applied in turn to improve the annotation of new and existing RefSeqs. Sequence analyses such as global alignment of genome neighbors to RefSeq are provided. Databases specific to viruses that have a large number of genome sequences generated from sequencing projects are created for easy data retrieval and analyses. Tools that facilitate viral genome annotation and classification are also available.

History

Like other organisms, the number of sequences for viruses increased dramatically in recent years thanks to the advances of sequencing technologies. There are over 500,000 sequences in the INSDC databases for human immunodeficiency virus 1 alone. This makes it very difficult for researchers to efficiently work with these sequences directly from the databases. Also, many sequences are very short, which users may not want to include in their searches or analyses. So a collection of complete viral genome sequences is desired. Viruses are unique compared to other organisms in that there is significant variability in the forms of their genomes—linear or circular, single-stranded or double-stranded, DNA or RNA. The genome organization and expression strategy can vary dramatically from virus to virus. The taxonomic classification standards for some viruses are not very well established. All these factors contribute to sequence submission errors,

¹ NCBI; Email: bao@ncbi.nlm.nih.gov; Email: jamesbr@ncbi.nlm.nih.gov; Email: blinkova@ncbi.nlm.nih; Email: akoadjei@ncbi.nlm.nih; Email: chetvern@ncbi.nlm.nih.

[✉] Corresponding author.

such as wrong molecular information, incorrect/missing gene/protein annotation, and chaos of taxonomic assignment in some viral genome sequence records. As part of the NCBI's Reference Sequence (RefSeq) database, the NCBI Viral Genome Project was created to cope with the issues described above.

Data Model

Viral Genome Reference Sequence and Genome Neighbors

From all complete viral genome sequences (including Viroids), one RefSeq record (or a set of RefSeq records for segmented viruses) is created for each species. Occasionally, more than one RefSeq record is created in a species to represent different subgroups of the virus (e.g., Dengue virus 1, 2, 3, and 4). All other complete viral genome sequences in the same species as the RefSeq become "neighbors" to the RefSeq. Both the RefSeq and neighbors can be retrieved from the NCBI's Entrez database. Please note that genome neighbors are not the same as GenBank related sequences, which represent records selected by sequence similarity.

Virus Taxonomy

The Viral Genomes Project is tightly linked with the Taxonomy database. The names and classifications of viruses in the Taxonomy database follow, to a large extent, the most recent report of the International Committee on the Taxonomy of Viruses (ICTV, <http://www.ictvonline.org>). As the ICTV reports appear infrequently, the NCBI Taxonomy database attempts to stay current by also accepting new names and classification schemes on a case-by-case basis as provided in the reports of the ICTV executive meetings, taxonomic proposals approved by ICTV, and based on the advice of outside experts.

However, many sequence submissions are for viruses that are not listed in the ICTV report and sometimes not even described in the published literature. In spite of this, the taxonomy database can index these organisms and associated records, and these names are placed under an "unclassified" node to distinguish them from the ICTV-approved names.

Resources for Specific Viruses

Databases for specific viruses (e.g., Influenza, Dengue, and West Nile) that have a large number of genome sequences generated from sequencing projects are created for easy data retrieval and analyses. See the Virus Variation chapter for more information.

Databases such as the Coronavirus Resource are also made for viruses of medical importance.

Dataflow

Scan INSDC Databases for Complete Viral Genome Candidates

For the NCBI viral genome collection, a complete genome is one that contains all coding regions of the virus. Newly released viral sequences in the INSDC databases are constantly screened for complete genomes by an automated procedure. A sequence is considered a candidate for a complete genome if either of the following two criteria is met: (i) The topology of the sequence is circular or (ii) the definition of the sequence contains any of the following phrases: “complete genome”, “complete chromosome”, “sequence of the genome of”, or “complete genomic sequence”.

Some complete viral genome sequences are not detected by this automatic procedure because the source record either does not correctly indicate a circular topology or does not include the keywords listed above. To overcome this problem, viral sequences undergo an additional screening based on the sequence length. Only sequences longer than 85% of the length of the reference sequence in the species are selected as the complete genome candidates.

Additionally, complete viral genome sequences are identified with the aid of external scientific advisors, experts on particular families or groups of viruses, who also assist in the curatorial process. The list of advisors and their contact information is available at <http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239&hopt=advisors>.

Accept Complete Viral Genome Candidates

NCBI staff manually review the complete viral genome candidates, and if satisfied, accept them as the reference sequence if there was not one in the species, or otherwise as neighbors to existing reference sequences.

When more than one complete genome is available for reference sequence, preference is given to the sequence of a well-studied and practically important virus isolate, and/or the one that has the best annotation.

The taxonomic classifications of the viruses where the new genomes are obtained are rigorously checked at this step. It is not uncommon that the GenBank submitter gives a sequence a new species name when it really belongs to an existing species. If the sequence is accepted without verification, an undesired reference sequence record will be created when it should actually be considered as a neighbor to another reference sequence.

Finding an appropriate taxonomic position for a virus usually involves comparative sequence analysis, using tools like BLAST and PASC (described below). The ICTV Study Groups are frequently consulted for taxonomy issues. When in doubt, a complete genome candidate will have a “wait” status until the issues are resolved, in which case, there is a delay in the creation of the reference sequence.

Segmented Viruses

Segmented viruses are those with more than one genome component (segment). Candidates for complete sequences of individual segments are determined using similar criteria as the ones described above for single-component viruses. One sequence is selected for each segment to form a set of reference sequences that covers all segments of

the genome. The reference genome set is manually assembled by matching strain and isolate information for available sequences of complete components. When several sequences are available for the same segment of the same strain and/or isolate, preference is given to a sequence obtained in the same laboratory as those of the other components. Other complete sequences become neighbors to the reference sequence of the same segment, and a segment name provided by NCBI staff is used to connect neighbors to the corresponding reference sequence.

RefSeq Creation

A RefSeq record is created from the INSDC sequence accepted by NCBI staff. Accession numbers unique to RefSeq records are assigned to the nucleotide (NC_XXXXXX) and protein (NP_XXXXXX, YP_XXXXXX or YP_XXXXXXXXXX) sequences. Gene [locus_tags](#) are assigned to the RefSeq as well.

RefSeq Curation

The curatorial process includes the correction and update of the record, along with the addition of relevant biological information taken from the literature, other sequence records, original submitters, and outside advisors. The most common corrections are made to the type and topology of the genomic molecule (double strand or single strand, linear or circular) as well as to taxonomy lineage.

A large part of the curatorial process involves improvement of genome annotation, which includes searches for missing genes, assignment of functional roles to protein products, correction of annotations for proteins expressed by frame shifting or read through, restoration of proteins disrupted by sequencing errors, and addition of post translational processing information. Some RNA viruses encode polyproteins that contain multiple functional domains and are cleaved by proteinases into mature peptides. NCBI staff adds mature peptide annotation (from polyproteins) to viral RefSeq records if they were not present in the original INSDC records (compare [NC_002532](#) and [X53459](#)). These mature peptides have RefSeq protein accession numbers and are thus indexed and retrievable as individual proteins.

RefSeq has established a number of collaborations in an effort to improve the accuracy of viral sequence records. In collaboration with Mark Borodovsky, the GeneMark program (<http://exon.gatech.edu/VIOLIN>) was used to predict open reading frames (ORFs) in some viral RefSeq genomes and to compare them with the original annotations. For example, the original GenBank record for the complete genomic sequence of a large double-stranded DNA virus—Sheppox virus ([AY077832](#)) contained no protein annotations. Subsequently, 147 protein coding genes were predicted by the GeneMark program in the genome, and added to the corresponding RefSeq record ([NC_004002](#)).

Another ongoing collaboration project is the revision and annotation of overlapping genes. Gene overlaps, which can be defined as having nucleotides coding for more than one protein by being read in multiple reading frames, are a common feature of viruses (2). Proteins created by gene overlaps are typically accessory proteins that play a role in viral

pathogenicity or spread (3, 4). Despite their importance, overlapping genes are difficult to identify and are often overlooked. Carefully annotated and curated data on overlapping genes in viral genome RefSeqs allow researchers to conduct studies on evolution and informational characteristics of overlapping genes as well as on functionality of corresponding products. With the help of Andrew Firth from the University of Cambridge, Cambridge, UK, and David Karlin from the University of Oxford, Oxford, UK, we have been working on adding (or correcting) missing overlapping genes and corresponding proteins in virus RefSeqs. At the present time, at least one RefSeq representative for each genus from the 14 selected virus families (*Arteriviridae*, *Arteriviridae*, *Bunyaviridae*, *Caliciviridae*, *Circoviridae*, *Disistroviridae*, *Flavoviridae*, *Luteoviridae*, *Paramixoviridae*, *Parvoviridae*, *Picornaviridae*, *Potyviridae*, *Reoviridae*, *Togaviridae*) was corrected based on experimental or predictive analysis. For each new protein the position of start and end codon is determined based on the experimental data or according to comparative analysis described in the literature. Protein names, their functions (if known), experimental data and literature links are added for each protein. The frameshifting sites (if present) and the nature of a frameshift are added to the genome annotations based on the most recent literature data. RefSeq NC_001479 is an example of a sequence with recently discovered gene overlaps. It represents the encephalomyocarditis virus (EMCV) species from the family *Picornaviridae*. According to the experimental analysis performed by Loughran et al. (5) a conserved ORF overlaps the 2B-encoding sequence of EMCV in the +2 reading frame. A previously overlooked ORF is translated as a 128-129 amino acid transframe fusion (2B*) with the N-terminal 11-12 amino acids of 2B, via ribosomal frameshifting. To represent the results of this study, we added the truncated version of polyprotein (CDS positions: 834-3998, 3998-4351) and 2B* protein (mature peptide with the coding region positions: 3966-3998, 3998-4348). Another example of overlapping genes is RefSeq NC_008311 belonging to the Murine norovirus (MNV) species from the family *Caliciviridae*. We updated the annotation to add a recently discovered (6) virulence factor 1 protein (VF1) encoded by subgenomic RNA (CDS coordinates: 5069-5710) in an alternative reading frame overlapping the VP1 coding region.

RefSeq proteins are clustered on the basis of sequence homology within the [Protein Clusters](#) resource and curated in aggregate by NCBI staff (also see the Protein Clusters chapter). This curation includes the assignment of functional protein names to clusters that can in turn be propagated to individual protein records in RefSeq, yielding consistent, informative names among clustered proteins. RefSeq staff work in collaboration with a number of stakeholders including SwissProt, ICTV, sequencing centers, and scientific communities to develop annotation and protein naming standards. The goal is to improve the quality and consistency of viral genome annotation in both RefSeq and INSDC databases.

HIV-1, Human Protein Interaction Database

Although numerous advances have been made in the fields of retrovirology and AIDS research, much of the biological processes that underlie infection, replication, and

immune evasion are unknown. Similarly, the mechanisms that orchestrate cellular restriction to infection as well as those that potentiate the innate and adaptive immune systems are poorly understood. The human immunodeficiency virus type 1 (HIV-1) RNA genome ([NC_001802](#)) encodes three major genes from which the major proteins—group specific antigen (Gag), polymerase (Pol), and envelope (Env)—are transcribed. Through various combinations of overlapping reading frames, differential splicing, and proteolytic cleavage, several HIV-1 proteins with regulatory and auxiliary roles are also expressed. These include the proteins transactivator (Tat), regulator of viral protein expression (Rev), negative factor (Nef), viral protein R (Vpr), viral infectivity factor (Vif), and viral protein U (Vpu). Tat and Rev regulate transcription and HIV-1 nuclear RNA export, respectively, while the accessory proteins Nef, Vpr, Vif, and Vpu are dispensable for replication in certain cell types (7).

A large number of protein-protein interactions involving viral and cellular proteins are required for both cellular immunity and competent viral infection. Information about protein-protein interactions is critical to advancements in vaccine research, therapeutic drug discovery, and cell biology. In order to facilitate these advancements, the HIV-1, Human Protein Interaction Database, was established to catalog all data published in peer-reviewed journals regarding HIV-1 and human protein interactions. Included in this database are brief descriptions of the respective interactions, National Library of Medicine (NLM) PubMed identification numbers (PMIDs) for articles describing the interaction, NCBI Reference Sequence (RefSeq) protein accession numbers, NCBI Entrez Gene ID numbers, and keywords that facilitate interactions searches.

The database is organized in a fashion that provides downloadable or onsite-viewable reports for the HIV-1 proteins. The protein interactions are categorized by 43 interaction keywords including binds, cleaved by, degrades, stimulates, co-localizes with, and recruits. By utilizing these keywords in drop-down menus, a user can narrow down a search for particular interaction types for specific HIV-1 proteins. For instance, the viral protein, Vif, binds the mammalian cellular protein apolipoprotein B mRNA editing enzyme (APOBEC3G) and targets it for proteasomal destruction. In the absence of Vif, APOBEC3G incorporation into HIV-1 virions leads to G-to-A hypermutation of the viral genome and markedly reduced replicative potential (8). By clicking on “vif” and then using the drop-box to choose “degrades” and hitting the “view” button, a report containing APOBEC3G and several similar interactions can be obtained. These reports can then be downloaded as a text file if the user chooses. Because the HIV-1 interaction data is integrated into the Entrez Gene database, information about protein domain structure, genomic context, synonymous names, gene loci, and links to order clones of the human genes may be also obtained. The availability of resources such as this can provide insights into the many biological processes that involve HIV-1 infection, replication, and evolution. Likewise, they provide data that may one day permit predictive modeling and/or construction of structural interaction networks (9).

This database is made available through the National Library of Medicine at <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions> (10) and a set of all the current

interactions can be obtained by FTP from <ftp://ftp.ncbi.nih.gov/gene/GeneRIF> under the file, hiv_interactions.gz.

Access

Viral Genome Resource Homepage

All NCBI virus-related data can be accessed through the Viral Genome Resource homepage (<http://www.ncbi.nlm.nih.gov/genomes/GenomesHome.cgi?taxid=10239>). On the left hand side bar, users can find links to information about the viral genome resource, list of viral genomes in alphabetical order or by taxonomic groups, viral RefSeq genome and protein records, sequence records of genome neighbors, ftp site for viral RefSeq data, virus-related tools, and resources for specific viruses. The main page lists large groups of viruses, which can be opened and moved down to lower level taxonomic lineages that contain RefSeqs. A search box is also provided for quick location of viral genomes belonging to a particular group (e.g., family).

The Viral Genome Presentation

The top level Viral Genome Presentation lists all viruses that have RefSeqs and is available at <http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239>. A list of the next level sub-lineages (e.g., ssRNA viruses) is provided so users can easily move down the taxonomic hierarchy to find viral genomes of their interest. An example viral genome presentation page for Potyviridae ([taxid=39729](#)) is shown in Figure 1.

Global Alignment of Genome Neighbors

Pairwise global alignment is performed between each genome neighbor and their corresponding RefSeq, using the "band" version of the Needleman-Wunsch algorithm. A graphical representation of the alignment is available when the number under the "Nbrs" column on viral genome presentation pages (Figure 1C) is clicked. For segmented viruses, the alignments for each genome segment are sequentially displayed. An example of the graphical view is shown in Figure 2.

The Genome Browser

Viral RefSeqs are also presented with all genome records of other organisms in the NCBI Genome Browser (<http://www.ncbi.nlm.nih.gov/genome/browse>), which can be filtered by large groups (dsDNA viruses, ssRNA viruses, etc.), subgroups (mostly viral families), and hosts.

Viral Genome Data Through ftp

As part of the bi-monthly RefSeq releases, all nucleotide and protein sequences and GenBank flat files for viral genomes are available at <ftp://ftp.ncbi.nih.gov/refseq/release/viral>. Various format of genome data for individual viruses are available at <ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/>.

The screenshot shows the NCBI Entrez Genome interface for the family *Potyviridae*. The main content area displays a table of viral genomes, with several rows highlighted by orange circles labeled A through E.

- A:** A list of sub-lineages one level below the *Potyviridae* node. The list includes: Brambyvirus [1], Bymovirus [4], Ipomovirus [5], Macravirus [1], Poacevirus [3], Potyvirus [86], Rymovirus [3], and Tritimovirus [4].
- B:** Organism names of viral genomes. The table lists various viruses under their respective genera, such as Brambyvirus, Bymovirus, Ipomovirus, Macravirus, Poacevirus, and Potyvirus. Some entries are in blue (ICTV approved) and some are in copper (unassigned/unclassified).
- C:** The number of genome neighbors to the RefSeq. The numbers are linked to the global alignment of genome neighbors with the RefSeq (Figure 2).
- D:** Links to show all RefSeq nucleotide or protein records for the viruses displayed.
- E:** Links to download a table with the information displayed or a list of accession numbers of the RefSeq records.

Figure 1. Viral genome collections in the family *Potyviridae*. A. A list of sub-lineages one level below the *Potyviridae* node in NCBI's Taxonomy database. The numbers in square brackets represent the number of genomes within the taxonomic node. B. Organism names of viral genomes. The ICTV approved names are in blue, and the unassigned/unclassified ones are in copper. C. The number of genome neighbors to the RefSeq. The numbers are linked to the global alignment of genome neighbors with the RefSeq (Figure 2). D. Links to show all RefSeq nucleotide or protein records for the viruses displayed. E. Links to download a table with the information displayed or a list of accession numbers of the RefSeq records.

Related Tools

FLAN

The influenza virus genome annotation tool (FLAN, for FLu ANnotation) was created as a result of NCBI's participation in the Influenza Genome Sequencing Project (11) which was initiated by the National Institute of Allergy and Infectious Diseases. Under this project, influenza virus samples provided by collaborators worldwide are sequenced by the J Craig Venter Institute, submitted to NCBI for genome annotation, and released immediately in GenBank. Since the beginning of this project in 2005, more than 11,000

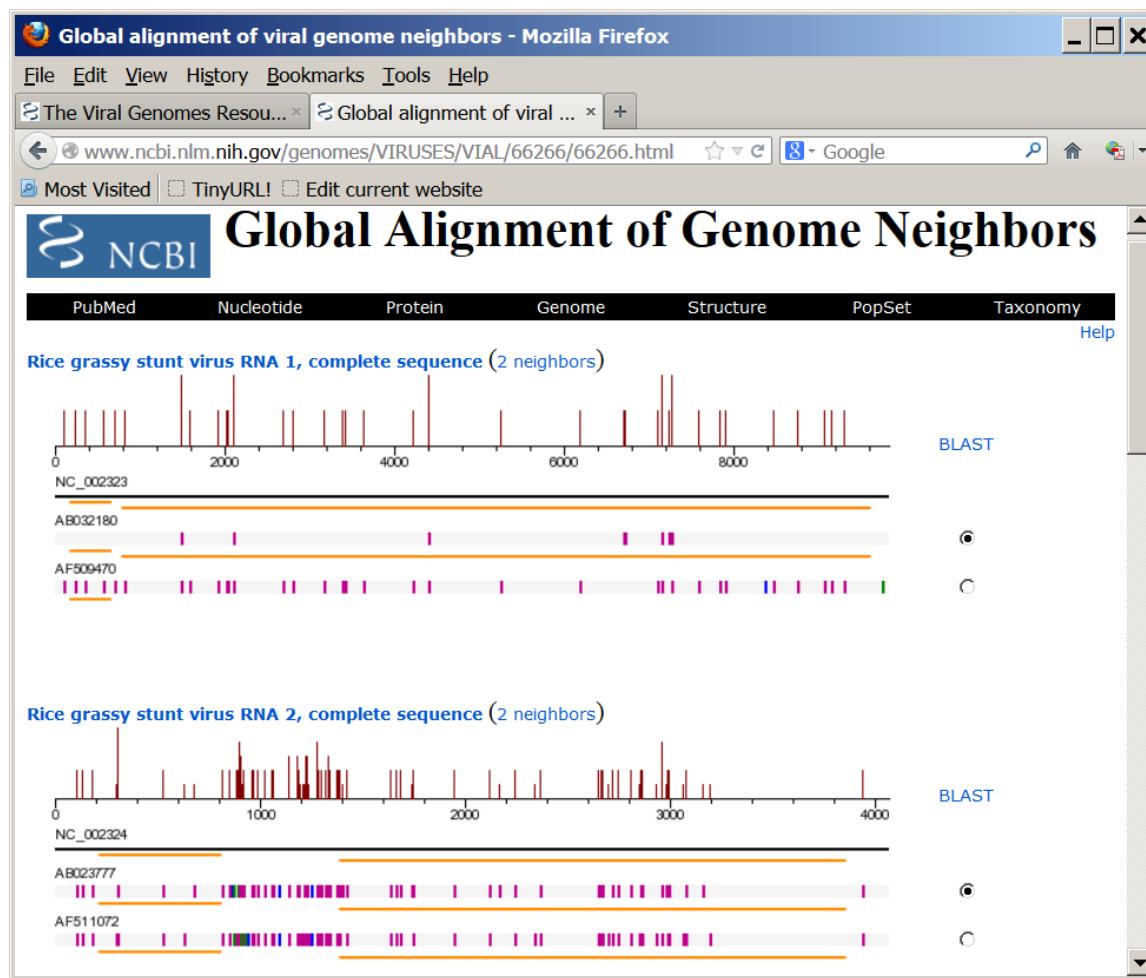


Figure 2. Global alignment of genome neighbors with the RefSeq of Rice grassy stunt virus ([NC_002323](#) and [NC_002324](#)). The magenta, blue, and green bars represent differences, deletions, or insertions in sequences, compared with the reference sequence. The orange lines represent coding regions annotated in the genome records. The histogram shows the average density of nucleotide changes (excluding gaps, insertions, and undetermined nucleotides) in all genome neighbors for each reference sequence segment.

genomes of influenza viruses have been sequenced and published in GenBank. Because of the large number of sequences, an automated genome annotation pipeline is required.

FLAN is an application for user-provided Influenza A virus, Influenza B virus, and Influenza C virus sequences. It can predict protein sequences encoded by a flu sequence and produce a feature table that can be used for sequence submission to GenBank, as well as a GenBank flat file.

The type/segment/subtype of an input influenza sequence is first determined by BLAST, and then aligned against a corresponding reference protein set with a "Protein to nucleotide alignment tool"—ProSplign (<http://www.ncbi.nlm.nih.gov/sutils/static/>

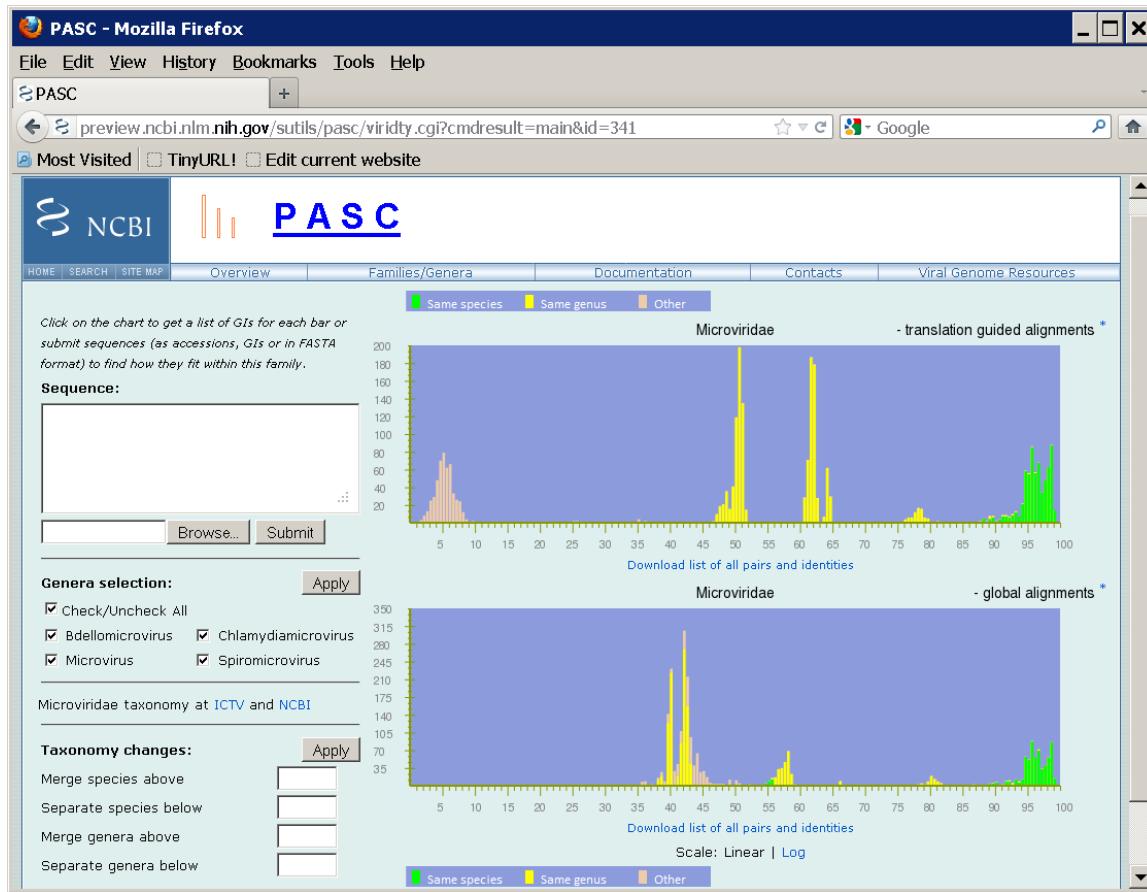


Figure 3. Frequency distribution of pairwise identities from the complete genome sequence comparison of 71 microviruses.

prosalign/prosalign.html). The translated product from the best alignment to the sample protein sequence is used as the predicted protein encoded by the input sequence.

In addition to the creation of the feature table, FLAN can also determine and report the following properties of flu sequences: influenza virus species (A, B, or C), length, genome segment, subtype of the HA and NA segment, common drug-resistant mutations in segment NA and M, mutations in the PB2 segment that might confer high virulence of influenza viruses, possible contaminated/vector sequences at the ends, the completeness of the nucleotide, protein and coding regions, insertion/deletion that will disrupt the coding regions, and premature stop codons in the coding regions. These capabilities of FLAN are used to populate certain fields in the Influenza Virus Sequence Database (12). They also make FLAN a useful tool for flu sequence validation to identify possible sequencing errors or human errors in segment/subtype assignment.

Internally, FLAN is implemented in a NCBI-developed framework which allows the execution of background CGI tasks for more than 30 s (default WEB frontend timeout). This allows the online interface of FLAN to process hundreds of flu sequences at a time.

In an effort to maintain consistent and high quality annotations of flu sequences, FLAN is recommended by GenBank as the tool to generate feature tables that can be used for flu sequence submissions to GenBank through the recently implemented “virus wizard” in Sequin.

FLAN (13) is available at <http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/annotation.cgi>.

PASC

Viruses are classified based on their properties such as morphology, serology, host range, genome organization, and sequence. The dramatic increase of virus sequences in public databases makes sequence-based virus classification more feasible.

The most commonly used virus classification tool based on sequence is phylogenetic analysis. The classification of about 70% of families and floating genera described in the ninth Report of ICTV are supported by phylogenetic trees (14). Despite its wide-spread use, phylogenetic analysis is usually computationally intensive, and requires expertise to interpret the results.

Recently, a novel method using natural vector based on the distributions of nucleotide sequences was reported on virus classifications (15).

Another sequence-based molecular classification method for viruses is to calculate pairwise identities of virus sequences within a virus family, and the number of virus pairs at each percentage is plotted. This will usually produce peaks that represent different taxonomic groups such as variants, species, and genera, and the percentages at borderlines of the peaks can be used as demarcation criteria for different taxa. This method has been applied to a few viral families including *Coronaviridae* (16), *Geminiviridae* (17), *Papillomoviridae* (18), *Picornaviridae* (19), and *Potyviridae* (20). A major drawback of this method is the inconsistency of the results when different protocols are used to calculate the pairwise identities. It is very difficult, if not impossible, for researchers to use the exact algorithm and parameters to test their own sequences as those used to establish the demarcation criteria (usually) by others. The identities obtained from the two systems are therefore not comparable. To overcome this problem, NCBI created the PASC (Pairwise Sequence Comparison) resource (21), where the same protocol is used for both the establishment of the demarcation criteria and the testing of new viral sequences. Many viral groups were included in the resource.

For a given virus family/group, complete genome sequences are retrieved from the NCBI viral genomes collection described in this chapter, which include both RefSeqs and neighbors. These sequences, together with their lineages in the NCBI Taxonomy database, are stored in a database. The database is updated every day to add new genome sequences and reflect taxonomy changes.

Traditionally, genome identities were calculated based on pairwise global alignments in PASC. Although this method works well for some virus families/groups such as

papillomaviruses and potyviruses, the results are not optimized for others mainly for the following reasons:

1. In some viruses with circular genomes such as the circoviruses, there is an inconsistency in the designation of the first nucleotide of the genome sequences in public databases.
2. In some viruses, particularly those with negative-strand RNA genomes, the opposite strand of the genome are sometimes submitted to the public databases. When genome identities are calculated based on the global alignment of two genomes in the opposite strand, the result is usually lower than what they should be.
3. For viruses that are distantly related, the identities obtained by global alignment are usually misleading, because the minimum identity of any two random genome sequences of the same size is 25%.

To overcome these issues, a BLAST-based alignment method is used. Two sets of BLAST (22) are performed on each pair of genome sequences. In the first set, the translated protein sequences of one genome in six frames are searched against the nucleotide sequence of the other genome using tblastn. The amino acid alignments in the tblastn results are converted back to nucleotide alignments. In the second BLAST set, pairwise blastn is carried out on the nucleotide sequences of the genomes. We then select a consistent set of hits from the two sets of BLAST results, preferring higher identity hits and trimming overlaps out of lower identity hits. This process will select blastn hits for close genomes, but most likely tblastn hits for distant ones. A mixture of blastn and tblastn hits might be used in some cases. Pairwise identities are calculated as the total number of identical bases in local hits divided by the mean sequence length of the genome pair. This method greatly improves the performance of PASC in some virus families (see Figure 3 for an example).

The identity distribution chart is plotted based on pairwise alignments computed between every member of the selected virus family or group. The pair is represented in green color if both genomes belong to the same species according to their assignment in NCBI's Taxonomy database; in yellow color if the two genomes belong to different species but the same genus; and in peach color if they belong to different genera. Both linear and log scales are available for the Y-axis (number of pairs).

To compare external genomes against existing ones in the database, specify the query genomes in the "Sequence" box, using either their GenBank Accession/GI numbers, entering the raw sequence in FASTA format, or uploading a file containing the sequences by clicking the "Browse" button. Up to 25 sequences can be added in one submission. After sequences are submitted, PASC will start computing pairwise identities between user-provided genomes and the existing genome sequences of the family. At the end of the process, for each input genome, PASC produces a list of pairwise identities, from the highest to the lowest, between this input genome and 1) the rest of input genomes (if there are more than one), and 2) 5 to 10 closest matches to existing genomes within the family.

The identity distribution chart will depict the currently selected genome with a different color. One can click on each genome's number to make it current, or can click the identity to see details of the alignment.

PASC can be used to:

1. Establish demarcation criteria for taxonomic classification of certain viruses, as demonstrated in the *Filoviridae* family (23).
2. Identify viruses that were incorrectly assigned in the taxonomy database.
3. Classify viruses with newly sequenced genomes.

PASC can be accessed through <http://www.ncbi.nlm.nih.gov/sutils/pasc>. It currently covers more than 52 virus families/groups, which are listed at <http://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi?textpage=main>.

Genotyping Tool

The retroviral family, *Retroviridae*, is composed of numerous enveloped RNA viruses from which scientific study has revealed many interesting biological principles. Because of both its historical and present health implications, the human immunodeficiency virus 1 (HIV-1) has been of major interest in the scientific and medical communities. As a result, the ability to quickly and efficiently identify HIV genotypes is critical to several areas of scientific and medical research. For instance, because of the almost unavoidable trend of HIV-1 drug resistance in infected individuals, the medical treatment of HIV-1 infected individuals is particularly driven by genotypic studies (24). Likewise drug discovery trials and epidemiological studies are also motivated by similar concerns.

By comparisons to preexisting alignments and trees, phylogenetic analysis can be used to discriminate between viral genotypes as well as to determine the subtype of new isolates. This can pose a particular problem with respect to viruses as coinfection and superinfection leading to inter-subtype recombination is not entirely uncommon (25). Because phylogenetic analyses cannot always distinguish such recombinants and new subtypes, several methods that analyze segments of the genome have been designed (26). The high selective pressure, and high error and replication rate of RNA viruses such as HIV-1 often makes it difficult to align viral sequences automatically.

The NCBI Genotyping tool (27), utilizes an algorithm that uses scored BLAST (22) pairwise alignments between overlapping segments of the query and reference sequences for each virus. The algorithm uses a “sliding window” along a query sequence that processes each window-sequence and segment separately. By comparing each segment to a set of reference sequences from BLAST-derived analyses, a similarity scores for each local alignment is obtained. Each query segment is assigned the reference sequence genotype that matches the query with the highest BLAST similarity score. This process is repeated for every subsequent “window” in the same manner until the entirety of the query sequence is covered by overlapping BLAST alignments. The results from all windows are combined and displayed graphically. By displaying the results of multiple segments in a computer generated graphical format, it is easier for the end-user to

determine the genotype of a query sequence. Likewise, because of the manner in which the results are obtained, recombinant genotypes and recombination breakpoints can also be identified. Currently, the NCBI Genotyping tool utilizes reference sets from HIV-1, hepatitis B virus (HBV), hepatitis C virus (HCV), human T-lymphotropic virus 1 and 2 (HTLV-1 and HTLV-2), simian immunodeficiency virus (SIV) and poliovirus (PV). The tool is located at <http://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi>.

References

1. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, Rozanov M, Tatusov R, Tatusova T. National center for biotechnology information viral genomes project. *J Virol.* 2004 Jul;78(14):7291–8. PubMed PMID: 15220402.
2. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J Virol.* 2009 Oct;83(20):10719–36. PubMed PMID: 19640978.
3. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. *Proc Biol Sci.* 2010 Dec 22;277(1701):3809-17.
4. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol.* 2012 Dec;29(12):3767–80. PubMed PMID: 22821011.
5. Loughran G, Firth AE, Atkins JF. Ribosomal frameshifting into an overlapping gene in the 2B-encoding region of the cardiovirus genome. *Proc Natl Acad Sci U S A.* 2011 Nov 15;108(46):E1111–9. PubMed PMID: 22025686.
6. McFadden N, Bailey D, Carrara G, Benson A, Chaudhry Y, Shortland A, Heeney J, Yarovinsky F, Simmonds P, Macdonald A, Goodfellow I. Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. *PLoS Pathog.* 2011 Dec;7(12):e1002413. PubMed PMID: 22174679.
7. Hughes, SH, Varmus, HE. Retroviruses. New York: CSHL Press; 1997.
8. Sheehy AM, Gaddis NC, Malim MH. The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nat Med.* 2003 Nov;9(11):1404–7. PubMed PMID: 14528300.
9. Franzosa EA, Garamszegi S, Xia Y. Toward a three-dimensional view of protein networks between species. *Front Microbiol.* 2012;3:428. PubMed PMID: 23267356.
10. Fu W, Sanders-Bear BE, Katz KS, Maglott DR, Pruitt KD, Ptak RG. Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D417–22. PubMed PMID: 18927109.
11. Ghedin E, Sengamalay NA, Shumway M, Zaborsky J, Feldblyum T, Subbu V, Spiro DJ, Sitz J, Koo H, Bolotov P, Dernovoy D, Tatusova T, Bao Y, St George K, Taylor J, Lipman DJ, Fraser CM, Taubenberger JK, Salzberg SL. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature.* 2005 Oct 20;437(7062):1162–6. PubMed PMID: 16208317.
12. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. The influenza virus resource at the National Center for Biotechnology Information. *J Virol.* 2008 Jan;82(2):596–601. PubMed PMID: 17942553.

13. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Tatusova T. FLAN: a web server for influenza virus genome annotation. *Nucleic Acids Res.* 2007 Jul;35(Web Server issue):W280-4.
14. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ. Virus taxonomy—ninth report of the International Committee on Taxonomy of viruses. London: Elsevier/Academic Press; 2011.
15. Yu C, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, Yang J, Yau SS. Real time classification of viruses in 12 dimensions. *PLoS One.* 2013 May 22;8(5):e64328. PubMed PMID: 23717598.
16. González JM, Gomez-Puertas P, Cavanagh D, Gorbatenya AE, Enjuanes L. A comparative sequence analysis to revise the current taxonomy of the family Coronaviridae. *Arch Virol.* 2003 Nov;148(11):2207–35. PubMed PMID: 14579179.
17. Fauquet CM, Briddon RW, Brown JK, Moriones E, Stanley J, Zerbini M, Zhou X. Geminivirus strain demarcation and nomenclature. *Arch Virol.* 2008;153(4):783–821. PubMed PMID: 18256781.
18. Bernard HU, Burk RD, Chen Z, van Doorslaer K, Hausen H, de Villiers EM. Classification of papillomaviruses (PVs) based on 189 PV types and proposal of taxonomic amendments. *Virology.* 2010 May 25;401(1):70–9. PubMed PMID: 20206957.
19. Oberste MS, Maher K, Kilpatrick DR, Pallansch MA. Molecular evolution of the human enteroviruses: correlation of serotype with VP1 sequence and application to picornavirus classification. *J Virol.* 1999 Mar;73(3):1941–8. PubMed PMID: 9971773.
20. Adams MJ, Antoniw JF, Fauquet CM. Molecular criteria for genus and species discrimination within the family Potyviridae. *Arch Virol.* 2005 Mar;150(3):459–79. PubMed PMID: 15592889.
21. Bao Y, Kapustin Y, Tatusova T. Virus Classification by Pairwise Sequence Comparison (PASC). In: Mahy BWJ, Van Regenmortel MHV, Editors. *Encyclopedia of Virology*, 5 vols. Oxford: Elsevier; 2008. Vol. 5, 342-348.
22. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997 Sep 1;25(17):3389–402. PubMed PMID: 9254694.
23. Bao Y, Chetvernin V, Tatusova T. PAirwise Sequence Comparison (PASC) and Its Application in the Classification of Filoviruses. *Viruses.* 2012 Aug;4(8):1318–27. PubMed PMID: 23012628.
24. Gallant, JE. "Antiretroviral drug resistance and resistance testing." *Topics in HIV medicine: a publication of the International AIDS Society*, USA. 2005;13.5:138.
25. Ramos A, Hu DJ, Nguyen L, Phan KO, Vanichseni S, Promadej N, Choopanya K, Callahan M, Young NL, McNicholl J, Mastro TD, Folks TM, Subbarao S. Intersubtype human immunodeficiency virus type 1 superinfection following seroconversion to primary infection in two injection drug users. *J Virol.* 2002 Aug;76(15):7444–52. PubMed PMID: 12097556.
26. Siepel AC, Halpern AL, Macken C, Korber BT. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res Hum Retroviruses.* 1995 Nov;11(11):1413–6. PubMed PMID: 8573400.

27. Rozanov M, Plikat U, Chappay C, Kochergin A, Tatusova T. A web-based genotyping resource for viral sequences. *Nucleic Acids Res.* 2004 Jul 1;32 (Web Server issue):W654-9.

Virus Variation

J. Rodney Brister, Ph.D.¹ and Yiming Bao, Ph.D.¹

Created: November 14, 2013.

Scope

As the number of large scale virus genome sequencing projects has grown, so too has the need for specialized resources designed to enhance the accessibility and utility of large sequence datasets. Virus Variation is a comprehensive resource designed to support search, retrieval, and display of large virus sequence data sets—providing users with the functionalities necessary to facilitate discovery activities.

This resource includes a search interface through which users can search and retrieve sequences based on a number of biological and clinical criteria. The selected sequences can then be downloaded or analyzed using a suite of Web-based tools and displays.

Currently, three viruses are included within Virus Variation—Dengue, West Nile, and Influenza—with more than 260,000 individual sequences between them. The resource is expanding and new viruses will be added in response to sequencing efforts and public health demand.

History

The Virus Variation Resource is an outgrowth of the NCBI Influenza Virus Resource originally created in 2004 to support the thousands of Influenza virus genomes sequenced during the National Institute of Allergy and Infectious Diseases (NIAID)-initiated Influenza Genome Sequencing Project (1). The goal of the resource then as now was to provide a suite of interfaces and tools designed specifically for large sequence datasets.

The first iteration of the Virus Variation resource was developed around Flaviviruses, with Dengue Virus added in 2009, and West Nile virus two years later (2). The current implementation combines the previous resources into a single comprehensive construct—building upon historic functionalities but flexible enough to accommodate a broad range of viruses.

Data Model

The Virus Variation Resource is comprised of three components: a specialized database, a unique search interface, and a group of sequence displays. The database is loaded with data processed from GenBank records, and virus-specific annotation pipelines are used to produce standardized, consistent protein and gene annotation across all sequences from a

¹ NCBI; Email: jamesbr@ncbi.nlm.nih.gov; Email: bao@ncbi.nlm.nih.gov.

given species. Automated and manual procedures capture descriptors—metadata—from sequence records, literature, and other databases, then map these to a common vocabulary, and store them with the sequences they describe.

Stored, standardized sequence data and related metadata provide infrastructure for an enhanced search interface that allows users to retrieve and download protein and nucleotide sequence sets based on a variety of biological criteria—like protein or gene of interest, genotype, host, collection country or region, disease severity, and collection date—as well as sequence patterns and key word searches. Specialized tools including a multi-sequence alignment viewer and phylogenetic tree builder use precalculated alignments to rapidly analyze sequences selected by the user and retrieved from the database.

Dataflow

Sequence Annotation Pipeline

Annotation vagaries and inconsistencies are a major impediment to sequence analysis. Virus Variation mitigates this problem using standardized sequence annotation pipelines that provide consistent annotation across all sequences belonging to a given viral species. Reference sequence sets are used to annotate proteins and other biologically and clinically relevant features. For example, the flu annotation pipeline generates information about drug-resistance mutations and the completeness of nucleotide and coding region sequences; both are stored in the database.

In general, the pipelines for each virus loaded into Virus Variation use a common backbone but unique reference protein sets and parsing strategies. For example, in the Dengue annotation pipeline the incoming sequence is initially assigned a genotype using megaBlast and a reference sequence set. That genotype assignment then points the annotation pipeline to a specific set of reference proteins that are used to annotate the new sequence. The annotation pipelines are used for both internal database loading and as a public resource for Influenza virus—providing standardized annotation for some GenBank submissions.

GenBank Submission Pipeline for Influenza Viruses

NCBI is a collaborator with the NIAID Influenza Genome Sequencing Project and has been tasked with gathering Influenza sequences and related metadata from J. Craig Venter Institute (JCVI), annotating the sequences, and releasing them in GenBank. NCBI has created an automated pipeline to facilitate the large number of sequences generated from the project.

In the pipeline, metadata are retrieved and updated daily from a JCVI ftp site and loaded onto an internal NCBI database. NCBI works closely with JCVI, viral sample providers, and the influenza virus research community in establishing the minimum and optional metadata sets to be incorporated into GenBank records. Organism names for new virus

isolates are entered in the NCBI Taxonomy Database. NCBI staff also manually review the metadata and communicate with data providers if there are any issues.

Sequencing data are assembled at JCVI and consensus sequences are verified with the Influenza Virus Genome Annotation Tool (FLAN, for FLu Annotation, see the Virus Genome Processing and Tools Chapter) and then submitted to NCBI through ftp once they are error free. At NCBI, the sequences are processed by FLAN and feature tables generated. These are combined with associated metadata to create GenBank files. In the past 8 years, nearly 11,200 complete influenza virus genomes have been generated from the NIAID project, and published in GenBank.

Database Loading Pipeline

The database loading pipeline is an automated process that parses data from records available in GenBank and maps them to fields used in the Virus Variation database. This process uses generalized parsing strategies to capture both common biological data like host and country of origin, as well as individualized strategies to capture more specific—often clinically relevant—data associated with particular viruses.

The loading pipelines are dependent on vocabulary lists that allow mapping of data parsed from records to controlled descriptors used within the database and displays. For example, host names—including common names and misspelled names—are mapped with these vocabulary lists to scientific names associated with taxonomy IDs in the NCBI Taxonomy Database and host group names like “birds” or “mammals” used in the search pages.

These automated processes are augmented by manual operations based on literature and semi-automated procedures used to capture third-party data releases. Annotation and data capture is also facilitated by community outreach efforts that seek to develop standard, experimentally-driven gene models and reference protein sets. These efforts also encourage the inclusion of rich metadata sets in public database submissions, as well as metadata sharing.

Database

The Virus Variation database stores sequence information derived from the annotation pipeline and associated metadata describing the sample in standardized formats. To balance storage flexibility with efficient data retrieval, Virus Variation combines a relational database with documents containing raw data.

Curation Interface

Since the Virus Variation database loading procedure uses a hybrid of automated and manual procedures, it is important that NCBI staff have the ability to review sequences with loading errors as well as enter data manually into the database. The Virus Variation curation interface enables curators to filter and sort sequences based on virus type, loading errors, and on a number of descriptors like sequence length. A number of editable

fields are displayed for each sequence—such as country, isolation date, and host—in a generalized format that is the same for each virus. Curators can review data associated with a given sequence and enter data manually into these fields as guided by literature or other sources. Additionally, there is the ability to adjust the displayed fields and messages to fit the needs of specific viruses and/or database loading procedures.

Access

The Virus Variation Resource can be accessed at <http://www.ncbi.nlm.nih.gov/genomes/VirusVariation/>. This home page includes links to virus-specific modules.

Search Interface

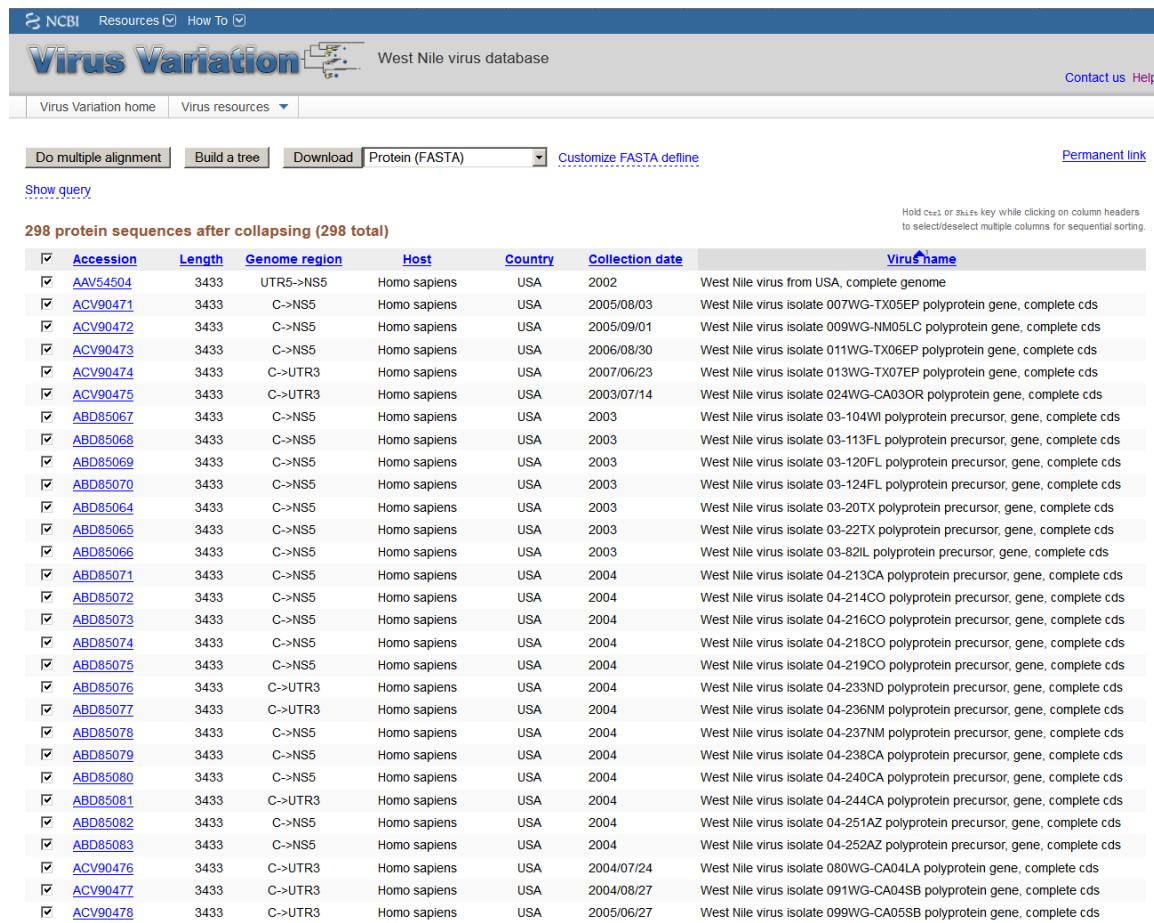
The unique search interface allows users to construct database queries based on a number of criteria including gene or protein region, GenBank accessions, and keywords, as well as biologically relevant descriptors like disease associations, host organism, and geographic information about the sample. Although the same basic interface design is used throughout the resource, the interface is customized to include specific search fields for individual viruses. The number of sequence records retrieved by a search is displayed with the query builder frame of the page—so that the user can modify search parameters. Once the desired query is built, retrieved sequences can be downloaded in a variety of formats directly or can be displayed within the results page.

The screenshot shows the Virus Variation search interface for the West Nile virus database. The top navigation bar includes links for NCBI, Resources, How To, Contact us, and Help. Below the header, there are links for Virus Variation home and Virus resources. The main search area has a section titled "Get sequences by accession" with a text input field for comma-separated accessions or a file upload option. An "Add query" button is also present. The next section, "Select sequence type:", includes radio buttons for Protein (selected) and Nucleotide. A "Search for keyword:" section contains a Keyword input field, a "Search in" dropdown set to "sequence pattern", and a "Define search set:" section. The search set interface includes tabs for Structural and Non-structural genes, with specific regions listed: UTR5, C, prM, E, NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5, and UTR3. Below these are dropdown menus for Host (any, Amphibian, Bird, Human, Mammal), Region/Country (regions, Africa, Asia, Europe, North America), Genome Region (any, UTR5, C, prM, E), and collection/release dates. A checkbox for "Full-length genomes only" is available. At the bottom of the search form are "Add query" and "Clear form" buttons.

Figure 1. The Virus Variation search interface. Users can use a number of search criteria including sequence patterns, host, geographic region, and collection date to retrieve either protein or DNA sequences from specified genome regions.

Results Page

The results page displays all the sequences retrieved in a given search where individual sequences can be selected prior to subsequent analysis or download. Individual records can be sorted by a variety of descriptors, selected or deselected, downloaded, sent to the multi-sequence alignment viewer, or sent to the phylogenetic tree viewer.



The screenshot shows the Virus Variation search results page. At the top, there are navigation links for NCBI, Resources, How To, Virus Variation (with a logo), West Nile virus database, Contact us, and Help. Below the header, there are buttons for 'Do multiple alignment', 'Build a tree', 'Download', 'Protein (FASTA)', and 'Customize FASTA define'. A 'Permanent link' is also present. A 'Show query' link is located below the buttons. The main content area displays a table titled '298 protein sequences after collapsing (298 total)'. The table has columns: Accession, Length, Genome region, Host, Country, Collection date, and Virus name. Each row contains a checkbox followed by the sequence details. A note at the top right of the table says: 'Hold Ctrl or Shift key while clicking on column headers to select/deselect multiple columns for sequential sorting.'

	Accession	Length	Genome region	Host	Country	Collection date	Virus name
<input checked="" type="checkbox"/>	AAV54504	3433	UTR5->NS5	Homo sapiens	USA	2002	West Nile virus from USA, complete genome
<input checked="" type="checkbox"/>	ACV90471	3433	C->NS5	Homo sapiens	USA	2005/08/03	West Nile virus isolate 007WG-TX05EP polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90472	3433	C->NS5	Homo sapiens	USA	2005/09/01	West Nile virus isolate 009WG-NM05LC polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90473	3433	C->NS5	Homo sapiens	USA	2006/08/30	West Nile virus isolate 011WG-TX06EP polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90474	3433	C->UTR3	Homo sapiens	USA	2007/06/23	West Nile virus isolate 013WG-TX07EP polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90475	3433	C->UTR3	Homo sapiens	USA	2003/07/14	West Nile virus isolate 024WG-CA03OR polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ABD85067	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-104WI polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85068	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-113FL polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85069	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-120FL polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85070	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-124FL polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85064	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-20TX polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85065	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-22TX polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85066	3433	C->NS5	Homo sapiens	USA	2003	West Nile virus isolate 03-82IL polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85071	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-213CA polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85072	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-214CO polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85073	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-216CO polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85074	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-218CO polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85075	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-219CO polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85076	3433	C->UTR3	Homo sapiens	USA	2004	West Nile virus isolate 04-233ND polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85077	3433	C->UTR3	Homo sapiens	USA	2004	West Nile virus isolate 04-236NM polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85078	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-237NM polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85079	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-238CA polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85080	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-240CA polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85081	3433	C->UTR3	Homo sapiens	USA	2004	West Nile virus isolate 04-244CA polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85082	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-251AZ polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ABD85083	3433	C->NS5	Homo sapiens	USA	2004	West Nile virus isolate 04-252AZ polyprotein precursor, gene, complete cds
<input checked="" type="checkbox"/>	ACV90476	3433	C->UTR3	Homo sapiens	USA	2004/07/24	West Nile virus isolate 080WG-CA04LA polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90477	3433	C->UTR3	Homo sapiens	USA	2004/08/27	West Nile virus isolate 091WG-CA04SB polyprotein gene, complete cds
<input checked="" type="checkbox"/>	ACV90478	3433	C->UTR3	Homo sapiens	USA	2005/06/27	West Nile virus isolate 099WG-CA05SB polyprotein gene, complete cds

Figure 2. The Virus Variation search results page. Records retrieved during a search can be displayed within the results page where individual sequences can be selected for download or further analysis.

Multi-sequence Alignment Viewer

The multi-sequence alignment viewer allows users to display alignments of selected protein or nucleotide sequences. Alignments are precalculated to save processing time. The viewer is based on the Genome Workbench alignment viewer and includes a number of advanced features including multiple display and scoring options. In the default view a consensus sequence is displayed as the anchor, and a reference feature table is used to define protein (or gene) positions and other important landmarks. The displayed features facilitate navigation along the alignment and allow users to hone in on regions of interest.

The anchor sequence can be changed from a consensus sequence to any in the alignment to facilitate greater scrutiny of specific sequences within the alignment.

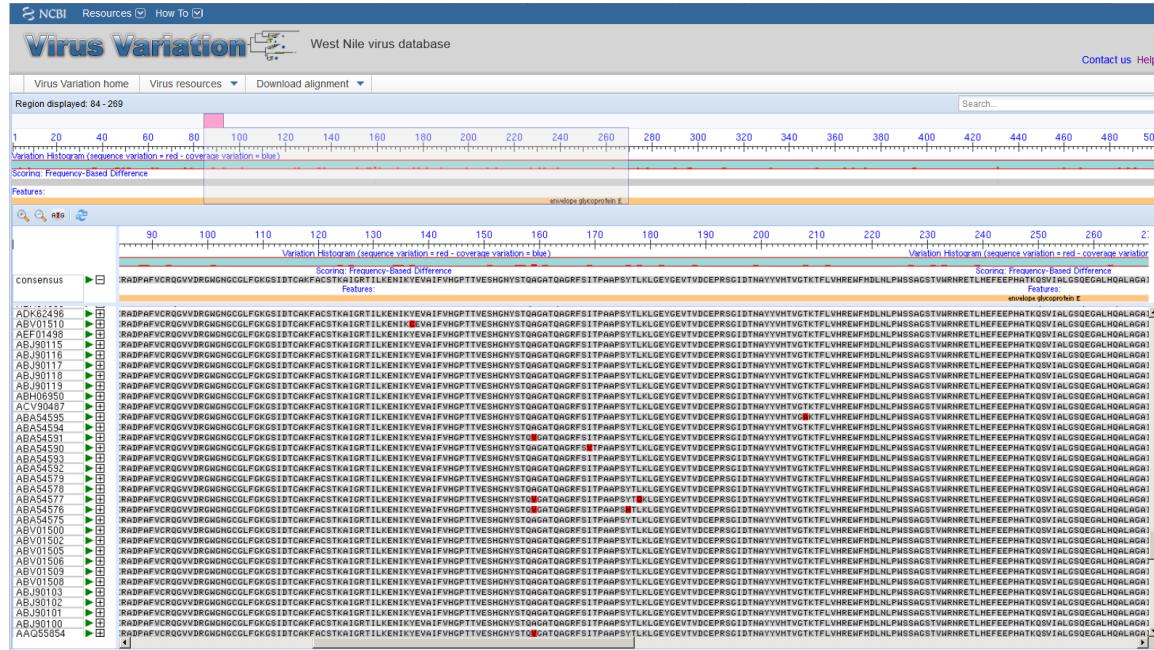


Figure 3. The Virus Variation multiple sequence alignment viewer. Selected protein or nucleotide sequences can be displayed in precalculated alignments allowing rapid comparison of sequences.

Phylogenetic Tree Viewer

The phylogenetic tree viewer displays phylogenetic trees built from alignments of sequences selected in the results page. The current viewer includes collapsible leaves, which allows a user to adjust the resolution of a selected subtree to improve viewing of large data sets (3). Users can also markup sequences based on date range and search for and tag sequences based on country, host, accession, and other descriptors. This feature provides a graphic representation of the metadata associated with sequences, enhancing the user's ability to make associations between phylogenetics and sequence descriptors.

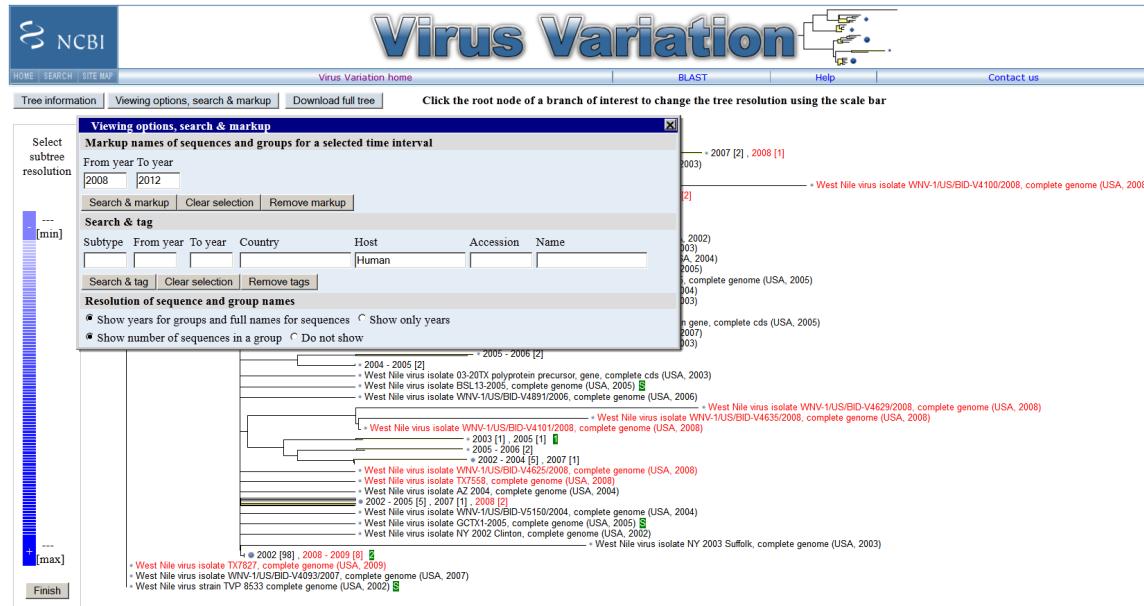


Figure 4. The Virus Variation phylogenetic tree viewer. Selected nucleotide and protein sequences can be quickly displayed on phylogenetic trees using precalculated alignments and a variety of clustering and distance algorithms. Sequences can be searched by metadata, such as country, host, and accession, and marked up in green. Sequences from specific dates can also be highlighted in red.

References

1. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D. The influenza virus resource at the National Center for Biotechnology Information. *J Virol*. 2008 Jan 8;2(2):596–601. PubMed PMID: 17942553.
2. Resch W, Zaslavsky L, Kiryutin B, Rozanov M, Bao Y, Tatusova TA. Virus variation resources at the National Center for Biotechnology Information: dengue virus. *BMC Microbiol*. 2009 Apr 2;9:65. PubMed PMID: 19341451.
3. Zaslavsky L, Bao Y, Tatusova TA. Visualization of large influenza virus sequence datasets using adaptively aggregated trees with sampling-based subscale representation. *BMC Bioinformatics*. 2008 May 16;9:237. PubMed PMID: 18485197.

Variation

Variation Overview

Deanna Church, PhD, Stephen Sherry, PhD, Lon Phan, PhD, Minghong Ward, MS, Melissa Landrum, PhD, and Donna Maglott, PhD.¹

Created: November 14, 2013.

Scope

This chapter provides an overview of the representation of sequence variation in NCBI's databases and a summary of the tools that are available to access and use these data. The resource-specific chapters in this section provide all the details; this overview ties these chapters together and fills in gaps where chapters are not yet available. The variation home page (<http://www.ncbi.nlm.nih.gov/variation>) is NCBI's portal to both databases and tools related to variation.

History

The major databases representing variation at NCBI are the databases archiving information about the location and types of variation, namely dbSNP for variation less than about 50 base pairs (bp), and dbVar for longer, structural variation. Those data are then made accessible from several sites at NCBI (e.g., Gene, Nucleotide, RefSeq) or have value added by establishing connections between variations and multiple types of phenotypes including disease names, clinical features, and gene expression (ClinVar, dbGaP, and PheGenI). Representation of variation at NCBI includes all taxa for which submissions have been received. This ranges from viruses to bacterial pathogens to human. The variation does not have to be heritable; sequence variation that has been observed in tumors or other somatic sources is also represented.

Although information about variation is maintained in distinct databases, the representation in those databases is being standardized to improve searching, reporting, evaluation, and analysis. For example, representation of types of variation (single nucleotide, insertion, copy number gain), of consequences of that variation (nonsense, missense, frameshift), and functional consequences (exon loss) are harmonized to terms from Sequence Ontology (<http://sequenceontology.org/>). With the launch of ClinVar in 2013, standardization reporting of clinical significance is being shifted from the archival databases to ClinVar.

dbSNP and short variation

A major focus of sequencing projects is to identify variations and evaluate their consequences. Beginning in 1998, the National Center for Biotechnology Information

¹ NCBI.

(NCBI) established a database, dbSNP (<http://www.ncbi.nlm.nih.gov/snp>), to manage information about human variation. Even then, the scope of the database was not limited to storing information about single nucleotide polymorphisms (SNPs), rather submission of all types of variation were accepted without restriction by allele frequency.

From the beginning, dbSNP has assigned an accession to each submitted variation (the submitted SNP or ss identifier). Multiple submissions for the same variation and their attributes are integrated to create a reference record (reference SNP or rs identifier. Allele frequency observed in particular populations is also accepted, but not required.

The overwhelming majority of early submissions to dbSNP was in support of HapMap (<http://hapmap.ncbi.nlm.nih.gov>) and represented single nucleotide variations that were indeed polymorphic according to the definition of a minor allele frequency of at least 1%. Thus there is a common misconception that if a variation were in dbSNP, it was indeed a polymorphic single nucleotide change.

dbSNP continues to support tools and reports geared to diverse users, from population geneticists to medical geneticists. An example is the NCBI Variation Viewer tool that has recently been completely rewritten to report all types of human variation.

dbVar and structural variation

dbVar (<http://www.ncbi.nlm.nih.gov/dbvar>) archives information about genomic structural variation from studies submitted for any organism. In general, these variants are longer than 50 bp. Each variant instance is assigned an identifier beginning with nssv. One or more variant instances at the same location are assigned an identifier beginning with ns. This identifier marks a region of the genome that a submitter has defined as containing structural variation. Variant regions point to sets of exemplar variant instances which support the assertion that the region contains variation. Because dbVar exchanges data with DGVa (1), some records may have accessions beginning with essv or esv, for instances and regions respectively.

As the archival databases (dbSNP and dbVar) became established, more and more data were being generated to use that variation to improve our understanding of population genetics, identify regions of the genome that affect rare and common disorders, and identify the effect of variation on gene expression. Thus dbSNP, which originally archived all those data, began to spin off or collaborate with studies having specific scopes. These include the resources in the table below:

HapMap	www.hapmap.org ; now housed at NCBI	Human population structure; identification of blocks of linkage disequilibrium and common variation
1000 Genomes	International project; dbSNP and dbVar maintain identifiers for locations where variation has been observed	Understanding of variation in more populations of apparently healthy individuals

Table continues on next page...

Table continued from previous page.

Genotypes	No dedicated interface	Maintains the genotype information from 1000 Genomes and GO-ESP. Supports displays in our 1000 Genomes and Variation View browsers
PheGenI	http://www.ncbi.nlm.nih.gov/gap/phegeni	Interface to view associations of phenotype to common variation
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar	History of interpretation of medically important variation

Data Model

Archive submissions

A major function of dbSNP and dbVar is to archive submissions. Thus each manages information about the submitter, the date of the submission, the study that generated the data, as well as the content. Part of the archival function includes validating the submission, for example determining if the data are consistent with the genome for which the submission was provided. These archives are accessioned, assigned ss identifiers by dbSNP and nssv by dbVar as appropriate.

ClinVar also archives submissions, namely the interpretation of sequence variation relative to health status. These submissions are assigned a 12-character accession beginning with the letters SCV and followed by numbers padded to 9 places. If submitters submit an update, the accession is assigned a new version.

Aggregate data

dbSNP aggregates data from multiple submissions by location on the genome and type of variation. The result of this aggregation is assigned a refSNP (rs) identifier, which is commonly used to reference that variant location in subsequent studies and publications. It must be emphasized that the rs identifier does not indicate the explicit sequence change at a location. In other words, one rs is assigned to a location on the genome where there is single nucleotide variation, even if all 4 nucleotides have been observed at that location.

ClinVar aggregates data based on the combination of the variation and phenotype. These aggregates are assigned a 12-character accession beginning with the letters RCV and followed by numbers padded to 9 places. The RCV accessions are versioned, with a new version assigned if an SCV is updated or a new SCV is added to the set.

Interpret data

Variation resources do compute some interpretations of the archived information. For example, for human variants, dbSNP and dbVar determine the HGVS representations of variants (<http://www.ncbi.nlm.nih.gov/variation/hgvs>). dbVar and ClinVar compute ISCN coordinates based on sequence location; the variation group calculates the molecular

consequences of a sequence change based on an NCBI Annotation Release. Other interpretations, such as clinical significance, functional consequence of variation, or association results are represented only as submitted.

Access

The variation resources provide interactive access via the Web, application-based access via E-Utilities or other API, data extractions for FTP transfer, and specialized downloads. The variation portal page is the recommended starting point to discover variation information of interest.

Related Tools

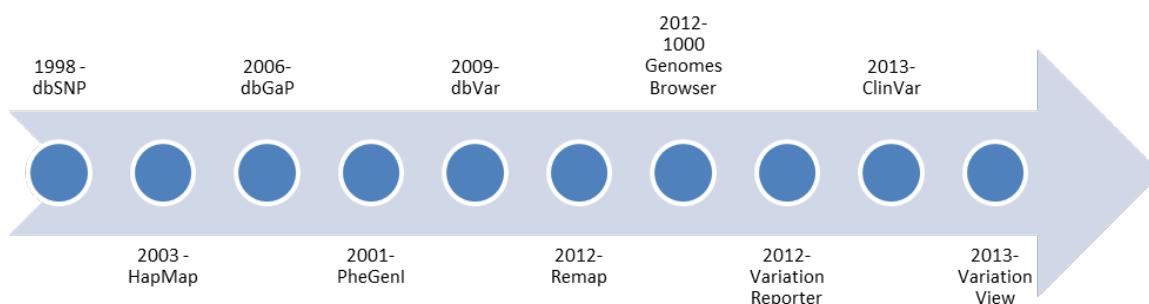


Figure 1. Overview of the introduction of variation-related resources at NCBI.

Clinical Remap

The [NCBI Clinical Remapping Service](#) projects variant coordinates that are on a RefSeqGene or LRG to an assembly, or coordinates from an assembly to any available RefSeqGene or specified list of target RefSeqGenes or LRG.

Because the Clinical Remap Service accepts BED, GVF, HGVS and VCF formats, it can be used to view novel variations with asserted positions in a larger genomic context. When queried, the Clinical Remap service provides a full mapping report, a variation report that shows known dbSNP variants (including those with predicted consequences or with associated clinical information) that map to the same position, as well as an Annotation Data Report and Genome Workbench Files that you can use for further data analysis.

ClinVar

ClinVar is a database that archives the relationship between variations and their possible phenotypes by collecting variations found in patient samples, clinical assertions made about each variant, and the associated data supporting a particular clinical assertion. The ClinVar database can be used as a gateway to additional phenotypic information for a variant including:

- Condition(s) asserted to be associated with the variant

- Available evidence supporting a particular clinical assertion for the variant
- Current interpretation of clinical significance

Links from ClinVar Summary and Individual Accession Reports to related dbSNP records are located in the “Allele Description” section and the SNP track “Sequence View” section.

Although refSNP report pages provide an assertion of clinical significance when available, they currently do not link to ClinVar records as of this writing. dbSNP anticipates reciprocal links to ClinVar from rs report pages in the future.

dbGaP

dbGaP archives and distributes data from studies that examine the relationship between phenotype and genotype. Such studies include Genome-Wide Association Studies (GWAS), medical sequencing, and molecular diagnostic assays. dbGaP allows open access for non-sensitive data, including study documents, phenotypic variables, and genotype-phenotype analyses, but controlled access for restricted data that include pedigrees, pre-computed genotype/phenotype associations, as well as de-identified phenotypes and genotypes of study participants.

Links are available from dbGaP controlled-access records to related variation data in dbSNP, but there are no reciprocal links from dbSNP records to dbGaP since dbGaP data security measures prohibit access to dbGaP individual-level data from any external resource. The refSNP report “Association” section will link to association results from NHGRI_GWAS and/or PheGenI when association data is available.

dbMHC

dbMHC provides a platform where users can access, submit, and edit data related to the human Major Histocompatibility Complex, also called the HLA (Human Leukocyte Antigen).

Both dbMHC and dbSNP store the underlying variation data that define specific HLA alleles. dbMHC provides access to related dbSNP records at the haplotype and variation level, whereas dbSNP provides access to related dbMHC records at the haplotype level.

Gene

The Gene database is the NCBI resource for gene-related data. Gene provides a Variation section that provides links to variation data and tools when data are known to be available.

HapMap

The International HapMap Project site allows access to its catalog of statistically related variations, also known as haplotypes, for a number of different human populations, and is a useful resource for those researchers looking for variations associated with a particular gene. HapMap haplotypes can be searched by a landmark such as a refSNP number or

gene symbol, as well as by sequence region or chromosome region. The resulting HapMap report includes an ideogram with various tracks that can be altered to provide required data, and appropriate tracks in the report will provide direct links to refSNP cluster records.

Phenotype-Genotype Integrator (PheGenI)

PheGenI allows an investigator to use clinical or physical traits, gene name, chromosomal location, or a dbSNP ID number (ss or rs) to search for, view, or download data from various NCBI resources in a single response page. A PheGenI search result can include the following data, if available:

- Related association results from Genome-Wide Association Studies (GWAS) that include links to associated PubMed abstracts
- Related dbGaP study data
- Related expression Quantitative Trait Loci (eQTL) data
- An annotated table of related genes that includes associated OMIM links
- An annotated table of variations from dbSNP
- An interactive view of the genome decorated with search results

PheGenI can be accessed from dbSNP using the PheGenI link, located in the refSNP report “association” field. This link will not be available if the record does not have related association data. PheGenI is also accessed from Gene, via a link provided at the top of the Phenotypes section, labeled *Review eQTL and phenotype association data in this region using PheGenI*.

PubMed

PubMed may provide access to information about variation that is not yet in a specific database. For example, this query [http://www.ncbi.nlm.nih.gov/pubmed/?term=%28mutation\[title\]+OR+variation\[title\]%29+AND+novel\[title\]](http://www.ncbi.nlm.nih.gov/pubmed/?term=%28mutation[title]+OR+variation[title]%29+AND+novel[title]) will retrieve articles with either mutation or variation in the title and novel in the title, and even be restricted to those for which free full text is available ([http://www.ncbi.nlm.nih.gov/pubmed/?term=%28mutation\[title\]+OR+variation\[title\]%29+AND+novel\[title\]%20AND%20%22loatrfree%20full%20text%22\[sb\]](http://www.ncbi.nlm.nih.gov/pubmed/?term=%28mutation[title]+OR+variation[title]%29+AND+novel[title]%20AND%20%22loatrfree%20full%20text%22[sb])).

Variation Batch Submission (VarBatch)

VarBatch is an online, spreadsheet-based interface to facilitate submitting and updating information about human variations described as HGVS expressions. When an asserted clinical variation is processed through VarBatch, it is assigned both a dbSNP submitted SNP (ss) accession as well as a ClinVar accession (format: SCV000000000.0), since the ClinVar accession represents the asserted variation/phenotype relationship.

Variation Reporter

[Variation Reporter](#) matches submitted variation calls to variants in NCBI's databases, and reports back metadata that NCBI has for matching variants. If a variant is novel to NCBI, and the variation is near a feature annotated by NCBI, Variation Reporter will report the predicted molecular consequence based on changes to that annotated feature.

Variation Viewer

Variation Viewer allows a user to review variation in the context of multiple types of sequence features and filter results to a subset of interest. The user can select the assembly; search for features such as genes across the genome; upload local data to view in the context of what is available from NCBI; navigate by gene symbol, exon, rs#, nssv accession, cytogenetic band; retain a history; and filter displays by Variant type, Molecular consequence, minor allele frequency from 1000 genomes (1000 Genomes MAF), minor allele frequency from GO-ESP, and representation in dbSNP, dbVar, and ClinVar. The sequence display, based on NCBI's graphical viewer, provides multiple track options including segmental duplications, paralogous regions, annotation releases from NCBI and Ensembl, somatic variants, common variants, RNAseq support of intronic features, and repeats. The track selection is designed to support evaluation of variations in the region of interest.

1000 Genomes Browser

The [1000 Genomes Browser](#) provides access to 1000 Genomes data, including variations, genotypes, and sequence read alignments within the context of GRCh37, the reference assembly used by the 1000 Genomes Project (2) for analysis. The browser allows you to configure the display to include multiple data tracks of interest, and provides links to related data housed in various NCBI resources. The 1000 Genomes Browser allows users to quickly view variation, allele frequencies or counts by population group, and alignments of reads to support review of the evidence used in calling the sequence. Detailed [instructions](#) about how to use the browser are provided.

References

1. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flliceck P, Church DM. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* 2013;41:D936–D941. PubMed PMID: 23193291.
2. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65. PubMed PMID: 23128226.

The Database of Genotypes and Phenotypes (dbGaP) and PheGenI

Kimberly A Tryka,¹ Luning Hao,¹ Anne Sturcke,¹ Yumi Jin,¹ Masato Kimura,¹ Zhen Y Wang,¹ Lora Ziyabari,¹ Moira Lee,¹ and Michael Feolo¹

Created: August 15, 2013.

Scope

The Database of Genotypes and Phenotypes (dbGaP) is a National Institutes of Health (NIH) sponsored repository charged to archive, curate and distribute information produced by studies investigating the interaction of genotype and phenotype (1). It was launched in response to the development of NIH's [GWAS policy](#) and provides unprecedented access to very large genetic and phenotypic datasets funded by National Institutes of Health and other agencies worldwide. Scientists from the global research community may access all public data and apply for controlled access data.

The information contained in dbGaP includes individual level molecular and phenotype data, analysis results, medical images, general information about the study, and documents that contextualize phenotypic variables, such as research protocols and questionnaires. Submitted data undergoes quality control and curation by dbGaP staff before being released to the public.

Information about submitted studies, summary level data, and documents related to studies can be accessed freely on the dbGaP website (<http://www.ncbi.nlm.nih.gov/gap>). Individual-level data can be accessed only after a Controlled Access application, stating research objectives and demonstrating the ability to adequately protect the data, has been approved (<https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login>). Public summary data from dbGaP are also accessed without restriction via the PheGenI tool, as detailed in the Related tools section.

History

Planning for the database began in 2006 and the database received its first request for data in mid-2007. The initial release of dbGaP contained data on two Genome-Wide Association Studies (GWAS): the Age-Related Eye Diseases Study (AREDS), a 600-subject, multicenter, case-controlled, prospective study of the clinical course of age-related macular degeneration and age-related cataracts supported by the National Eye Institute

¹ NCBI; Email: trykak@ncbi.nlm.nih.gov; Email: hao@ncbi.nlm.nih.gov; Email: kianga@ncbi.nlm.nih.gov; Email: jinyu@ncbi.nlm.nih.gov; Email: kimurama@ncbi.nlm.nih.gov; Email: jawang@ncbi.nlm.nih.gov; Email: ziyabarl@ncbi.nlm.nih.gov; Email: leemoira@ncbi.nlm.nih.gov; Email: feolo@ncbi.nlm.nih.gov.

(NEI), and the National Institute of Neurological Disorders and Stroke (NINDS) Parkinsonism Study, a case-controlled study that gathered DNA, cell line samples and detailed phenotypic data on 2,573 subjects. The data from the Genetic Association Information Network (GAIN) (2) was released soon after.

Although initially designed for GWAS, the scope of dbGaP has expanded to facilitate making individual level information accessible to research communities and to provide data needed to understand the manifestation of disease and how that relates to the genome, proteome and epigenome. The dbGaP has been growing rapidly since its inception. See the dbGaP [home page](#) for current content.

Data Model

Accessioned Objects

The data in dbGaP are organized as a hierarchical structure of studies. Accessioned objects within dbGaP include studies, phenotypes (as variables and datasets), various molecular assay data (SNP and Expression Array, Sequence, and Epigenomic marks), analyses, and documents (Figure 1). Each of these is described in its own section below.

Studies

The data archived and distributed by dbGaP are organized as studies. Studies may be either stand-alone or combined in a “parent study/child study” hierarchy. Parent or “top level” studies may have any number of child studies (also referred to as substudies). However, study hierarchy is limited to two levels (parent and child only). In other words, substudies may not have substudies. Studies, whether parent or child, can contain all types of data ascertained in genetic, clinical or epidemiological research projects such as phenotype and molecular assay information that are linked via subject and sample IDs. Studies often contain documents, such as questionnaires and protocols, which help contextualize the phenotype and genotype data. Study data are distributed by consent groups, each of which contains all data from a set of study participants who have signed the same consent agreement. In other words, the data delivered for a single consent group will all have the same Data Use Limitations (DULs) for future research use.

Each study is assigned a unique accession number which should be used when citing the study. The general dbGaP accession format for a study is phs#####.v#.p#. The first three letters [phs] denote the object type ('s' denotes study), followed by 6-digit, 0-padded object number (#####) which is consecutively assigned by dbGaP. The version number (.v#) indicates updates of the object, where # is initially 1 and increments by 1 as the object is updated. The version number is followed by participant group (.p#) where # is initially 1 and increments by 1 as the participant group changes. The version number of a study will increment any time the version of an object contained by the study (such as a phenotype variable, genotype data, or a sub study) is updated. The participant group of a study will change when existing subjects are removed, or when an existing subject changes from one consent group to another, but not when additional subjects are added.

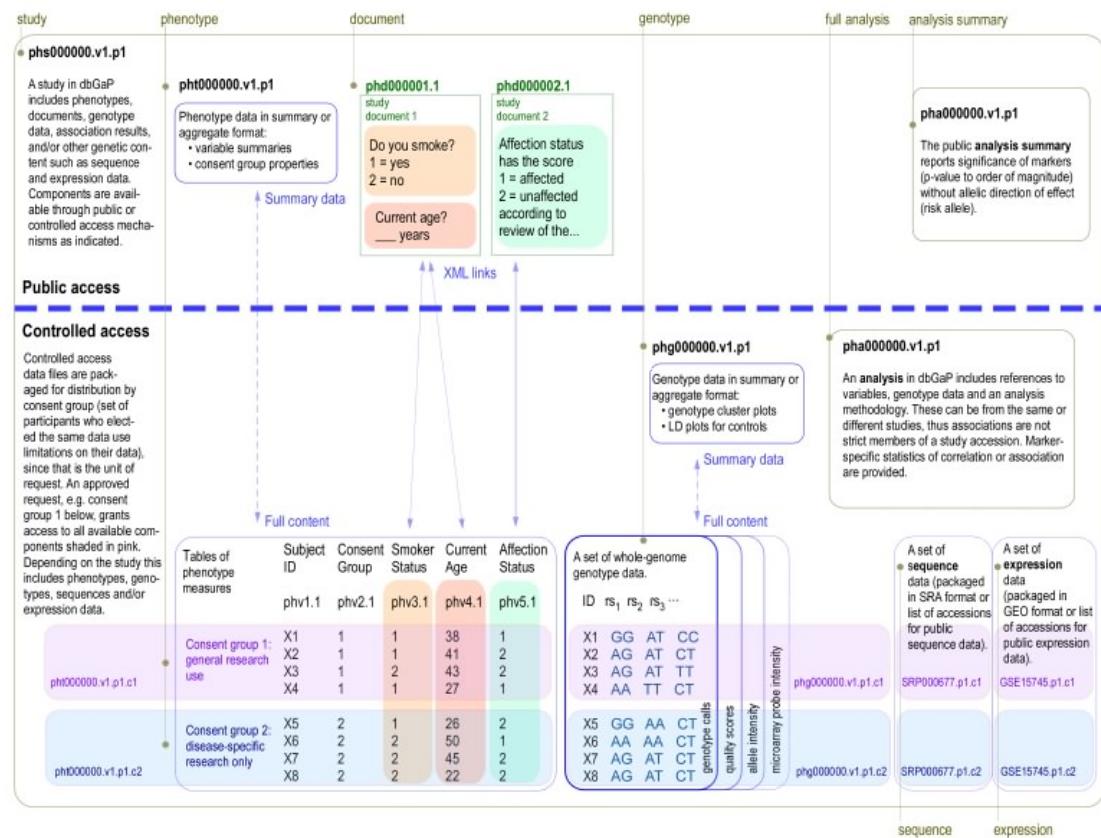


Figure 1. This figure shows the relationships between dbGaP accessioned objects and whether they are available publicly or only through Controlled Access. This is an updated version of a figure that originally appeared in Mailman, et al. 2007.

While the data found in studies can vary widely based on both the number of participants and the variety of deposited phenotypic and molecular data, all studies include basic descriptive metadata such as study title, study description, inclusion/exclusion criteria, study history, disease terms, publications related to the study, names and affiliations of the principal investigators, and sources of funding. This information is publicly available on the study's report page at the [dbGaP website](#).

Datasets and Variables

Phenotypic data values are submitted to dbGaP as tabular files or datasets (accessioned with pht#, where 't' denotes table), where columns represent phenotypic variables (accessioned with a phv#, where 'v' denotes variable) and rows represent subjects. A dbGaP phenotype variable consists of two parts: the data values and the description of the data in the accompanying data dictionary. Each cell (value) in a dataset is stored in a relational database and is mapped to the appropriate phenotype variable and subject. Phenotype variable metadata are provided by the submitter via a data dictionary for each dataset and include: variable name, variable description, units, and a list of any coded

responses. The variable's data type (text string, integer, decimal or date) is automatically determined by calculating which type is in the majority. Conflicts between submitted and calculated data types, or other discrepancies in the data, are reconciled by dbGaP curators in consultation with the data submitter.

Variables are created from the columns of the dataset; each variable and dataset is accessioned using the general dbGaP format ph(v|t)#####.v#.p#. A variable's version (v#) will change when either values of data change or its entry in the data dictionary changes. A dataset's version will change when a variable inside the dataset is added, updated or deleted. For both variables and datasets the participant set (p#), is inherited from the study to which it belongs. Variables, and sometimes datasets, are linked to appropriate sections of documents (please see the Website section below for details).

Individual level phenotype data is only available through the dbGaP [Authorized Access System](#). Public summary-level variable information is available on the dbGaP [website](#) and [ftp site](#).

Genotype data

Genotype data hosted at the dbGaP consist of individual level genotypes and aggregated summaries, both of which are distributed through the dbGaP [Authorized Access System](#). The types of data available include DNA variations, SNP assay, DNA methylation (epigenomics), copy number variation, as well as genomic/exomic sequencing. RNA data types such as expression array, RNA seq, and eQTL results are also available. For details about the accepted format of submitted genotype files please see the dbGaP [submission guide](#).

Genotype data are accessioned based on their data type and use the general dbGaP accession format ph(g|e|a)#####.v# where 'g' denotes GWAS, 'e' expression, and 'a' analysis. Versioning of genotype data is triggered by addition or withdrawal of samples, sample consent status change, or error correction.

Genotype data files are compressed and archived into tar files for distribution. The files are explicitly named to indicate file content, such as image data (cel and idat), genotype calls (genotype), and locus annotations (marker info). Genotype calls are usually clustered according to file format and genotyping platform, including one sample per file (indfmt), multiple-sample matrix (matrixfmt) and pre-defined variant call format ([vcf](#)). They will be accompanied by sample-info file for subject lookup and consent status. The consent code and consent abbreviation are also embedded in the file name.

Examples of genotype data file names:

phe000005.v1.FHS_SHARe_project4_miRNA.sample-info.MULTI.tar

phg000006.v6.FHS_SHARe_Affy500K.genotype-calls-matrixfmt.c2.HMB-NPU-MDS-IRB.tar

The various pieces of the names can be parsed to extract meaningful content: the genotype accession (phe000005.v1 and phg000006.v6); the study (FHS_SHARe in both cases); the molecule type (miRNA); the platform/chip information (Affy500K); the content type (sample-info or genotype-calls-matrixfmt); the consent code (c2); and the consent abbreviation (HMB-NPU-MDS-IRB).

Analyses

Because of the large volume of data generated and concerns regarding participant confidentiality many genetic epidemiological analyses have not been published. But, because individual-level data is only accessed through Controlled Access, dbGaP can archive, integrate and distribute these results.

Analyses can either be provided by submitters or be pre-computed by dbGaP staff, though pre-computes account for a small number of the total analyses. Submitted analysis results are accessioned with the prefix “pha”. After removing identifiable elements, like counts and frequencies, analysis results are displayed in the public dbGaP browser that dynamically links to NCBI annotation resources, like [dbSNP](#), [Gene](#), [RefSeq](#). These public views can be found through the “Analysis” link on the study page and they can be downloaded from the FTP site. The original submitted analyses, including updated marker info, are fully accessible through dbGaP Controlled Access.

Analysis files are typically formatted by population, trait and analysis method (typical), however some include surveys across multiple populations, and may use either SNPs or genes as the loci analyzed. However, in general, they all contain the following three parts which are also required for dbGaP submission.

1. Metadata, which includes trait, population, sample size and brief descriptions on Analysis and Method.
2. Marker information and genotyping summary, such as identifiers of loci (variation, gene and structure variant), alleles, genotype counts (frequency), call rate and p-value from Hardy-Weinberg-equilibrium testing.
3. Testing statistics, including p-value, effect size (odds ratio/regression coefficient/relative risk) and direction (coding allele) if association results.

With these resources, other scientists can verify the discoveries, recalculate statistics under various genetic models, develop new hypotheses, and more importantly, construct a meta-analysis even though individual-level data are inaccessible. The details of data fields are listed in dbGaP submission guides and we welcome suggestions and comments from the scientific community. An interactive view of the analysis results submitted to dbGaP is described in The dbGaP Genome Browser in the Related Tools section of this chapter.

Documents

The dbGaP encourages investigators to submit documents related to their studies, such as protocols, patient questionnaires, survey instruments and consent forms, along with their

data. These documents provide valuable information and context for subsequent researchers who will apply for and download datasets. All submitted documents are available publicly and can be used by anyone interested in gaining a better understanding of the phenotypic data found in a study.

Each document is accessioned using the general dbGaP format phd#####.v# where “d” indicates document. A document’s version (v#) will change when the variables annotated on the document change, or when the document itself is changed significantly. (For example, fixing a typo would not be considered a significant change unless it were to change the meaning of the document.)

Documents submitted to dbGaP are represented in a common XML format. Converting documents into a common format allows all documents to be treated uniformly in the database (aiding indexing and discovery) and to be displayed in a single HTML style. Additionally, the XML format allows curated information to be added to the documents. This curated information is used to create live links between the documents and other portions of the dbGaP website, such as variable report pages. Linking between documents and other objects will be discussed further in the section about the dbGaP website.

Documents are generally viewable on the dbGaP website in both HTML and PDF format (the PDF for a document may be the originally submitted object, if it was sent as PDF, or could be a PDF representation of another format such as Microsoft Word or Excel or a plain text file).

The XML used by dbGaP is an extension of NLM’s Archiving and Interchange Tag Set Version 2.3 (<http://dtd.nlm.nih.gov/archiving/2.3/>). The extension adds structures to code questionnaires and adds a number dbGaP-specific attributes to common document structures (such as sections, tables, and lists) to facilitate curation. A copy of our extension is publicly available at <http://dtd.nlm.nih.gov/gap/2.0/wga-study2.dtd>, and documentation for the extension is located at <http://dtd.nlm.nih.gov/gap/2.0/doc/wga-document.html>.

Dataflow

Submissions

The NIH strongly supports the broad sharing of de-identified data generated by NIH-funded investigators and facilitates data sharing for meritorious studies that are not NIH-funded. Decisions about whether non-NIH-funded data should be accepted are made by individual NIH Institutes and Centers (IC); ICs will not accept data unless the submission is compatible with NIH’s GWAS policy.

NIH-Funded Studies

Institutional certification, as well as basic information about the study, is required when submitting data to dbGaP.

- **Institutional certification** consists of a letter signed by the principal investigator and an institutional official that confirms permission to submit data to dbGaP. NIH has developed [Points to Consider for IRBs and Institutions](#) to assist institutions in their review and certification of an investigator's plan for submission of data to dbGaP.
- **Basic information** consists of items like the title of the study, a description and history of the study, inclusion and exclusion criteria, listing of previous and certification of PI's data
- Principal investigators (PIs) and funding information.

Principal investigators (PIs) should familiarize themselves with the "[NIH Points to Consider](#)" document that provides information about: the NIH GWAS Data Sharing Policy; benefits of broad sharing of data through a central data repository; risks associated with the submission and subsequent sharing of such data; safeguards designed to protect the confidentiality of research participants; and specific points for institutional review boards (IRBs) to consider during review and certification of PIs' data submission plans.

The principal investigators must contact their NIH program official (PO) to begin the submission process. If the study was not funded by the NIH, PIs should contact dbGaP-help@ncbi.nlm.nih.gov for guidance.

Non-NIH-Funded Studies

To submit non-NIH-funded data to dbGaP the following information will need to be provided:

- **Institutional certification** as described in the last section. To provide this, someone from the institution or organization will need to be registered in eRA Commons. Information regarding registration is available from the [eRA Commons website](#). (Note: The review of a PI's request can be initiated without the certification, but the review process will be expedited if GWAS staff receives the certification at time of submission.)
- **Basic information** about the study, as described in the previous section.
- **The NIH IC** that most closely aligns with the research. A list of ICs can be found at <http://www.nih.gov/icd/>.
- **Whether the study has been published or accepted for publication.** If it has the PI should provide documentation (i.e., the publication citation or a copy of any correspondence indicating that an article about the study has been accepted for publication).

The PI should submit all information and the certification to GWAS@mail.nih.gov. Once GWAS staff receives the documents, they will forward them to the appropriate IC program administrator for consideration. The IC program administrator will contact the PI with any questions and/or to notify you of the IC's decision.

The PI is encouraged to consult with the Program Officer/Director (PO/PD) and/or **IC GWAS Program Administrator (GPA)** at an NIH Institute or Center (IC) to discuss the

project, data sharing plan, and data certification process (non NIH funded projects should contact dbGaP Help) to complete the registration process.

Instructions for submitters

Study Registration

Before data can be submitted to dbGaP, the study must be registered in the dbGaP Registration system following these steps:

- The GPA from the sponsoring IC gathers study registration information from the PI.
- Completes the study registration in the dbGaP Registration System by providing:
 - Study details
 - Signed Institutional Certification
 - Approved Data Use Certification (DUC)
 - Consent groups and Data Use Limitations (DUL)
- The Registration System sends an automated email to the investigator upon completion of the study registration acknowledging the study registration and giving further instructions on how to submit data to dbGaP.

Data submission

The PI will be provided with the [dbGaP Submission Guide](#). The packet contains templates and instructions on how to format the data files for submission to dbGaP. The expected files for each single study are:

- 1_dbGaP_StudyConfig*
- 2a_dbGaP_SubjectPhenotypesDS
- 2b_dbGaP_SubjectPhenotypesDD
- 3a_dbGaP_SampleAttributesDS*
- 3b_dbGaP_SampleAttributesDD*
- 4a_dbGaP_SubjectDS*
- 4b_dbGaP_SubjectDD*
- 5a_dbGaP_SubjectSampleMappingDS*
- 5b_dbGaP_SubjectSampleMappingDD*
- 6a_dbGaP_PedigreeDS**
- 6b_dbGaP_PedigreeDD**
 - * Required

** Required if there are related subjects

Note for studies that expect/involve SRA (Sequence Read Archive) file submission.

Once the required files (listed above) are received by dbGaP, passed dbGaP QCs, and the subject consents and subject sample mapping have been loaded into dbGaP and provided

continues on next page...

continued from previous page.

to BioSample, the PI will be provided with a study accession number and a link to the corresponding study sample status page. The SRA submitter can then apply for an SRA submission account (Aspera account) and submit SRA files to dbGaP.

Data processing

Data received at dbGaP undergoes a sequence of processing steps. Figure 2 illustrates the steps involved in moving a study through dbGaP. The first step is getting the study registered (as has been discussed above), and occurs before data transmission (shown in gray). The dbGaP data processing occurs in two pipelines, the phenotype curation (blue) and genotype curation (purple). These pipelines are largely processed in parallel but converge prior to data release. The final step is preparing the study for release to the public (green). Particulars of the phenotype and genotype curation will be discussed below.

Phenotype processing

The phenotypic data are subjected to both automated and human-mediated assessment.

Before the data are loaded into the database, scripts are used to evaluate the following:

- **Poor formatting** - Each dataset submitted to dbGaP should be a rectangular table showing variables in columns and subject or sample IDs in rows.
- **HIPAA violations** - The datasets submitted to dbGaP should follow the Health Insurance Portability and Accountability Act (HIPAA) rules in order to protect the privacy of personally identifiable health information.
- **Issues with Subject and Sample IDs** - Submitters are required to use the subject consent file to list all the subjects who participated in, or are referred to by, the study, with their consent values. Subjects who had not directly participated in the study, e.g., parents of participants included in the pedigree file, should also be included in the consent file with consent value 0.
- **Missing information in Data Dictionaries** - The submitters are required to submit data dictionaries, in addition to datasets, to explain the meanings of the variables and data values. For each value in the datasets, dbGaP requires that the submitters provide a variable description, variable type, units of values for numerical variables, and logical minimum and maximum values if available. For each encoded value, a code meaning should be included in the data dictionary.
- **Issues with Pedigree files** - A pedigree file submitted to dbGaP should include the following columns: family ID, subject ID, father ID, mother ID, sex, and twin ID if available. We also require that all the subjects who appear in father ID or mother ID columns also be included in the subject ID column.

More detailed information about the particular automated testing performed to catch errors in the broad categories listed above can be found in Appendix – Phenotype Quality Control.

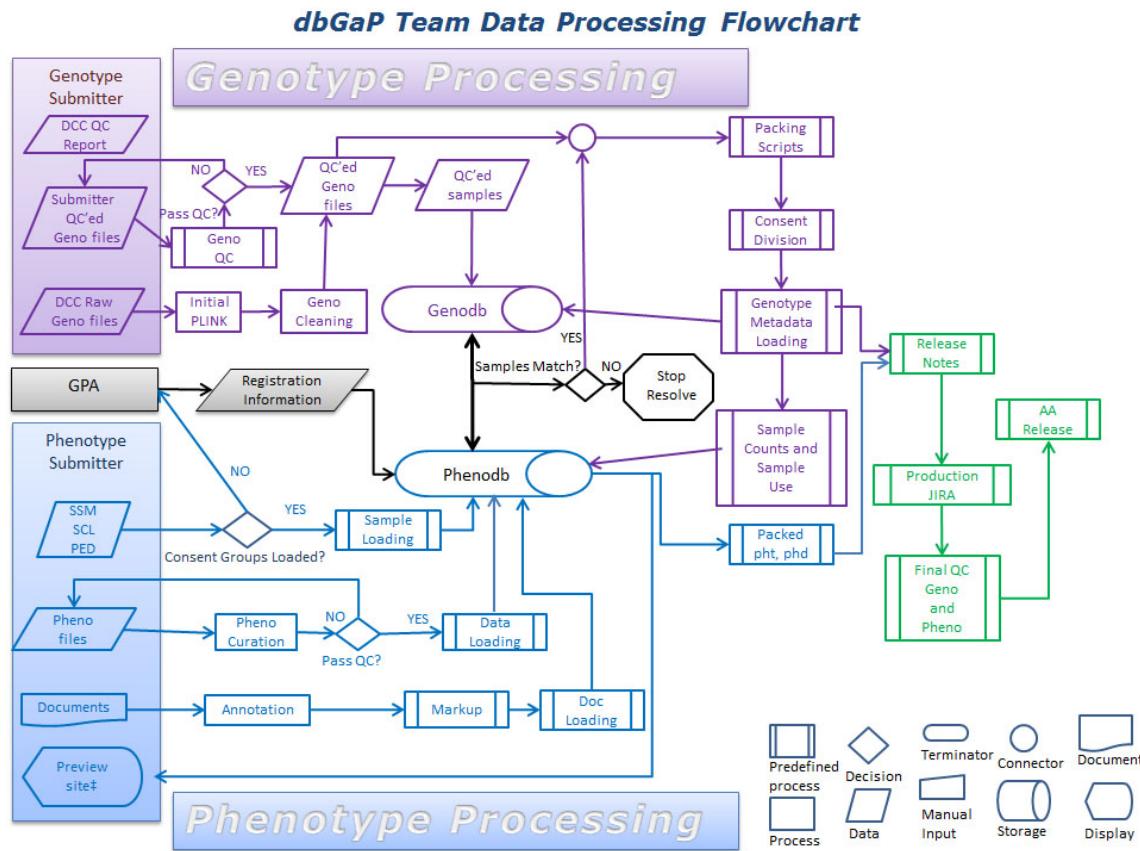


Figure 2. This figure shows the complex processing that occurs after data has been submitted to dbGaP. Of particular note is the step, in the center of the chart in black, where samples are matched.

Reports from the automated quality control (QC) scripts are reviewed by curatorial staff. If necessary, curators will communicate with the submitters and ask that new files be submitted correcting the errors. Even if the automated QC checks do not detect problems, the curatorial staff check all data dictionaries manually to see if there are any problems that were not identified by automated checks.

Genotype processing

The genotype data processing and QC process consists of the following steps:

1. Check for availability of Sample-Subject Mapping file and validate against subject list.
2. Check for availability of sample genotype file and validate against SSM.
3. Process and genotype file and generate PLINK (3) set.
4. Check for data consistency in the submitted QC component against data from other submissions for the study.
5. Conduct QC checks and generate genotype-QC component. This step includes checking:

- a. Missing call rates per sample/per marker.
 - b. Minor allele frequency.
 - c. Mendelian error rate when trios are available.
 - d. Duplicate concordance check (Generate SNP and subject filters based on tests results
 - e. when dups are available).
 - f. Gender check.
 - g. IBD analysis.
6. Generate SNP and subject filters based on tests results.
 7. Verify genotype and phenotype data using NCBI GWAS pre-compute against similar pre-compute provided by PI/analysis group.
 8. Split PLINK sets according to consent and generate genotype-calls-mtrxfm components.
 9. Generate sample-info and marker-info release components.
 10. Partition and pack build of individual genotype files according to subject consent information and data type.

The quality of the genotype data is checked at both the genotype data file and the genotyping level. Typically at the file level a genotype release contains individual level data in both individual (one file per sample) and matrix (one matrix with all samples) formats. The genotype matrices are generated by dbGaP curators from submitted individual genotype files and subject related information, such as gender and pedigree data. These matrices then are used to generate pre-computes/metrics/QC-filters, which are further verified against similar pre-computes submitted by the investigators. When necessary, the genotyping quality of each sample is also verified using a B-Allele frequency (BAF) analysis pipeline which calculates and processes BAF values to identify samples with extremely “noisy” or failed genotyping.

The following quality assurance steps are implemented to facilitate cross-study and cross-technology data merging and analysis:

1. Identify duplicated genotypes across all studies as well as generating data.
2. Check data formats, annotation, and QC- metrics for genotype data derived using different technologies.
3. Check ID reconciliation, gender, missing call rate, duplicate concordance, identity-by-descent, and Mendelian error rate at sample as well as SNP level.

Subjects and Samples

In dbGaP there are two similar yet distinct concepts that describe the participants in a study: subject and sample. A subject corresponds to an individual human. A sample in dbGaP corresponds to each analyte (DNA/RNA) that is put in the machine or on the chip, rather than the physical result of obtaining a tissue or blood sample, though this is accepted as well. Modeling samples this way allows dbGaP to track duplicates, centers, plates, wells, and any sequence of aliquots that precedes the actual aliquot used to produce

the molecular data finally submitted to dbGaP. The tracking method also enables dbGaP to easily add, rename, or redact samples over time.

Example.

Consider a case in which a single subject has both a blood draw and a cheek swab. DNA is extracted from both samples. The DNA extracted from the blood is stored for years after being drawn, and then sequenced on two different platforms, and the DNA extracted from the cheek swab is also used on a GWAS chip. In this scenario, dbGaP prefers to receive three samples belonging to the same subject (even though there were two intermediate physical samples).

Note: The information about intermediate samples may be informative and can be included as one or more variables in the sample attribute file.

Submitters are required to assign de-identified IDs to subjects and samples; these are submitted subject and sample ids. However, dbGaP also assigns a unique id to samples and subjects; these are the dbGaP Subject and Sample IDs and are included in the final phenotype files available through controlled access. The dbGaP assigns its own IDs to accurately represent cases where a single subject (person) has participated in more than one study. In such a case the two submitted subjects will be assigned the same dbGaP Subject ID. This can only be done if the submitter provides the information that a subject in their study is the same subject as in an existing dbGaP study. Similarly, cell repository, or otherwise readily available samples such as Coriell samples, used as controls in multiple studies will typically receive the same dbGaP Sample ID.

All phenotype and molecular data are connected through the Subject Sample Mapping file.

Data ID mapping

The data submitted to the dbGaP are de-identified. The phenotype and genotype data are connected through the subject sample mapping file in which one sample is mapped to exactly one subject and one subject is mapped to any number of samples. The following is a partial list of the IDs and attributes included in dbGaP phenotype and molecular data files.

1. **SUBJECT_ID:** This is the submitted Subject ID and it is included in the Subject Consent Data File, the Subject Sample Mapping Data File, the Pedigree Data File (if applicable), and all Subject Phenotype Data Files. SUBJECT_ID should be an integer or string value consisting of the following characters: English letters, Arabic numerals, period (.), hyphen (-), underscore (_), at symbol (@), and the pound sign (#). In addition to the submitted Subject ID, dbGaP will assign a dbGaP Subject ID that will be included in the final phenotype dump files along with the submitted Subject ID.

2. **SAMPLE_ID:** This is the submitted Sample ID and is included in the Subject Sample Mapping Data File and the Sample Attributes Data File. This ID should be used as the key for the individual level molecular data. Each sample should be submitted with a single, unique, de-identified Sample ID. The acceptable characters in Sample IDs are the same as those in the Subject IDs. In addition to the submitted Sample ID, dbGaP will assign a dbGaP Sample ID that will be included in the final phenotype dump files along with the submitted Sample ID. The SAMPLE_IDs listed in the Subject Sample Mapping Data File should be identical to the samples found in the genotype, SRA, and other molecular data.
3. **dbGaP_SAMPLE_ID:** This is the dbGaP assigned unique identifier assigned to the submitted Sample ID. The dbGaP Sample ID is included as a column in the final phenotype dump files whenever there is a submitted sample ID column.
4. **dbGaP SUBJECT_ID:** This is the dbGaP unique identifier assigned to the submitted Subject ID. The dbGaP Subject ID is included as a column in the final phenotype dump files whenever there is a submitted subject ID column. The dbGaP Subject ID is unique cross all dbGaP studies, which means that if a subject is known to have participated in multiple studies that have been submitted to dbGaP, the same dbGaP Subject ID will be assigned to the individual across multiple studies, though the submitted subject ID may be different.
5. **SOURCE SUBJECT_ID and SUBJECT_SOURCE:** The Source Subject ID (SOURCE SUBJECT_ID) is the de-identified alias Subject ID used in the public repository, consortium, institute, or study from where the subject has been obtained. The Subject Source (SUBJECT_SOURCE) is the name of the third party source, public repository, consortium, institute, or study that corresponds to the subject. For subjects originating from a shared source (such as a public repository, consortium, institute, study, etc.) or for subjects with alias IDs, these 2 variables will be included in the Subject Consent Data File. The SOURCE SUBJECT_ID maps to the SUBJECT_ID. For referencing HapMap subjects from Coriell, the SUBJECT_SOURCE value is written as "Coriell." The SOURCE SUBJECT_ID should be written as the de-identified subject ID assigned by Coriell (e.g., NA12711).
6. **FAMILY_ID:** The Family ID is a column of de-identified Family IDs in the pedigree file. The Family ID is also referred to as the Pedigree ID. The family ID should be the same for individuals belonging in the same biological family. The family ID is found in the pedigree file if a pedigree file is available.
7. **SEX:** The gender variable can be included in a subject phenotype data file or in a pedigree file if a pedigree file is available.
8. **FATHER and MOTHER:** In the pedigree file, FATHER and MOTHER are the two columns of the unique, de-identified subject IDs of the participant's biological father and mother. The Father ID and Mother ID may not be identical. 0 (zero) or blank is filled in for founders or marry-ins (parents not specified) in a pedigree. Each unique Father ID and unique Mother ID is also listed in the Subject ID column of both the Pedigree Data File and the Subject Consent Data File.

9. **TWIN_ID:** Monozygotic twins or multiples of the same family have Twin IDs. Twins or multiples of the same family share the same TWINID, but are assigned different SUBJECT_IDS.
10. **CONSENT:** Every subject that appears in a Subject Phenotype Data File must belong to a single consent group and every sample that appears in a Subject Sample Mapping File and in a Sample Attribute Data File must belong to a consented subject. The consent information is listed in the Subject Consent Data File. Consents are determined by the submitter, their IRB, and their GPA (GWAS Program Administrator) along with the DAC (Data Access Committee). All data is parsed into its respective consent groups for download.

Curatorial document annotation

One thing that sets dbGaP apart from similar databases is the extent of curatorial work done with the data and documentation we receive. For documents, this involves making connections between appropriate portions of text and other accessioned objects (such as variables, data tables, and other documents) and creating links to external resources. We refer to this process as “document annotation” and it involves embedding references into the XML for the documents.

Documents can either be annotated by the submitter or by the dbGaP curator responsible for a study. Types of annotations include adding variable IDs to particular sections of text so that the text can be linked to the variable report page or adding references so that hyperlinks can be made between chapters of a protocol document.

Access

Public data (unrestricted)

dbGaP Website

Report types

The web site provides reports specific to the objects in dbGaP. These reports are explained in the following.

Study

- the study's accession, name, description, history, inclusion/exclusion criteria, a summary of the molecular data collected, a list of related publications, and a list of relevant phenotypes selected by the PI;
- links to Authorized Access, description of data use limitations and use restrictions, release date, embargo release date, and a list of users and their public and technical research use statements who have been authorized to access individual-level data;
- links to publicly available information – including a study manifest -- via a public ftp site;
- links to other related NCBI resources (e.g. BioSample, SRA, BioProject, MeSH);

Variable

- the variable's name, accession, description, comments;
- a statistical summary of the variable's values;
- a curated list of excerpts from study documents that relate to the variable

Document

- the document's name and accession;
- the document's contents in HTML format; note that the red question marks link particular excerpts of the document to other study objects. For example, clicking on a red question mark near a protocol description might list the phenotype variables that were measured using that protocol;
- a link to a PDF version of the document

Analysis

- the analysis' name, accession, description, and a brief synopsis of the methods used;
- relevant summary plots (*e.g.* Manhattan plots of p-values; Log QQ p-value plot);
- a link to the Genome Browser, where analysis results can be examined in greater detail.

Dataset

- the dataset's name, accession, and description;
- the dataset's release date and embargo release date;
- list of variables contained in the dataset;
- links to summary report and data dictionary

Searching dbGaP

All publicly released dbGaP studies can be queried from the search box on the top of the dbGaP homepage. Queries can be very simple, just keywords of interest (“cancer”), or complex, making use of search fields and Boolean operators (“cholesterol[variable] AND phs000001”). More complex searches can be facilitated by using the “Advanced Search” which helps create queries via a web form.

There are many search fields available in dbGaP. Table 1 shows a selection of the most commonly used fields, explains what they search for, and gives an example of how the search would be formed.

Additionally, complex queries can also contain Boolean operators. For example:

Cancer[Disease] AND True[Study Has SRA Components]

returns a list of all studies having SRA data and where the PI has assigned the keyword “cancer” as a disease term.

The screenshot shows the dbGaP search results for the query "diabetes". The search interface includes a search bar, filter options, and tabs for different types of study results.

Search Results: 1 to 20 of 148

Search results: 108703 Variables, 299 Analyses, 906 Documents, and 277 Datasets in 148 Studies

Studies (148) | **Variables (108703)** | **Study Documents (906)** | **Analyses (299)** | **Datasets (277)**

Filter your results:

- All (148)
 - dbgap_type_studies (148)
 - Studies having SRA data (27)
 - SHARE project (6)

Find related data:

Database: Select

Recent activity

- Turn Off Clear
- diabetes AND 1[s_discriminator] (148) dbGaP
- The Entrez Search and Retrieval System Bookshelf
- Querying and Linking the Data - The NCBI Handbook Bookshelf
- The NCBI Handbook Bookshelf

Study	Embarго Release	Details	Participants	Type Of Study	Links	Platform
phs000256.v3.p2 The Vaginal Microbiome: Disease, Genetics and the Environment	Versions 1-2: passed embargo Version 3:	V D A S	3474	Twin, Clinical Cohort		454 GS FLX Titanium 454 GS FLX Titanium
phs000293.v1.p1 The Familial Intracranial Aneurysm Linkage Study (FIA)	Version 1: 2014-07-26	V D A S	2507	Family Linkage	Links	Infinium II HumanLinkage-12
phs000546.v1.p1 NHLBI GO-ESP; Heart Cohorts Exome Sequencing Project (ISGS)	Version 1: 2013-10-10	V D A S	75	Case-Control	Links	HiSeq 2000

Figure 3. This shows the returns for a simple search on the word "diabetes". Note that specific results for Studies, Variables, Study Documents, Analyses, and Datasets can be accessed by choosing the appropriate tab.

As with all other NCBI resources, the searches in dbGaP are performed using the Entrez search and retrieval system. Please see the [Entrez chapter](#) of the NCBI handbook for general guidance on forming Entrez queries.

Once a search query is executed and results returned (Figure 3), clicking on an item's name or accession will lead to a page listing more specific information about that object. This information is of particular importance to those users who want to find out more about a study before deciding whether or not to apply for Authorized Access. (Note that on each of the different pages, one can examine other objects in the study by using the navigational aid along the right-hand edge of the page.)

Table 1. This table lists fields in the dbGaP advanced search that are likely to be useful to most searchers.

Search Field Name	Purpose	Example	Interpretation
Disease	Find all studies where the PI has assigned the indicated disease keyword	hypertension[Disease]	Find all studies where the PI has assigned the keyword "hypertension" as a disease term

Table 1. continues on next page...

Table 1. continued from previous page.

Search Field Name	Purpose	Example	Interpretation
Genotype Platform	Find all studies that use the indicated genotype platform	HumanOmni1_Quad_v1-0_B[Genotype Platform]	Find all studies that use the HumanOmni1_Quad_v1-0_B genotype platform.
Project	Find all studies that are associated to the indicated project	eMERGE[Project]	Find all studies that are associated to the eMERGE project.
Attribution	Find all studies that have the indicated keywords within the attribution section of a study	Johnson[Attribution]	Find all studies where “Johnson” is listed somewhere in the attribution.
Analysis	Find all analyses where the keyword is contained in an the analysis’s title or description	cancer[Analysis]	Find all analyses where the keyword “cancer” appears in the title or description.
Study Has SRA Components	Find all studies having SRA data or that are scheduled to have SRA data	True[Study Has SRA Components]	Find all studies having SRA data or scheduled to have SRA data.
Variable	Find all variables where the keyword is contained in the variable’s	diabetes[Variable]	Find all variables where the keyword “diabetes” appears in the name or description.

Table 1. continues on next page...

Table 1. continued from previous page.

Search Field Name	Purpose	Example	Interpretation
	name or description		
Dataset	Find all datasets where the indicated keyword is contained in the dataset's name or description	visit[Dataset]	Find all datasets where the keyword "visit" appears in the name or description.
Document	Find all documents where the indicated keywords appear in the document's name or content	protocol[Document]	Find all documents where the keyword "protocol" appears in the content.
Study	Find all studies where the indicated keywords appear somewhere within the study.	glaucoma[Study]	Find all studies where the keyword "glaucoma" appears in at least one object associated to that study.

Variables on the public website

Phenotype variables can either be found by doing a search on the dbGaP home page, and then linking to an individual variable page (see Figure 4 and Figure 5 for examples), or they can be found by choosing the “Variables” tab if you are already looking at the website of a study. If you are using the “Variables” tab the phenotypes are generally grouped into broad categories for ease of browsing. These categories can be found to the right-hand side of any variable report web page (see Figure 4). Most studies use the following categories, as appropriate:

- Affection Status
- Sociodemography and Administration
- Medical History
- Physical Observations

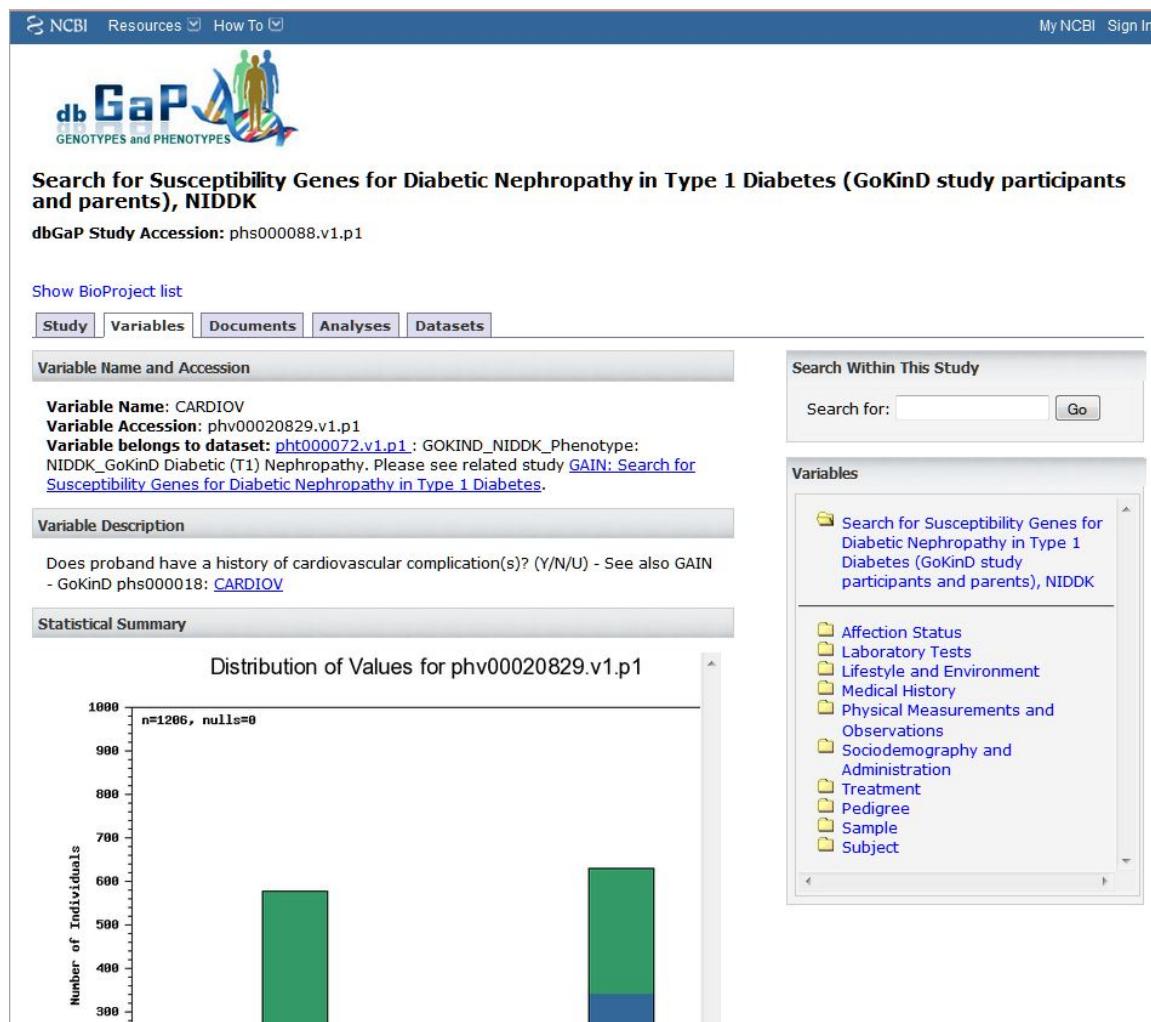


Figure 4. Top portion of the variable report for the variable phv00020829, CARDIOV. Variables can be browsed by category using the navigation to the right hand side of the page.

- Lab Measurements
- Psychological and Psychiatric Observations
- Lifestyle and Environment
- Treatment

Exceptions to this grouping method are found in large studies such as the [Framingham Cohort](#) which have their own long-standing system for grouping data. When searching for variables in large studies, or if you have a very specific query, it can be more efficient to search for variables using the search box on the right side of the variable report page (or from the dbGaP home page if you want to perform a cross-study search), rather than attempting to browse through the hierarchy of folders.

Document Parts Related to Variable

- **Document Name:** GoKinD Study Diabetic Offspring
 - [See document part in context](#)

4. IF YES, PLEASE INDICATE YEAR

Have you ever had a heart attack?

NO YES

YEAR

Have you ever been hospitalized due to a heart attack?

- **Document Name:** Medical History and Physical Examination
 - [See document part in context](#)

4. CARDIOVASCULAR

Does the proband/relative have a history of any of the following?

a. History of Hypertension (defined as systolic \geq 140 or diastolic \geq 90)

No Yes

b. Angina

- **Document Name:** GoKinD GW Clinics Derived variables SAS code documentation
 - [See document part in context](#)

```
/* -----
/* CARDIOV - calculate cardiovascular complications */
/* for GW clinic offspring */
```

Figure 5. Top portion of the variable report for the variable phv000200829, CARDIOV. This shows how variables are linked to appropriate sections of documents. If you follow the first link, you will be taken to the portion of the document show in Figure 6.

Documents on the public website

There are multiple pathways to find documents through the dbGaP web site. On the dbGaP [home page](#) the newest studies are listed under the “Latest Studies” heading, with the most direct route to documents being the orange “D” icons. A gray icon means there

are no documents associated with the study. Beneath that section, the "List Top Level Studies" link leads to a searchable listing of all studies and documents, with an advanced search option available for building document-specific queries. On a study page, clicking the Documents tab will open the study's default document, with a folder tree on the right to explore the rest, and a "Search Within This Study" box that will search document text (Figure 5). Variable pages may also link to documents in which they are annotated, through the "See document part in context" links. Documents for each study are also available on the dbGaP ftp site as a downloadable zip file, which includes the pdfs, xml, and images. The ftp site is accessible from study pages by clicking the link under "Publicly Available Data."

Using document annotation

As noted previously, curators establish connections between appropriate portions of text and other accessioned objects. As an example of the types of functionality that annotations can provide, imagine looking at a variable report page having to do with whether subjects take a multivitamin. If you scroll down to the bottom of the variable page, there is a section labeled "Document parts related to the variable" which is shown in Figure 5.

This shows that there are two documents, a Coding Manual and an Annotated Form, that have text which has been associated to the multivitamin variable. If you click on the "See document part in context" link you will be taken to the appropriate portion of the document. If you follow the link for the Annotated form, you would get taken to a page that looks like Figure 6.

In the image of the questionnaire shown in Figure 6, the icon of a red circle with a white question mark, ②, indicates that a section of the document is associated with one or more accessioned objects in a study. The accessioned objects are generally variables, but can include data tables. Clicking on the icon will either take the user to a variable report page (if only a single variable is associated with the icon) or to an Entrez search result page (if there are multiple objects associated with that icon).

dbGaP ftp site

The [ftp site](#) includes a directory for every study, which contains a directory for every version of a study, as well as a directory where analyses are found. Currently, each version of a study contains directories for documents, phenotype variable summaries, manifests, and release notes (Figure 7). Manifests describe the files available in each consent category while release notes describe the history of the released files as well as giving details of any changes made from previous versions.

Please note that older versions of studies may have a different directory structure although they contain similar information.

② 4. IF YES, PLEASE INDICATE YEAR

Have you ever had a heart attack?

NO YES

YEAR

Have you ever been hospitalized due to a heart attack?

NO YES

YEAR

Have you ever had coronary bypass surgery?

NO YES

YEAR

Have you ever had angioplasty?

NO YES

YEAR

Have you ever had a stroke or TIA (transient ischemic attack)?

NO YES

YEAR

② 5. Has a doctor ever said that you have retinopathy or eye problems related to diabetes?

NO YES

IF YES, please specify and indicate year of first diagnosis or treatment (if applicable):

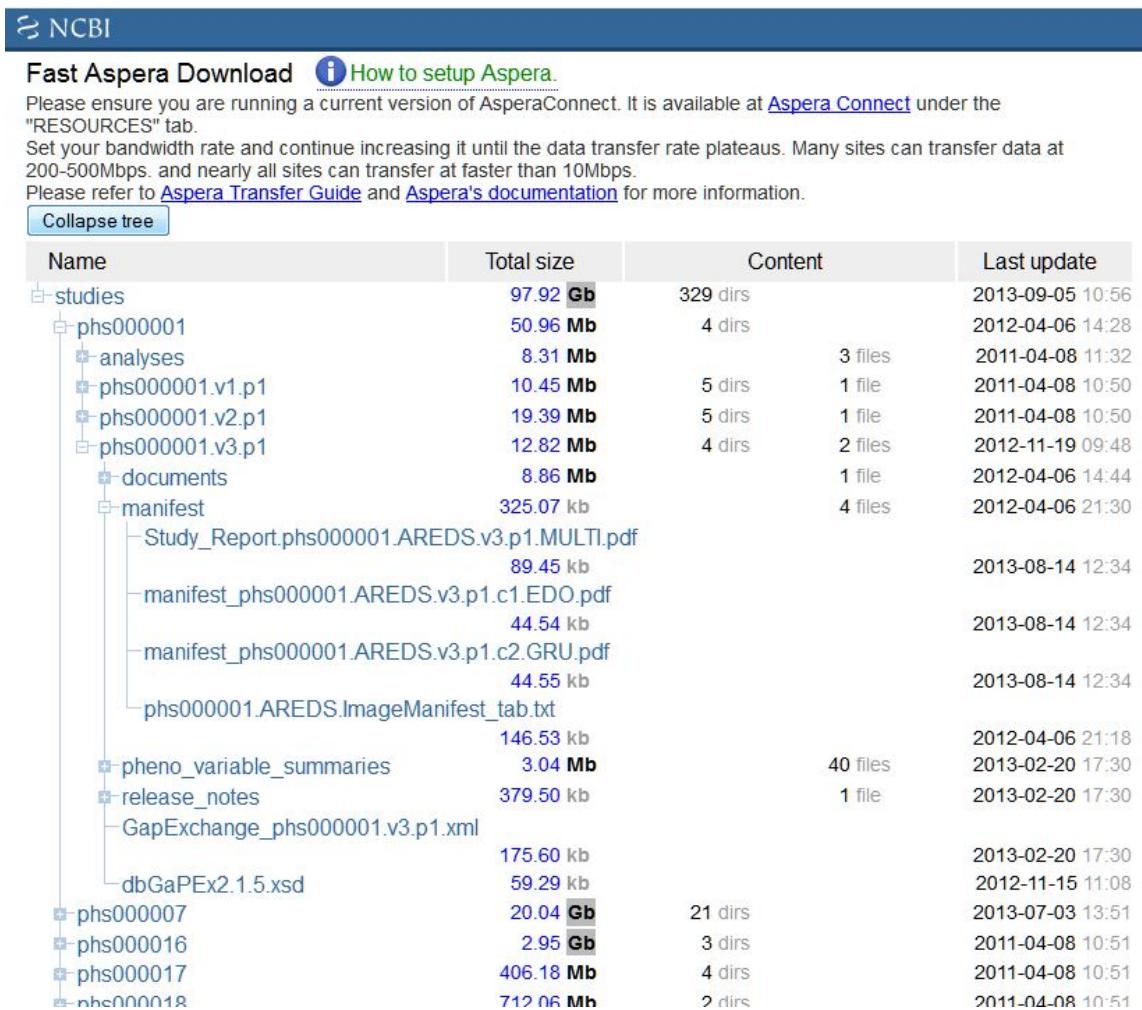
DIAGNOSIS	Yes/No	Year
-----------	--------	------

Non-Proliferative Retinopathy

Yes

Figure 6. This show the portion of the web page for the document “GoKinD Study Diabetic Offspring” (phd000152.2) which you would be taken to if you clicked the first link from Figure 5.

The variable summaries and the data dictionaries are delivered as XML files with an accompanying XSL file which produces the HTML rendering of the file that can be viewed on a browser.



The screenshot shows the NCBI Aspera Download interface. At the top, there's a header with the NCBI logo and a link to "How to setup Aspera". Below the header, a message says to ensure you are running a current version of AsperaConnect, available at [Aspera Connect](#). It also notes that bandwidth rate should be set and increased until it plateaus, with many sites capable of 200-500Mbps. A note also refers to the [Aspera Transfer Guide](#) and [Aspera's documentation](#).

A "Collapse tree" button is visible. The main area is a table with columns: Name, Total size, Content, and Last update. The table lists files and directories under the "studies" root. The "studies" directory has a total size of 97.92 Gb and contains 329 dirs. One of its sub-directories, "phs000001", is expanded, showing its contents:

Name	Total size	Content	Last update
studies	97.92 Gb	329 dirs	2013-09-05 10:56
phs000001	50.96 Mb	4 dirs	2012-04-06 14:28
analyses	8.31 Mb	3 files	2011-04-08 11:32
phs000001.v1.p1	10.45 Mb	5 dirs	2011-04-08 10:50
phs000001.v2.p1	19.39 Mb	5 dirs	2011-04-08 10:50
phs000001.v3.p1	12.82 Mb	4 dirs	2012-11-19 09:48
documents	8.86 Mb	1 file	2012-04-06 14:44
manifest	325.07 kb	4 files	2012-04-06 21:30
Study_Report.phs000001.AREDS.v3.p1.MULTI.pdf	89.45 kb		2013-08-14 12:34
manifest_phs000001.AREDS.v3.p1.c1.EDO.pdf	44.54 kb		2013-08-14 12:34
manifest_phs000001.AREDS.v3.p1.c2.GRU.pdf	44.55 kb		2013-08-14 12:34
phs000001.AREDS.ImageManifest_tab.txt	146.53 kb		2012-04-06 21:18
pheno_variable_summaries	3.04 Mb	40 files	2013-02-20 17:30
release_notes	379.50 kb	1 file	2013-02-20 17:30
GapExchange_phs000001.v3.p1.xml	175.60 kb		2013-02-20 17:30
dbGaPEx2.1.5.xsd	59.29 kb		2012-11-15 11:08
phs000007	20.04 Gb	21 dirs	2013-07-03 13:51
phs000016	2.95 Gb	3 dirs	2011-04-08 10:51
phs000017	406.18 Mb	4 dirs	2011-04-08 10:51
phs000018	712.06 Mb	2 dirs	2011-04-08 10:51

Figure 7. This figure shows the basic organization of the ftp site, with the study phs000001.v3.p1 opened up to show portions of the hierarchy of directories and files.

The documents directory contains at least one .zip file that holds the xml files, images, and pdf versions of the documents in a study. In cases where there are a large number of documents the files may be separated into separate .zip files for xml, images, and pdf.

dbGaP Authorized Access

Data distribution by dbGaP is governed by the NIH's policies and procedures for managing Genome Wide Association Study (GWAS) data. Information related to these policies can be found on the [NIH GWAS website](#). Questions related to GWAS policy can be directed to GWAS@mail.nih.gov.

The individual level data is only available to authorized users. Requests for data and data downloads are managed through the dbGaP [Authorized Access System](#) (dbGaP-AA), a web platform that handles request submission, manages reviewing and approval processes

carried out by signing officials (SOs) and Data Access Committees (DACs), and facilitates secured, high speed downloads of large data sets for approved users.

The dbGaP data are organized and distributed by consent groups. That is, the data are grouped by subjects that have agreed to the same set of data use limitations. The data can only be selected by consent group when making data access requests. There are no overlapping subjects between the consent groups within a study. The data requests are also reviewed and approved by consent group. Therefore it is very important that requesters understand the Data Use Limitations of consent groups before they apply for dbGaP data access.

Each data file distributed through the dbGaP has an embargo release date. The data access policy requires that the results obtained from analyzing the dbGaP data are not published before the embargo release date.

To access the Authorized Access system, non-NIH users must have an [NIH eRA Commons](#) account with a Principal Investigator (PI) role. The login username and password of a user's dbGaP-AA account are the same as those of a user's eRA account. NIH users need to be registered in the dbGaP system by the GWAS Project Administrator (GPA) of an affiliated institute before gaining access to the dbGaP-AA. After being registered, the NIH user can login to the dbGaP-AA account using the login username and password of their NIH CIT (or email) account.

A data access request (DAR) is made by filling out forms inside dbGaP-AA. The request includes a Research Use Statement and a Non-technical Summary. The DAR must also designate an institutional Signing Official and IT director for their project. If any of requested datasets has an IRB (Institutional Review Board) approval requirement, an IRB approval document should be uploaded to the system before submitting the request. By signing the application form, the data requester agrees to obey terms and conditions laid out in the governing Data Use Certification (DUC) document.

The DAR will first be reviewed by the SO. If approved, it will be passed on to the appropriate Data Access Committee or committees. A DAC is a committee appointed by an NIH institute (or group of institutes) which evaluates DARs requesting access to studies from their portfolio. Each DAC evaluates whether requests conform to NIH policies and procedures including whether the proposed research is consistent with the Data Use Limitations stipulated for each study. If approved, the requester must agree to obey data use restrictions dictated by participant informed consent agreements and to comply with data use, sharing, and security policies laid out in a governing DUC. At that point the data can be downloaded by the requester.

The dbGaP system manages data downloads using Aspera, a system designed to expedite high-speed data transfers. Use of Aspera requires that Aspera Connect, a browser plugin available through the Aspera website, is installed on the downloading machine. Data download can be carried out through either Aspera Connect's web-interface or by using Aspera ASCP on the command line. For SRA (Sequence Read Archive) data distributed

through the dbGaP data download can be done directly through the sra-toolkit, which allows transfer based on http protocols. Detailed information about sra-toolkit can be found from the [SRA toolkit documentation](#).

All data downloaded from the dbGaP are encrypted. The downloaded data, with the exception of SRA data, need to be decrypted before being used. For SRA data, we suggest that users work directly with the data dump utilities that are available through the NCBI sra-toolkit without decryption. The NCBI decryption tools and sra-toolkit are available from the [SRA software download site](#).

Approved data users are required to submit an annual project progress report to all the DACs from which they received approval. A project close-out request should be filed if the project is finished. Most dbGaP data requests have a one year approval period. To renew a project the PI needs to revise and resubmit the DAR, as well as submit the annual report. The resubmitted project will go through the SO and DAC review process again. During this process, only expired data requests under the project will be re-reviewed. Previously approved data requests that have not expired will remain approved.

A data request is not transferrable. If a PI leaves the institution listed in the DAR, all the dbGaP requests sponsored by the institution should be closed out. As a part of the close-out process, all data downloaded through the project need to be destroyed and the process has to be confirmed by the IT director and SO. The PI will need to reapply for the data once they have settled at their new location.

Related Tools

Phenotype-Genotype Integrator (PheGenI)

Scope

The [Phenotype-Genotype Integrator](#) (PheGenI), (4) merges NHGRI genome-wide association study (GWAS) catalog data with several databases including [Gene](#), [dbGaP](#), [OMIM](#), [GTEx](#) and [dbSNP](#). This phenotype-oriented resource, intended for clinicians and epidemiologists interested in following up results from GWAS, can facilitate identification and ranking of variants that may warrant additional study.

History

PheGenI was first released in 2011. The major functionality has not changed, *i.e.* modes of search and categories of display, but functions have been added to improve both queries and data processing. For example, an autocomplete function was added to facilitate the phenotype queries, and download functions were added to the ideogram and tabular results sections. PheGenI is under active development, with contents and displays scheduled to be more closely integrated with additional web resources.

Data Flow

PhenGenI is populated automatically via feeds from NHGRI, dbSNP, dbGaP and NCBI's genome annotation pipeline. Please note that PheGenI does not display all p-values from each dbGaP-hosted analysis. Specifically, only p-values $<10^{-4}$, and/or the lowest 100 p-values are included for each analysis. Currently, the phenotype search terms are based on MeSH, but will be enhanced with additional options in the future.

Access

Users can search based on chromosomal location, gene, SNP, or phenotype and then view and download results. Association results can be filtered by p-value, and genotype data can be filtered by location of variant site relative to gene annotation. The results are separated into several categories, including association results, genes, SNPs, eQTL data, a dynamic genome view and dbGaP studies. Each section provides a download function.

As a tool to find data in dbGaP, the view of all analysis results is accessed by clicking on the dbGaP link in the source column of the Association Results table. For full analysis and aggregate statistics such as allele frequencies, apply for controlled access.

Gene's Phenotypes section also provides links to PheGenI, via the anchor "Review eQTL and phenotype association data in this region using PheGenI".

The dbGaP Genome Browser

The genome wide association results hosted at the dbGaP are displayed through the dbGaP genome browser, where they can be viewed along the human genome.

The dbGaP genome browser can be accessed through the analysis page of a given dbGaP study. For example, under the "Analyses" tab of the dbGaP study [phs000585.v1.p1](#). If there are multiple analyses, you can select one from the right panel. The link named [View association results in Genome Browser](#) leads to the chromosomal viewer and each region (block) there contains results from all tested loci within (Figure 8). The color is coded for the smallest p-value in that block.

The genome browsing page (Figure 9) is opened by clicking on the region. The testing results are tabularized in the middle. The genome track on the top allows zooming in to see more detailed genomic location and linkage disequilibrium structure. GWAS Catalog (NHGRI) data, or an added analysis, can be aligned with the current track under the same coordinates, which allows viewers to compare results from different studies. The sequencing view (bottom) shows genome annotations (gene, transcript and protein) at that region. Each object in this page is linked to its annotated database, which helps scientists to study biological function behind the genetic variations.

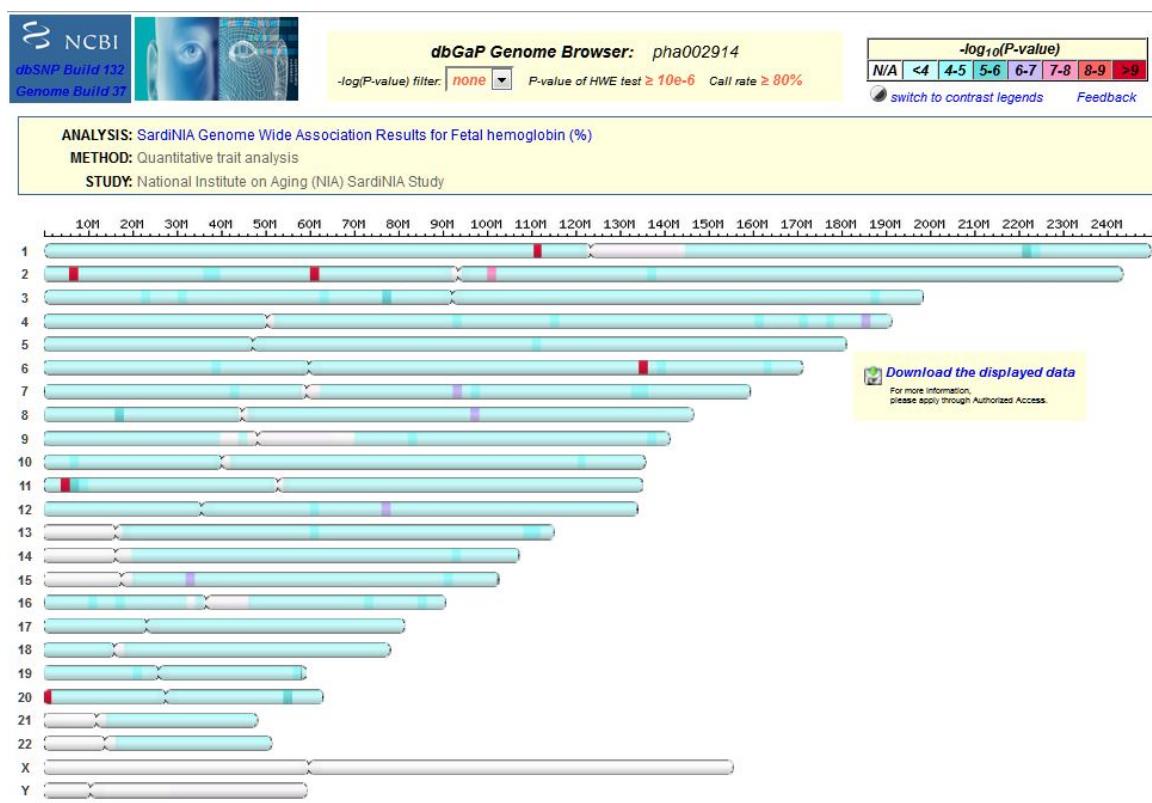


Figure 8. Genome Browser showing pha002914.

Focus on Specific Genomic Regions through Browser

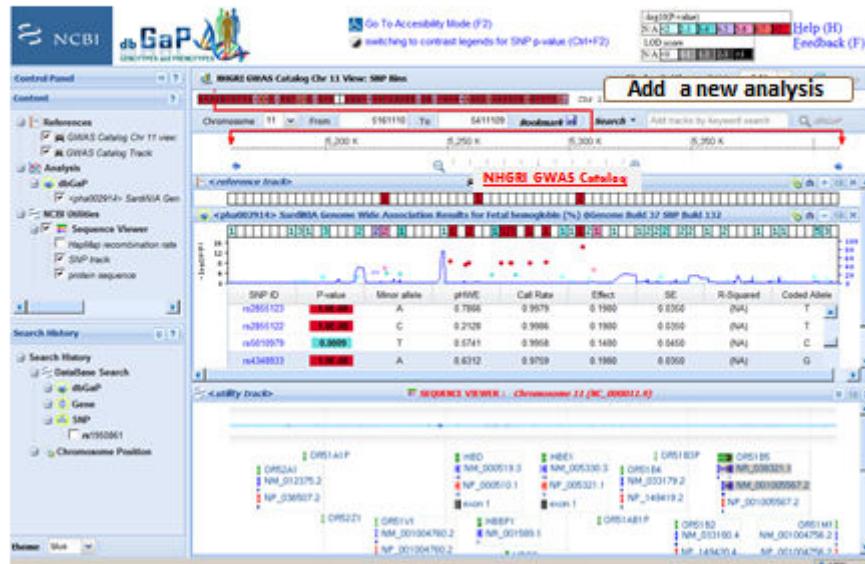


Figure 9.

References

1. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181–6. PubMed PMID: 17898773.
2. GAIN Collaborative Research Group. Manolio TA, Rodriguez LL, Brooks L, Abecasis G; Collaborative Association Study of Psoriasis, Ballinger D, Daly M, Donnelly P, Faraone SV; International Multi-Center ADHD Genetics Project, Frazer K, Gabriel S, Gejman P; Molecular Genetics of Schizophrenia Collaboration, Guttmacher A, Harris EL, Insel T, Kelsoe JR; Bipolar Genome Study, Lander E, McCowin N, Mailman MD, Nabel E, Ostell J, Pugh E, Sherry S, Sullivan PF; Major Depression Stage 1 Genomewide Association in Population-Based Samples Study, Thompson JF, Warram J; Genetics of Kidneys in Diabetes (GoKinD) Study, Wholley D, Milos PM, Collins FS. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007;39(9):1045–51. PubMed PMID: 17728769.
3. PLINK.: <http://pngu.mgh.harvard.edu/~purcell/plink/>
4. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, Feolo M, Hindorff LA. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet*. 2013. PubMed PMID: 23695286.

Appendix – Phenotype Quality Control

Each submitted dataset should have a corresponding data dictionary with information describing the variables and their values. Additionally, dbGaP requires the following three special datasets to be submitted:

1. Subject Consent
2. Subject Sample Mapping
3. Pedigree

After receiving the data submissions, QC scripts are executed to check the files for potential errors. The results are manually checked and errors are reported to submitters for clarification or resubmission of data. There are usually a few iterations prior to the study being able to be loaded.

The format of the datasets

Each dataset submitted to dbGaP should be a rectangular table showing variables in columns and subject or sample IDs in rows. Each dataset should be a single tab-delimited plain text file. Microsoft Excel files are also accepted, but are converted to tab-delimited plain text files for processing. Once the files pass all the qc checks, they are loaded into dbGaP databases and distributed as tab-delimited plain text files to approved Authorized Access users. The following formatting requirements will be ensured by running QC scripts:

1. Each column has a unique, non-blank column header (variable name).
2. Each dataset has a subject (or sample) ID column.
3. Each row has a subject (or sample) ID value.
4. There are no duplicated rows in the table.
5. Datasets do not include any characters that will not be rendered correctly on the web pages.
6. Duplicated subject IDs in different rows are acceptable, but will be reported in the qc checks so that curators can manually verify that there are no errors.
7. Variables without any values (with only column header) are acceptable but will be reported in the qc checks.

Subject and Sample IDs

A dbGaP phenotypic dataset is a collection of variable values of individuals (subjects) or samples of individuals. Each subject should be submitted with a distinct subject ID; each sample should be submitted with a distinct sample ID. Submitters can use multiple subject alias IDs for a single subject. In addition to the submitted subject ID dbGaP assigns a single dbGaP subject ID for each submitted subject (individual person), even if the subject has multiple alias IDs. The dbGaP subject ID different from the submitted subject ID. Submitters are required to use the subject consent file to list all the subjects who participated in, or referred to by, the study, with their consent values. Subjects who did

not participate in the study, namely HapMap controls and parents of participants included in the pedigree file, should also be included in the consent file with consent value 0.

The subject and sample QC scripts check the following:

1. Each subject, who might be represented by multiple aliases, in the consent file has exactly one consent value (an integer).
2. Each sample, which might be represented by multiple aliases, in the sample mapping file maps to exactly one subject in the subject consent file.
3. All subjects in all the datasets (including subjects that have molecular data only and no phenotype data and relatives in the pedigree file) are included in the subject consent file. Additional subjects who are in the subject consent file, but are not found in any of the phenotype datasets are flagged and reported, but are not considered an error if the subject's data will be submitted at a later time.
4. Samples that are not found in the molecular data, but are found in the subject sample mapping file will be flagged and reported, but are not considered an error if the sample will be submitted at a later time.
5. Multiple alternate names (aliases) for a single subject within a single or across multiple studies is assigned only one dbGaP subject ID. If the subject does not have a dbGaP subject ID, a unique ID will be assigned to the subject when the dataset is loaded into the database. Currently, a single dbGaP subject ID is assigned to a Subject only when the submitter provides the linking information. This is true at the sample level as well. The case of alternate names for samples should be less common than subjects, since dbGaP considers sample IDs to refer to the final analyte (DNA/RNA) that is put in the machine or on the chip, rather than the physical result of obtaining a tissue or blood sample, though this is accepted as well.
6. If the gender of a subject is reported in multiple places (different datasets or different rows of the same dataset) the gender values should be the same.
7. If a subject, or an alias of this subject, is already found in the dbGaP database, the gender of the subject in the dataset should be the same as that in the database.
8. If a sample, or an alias of this sample, is already found in the dbGaP database, the sample in the dataset should map to the same subject as in the database.
9. Each subject within a single study should not have conflicting case-control status, especially in the scenario when the same case control variable appears in multiple datasets.
10. The number of subjects and samples are consistent between iterative submissions. When the counts are different, they are reported to the submitter for confirmation or resubmission.

HIPAA violations

The datasets submitted to dbGaP should follow the Health Insurance Portability and Accountability Act (HIPAA) rules in order to protect the privacy of personally identifiable health information. Due to the complexity of HIPAA rules, it is impossible to write a

program to report all HIPAA violations without turning up false positives. It is also impractical to manually check all the data values and find all the HIPAA violations. QC scripts have been created to check variable names, descriptions, and values, and to flag variables that are likely to have sensitive information. dbGaP curators then manually check the flagged variables to determine whether these are HIPAA violations. The QC scripts first report all variables whose names or descriptions contains the following key words (case insensitive except for ‘IP’ and ‘DOB’):

1. name
2. address
3. zip
4. phone
5. telephone
6. fax
7. mail
8. email
9. social
10. ssn
11. ss#
12. birth
13. DOB
14. license
15. account
16. certificate
17. vehicle
18. url
19. IP

Only the names or descriptions containing a whole-word match with at least one of the above key words are reported. A word in the variable name or description that contains a key word as a substring is not considered a match. For example, “Leave your email/phone.” is reported as a match since it contains key words “email” and “phone”, but “zipper” is not reported because it only contains key word “zip” as a substring.

In many cases the variable names or descriptions do not have any indication that the variable might have HIPAA incompatible information. To work around this, the QC scripts also check variable data values for sensitive information. Data values are much harder to check than variable names and descriptions due to the sheer number of individual values and the great variety of errors. Fortunately, almost all of the HIPAA violations in the datasets submitted to dbGaP database are related to dates, including dates as separated values and dates embedded in longer texts. Below are some of the examples of the dates found in the datasets submitted to dbGaP:

- 11-JUN-1970
- 01-SEP-65

- SEPT-NOV 1995
- 2004.05.10
- 2/2/85
- 8-11-83
- 10/1974
- JAN '93
- 3/04
- SEPT 85
- NOV-89-
- FEB64
- NOVEMBER 1992
- "DEC" "92"
- Feb.4
- Jan – 1996
- March, 2004
- (Nov,2005)
- APR. "85
- APRIL 91
- apr 2000
- (APRIL 1997)
- (3/00)
- DEC.1992
- 1998-May
- October-September, 2004
- Jan. 1
- May 3rd
- xxxxxIN2002.03.01
- 19941122
- 112004

Most of the above values, e.g., "01-SEP-65", "2004.05.10", "DEC.11992", are obviously date values and not HIPAA compatible. It is hard to write programs to find dates in all the different formats without generating too many false positives. However, some of them are not so obvious and need manual confirmation using variable descriptions and context of the values. For example, "3/04" could mean "March 2004", or "3 out of 4"; 19941122 could be "Nov. 22, 1994" or the number 19941122; "112004" could be "Nov. 2004", "Nov. 20, 2004" or the number 112004. If we report all the 6-digit numbers as potential date values, we will generate a great amount of false positives. More complicated algorithms are needed to detect date values with high sensitivity without sacrificing too much specificity. We use the following algorithm to detect the date values in the datasets:

1. Two 1 or 2-digit numbers and a 2 or 4-digit number, in this order, separated by "/", "-" or ":", e.g., "3/5/1994" or "12-28-03".
2. One 4-digit number and two 1 or 2-digit numbers separated by "/", "-" or ":", e.g., "1994.2.13".

3. A 1 or 2-digit number and a 4-digit number starting with 19 or 20 separated by “/”, e.g., “10/1994” (but not “10.1994”).
4. A 1 or 2-digit number followed by a “/” and a 2-digit number starting with 0, e.g., “3/04” (but not “3/94”).
5. A month name or short name and a 1, 2, or 4-digit number, in either order, separated by some non-letter, non-number characters or not separated, e.g., “JAN ‘93”, “FEB64”, “May 3rd” (but not “may be 14”). An example of a false positive is “4 (may be under reporting)”.
6. A 6-digit number is considered to be a potential date value if its first four digits make a valid date in mmdd format (i.e., first two digits read as month second two as day of the month). For example, 122876 is considered to be a potential date value since 1128 is a valid date (Nov. 28) in mmdd format; 231208 or 113198 is not a potential date since 2312 or 1131 is not a valid date in month/day format. If all of the values, or first 10 values, of a variable are 6-digit potential dates, this variable together with its potential date values will be reported by the scripts.
7. An 8-digit number is considered to be a potential date value if it makes a valid date in the 20th or 21st century in either mmddyyyy or yyyyymmdd format. For example, 19940822 is considered to be a potential date since it can be read as 1994/08/22 (Aug. 22, 1994). 10312005 is a potential date value since it can be read as 10/31/2005 (Oct. 31, 2005). “19080230” is not considered to be a potential date since neither 1908/02/30 nor 19/08/0230 is a valid date in the 20th or 21st century. If all of the values or the first 10 values of a variable are 8-digit numbers of potential date values, the variable will be reported as containing potential HIPAA violations.

In addition, the QC scripts also report values that look like social security numbers (e.g., “123-45-6789” or “123456789”), phone numbers (e.g., “321-456-7890” or “(301)456-7890”), zip codes (e.g., “MD 20892”), etc. A few cases of this kind of information have been detected by the QC scripts. However, other cases like names of people are not found by the QC scripts, but by human curation.

Extreme values that might be used to identify individual participants (ages over 90, extremely heavy body weights, families with extraordinary large numbers of children) are also HIPAA violations. The QC scripts infer age variables from variable names, descriptions, and units and report ages over 89 as potential HIPAA violations. For other extreme values, since the HIPAA rules don’t specify particular cut-off values, we check the value distribution curves by hand and decide whether we need to hide the extreme values on a case-by-case basis.

Data dictionaries

Data dictionaries are required to be submitted along with every dataset, to explain the meanings of the variables and data values. For each value in the datasets, dbGaP requires that the submitters provide a variable description, variable type, units of values for numerical variables, as well as logical minimum and maximum values if available. For

each encoded value, a code meaning should be included in the data dictionary. Since the data dictionaries submitted to dbGaP vary in format, many of which are not quite machine-readable, curators spend a good deal of time understanding the data dictionaries, correcting errors, and making other modifications so that they can be read by computer programs. Then QC scripts are executed to compare the data dictionaries with the corresponding datasets. The QC scripts report variables in the datasets that are missing required information (such as descriptions) in the data dictionary, as well as variables described in the data dictionaries but not found in the datasets. Many of these mismatches are caused by typos, such as “0” for “O” and vice-versa. A number of the numerical variables submitted to dbGaP are missing units. Often the units are implied in a variable’s description or in other documents. The QC scripts try to add the units back by checking the variable descriptions, which is then verified by manual curation.

In addition to missing descriptions and units, many datasets submitted to dbGaP have missing, or incomplete, code/value pairs. Some of these errors are easy to detect, e.g., the values of a categorical variable are encoded by integers and all of the code meanings are provided except for one code. However, if the variables contain both numerical values and numerically encoded categorical values, the errors of missing code meanings are hard to detect. Usually in this case, the submitters would use numbers beyond logical value range to encode for non-numerical meanings. For example, if the variable is age of patient, they would use code values like -1, 999 for meanings like “N/A” or “unknown”. If the submitters provided logical minimum and maximum values to us, it would be easy for us to find all the missing code meanings. However, in most of the cases the logical minimum and maximum values are either not available or incorrect. Unreasonable or suspicious values found automatically or manually are reported to submitters to clarify and correct.

QC scripts are executed to compare each submitted dataset to its corresponding data dictionary. The QC scripts report the following errors or potential errors:

1. Variables missing descriptions in data dictionary.
2. Variables with descriptions in data dictionary but not found in dataset (usually due to manual typos in variable names).
3. Potential errors or missing information in value code meanings.

The following algorithm is used to detect potential errors in the code meanings of each variable:

1. If the variable is labeled as code-value type by the submitter, report all values in dataset without code meanings in data dictionary.
2. If all the values are numbers,
 - i. Report extreme values beyond $5 \times SD$ as potential encoded values. Exclude 5 largest numbers when calculating SD.
 - ii. Report rare negative numbers as potential encoded values. A negative number is considered to be rare if only 1 or 2 out of total more than 10 distinct values, or less than 1% of the distinct values are negative numbers.
3. If all the values are non-number texts,

- i. Report all the values without code meanings in the data dictionary if more than half of the distinct values have code meanings.
 - ii. Report all the values in the dataset that differ from a code in the data dictionary only by case. For example, if the data dictionary includes code meaning “UNK=Unknown”, but the dataset has a value “Unk” instead of “UNK”, the scripts report the case mismatch.
4. If some of the values are numbers but some are non-numbers, separate the values into a set of numerical values and a set of text values, then report the potential encoded values using the above rules.
 5. Report all the code values in data dictionary but not used in the dataset.

Again there is a trade-off between sensitivity and specificity. The QC scripts allow the curator to set some parameters like cut-off number of SDs to adjust the sensitivity and specificity. For example, if too many real extreme values are reported as potential encoded values, i.e., the false positive rate is high, we can set the parameter to let the QC scripts report only the extreme values beyond $6\times\text{SD}$ or more.

Pedigree file

If there are related individuals in the study, a pedigree file should be submitted to dbGaP. Pedigrees can be quite complex depending on the number of vertical and horizontal relationships, however all relationships can be summarized using the following five required columns: family ID, subject ID, father ID, mother ID, and sex. dbGaP also collects twin IDs, where these IDs can be expanded to include multiples. An additional column can be included to differentiate monozygotic and dizygotic twins, and twin ID if available. All subjects who appear in the father ID or the mother ID columns should also be included in the subject ID column. QC scripts were created to check the pedigree files and report the following errors or potential errors:

1. Any of the above required columns is missing.
2. Subject IDs appearing more than once in the subject ID column.
3. Father or mother IDs that are not found in the subject ID column.
4. Subjects missing family IDs.
5. Subjects missing sex values.
6. Male subjects shown in the mother ID column and female subjects as fathers.
7. Subjects with non-null but same father and mother IDs.
8. Subjects having children with their parents or grandparents.
9. Subjects having children with their sibling or half siblings.
10. Subjects having children with their uncle or aunts.
11. Subjects having children with their cousins (Usually these are not errors. We flag them out just to make sure the data is correct.)

The Database of Short Genetic Variation (dbSNP)

Adrienne Kitts, MS, Lon Phan, PhD, Minghong Ward, MS, and John Bradley Holmes, PhD

Created: June 30, 2013; Updated: April 3, 2014.

Scope

Sequence variation is of scientific interest to population geneticists, genetic mappers, and those investigating relationships among variation and phenotype. These variations can be of several types, from simple substitutions that do not affect sequence length, to those that result in minor length differences, to those that affect multiple genes and multiple chromosomes. Variations can also be categorized with respect to their frequency within a population, from a variation with a single allele to a variation that is highly polymorphic.

Although SNP is the abbreviation for “single nucleotide polymorphism,” dbSNP is a public archive of all short sequence variation, not just single nucleotide substitutions that occur frequently enough in a population to be termed polymorphic. dbSNP includes a broad collection of simple genetic variations such as single-base nucleotide substitutions, small-scale multi-base deletions or insertions, and microsatellite repeats. Data submitted to dbSNP can be from any organism, from any part of a genome, and can include genotype and allele frequency data if those data are available. dbSNP accepts submissions for all classes of simple sequence variation, and provides access to variations of germline or somatic origin that are clinically significant.

In order to emphasize the comprehensive nature of dbSNP’s content, the full name of the database was changed from “database of Single Nucleotide Polymorphism” to the more inclusive “database of Short Genetic Variation” in July of 2011. The acronym that represents the database will remain “dbSNP” to avoid any confusion that might arise from a complete name change.

Each record in dbSNP includes the sequence context of the variant, the frequency of the polymorphism in a population if available, its zygosity if available, and the experimental method(s), protocols, and conditions used to assay the variation by each submitter. Individual submissions are clustered into dbSNP reference records (rs#) that contain summary data which may include clinical significance from [ClinVar](#), association with phenotype from [dbGaP](#), variation false positive status, allele origin (germline or somatic), and submitter attributes.

The dbSNP has been designed to support submissions and research into a broad range of biological problems that include the identification of genotype-phenotype relationships, genetic and physical mapping, functional analysis, pharmacogenomics, and association studies.

Medical Genetics

Advances in next-generation sequencing technologies allow researchers to generate massive amounts of sequence data. When clinical samples are sequenced using these technologies, novel variants that have causative roles in disease may be identified. dbSNP's role is to manage information on the location and type of these novel variants, while ClinVar manages the current interpretation of the variants' clinical phenotype.

dbSNP integrates clinical attribute data from ClinVar (i.e., clinical assertions and asserted allele origin) into new and existing human refSNP records, as well as into additional curated attribute data that includes minor allele frequency and variation false positive status.

[VCF files](#) generated from dbSNP's curated records can be used to filter (subtract) known variants from a set of variant calls to identify novel variants or narrow a list of potential causative variants that might warrant further evaluation.

Genome Mapping

Variations are used as positional markers in genetic and physical mapping of nucleotide sequences when they map to a unique location in a genome. In other words, the variations represented by dbSNP records can serve as stable landmarks in the genome even if the variation is fixed for one allele in a sample. When multiple alleles are observed in a sample pedigree, pedigree members can be tested for variation genotypes as in traditional genetic mapping studies. To aid in such mapping efforts, dbSNP updates variation mapping and annotation for each organism with the release of every new genome assembly.

Molecular and Functional Consequences

Variations that occur in functional regions of genes or in conserved non-coding regions might affect transcription, post-transcriptional processing, or a protein product. dbSNP computes the molecular consequence of any sequence change, based on NCBI's annotation of the genome. Functional consequences may be reported from submitters.

Association Studies

dbSNP annotates variations with significant association to phenotype from Genome Wide Association Studies (GWAS) as reported by dbGAP and provides a detailed catalog of common variations. dbSNP's GWAS annotations and common variation catalog are used to inform the design of GWAS studies, the creation of variation arrays used in GWAS studies, and the interpretation of GWAS study results.

History

Creation and Growth

dbSNP was established in September, 1998, to address the need for a general catalog of genomic variation that would facilitate the scientific community's efforts in genetic association studies, gene mapping, and evolutionary biology. Initially, dbSNP was composed of small-scale locus specific submissions defined by flanking invariant sequence. Following the advent of high-throughput sequencing and the availability of complete genome assemblies for many organisms, however, dbSNP now receives a greater number of variants defined by sequence change at asserted locations on a reference sequence.

Evolution in Submitted Content

Because dbSNP was developed before a human reference assembly was available, initial submissions were primarily from human and defined a variant sequence in the context of flanking sequence. There was often little supporting evidence or validation data. As sequencing and other discovery technologies have changed, dbSNP has grown apace, and now includes data from over 300 organisms as well as ample validation data, including multiple independent submissions, frequency data, genotype data, and allele observations. To meet community needs for a centralized variation database, in the spring of 2008, dbSNP began accepting clinical assertions for new and existing variations as well as asserted locations for variation placement. The [ClinVar](#) database, now has the role of accepting variation clinical assertion data, and following its own accession process, will route novel variation positions to dbSNP for the assignment of ss (submitted SNP) and rs (refSNP) numbers.

In addition to integrating clinical assertions into refSNP records, dbSNP has introduced other curated attributes into refSNP records, such as minor allele frequency, asserted allele origin, and potential false positive status. It also uses the curated records to generate VCF files that can be employed to filter variation calls for the presence of novel variations and identify potential causative variants.

Usage Evolution

Originally, the data in dbSNP was used only to populate sequence maps since polymorphic marker density was too low to allow further application of the data. By 2007, however, marker density had increased enough to allow for the use of variation data in association studies, high resolution mapping, and a host of other applications, including population evolution and phylogeny studies that continue to further our understanding of genetic relationships and the genomic basis of traits.

dbSNP's current integration of clinical information into dbSNP records will allow greater application of dbSNP data to the fields of molecular medicine and pharmacogenetics, as well as emerging fields study such as pharmacometabolomics and precision medicine.

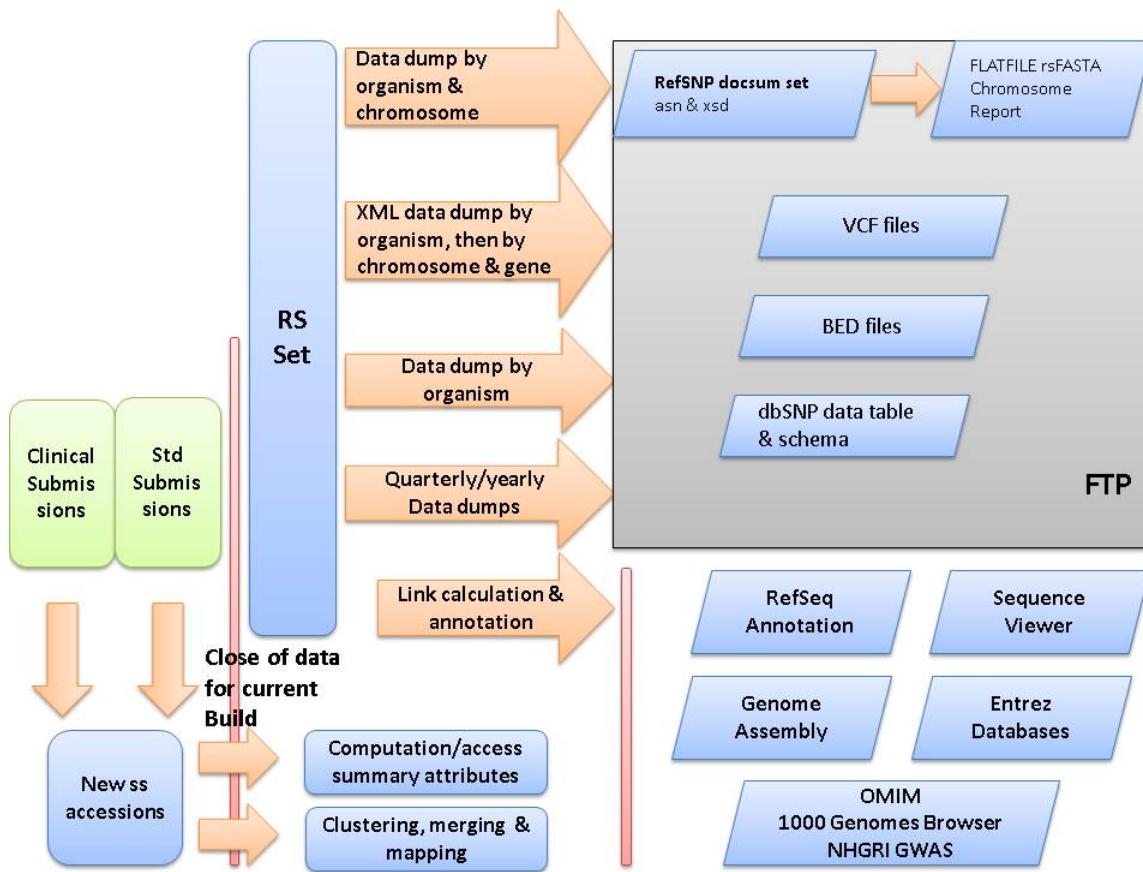


Figure 1. SNP Build Cycle. The dbSNP build cycle starts with close of data for new submissions. dbSNP calculates summary attributes and provide submitter asserted summary attributes for each refSNP cluster. These attributes include genotype, false positive status, minor allele frequency (MAF), asserted allele origin, asserted clinical significance, and many others. dbSNP maps all data, including existing refSNP clusters and new submissions, to the available reference genome sequence for the organism. If no genome sequence is available, dbSNP maps the data to non-redundant DNA sequences from GenBank. dbSNP uses map data on co-occurrence of hit locations to either merge submissions into existing clusters or to create new clusters. dbSNP then annotates the new non-redundant refSNP (rs) set on reference sequences and dump the contents of dbSNP into a variety of comprehensive formats on the dbSNP FTP site for release with the online build of the database. dbSNP then creates links from each record to internal and external resources that can provide additional data for each record.

Data Model

The dbSNP data model will evolve to capture new content, but currently has two major classes of data. The first class is submitted data, namely original observations of sequence variation, which are accessioned using a “submitted SNP” (ss) identifier. The second class is generated during the dbSNP build cycle (Figure 1) by aggregating data from multiple submissions as well as data from other sources and is identified with a “reference SNP” (rs) number. The rs identifier represents aggregation by type of sequence change and

location on the genome if an assembled genome is available, or aggregation by common sequence if a genome is not available.

It is important to note that no matter how the data are aggregated, the rs identifier is an identifier for a location and type of variation—it is not an identifier for every sequence that may have been observed at that location. In other words, if there is a single nucleotide variation in which alleles A, C, G, and T have all been observed, they all have the same rs identifier. And, if at a location, there is a single nucleotide variation and a length variation, then multiple rs identifiers will be assigned, one for each variation type.

Note: dbSNP is in the process of updating its assembly process. Further information about assembly changes will be available on dynamic SNP documentation currently under construction.

Submitted Content

dbSNP accepts submissions from public laboratories and private organizations. dbSNP does not accept synthetic mutations or variations ascertained from cross-species alignments and analysis. Variations > 50 nucleotides in length should be submitted to the Database of Genomic Structural Variation ([dbVAR](#)).

dbSNP will not hold data to be released on a particular date or in a particular dbSNP build. If, however, you are submitting non-clinical human data or non-human data and your manuscript requires dbSNP accession numbers (ss numbers) for the review process, we can hold the submitted data until the publication is accepted and you have notified us that dbSNP can release the data. Once notification has been given, dbSNP will release the data during the next build release cycle. See the [ClinVar Submission documentation](#) for the asserted clinical variation data hold policy.

A short tag or abbreviation called a “submitter HANDLE” uniquely defines each submitting laboratory and groups the submissions within the database. See dbSNP’s online [submission instructions](#) for help preparing a submission.

The 10 major data elements of a submission include:

Sequence Context

An essential component of a submission to dbSNP is an unambiguous definition of sequence context of the variation being submitted. dbSNP no longer accepts sequence context as a variant sequence within a flanking sequence, and now minimally requires that sequence context be submitted as an asserted position on RefSeq or [INSDC](#) sequences.

Asserted Positions

Asserted positions are statements based on experimental evidence that a variant is located at a particular position on a sequence accessioned in a public database. dbSNP prefers that all variant asserted positions are submitted on a sequence accession that is part of an

assembly housed in the NCBI [Assembly Resource](#). If no assembly is available, dbSNP will accept data on a RefSeq or [INSDC](#) sequence for an asserted position that is not associated with an assembly.

For those variations that have asserted positions not associated with an assembly, the rs of that variation cannot be annotated to the assembly, and therefore will not appear on maps or graphic representations of the assembly. If, however, at some future date, a new assembly is created in the Assembly Resource to which the sequence aligns, the reported variant will be assigned an rs number at that time. Once an rs number is assigned to the variant, the variant will appear on maps or graphic representations of the assembly.

Flanking Sequence

dbSNP no longer accepts flanking sequence on a routine basis, and now requires that variant positions are reported as asserted positions on a sequence that is part of an assembly housed in the [NCBI Assembly Resource](#).

Flanking sequence can only be used to report sequence context for those variants whose location could not be placed using an asserted position. Variations submitted with flanking sequence will be assigned a submitted SNP (ss) number that can be accessed by using the dbSNP homepage “ID search” tool or through an FTP download.

Because variants submitted with flanking sequence will be assigned an ss ID only, they will not appear on maps or graphic representations of the assembly. If, however, an assembly becomes available at a later date that allows us to map by BLAST, we will assign an rs to the variant if it is possible. dbSNP cannot predict when such an assembly will become available, or when mapping by BLAST will occur.

If a variant must be submitted with flanking sequence, dbSNP accepts variation flanking sequence as either genomic DNA or cDNA, and has a minimum length requirement of 25 bp on either side of the variation to maximize the specificity of the sequence in larger contexts.

Note: dbSNP structures its submissions so that a user can distinguish regions of assayed sequence actually surveyed for variation from those regions that are cut and pasted from a published reference sequence to satisfy dbSNP’s minimum-length requirements.

Note: SS numbers can be used in publications describing Assay Variants.

Alleles

Alleles define each variation class (Table 1). dbSNP defines single nucleotide variants in its submission scheme as G, A, T, or C, and does not permit ambiguous IUPAC codes, such as N, in the allele definition of a variation. In cases where variants occur close to one another, dbSNP permits IUPAC codes such as N, and in the flanking sequence of a variation, actually encourages them. See (Table 1) for the rules that guide dbSNP post-submission processing in assigning allele classes to each variation.

Table 1. Variation class organizes submissions by allele definition

Note: dbSNP has an allele length limitation of <=50bp. Submit alleles >50 nucleotides in length to the Database of Genomic Structural Variation ([dbVAR](#)).

<i>Variation Class^{a, b}</i>	<i>Allele Class Assignment Rules</i>	<i>Sample Allele Definition</i>	<i>Class Code^c</i>
Single Nucleotide Variation (SNV) ^a	Single base substitutions involving A, T, C, or G.	A/G	1
Deletion/Insertion Variations (DIVs) ^a	Designated by the full sequence of the insertion as one allele, and either a fully defined string for the variant allele or a “-” character to specify the deleted allele. This class will be assigned to a variation if the variation alleles are of different lengths or if one of the alleles is deleted (“-”).	-/AA/CCT/GCC/GCCTG ss149071	2
Heterozygous ^a	The term heterozygous is used to specify a region detected by certain methods that do not resolve the variation into a specific sequence motif. In these cases, a unique flanking sequence must be provided to define a sequence context for the variation.	(heterozygous)	3
Microsatellite or short tandem repeat (STR) ^a	Alleles are designated by providing the repeat motif and the copy number for each allele. Expansion of the allele repeat motif designated in dbSNP into full-length sequence will be only an approximation of the true genomic sequence because many microsatellite markers are not fully sequenced and are resolved as size variants only.	(CAC)8/9/10/11	4
Named ^a	Applies to insertion/deletion variants of longer sequence features, such as retroposon dimorphism for Alu or line elements. These variations frequently include a deletion “-” indicator for the absent allele. Observed field starts with '(', but is not class 3 or 4	(alu) / -	5
NoVariation ^a	Reports may be submitted for segments of sequence that are assayed and determined to be invariant in the sample.	(NoVariation)	6

a) Seven of the classes apply to both submissions of variations (submitted SNP assay, or ss#) and the non-redundant refSNP clusters (rs#'s) created in dbSNP. b) The “Mixed” class is assigned to refSNP clusters that group submissions from different variation classes. c) Class codes have a numeric representation in the database itself and in the export versions of the data (VCF and XML).

Table 1. continues on next page...

Table 1. continued from previous page.

<i>Variation Class^{a, b}</i>	<i>Allele Class Assignment Rules</i>	<i>Sample Allele Definition</i>	<i>Class Code^c</i>
Mixed ^b	The refSNP cluster contains submissions from 2 or more allelic classes	Mix of allelic classes	7
Multi-Nucleotide Variation (MNV) ^a	Multi-base variations of a single, common length.	AT/GA ss2421179	8
Exception	The submitted variation needs to be checked	The submitted variation does not contain “/” to indicate presence of variant.	9

a) Seven of the classes apply to both submissions of variations (submitted SNP assay, or ss#) and the non-redundant refSNP clusters (rs#'s) created in dbSNP. b) The “Mixed” class is assigned to refSNP clusters that group submissions from different variation classes. c) Class codes have a numeric representation in the database itself and in the export versions of the data (VCF and XML).

Method

Each submitter defines the methods in their submission as either the techniques used to assay variation or the techniques used to estimate allele frequencies. dbSNP groups methods by method class (Table 2) to facilitate queries using general experimental technique as a query field. The submitter provides all other details of the techniques in a free-text description of the method. Submitters can also use the METHOD_EXCEPTION field to describe changes to a general protocol for particular sets of data (batch-specific details). Submitters generally define methods only once in a submission.

Table 2. Method class organizes submissions by methodological or experimental approach

Method Class	Class Code
Denaturing high pressure liquid chromatography (DHPLC)	1
DNA hybridization	2
Computational analysis	3
Single-stranded conformational polymorphism (SSCP)	5
Other	6
Unknown	7
Sequence	9
Clinical Submission; DHPLC	101
Clinical Submission; Hybridization	102
Clinical Submission; Computation	103
Clinical Submission; SSCP	105
Clinical Submission; Other	106

Table 2. continues on next page...

Table 2. continued from previous page.

Method Class	Class Code
Clinical Submission; Unknown	107
Clinical Submission; RFLP	108
Clinical Submission; Sequence	109

Asserted Allele Origin

A submitter can provide a statement (assertion) with supporting experimental evidence that a variant has a particular allelic origin. Assertions for a single refSNP are summarized and given an attribute value of germline or unknown. Variants of somatic origin should be submitted to [ClinVar](#). Additional attributes (e.g., paternal) will be added in the future.

Population

Each submitter defines population samples either as the group used to initially identify variations or as the group used to identify population-specific measures of allele frequencies. These populations may be one and the same in some experimental designs. Although dbSNP has assigned populations to a population class based on the geographic origin of the sample, we will phase out this practice in the near future since most population descriptions are now submitted to [BioSample](#). We are encouraging dbSNP submitters to start registering their samples with BioSample to obtain an assigned accession that they can use in their dbSNP submission.

Sample Size

There are two sample-size fields in dbSNP. One field, SNPASSAY SAMPLE SIZE, reports the number of chromosomes in the sample used to initially ascertain or discover the variation. The other sample size field, SNPPOPUSE SAMPLE SIZE, reports the number of chromosomes used as the denominator in computing estimates of allele frequencies. These two measures need not be the same.

Population-specific Allele Frequencies

Alleles typically exist at different frequencies in different populations; a very common allele in one population may be quite rare in another population. Also, allelic variants can emerge as private polymorphisms when particular populations have been reproductively isolated from neighboring groups, as is the case with isolated or remote populations.

Frequency data are submitted to dbSNP as allele counts or binned frequency intervals, depending on the precision of the experimental method used to make the measurement. dbSNP contains records of allele frequencies for specific population samples that are defined by each submitter and used in validating submitted variations. See Table 3 for use of allele frequencies in validation.

Table 3. Validation status codes summarize available validation data

Validation evidence	Description	Code in database for ss#	Code in FTP dumps for ss#	Code in database for rs#	Code in FTP dumps for rs#
Not Validated	Validation code for an ss, where there is no BatchUpdate information, no 0 or 1 frequency data, and no non-computational validation method. Validation code for an rs, "" or ""ed from ss code. If the rs has single ss with code 1, then the rs code is set to 0.	0	Not present	0 ^a	Not present
By Cluster	Validation code for an rs that has at least two ss, and at least one of those ss was validated by a non-computational method. Validation code for an ss, if the method is non-computational. Validation code for a single member rs where the ss validation_status is 1, the rs validation_status is set to 0.	1	1 ^b	1,0 ^b	1
By Frequency	Validation code for a variation that has frequency or genotype data and has a minor allele count of at least 2.	2	2	2	2
By Submitter	Validation code for a variation that had a BatchUpdate with a second validation method submitted by the original submitter.	4	4	4	4
by DoubleHit	Validation code for a variation where every allele has been observed in at least two chromosomes.	8	8	8	8
HapMap	Validation code for a variation that has genotype frequency from HapMap.	16	16	16	16

a) If the rs# has a single ss# with code 1, then rs# is set to code 0. b) For a single member rs where the ss validation status = 1, the rs# validation status is set to 0.

Note: 57 Additional validation codes defined by bitstring are available in the [SNPValidationClass](#) file, location in the [shared_data](#) directory of the dbSNP FTP site.

Population-specific Genotype Frequencies

Similar to alleles, genotypes have frequencies in populations that can be submitted to dbSNP, and are used in validating submitted variations.

Individual Genotypes

dbSNP accepts individual genotypes from samples provided by donors that have consented to having their DNA sequence housed in a public database (e.g., HapMap or the 1000 Genomes project). Genotypes reported in dbSNP contain links to population and method descriptions. General genotype data provide the foundation for individual haplotype definitions and are useful for selecting positive and negative control reagents in new experiments.

Validation Information

dbSNP accepts individual assay records (ss numbers) without validation evidence. When possible, however, dbSNP tries to distinguish high-quality validated data from unconfirmed (usually computational) variation reports. Assays validated directly by the submitter through the VALIDATION section show the type of evidence used to confirm the variation. Additionally, dbSNP will flag an assay variation as validated (Table 3) if:

- There are multiple independent submissions to the refSNP cluster with at least one non-computational method,

OR

- The variation was genotyped by the HapMap project, sequenced by the 1000 Genomes project, or other large sequencing projects.

Computed Content

dbSNP releases its content to the public in periodic “builds” that are synchronized with the release of new genome assemblies for each organism (Handbook: Eukaryotic Genome Annotation Pipeline). The dbSNP build process proceeds as follows:

1. Cluster variations (ss) submitted since the previous build into RefSNPs (rs).
2. Map the refSNP clusters to the appropriate assembly.
3. Merge co-locating refSNP clusters when appropriate.
4. Mark suspected false positive variants (see the “Suspect Variations” section of this chapter for more information on false positive selection).
5. Compute a functional context for the mapped variants.
6. Compute minor allele frequency as well as average heterozygosity and standard error.
7. Compute links to other relevant NCBI resources such as Gene, PubMed, and RefSeq for RefSNP clusters.
8. Map all clustered variations to RefSeq sequences.

See Figure 1 for a complete graphic description of the dbSNP Build process.

Dataflow

New Submissions and the Start of a New Build

Each build starts with a “close of data” that defines the set of new submissions that will be mapped to genome sequence for subsequent annotation and grouping of variations into refSNPs. The set of new data entering each build typically includes all submissions received since the close of data in the previous build.

Submitted SNPs and Reference SNP Clusters

When a new variation is submitted to dbSNP, it is assigned a unique submitted SNP ID number (ss#). If the variation is submitted with an asserted position, once the ss number is assigned, the asserted position coordinates are remapped to the corresponding coordinates on the current assembly. If the variation is submitted with flanking sequence, dbSNP aligns the flanking sequence of each submitted SNP to its appropriate genomic location(s).

When multiple submissions of the same variation class (Table 1) that have the same weight (uniqueness) map to the same position on the assembly, dbSNP clusters the ss, defines the “reference SNP cluster,” or “refSNP,” and provides the cluster with a unique RefSNP ID number (rs#). If submitted SNPs of more than one variation class map to a single position, then an rs number will be assigned for each variation class at that position. If only one submission maps to a specific position, then its ss is assigned an rs number and is the only member of its RefSNP cluster until another submitted SNP of the same variation class is found that maps to the same position.

A refSNP cluster has a number of summary attributes that are computed over all cluster members (Figure 2), and are used to annotate variations contained in other NCBI resources. See Figures 2A, 2B, 2C and 2D for the location of all summary attributes and internal/external resource links in a refSNP cluster report.

dbSNP exports the entire dbSNP refSNP set in many report formats to its [FTP](#) site, and delivers them as sets of results when a user conducts a dbSNP batch query.

Note: Summary properties derived from asserted data (e.g., clinical assertions, asserted positions, asserted allele origin) are based on experimental evidence and cannot be seen as a confirmation of a particular clinical phenotype, genomic position, or allele origin since NCBI does not independently verify assertions and cannot endorse their accuracy.

Mapping to a Genome Sequence

When a new genome build is ready, dbSNP uses assembly-assembly alignments to remap ss asserted locations and rs locations from the old assembly to the new assembly.

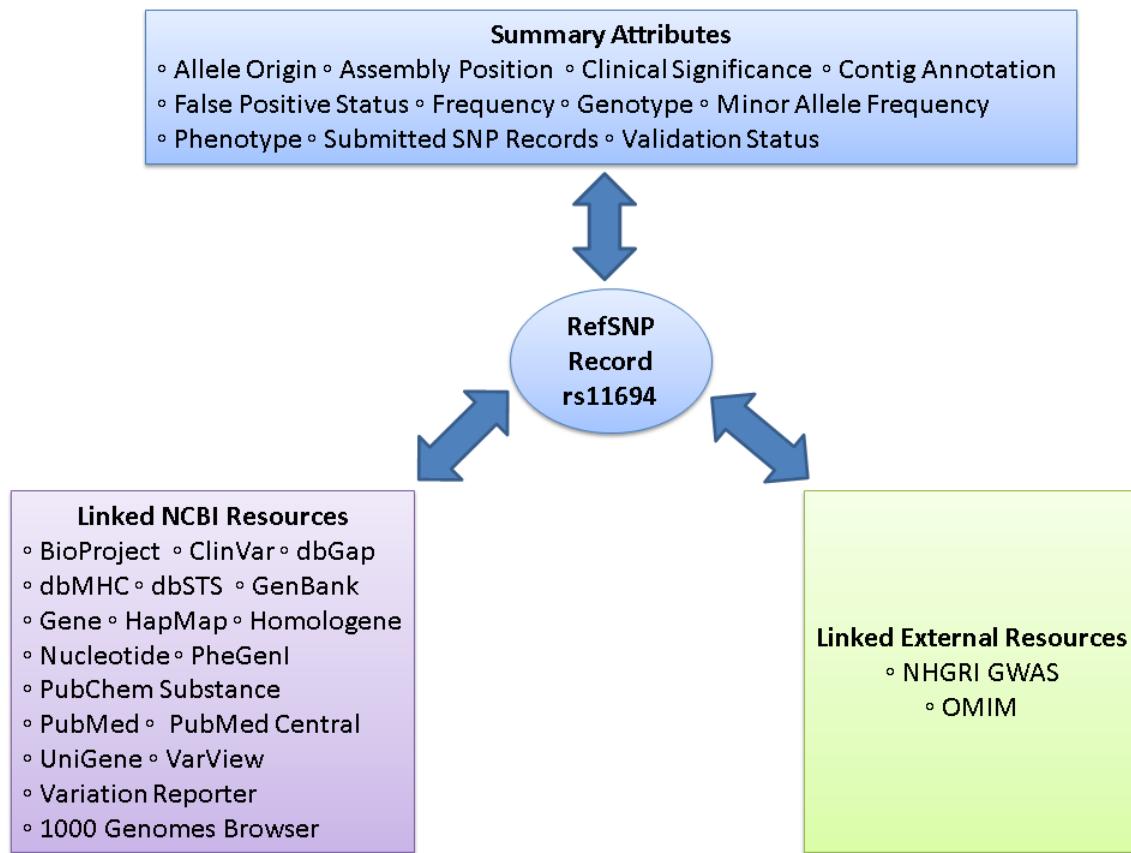


Figure 2. The refSNP Summary Record (refSNP Cluster Report). The refSNP Summary Record, also known as the refSNP Cluster Report, provides the user with extensive summary attributes that are supplied by the submitter, calculated by dbSNP from submitted data, or contributed by another NCBI resource. Links within the refSNP summary record direct the user to additional information from any of 20 possible internal and external websites. These linked sites may contain supplementary data that were used in the initial variation call, or may provide detailed phenotypic or clinical assertion data that may suggest variation function. Figures 2A through 2D illustrate the location of the summary attributes as well as the internal and external resource links.

To map submissions without asserted locations to a genome assembly, dbSNP obtains FASTA files for those submitted SNPs submitted prior to the “close of data,” as well as FASTA files for refSNPs in the current build that can’t be remapped, and then maps the submitted SNPs and refSNPs to the genome sequence using the BLAST procedure described in Appendix 2.

If an organism is represented by multiple assemblies then each assembly is annotated. For example, dbSNP annotates two major human assemblies: the Genome Reference Consortium (GRC) Reference assembly, and the haploid hydatidiform mole (CHM1) assembly.



Figure 2A. The refSNP Summary Report: Allele Summary and Integrated Maps Sections. The Allele Summary section of the refSNP report provides clinical significance (A), where the phenotype may be viewed by clicking on either the VarView or the OMIM buttons; the allele origin (B), indicated as Germline or Somatic for each allele; the Minor Allele frequency (C); Validation Status (D), where definitions for graphic icons indicating validation class are viewed by clicking the “Validation Status” column header link (D); and links to both internal and external resources (E) that provide additional data. The Integrated Maps section of the refSNP report provides a summary of genome mapping information for the variation (F), which can be accessed on the NCBI Sequence Viewer by clicking on any value in the Chromosome Position (Chr Pos) column or the Contig Position (Contig Pos) column. The magnifying glass icon (G) links to a view of the variation in the 1000 Genomes Browser.

refSNP Clustering and refSNP Orientation

The orientation of a refSNP, and hence its sequence and allele string, is set by the first submitted SNP (ss) used to create a refSNP (rs) cluster. By convention, the cluster exemplar is the member of a refSNP cluster that has the longest flanking sequence or is the first variant with an asserted location assigned in the cluster. If in a later build, a new variant added to the cluster becomes the exemplar and happens to be in reverse orientation to the current orientation of the refSNP, dbSNP preserves the orientation of the refSNP by using the reverse complement of the cluster exemplar to set the orientation of the refSNP sequence.

For those variants that are submitted with an asserted position rather than a flanking sequence, once dbSNP maps and verifies the asserted position, the flanking sequence is derived from the asserted position and used to determine the variant orientation.

Once the clustering process determines the orientation of all member sequences in a cluster, it will gather a comprehensive set of alleles for a refSNP cluster.

Following the dbSNP redesign, dbSNP will keep the use of flanking sequence and exemplars for those organisms that don't have a mature assembly, but will phase out the “exemplar” concept once creation of asserted positions for those variations that were submitted with flanking sequence has begun.

Note: dbSNP reports all variants and variant alleles on the + strand of the assembly.

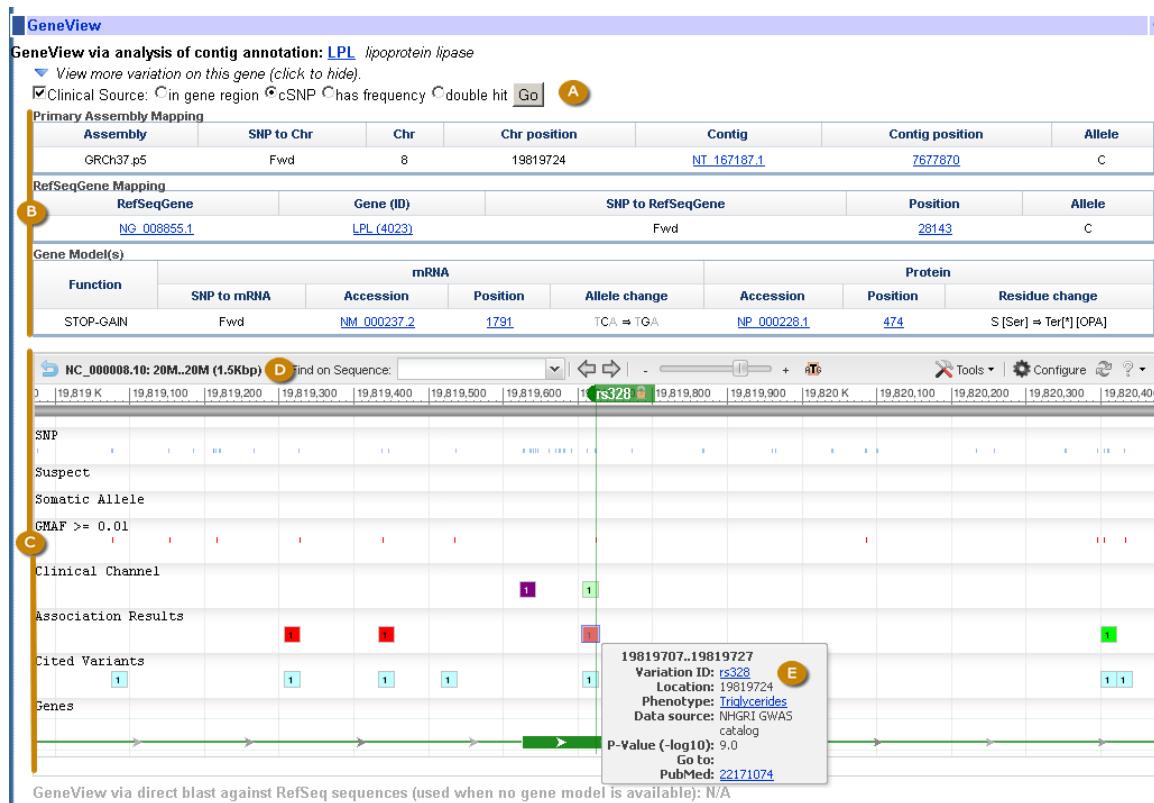


Figure 2B. The refSNP Summary Report: the GeneView Section. The GenView section has a Display menu at the top. Once menu choices have been made, clicking the “Go” button (A) will generate the GeneView Display. The default setting (“Clinical Source” and “cSNP”) for this menu generates a tabular GeneView display (See figure 2C). The tables that follow (B) summarize variation mapping information and protein changes, and the section ends with a Sequence Viewer display of the variation on the latest genome assembly (C). Clicking on the reference sequence accession number at the top left of the display (D) allows you to view the accession on the human assembly in one of several views. Mousing over an icon in the display generates a pop up view of additional information (E).

Re-Mapping, refSNP Merging and refSNP Splitting

Re-Mapping and refSNP Merging

RefSNPs are operationally defined as a variation at a location on a reference assembly. Every time there is a genomic assembly update, the interim reference sequence may change, so the refSNPs must be updated or re-clustered.

The re-clustering process begins when NCBI updates the genomic assembly. All existing refSNPs (rs) and newly submitted SNPs (ss) are mapped to the genome assembly using assembly-assembly remap or multiple BLAST and MegaBLAST cycles as delineated in Appendix 2.

dbSNP clusters variations that co-locate at the same place on the genome into a single refSNP. Newly submitted variations can either co-locate to form a new refSNP cluster, or

Clinical Source in gene region cSNP has frequency double hit refresh

gene model (contig mRNA transcript)		Contig Label	Contig	mRNA	protein	mRNA orientation	transcript SNP count									
Region	Chr. position	mRNA pos	dbSNP rs# cluster id	Heterozygosity	A Validation	MAF	Allele origin	3D	Clinically Associated	Clinical Significance	Function	dbSNP allele	Protein residue	Codon pos	Amino acid pos	PubMed
19797034	453	rs11570895	N.D.								missense	T	Val [V]	2	28	
											contig reference	C	Ala [A]	2	28	
19805707	475	rs145405273	0.001								synonymous	T	Ile [I]	3	35	
											contig reference	C	Ile [I]	3	35	
19805708	476	rs1801177	N.D.								missense	A	Asn [N]	1	36	F
											contig reference	G	Asp [D]	1	36	F

Figure 2C. The refSNP Summary Report: The GeneView Display. The default GeneView display provides a tabular summary of variations mapped to splice variants of a particular gene. The variation summary is arranged in the sequential order that the variations appear in the genome and are color coded by functional class: in the example above, red indicates non-synonymous change and green indicates synonymous change. Additional colors not seen in the above example include white (in gene region), orange (UTR), blue (frame shift), yellow (intron). Attributes provided in the GeneView display include validation class (A), where definitions for graphic icons indicating validation class are viewed by clicking the “Validation” column header link (A); Minor Allele Frequency (if available) (B); allele origin (C) indicated as Germline or Somatic; a link to the Structure record (D) if 3D protein structure information is available; a link to the OMIM record (E) for those variations that have a clinical assertion; and a link to a list of Pubmed articles (F) that cite the variation.

may instead cluster with an already existing refSNP. When newly submitted variations cluster among themselves, they are assigned a new refSNP number, and when they cluster with an existing refSNP, they are added to that refSNP cluster.

Sometimes an existing refSNP will co-locate (identical coordinates) with another refSNP when dbSNP improves its clustering algorithm, when submissions are corrected, or when genome assemblies change between dbSNP builds. When existing refSNPs co-locate, the refSNP(s) with higher refSNP number(s) are retired (never to be reused), and all the submitted SNPs from the retired cluster(s) are re-assigned to the retained refSNP. The re-assignment of the submitted SNPs from a higher refSNP number to a refSNP cluster with a lower refSNP number is called a “merge,” and occurs during the “rs merge” step of the dbSNP mapping process. Merging is only used to reduce redundancy in the catalog of rs numbers so that each position has a unique identifier. All “rs merge” actions that occur are recorded and tracked.

Note: Originally, refSNPs clusters included variations of different class types because the submitted variations happen to map to the same location (true SNP, indels, mixed). dbSNP found that due to ballooning data submissions, however, refSNPs were becoming increasingly difficult to interpret because of the multiple variation class types present in each cluster. Since different variation classes represent different genetic events, dbSNP has since altered refSNP clusters to include only a single variation class type.

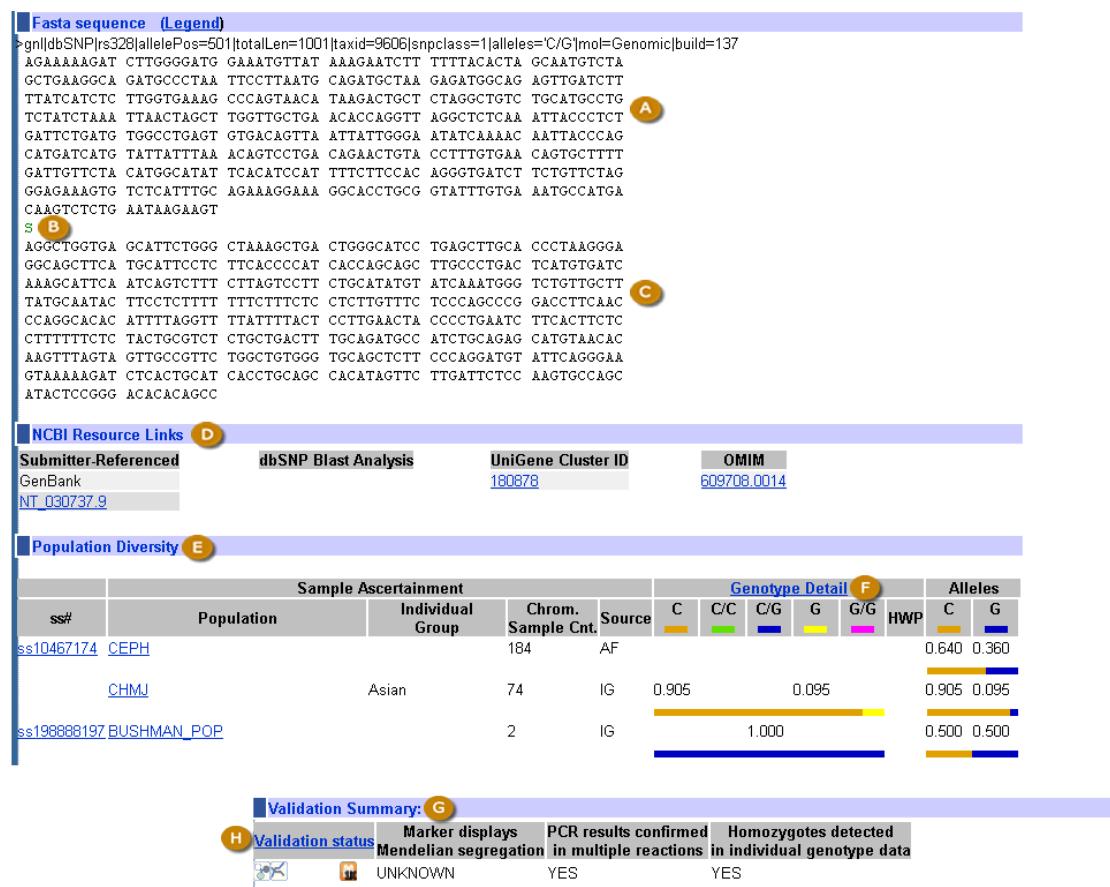


Figure 2D. The refSNP Summary Report: FASTA, Resource Links, Population Diversity and Validation Summary Sections. The FASTA section provides the variation 5' flanking sequence (A), the allele (B), and the 3' flanking sequence (C) as provided by the submission record or determined from an asserted position. The NCBI Resource Links section (D) provides links to additional information (if available) from Genbank, a BLAST analysis, UniGene, OMIM, or Structure (not shown). The Population Diversity section (E) provides a table of genotypes and allele frequencies for the variant from different populations and studies. Click on the Genotype Detail link (F) to see a “Genotype and Allele Frequency” report that provides detailed genotype and allele frequency information for each submitted variation of the cluster. The Validation Summary section (G) is the final section of the refSNP summary report, and provides a summary of the validation status for the variation. To see definitions for the icons used in the Validation Status graphical summary, click on the “Validation status” column header link (H).

refSNP Splitting

Due to assembly changes or software updates, submissions previously calculated to be in the same cluster can be differentiated. In such cases, dbSNP separates or “splits” the cluster into two or more refSNP clusters depending on the particular circumstance. dbSNP may also split a cluster if newly submitted evidence indicates that two or more variation classes are clustered within a single refSNP number.

When an existing refSNP is split, those submitted SNP (ss) numbers that were most recently added to the cluster will be “split away” to form a new cluster. When this happens, the remaining ss numbers in the original cluster retain the old rsID number, while the ss numbers that are “split away” either cluster to another existing refSNP if they map to it, or are assigned a new refSNP number. The number of clusters that emerge from a split depends upon the number of distinct locations and types of variation classes that can now be identified.

RefSNP Number Stability

If a refSNP number has been merged into or split away from another refSNP number, it is very easy to use a retired refSNP number to find the current one (see hint below). In other words, a refSNP number can be termed stable because merged or split refSNP numbers can always be traced to a previous refSNP number.

Hint:

There are three ways the you can locate the partner numbers of a merged refSNP, and one way to locate the partner of a split refSNP:

- If you enter a retired rs number into the “Search for IDs” search text box on the [dbSNP home page](#), the response page will state that the SNP has been merged, and will provide the new rs number and a link to the refSNP page for that new rs number.
- You can retrieve a list of merged rs numbers from [Entrez SNP](#). Just type “mergedrs” (without the quotation marks) in the text box at the top of the page and click the “go” button. You can limit the output to merged rs numbers within a certain species by clicking on the “Limits” tab and then selecting the organism you wish from the organism selection box. Each entry in the returned list will include the old rs numbers that has merged, and the new rs number it has merged into (with a link to the refSNP page for the new rs number).
- You can also review the [RsMergeArch table](#) for the merge partners of a particular species of interest, as it tracks all merge events that occur in dbSNP. This table is available on the dbSNP FTP site, a full description of it can be found in the [dbSNP Data Dictionary](#), and the column definitions are located in the dbSNP_main_table.sql.gz, which can be found in the [shared_schema](#) directory of the dbSNP FTP site.
- You can locate the partner of a split refSNP only by using SQL:

```
SELECT *
FROM [human_9606].[dbo].[RsSplitArch] where rs_2_split = 26
rs_2_split rs_new_split split_build_id create_time last_updated_time
26 78384355 132 2010-08-19 23:38:00 2010-08-19 23:38:00
```

If, however, what is meant by “stable” is that the refSNP number of a particular variation always remains the same, then one should not consider a refSNP entirely stable, as a refSNP number may change if two refSNP numbers merge or split. Merging can occur if new evidence suggests that two refSNPs at a single sequence location are of the same variation type, and splitting will occur if mixed variation classes (e.g., SNV and indel) are clustered in a single refSNP. For more detail on merging and splitting, see the “Re-Mapping and RefSNP Merging” section and the “refSNP splitting” sections above.

A refSNP number may also change if:

- All of the submitted SNP (ss) numbers in a refSNP cluster are withdrawn by the submitter.
- dbSNP breaks up an existing cluster and re-instantiates a retired rs number based on a reported conflict from a dbSNP user.

Suspect Variations

Currently, a variant is flagged as a “suspect”, i.e., a potential false positive when the presence of a paralogous sequence in the genome (1,2) could cause mapping artifacts or if there is evidence suggesting sequencing errors or computation artifacts.

dbSNP will be updating its suspect false positive flagging system in the near future to rank suspect variants according to the amount of supporting evidence available for each refSNP. Those refSNP clusters that are suspect, but have data from multiple submitters indicating a heterozygous state exists, will rank more highly in dbSNP’s new system as a variation to be trusted than a suspect refSNP that has data from a single submission, has multiple submissions that show no evidence of heterozygosity, or with conflicting evidence of heterozygosity.

Molecular Class

dbSNP computes a molecular context for sequence variations by inspecting the flanking sequence for gene features during the contig annotation process, and does the same for RefSeq/GenBank mRNAs.

dbSNP has adopted [Sequence Ontology](#) (SO) terms to define its variation molecular classes so as to conform to the standard set by the biological community. The subset of SO terms dbSNP uses as functional classes can be found in Table 4.

A variation may have multiple functional classes. Multiplicity will result, for example, when a variation falls within an exon of one transcript and an intron of another for the same gene.

Table 4. Molecular codes for refSNPs in gene features

Functional class	Definition/Example	dbSNP code	Sequence Ontology Code (term definition)
cds-synon	Synonymous change. Example: rs248, GAG->GAA, both produce amino acid: Glu	3	SO:0001819

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Table 4. continues on next page...

Table 4. continued from previous page.

Functional class	Definition/Example	dbSNP code	Sequence Ontology Code (term definition)
intron	Example: rs249	6	SO:0001627
cds-reference	Contig reference	8	
synonymy unknown	Coding: synonymy unknown. Not used since 2003	9	
nearGene-3	Within 3' 0.5kb to a gene. Example: rs3916027 is at NT_030737.9 pos7669796, within 500 bp of UTR starts 7669698 for NM_000237.2	13	SO:0001634
nearGene-5	Within 5' 2kb to a gene. Example: rs7641128 is at NT_030737.9 pos7641128, with 2K bp of UTR starts 7641510 for NM_000237.2	15	SO:0001636
intergenic	Variant between two genes, outside of 2Kb upstream and 500bp downstream of a gene	20	SO:0001628
ncRNA	Variant on non-coding RNA(NCBI Refseq prefix NR,XR)	30	SO:0001619
STOP-GAIN	Changes to STOP codon. Example: rs328 , TCA->TGA, Ser to terminator	41	SO:0001587
missense	Alters codon to make an altered amino acid in protein product. Example: rs300 , ACT->GCT, Thr->Ala	42	SO:0001583
STOP-LOSS	Changes STOP codon to other non stop codon	43	SO:0001578
frameshift	Indel SNP causing frameshift	44	SO:0001589
cds-indel	Indel SNP with length of multiple of 3bp, not causing frameshift	45	SO:0001650
UTR-3	3 prime untranslated region. Example: rs3289	53	SO:0001624

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Table 4. continues on next page...

Table 4. continued from previous page.

Functional class	Definition/Example	dbSNP code	Sequence Ontology Code (term definition)
UTR-5	5 prime untranslated region. Example: rs1800590	55	SO:0001623
splice-3	3 prime acceptor dinucleotide. The last two bases in the 3 prime end of an intron. Most intron ends with AG. Example: rs193227 is in acceptor site.	73	SO:0001574
splice-5	5 prime donor dinucleotide. 1st two bases in the 5 prime end of the intron. Most intron starts is GU. Example: rs8424 is in donor site.	75	SO:0001575

Most gene features are defined by the location of the variation with respect to transcript exon boundaries. Variations in coding regions, however, have a functional class assigned to each allele for the variation because these classes depend on allele sequence.

Clinical Assertions

Novel variations and experimental evidence supporting clinical assertions (Table 5) are submitted to ClinVar. dbSNP and ClinVar will continue to support the Human Variation Batch Submission site as a Web-based tool that can be used to submit or update medically important variation submissions.

When novel variants with supporting clinical evidence are received from ClinVar, dbSNP remaps the asserted positions of the variants to the corresponding coordinates on the current assembly, as well as to cDNA, protein, and RefSeqGene sequences. Once the variants are mapped, dbSNP assigns ssIDs and rsIDs to each.

Data submitted through the Human Variation Batch Submission site that supports a clinical assertion are processed and extracted by ClinVar, which uses the data to assign clinical attributes to novel variants or to update the clinical attributes of existing variants (LSDB).

Once the submitted variants are mapped and their attributes assigned, these data are made available to other NCBI resources, including VarView, ClinVar, and Variation Reporter.

Table 5. Clinical Significance organizes submissions by clinical assertion type

Class Code	VCF and ASN.1 Terms	ClinVar Display Terms
0	unknown	Uncertain significance
1	untested	not provided
2	non-pathogenic	Benign
3	probable-non-pathogenic	Likely benign
4	probable-pathogenic	Likely pathogenic
5	pathogenic	Pathogenic
6	drug-response	drug response
7	histocompatibility	
255	other	other
		confers sensitivity
		risk factor
		association
		protective

¹ Variations for which there is not yet an enumerated clinical significance class. These variations are grouped in a clinical significance class called "other", which includes: Variations* that are found only in somatic cells and are with or without known trait of phenotype; Somatic or germline variations that are disease risk factors; Somatic or germline variations that act to protect a disease state (protective variants)

* Note: If a variant's source is not asserted during submission, dbSNP assumes that the source of the variant is germline. Those variants submitted with the clinical phrase (clinic_phrase) tag set to "cancer" are reported as somatic.

Note: As assertion categories may change, see [ClinVar](#) for up-to-date clinical assertion definitions.

Note: The clinical significance terms presented in table 5 are based on terminology recommended by the American College of Medical Genetics and Genomics (ACMG). ACMG revisions are adopted by NCBI as quickly as possible. See <http://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/#standard> for the most recent clinical significance terms used in NCBI's reporting.

Population Diversity Data

Average Heterozygosity

The best single measure of a variation's diversity in different populations is its average heterozygosity. This measure serves as the general probability that both alleles are in a diploid individual or in a sample of two chromosomes. Estimates of average heterozygosity have an accompanying standard error based on the sample sizes of the underlying data, which reflects the overall uncertainty of the estimate. dbSNP's computation of average heterozygosity and standard error for RefSNP clusters is available [online](#). Please note that dbSNP computes heterozygosity based on the submitted allele frequency for each variation. If the frequency data for a variation is not submitted, dbSNP cannot compute the heterozygosity value, and therefore the refSNP report will show no heterozygosity estimate.

Additional population diversity data calculated for refSNP records includes population counts, individuals sampled for a variation, genotype frequencies, and Hardy Weinberg probabilities.

Minor Allele Frequency (MAF)

Minor Allele Frequency is the allele frequency for the 2nd most frequently seen allele. dbSNP aggregates the minor allele frequency for each refSNP cluster over multiple submissions to help users distinguish between common polymorphisms and rare variants.

Consider a variation with the following alleles and allele frequencies:

Reference Allele = G; frequency = 0.600

Alternate Allele = C; frequency = 0.399

Alternate Allele = T; frequency = 0.001

Based on the MAF guideline mentioned above, the minor allele is "C," so the minor allele frequency (MAF) is 0.399. Allele "T" with frequency 0.001 is considered a rare allele rather than a minor allele.

1000 Genomes Minor Allele Frequency (1000G MAF)

"1000G MAF" is the minor allele frequency (see above) based on genotype data from the 1000 Genomes Project [phase 1] global population of 1094 individuals.

Build Integration

dbSNP annotates the non-redundant set of variations (refSNP cluster set) on reference genome genomic sequences, chromosomes, mRNAs, and proteins as part of the NCBI RefSeq project. dbSNP computes summary properties for each refSNP cluster, which are then used to build fresh indexes for dbSNP in the Entrez databases, and to update the variation map in the NCBI Map Viewer. Finally, dbSNP updates links between dbSNP and BioProject, dbVar, dbGaP, Gene, HomoloGene, Nucleotide, OMIM, Protein, PubChem Substance, PubMed, PubMed Central, VarView, and Variation Reporter.

Public Release

Public release of a new build involves an update to the public database and the production of a new set of files on the dbSNP FTP site. dbSNP makes an announcement to the [dbsnp-announce](#) mailing list when the new build for an organism is publicly available.

The dbSNP Redesign: Changes to Clustering

As of this writing, dbSNP is planning a redesign that will introduce a number of fundamental changes to dbSNP's dataflow. Users are encouraged to review the proposed changes below and to submit any comments and suggestions to snp-admin@ncbi.nlm.nih.gov. Among these changes is a new clustering algorithm.

While improvements continue to be made in genome assemblies, artificial sequence duplications will resolve and collapse, artificially collapsed regions of a genome will expand, and missing sequence regions will be added. In the case of a sequence collapse, the number of hits that a variation has on the assembly may be reduced, while in the case of a sequence expansion or addition, the number of hits that a variation has to the genome may increase.

In order to have the flexibility necessary to deal with these genome assembly changes, dbSNP 2.0 will have a new clustering algorithm that will change the concept of a refSNP as we know it. This new clustering algorithm will change the rules that govern clustering to invert the ss to rs relationship as it currently exists.

At present, multiple ss numbers can exist in the same rs cluster and each ss number in that cluster links to its one corresponding rs number. The new clustering rules, however, will change this relationship in that each rs will represent a unique location. Under these new rules, a single ss number could link to multiple rs numbers if the ss number maps to more than one location.

Access

The SNP database can be queried directly from the search bar at the top of the [dbSNP homepage](#), by using the links to dbSNP resources and search options located on the homepage, or by accessing related NCBI resources that link to dbSNP data.

dbSNP Home Page

dbSNP is a part of the Entrez integrated information retrieval system and may be searched either by using an ID number query, or by using combinations of different search fields and qualifiers.

Single Record Query

Use the search bar at the top of the [dbSNP homepage](#) to find variations using dbSNP record identifiers. The record identifiers currently supported for single record queries are the reference SNP (refSNP) cluster ID number (rs#), the submitted SNP accession number (ss#), and the local (or submitter) ID number.

Complex Entrez Query

Use the [SNP Advanced Search Builder](#) page to construct a complex search using combinations of different search fields and qualifiers. The Advanced Search Builder allows you to construct a query by selecting multiple search terms from a large number of fields and qualifiers. See the [Advanced Search Builder video tutorial](#) for information about how to find existing values in fields and combine them to achieve a desired result.

dbSNP Batch Query

dbSNP Batch Query allows you to query using variation IDs (rs ID, ss ID, or local IDs), collected in a primary search to retrieve a large quantity of variations at the same time in a selected report format. Available report formats include ASN.1, BED, chromosome, FASTA, Flat File, genotype report, rs cluster report, ss detail report, and XML.

Variation Reporter

Variation Reporter matches submitted variation calls to variants housed in dbSNP or dbVar, thereby allowing access via a Web search or through an application programming interface (API) to all data and metadata that dbSNP has for the matching variants. If you submit novel variants and there are no matches between your data and the variants housed in dbSNP or dbVar, the Variation Reporter will provide the predicted consequence of each submitted variant.

BLAST

BLAST can be used to match submitted variations with asserted positions to matching dbSNP records (See the instructions at: ftp://ftp.ncbi.nih.gov/pub/factsheets/HowTo_Finding_SNP_by_BLAST.pdf). Query BLAST with the sequence or clone that contains the asserted position of the variant, and then select an appropriate reference database as the BLAST target. The BLAST algorithm will find any existing SNP records that map to the queried sequence, and thence to the variant of interest, if a dbSNP record happens to match it.

Note: If BLAST fails to find a matching dbSNP record for a variation of interest on a queried sequence:

1. You cannot assume that the variant is novel without further study, because there are several reasons why an existing variant may not yet have a dbSNP record:
 - a. The sequence location of the existing variant may be missing from the reference assembly, or the transcript location of the variant has not yet been sequenced.
 - b. The existing variant may have been submitted with low sequence quality or ambiguous base calls, which would inhibit placement on the reference assembly.
 - c. The variant may exist in the literature, and has not yet been submitted by the author for inclusion in dbSNP. This is particularly true for those variants that were reported in historic literature.
2. You can use Variation Reporter to get a predicted consequence of human variations to help you in your analysis if the variations have known sequence locations.

SNP Submission Information Queries

If refSNP(rs) or submitted SNP (ss) numbers are not available to use in a search for a dbSNP record, use the “[Submission Information](#)” module to construct a query that will select dbSNP variation records based on other available information associated with a submitted variation:

- Information associated with the submitter
- Information about the submitted batch that contains the variation of interest
- Information associated with the method used to assay for the variation (Table 2)
- Information associated with the submitted population
- Information associated with the publication reporting the variant

Search via ClinVar, Gene, or PubMed

There are multiple databases in NCBI that maintain links to dbSNP. Related records in dbSNP can be identified by following the Find related data on the Summary display, or following the links in the Related information section of a single record.

Entrez Programming Utilities (Eutils)

Use Entrez Programming Utilities (E-utilities or Eutils) to query dbSNP and retrieve information via Web services. You can test an Entrez query interactively and then execute that query using Eutils. There are a number of available Eutil programs that cover a wide range of query types. See the [Entrez Programming Utilities help documentation](#) for more information.

dbSNP FTP Site

NCBI supports the public distribution of dbSNP data by providing compressed data dumps in a number of different formats. Access to the NCBI FTP site is available via the World Wide Web (<ftp://ftp.ncbi.nih.gov/snp/>) or anonymous FTP (host `ftp.ncbi.nih.gov` `cd.snp`). In addition to the data formats described on the [FTP README file](#), which include ASN.1, FASTA, and XML, dbSNP FTP offers two additional formats:

VCF Format

The Variant Call Format, or VCF, was developed for the [1000 Genomes Project](#) as a standardized format for storing large quantities of sequence variation data (SNPs, indels, larger structural variants, etc.) and any accompanying genotype data and annotation. A VCF file contains a header section and a data table section. Since the metadata lines in the header section can be altered to fit the requirements of the data to be submitted, you can use VCF to submit many different kinds of common variations (as well as their associated genotypes and annotation) that are contained within one reference sequence. VCF files are compressed (using bgzip) and are easily accessed. See Danecek, et. al. for a concise overview of VCF (3), and the official 1000 Genomes site for a [detailed description of the VCF format](#). Submissions to dbSNP currently use VCF format [version 4.1](#).

BED Format

The Browser Extensible Data (BED) format was developed by [UCSC Genome Bioinformatics](#) as a means of displaying data lines for genome browser annotation tracks. Each line of the BED format represents a single annotated feature that is described using required and optional fields. dbSNP BED files are derived from dbSNP RS Docsum ASN.1 (ftp://ftp.ncbi.nih.gov/snp/specs/docsum_3.4.xsd) and use the three required fields found in the [standard BED format](#) as well as three of the nine optional fields (name, score, strand). The dbSNP BED format has been QA tested and is compatible with standard BED tools and genome browser uploads such as the NCBI Remap Service (<http://www.ncbi.nlm.nih.gov/genome/tools/remap>), the UCSC Genome browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>), and the EBI Genome Browser (<http://www.ensembl.org>).

ADA Section 508-Compliance Link

All links provided on the dbSNP homepage are also provided in text format at the bottom of the page to support browsing by text-based Web browsers. Suggestions for improving database access by disabled persons should be sent to the dbSNP development group at snp-admin@ncbi.nlm.nih.gov.

Local Copies of dbSNP

If you wish to create a SQL copy of dbSNP on a local server for direct access, use the directions in Appendix 3 of this chapter to create the tables and indices for dbSNP from the dbSNP schema, data, and SQL statements.

Note: We will be phasing out the relational database architecture of dbSNP during the dbSNP redesign, and are considering replacing it with Service Oriented Architecture (SOA) and a BLOB/CLOB store system in dbSNP 2.0. Storage technology and object schemas, however, are still under design. Since dbSNP 2.0 may not be an SQL based system, we will provide users with an API to access bulk dumps of data for those wanting to create a local copy of dbSNP. Check or subscribe to the [dbSNP News and Announcements](#) site for updates regarding the redesign and availability of the data as relational tables or as objects.

Related Tools and Studies

There are multiple tools related to processing or learning more about short sequence variations. These are described in depth in the variation overview section of the Handbook. In brief, they support the following use cases:

Converting a Location on One Assembly or Sequence to Another

[NCBI's Genome Remapping Service](#) (Remap) allows you to convert locations from one sequence to another based on alignments. Use Remap if you have identified the location

of variation on an assembly, or on a RefSeqGene/LRG, and want to determine the location on a different assembly (or on the genome in the case of the RefSeqGene).

History of Interpretation of the Medical Importance of an Allele

ClinVar archives the relationship reported between variations and phenotype by accessioning and versioning submissions.

Association Studies

dbGaP archives and distributes data from studies that examine the relationship between phenotype and genotype. Such studies include Genomewide Association Studies (GWAS), medical sequencing, and molecular diagnostic assays. Links are available from dbGaP controlled access records to related variation data in dbSNP, but there are no reciprocal links from dbSNP records to dbGaP unless the aggregate data are public. The refSNP report “Association” section will link to association results from the [NHGRI GWAS Catalog](#) and/or [PheGenI](#) when association data are available.

Histocompatibility

dbMHC provides a platform where users can access, submit, and edit data related to the human Major Histocompatibility Complex, also called the HLA (Human Leukocyte Antigen).

Both dbMHC and dbSNP store the underlying variation data that define specific HLA alleles. dbMHC provides access to related dbSNP records at the haplotype and variation level, whereas dbSNP provides access to related dbMHC records at the haplotype level.

Haplotypes

The [International HapMap Project](#) site allows access to its catalog of statistically related variations, also known as haplotypes, for a number of different human populations, and is a useful resource for those researchers looking for variations associated with a particular gene. HapMap haplotypes can be searched by a landmark such as a refSNP number or gene symbol, as well as by sequence region or chromosome region. The resulting HapMap report includes an ideogram with various tracks that can be altered to provide required data, and appropriate tracks in the report will provide direct links to refSNP cluster records.

Variation as Related to Citations, Genes, Phenotypes, and other NCBI Databases

Multiple databases in NCBI can be used to identify variation that meets certain criteria. They may either reference rs numbers explicitly, or provide links from their records to records in dbSNP.

Variation Batch Submission (VarBatch)

VarBatch is an online submission resource for both clinical and non-clinical human variations, and allows the update and annotation of previously submitted variations. When an asserted clinical variation is processed through VarBatch, it is assigned both a dbSNP submitted SNP (ss) accession as well as a ClinVar accession (format: SCV000000000.0), since the ClinVar accession represents the asserted variation/phenotype relationship.

Note: Since VarBatch does not accept frequency, genotype, or population data, submit these data to dbSNP as updates to your VarBatch submission using the dbSNP VCF or Flat File format via email or through a pre-arranged FTP upload once ss numbers are assigned to your submitted variations.

Variation Reporter

Variation Reporter matches submitted variation call data to variants housed in dbSNP or dbVAR, allowing access to all data and metadata that dbSNP has for any known matching variants. If you submit novel variants to the Variation Reporter, and there are no matches between your data variants housed in dbSNP or dbVAR, the Variation Reporter will provide the predicted consequence of each submitted variant.

VarView

VarView reports display detailed variation information associated with a particular gene and are created only for those genes that have asserted clinical variations. VarView can be accessed in two ways:

1. Through Gene by using the query “*gene.snp_clin[filter]*” to identify gene records that have a VarView report.
2. Through dbSNP either by using the “VarView” link  displayed in refSNP reports for variations that have asserted clinical significance, or by querying dbSNP using “*snp_gene_clin[filter]*” to identify variants that have a VarView report.

Once a Gene or dbSNP record has been selected, and the VarView link on the record has been activated, a VarView report will appear that includes:

- A brief description the gene
- A list of all observed rs variants of the gene
- Links to both internal and external resources including locus specific databases (LSDB), OMIM, Gene, and PubMed.

When one of the listed rs variations in the VarView report is selected, the “submission details” section of the report provides a list of ss numbers associated with the selected rs number as well as links to submitter sites and each ss report.

Note: VarView will be replaced by a new Variation Gene Viewer in April 2014. This new resource will allow users to access all of NCBI's variation data (i.e., dbSNP, dbVar, ClinVar) in a gene-centric fashion.

1000 Genomes Browser

The [1000 Genomes Browser](#) provides access to 1000 Genomes data including variations, genotypes, and sequence read alignments within the context of GRCh37, the reference assembly used by the 1000 Genomes Project for analysis. The browser allows you to configure the display to include multiple data tracks of interest and provides links to related data housed in various NCBI resources. The 1000 Genomes Browser allows users to quickly view alignments supporting a particular variant call and can be used to download and read variant data for small genomic regions of interest.

Access the 1000 Genomes Browser from dbSNP using the 1000 Genomes Browser link  in the refSNP report “Integrated Maps” section.

Access dbSNP from the 1000 Genomes Browser using the “hover” feature in either the “Clinical Channel” or “Cited Variant” tracks. Click on the variation rsID that appears.

References

1. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum Mutat.* 2010 Jan;31(1):67–73. PubMed PMID: 19877174.
2. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tselenko A, Sampas N, Bruhn L, Shendure J. 1000 Genomes Project, Eichler EE. Diversity of human copy number variation and multicopy genes. *Science.* 2010 Oct 29;330(6004):641–6. PubMed PMID: 21030649.
3. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics.* 2011 Aug 1;27(15): 2156–8. PubMed PMID: 21653522.

Appendices

Appendix 1. dbSNP Report Formats

ASN.1

The [docsum_3.4.asn](#) file is the ASN structure definition file for ASN.1 and is located in the `/specs` subdirectory of the dbSNP FTP site. The [00readme file](#), located in the main dbSNP FTP directory, provides information about ASN.1 data structure and data exchange. ASN.1 text or binary output can be converted into one or more of the following formats: Flat File, FASTA, DocSum, Chromosome Report, RS/SS, and XML.

Note: ASN.1 data must be retrieved programmatically by using [eUtils](#) or by using the [dbSNP Batch Query Service](#).

BED

The Browser Extensible Data (BED) format was developed by [UCSC Genome Bioinformatics](#) as a means of displaying data lines for genome browser annotation tracks.

Each line of the BED format represents a single annotated feature that is described using required and optional fields. dbSNP BED files are derived from dbSNP RS DocSum ASN.1 (ftp://ftp.ncbi.nih.gov/snp/specs/docsum_3.4.xsd) and use the three required fields found in the [standard BED format](#) as well as three of the nine optional fields (name, score, strand).

The dbSNP BED format has been QA tested and is compatible with standard BED tools and genome browser uploads such as the NCBI Remap Service (<http://www.ncbi.nlm.nih.gov/genome/tools/remap>), the UCSC Genome browser (<https://genome.ucsc.edu/cgi-bin/hgGateway>), and the EBI Genome Browser (<http://www.ensembl.org>).

Chromosome Report

The Chromosome Reports format provides an ordered list of RefSNPs in approximate chromosome coordinates and contains a great deal of information about each variation. Since the coordinate system used in this format is the same as that used for the NCBI Genome Map Viewer, Chromosome Reports contains information helpful in the identification of variations that can be used as markers.

A full description of the information provided in the Chromosome Reports format is available in the [00readme](#) file, located in the SNP directory of the [SNP FTP](#) site.

Note: A Chromosome Report's directory may contain any of the following files:

- **chr_AltOnly.txt.gz:** List of variations that map to a non-reference (alternate) assembly (e.g., a human refSNP maps to HuRef or TCAGChr7, but not to GRC)
- **chr_MT.txt.gz:** List of variations that map to the mitochondria chr_Multi.txt.gz: List of variations that map to multiple chromosomes
- **chr_NotOn.txt.gz:** List of variations that did not map to any chromosomes
- **chr_PAR.txt.gz:** List of variations on the pseudoautosomal regions of human or great ape X and Y chromosomes.
- **chr_UN.txt.gz:** List of mapped variations that are on unplaced chromosomes

FASTA: ss and rs

The FASTA report format provides the flanking sequence for each report of variation in dbSNP, as well as for all the submitted sequences that have a report of “no variation.” The FASTA data format is typically used for sequence comparisons using [BLAST](#).

Online BLAST is useful for conducting a few sequence comparisons in the FASTA format, whereas multiple FASTA sequence comparisons require the installation of a local stand-alone version of BLAST, and the construction of a local database of FASTA formatted data.

A full description of the information provided in the FASTA report format is available in the [00readme](#) file, located in the SNP directory of the [SNP FTP](#) site.

Gene Report

The dbSNP Gene report is a text report that provides a list of all refSNPs currently known to be located in a particular gene, as well as a summary of general and clinical information for each listed variation. The file naming convention for gene_report is “XXXXX_gene_report.txt.gz,” where “XXXX” represents the gene symbol (e.g., LPL, the gene symbol for lipoprotein lipase).

A full description of the information provided in the gene_report format is available in the [00Gene_report_format_README](#), located in the human [gene_report directory](#) of the [SNP FTP](#) site.

Genotype Report

Since the massive amount of genotype data we receive from large sequencing projects (e.g., 1000 Genomes) makes it difficult for NCBI to maintain and query the dbSNP SQL tables, we will no longer provide genotype data or reports.

NCBI is currently developing a new service (Genotype Server) that will more efficiently store and serve genotype and frequency data using API, the internet, and FTP. It should be available sometime in 2014.

The genotype XML, on the [dbSNP FTP server](#), is still available and provides submitter and genotype information for many submitted SNPs. It is organized in chromosome specific files under each organism directory in the “genotype subdirectories” (e.g., human genotype XML files are located in ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/genotype/). Users should be aware, however, that the genotype XML is also in the process of being phased out.

Note: Until NCBI’s new Genotype Server is released, genotype data can be queried and downloaded at these two alternative sites:

1000 Genomes: <http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>

HapMap: <http://hapmap.ncbi.nlm.nih.gov/>

rs docsum flatfile

The rs docsum flatfile report is generated from the ASN.1 datafiles and is provided in the files whose naming convention is "/ASN1_flat/ds_flat_chXX.flat". Files are generated per chromosome (chXX in file name), as with all of the large report dumps.

Because flatfile reports are compact, they will not provide you with as much information as the ASN.1 binary report, but are useful for manual scanning of human SNP data because they provide detailed information at a glance.

A full description of the information provided in the rs docsum flatfile format is available in the 00readme file, located in the SNP directory of the [SNP FTP](#) site.

VCF

The Variant Call Format, or VCF, was developed for the [1000 Genomes Project](#) as a standardized format for storing large quantities of sequence variation data (SNPs, indels, larger structural variants, etc.) and any accompanying genotype data and annotation.

A VCF file contains a header section and a data table section. Since the metadata lines in the header section can be altered to fit the requirements of the data to be submitted, you can use VCF to submit many different kinds of common variations (as well as their associated genotypes and annotation) that are contained within one reference sequence. VCF files are compressed (using bgzip), and are easily accessed.

See Danecek, et. al. for a concise overview of VCF (3) and the official 1000 Genomes site for a [detailed description of the VCF format](#). Submissions to dbSNP currently use VCF format version 4.1.

XML

The XML format provides query-specific information about refSNP clusters, as well as cluster members in the NCBI SNP Exchange (NSE) format. The XML schema is located in the [docsum_3.4.xsd](#) file, which is housed in the [/specs](#) subdirectory of the dbSNP FTP site. A human-readable text form of the NSE definitions can be found in [docsum_3.4.asn](#), also located in the [/specs](#) subdirectory of the dbSNP FTP site.

Note: XML data must be retrieved programmatically by using [eUtils](#) or by using the dbSNP Batch Query Service.

Appendix 2. Rules and Methodology for Mapping

The appearance of FASTA-formatted genome sequence for a new build of an assembly or the significant accrual of newly submitted SNP data for an organism will initiate a cycle of MegaBLAST and BLAST alignment of the variations to the NCBI genome assembly of the organism.

Variation Placement by Remapping

dbSNP uses sequence alignments to map asserted locations and underlying features on to reference sequences. During the build process, dbSNP performs three types of remapping: Mapping up, Mapping Down, and Assembly to Assembly remapping.

Mapping Up

“Mapping Up” refers to the process of mapping a submitted variation whose location is based on a reference sequence, cDNA, or protein to the current genome build and to RefSeqGene using sequence alignments.

Mapping up from cDNA to Genomic Sequence

If the provided location is in an exon, dbSNP maps the input coordinates directly to the genome through available alignments. If the provided location is in an intron, dbSNP maps the exon boundary coordinate that is closest to the intron position, again using available alignments.

Mapping up from Protein to cDNA

dbSNP aligns the protein accession and location as well as the asserted location of the variation on the protein to cDNA. This alignment generates up to three possible sequence locations of the variation at the nucleotide level, where it is possible to discern the stated variation at the protein level.

Mapping Down

“Mapping Down” refers to the process of using genomic alignments to map information found on a genomic sequence to transcript sequences and proteins.

Assembly to Assembly Remapping

Assembly-Assembly remapping allows the projection of features from one assembly coordinate system to another using genomic alignments. dbSNP performs a base by base analysis of each feature on the source sequence in order to project the feature through the alignment to the new sequence.

Variation Placement by BLAST

When an asserted location for a submitted variant is not available, dbSNP will attempt to place the variation on the genome by BLASTing submitted variation flanking sequences against a genomic assembly. This mapping process is a multi-step, computer-based procedure that begins when refSNP and submitted SNP FASTA sets are aligned to the most recent genome assembly using BLAST or MegaBLAST. The quality of each alignment is determined using an Alignment Profiling Function.

The BLAST/MegaBLAST output of the ASN.1 binary files of local alignments is analyzed by an algorithm to create a group of local alignments that lay close to one another on a sequence. If the global alignment is greater than or equal to a pre-determined percentage of the flanking sequence, it is accepted as a true alignment between the refSNP or submitted SNP and the genome assembly.

This group of close local alignments is then processed to define alleles and LOC types for each hit and to establish the hit location. The output is filtered to remove paralogous hits and to select those variations that have the greatest degree of alignment to a particular

contig. The output is then placed into a file and processed to create an MD5 positional signature for each variation. These signatures are then placed in the SNP MAP INFO file and loaded into dbSNP.

Once all the results from previous steps are loaded into dbSNP, dbSNP looks for clustering candidates. If an MD5 signature for a particular SNP is different from the MD5 signature of another SNP, then each SNP will have a unique hit pattern and need not be clustered. If an MD5 signature of a particular SNP is the same as that of another SNP, the two SNPs may have the same hit pattern, and if after further analysis, the hit patterns are shown to be the same, the two SNPs will be clustered.

Appendix 3. How to Create a Local Copy of dbSNP

How to Create a Local Copy of dbSNP

Currently, dbSNP is a relational database that contains hundreds of tables. Since the inception of build 125, the design dbSNP has been altered to a "hub and spoke" model, where the dbSNP_Main_Table acts as the hub of a wheel, storing all of the central tables of the database, while each spoke of the wheel is an organism-specific database that contains the latest data for a specific organism. dbSNP exports the full contents of the database for the public to download from the dbSNP [FTP](#) site. During the dbSNP redesign, however, we will be phasing out the relational database architecture of dbSNP, and are considering replacing it with Service Oriented Architecture (SOA) and a BLOB/CLOB store system in dbSNP 2.0.

Due to security concerns and vendor endorsement issues, dbSNP cannot provide users with direct dumps of dbSNP. The task of creating a local copy of dbSNP can be complicated and should be left to an experienced programmer. The following sections will guide you in the process of creating a local copy of dbSNP, but these instructions assume knowledge of relational databases, and were not written with the novice in mind.

If you have problems establishing a local copy of dbSNP, please contact dbSNP at snp-admin@ncbi.nlm.nih.gov.

Schema: The dbSNP Physical Model

A schema is a necessary part of constructing your own copy of dbSNP because it is a visual representation of dbSNP that shows the logical relationship between the data. It is available as a printable PDF [file](#) from the dbSNP FTP site.

Data in dbSNP are organized into "subject areas" depending on the nature of the data. The [data dictionary](#) includes a description of the tables in dbSNP as well as tables of columns and their properties. Foreign keys are not enforced in the physical model because they make it harder to load table data asynchronously. In the future, dbSNP will add descriptions of individual columns. The [data dictionary](#) is also available online from the dbSNP website.

Resources Required for Creating a Local Copy of dbSNP

Software:

- **Relational database software.** If you are planning to create a local copy of dbSNP, you must first have a relational database server, such as Sybase, Microsoft SQL server, or Oracle. dbSNP at NCBI runs on an MSSQL server version 2000, but there are users who have successfully created their local copy of dbSNP on Oracle.
- **Data loading tool.** Loading data from the dbSNP FTP site into a database requires a bulk data-loading tool, which usually comes with a database installation. An example of such a tool is the bcp (bulk-copy) utility that comes with Sybase, or the “bulkinser” command in the MSSQL server.
- **winzip/gzip to decompress FTP files.** Complete instructions on how to uncompress *.gz and *.Z files can be found on the dbSNP [FTP](#) site.

Hardware:

- **Computer platforms/OS**

Databases can be maintained on any PC, Mac, or UNIX with an internet connection.

- **Disk space**

To ascertain the disk space needed for a complete copy of dbSNP for a particular organism, determine the total download file size for the organism as a starting point. You need a minimum of three times of the data file size to have space for creating indices and storing your own working tables. The allocated size of dbSNP human B137 on dbSNP’s internal server is 3TB, while mouse B137 size is about 700GB.

- **Memory**

The minimum amount of memory required is approximately **4GB**.

- **Internet connection**

dbSNP recommends a high-speed connection to download such large database files.

dbSNP Data Location

The [FTP database directory](#) in the dbSNP FTP site contains the schema, data, and SQL statements to create the tables and indices for dbSNP:

- The [shared_schema](#) subdirectory contains the schema DDL (SQL Data Definition Language) for the dbSNP_main_table.
- The [shared_data](#) subdirectory contains data housed in the dbSNP_main_table that is shared by all organisms.
- The [organism_schema](#) subdirectory contains links to the schema DDL for each organism specific database.
- The [organism_data](#) subdirectory contains links to the data housed in each organism specific database. The data organized in tables, where there is one file per table. The file name convention is: <tablename>.bcp.gz. The file name convention for the

mapping table also includes the dbSNP build ID number and the NCBI genome build ID number. For example, B125_SNPContigLoc_35_1 means that during dbSNP build 125, this SNPContigLoc table has SNPs mapped to NCBI contig build 35 version 1. The data files have one line per table row. Fields of data within each file are tab delimited.

dbSNP uses standard SQL DDL(Data Definition Language) to create tables, views for those tables, and indexes. There are many utilities available to generate table/index creation statements from a database.

Hint

If your firewall blocks passive FTP, you might get an error message that reads: "Passive mode refused. Turning off passive mode. No control connection for command: No such file or directory." If this happens, try using a "smart" FTP client like NCFTP (available on most UNIX machines). Smart FTP clients are better at auto-negotiating active/passive FTP connections than are older FTP clients (e.g., Sun Solaris FTP).

Stepwise Procedure for Creating a Local Copy of dbSNP

1. Prepare the local area

(check available space, etc.)

2. Download the schema files

- a. Download the following files from the dbSNP [shared_schema](#) subdirectory: dbSNP_main_table, dbSNP_main_index_constraint, and all the files in the [shared_data](#) subdirectory. Together, the files from both of these subdirectories will allow you to create tables and indices for the dbSNP_main_table.
- b. Go to the [organism_schema](#) subdirectory and select the organism for which you wish to create a database. For the purpose of this example, human_9606 has been selected. Once human_9606 is selected, you will be directed to the [human organism_schema](#) subdirectory. Download all of the files contained in this subdirectory.
- c. Go to the [organism_data](#) subdirectory, and select the organism for which you wish to create a database. For the purpose of this example, human_9606 has been selected. Once you select human_9606, you will be directed to the [human organism_data](#) subdirectory. Download all of the files contained in this subdirectory.

A user must always download the files located in the most recent versions of the shared_schema and shared_data subdirectories in addition to any organism specific content.

Save all the files in your local directory and decompress them.

Hint:

On a UNIX operating system, use gunzip to decompress the files: dbSNP_main_table and dbSNP_main_index_constraint.

The files on the SNP FTP site are UNIX files. UNIX, MS-DOS, and Macintosh text files use different characters to indicate a new line. Load the appropriate new line conversion program for your system before using bcp.

3. Create the dbSNP_main_table

- a. From the [shared_schema](#) subdirectory, use the dbSNP_main_table file to create tables, and use the dbSNP_main_index_constraint files to create indices for the dbSNP main database.
- b. Load all of the bcp files located in the [shared_data](#) subdirectory into the dbSNP_main_table you just created using the data-loading tool of your database server (e.g., bcp for Sybase). See the sample FTP protocol and sample Unix C Shell script (below) for directions.
- c. Create indices by opening the dbSNP_main_index_constraint.sql file. If you are using a database server that provides the isql utility, then use the following command:

```
isql -S <servername> -U username -P password -i  
dbSNP_main_index_constraint.sql
```

Hint:

The “bcp” files in the shared_data and organism_data subdirectories may be loaded into most spreadsheet programs by setting the field delimiter character to “tab”.

4. Create the organism specific database

Once the dbSNP_main_table has been created, create the organism specific database using the files in your specific organism’s organism_schema and organism_data subdirectories. Human_9606 will be used for the purpose of this example:

- a. Create the human_9606 database using the following files found in the human_9606 [organism_schema](#): human_9606_table.sql.gz, human_9606_view.sql.gz, human_9606_index_constraint.sql.gz, and human_9606_foreign_key.sql.gz
- b. Load all of the bcp files located in the [shared_data](#) subdirectory into the human_9606 database you just created using the data-loading tool of your database server (e.g., bcp for Sybase). See the sample FTP protocol and sample Unix C shell script (below) for directions.

Hint:

Use “ftp -i” to turn off interactive prompting during multiple file transfers to avoid having to hit “yes” to confirm transfer hundreds of times.

Hint:

To avoid an overflow of your transaction log while using the bcp command option (available in Sybase and SQL servers), select the "batch mode" by using the command option: -b number of rows. For example, the command option -b 10000 will cause a commit to the table every 10,000 rows.

5. Sample FTP Loading protocol

- a. Type ftp -i ftp.ncbi.nih.gov (Use "anonymous" as user name and your email as your password).
- b. Type: cd.snp/database
- c. To get dbSNP_main for shared tables and shared data: Type ls to see if you are in the directory with the right files. Then type "cd shared_schema" to get schema file for dbSNP_main, and finally, type "cd shared_data" to get the data for dbSNP_main.
- d. Type binary (to set binary transfer mode).
- e. Type mget *.gz (to initiate transfer). Depending on the speed of the connection, this may take hours since the total transfer size is gigabytes in size and growing.
- f. To decompress the *.gz files, type gunzip *.gz. (Currently, the total size of the uncompressed bcp files is over 10 GB).

6. Use scripts to automate data loading.

- a. Located in the [loadscript](#) subdirectory of the dbSNP FTP site, there is a file called cmd.create_local_dbSNP.txt that provides a sample UNIX C shell script for creating a local copy of dbSNP_main and a local copy of a specific organism database using files in the shared_schema, and the organism_schema subdirectories.
- b. Also in the [loadscript](#) subdirectory of the dbSNP FTP site, there is a file called cmd.bulkinsert.txt that provides a sample UNIX C shell script for loading tables with files located in shared_data and organism_data subdirectories.

7. Data integrity (creating a partial local copy of dbSNP)

dbSNP is a relational database. Each table has either a unique index or a primary key. Foreign keys are not reinforced. There are advantages and a disadvantage to this approach. The advantages are that this approach makes it easy to drop and recreate the table using the dbSNP_main_table, which then makes it possible to create a partial local copy of dbSNP. For example, if you are interested only in the original submitted SNP and their population frequencies, and not in their map locations on NCBI genome contigs or GenBank Accession numbers (both are huge tables), then these tables can be skipped (i.e., SNPContigLoc and MapLink). Please remember that mapping tables such as SNPContigLoc will have a build ID prefix and suffix included in its file name. (e.g., SNPContigLoc will be

b125_SNPContigLoc_35_1for SNP build 125, and NCBI contig build 35 version 1). Of course, to select tables for a particular query, the contents of each table and the dbSNP entity relationship (ER) diagram need to be understood. The disadvantage of unreinforced references is that either the stored procedures or the external code needs to be written to ensure the referential integrity.

dbVar

Adrienne Kitts, M.S., Deanna Church, Ph.D., Tim Hefferon, Ph.D., and Lon Phan, Ph.D.

Created: October 26, 2014.

Scope

The Database of Genomic Structural Variation (dbVar) is a National Center for Biotechnology Information (NCBI) archival database that manages sequence variation. dbVar complements dbSNP by archiving copy number variants (CNV), insertions, deletions, inversions, and translocations (1) that are longer than 50 base pairs (bp). The database is organized around the studies that have identified these variants, and includes variations from research-based and clinical submissions. Structural variants that have asserted germline or somatic clinical significance should be submitted to [Clinvar](#) which will forward appropriate portions of the data to dbVar for accessioning.

dbVar and the European Bioinformatics Institute's (EBI's) [Database of Genomic Variants archive](#) (DGVa) share the same data model and exchange data regularly. Together they represent the largest and most comprehensive archive of structural variation in the world.

Structural variation can be detected with a variety of experimental methods. Most of the data dbVar receives have been generated by Next-Generation Sequencing (NGS) or microarray (either oligo or SNP array) technologies. These methods alone can detect a wide variety of variant types, but they vary in the degree of precision and certainty they can provide with respect to breakpoint location and copy number change. [A complete list of structural variation detection methods and analysis types](#) can be found in dbVar's online documentation.

dbVar's primary tasks are to support the submission and organization of structural variations to aid researchers in the study of a wide range of biological problems . dbVar assigns stable, traceable identifiers to the structural variants found in each study, calculates locations on newer assemblies as appropriate, provides some validation of content, and integrates the structural variant data submitted to us with a wide array of NCBI tools and data.

Medical Genetics

Advances in next-generation sequencing technologies have allowed researchers to generate massive amounts of sequence data. When clinical samples are sequenced using these technologies, novel structural variants that have causative roles in disease may be identified.

Structural variants have been implicated in complex human disorders that include cancer, neurological diseases, and developmental disabilities such as Down, Turner, and Prader-Willi Syndromes, as well as increased susceptibility to disorders including HIV, Crohn

disease, and lupus(4). dbVar accepts submissions from disease resources to maximize our representation of complex genetic disorders. One example are the submissions from the International Collaboration for Clinical Genomics (ICCG), which is part of ClinGen, a larger NIH-funded effort.

dbVar manages and organizes structural variation data such as location and variation type from all submitters and provides clinicians and researchers a point of access to both clinical cases and control subjects. dbVar also provides value-added data such as confirmed validation status and current clinical phenotype interpretation through our integrated relationship with ClinVar.

Association Studies

Structural variant submissions that contain sensitive clinical information or do not have informed consent from the originating sample donor are submitted to NCBI's [Database of Genotypes and Phenotypes](#) (dbGaP). Variants are assigned dbGaP accessions, stripped of identifying information, and then submitted to dbVar. dbVar in turn, provides users access to variant location, variant type, and summary variant data stripped of identifying information. All sensitive information contained in the dbGaP submission remains in dbGaP and can only be retrieved through dbGaP's controlled access system.

dbVar's annotated records and catalog of common structural variations can be used to inform the design of Genome-wide Association Studies (GWAS), create variation arrays used in these studies, and interpret GWAS study results.

Evolutionary Biology

Although it is known that structural variation plays an important role in species and disease evolution(5), until recently, the technology required to produce structural variant maps with the degree of resolution needed to trace the evolutionary history of a species or gene has not been available. The advent of high throughput sequencing technologies has made the comparison of structural variation between related species possible (6), so the time when evolutionary analysis of structural variation will be more commonplace is approaching. dbVar currently houses studies that have evolutionary implications, including studies that compare structural variation observed in different breeds of dog (nstd10, nstd13), among the Great Apes (nstd82, [estd193](#)), and structural variant studies that have implications in gene and gene function evolution ([estd199](#), [nstd78](#)). dbVar's variant catalog can be used to inform evolutionary studies through its use in the selection of candidate variants for both species and gene evolutionary studies.

History

Creation and Growth

Following the discovery that healthy human subjects contained copy number variants (CNV) wide spread throughout the genome, the [Database of Genomic Variants](#) (DGV)

was founded in 2004 at the Center for Applied Genomics in Toronto, Canada, to organize, manage, and provide access to the initial data produced by early structural variation studies (7). It was soon discovered that a more comprehensive and permanent archive would be needed to work in conjunction with DGV. The National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) collaboratively launched dbVar and DGVa (Database of Genomic Variants archive) in 2009 to meet this need.

The collaboration between dbVar and DGVa was designed to be close; the two resources communicate and exchange data on a regular basis, and share a uniform data model as well as similar database schemas. dbVar and DGVa together assumed DGV's role in the organization and management of structural variation data and augmented this role by providing increased capacity for structural variant storage as well as integration with a host of other genetic databases and tools.

Database Development Milestones

After its initial launch in 2009, dbVar grew slowly in size and began integrating its data with that of other NCBI resources. By 2012, dbVar underwent its first major improvement with a database schema overhaul conducted in collaboration with DGVa, which vastly improved mutual data exchange capabilities. 2012 was also the year that dbVar released the [dbVar Genome Browser](#), which dramatically improved our users' ability to interpret and analyze dbVar data by allowing a side-by-side comparison of variants from any study in dbVar superimposed on an annotated set of chromosome ideograms from multiple assemblies. By 2013, dbVar saw a great advance in its ability to provide its users with deeper clinical insight when dbVar became fully integrated with [ClinVar](#). The integrated relationship between dbVar and ClinVar provides dbVar users with access to in-depth information about clinical assertions associated with structural variants. With the release of the [Variation Viewer](#) upgrade in 2013, dbVar users could also explore structural variants side by side with small-scale variants (dbSNP) and clinical variations (ClinVar) in a genomic context. dbVar also integrated with NCBI's [Variation Reporter](#) at about the same time, allowing users to submit variation calls to find metadata for known variants, or see the predicted molecular consequence of the call if the submitted variant is novel.

The discovery of structural variants that have multiple breakpoints derived from complex genetic rearrangements spurred dbVar's development of a database model to capture and display complex structural variants associated with cancer and other illnesses. dbVar's ability to capture both simple and complex structural variants was greatly improved in the fall of 2014 when it began accepting dbVar submissions in [Variant Call Format \(VCF\)](#). dbVar's VCF submission specification is very similar to the [1000 Genomes VCF specification v.4.2](#), which allows for annotation of both simple and complex structural variants. dbVar's VCF specification contains modifications that allow our users to submit additional information for their structural variants with ease.

Evolution in Submitted Content

Upon its initial release, dbVar was populated with historical structural variation data that was mined from available research literature. Because the first paper describing the prevalence of genomic structural variation data in normal human subjects was released in 2004 (8), there wasn't very much historic data available, which necessarily meant that the initial population of structural variants in dbVar was limited to human. Initial submissions to dbVar in 2009 included data from human as well as from model species such as *Mus musculus*, and agriculturally important species such as *Bos taurus* and *Sus scrofa*.

Currently, dbVar archives and provides access to 4.3 million variant regions from 127 studies and 16 taxa that include plants, invertebrates, and vertebrates.

Data Model

dbVar stores data in a three-level nested hierarchy:

The topmost level in the dbVar nested hierarchy is called the Study (**std**), which can be either a particular publication that produced a set of data, or a community resource that submits new data on a regular basis. Regardless of which it is, a study is a record that serves to indicate a group of data (sv and their supporting ssv) generated in a particular series of experiments, and provide general information about those experiments.

The next level down in the hierarchy is called a Variant Region (**sv**: structural variant), which is formed when multiple Variant Calls (ssv) in a particular region from one study are grouped together under the same identifier.

Note: dbVar does not currently integrate data across studies, so what appears to be the same region will not have the same sv identifier if those regions were identified in different studies (std).

The base level of dbVar's nested data storage hierarchy is called a "Variant Call" or **ssv** (supporting structural variant). Variant Calls are the individual variant placements that contain the actual data used to place submitted structural variants.

Accessions

If the Study, the Variant Region, or Variant Call was originally submitted to dbVar at NCBI, the accession numbers are given an "n" prefix (i.e., nstd, nsbv, and nssv, respectively). If the Study, Variant Region or Variant Call was originally submitted to DGVa at EBI, the accession numbers are given an "e" prefix (i.e., estd, esbv, and essv).

The quality of the data in dbVar depends on the quality of the data submitted to us, so we provide the following consistency checks:

- dbVar performs data validation during submission processing that will catch certain types of errors, such as inconsistent data, invalid placements, or invalid

entries. Serious errors will cause the validation processes to stop submission loading, and dbVar will determine at that point whether it is necessary to contact the submitter for corrections.

- If a variant submission has a noticeably incorrect placement (e.g., coordinates located at the end of a chromosome or within an assembly gap), we will return the submission and ask the submitter to check and/or correct the location.

Since dbVar is minimally curated and reviews submission for obvious errors only, it is important that all submissions to dbVar contain high quality data, and the responsibility for maintaining data quality rests ultimately with our submitters.

Note: 1000Genomes records in dbVar now contain a data quality indicator.

Study (std, nstd, or estd)

A dbVar Study is a record that serves to group together Variant Region (sv) and Variant Call (ssv) data with descriptive metadata including organism, study type, submitter, project, and any associated publications.

Although the fields that characterize a study rarely change, the Variant Regions and Variant Calls in studies submitted by an ongoing project such as 1000Genomes or ICCG (International Collaboration for Clinical Genomics) may change if the submitter updates a Variant Region or submits new Variant Calls.

Study Submitted Content

Content for a study includes general information including author contact information, study identifiers, a brief description of the study, the study type (e.g., control set, case-control, matched-normal, etc.), associated IDs from PubMed, Taxonomy, the Entrez Genome Project, and dbGaP, as well as a description and identifiers for all samples and sample sets used in the study. Complete information regarding methods and analyses used for variant discovery must also be included in a study submission as well as any validation data generated for the experiment.

The Study is a dbVar user's entry point to structural variation data, which is why the primary search mode of dbVar is the “[Study Browser](#)”. Because data contained within a Study record are found at the same time by the same authors in the same laboratory (or laboratories) using a specific set of methods and analyses, you can use any of these characteristics as search terms in the dbVar Study Browser to find studies, variant regions, and variant calls whose metadata contain the entered search term.

Variant Region (sv, nsv, or esv)

A Variant Region (sv) record is composed of Variant Calls (ssv) located at or near the same location that have been grouped together, or “merged”—either by the submitter at his or her discretion, or by dbVar in the case of calls with identical coordinates and call types. You can consider the Variant Region record as a parent to the grouped ssv records

that are its children in the dbVar nested record hierarchy. It should be noted that if two Variant Regions from a study overlap, the submitter can merge them into a single Variant Region. Each submitted Variant Region is assigned a unique sv (structural variant) ID.

[Variant Region records](#) provide variant location and placement information, the evidence used to place the variant, available validation and clinical assertion data, as well as links to associated publications.

Reference Variants

As stated in the previous section, a Variant Region, like a refSNP, is a marker on the genome that denotes a group of variants found in a specific region, but this is where the similarity ends. Variant Regions group variations of different types that may occur in the same position, but because of breakpoint uncertainty in the identification of variation boundaries, the variations may also occur scattered throughout the same identified genomic span. Because an exact location for each structural variant cannot be ascertained, true reference variants cannot be established at this time.

We will move closer to the identification of reference variants as sequencing continues, the comparison of genomes continues to expand, and current breakpoint ranges are narrowed down until they are much more defined.

Submitted Content

Merging Variant Calls into Variant Regions

Consolidation of similar or identical Variant Calls into a Variant Region, or merging overlapping Variant Regions into each other can be performed during submission of structural variants to dbVar, but submitter-performed merging is completely optional.

dbVar will merge those Variant Calls submitted without merge data with other Variant Calls at the same location and of the same type and will assign a Variant Region ID during submission processing. Likewise, dbVar will merge overlapping Variant Regions into each other if a submitter chooses not do so during submission.

Linked Variant Call / Linked Variant Region IDs

Submitters who choose to merge their Variant Calls or Variant Regions will need to provide a list of IDs for previously submitted Variant Calls or a list of Variant Region IDs that are to be merged. As an alternative, in the case of a Variant Call merge, submitters can provide the ID of the Variant Region parent in the Variant Call portion of the submission, while in the case of a Variant Region merge, submitters can provide the ID of the parent Variant Region instead of a list of the merging Variant Region IDs.

Assertion Method

Submitters who choose to merge Variant Calls or Variant Regions will need to provide details of the method they used to assert that a group of Variant Calls or Variant Regions require merging. Submitters can provide any algorithms they used to establish that the

Variant Calls or Variant Regions require merging, or they can provide a simple statement such as “reciprocal 50% overlap”, “Calls with identical coordinates merged”, or “Region is identical to call, no merge performed”.

Note: Since dbVar is an archive, we do not validate submitted assertion methods, and as such, we rely on our submitters to use reliable assertion methods for merging groups of Variant calls into Variant Regions, or merging Variant Regions together.

Genotype data

dbVar has begun accepting genotype data in [VCF format](#), and is already in receipt of Genotype data from the 1000Genomes Project submitted via VCF. dbVar will soon begin accepting genotypes in other formats via an updated submission template. If you have questions about submitting structural variation genotype data to dbVar, please contact us at: dbvar@ncbi.nlm.nih.gov

Variant Call (ssv)

A Variant Call record represents an independent instance of a variation produced by an experiment as well as its subsequent analysis. It includes data indicating the location, type, and size of a detected structural variant. Each submitted Variant Call is assigned an ssv (supporting structural variant) ID. If the variant call was submitted to dbVar at [NCBI](#), the ssv ID is given the prefix “nssv”. If the variant call was submitted to DGVa at [EBI](#), the ssv ID is given the prefix “essv”.

Variant Calls are comparable in nature to alleles, but depending on the particular experiment, a variant may or may not actually be an allele. For example, an experiment may yield a variant call that is an allele, but a second, different analysis may yield a completely different call for that variant. In such a case, the variant isn’t actually an allele—it is an artifact. Therefore, given this possible difference in analytical outcome, if a call was generated from a pooled sample, it may or may not be an allele.

Sequence Requirements

dbVar requires that all Variation Calls must be made on an assembly sequence that has already been submitted to an International Nucleotide Sequence Database Collaboration ([INSDC](#)) database, which includes Genbank, the European Nucleotide Archive, or the DNA Database of Japan (DDBJ).

Variant Call Boundaries

Structural variation can be difficult to represent because current structural variation detection technologies seldom provide the base pair resolution necessary to determine variant breakpoints. This introduces an element of uncertainty into the identification of breakpoint boundaries. The extent of this uncertainty depends on the experimental methods that were used to detect the variant, which in turn influences what data submitters will provide to us:

- Detection methods such as arrayCGH and SNP array produce only a range of coordinates within which the breakpoints likely occur, so the submitter can define a minimal region that is definitely involved in the variation, but cannot define precise breakpoints.
- Detection methods such as Paired-End Mapping and Optical Mapping will produce just the precise location for the outer boundaries between which the variant breakpoints must fall, so the submitter can define the region of the genome known to contain the variant, but not the exact location of the variant or its breakpoints.
- Detection methods such as long read sequencing technology or 2nd generation sequencing reads may or may not provide break point resolution. In those cases where breakpoint resolution is achieved, the submitter provides the breakpoint coordinates. In those cases where sequence detection does not give precise breakpoints, the submitter can provide a range of breakpoint coordinates.

When structural variants are submitted to dbVar, we ask the submitter to provide a specific set of data that will capture all the available information we need—including the degree of breakpoint uncertainty present— regardless of the detection method used to find the variant.

Validation

Because dbVar is an archive, we report variants as they are submitted to us and accept (but do not require) validation data used to confirm variant calls. To be considered “validated”, a variant must be confirmed as valid by one or more separate methods. The number of calls validated for a variant region, validation methods and analysis will be part of the Study (std) page, while the Variant Region (sv) and Variant Call (ssv) records will contain summary validation data.

Dataflow

New Submissions and the Start of a New Release

Submissions to dbVar are currently accepted via email to dbvar@ncbi.nlm.nih.gov, and will eventually be accepted through a direct upload using the [NCBI Variation Submission Portal](#), which will allow submitters to track the progress of their submissions and will allow for direct communication between dbVar and the submitter should an error be found during submission processing.

Most data submitted to dbVar are data associated with a recently published study, or a study associated with a publication currently in review. Submission updates to existing studies are generally limited to large ongoing studies like 1000Genomes.

Data Conversion

Data submitted to dbVar is received in the form of an Excel spreadsheet and is converted from Excel to dbVar’s XML format. dbVar’s converter software first converts the

Submitted Excel spreadsheet into a series of tab separated, text-based files, and then during a subsequent step, the text files are converted into dbVar XML submission files.

The dbVar data converter contains a series of validation steps that scan the data for errors during each step in the conversion process. If the validation processes finds minor errors during either conversion step (e.g., data is not in the right form, etc.), the dbVar submissions team will correct the error in the original submission file and put it through the conversion process again.

If an error found during the conversion process is more complex (e.g., coordinates that extend beyond the length of the chromosome), then a member of the dbVar team will contact the submitter, explain the issue and ask the submitter to fix it and resubmit. When the corrected data are received from the submitter, dbVar loads the corrected data through the converter process again.

This process is repeated until the conversion process no longer generates error messages for the submission.

Data Testing

Once the submission has been successfully converted to dbVar XML submission files, the files are loaded into a test version of dbVar. The loading process itself has its own set of validations, which scan the data for errors. If the detected errors are simple, the dbVar submission team will correct them, and if the errors are complex, a member of the dbVar submission team will contact the submitter, explain the issue and ask the submitter to fix it and resubmit.

Once the submission loads to the dbVar test site successfully, the data are reviewed to verify that the data is appropriate to the submission and that the data are being displayed correctly. If the data as shown on the test site is incorrect or does not display properly, dbVar will troubleshoot any difficulties with the data.

It should be noted that we are in the process of shifting the validation processes so that all validations will take place during the conversion process.

Merging Calls into Regions

Each submission to dbVar will contain calls (nssv) which the submitter may or may not have grouped into regions (nsv) since the region portion of the submission is optional. If the submitter decides to create and submit a region or regions, dbVar will check the method used to create the region to insure that the grouping is accurate.

If the submission does not contain any regions, the converter will automatically merge all calls that are the same type and have the same coordinates.

If a call or calls within the submitted region extend beyond the coordinates defined by the submitted range, an error warning will be generated, and the dbVar staff will review the error. If the merging error is simple, the dbVar team will resolve the issue, but if the

merging error is complex, a member of the dbVar team will contact the submitter, explain the problem and send the submission back to be fixed and resubmitted.

Clinical Assertions

Structural variants with clinical significance are submitted to ClinVar, which will then process and accession the data, and sends the data in the ClinVar XML format to dbVar. dbVar maps the ClinVar XML formatted data to dbVar XML format, validates the data, which then proceeds through the dbVar test site load process.

It should be noted that the integration between dbVar and ClinVar is still relatively new, so dataflow between these two resources is still in process and may change.

Association Studies

Structural variants contained within an association study are submitted to dbGaP. Before sending structural variant data from an association study to dbVar, dbGaP removes any sensitive data, and sends the data as tab files to dbVar.

Data Exchange with DGVa

dbVar exchanges data with DGVa every month using an exchange XML format agreed upon by both databases. Because of this shared database schema, the recipient archive usually experiences very few problems loading to their own database, and the exchanged data are available for viewing and download on the recipient's site within a week of their exchange.

Quality Control

If the data displayed on the dbVar test site is approved, the submission is loaded to our quality assurance (QA) database and cross-checked for errors. If the results of the QA testing show no errors, then the data is released to the dbVar public site.

Remapping

dbVar annotates Variant Regions (the non-redundant set of variations) on reference genome genomic sequences, chromosomes, mRNAs, and proteins as part of the NCBI RefSeq project. We then use Assembly-Assembly remapping to project features from the reference assembly coordinate system to selected assembly coordinate systems using genomic alignments. dbVar performs a base-by-base analysis of each feature on the source sequence in order to project the feature through the alignment to the new sequence.

Build Integration

dbVar updates the links between dbVar and BioProject, ClinVar, dbGaP, dbSNP, Gene, HomoloGene, MedGen, Nucleotide, OMIM, Protein, PubMed, PubMed Central, Taxonomy Variation Viewer, and Variation Reporter.

dbVar also maintains links from dbVar records to resources outside of NCBI's Entrez system such as ClinGen, GeneReviews, HPO, and OMIM. Links to all resources related to a particular dbVar record can be found in the upper right-hand corner of the dbVar record under "Links".

Public Release

A dbVar public release involves an update to the public database and the production of a new set of files on the dbVar FTP site. dbVar currently makes an announcement on the dbVar News and Announcements RSS feed and NCBI News (<http://www.ncbi.nlm.nih.gov/news>) when a dbVar release is made publicly available, but intends to eventually move its public release announcements to the ncbi-announce list and will make these announcements on a monthly basis in the future. dbVar announcements are also posted on NCBI's Twitter account and Facebook page to take advantage of the visibility social networking can provide.

Access

dbVar can be queried directly from the search bar at the top of the [dbVar homepage](#), by using the links to dbVar resources and search options located on the homepage (including FTP), or by accessing related NCBI resources that link to dbVar data.

dbVar Home Page

dbVar is a part of the [Entrez](#) integrated information retrieval system and may be searched either by using an ID number query, or by using combinations of different search fields and qualifiers.

Single Record Query

Use the search bar at the top of the [dbVar homepage](#) to find variations using dbVar record identifiers. The record identifiers currently supported for single record queries are the study ID (nstd or estd), the Variant Region ID (nsv or esv) and the variant call ID (nssv or essv).

Complex Entrez Query

Use the [dbVar Advanced Search Builder page](#) to construct a complex search using combinations of different search fields and qualifiers. The Advanced Search Builder allows you to construct a query by selecting multiple search terms from a large number of fields and qualifiers. See the [Advanced Search Builder video tutorial](#) for information about how to find existing values in fields and combine them to achieve a desired result.

Study Browser

The [dbVar Study Browser](#) displays all available dbVar studies, and allows the displayed studies to be filtered by organism, study type, method, and variant size. Once the number

of available studies has been narrowed, links to publications and individual study pages allow a more in-depth search of available data.

Genome Browser

The dbVar Genome Browser searches dbVar data within the framework of a selected genomic assembly using a location, gene name, or phenotype as search terms and presents the result in a graphic display showing the variant in relation to genes located in the region. All dbVar variant region (nsv or esv) report pages are linked to the dbVar Genome Browser via the “Genome View” tab, which presents a graphic display of variant regions from other studies that overlap the variant displayed in the variant region report.

Variation Reporter

Variation Reporter matches submitted variation calls to references in ClinVar and to variants housed in dbVar or dbSNP, thereby allowing access via a Web search or through an application programming interface (API) to all data and metadata that dbVar has for the matching variants. If you submit novel variants and there are no matches between your data and the variants housed in dbVar or dbSNP, the Variation Reporter will provide the predicted consequence of each submitted variant.

Variation Viewer

Variation Viewer allows users to access variation data from dbVar, dbSNP, and ClinVar in relation to a specific gene or chromosomal location, and will allow the user to display data from any of these sources in an integrated navigable map. Users can search by dbVar accessions, gene, phenotype or disease, and chromosome positions (<http://www.ncbi.nlm.nih.gov/variation/view/help/#search>).

Search via ClinVar, Gene, or PubMed

There are multiple databases in NCBI that maintain links to dbVar. Related dbVar records are located by following links in the “Related Information” section of a record.

dbVar FTP Site

NCBI supports the public distribution of dbVar data by providing compressed data dumps in a number of different formats. Access to the NCBI FTP site is available via the World Wide Web (<ftp://ftp.ncbi.nlm.nih.gov/pub/dbVar/data/>) in data formats that include CSV, GVF, TAB, VCF, and XML.

ADA Section 508-Compliance

All links provided on the dbVar homepage are also provided in text format at the bottom of the page to support browsing by text-based Web browsers. Suggestions for improving database access by disabled persons should be sent to the dbVar development group at dbvar@ncbi.nlm.nih.gov

Related Tools and Studies

Remapping

NCBI's Genome Remapping Service (Remap) supports the conversion of genomic locations from one sequence to another based on alignments. Use Remap if you have identified the location of variation on an assembly, or on a RefSeqGene/LRG, and want to determine the location on a different assembly (or on the genome in the case of the RefSeqGene). dbVar remaps data in all its submissions to and from recent assemblies (e.g., from GRCh37 to GRCh38).

Association Studies

dbGaP archives and distributes data from studies that examine the relationship between phenotype and genotype. Such studies include Genome-wide Association Studies (GWAS), medical sequencing, and molecular diagnostic assays. Links are available from dbGaP controlled access records to related variation data in dbVar, but there are no reciprocal links from dbVar records to dbGaP unless the aggregate data are public.

Variation as Related to Citations, Genes, Phenotypes, and other NCBI Databases

Multiple databases in NCBI can be used to identify variation that meets certain criteria. They may either reference dbVar ID numbers explicitly, or provide links from their records to records in dbVar.

Variation Reporter

Variation Reporter matches submitted variation call data to variants housed in dbVar or dbSNP, allowing access to all data and metadata that dbVar has for any known matching variants. If you submit novel variants to the Variation Reporter, and there are no matches between your data variants housed in dbVar or dbSNP, the Variation Reporter will provide the predicted consequence of each submitted variant.

Variation Viewer

Variation Viewer allows users to access variation data from dbVar, dbSNP, and ClinVar in relation to a specific gene or chromosomal location, and will allow the user to display data from any of these sources in an integrated navigable map.

1000 Genomes Browser

The 1000 Genomes Browser provides access to 1000 Genomes data including variations, genotypes, and sequence read alignments within the context of GRCh37, the reference assembly used by the 1000 Genomes Project for analysis. The browser allows you to configure the display to include multiple data tracks of interest and provides links to

related data housed in various NCBI resources. The 1000 Genomes Browser allows users to quickly view alignments supporting a particular variant call and can be used to download and read variant data for small genomic regions of interest.

References

1. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Gen.* 2006 Feb;7(2):85–97. PubMed PMID: 16418744.
2. She X, Cheng Z, Zöllner S, Church DM, Eichler EE. Mouse segmental duplication and copy number variation. *Nat Genet.* 2008 Jul;40:909–914. PubMed PMID: 18500340.
3. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science.* 2008 Jun 20;320(5883):1629–3. PubMed PMID: 18535209.
4. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Gene.* 2009;10:451–81. PubMed PMID: 19715442.
5. Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.* 2012 Jan;28(1):43–53. PubMed PMID: 22094265.
6. Sindi SS, Raphael BJ. Identification of Structural Variation, In Maria Poptsova, *Genome Analysis: Current Procedures and Applications*. Caister Academic Press, pp. 1-19, 2014
7. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 2014 Jan;42(Database issue):D986–92. PubMed PMID: 24174537.
8. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the Human Genome. *Nat Genet.* 2004 Sep;36(9): 949–51. PubMed PMID: 15286789.
9. Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, Feuk L, Flicek P. Public data archives for genomic structural variation. *Nat Genet.* 2010 Oct;42(10):813–4. PubMed PMID: 20877315.
10. Lappalainen I, Lopez J, Skipper L, Hefferon T, Spalding JD, Garner J, Chen C, Maguire M, Corbett M, Zhou G, Paschall J, Ananiev V, Flicek P, Church DM. DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D936–41. PubMed PMID: 23193291.

ClinVar

Melissa Landrum, PhD, Jennifer Lee, PhD, George Riley, PhD, Wonhee Jang, PhD, Wendy Rubinstein, MD, PhD, Deanna Church, PhD, and Donna Maglott, PhD

Created: November 21, 2013.

Scope

It is increasingly easy to determine where an individual's nucleotide sequence may differ from a reference standard. It is much more difficult to determine which if any of those sequence variants has an effect on health. ClinVar has been developed to facilitate the evaluation of variation-phenotype relationships by archiving submitted interpretations of these relationships with supporting evidence, by aggregating data from multiple groups such as laboratories to determine if there is a consensus about the interpretation, and by making summary data freely available. ClinVar differs from NCBI's variation archives, namely dbSNP and dbVar, which have the responsibility of maintaining information about the types and locations of all sequence variation. In contrast, ClinVar provides a curated layer on top of these resources, focusing on the subset of all variation that may be medically relevant.

ClinVar integrates and cross-references data from multiple databases at NCBI. In addition to dbSNP and dbVar, ClinVar depends on MedGen to represent phenotype, Gene to represent genes, and on human RefSeqs to represent the location of sequence variation.

History

As a public database, ClinVar is young, having been fully released for the first time in April, 2013. However, ClinVar has been in development for several years, growing out of discussions of the Variome project about the benefits of centralizing information about rare human variation and its relationship to health. In 2008 dbSNP launched several tools to make it easier to submit such data, <http://www.ncbi.nlm.nih.gov/SNP/tranSNP/tranSNP.cgi> for single alleles and <http://www.ncbi.nlm.nih.gov/SNP/tranSNP/vsub.cgi> for spreadsheets. An application was developed to provide gene-specific views of such submissions ([VarView](#)); the single record display indicated if the location was submitted via the clinical channel; and our sequence displays provided a Clinical channel track. Several locus-specific databases used this functionality to submit data about rare human variation.

In addition to submissions from external groups, the RefSeqGene staff shepherded data from GeneReviews and OMIM into dbSNP to augment the connections among the published literature, other databases, and the variation archives. Based on this foundation, and NCBI's maintenance first of GeneTests, and now the NIH Genetic Testing Registry (GTR), NCBI was approached by several stakeholders to develop what is now called

ClinVar. The genetic testing community was seeking a comprehensive, up-to-date, freely accessible resource in which to share data and pool resources to evaluate human variation.

Data Model

ClinVar's data model is based on five major categories of content: submitter data for attribution, definition of the variation, characterization of the phenotype, evidence about the effect of variation on health, and interpretation of that evidence. Whenever possible, the content is highly structured rather than free text, and is harmonized to controlled vocabularies or other data standards.

ClinVar is a submitter-driven resource that maintains an archive of what has been received and processed. Data from submitters is assigned an accession of the format SCV123456789 (SCV) and data from multiple submitters about the same variation/phenotype combination is aggregated and assigned an accession of the format RCV123456789 (RCV). Content is versioned, i.e., the first submission is assigned version 1 and any updates to a submission is represented as an incremented version of the same accession. The RCV record also includes content added by NCBI, such as accessions from other databases, standard terminology, and analysis of related submissions.

Submitter

ClinVar represents submitters as both organizations and individuals. The infrastructure supporting this content is shared with the NIH Genetic Testing Registry (GTR), dbSNP, and dbVar. Submitters have the right to request anonymity, although to date no submitter has requested this option. Summary data about submissions are provided on the website (<http://www.ncbi.nlm.nih.gov/clinvar/submitters/>).

Variation

Variation is a key component of ClinVar's data model, especially to be able to represent variation's relationship to phenotype. Variation is thus represented both as the sequence at a particular location, or a combination of sequence changes. In other words, ClinVar can represent the interpretation of a single allele, compound heterozygotes, haplotypes, and combinations of alleles in different genes. Variation is modeled in the database as a set of variations, but currently most sets have only one member. The goal is to represent each variation on a reference sequence, but the data flow from some submitters is not amenable to establishing this immediately. Thus free text is accepted.

Variations submitted to ClinVar are compared to variations accessioned by dbSNP or dbVar. If known, ClinVar adds the rs# (dbSNP) or variant call identifier (dbVar) to the RCV record. If novel, the information is submitted to the appropriate variation database to be accessioned, so that the identifiers can be added to ClinVar. In other words, ClinVar does not create new identifiers for locations of variation. Also the archival databases do not note the number of submitters that have contributed information to ClinVar about a variation. That said, to support internal data flows and some public reports ClinVar does

assign an internal unique identifier to the sequence change at each location, which is reported in the XML and tab-delimited exports as an integer identifier (Table 1).

ClinVar reports multiple types of attributes for each variant. HGVS expressions are reported based on the current reference assembly, [RefSeqGenes](#), cDNAs and proteins as appropriate. When there are multiple transcripts for a gene, ClinVar selects one HGVS expression to display as the preferred name. By default, this selection is based on the first reference standard transcript identified by the RefSeqGene/Locus Reference Genomic (LRG) collaboration, but can be overridden upon request.

Some of the data ClinVar reports related to variation are values added by NCBI. These are reported only as part of the RCV record (because the SCV accession is what the submitter provides), and can include alternate HGVS expressions, allele frequencies from the 1000 Genomes project or GO-ESP, identifiers from dbSNP or dbVar, molecular consequences (e.g., nonsense/missense/frameshift), location data (splice site, UTR's, cytogenetic band, genes), and confidence in variation calls at that location.

Table 1. Identifiers used by ClinVar

Name	Scope	Examples
SCV accession.version	Assigned to a submission	SCV000065090.1
RCV accession.version	Assigned to an aggregation	RCV000008391.2
AlleleID	Assigned to one variation	22968
rs#	Identifier assigned by dbSNP to a type of variant at a location on an assembly	rs238
nsv#	Identifier assigned by dbVar to a variant region	nsv513782

Phenotype

ClinVar represents phenotype as concepts identified in MedGen. Similar to management of variation, these concepts can be single or sets of multiple values. Sets are used primarily to report a combination of clinical features; single values are used to represent diagnostic terms or indications. Submitters are encouraged to submit phenotypic information via identifier, e.g., MIM number, MeSH term, or identifier from the [Human Phenotype Ontology](#) (HPO). Free text is accepted and ClinVar staff will work with submitters to determine if that text can be mapped to current standardized concepts. If not, ClinVar establishes a new identifier to be represented in MedGen and adds that MedGen identifier to the RCV record.

Interpretation

All interpretations of the relationship between variation and phenotype in ClinVar are supplied by submitters. ClinVar reports [clinical significance](#), the date that clinical significance was last interpreted by the submitter, and functional significance. To support interpretation, mode of inheritance of a variation relative to a disorder and qualification

of severity of phenotype are also represented. Terms for clinical significance are those recommend by the American College of Medical Genetics (ACMG). If submitters disagree on the interpretation of the clinical significance of any variation, that record is marked in the aggregate report as having conflicts. If one submitter does not provide this information, and another does, that is not marked as conflicting.

Comparison of clinical significance provided by multiple submitters is computed by two methods. One is a strict interpretation, per RCV accession, of any difference. In other words, pathogenic and likely pathogenic are reported as being in conflict. The second is more relaxed, and based only on the variation and not the variation as related to a specific phenotype. In this mode, the conflicts are reported only at the extremes, i.e., differences between pathogenic/likely pathogenic, benign/likely benign, and uncertain significance.

Evidence

Evidence that supports an interpretation of the variation-phenotype relationship can be highly structured and/or a free-text summary discussing how the evidence was evaluated. When structured, content includes the description of how the variants were called and in what context (genetic testing, family studies, comparison of tumor/normal tissue, animal models, etc.) Based on that context, the results can be represented as number of independent observations per person or chromosome, number of segregations observed, number of times other rare variations were identified in the same gene or other genes, etc. At present, most structured data are reports of number of individuals in which non-somatic variation was observed, sometimes with indication of number of families.

Dataflow

Initial records

The major data flows for ClinVar are diagrammed in Figure 1. Direct submissions are validated, converted to XML, and accessioned. If any content does not validate, submitters are contacted and corrections are requested. When valid, the records are assigned accessions (SCV) and processed. Submitters are provided reports including the accessions assigned to their data and indications as to whether any of their data conflicted with current public submissions.

Data that NCBI processes from OMIM or GeneReviews are managed slightly differently. Data from OMIM are updated daily from automatic feeds, and bypass the validation assigned to direct submissions. If possible, novel variations are converted to sequence coordinates by testing possible reference sequences and determining if the data in the text of OMIM's description of the allele are consistent with reported sequence changes. As resources permit, NCBI staff reviews recent records from OMIM that cannot be processed automatically. Data from GeneReviews are extracted from the tables embedded in the GeneReview, as well as attached tables provided by the submitter. Any questions that arise in processing data from GeneReviews are reported to GeneReviews staff for review.

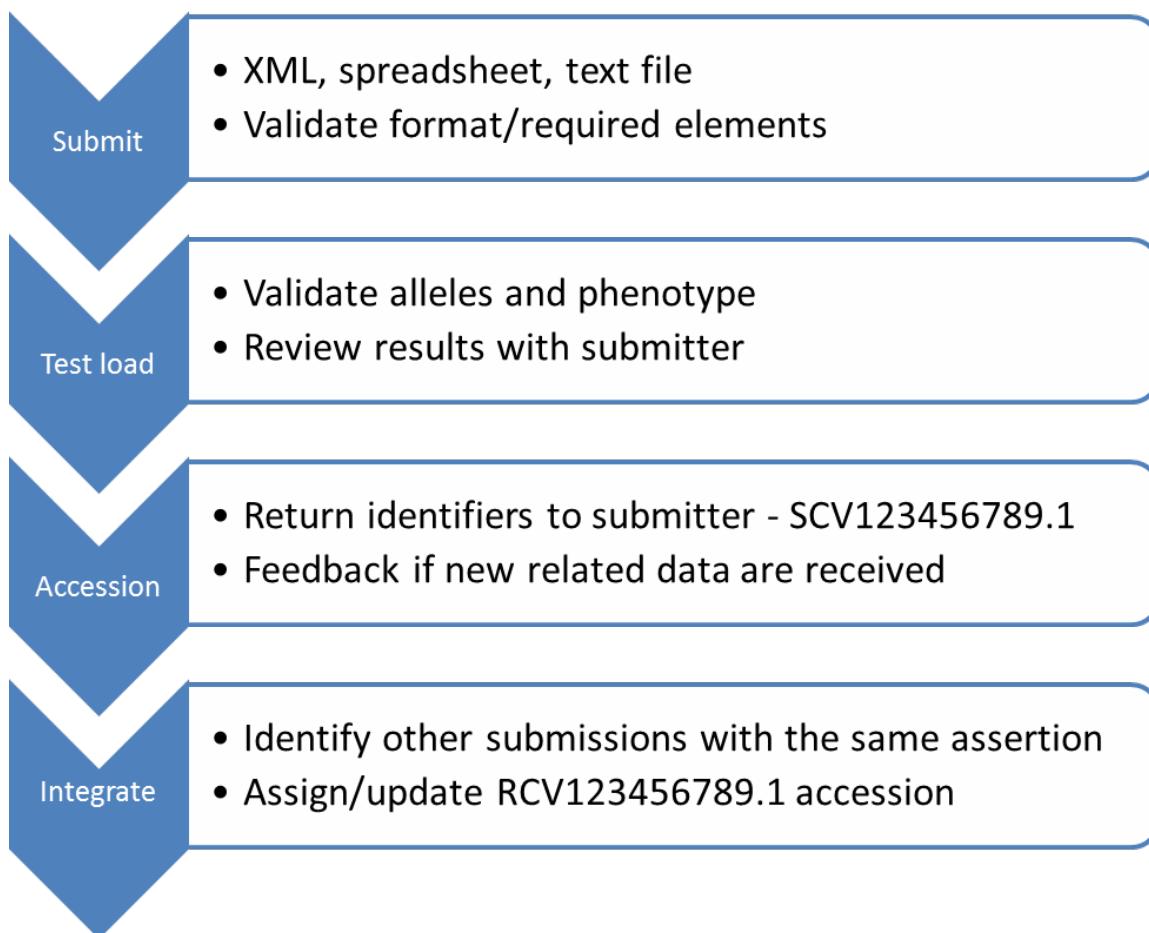


Figure 1. Overview of the flow of information through ClinVar. ClinVar validates content and looks for differences relative to previous submissions and returns reports to the submitter before the data are released to the public.

Updates

Submitters may update their submissions at any time. With an update, the accession is assigned a new version. Thus if a unit record in a submission were assigned the accession SCV000000001, with an update the version would be incremented, in this case to 2 (SCV000000001.2).

RCV-specific processing

Data associated with an RCV accession can change in one of two ways. One is represented by an increment of a version. Again, if there are multiple submissions about the same variation-phenotype relationship, these are aggregated into one RCV accession and versioned. The version of an RCV accession is incremented if a new submission is received for the same variation-phenotype relationship (i.e., a new SCV accession is added

to the set represented by the RCV accession), or if any SCV accession in the set is itself updated and assigned a new version.

The content of an RCV accession can also change without that being reflected in a new version. If a genomic assembly changes, if genomic coordinates are established for a variation for the first time, if database identifiers such as rs#, nsv#, or PubMed ids are added, if preferred terms are redefined, then the content will be updated without assigning a new version, but with a new unique identifier. These snapshots of content are calculated weekly, and the unique integer identifier is detected when accessing ClinVar via E-Utilities.

Access

Web

ClinVar's website, <http://www.ncbi.nlm.nih.gov/clinvar>, is part of NCBI's Entrez system and thus is searchable with the standard query interface and Advanced query options. ClinVar supports retrieval by any text in the RCV record, including descriptions of variation (HGVS expression, rs, nsv, nssv, OMIM allelic variant identifier, identifier used in a locus-specific database or LSDB), genes (symbol or full name), disease (names and identifiers), submitter names, and clinical significance. To facilitate a common search strategy, a query that is detected to be a human gene symbol displays a link to make it easier to limit your query results by that symbol. The default result set is a table of 20 rows, but that can be altered using Display Settings (Figure 2). When multiple results are returned from a query, filters are provided at the left that reflect the content of the retrieval set (values and counts of each). Clicking on one of those options removes all but the selection from the display, a restriction that can be reversed by using the Clear option.

The full record is accessed by clicking on See details in the first column of the tabular display, or the title row if the summary display option is used. At present, the detailed display corresponds to content of an RCV accession (Figure 3). The Clinical significance, Allele description, Condition(s) sections, and the Genome view report aggregate data; the Clinical Assertions are submitter-specific, and the Evidence (not shown) is provided both in aggregate and submitter-specific sections.

Before the end of 2013, a new display will be provided via See details, quite similar to the RCV report but aggregated per single variation rather than variation-phenotype combination. This new display allows users to see all data for a variation even when submitters' representation of phenotype differs.

Data in ClinVar can also be discovered via other NCBI databases, based on the links that are built when content is shared. Examples include dbSNP, dbVar, Gene, MedGen, Nucleotide, and PubMed. Locations of variation represented in ClinVar are annotated on RefSeqs and are visible in the graphical sequence displays (e.g., <http://www.ncbi.nlm.nih.gov/nuccore/125662814?report=graph>), and browsers such as 1000 Genomes (<http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>). ClinVar also

Results: 1 to 20 of 81								
	Gene	Variation	Freq	Phenotype	Clinical Significance	Review Status	Chr	Location (GRCh37.p10)
See details	RYR1	c.11941C>T (p.His3981Tyr)	GMAF: 0.0051	Minicore myopathy with external ophthalmoplegia	pathogenic	classified by single submitter	19	39034444
See details	RYR1	c.13603G>A (p.Glu4535Lys)		melanoma	not provided	not classified by submitter	19	39058501
See details	RYR1	c.12335C>T (p.Ser4112Leu)		melanoma	not provided	not classified by submitter	19	39051805
See details	RYR1	c.8134C>T (p.Pro2712Ser)		melanoma	not provided	not classified by submitter	19	38995454
See details	RYR1	c.2611G>A (p.Glu871Lys)		melanoma	not provided	not classified by submitter	19	38954096
See details	RYR1	c.487C>T (p.Arg163Cys)		melanoma	not provided	not classified by submitter	19	38934851
See details	RYR1	c.97A>G (p.Lys33Glu)		King Denborough syndrome	pathogenic	classified by single submitter	19	38931436
See details	RYR1	c.10579C>T (p.Pro3527Ser)		Central core disease, autosomal recessive	pathogenic	classified by single submitter	19	39016095

Figure 2. Tabular results of a ClinVar search.

provides specialized pages for certain types of access. One is the list of genes and disorders for which ACMG recommends that incidental findings be reported (1) (<http://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/>); another is the listing of submitters and all their submissions (<http://www.ncbi.nlm.nih.gov/clinvar/submitters/>).

FTP

Data from ClinVar are reported from several directories at NCBI and in several formats. The README file (<ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/README.txt>) provides a comprehensive list. Current content includes the file converting MIM numbers, GeneIDs, and MedGen concepts ids on Gene's FTP site ([mim2gene_medgen](#)), the listing of standard terms used by ClinVar at GTR's FTP site, and the tab-delimited, XML, and VCF files from ClinVar. The VCF files are available from dbSNP (with the symbolic link from ClinVar).

E-Utilities

ClinVar supports programmatic access via E-Utilities as esearch, esummary, and elink. E-fetch is not enabled. Please note that esearch (e.g., [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=clinvar&term=brca1\[gene\]&retmax=1000](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=clinvar&term=brca1[gene]&retmax=1000))

returns the unique identifiers for an RCV record, which does not correspond 1:1 with an accession.version. The unique identifiers represent an instance of that record, which may change without a version change if NCBI adds data to the record such as an rs# or a

RYR1:c.5333C>A (p.Ser1778Ter) AND Congenital myopathy with fiber type disproportion

Clinical significance:	pathogenic (Last evaluated: Apr 11, 2013)	Help					
Review status:							
Based on:	1 submission [Details]						
Record status:	current						
Accession:	RCV000034928.1						
Allele description							
Gene:	RYR1:ryanodine receptor 1 (skeletal) [Gene OMIM]						
Variant type:	single nucleotide variant						
Genomic location:	Chr19:38976628 (on Assembly GRCh37)						
Preferred name:	RYR1:c.5333C>A (p.Ser1778Ter)						
Protein change:	S1778*						
HGVS:	NC_000019.9:g.38976628C>A NG_008866.1:g.57289C>A NM_00540.2:c.5333C>A NP_00531.2:p.Ser1778Ter						
Links:	GeneReviews: NBK1259 ; dbSNP: 367543055						
1000Genome:	rs367543055						
Molecular consequence:	NM_00540.2:c.5333C>A: STOP-GAIN [Sequence Ontology: SO:0001587]						
Suspect:	Not available						
Condition(s)							
Name:	Congenital myopathy with fiber type disproportion (CFTD)						
Synonyms:	Congenital fiber type disproportion (CFTDM)						
Identifiers:	GeneReviews: NBK1259 ; MedGen: C0546264 ; OMIM: 255310 ; Orphanet: 2020						
Age of onset:	Neonatal/infancy						
Clinical Assertions Genome View Evidence Help							
Submission Accession	Submitter	Review Status	Clinical Significance (Last evaluated)	Origin	Method	Consequence	Citations
SCV000058535	GeneReviews		pathologic (Apr 11, 2013)	not provided	curation		

Figure 3. Detailed display of an RCV record. This is currently the default display.

ConceptUID from MedGen. A record retrieved by an outdated ID provides a link to the current record.

Related Tools

The data for which ClinVar is responsible, namely the archive of interpretations of clinical significance, is integrated into the various tools NCBI maintains to manage recalculation of sequence coordinates ([Clinical Remap](#)) and to report what is known about human variation at a genomic location ([1000 Genomes Browser](#), [Variation Reporter](#), [Variation View](#)). These data are integrated monthly on the first Thursday.

References

- Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, McGuire AL, Nussbaum RL, O'Daniel JM, Ormond KE, Rehm HL, Watson MS, Williams MS, Biesecker LG. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med.* 2013;Jul15(7):565–74. PubMed PMID: 23788249.

Health

MedGen

Maryam Halavi, MD, PhD^{✉1} Donna Maglott, PhD,¹ Viatcheslav Gorelenkov, MS,¹ and Wendy Rubinstein, MD, PhD¹

Created: May 28, 2013.

Scope

MedGen is NCBI's portal to information about human disorders and other phenotypes having a genetic component. MedGen is structured to serve health care professionals, the medical genetics community, and other interested parties by providing centralized access to diverse types of content. For example, because MedGen aggregates the plethora of terms used for particular disorders into a specific concept, it provides a Rosetta stone for stakeholders who may use different names for the same disorder. Maintaining a clearly defined set of concepts and terms for phenotypes is essential to support efforts to characterize genetic variation by its effects on specific phenotypes. The assignment of identifiers for those concepts allows computational access to phenotypic information, an essential requirement for the large-scale analysis of genomic data.

Once a concept is defined, MedGen offers a growing collection of attributes about that concept including a definition or description, clinical findings, causative genetic variants and the genes in which they occur, available clinical and research tests, molecular resources, professional guidelines, original and review literature, consumer resources, clinical trials, and Web links to other related NCBI and non-NCBI resources. Convenient access to such a range of supporting data allows MedGen's users to synthesize and apply the latest knowledge to important clinical and biological questions.

History

There are multiple public databases, ontologies, or tools that provide terms, definitions, and other information about human diseases and phenotypes. None, however, is focused on maintaining an up-to-date information resource that both harmonizes terminology data from others and also provides an interface to use those harmonized terms to identify related information in the current public arena.

Recognizing this broad challenge and the collective interest to address it, MedGen was initiated in 2012 as a public resource in the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH), National Library of Medicine (NLM). Seeded from terms established from NLM's Unified Medical Language

¹ NCBI; Email: halavim@ncbi.nlm.nih.gov; Email: maglott@ncbi.nlm.nih.gov; Email: viatcheslav.gorelenkov@nih.gov; Email: rubinstw@ncbi.nlm.nih.gov.

[✉] Corresponding author.

system (UMLS®), the NIH Genetic Testing Registry (GTR®), and ClinVar; MedGen continues to add terms, types of content, and services.

As of August 2013, more than 170,000 concepts had been integrated in MedGen. Many aspects of MedGen and its data gathering procedures are evolving, so users' suggestions and comments are welcomed.

Data Model

MedGen has a very simple data model. Once a concept is identified, categories of information that relate to that concept are identified and reported. These categories may be descriptors, links to other databases, or concept-concept relationships.

A MedGen record has several important components including:

Concept-related

Concept: An aggregation of terms from multiple vocabulary sources, which have been determined to be comparable (i.e., represent the same concept).

ConceptID (CUI—Concept Unique Identifier): A unique stable identifier assigned to each concept.

Semantic type: A defined set of categories of concepts, so that concepts that share terms can be differentiated by scope. An example is the distinction between 'autism' as a synonym for the diagnosis Autism spectrum disorders (C1510586) and 'autism' as a clinical feature or finding ([CN000674](#)) identified in multiple disorders.

Descriptors

MedGen maintains multiple types of descriptors, including names, acronyms, abbreviations, sources of descriptors, attribution for and identifiers used by those sources, cytogenetic locations, mode of inheritance, textual definitions or descriptions, types of genetic testing registered in the NIH Genetic Testing Registry (GTR®), genes for which aberrant function may be related to the disorder, related variations, professional guidelines, consumer resources, and more as detailed below (see Table 1). Sources of terms are called 'vocabularies', consistent with the usage of UMLS® (1). The descriptors are often presented in the query interface as distinct **fields** or as **concepts' properties**.

Table 1. A list of data elements aggregated in MedGen and their sources

#	Data Type	Source
1	Clinical and research tests	GTR®
2	Clinical features	HPO
3	Clinical trials®	ClinicalTrials.gov

Table 1. continues on next page...

Table 1. continued from previous page.

#	Data Type	Source
4	Concept ID	UMLS® or GTR®
5	Consumer resources	Genetics Home Reference, Genetic Alliance, Genetic and Rare Diseases Information Center, MedlinePlus
6	Cytogenetic location	NCBI annotation
7	Gene	NCBI's Gene
8	Links to other NCBI's resources	NCBI's resources such as Gene, MeSH®, ClinVar, Bookshelf, BioSystems, etc.
9	MedGen Identifier	MedGen
10	Medical encyclopedia	A.D.A.M Medical encyclopedia via PubMed Health
11	Mode of inheritance	OMIM® /ClinVar/GTR®
12	Molecular resources	Coriell Institute for Medical Research
13	Professional Guidelines	NCBI curation
14	RefSeqGene	RefSeqGene
15	Reviews	GeneReviews™, PubMed Clinical Queries
16	Semantic type	UMLS®
17	SNOMED CT® terms	SNOMED CT®
18	Source Identifiers	Various sources, such as OMIM®, HPO, etc.
19	Terms definitions	GeneReviews™, Medical Genetics Summaries, etc.
20	Terms, acronyms, synonyms	Defined vocabularies
21	Terms hierarchies	GTR®, MedGen
22	Variations	ClinVar

Definitions/Descriptions

Terms imported to MedGen may have been associated with a definition by their source or in other sources. MedGen includes a single definition for each concept to be displayed in the search summaries. In case of multiple definitions, a simple prioritization is used: the first priority is assigned to the definitions from GTR®, followed by SNOMED CT®. Subsequent to that order, MedGen adheres to the UMLS® prioritization of source vocabularies and term types. In the full report for a concept MedGen may report multiple definitions. Any definition displayed in MedGen includes attribution and a link to the source.

Identifiers from source databases

MedGen maintains and displays identifiers from source databases not only to provide attribution, but also to support interactive and programmatic searching and links to

sources' websites. MedGen also maintains alternative IDs from the Human Phenotype Ontology (HPO) in the current record of that HPO concept.

Gene-phenotype relationships

A disease concept in MedGen reports the symbol(s) of genes that are reported to be causative. Links are also provided to NCBI's Gene database and OMIM®.

Cytogenetic locations

Cytogenetic locations for each concept are reported from ClinVar based on location of the genes that contribute to that disorder.

Inter-concept relationships

MedGen maintains two major types of inter-concept relationships. The first is the diagnosis-clinical feature relationship, so that the user can see all features reported for a disorder and find all disorders that share a clinical feature. The second type is hierarchies. MedGen currently provides three types:

- Clinical features, from top-level to each child, no matter what the final level. This is used to group clinical features by type
- Disorders: computed from parent-child relationships
- Disorders: curated by GTR®/ClinVar staff.

Published literature

MedGen's Web interface provides information about related publications in multiple ways. One is curation, by NCBI staff, of professional guidelines related to a disorder. The second is aggregation of publications from contributors ClinVar and GTR®, which are reflected in the links to PubMed. The Recent clinical studies section uses PubMed's Clinical Queries logic. Finally, the preferred name of the MedGen record is used to query PubMed and identify highly related literature.

Dataflow

Data in MedGen are acquired both programmatically and manually via curation, depending in part on the type of information and the sources for that information. This section summarizes those flows organized by type of information.

The first step in organizing the information is to establish a concept that defines a disorder or phenotype, classify that concept by type, and then assign that concept a stable unique identifier. With that framework established, data are then aggregated around that concept. The extent of metadata attributed to each concept may vary based on the availability. However, to ensure maximized benefit of having current, correct, and complete metadata, relevant metadata are actively managed in MedGen.

Concept UID

Concepts' unique identifiers (Concept UID or CUI) are assigned to each concept to facilitate connecting different terms from various vocabularies to that concept. The Concept UIDs in MedGen are either derived from UMLS® or assigned by MedGen (starting with CN) if a match based on term and semantic type in UMLS® cannot be identified. UMLS® is maintained by NLM and provided to researchers on the terms of license agreement without any charge. Terms in UMLS® are classified based on broad categories of semantic types and term relations (2). Unique identifiers assigned in UMLS® are permanent Concept UIDs, however, in each semi-annual UMLS® release, some of the Concept UIDs can be merged or deleted. If so, concept UIDs calculated by MedGen are deprecated in favor of those from UMLS® for that concept. Concept UIDs also may be merged or deleted either because of vocabulary changes or because of NCBI internal curation. If a concept semantic type is in the scope for MedGen, Concept UID is imported to MedGen programmatically. MedGen maintains the history of the UMLS® Concept UID merges and deletions to ensure stable and permanent identifiers.

Names, acronyms, abbreviations (terms)

MedGen integrates large sets of terms, their relationships, and their definitions (if available), as well as additional supplementary information from a variety of sources (termed vocabularies).

- [ClinVar](#), daily
- Human Phenotype Ontology ([HPO](#)), weekly
- Genetic Testing Registry ([GTR®](#)), daily
- Medical Subject Headings Thesaurus ([MeSH®](#)), semi-annually via UMLS®
- National Cancer Institute Thesaurus ([NCIt](#)), semi-annually via UMLS®
- Online Mendelian Inheritance in Man ([OMIM®](#)), daily
- Systemized Nomenclature of Medicine—Clinical Terms ([SNOMED-CT®](#)), semi-annually via UMLS®

Names and acronyms for each concept are aggregated from these vocabulary sources. The preferred name and preferred acronym are either internally curated or selected based on the UMLS® standards. Alternate terms derived from other vocabularies are reported as synonyms for each concept.

MedGen restricts the processing of concepts from UMLS® to a subset of semantic types (disease or syndrome, abnormality and dysfunction, sign and symptom, finding, molecular and pathological function, pharmacologic substance, neoplastic process, etc.). However, one concept can be associated with more than one semantic type and reported more than once with different semantic types. A semantics-aware mapping approach is used to maintain useful associations between concepts and support their bi-directional multi-level relationships.

Achondroplasia (ACH)
MedGen UID: 1289 • Concept ID: C0001080 • Disease or Syndrome
Modification Date: 13 Jul, 2013

Synonyms: ACH; Achondroplastic dwarfism; Chondrodystrophy fetalis; Chondrodystrophy syndrome; Congenital osteosclerosis; Dwarf, achondroplastic; Osteosclerosis congenita

Modes of inheritance: Autosomal dominant inheritance

SNOMED CT: Achondroplasia (86268005); Chondrodystrophy fetalis (86268005); Achondroplastic dwarf (86268005); Osteosclerosis congenita (86268005); Congenital osteosclerosis (86268005); Achondroplastic dwarfism (86268005)

Gene: FGFR3
Cytogenetic location: 4p16.3
OMIM: 100800

Figure 1. Names, acronyms, identifiers, MOI, and Cytogenetic location displayed on MedGen website. Top section of the full report in MedGen includes the title for the concept, IDs associated with the concept, semantic type, synonyms, mode of inheritance, related terms from SNOMED CT®, gene, cytogenetic locations, and if available MIM identifier from OMIM®.

Names from OMIM® (3) are processed from both UMLS® and from daily updates directly from OMIM®. Terms from HPO, as a primary source for clinical features of Mendelian disorders, are updated weekly. Terms from GTR® are mainly based on what was provided by the submitters during a test registration, but curators will review the evidence for each submission. Because SNOMED-CT® has many concepts that are not in scope for MedGen, MedGen does not represent all of SNOMED CT®, but only SNOMED CT's terms for concepts in MedGen.

If a new vocabulary source is identified, MedGen will integrate the data based on evaluation of terms and semantic types, maintaining the source of the data (vocabulary) and the identifier used by the source. If a term matched an existing concept, the term and source will be added; otherwise a new CUI will be established.

Genetic/genomic characteristics (MOI, cytogenetic location)

All modes of inheritance for a disorder are extracted from the resources reporting the term. Gene symbols and cytogenetic locations associated with a disorder concept are derived from NCBI's Gene database (which uses the HUGO Gene Nomenclature Committee (HGNC) standard) based on the gene-disorder relationships. Figure 1 shows an example of how these data are displayed on the website.

Clinical feature-disorder relationships

Clinical features describing the sign and symptoms characteristic of a disorder are provided from HPO (4) and updated weekly. The phenotype abnormalities in HPO are categorized in 20 organ abnormality groups and MedGen uses the same categorization for its reporting of clinical features in a hierarchical format. MedGen enhances access to all conditions with this clinical feature by providing an option to search on a specific clinical feature. Figure 2 illustrates the Clinical features section in MedGen.

Clinical features

Show all Hide all

- ▼ Abnormality of head and neck
 - Abnormality of the teeth
 - Depressed nasal bridge
 - [Foramen magnum stenosis](#)
 - Frontal bossing
 - Macrocephaly
 - Megalencephaly
- ▶ Abnormality of the cardiovascular system
- ▶ Abnormality of the ear
- ▶ Abnormality of the immune system
- ▶ Abnormality of the integument
- ▶ Abnormality of the musculature
- ▶ Abnormality of the nervous system
- ▶ Abnormality of the respiratory system
- ▶ Abnormality of the skeletal system
- ▶ Increased upper to lower segment ratio

Foramen magnum stenosis
MedGen UID: 505811 • Concept ID: CN004847 • Finding

An abnormal narrowing of the foramen magnum.

See: Feature record | Search on this feature

Figure 2. Clinical feature section in MedGen. Clinical features are reported under top nodes established by HPO (top nodes are immediate children of the HPO term Phenotypic abnormality - HP:0003812). Each node can be expanded or collapsed to improve viewing of all items. A click on any item in the list displays a pop-up window with the definition of that feature and links to either the full report in MedGen for that feature (Feature record) or other conditions reported to have that feature (Search on this feature).

Term Hierarchy

GTR MedGen

C Clinical test, **R** Research test, **O** OMIM, **G** GeneReviews

C R O G • Achondroplasia

A

Term Hierarchy

GTR MedGen

- ▼ Disorder of musculoskeletal system
 - ▼ Disorder of bone
 - ▼ Disorder of bone development
 - ▼ Dwarfism
 - ▼ Achondroplasia
 - ▼ Thanatophoric dysplasia
 - Thanatophoric dysplasia, type 1
 - Thanatophoric dysplasia, type 2

B

Figure 3. Term hierarchies in MedGen. The hierarchies are illustrated in tabbed navigation format. (A) Illustrate the GTR® hierarchy alongside with icons, which provide links to corresponding Clinical tests, Research tests, OMIM®, or GeneReviews™ Records. (B) Illustrate the hierarchy reported in MedGen. Small arrows on the left allow for expansion and contraction of the branches of a large hierarchical tree to ease navigation.

Term hierarchies

Term hierarchies are constructed based on relationships reported for each concept as direct or indirect links between terms from the vocabulary sources. This enables users to

expand their search queries and browse terms relationships. MedGen represents hierarchies as trees in which the terms are arranged having a root (top level node) and many branches (children), which can be in the same level or below their parent. The MedGen hierarchy is constructed based on either direct links for each concept or extending links vertically on the hierarchy tree toward the parents and the children (traveling 3 levels upward to find a common parent and then downward to find related children). Alternative hierarchies such as GTR® are also provided by a tabbed navigation, as it is shown in Figure 3. The concepts in GTR® hierarchy are displayed alongside any available links to Clinical tests, Research tests, OMIM®, or *GeneReviews*™. Figure 3 illustrates the display of both GTR® hierarchy (A) and MedGen hierarchy (B) in the Term Hierarchy section.

Available testing

The clinical tests and research tests registered in the Genetic Testing Registry (GTR®) are mapped to each relevant disorder concept and hyperlinked to full test records in GTR®. This convenient access, as displayed in Figure 4, improves users' ability to view the test's purpose, methodology, validity, evidence of a test's usefulness, and laboratory contacts and credentials. The tests are grouped according to the primary method used, as reported to GTR® (5).

Professional guidelines

The relevant clinical practice guidelines, position statements, and recommendations from various sources, such as American College of Medical Genetics and Genomics (ACMG), Evaluation of Genomic Applications in Practice and Prevention (EGAPP), American Congress of Obstetricians and Gynecologists (ACOG), The Clinical Pharmacogenetics Implementation Consortium (CPIC), The National Society of Genetic Counselors (NSGC), etc., are curated and associated with related concepts (currently, 258 guidelines have been curated for 462 conditions). To facilitate access to these guidelines, MedGen has a dedicated section in its full report (Professional guidelines section), which provides hyperlinks to either PubMed or PMC (if available) or to other online source for each guideline.

Publications

Highly relevant literature and publications are useful in assessing the nature and importance of a concept as well as expanding a user's perspective about a concept. In MedGen, these highly relevant literatures are aggregated through use of comprehensive and specialized queries on PubMed and PubMed Central (PMC). The citations are not directly stored in MedGen, but are retrieved dynamically to keep the content and links up-to-date.

Direct queries to PubMed and PMC are executed either based on the relevant terms from Medical Subject Headings (MeSH®) as query term; or if there is no MeSH® term connection, by using the preferred name of the concept, its synonyms, or its acronyms as

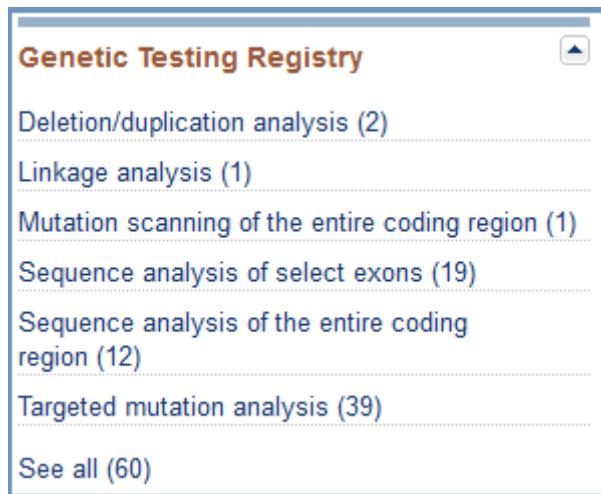


Figure 4. The clinical and research tests associated with each concept. The tests are listed according to the primary method used. The total number of available tests is shown and the option “See all” enables access to all test records registered for each concept.

query terms. Both PubMed and PMC use a ranking system for pushing more relevant results to the top of the results set. However, to increase the specificity of the results MedGen uses filters such as English language, human species, having abstract, genetics subheading, etc. If the number of the articles returned by these direct queries is very large, an arbitrary number may be used as a cutoff point to keep the results manageable. If no result is generated by using these queries, then a graphical model is used. In this graphical model MedGen combines logical structures and probabilities to create a flexible framework for finding related articles based on relationship to associated disorders, genetic mechanism of the disorders, body sites, or unique ingredients for drugs.

Although not strictly part of its dataflow, MedGen also provides access to publications at the time of the display of a full record. In other words, the name of the record is submitted to PubMed via [PubMed Clinical Queries](#). These clinical queries are filtered by five different Clinical Study Categories (Etiology, Diagnosis, Prognosis, Therapy, and Clinical prediction guides) and Systematic Reviews. Use of the broad scope option provides maximum coverage of the relevant literature and additional filters (namely English language and human, and not comment publication types or letter publication types) increase specificity. The results are displayed in “Recent clinical studies” and “Recent systematic reviews” sections, respectively. In each section, the title, list of authors, and journal information is displayed. Figure 5 illustrates the display of the literature in various sections.

Gene-disorder relationships

Gene symbols in MedGen are limited to current or previous official symbols from the HGNC ([Data sources for Gene](#)). If there is no official symbol from the HGNC, then NCBI

The screenshot shows the NCBI MedGen interface for the disease Achondroplasia (ACH). The top navigation bar includes links for NCBI Resources and How To. The search bar is set to 'MedGen'. Below the search bar are 'Limits' and 'Advanced' buttons.

Display Settings: Full Report **Send to:**

Achondroplasia (ACH)
MedGen UID: 1289 • Concept ID: C0001080 • Disease or Syndrome
Modification Date: 26 Sep, 2013

Synonyms: ACH; Achondroplastic dwarfism; Chondrodystrophy fetalis; Chondrodystrophy syndrome; Congenital osteosclerosis; Dwarf, achondroplastic; Osteosclerosis congenita

Modes of inheritance: Autosomal dominant inheritance

SNOMED CT: Achondroplasia (86268005); Chondrodystrophy fetalis (86268005); Achondroplastic dwarf (86268005); Osteosclerosis congenita (86268005); Congenital osteosclerosis (86268005); Achondroplastic dwarfism (86268005)

Gene: FGFR3
Cytogenetic location: 4p16.3
OMIM: 100800

Disease characteristics **Go to:**

Additional descriptions **Go to:**

Clinical features **Go to:**

Term Hierarchy **Go to:**

Professional guidelines **A**

PubMed
Statement on guidance for genetic counseling in advanced paternal age.
Torriello HV, Meck JM, Professional Practice and Guidelines Committee
Genet Med 2008 Jun;10(6):457-60. doi: 10.1097/GIM.0b013e318176fabb. PMID: 18496227 **Free PMC Article**

Recent clinical studies **B**

Etiology
Sagittal spinopelvic parameters in children with achondroplasia
Karikari IO, Mehta AI, Solakoglu C, Bagley CA, Ain MC, Gottfried O
J Neurosurg Spine 2012 Jul;17(1):57-60. Epub 2012 Apr 27 doi: 10.1227/JNS.0b013e318176fabb

Diagnosis
Physeal growth arrest after tibial lengthening in achondroplasia
Cai Y, Zhao H, Liu Z, Liu C, Luan J, Zhou X, Han J
Orphanet J Rare Dis 2012 Aug 22;7:55. doi: 10.1186/1750-1172-7-55. [Epub ahead of print] PMID: 22913777 **Free PMC Article**

Therapy
Prognosis
Clinical prediction guides

See all (223)

Recent systematic reviews **C**

A systematic review of genetic skeletal disorders reported in Chinese biomedical journals between 1978 and 2012.
Cui Y, Zhao H, Liu Z, Liu C, Luan J, Zhou X, Han J
Orphanet J Rare Dis 2012 Aug 22;7:55. doi: 10.1186/1750-1172-7-55. [Epub ahead of print] PMID: 22913777 **Free PMC Article**

Figure 5. The professional guidelines and clinical literature sections in MedGen. (A) For each professional guideline the title, list of authors, and the journal information are presented in a “Professional guidelines” section, which include PMID and links to PMC, if available. The results from clinical queries in PubMed are presented in a (B) “Recent clinical studies” section under five Clinical Study Categories: Etiology, Diagnosis, Prognosis, Therapy, and Clinical prediction guides. The results of using the systematic reviews filter in the PubMed Clinical Queries are displayed in (C) Recent Systematic Reviews section with similar format. If available the link to the free article in PMC is included.

Gene's preferred symbol is used. The data supporting the gene-phenotype relationship is built primarily from OMIM® with review from NCBI staff.

Molecular resources

Concepts in MedGen are mapped to several resources in the area of molecular medicine based on data in ClinVar and Gene. This allows users to explore additional molecular information such as related sequence, its location, variations, etc. For example, MedGen provides access to relevant genomic sequences, which are reference standards for well-defined genes reported in [RefSeqGene](#); link to relevant records in [Coriell Institute for Medical Research](#), which provides essential research reagents to the scientific community and links to ClinVar and Gene.

Consumer resources

MedGen also actively seeks and provides direct access to available consumer-friendly information related to each concept. Submitters provide URLs connected either directly or indirectly to Concept IDs. Figure 6 illustrates how these sections are displayed in MedGen.

Additional information

To enrich its content, MedGen provides links to the A.D.A.M. Medical Encyclopedia, which includes over 4,000 articles about diseases, tests, symptoms, injuries, and surgeries with an extensive library of medical photographs and illustrations. Concepts in MedGen are mapped to related Medical Encyclopedia articles via [PubMed Health](#). This resource on prevention and treatment of diseases and conditions for both consumers and clinicians is provided by the NCBI and is specialized in reviews of clinical effectiveness, with easy-to-read summaries for consumers as well as full technical reports. Particular links to PubMed Health may be provided under a Medical Encyclopedia section or under an Outreach and support section. Additionally, if available, links are provided to [ClinicalTrials.gov](#). This web-based resource is maintained by the National Library of Medicine (NLM) at the National Institutes of Health (NIH) and provides patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies of human participants conducted around the world. Figure 7 illustrates how these sections are displayed in MedGen.

MedGen data is further supplemented by providing links to curated relevant reviews for each concept. The reviews can be selected from [GeneReviews™](#) or reviews in PubMed. The [GeneReviews™](#) are exclusively published online on NCBI's Bookshelf and are expert-authored, peer-reviewed disease descriptions presented in a standardized format and focused on clinically relevant and medically actionable information on the diagnosis, management, and genetic counseling of patients and families with specific inherited conditions. [PubMed Clinical Queries](#) provides an interactive interface to discover citations for medical genetics content such as systematic reviews, meta-analyses, reviews of clinical trials, evidence-based medicine, consensus development conferences, and

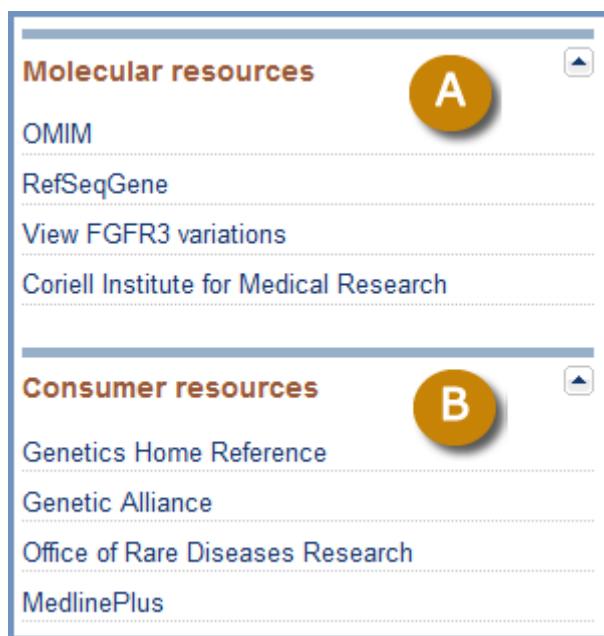


Figure 6. The Molecular and Consumer resources in MedGen. If available, concept specific links to (A) Molecular resources and (B) Consumer resources are provided in the discovery panel at the right.

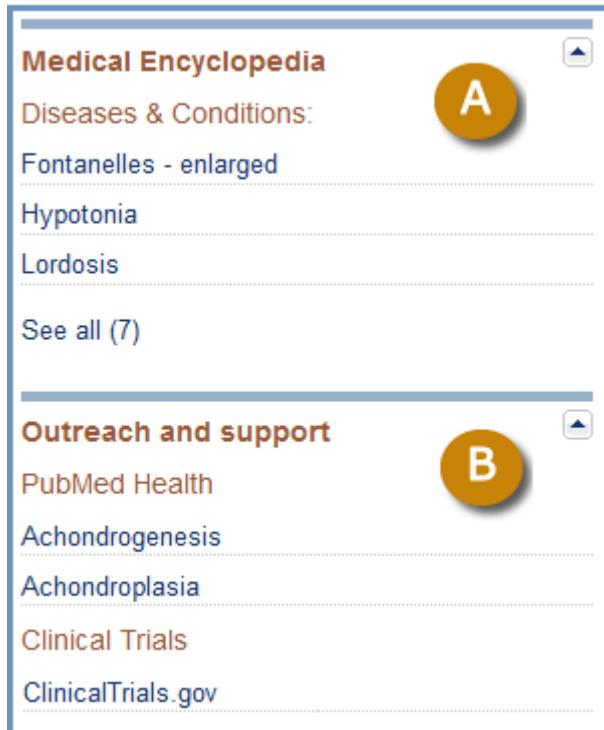


Figure 7. The Medical Encyclopedia and Outreach and support in MedGen. If available, concept specific links to articles in (A) Medical Encyclopedia and (B) PubMed Health and Clinical Trials are provided in the discovery panel at the right.

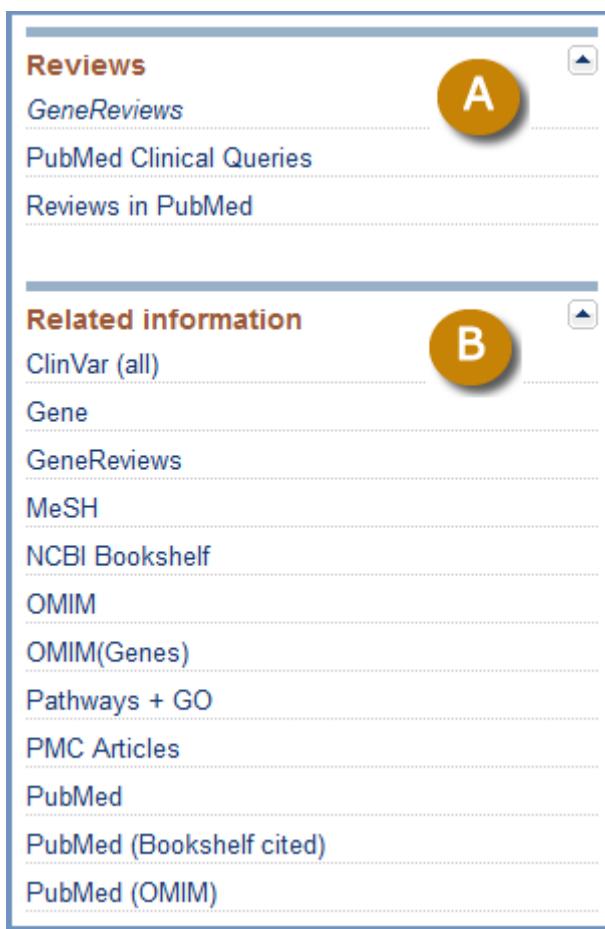


Figure 8. The Reviews and Related information in MedGen. (A) Concept specific Reviews are presented in three different groups as *GeneReviews*, PubMed Clinical Queries, and Reviews in PubMed. (B) Related information displays a list of available links to various NCBI databases, which varies depending on the concept.

guidelines. Other reviews in PubMed are retrieved by querying the concept term in PubMed and limiting the results to human species and setting review as publication type.

Furthermore, MedGen facilitates user's access to related information for a single topic in various NCBI resources by creating reciprocal links between MedGen and other NCBI databases. These links are computed by NCBI's query retrieval system and offer ample opportunities to explore different aspects of a topic based on available related information, such as sequence variations and its relationship to human health (i.e., clinical assertions for a particular disease or particular gene) in ClinVar; gene-specific connections for map, sequence, expression, structure, function, citation, and homology data in Gene; interacting genes, proteins, biomarkers, drugs, and small molecules in Pathways; etc. Some of these links are computed by using a third resource to streamline the connection. For example, the links to PubMed are generated based on general queries to PubMed as well as special queries derived from citations in other sources such as

GeneReviews[™], Medical Genetics Summaries, OMIM[®], etc. Figure 8 illustrates how Reviews and Related information sections are displayed in MedGen.

Access

MedGen can be accessed on the web at <http://www.ncbi.nlm.nih.gov/medgen/>, which gives users options to find, view information, and learn more about medical genetics by conducting basic and advanced searches. MedGen data can be downloaded based on user specific interest at [MedGen FTP site](#) or accessed programmatically via E-utilities.

Web interface

MedGen uses a powerful search and retrieval system developed by NCBI (Entrez) to search and retrieve data from MedGen and other integrated databases at NCBI. Basic queries can simply be submitted either by entering free text or entering field values followed by a proper field qualifier such as [gene], [mim], [moddate], etc., in the search bar. MedGen supports constructing complex queries by providing [Limits](#) and [Advanced](#) options. Users also can take advantage of a spell-check dictionary and Search History function to refine their search terms more precisely. Users can select terms from the suggested list of search terms provided by dictionary as they enter their search keywords. They can use [Limits](#) to restrict their query by chromosome, relationships to other NCBI databases, types of content, and/or sources of the terms. Or use the [Advanced](#) function to combine concepts and/or different fields in their search terms. Search History allows the users to take advantage of adding more restrictions step by step to a previous query.

Detailed instructions on using the search interface are provided in the [MedGen Help Documentation](#).

Search results in MedGen are first presented in a summary format, in order of relevance and with 20 items per page as default. The format is flexible and can be modified from summary to UI List, XML, or text. The number of items per page also can be adjusted based on users' preferences. An easy access to the unique MedGen identifier, Concept ID, and semantic type for each concept is ensured by displaying the identifiers below the name. In a full report users can access all metadata available for a concept and explore all links and supplementary information aggregated in MedGen. The table of contents in the upper right corner and collapsible sections ease the navigation in accessing desired data (Figure 9).

If an old Concept ID or term is used as a search term, information for the new merged record will be retrieved and displayed.

Clicking on the title of a search result provides a full display. For more details of what is accessed from the full display, please see the figures in the Dataflow section.

The screenshot shows a "Table of contents" window with a blue header bar. Below the header, there is a vertical list of seven items, each preceded by a small blue downward arrow indicating it is a link:

- Disease characteristics
- Additional description
- Clinical features
- Term Hierarchy
- Professional guidelines
- Recent clinical studies
- Recent systematic reviews

Figure 9. The Table of contents in MedGen. Content of each full report is organized in several sections listed in the table of contents and hyperlinked for easy access.

FTP Download

All Concepts in MedGen can be downloaded via MedGen's [FTP site](#), which is updated weekly every Wednesday. The metadata included with each concept are presented in multiple files. All of these files are in text format and provided as compressed zip files. Vertical bar (|) is used as the delimiter and the column names are declared in the first line of each file.

To assist users in tracking retired or merged Concepts IDs these concepts are reported as paired IDs in MERGED.RRF file. Semantic types for each concept are reported in MGSTY.RRF file and concept definitions alongside of the source of the definition are provided in MGDEF.RRF file. Concept names are stored in NAMES.RRF file. In order to verify if a term is a preferred term from a vocabulary source, the users can look up the ISPREF field (either "Y" or "N") in MGCONSO.RRF file. In this file one can also find any identifier asserted by the source, abbreviation for the source, type of term as defined by the source, etc.

The relationships between concepts are provided in MGREL.RRF. Concepts may have one or multiple relationship labels (i.e. one concept can be a child, a parent, a sibling, etc.). Summary data for each concept identifier is provided in MGCONSO.RRF. Each concept may have many attributes, which all are summarized in MGSAT.RRF file.

By combining data from MGSAT.RRF and MGCONSO.RRF users can work out paths to many of the connections that MedGen has made with its external resources. For example, connections between HPO Primary IDs and MedGen concepts are maintained in MGCONSO.RRF (using the first column (CUI), the 9th column (SAB = HPO), and the 8th column (SDUI) for the ID asserted by HPO). When HPO staff retire or merge a Primary ID in HPO and report it as an Alternative ID, MedGen will report the Alternative ID as an attribute (HPO_ALT_ID) for that concept. Therefore, users can find

those Alternative IDs in MGSAT.RRF under ATV in the 8th column. For example, “HP:0003122” is reported as an alternative ID for Glycosuria (“HP:0003076”) by HPO. MedGen reports this Alternative ID in MGSAT.RRF as an attribute for Glycosuria (“C0017979”). Thus, in MGSAT.RRF users can find a line for “C0017979”, which has HPO_ALT_ID as attribute name (ATN, the 6th column) and “HP:0003122” as attribute value (ATV, the 8th column).

Another example is the connection made between OMIM® IDs and CUIs. OMIM® IDs are reported for concepts having OMIM® as their source vocabulary and either of “term types” (TTY) of “Preferred name” (PT), “synonym” (SYN), or “acronym” (ACR). However, some of the records in OMIM® have a Gene Phenotype Relationships section, which reports MIM numbers for genes associated with that record. Since genes are not in the scope for MedGen, the connections are maintained by gene symbols via the NCBI Gene database.

MedGen assist users in obtaining the mapped data connecting MedGen concepts to HPO and MedGen concepts to HPO and OMIM® by providing two additional data files (MedGen_HPO_Mapping.txt and MedGen_HPO_OMIM_Mapping.txt respectively).

Users also can download a complete list of links between MedGen concepts and literature from PubMed in medgen_pubmed file (i.e., CUI, PMID connection). This allows users to have access to all relevant literature regardless of the rules used to create these connections.

MedGen's updates and major releases can be followed through [MedGen RSS feed](#) announcements.

E-utilities

Entrez provides a series of programming utilities as a stable interface into the Entrez query and database system. These utilities allow using a fixed URL syntax, which translates the input parameters into a query request for search and retrieval. MedGen has enabled use of E-utilities for its database via esearch and summary, but not efetch. For a basic text search users can simply place their query term at the end of the following URL (replacing <query_term>): http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=medgen&term=<query_term>

References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267–70. PubMed PMID: 14681409.
2. Fung KW, Bodenreider O. Utilizing the UMLS for semantic mapping between terminologies. AMIA Annu Symp Proc. 2005.:266–70. PubMed PMID: 16779043.
3. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009 Jan;37(Database issue):D793–6. doi: 10.1093/nar/gkn665. doi:pub 2008 Oct 8. PubMed PMID: 18842627.

4. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* 2008 Nov;83(5):610–5. doi: [10.1016/j.ajhg.2008.09.017](https://doi.org/10.1016/j.ajhg.2008.09.017). doiEpub 2008 Oct 23. PubMed PMID: 18950739.
5. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, Hem V, Gorelenkov V, Song G, Wallin C, Husain N, Chitipiralla S, Katz KS, Hoffman D, Jang W, Johnson M, Karmanov F, Ukrainchik A, Denisenko M, Fomous C, Hudson K, Ostell JM. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Research.* 2013;41(D1):D925–D935. doi: [10.1093/nar/gks1173](https://doi.org/10.1093/nar/gks1173). PubMed PMID: 23193275.

Genes and Gene Expression

Genes and Gene Expression

Donna Maglott, PhD,¹ Tanya Barrett, PhD,¹ Terence Murphy, PhD,¹ Michael Feolo, PhD,¹ Lukas Wagner, PhD,¹ and Richa Agarwala, PhD¹

Created: November 7, 2013.

Scope

Gene Overview

NCBI maintains information about genes primarily in two contexts. One context is defined by public sequence information, such as annotation of RefSeqs (see RefSeq chapter) or linking with records in the International Nucleotide Sequence Database Consortium or INSDC (see Genome Reference Consortium chapter). Connection of sequence information to a GeneID or a UniGene cluster identifier is critical to any analysis of gene expression. The second context for defining a gene is by mapped phenotype. GeneIDs are not assigned to mapped loci for all taxa, but when they are, the expectation is that the genes will eventually be connected to sequence as the molecular basis for the phenotype is defined.

Once an identifier is assigned to the concept of a gene, multiple databases connect information to that concept. Within NCBI, these databases include Gene, for primary data about the gene and portals to information about its expression, products, homologs, and phenotypes; BioSystems for pathways involving its products; GEO (see GEO chapter) and UniGene for information about expression; Bookshelf, PubMed and PubMedCentral for publications; dbGaP, PheGenI, MedGen and OMIM for phenotypes; HomoloGene for homology; dbSNP, dbVar, and ClinVar for variation; and Taxonomy for information about the organism. In other words, there are many resources at NCBI that maintain information about genes, but this section focuses on these:

- Gene
- HomoloGene
- UniGene

Expression Overview

Several resources at NCBI maintain primary data about the tissues, health states, and developmental stage or age in which genes are expressed, or the sequence variation that affects their expression. The data archives reflect the primary methods by which these data have been generated, starting with sampling cDNA sequences from non-normalized libraries in UniGene, through array or RNAseq based approaches in GEO, to association

¹ NCBI.

data in GTex. These resources also maintain tools to analyze the datasets, which can be quite large.

History

Genes

Representation of genes as objects with stable identifiers began at NCBI in 1995 with the clustering of 3' untranslated regions from GenBank release 88 into gene-specific sets as UniGene. (1) . When the RefSeq project got underway in 1998, a collaboration developed among the human nomenclature committee (now HGNC), OMIM, and the RefSeq team to aggregate gene-specific information for tracking. This developed into LocusLink (2) which evolved into Gene in 2004(3). In the late 1990's and into the early 2000's, most eukaryotic genes identified by sequence information were characterized by sequences assumed to represent gene expression, namely cDNAs. Now, however, many genes are first identified computationally rather than by direct sequence evidence, namely by gene prediction software that may use direct experimental results, but may also calculate which regions of a genomic sequence are likely to be a gene based on comparison to related species or analysis of predicted proteins.

HomoloGene

Examination of a gene's function is facilitated by evaluation across multiple species, HomoloGene (see HomoloGene chapter), launched as a distinct resource in 2000, is designed to facilitate these analyses by grouping genes according to homology, providing tools for comparisons, and aggregating data for these homology groups.

Gene Expression

Methods of identifying regions of the genome that are transcribed have changed over the years. Large-scale cDNA projects such as I.M.A.G.E. (4) and the Mammalian Gene Collection (MGC) (5) determined what sequences were expressed based on cDNA cloning and sequencing of those clones. Given those sequences, array-based techniques were used to compare expression of sequences under different experimental conditions. Now, with the power of RNAseq, expression can be analyzed qualitatively and quantitatively without requiring a cloning step.

Sequence variation that affects gene expression is also being evaluated (6). PheGenI provides a window to these data.

Data Model

When a gene is defined by sequence, map location, or a nomenclature group, it is assigned a stable identifier for tracking, the GeneID. The connection between the GeneID and nucleotide or protein sequence is used by many of NCBI's databases to represent their data in the context of a gene. Thus you will see links to Gene, or GeneIDs annotated, in

numerous sites at NCBI. HomoloGene, by grouping genes computed or curated to be homologs, is one of those sites.

Dataflow

Establishment of records in Gene, and evaluation of sequence expression, occur independently. In other words, there are many sequences in GEO, Nucleotide, Protein, or UniGene that do not cross reference a record in Gene. For genomes annotated by NCBI (see Eukaryotic and Prokaryotic genome annotation chapters) connections between sequence and GeneIDs are evaluated with each Annotation Release, thus narrowing some of the gaps.

References

1. Boguski MS, Schuler GD. ESTablishing a human transcript map. *Nat Genet*. 1995 Aug; 10(4):369–71. PubMed PMID: 7670480.
2. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. *Nucleic Acids Res*. 2000 Jan 1;28(1):126–8. PubMed PMID: 10592200.
3. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2005 Jan 1;33(Database issue):D54–8. PubMed PMID: 15608257.
4. Lennon G, Auffray C, Polymeropoulos M, Soares MB. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics*. 1996 Apr 1;33(1):151–2. PubMed PMID: 8617505.
5. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Snead A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnurch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J; MGC Project Team. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian GeneCollection (MGC). *Genome Res*. 2004 Oct;14(10B):2121–7. Erratum in: *Genome Res*. 2006 Jun; 16(6):804. Morrin, Ryan [corrected to Morin, Ryan]. PubMed PMID: 15489334.
6. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013 Jun;45(6):580–5. PubMed PMID: 23715323.

Gene Expression Omnibus (GEO)

Tanya Barrett, Ph.D.¹

Created: May 19, 2013.

Scope

The [Gene Expression Omnibus \(GEO\)](#) is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data sets (1). Approximately 90% of the data in GEO are gene expression studies that investigate a broad range of biological themes including disease, development, evolution, immunity, ecology, toxicology, metabolism, and more. The non-expression data in GEO represent other categories of functional genomic and epigenomic studies including those that examine genome methylation, chromatin structure, genome copy number variations, and genome–protein interactions. A breakdown of GEO data types and technologies is provided on the repository [Summary](#) page.

Data in GEO represent original research submitted by the scientific community in compliance with grant or journal provisos that require data to be made available in a public repository, the objective being to facilitate independent evaluation of results, reanalysis, and full access to all parts of the study. The resource supports archiving of all parts of a study including raw data files, processed data, and descriptive metadata, which are indexed, cross-linked, and searchable. While the principal role of GEO is to serve as a primary data archive, the resource also offers several tools and features that allow users to explore, analyze, and visualize expression data from both gene-centric and study-centric perspectives.

To summarize, the main goals of GEO are to:

- Provide a robust, versatile primary data archive database in which to efficiently store a wide variety of high-throughput functional genomic data sets.
- Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community.
- Provide user-friendly mechanisms that allow users to locate, review, and download studies and gene expression profiles of interest.

History

The post-genomic era has led to a multitude of high-throughput methodologies that generate massive volumes of gene expression and other types of functional genomic and epigenomic data. The GEO database was established in 2000 to archive the burgeoning

¹ NCBI.

volumes of microarray gene expression data beginning to be produced by the research community at that time (2). Furthermore, as microarrays became used routinely in almost every area of biological research, many journals adopted the requirement that the microarray data discussed in manuscripts should be deposited in a public repository so that anyone could freely access and critically evaluate the data. Today, GEO archives data for approximately 40,000 studies comprising a million samples, for over 2200 organisms, submitted by 15,000 laboratories from around the world. Users can track database growth on the GEO [History](#) page.

Since its inception, many aspects of the GEO database and operating procedures have undergone major revisions and development, including: increasingly stringent submission requirements, concomitant with developing community standards like MIAME (Minimum Information About a Microarray Experiment) (3); enhanced submission formats that ease the burden on submitters and promote well-annotated MIAME-compliant data; improved indexing and analysis tools that help users more easily locate information relevant to their interests; and database modifications to support evolving data types, including next-generation sequence data.

Data Model

The GEO database archives a wide variety of rapidly evolving, large-scale functional genomic experiment types. These studies generate data of many different file types, formats, and content that consequently present considerable challenges in terms of data handling and querying. The core GEO database has built-in flexibility to accommodate diverse data types. Notably, tabular data are not fully granulated in the core database. Rather, they are stored as plain text, tab-delimited tables that have no restrictions on the number of rows or columns allowed. However, some columns reserve special meaning, and data from these are extracted to secondary resources, including the GEO Profiles database, and used in downstream query and analysis applications such as [GEO2R](#). Accompanying raw data files are stored on an FTP server and linked from each record. Contextual biological descriptions, protocols, references, and other metadata are stored in designated fields within a relational MSSQL database.

An outline of the GEO data structure is presented in Figure 1. The data are organized into the following entities:

Platform

A Platform record contains a description of the array or sequencer and, for array-based Platforms, a data table defining the array template. The information within Platform records is supplied by submitters or commercial vendors. Each Platform record is assigned a unique and stable GEO accession number with prefix GPL. A Platform may reference many Samples that have been deposited by multiple submitters. Platforms are indexed and searchable using the Entrez GEO DataSets interface.

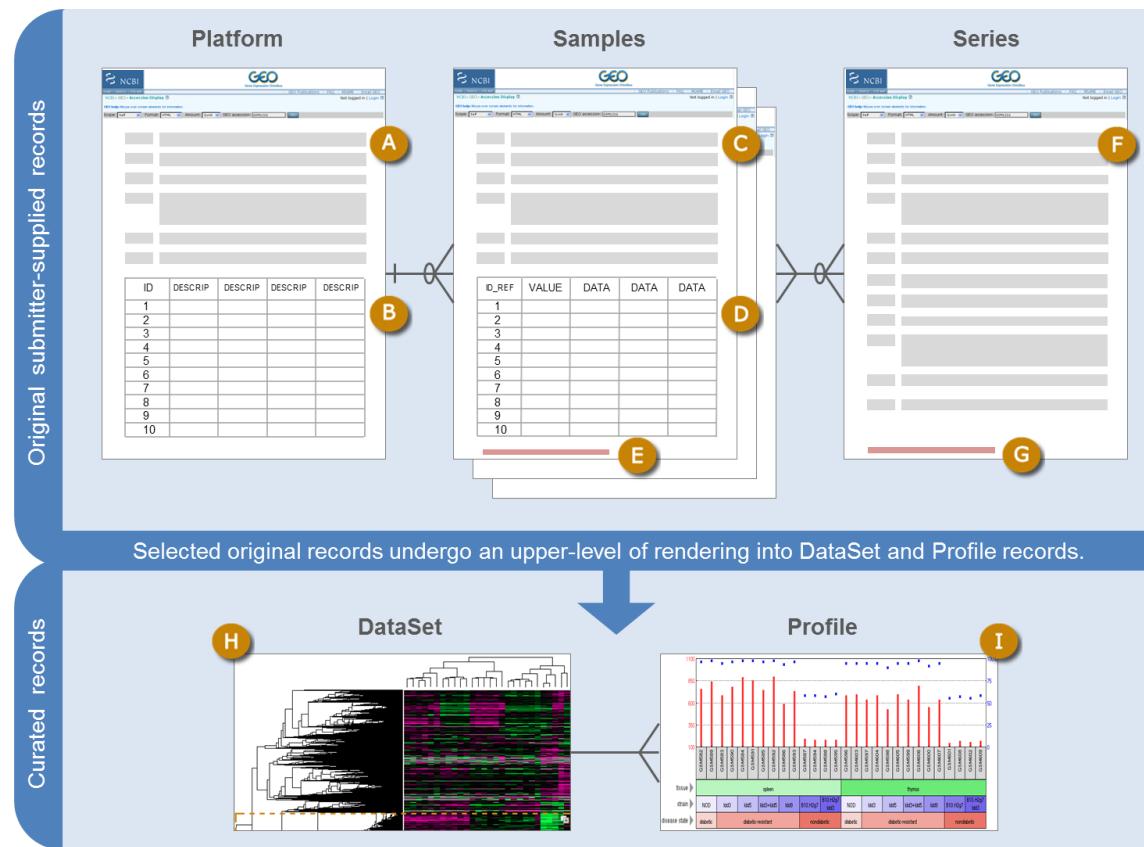


Figure 1. Sketch of GEO data organization. Platform, Sample, and Series records are created from the data supplied by submitters. These records contain: (A) a description of the array or sequencer; (B) a tab-delimited table of the array template definition; (C) a description of the biological sample and protocols to which it was subjected; (D) a tab-delimited table of processed array hybridization results; (E) links to original raw data file, or processed sequence data file; (F) a description of the overall study, study design, and citation information; (G) links for bulk downloading data for the entire study. At periodic intervals, selected records undergo further processing into curated GEO DataSet (H) and GEO Profile (I) records that help users analyze and visualize gene expression.

Sample

A Sample record contains a description of the biological material and the experimental protocols to which it was subjected. A data table with normalized abundance measurements for each feature on the corresponding Platform is usually included, as well as links to corresponding raw data files. The information within Sample records is supplied by submitters. Each Sample record is assigned a unique and stable GEO accession number with prefix GSM. A Sample entity must reference only one Platform and may be included in multiple Series. Samples are indexed and searchable using the Entrez GEO DataSets interface.

Series

A Series record links together a group of related Samples and provides a focal point and description of the whole study. The information within Series records is supplied by submitters. Each Series record is assigned a unique and stable GEO accession number with prefix GSE. Series are indexed and searchable using the Entrez GEO DataSets interface.

DataSet

The submitter-supplied Platform, Sample, and Series data are very heterogeneous with regards to the style, content, and level of detail with which the studies are described. But despite this diversity, all array-based gene expression submissions share a common core set of elements:

- sequence identity tracking information of each feature on the Platform
- normalized expression measurements within Sample tables
- text describing the biological source of the sample and the study aim

Through a procedure that employs both automated data extraction and manual curation, these three categories of information are captured from the submitter-supplied records and organized into an upper-level record called a curated GEO DataSet. A DataSet comprises a description of the study as well as consistently processed and comparable Samples that are categorized according to experimental variables. Each DataSet record is assigned a unique and stable GEO accession number with prefix GDS. DataSets are indexed and searchable using the Entrez GEO DataSets interface.

Profiles

GEO Profiles are derived from GEO DataSets. A GEO Profile is a gene-centered representation of the data that presents gene expression measurements for one gene across a DataSet. Profiles are indexed and searchable using the Entrez GEO Profiles interface.

Information about how to use and interpret GEO DataSets and Profiles is provided in the Access section.

Dataflow

Researchers typically initiate a data deposit to GEO before a manuscript describing the study has been submitted to a journal for review. Researchers use their MyNCBI account to login and register submissions. Several submission formats are supported including spreadsheets and XML, see the full [submission instructions](#).

All deposits undergo syntactic validation as well as review by a GEO curator to ensure that data are organized correctly and contain sufficient information to interpret the study. If content or structural problems are identified, the curator works with the submitter until the issue is resolved. Once the data pass review, stable GEO accession numbers are

assigned and can be cited in the manuscript. Researchers usually keep their GEO records private until the corresponding manuscript is published. During this period, researchers have the option to generate a reviewer URL that grants anonymous, confidential access to their private data, which can be provided to journal editors for review purposes.

Upon release, Platform, Sample, and Series records are indexed in the Entrez [GEO DataSets](#) database where users can query and download the data, or perform a gene expression analysis using the [GEO2R](#) comparison tool. Some components of GEO submissions are brokered to other NCBI databases, including original next generation sequence reads to SRA and study descriptions to BioProject, with reciprocal links back to GEO as appropriate.

At approximately monthly intervals, selected Series undergo further processing by curators to create GEO DataSet and GEO Profile records. GEO DataSet records represent curated collections of biologically and statistically comparable GEO Samples, and are indexed in the [Entrez GEO DataSets](#) database. GEO Profiles are derived from GEO DataSets and depict expression measurements for an individual gene across all Samples in that DataSet. Profiles are indexed in the [Entrez GEO Profiles](#) database.

Access

GEO has a suite of tools that allow users to browse, download, query, analyze, and visualize data relevant to their specific interests.

Browse

The [GEO repository browser](#) has tabs containing tables that list Series, Sample, Platform, and DataSet records. The tables include information that can be searched and filtered, as well as links to related records and supplementary file downloads. The tables can be exported and include further information not displayed on the browser, including corresponding PubMed identifiers and related SRA accessions.

Download

All the data in GEO can be downloaded in a variety of formats using a variety of mechanisms (see the [Download GEO data](#) documentation). Options include bulk data download directly from the [FTP site](#) or from links on records using the ‘Send to: File’ feature on Entrez GEO DataSet or Entrez GEO Profiles retrievals or programmatically using E-utilities.

Query

NCBI has a powerful search and retrieval system called Entrez that can be used to search the content of its network of integrated databases. GEO data are available in two separate Entrez databases referred to as [Entrez GEO DataSets](#) and [Entrez GEO Profiles](#). A typical workflow is for the user to first identify studies of interest by querying Entrez GEO DataSets, and then use either GEO2R or GEO Profiles to identify specific genes or gene

expression patterns within that study. Alternatively, the user can query the Entrez GEO Profiles database directly to retrieve the expression patterns of a specific gene across all curated GEO DataSets. A rich complement of Entrez links is generated to connect data to related information: inter-database links reciprocally connect GEO to other NCBI resources such as PubMed, GenBank, and Gene; intra-database links connect genes related by expression pattern, chromosomal position, or sequence. Both databases are extensively indexed under many separate fields, meaning that users can refine their searches by constructing fielded queries. The *Faceted search* tool, located on the left side of retrievals, can help users filter and refine their Entrez results, and the *Advanced* search tools enable generation of complex multipart queries or combinations of multiple queries to find common intersections in search results.

Query Entrez GEO DataSets

This database stores descriptions of the original submitter-supplied Platform, Sample, and Series records as well as curated DataSets. The Entrez GEO DataSets database can be searched using many different attributes including keywords, organism, study type, and authors. More information about how the results are displayed and supported query fields is provided at [About GEO DataSets](#) and [Querying GEO DataSets](#) pages. Example queries include:

Retrieve studies that investigate the effect of smoking or diet on non-human mammals <code>(smok* OR diet) AND (mammals[organism] NOT human[organism])</code>
Search for studies that examine gene expression using next-generation sequencing <code>"expression profiling by high throughput sequencing"[DataSet Type]</code>
Retrieve submissions that have Affymetrix CEL files <code>cel[Supplementary Files]</code>
Search for curated DataSets that have 'age' as an experimental variable <code>age[Subset Variable Type]</code>
Retrieve studies that consist of between 100 and 500 samples <code>100:500[Number of Samples]</code>
Retrieve studies that include 'Smith, A.' as an author <code>smith a[Author]</code>

Query Entrez GEO Profiles

This database stores gene expression profiles derived from curated GEO DataSets. Each Profile is presented as a chart that displays the expression level of one gene across all Samples within a DataSet. Experimental context is provided in the bars along the bottom of the charts making it possible to see at a glance whether a gene is differentially expressed

across different experimental conditions. The Entrez GEO Profiles database can be searched using many different attributes including keywords, gene symbols, gene names, GenBank accession numbers, or Profiles flagged as being differentially expressed. More information about how the results are displayed and supported query fields is provided at [About GEO Profiles](#) and [Querying GEO Profiles](#) pages. Example queries include:

Retrieve all gene expression profiles for CYP1A1
CYP1A1[Gene Symbol]
Retrieve gene expression profiles of CYP1A1 or ME1 in DataSets that investigate the effects of smoking or diet
(CYP1A1[Gene Symbol] OR ME1[Gene Symbol]) AND (smok* OR diet)
Retrieve gene expression profiles for all kinases in the DataSet with accession number GDS182
kinase[Gene Description] AND GDS182
Retrieve all gene expression profiles for genes that have the Gene Ontology(GO) term 'apoptosis' in the DataSet with accession number GDS182
apoptosis[Gene Ontology] AND GDS182
Retrieve gene expression profiles for genes that lie within base range 10000:3000000 on chromosome 8 in mouse
(8[Chromosome] AND 10000:3000000[Base Position]) AND mouse[organism]
Retrieve genes that exhibit differential expression in DataSets that examine the effect of an agent
agent[Flag Information] AND "value subset effect"[Flag Type]

GEO BLAST

Another way to query the GEO Profiles database is by nucleotide sequence similarity using [GEO BLAST](#). The GEO BLAST database contains all GenBank sequences represented on microarray Platforms that participate in curated DataSets. This BLAST interface performs a standard BLAST sequence search and retrievals point to corresponding gene expression profiles in the GEO Profiles database. This interface is helpful in identifying gene expression information for sequence homologs of interest, e.g., related gene family members or for cross-species comparisons.

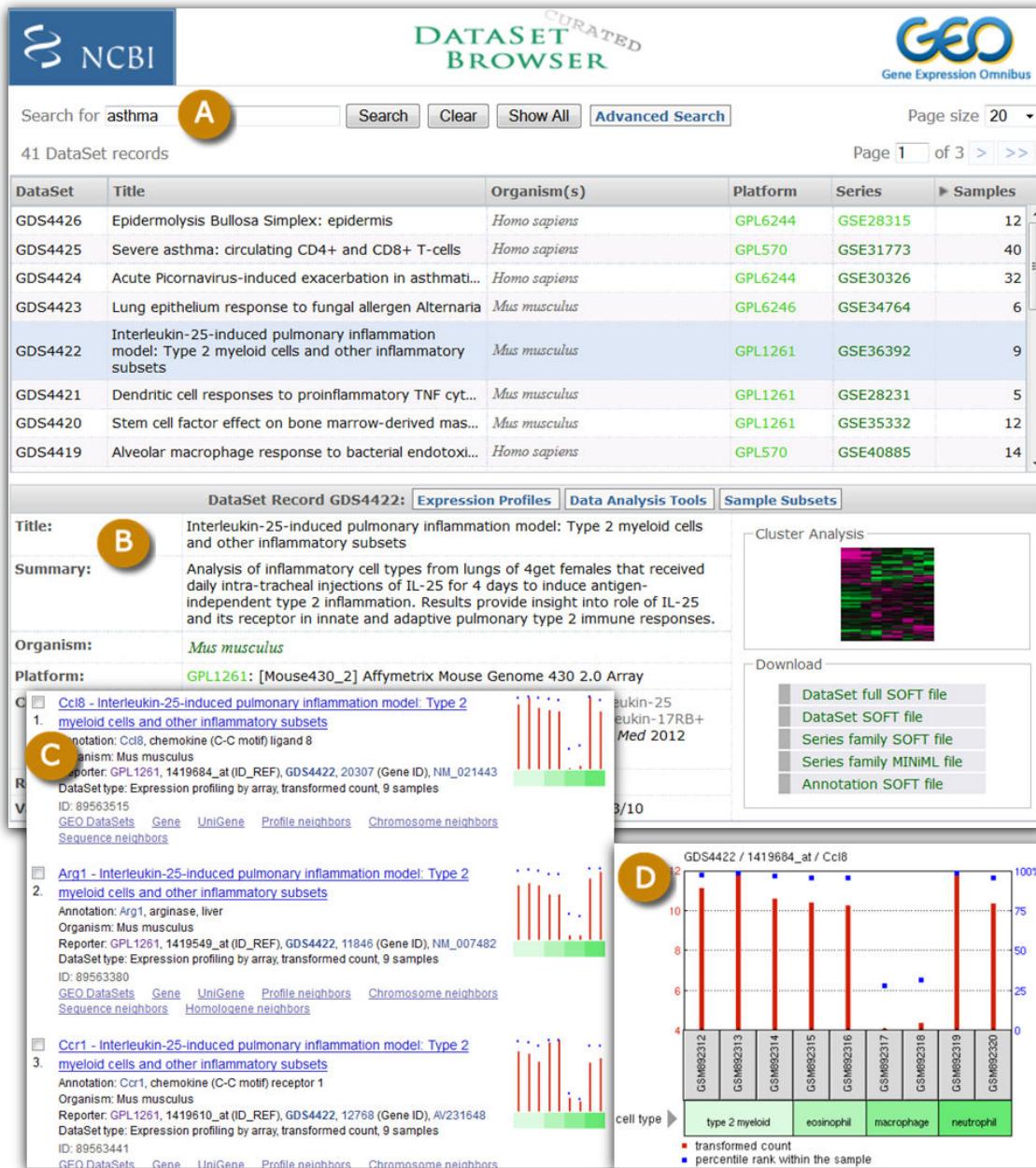


Figure 2. A selection of GEO DataSet and Profile screenshots. The DataSet Browser (A) enables simple keyword searches for curated GEO DataSets. When a DataSet is selected, a window appears (B) that contains detailed information about that DataSet, download options, and links to analysis features including gene expression profiles in Entrez GEO Profiles (C). Each expression profile chart can be viewed in more detail to see the activity of that gene across all Samples in that DataSet (D).

Analyze and visualize

In addition to being able to query and locate specific studies and genes of interest as described above, users may choose to examine those studies further and identify genes

that have particular expression characteristics such as being highly expressed in one type of experimental condition compared to another, or having similar expression patterns to a selected profile of interest. GEO provides several tools and graphical renderings that facilitate interpretation and visualization of microarray-based gene expression data. These tools do not require specialized knowledge of microarray analysis methods, nor do they require time-consuming download or processing of large sets of data.

DataSet Analysis Tools

Several features are provided on curated DataSet and Profile records to assist with identification of genes of interest. These include:

- Pre-calculated, interactive [cluster heatmap](#) images that help detect natural groups of coordinately regulated genes. Areas of the cluster can be selected and underlying expression values downloaded or exported to Entrez GEO Profiles.
- A ‘[compare 2 sets of samples](#)’ tool that offers rudimentary Student’s t-test analysis to locate differentially expressed genes between two sets of samples.
- A [Find genes](#) feature that retrieves genes that have been flagged as being differentially expressed according to specific experimental variables.
- [Boxplot images](#) that display the distribution of expression values together with experimental design are useful for quality control checks.
- Various categories of [Neighbors](#) on GEO Profiles, which, e.g., connect genes that show a similar expression pattern to the chosen Profile, or retrieve Profiles from across all DataSets that are related by Homologene group.

GEO2R

GEO2R is an interactive online tool that allows users to perform a sophisticated R-based analysis of GEO data to help identify and visualize differential gene expression. Unlike GEO’s DataSet Analysis Tools described above, GEO2R does not rely on curated DataSet records but rather interrogates original submitter-supplied data directly as soon as they are released. GEO2R uses established Bioconductor R packages (4, 5) to transform and analyze GEO data. The application allows users to designate up to 10 groups of Samples to compare, and offers several statistical parameters with which to perform the analysis.

Results are presented as a table of genes ordered by significance. The results table contains various categories of statistics, including P-values, t-statistics, and fold change, as well as gene annotations, including gene symbols, gene names, Gene Ontology (GO) terms, and chromosome locations. The expression pattern of each gene in the table can be visualized by clicking the row to reveal expression profile graphs or the complete set of ordered results can be downloaded as a table. Alternatively, if users are not interested in performing differential expression analysis but rather only want to see the expression profile of a specific gene within a study, they can bypass all the above and simply enter the Platform gene ID to visualize that profile. To assist users replicate their analyses, the native R script generated in each session is provided so it can be saved as a reference for how results were calculated, or used to reproduce GEO2R top genes results. A [YouTube video tutorial](#) demonstrating GEO2R functionality is available.

References

1. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* 2013;Jan41(Database issue):D991–5. PubMed PMID: 23193258.
2. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;Jan 30(1):207–10. PubMed PMID: 11752295.
3. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet.* 2001;Dec29(4):365–71. PubMed PMID: 11726920.
4. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics.* 2007;Jul 1523(14):1846–7. PubMed PMID: 17496320.
5. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 2004;5(10):R80. PubMed PMID: 15461798.

Gene

Donna Maglott, PhD, Kim Pruitt, PhD, Tatiana Tatusova, PhD, and Terence Murphy, PhD

Created: November 14, 2013.

Scope

NCBI's Gene database is designed to aggregate gene-specific information from multiple perspectives including sequence, mapping, publications, function of products, expression, evolution, and consequences of variation. Gene makes these data available for diverse scenarios, from occasional interactive access on the Web through computational access to selected or complete data sets.

Gene assigns an identifier (the GeneID) for each gene in each taxon either represented in the NCBI Reference Sequence (RefSeq) project, or under consideration by RefSeq. Usually this taxon is defined at the species level, but sometimes will be per isolate, strain or cultivar. Gene is closely coupled with RefSeq, in that genes annotated on RefSeq sequences are assigned GeneIDs for tracking. Not all records in Gene, however, are based on RefSeqs. Gene works closely with multiple groups that may identify a gene before it has been defined by sequence. In other words, some records in Gene are mapped traits or other phenotypes.

This document does not provide detailed instructions about how to use Gene or comprehensive details about how Gene is built from numerous data sources. For detailed, up-to-date documentation, please refer to Gene's [Help document](#).

History

The database currently known as Gene was first made public in 1999 as LocusLink (1). There was only one species represented (human) and little more than 9000 records. The Web interface supported links only to dbSNP, OMIM, RefSeq, GenBank, and UniGene within NCBI, as well as to the now defunct Genome Database (GDB) and a few other databases externally (Figure 1). By late 2003, when Entrez Gene was released, there were 10 species, almost 195000 records, and links computed to dbSNP, Ensembl, the HUGO Gene Nomenclature Committee (HGNC), GEO, Map Viewer, Mammalian Gene Collection (MGC), Nucleotide, Protein, PubMed, Taxonomy, UCSC, UniSTS, UniGene, and multiple species-specific model organism databases (Figure 2). Now Gene represents more than 11,000 taxa, more than 13,000,000 records, and more than 40 types of links to other NCBI databases.

In addition to the taxonomic scope, the number of records, and connectivity, the look and feel of Gene has changed over the years. The current database implementation provides a hierarchical Table of Contents to facilitate navigation, integration with MyNCBI to support [personalized display](#) of sections of the record, an embedded viewer of NCBI's annotation of any gene on one or more genomic RefSeqs, a page dedicated to the display

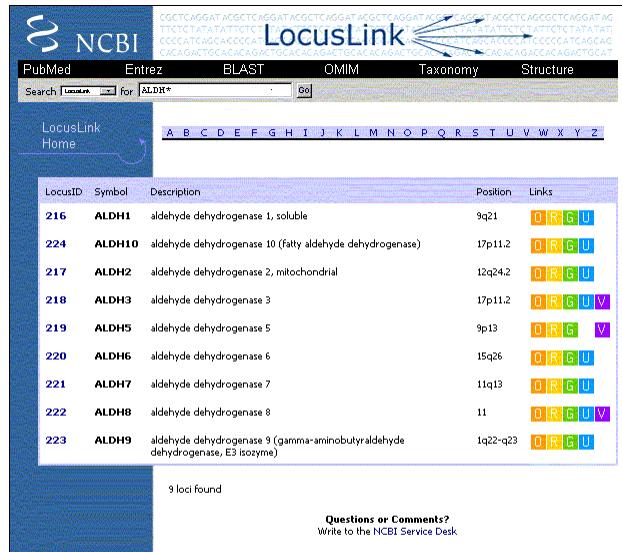


Figure 1. Representation of gene-specific information in LocusLink.

of GeneRIFs, and, especially for human, enhanced access to gene-specific variation and phenotype reports (Figure 3).

Data Model

Gene has a simple data model. Once the concept of a gene is defined by sequence or mapped location, it is assigned a unique integer identifier or GeneID. Then data of particular types are connected to that identifier. These types include sequence accessions, names, summary descriptions, genomic locations, terms from the Gene Ontology Consortium (2), interactions, related phenotypes, and summaries of orthology. For some of the commonly requested elements, and because of the simplicity of the data model, Gene provides tab-delimited files of content anchored on the GeneID.

The full extraction of Gene is exported as binary ASN.1 (ftp://ftp.ncbi.nih.gov/gene/DATA/ASN_BINARY/) with a tool provided to convert to XML (ftp://ftp.ncbi.nlm.nih.gov/asn1-converter/by_program/gene2xml). The ASN.1 representation of a Gene record (http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene asn) incorporates several objects used by other resources (Gene-ref, BioSource, RNA-ref, etc.), but also has several objects specific to Gene to represent the type of gene, map location, and properties. A major component of Gene's ASN.1 representation is the generic Gene-commentary that is used to represent content defined by type, heading, label, text, and source.

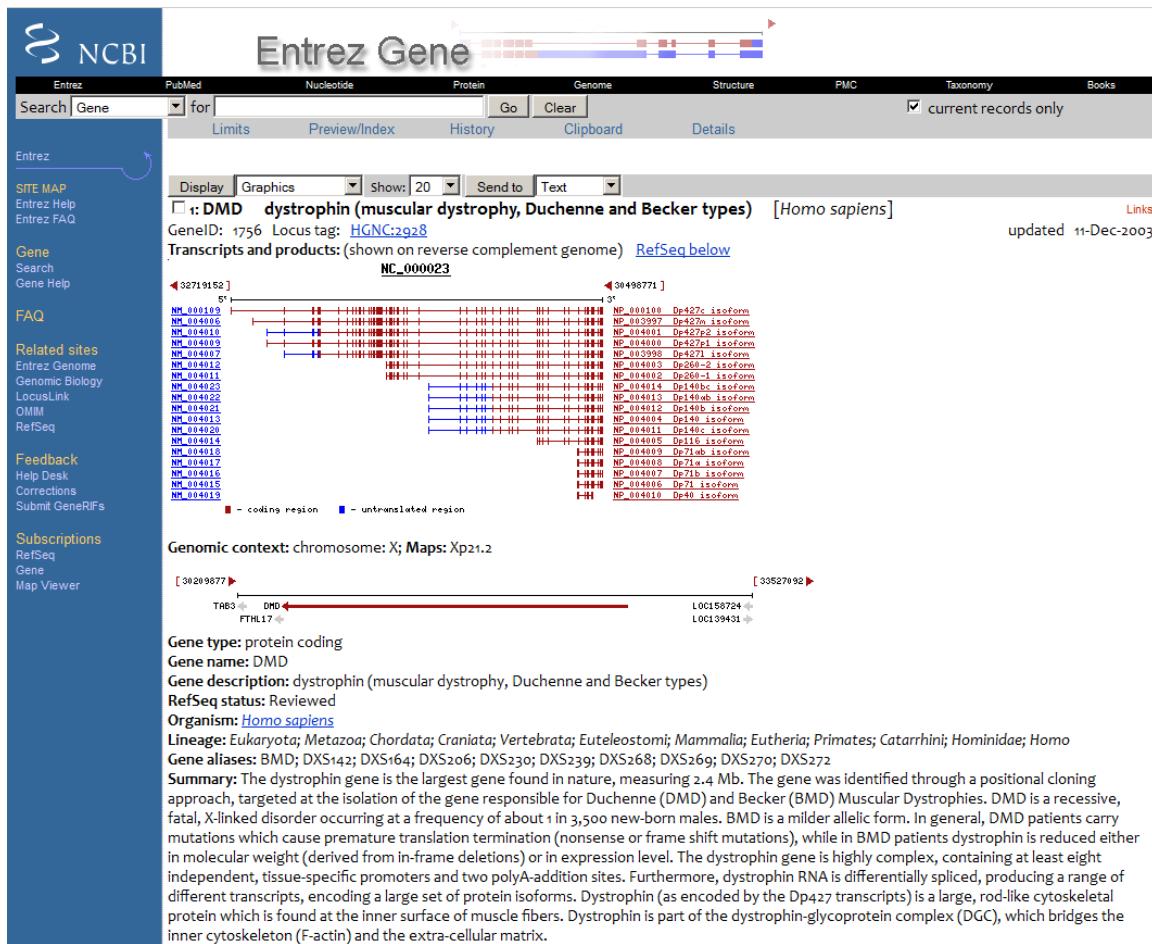


Figure 2. Gene in 2003. The diagram of the gene structure was idiosyncratic to Gene; the organization of the page followed the NCBI conventions of the time by using a blue sidebar at the left to provide general information about Gene and other resources. Links to related data in other databases was accessed by clicking on the Links menu at the upper right.

Dataflow

Gene is updated daily and incrementally. In other words, on any given day a record may be changed but not all records will be changed on the same day. The FTP site is refreshed comprehensively each day, except for special reports and documentation files.

Data are added to Gene by integrating automated and curatorial flows. For some taxa, primarily genomes submitted to NCBI with annotation of genes, data are loaded to Gene by extracting information annotated on the gene feature of the genomic sequence that was submitted. Those data may be supplemented by data from Gene Ontology (GO) based on identifiers in the sequence, according to rules reported by Gene in ftp://ftp.ncbi.nih.gov/gene/DATA/go_process.xml. The content of the Gene record for these species is thus updated only when a new annotation of the genome is supplied, or when supplementary data such as GeneRIFs, GO terms, or UniGene clusters are updated.

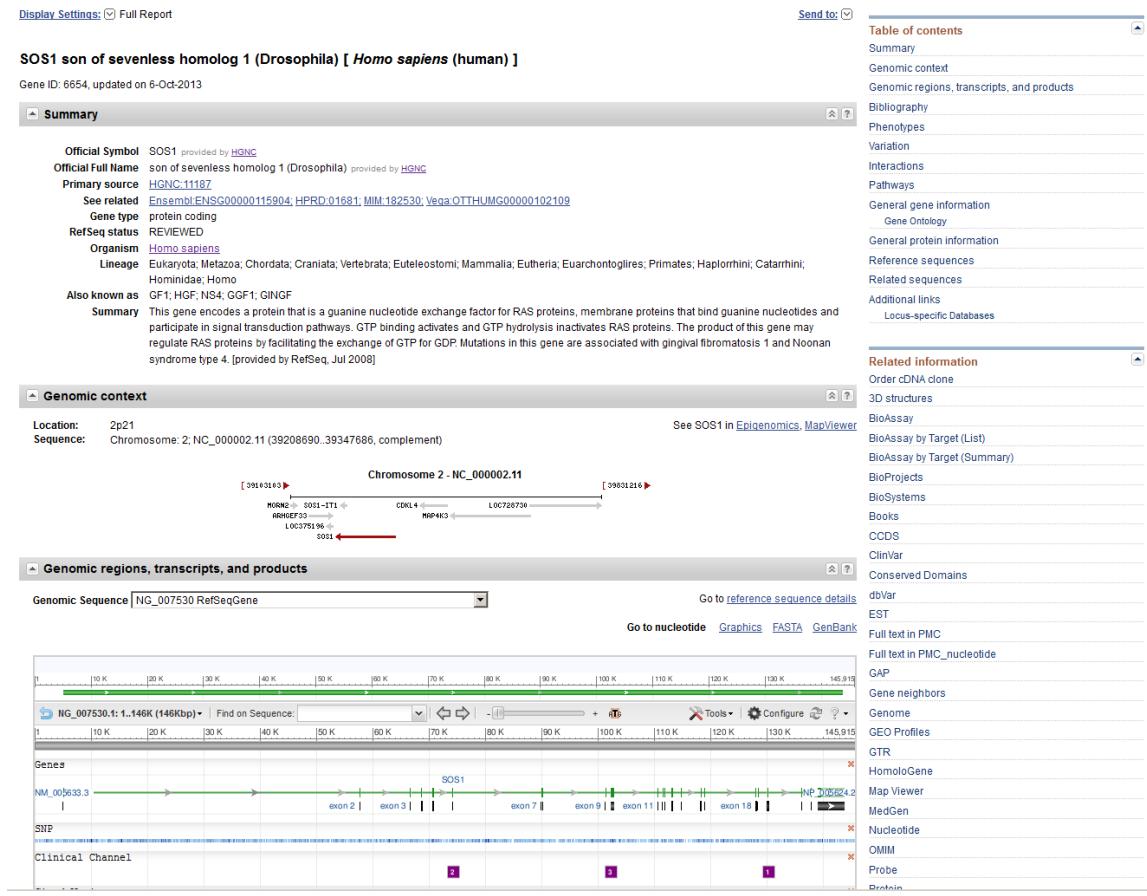


Figure 3. Gene in 2013. Partial display of a record in Gene showing content comparable to that in Figure 2, namely the summary section, the genomic context, and part of the embedded view of the annotation of the gene on a selected genomic sequence. In this example the genomic sequence is a RefSeqGene, and thus shows a more limited set of alternative transcripts, and report the exon numbering system defined by the RefSeqGene.

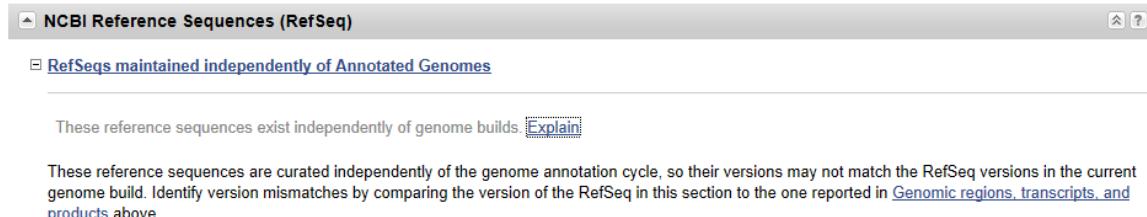


Figure 4. A record maintained independently of annotation releases. If this information is included in the Reference Sequences portion of the Gene record, other content of the record is also likely to change more often.

For the taxa included in RefSeq's curated set (see the RefSeq chapter for more information), updates may happen daily, and independently of a re-annotation of a

genome. There are automated flows to reconcile official gene symbols and full names, protein names, and database identifiers. Curators may modify summaries, add or redefine transcript or RefSeqGene RefSeqs, or add citations to the record. When this happens the Gene record is updated. Genes that are in scope for more frequent updates can usually be detected because the NCBI Reference Sequences section will include a subsection entitled RefSeqs maintained independently of Annotated Genomes (Figure 4).

More detailed information about the maintenance of information in Gene is provided in [Gene Help](#).

Access

Web

Gene

Gene is accessed on the Web via <http://www.ncbi.nlm.nih.gov/gene/>. If the GeneID is known, the path to a specific record is generated based on the root path plus the GeneID, e.g., <http://www.ncbi.nlm.nih.gov/gene/672> for human BRCA1, for which the GeneID is 672.

Gene's website is searched via NCBI's Entrez system. The fields, filters, and properties that support effective queries are documented in Gene's [Help](#) book. Among those that are used most often are the gene symbol ([gene]) and a sequence accession.

Other NCBI databases

Gene is also accessed from other databases at NCBI. For example, a query to sequence databases, ClinVar, MedGen, or PubMed will detect what looks like a Gene symbol, and provide a display summarizing what is available in Gene (Figure 5). Records in Gene related to other database entries can be identified by following the links to Gene displayed in the panel at the right.

FTP

Information about genes is accessible from any FTP site of NCBI that includes GeneIDs as part of the content. These will not be enumerated in this document; just be aware that if a record reports a GeneID or gene_id, that is the identifier from NCBI's Gene database.

Gene

Gene's FTP site (<ftp://ftp.ncbi.nlm.nih.gov/gene/>) is divided into DATA, GeneRIF, and tools sections. The <ftp://ftp.ncbi.nlm.nih.gov/gene/README> file describes all sections, reports maintenance details, and provides detailed information about files available from Gene, as well as the annotation-specific files provided from <ftp://ftp.ncbi.nih.gov/genomes>. The DATA subdirectory provides several comprehensive files, but also includes subdirectories for the full extractions (ASN_BINARY) and tab-delimited reports (GENE_INFO) that provide subsets of data divided by major taxonomic groups.

The screenshot shows a search interface for PubMed. At the top, there is a dropdown menu set to 'PubMed' and a search bar containing the query 'pcdhga12'. Below the search bar are links for 'RSS', 'Save search', and 'Advanced'. Underneath the search bar, the text 'Display Settings:' is followed by a dropdown menu showing 'Summary, Sorted by Recently Added'. To the right of this is a 'Send to:' button with a checked checkbox. A large box below contains the search results: 'See 2 articles about PCDHGA12 gene function' and 'See also: PCDHGA12 protocadherin gamma subfamily A, 12 in the Gene database'. It also lists 'pcdhga12' found in 'Homo sapiens | Mus musculus | Rattus norvegicus | All 33 Gene records'.

Figure 5. Gene sensor in PubMed. A query that matches a gene symbol provides the user with link to more information in Gene, as well as the listing of citations in PubMed that satisfy the query (<http://www.ncbi.nlm.nih.gov/pubmed/?term=pcdhga12>)

GFF

For those interested in the location of genes and exons in a genomic context, the genomes path provides a GFF directory for many species. The README_CURRENT_RELEASE file indicates the NCBI Annotation Release being reported and the dates on which data were frozen to support the annotation. In the GFF file, GeneID is reported as a cross-reference, e.g., Dbxref=GeneID:1080. NCBI uses the [GFF3 standard](#).

E-Utilities

Gene is fully accessible programmatically using NCBI's E-Utilities. The [tools section](#) on Gene's FTP site provides some sample perl scripts to extract information from Gene based on esummary and efetch and elink.

Related Tools

In addition to the scripts available from the tools directory of Gene's FTP site (<ftp://ftp.ncbi.nlm.nih.gov/gene/tools/README>), gene2xml (ftp://ftp.ncbi.nlm.nih.gov/asn1-converters/by_program/gene2xml/) supports conversion of Gene's ASN.1 representation to XML. Gene-related programming tips are included in [Gene Help](#) and [FAQ](#).

References

1. Maglott DR, Katz KS, Sicotte H, Pruitt KD. NCBI's LocusLink and RefSeq. Nucleic Acids Res. 2000;Jan 128(1):126–8. PubMed PMID: 10592200.
2. Gene Ontology Consortium. Blake JA, Dolan M, Drabkin H, et al Gene Ontology annotations and resources. Nucleic Acids Res. 2013;Jan41(Database issue):D530–5.doiEpub 2012 Nov 17 doi: [10.1093/nar/gks1050](https://doi.org/10.1093/nar/gks1050). PubMed PMID: 23161678.

UniGene

Lukas Wagner, PhD and Richa Agarwala, PhD

Created: November 14, 2013.

Scope

UniGene is a largely automated analytical system for producing an organized view of the transcriptome. By analyzing sequences known to be expressed, and the libraries or samples from which they were derived, it is possible to organize the data into gene-specific clusters, and, in some cases, evaluate the patterns of expression by tissue, health status, and age. In this chapter, we discuss the properties of the input sequences, the process by which they are analyzed in UniGene, and some pointers on how to use the resource.

History

The task of assembling an inventory of all genes of *Homo sapiens* and other organisms began more than two decades ago with large-scale survey sequencing of transcribed sequences. The resulting Expressed Sequence Tags (ESTs) continue to be an invaluable component of efforts to characterize the transcriptome of many organisms. These efforts which rely on ESTs include genome annotation (2-4), expression systems (5), and full-length cDNA cloning projects (6). In addition, targeted gene-hunting projects have benefited from the availability of these sequences and the physical clone reagents. However, the high level of redundancy found among transcribed sequences, not to mention a variety of common experimental artifacts, made it difficult for many people to make effective use of the data. This problem was the motivation for the development of UniGene.

Now that the genomes of many species have been sequenced completely, a fundamental resource expected by many researchers is a simple list of all of an organism's genes. However, many species of medical and agricultural importance do not yet have a complete annotated genome available. Furthermore, when the genomic sequence of an organism is made public, a collection of cDNA sequences provides the best tool for identifying genes within the DNA sequence. When the source material for cDNA sequences is drawn from diverse tissues, an approximate expression profile for the organism's transcriptome can be computed. This approximate expression profile can serve to at least identify transcripts of interest to researchers interested in a particular system, and at best to characterize the function of novel transcripts. Thus, we can anticipate that the sequencing of transcribed products will remain a significant area of interest well into the future.

Data Model

The data model for UniGene is straightforward. Identify sequences of RNA molecules, the source of those sequences (species, tissue, age, health status), compute when independent

sequences are derived from the same gene based on sequence similarity, and report the results. Historically this computation was based on ESTs (Extended Sequence Tags), but now the vast majority of sequences are either full-length clones or RNAseq data.

ESTs

The basic strategy of generating ESTs involves selecting cDNA clones at random and performing a single, automated, sequencing read from one or both ends of their inserts. This class of sequence is characterized by being short (typically about 400-600 bases) and relatively inaccurate (around 2% error). In most cases, there is no initial attempt to identify or characterize the clones. Instead, they are identified using only the small bit of sequence data obtained, comparing it to the sequences of known genes and other ESTs. It is fully expected that many clones will be redundant with others already sampled and that a smaller number will represent various sorts of contaminants or cloning artifacts. There is little point in incurring the expense of high-quality sequencing until later in the process, when clones can be validated and a non-redundant set selected.

Despite their fragmentary and inaccurate nature, ESTs were found to be an invaluable resource for the discovery of new genes, particularly those involved in human disease processes. After the initial demonstration of the utility and cost effectiveness of the EST approach, many similar projects were initiated, resulting in an ever-increasing number of human ESTs. In addition, large-scale EST projects were launched for several other organisms of experimental interest. In 1992, a database called dbEST was established to serve as a collection point for ESTs, which are then distributed to the scientific community as the EST division of GenBank.

Dataflow

The number of transcribed sequences is large enough that interactive analysis of each sequence by a researcher is impossible. A major challenge is to make putative gene assignments for these sequences, recognizing that many of these genes will be anonymous, defined only by the sequences themselves. Computationally, this can be thought of as a clustering problem in which the sequences are vertices that may be coalesced into clusters by establishing connections among them.

Experience has shown that it is important to eliminate low-quality or apparently artifactual sequences before clustering because even a small level of noise can have a large corrupting effect on a result. Thus, procedures are in place to eliminate sequences of foreign origin (most commonly *Escherichia coli*) and identify regions that are derived from the cloning vector or artificial primers or linkers. At present, UniGene focuses on protein-coding genes of the nuclear genome; therefore, those identified as rRNA or mitochondrial sequence are eliminated. Through the NCBI Trace Archive, an increasing number of EST sequences now have base-level error probabilities that are used to identify the highest quality segment of each sequence. Repetitive sequences sometimes lead to false alignments and must be treated with caution. Simple repeats (low-complexity

regions) are identified using a word-overrepresentation algorithm called DUST, and transposable repetitive elements are identified by comparison with a library of known repeats for each organism. Rather than eliminating them outright, subsequences classified as repetitive are soft-masked, which is to say that they are not allowed to initiate a sequence alignment, although they may participate in one that is triggered within a unique sequence. For a sequence to be included in UniGene, the clone insert must have at least 100 base pairs that are of high quality and not repetitive.

With a given a set of sequences, a variety of different sources of information may be used as evidence that any pair of them is or is not derived from the same gene. The most obvious type of relationship would be one in which the sequences overlap and can form a near-perfect sequence alignment. One dilemma is that some level of mismatching should be tolerated because of known levels of base substitution errors in ESTs, whereas allowing too much mismatching will cause highly similar paralogous genes to cluster together. One way to improve the results is to require that alignments show an approximate dovetail relationship, which is to say that they extend about as far to the ends of the sequences as possible. Values of specific parameters governing acceptable sequence alignments are chosen by examining ratios of true to false connections in curated test sets. It is important to note that the resulting clusters may contain more than one alternative-splice form.

Multiple incomplete but non-overlapping fragments of the same gene are frequently recognized in hindsight when the gene's complete sequence is submitted. To minimize the frequency of multiple clusters being identified for a single gene, UniGene clusters are required to contain at least one sequence carrying readily identifiable evidence of having reached the 3' terminus. In other words, UniGene clusters must be anchored at the 3' end of a transcription unit. This evidence can be either a canonical polyadenylation signal or the presence of a poly(A) tail on the transcript, or the presence of at least two ESTs labeled as having been generated using the 3' sequencing primer. Because some clusters do not contain such evidence (typically, they are single ESTs), not all uncontaminated sequences in dbEST appear in UniGene clusters. Of course, alternatively spliced terminal 3' exons will appear as distinct clusters until sequence that spans the distinct splice forms is submitted.

With the availability of genome sequence, a more stringent test of 3' anchoring is possible, because internal priming can be recognized. Clusters that satisfy this more-stringent requirement can be identified by adding the term has_end to any query. Specific query possibilities such as this one are listed under the rubric Query Tips on the UniGene homepage.

Access

The UniGene website allows the user to search for particular genes of interest, or to browse UniGene entries related by expression or sequence similarity. Each UniGene Web page includes a header with a query bar and a sidebar providing links to related online resources. UniGene is also the basis for other NCBI resources:

- ProtEST, a facility for browsing protein similarities;
- Digital DifferentialDisplay (DDD), for comparison of EST-based expression profiles; and
- a library browser and display, which support exploration of cDNA libraries from tissues of interest.

Looking for sequences expressed under a particular circumstance (a body site or developmental stage, for example) is a common method by which users identify individual genes or sets of genes that are of interest. There are several interfaces into UniGene's data to help users do this. Most broadly, there is a straightforward way to browse all cDNA libraries prepared with RNA from a particular biological source. Properties of individual libraries are summarized as well; the library submitter's description of source material and protocol, and a summary of the UniGene clusters expressed by the library's sequences.

UniGene Cluster Browser

The UniGene Cluster page summarizes the sequences in the cluster and additional derived information that may be used to infer the identity and in some cases function of the gene. Figure 1 shows an example of such a view for the human SERPINF2 gene. When available, links are provided to a corresponding entry in other NCBI resources (e.g., Gene, HomoloGene, OMIM) or external databases (e.g., Mouse Genome Informatics (MGI) at the Jackson Laboratory and the Zebrafish Information Network (ZFIN) at the University of Oregon). Additional sections on the page provide protein similarities, mapping data, expression information, and lists of the clustered sequences.

Possible protein products for the gene are suggested by providing protein similarities between one representative sequence from the cluster and protein sequences from selected model organisms with an annotated genome. For each model organism, the protein with the highest degree of sequence similarity to the nucleotide sequence is listed, with its title and GenBank accession. The sequence alignment is summarized using the percent identity and length of the aligned region. Also provided is a link (a popup menu attached to the protein accession) to other protein-oriented NCBI resources including ProtEST, which summarizes translating searches of the model organism protein against all organisms in UniGene.

The next section summarizes information on the aligned or inferred map position of the gene. For human and some other annotated genomes, the map position and link to the genomic neighborhood as represented in Map Viewer. Absent these aligned map positions, radiation hybrid (RH) maps have been constructed using Sequence Tagged Site (STS) markers derived from ESTs. In these cases, the UniGene cluster can be associated with a marker in the UniSTS database, and a map position can be assigned from the RH map. More recently, map positions have been derived by alignment of the cDNA sequences to the finished or draft genomic sequences present in the NCBI MapViewer. For example, the SERPINF2 gene in Figure 1 has a link to human chromosome 17 in the

UniGene

ORGANIZED VIEW OF THE TRANSCRIPTOME

Search | UniGene Go Clear

UGID: 155668 UniGene Hs.159509 Homo sapiens (human) SERPINF2 Order cDNA clone, Links

Serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2)

Human protein-coding gene SERPINF2. Represented by 121 ESTs from 46 cDNA libraries. EST representation biased toward fetus. Corresponds to 3 reference sequences (different isoforms). [UniGene 155668 - Hs.159509]

SELECTED PROTEIN SIMILARITIES

Comparison of cluster transcripts with RefSeq proteins. The alignments can suggest function of the cluster.

Best Hits and Hits from model organisms	Species	Id(%)	Len(aa)
NP_000925.2 SERPINF2 gene product	H. sapiens	100.0	490
NP_032904.1 Serpin2 gene product	M. musculus	80.2	490
NP_001087821.1 serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 precursor	X. laevis	46.9	309
Other hits (2 of 15) [Show all]			
XP_00315334.1 PREDICTED: alpha-2-antiplasmin-like	P. troglodytes	98.8	431
XP_003912120.1 PREDICTED: alpha-2-antiplasmin isoform 2	P. anubis	96.6	494

GENE EXPRESSION

Tissues and development stages from this gene's sequences survey gene expression. Links to other NCBI expression resources.

Restricted Expression: fetus [Show more like this]

EST Profile: Approximate expression patterns inferred from EST sources. [Show more entries with profiles like this]

GEO Profiles: Experimental gene expression data (Gene Expression Omnibus).

cDNA Sources: liver; testis; mixed; kidney; eye; lung; blood; mammary gland; prostate; spleen; intestine; muscle; uncharacterized tissue; brain; embryonic tissue; pancreas

MAPPING POSITION

Genomic location specified by transcript mapping, radiation hybrid mapping, genetic mapping or cytogenetic mapping.

Location on genome: Chromosome 17; NC_000017.10 (1646121..1658563) MapViewer

Map position: 17p13

SEQUENCES

Sequences representing this gene; mRNAs, ESTs, and gene predictions supported by transcribed sequences.

mRNA sequences (9)

D00116.1	Homo sapiens mRNA for alpha-2-plasmin inhibitor, partial cds	P
AK124987.1	Homo sapiens cDNA FLJ42997 fis, clone BRTHA2011351	
D00174.1	Homo sapiens mRNA for alpha-2-plasmin inhibitor, complete cds	P
NM_000934.3	Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 1, mRNA	P
BC031592.1	Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2, mRNA (cDNA clone MGC:34213 IMAGE:5190061), complete cds	PA
AK303763.1	Homo sapiens cDNA FLJ50942 complete cds, highly similar to Alpha-2-antiplasmin precursor	P
NM_001165921.1	Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 3, mRNA	PA
NM_001165920.1	Homo sapiens serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 (SERPINF2), transcript variant 2, mRNA	PA
J02654.1	Human alpha-2-antiplasmin mRNA, 3' end	P

EST sequences (10 of 121) [Show all sequences]

AA910278.1	Clone IMAGE:1520865	kidney	3' read A
AA929082.1	Clone IMAGE:1553235	kidney	3' read A
AI081521.1	Clone IMAGE:1555724	kidney	3' read A
BX107460.1	Clone IMAGp998B203938; IMAGE:1554355	kidney	P
AI249729.1	Clone IMAGE:1864250	kidney	3' read A
AI249743.1	Clone IMAGE:1864271	kidney	3' read A
CB156019.1	Clone L17N670205n1-2-E08	liver	5' read
CB157447.1	Clone L17N670205n1-9-G04	liver	5' read P
CB157473.1	Clone L17N670205n1-10-B05	liver	5' read P
CB162683.1	Clone L17N670205n1-30-A08	liver	5' read

Key to Symbols

- P Has similarity to known Proteins (after translation)
- A Contains a poly-Adenylation signal
- S Sequence is a Suboptimal member of this cluster
- M Clone is putatively CDS-complete by MGC criteria

Links

NLM NIH UniGene Privacy Statement Disclaimer NCB Help

Figure 1. Web view of a UniGene cluster.

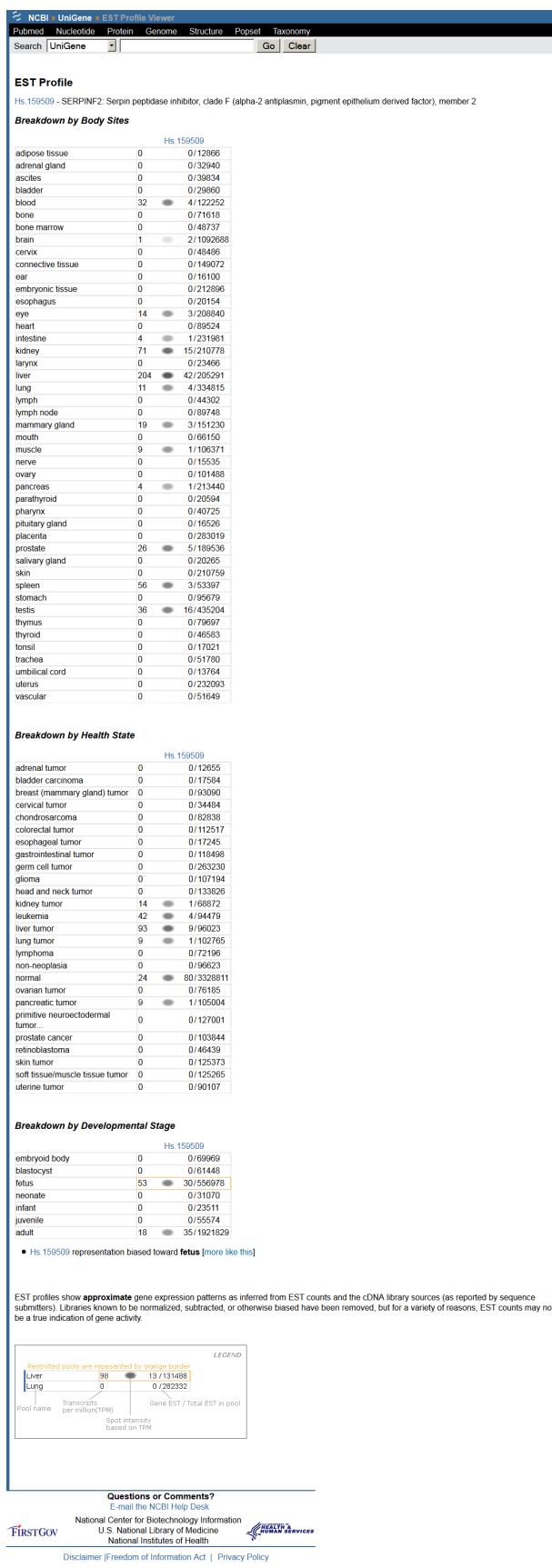


Figure 2. Expression profile view of a UniGene cluster.

Map Viewer. The map is initially shown with a few selected tracks that are likely to be of interest, but others may be added by the user.

Although ESTs are not an optimal probe of gene expression, both the total number of ESTs and the tissues from which they originated are often useful. In the Gene Expression section of the cluster browser, a link to a summary of the gene's expression in cDNA libraries is available, with an example shown in Figure 2. Expression in each body site or developmental stage available in ESTs (excepting those from normalized or subtracted libraries, also excepting those which come from libraries of mixed source material) is reported, expressed as counts of ESTs transcribed per million sequenced. Expression data is also available by FTP. UniGene clusters with similar expression profile are precomputed, and available under the link labelled "show more like this." This is most likely to be informative when expression differs markedly from uniform expression. Clusters that are predominantly expressed in a single body site or developmental stage are searchable in Entrez, by using the field "Restricted Expression." More specifically, these are clusters where 2/3 or more of the detected gene expression expressed in normalized units of transcripts per million is from a single source. Links to NCBI's GEO computed from the GenBank accessions in the UniGene cluster are also present in this section of the cluster view.

The component sequences of the cluster are listed, with a brief description of each one and a link to its UniGene Sequence page. The Sequence page provides more detailed information about the individual sequence, and in the case of ESTs, includes a link to its corresponding UniGene Library page. On the cluster page, the EST clones that are considered by the Mammalian Gene Collection (MGC) project to be putatively full length are listed at the top, whereas others follow in order of their reported insert length. At the bottom of the UniGene Cluster page is a button for users to download the sequences of the cluster in FASTA format.

An FTP representation of UniGene is available as well. Sequence sets as FASTA (both aggregate and best representative per cluster), a summary of the mapping of sequences to UniGene clusters with library of origin for ESTs, and an expression summary. A common use of UniGene is to use a single representative sequence from each cluster for primer design or as a BLAST database. In this case, researchers are advised to retain both the sequence accession number as well as the cluster identifier for later reference. This is because the cluster identifier is not guaranteed to be indefinitely stable. While most UniGene builds differ only through incremental changes to existing clusters or the addition of newly represented transcripts, new sequences or new genome mapping can provide information that leads to substantial reorganization of previously identified clusters.

Related Tools

The screenshot shows a UniGene cluster browser page for the gene SERPINF2. The main content area displays the gene's name and a brief description. Below this, a 'SELECTED PROTEIN SIMILARITIES' section contains a table of best hits from model organisms. A mouse cursor is hovering over the accession number NP_001087821.1, which triggers a context-sensitive dropdown menu. This menu includes links for 'Conserved domains (CDD)', 'Gene summary', 'Protein sequence', 'Protein/EST matches (ProtEST)', and 'Protein/protein matches (BLink)'. The 'Protein/EST matches (ProtEST)' link is highlighted with a red box.

Best Hits and Hits from model organisms		Species	Id(%)	Len(aa)
NP_000925.2	SERPINF2 gene product	<i>H. sapiens</i>	100.0	490
NP_032904.1	Serpinf2 gene product	<i>M. musculus</i>	80.2	490
NP_001087821.1	serpin peptidase inhibitor, clade F (alpha-2 derived factor),	<i>X. laevis</i>	46.9	309

		Species	Id(%)	Len(aa)
XP_003315334.1	like	<i>P. troglodytes</i>	98.8	431
XP_003912120.1	isoform 2	<i>P. anubis</i>	96.6	494

Figure 3. Popup menu in UniGene cluster browser providing links to information about proteins similar to the transcript sequences in the cluster.

Protein Similarity Browser

The ProtEST section of UniGene allows the user to explore precomputed alignments for a selected protein to the cDNA sequences found in any cluster. Especially for cases where looking at alignments to the same protein of transcripts from multiple organisms, this interface provides a single concise overview. In the cluster viewer's protein similarity section, this overview is under the "Protein/EST matches" link in the popup menu that appears on mouseover of a protein accession; this popup is shown in Figure 3. BLASTX has been used to compare each sequence in UniGene to selected protein sequences drawn from model organisms with a fully annotated genome. By default, only alignments to the organisms in the same broad taxonomic group as the original organism are shown (primates, rodents, etc), though alignments to a broader set of organisms can be selected from the protEST pulldown menu. These alignments include alignments both to RefSeqs, which are based on a sequenced and annotated mRNA, as well as RefSeqs, which are gene predictions.

The sequence alignments in ProtEST are summarized in tabular form (Figure 4). The first column is a schematic representation of the nucleotideprotein alignment. The width of the column represents the entire length of the protein, whereas the unaligned nucleotide

NCBI UniGene ORGANIZED VIEW OF THE TRANSCRIPTOME

PubMed Nucleotide Protein Genome Structure Popset Taxonomy

Search All Databases Go Clear

Protein / EST Matches (ProtEST)

NP_001087821.1 - serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 2 precursor [Xenopus laevis] (705 aa)

Use this form to change data being displayed. Help
 Taxonomic group: Mammals Species: all
 Maximum sequences per entry: 3 Apply

UniGene sequences from Mammals that match this protein

UniGene entry	EST/mRNA Sequence		Protein Sequence		Alignment Quality				
	accession	strand	coordinates	region	coordinates	score (bits)	ident. (%)	len. (aa)	
Chi.16680	<i>Capra hircus</i>								
	JO421145	+	340-1455		317-687	338	45.2	371	align
	JO421144	+	343-1458		317-687	338	45.2	371	align
	JO590481	+	3-623		460-665	207	50.7	206	align
Rn.15774	<i>Rattus norvegicus</i>								
	CO572664	+	28-780		400-650	246	48.2	251	align
	CO572497	+	15-722		446-680	238	50.0	235	align
	DY312856	+	5-694		421-650	230	48.7	230	align
Bt.9352	<i>Bos taurus</i>								
	CR553170	+	320-997		414-639	226	48.7	226	align
	CR552417	+	49-681		429-639	208	48.3	211	align
	CR551804	+	26-670		425-639	208	47.4	215	align
Hs.159509	<i>Homo sapiens</i>								
	D00116	+	1-720		465-703	225	47.1	239	align
	BX433337	-	25-756		461-703	220	45.9	243	align
	BM553329	+	20-547		395-570	173	46.0	176	align
Ssc.81738	<i>Sus scrofa</i>								
	BP445819	+	3-608		486-687	197	48.0	202	align
	AK232353	+	2-607		486-687	197	48.0	202	align
	BW963770	+	100-576		529-687	154	48.4	159	align
Oar.22364	<i>Ovis aries</i>								
	DY515275	+	1-522		471-644	175	50.0	174	align
Mm.279733	<i>Mus musculus</i>								
	BI143862	+	101-613		407-577	168	46.2	171	align
	CR758770	-	2-463		407-560	162	48.0	154	align
	CB953301	+	4-453		535-683	146	49.3	149	align
Mfa.15423	<i>Macaca fascicularis</i>								
	CO775724	+	6-560		520-703	163	46.0	184	align
	CO775781	+	6-560		520-703	163	46.0	184	align

NLM NIH UniGene | Privacy Statement | Disclaimer | NCBI Help

Figure 4. Protein-transcript alignment summary.

sequence is represented as a thin gray line and the aligned region is represented as a thick magenta bar. The alignment representation is a hyperlink to the full alignment regenerated on-the-fly using BLAST. Other information in the table includes the frame and strand of the alignment the UniGene cluster ID, the GenBank accession, and a summary of the aligned region and percent identity.

Digital Differential Display (DDD)

DDD is a tool for comparing EST-based expression profiles among the various libraries, or pools of libraries, represented in UniGene. These comparisons allow the identification of those genes that differ among libraries of different tissues, making it possible to determine which genes may be contributing to a cell's unique characteristics, e.g., those that make a muscle cell different from a skin or liver cell. Along similar lines, DDD can be used to try to identify genes for which the expression levels differ between normal, premalignant, and cancerous tissues or different stages of embryonic development.

As is UniGene, the DDD resource is organism specific and is available from the UniGene website for that organism. For those libraries that have sequences in UniGene, DDD lists the title and tissue source and provides a link to the UniGene Library page, which gives additional information about the library. From the libraries listed, the user can select two for comparison. DDD then displays those genes for which the frequency of the transcript is significantly different between the two libraries. The output includes, for each gene, the frequency of its transcript in each library and the title of the gene's corresponding UniGene cluster. Results are sorted by significance, with the genes having the largest differences in frequencies displayed at the top. Libraries can be added sequentially to the analysis, and DDD will perform an analysis on each possible library gene pair combination. Similarly, groups of libraries can be pooled together and compared with other pools or single libraries. An example comparing two pools of libraries with similar sequence counts from human muscle and human brain is shown in Figure 5.

DDD uses the Fisher Exact test to restrict the output to statistically significant differences ($P < 0.05$). The analysis is also restricted to deeply sequenced libraries; only those with over 1000 sequences in UniGene are included in DDD. These requirements place limitations on the capabilities of the analysis. Unless there are a large number of sequences in each pool, the frequencies of genes are generally not found to be statistically significant. Furthermore, the wide variety of tissue types, cell types, histology, and methods of generating the libraries can make it difficult to attribute significant differences to any one aspect of the libraries. These issues underscore the need for more libraries to be made public and the need for the comparisons to be made using proper controls.

cDNA Library Browser and UniGene cDNA Library Display

Researchers frequently wish to identify particular cDNA libraries that interest them. In addition to the gene-oriented resources described above, UniGene offers an overview of all libraries from an organism of interest in the library browser. Libraries are grouped by their source material (body site or developmental stage where these are described by the

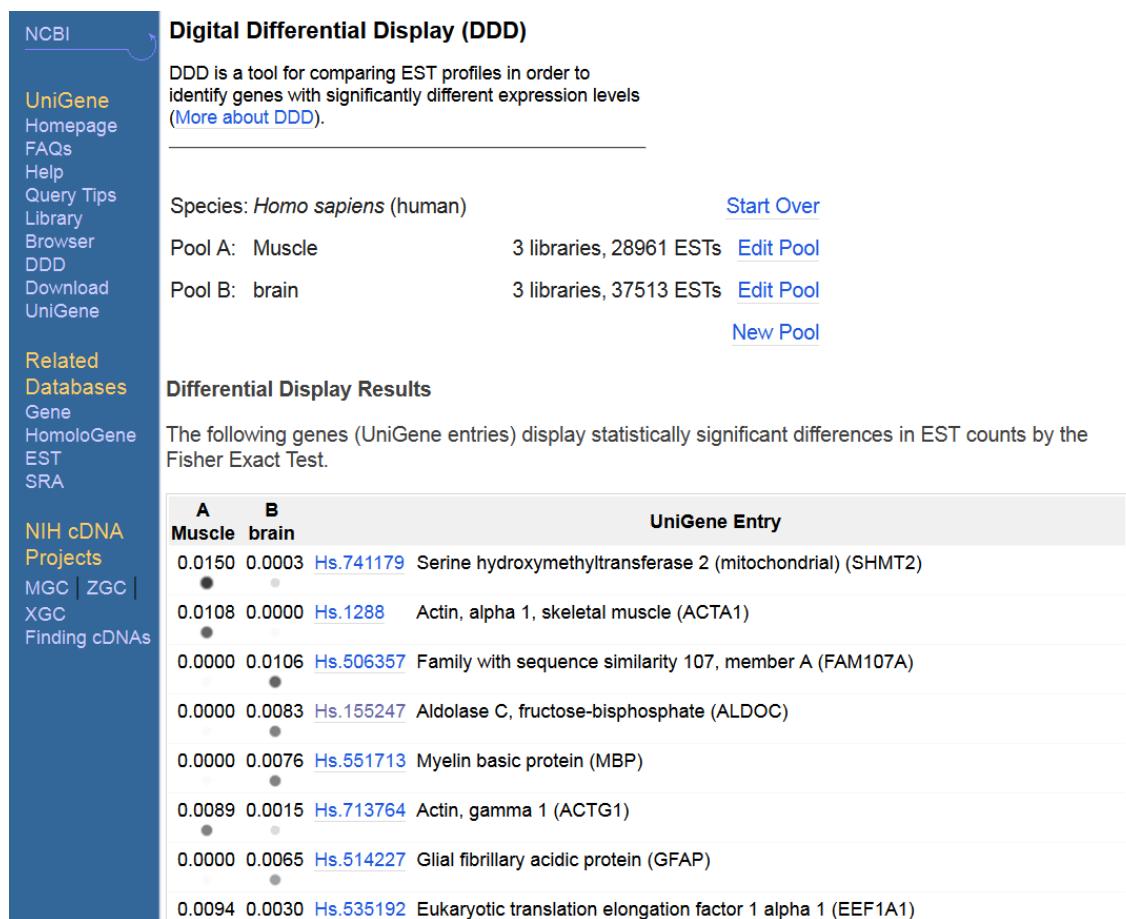


Figure 5. Differential expression assessment comparing libraries from muscle and from brain.

library's submitter), with the Web interface for browsing shown in Figure 6. For individual libraries, the library summary aggregates information provided by the submitter with the UniGene clusters that contain sequences from a library, with the Web interface for an individual library summary shown in Figure 7. Researchers may download all sequences from the library in FASTA format whether they are in any UniGene cluster or not from this page.

NCBI » UniGene » Homo sapiens » Library Browser

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Go Clear

Libraries for with minimum sequences Show

[Collapse All](#) | [Expand All](#)

Body Sites

▼ adipose tissue 18 libraries

Lib. ID	Library Name	Sequences
Lib.10983	Human Fat Cell 5'-Stretch Plus cDNA Library	9638
Lib.886	NCI_CGAP_Lip2	1740
Lib.16445	Sugano cDNA library, adipose tissue	1665
Lib.816	Adipose tissue, white II	1195

Not Shown: 14 libraries having fewer than 1000 sequences

► adrenal gland 30 libraries

► amniotic fluid 63 libraries

► ascites 17 libraries

▼ bladder 65 libraries

Lib. ID	Library Name	Sequences
Lib.5383	NIH_MGC_53	10515
Lib.8658	NIH_MGC_93	9864
Lib.18307	BLADE2	8611

Not Shown: 62 libraries having fewer than 1000 sequences

▼ blood 385 libraries

Lib. ID	Library Name	Sequences
Lib.13022	Homo sapiens T CELLS (JURKAT CELL LINE) COT 10-NORMALIZED	10909
Lib.9724	NIH_MGC_118	10515

Figure 6. UniGene library browser.

UniGene
ORGANIZED VIEW OF THE TRANSCRIPTOME

Search [UniGene] for [] Go Clear

NCBI > Databases > UniGene > Homo sapiens > cDNA libraries

Library:18307 (dbEST ID)

BLADE2

Library Description

Organism: *Homo sapiens*
Tissue: bladder
Vector: pME18SFL3
Source: Takao Isogai, Helix Research Institute

Gene Content Analysis

8,141 ESTs from this library were grouped into 3,187 UniGene entries (putative genes) [UniGene build #236, 09-Mar-2013]. EST counts for each entry may be used to calculate an approximate expression level in transcripts per million (TPM).

ESTs	TPM	UniGene Entry
291	35745	● Hs.586423 Eukaryotic translation elongation factor 1 alpha 1.
194	23830	● Hs.53985 Glycoprotein 2 (zymogen granule membrane).
86	10564	● Hs.644105 CD24 molecule.
78	9581	● Hs.520640 Actin, beta.
78	9581	● Hs.508113 Olfactomedin 4.
70	8598	● Hs.520348 Ubiquitin C.
61	7493	● Hs.446852 Eukaryotic translation initiation factor 3, subunit L.
53	6510	● Hs.298280 ATP synthase, H ⁺ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle.
48	5896	● Hs.567312 Regenerating islet-derived 3 alpha.
48	5896	● Hs.418167 Albumin.

EST Sequences
8,611 EST sequences from this library have been submitted to the public sequence database as of 09-Mar-2013.
Source: NCBI GenBank

Download Sequences
Click the button to save sequences to a file (FASTA format)

Keywords
bladder, normal, unknown developmental stage, uncharacterized preparation
Source: NCBI Clone Registry

References
Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Kimura K, et al., Genome Res 16, 55-65 (2006).

Figure 7. UniGene library summary.

References

- Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde RF, Moreno RF, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. 1991;252(5013):1651–1656. PubMed PMID: 2047873.
- Lu F, Jiang H, Ding J, Mu J, Valenzuela JG, Ribeiro JM, Su XZ. cDNA sequences reveal considerable gene prediction inaccuracy in the *Plasmodium falciparum* genome. *BMC Genomics*. 2007;8:255. PubMed PMID: 17662120.
- Shangguan L, Han J, Kayesh E, Sun X, Zhang C, Pervaiz T, Wen X, Fang J. Evaluation of genome sequencing quality in selected plant species using expressed sequence tags. *PLoS One*. 2013 Jul 29;8(7) PubMed PMID: 23922843.
- Head and Neck Annotation Consortium. Large-scale transcriptome analyses reveal new genetic marker candidates of head, neck, and thyroid cancer. *Cancer Res*. 2005 Mar 1;65(5):1693–9. PubMed PMID: 15753364.

5. Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol.* 2011 Oct;9(10) PubMed PMID: 22028629.
6. MGC Project Team. The completion of the Mammalian Gene Collection (MGC). *Genome Res.* 2009 Dec;19(12):2324–33. PubMed PMID: 19767417.

Nucleotide

GenBank

Ilene Mizrachi^{✉1}

Created: November 12, 2013.

Scope

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at the National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) Data Library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ). GenBank and its collaborators receive sequences produced in laboratories throughout the world from hundreds of thousands of distinct organisms. GenBank continues to grow at an exponential rate, doubling every 18 months. Release 197, produced in August 2013, contains over 154 billion nucleotide bases in more than 167 million sequences. GenBank is built by direct submissions from individual laboratories and from large-scale sequencing centers.

Submissions to GenBank include a number of data types: single gene or mRNA sequences, phylogenetic studies, ecological surveys, complete genomes, genome assemblies (WGS), transcriptome assemblies (TSA), and third-party annotation (TPA). In the past, transcript surveys (EST), genome surveys (GSS), and high-throughput genome sequences (HTGS) constituted a significant fraction of submissions, but with the emergence of next-generation sequencing technologies, we have seen a steady decrease in these data types. Submissions to GenBank are prepared using one of a number of submission tools and transmitted to NCBI. Upon receipt of a submission, the GenBank staff reviews the records, assigns an accession number and performs quality-assurance checks prior to release of the data to the public archive. Sequence records once released are retrievable by [Entrez](#), searchable by BLAST, and downloadable by FTP.

¹ NCBI; Email: mizrachi@ncbi.nlm.nih.gov.

[✉] Corresponding author.

Protein

NCBI Protein Resources

Eric Sayers, PhD^{✉1}

Created: November 12, 2013; Updated: November 21, 2013.

Introduction

Proteins are the machines of life. They perform almost all of the processes necessary to sustain life, and also form a variety of structural and connective materials that constitute the bulk of our bodies and those of all other organisms. While we can think of a single protein as a discreet polymer of amino acids, the functional form of many proteins in the cell is actually a complex of several individual polymer chains, all working in concert to do a particular task. At NCBI, we therefore provide several resources that represent these various aspects of proteins, ranging from the sequences of individual chains to functional classifications of large protein families.

Protein

The [Protein](#) database is the most fundamental NCBI resource for proteins. It contains text records for individual protein sequences derived from a variety of sources, including GenBank, the NCBI Reference Sequence (RefSeq) project and several external databases including UniProtKB/SWISS-Prot and the Protein Data Bank (PDB). It is important to understand that the sequences contained in almost all Protein records (with the exception of PDB) are conceptual translations of an RNA coding sequence, meaning that no one determined the protein sequence experimentally, but rather inferred the sequence from the corresponding RNA. Protein records are available in several formats (including FASTA and XML) and are linked to many other NCBI resources, allowing users to find relevant data such as literature, DNA/RNA sequences, genes, biological pathways and expression and variation data. We also provide pre-computed sets of identical and similar proteins for each sequence as determined by the BLAST algorithm. The BLAST Link (Blink) tool provides a graphical view of these pre-computed sets and provides a variety of filtering tools and links to multiple alignment views.

Structure

The [Structure](#) database contains 3D coordinate sets for experimentally-determined structures in PDB. At NCBI, we import these data from PDB and format the data for viewing in Cn3D, the NCBI structure viewer. We also calculate structural similarity between all of these records using the VAST algorithm and allow users to view

¹ NCBI; Email: sayers@ncbi.nlm.nih.gov.

[✉] Corresponding author.

superpositions of highly similar structures. We also link these structure records to the corresponding sequence records in the Protein database, to literature articles where the structure was reported, and to information about any ligands present in the structure.

Conserved Domains (CDD)

The [Conserved Domain](#) database (CDD) is a collection of sequence profiles that represent highly conserved domains within protein sequences. Very often these domains have a particular function that is shared between those sequences that contain it. Typically one identifies the presence of a conserved domain in a sequence using the CD-Search tool, and these results provide access to sequence alignments, distance trees, selected literature, and structural views that highlight important elements within the domain. While we curate our own set of domain records, CDD also contains records from external resources such as Pfam and SMART. NCBI provides links to precomputed CD-Search results for all Protein records and also displays such results on each record in the Structure database. CD-Search results are also provided in several other NCBI displays, including protein BLAST results and the graphical sequence viewer.

Protein Clusters

The [Protein Clusters](#) database contains sets of proteins annotated on RefSeq genomes from prokaryotes, viruses, fungi, plants, and organelles (mitochondria and chloroplasts). Each protein cluster record consists of a set of protein sequences clustered according to the maximum alignments calculated by BLAST between the individual sequences. We then hypothesize that the proteins within each set are homologous, and on this basis use these clusters to support functional annotation of new RefSeq genomes.

Protein Clusters

Tatiana Tatusova, PhD,¹ Leonid Zaslavsky, PhD,¹ Boris Fedorov, PhD,¹ Diana Haddad, PhD,¹ Anjana Vatsan, PhD,¹ Danso Ako-adjei, PhD,¹ Olga Blinkova, PhD,¹ and Hassan Ghazal, PhD^{1,2}

Created: September 14, 2014.

Scope

The Protein Clusters dataset consists of organized groups (clusters) of proteins encoded by complete and draft genomes from the NCBI Reference Sequence (RefSeq) collection of microorganisms: prokaryotes, viruses, fungi, protozoans; it also includes protein clusters from RefSeq genomes of plants, chloroplasts, and mitochondria. Clusters for each group are created and curated separately and given a different accession prefix. The primary goal of Protein Clusters is to provide the support to functional annotation of RefSeq genomes. Functional annotation of novel proteins is based on the assumption that proteins with high levels of sequence similarity are likely to share the same function. This oversimplified model of a linear evolution where similar proteins evolve from a single ancestor is further complicated by the events of gene duplication. Clusters of related (homologous) proteins include both orthologs and paralogs. Orthologs are genes in different organisms (species) that evolved from a common ancestral gene by speciation; paralogs are genes related by duplication within a genome. The definition was first introduced by Fitch in 1970 (7). The analysis of protein families from various organisms has shown that this definition does not embrace all the complexity of relationships of genes from different organisms. For more details see Koonin *et al.* 2005 (16). The NCBI Protein Clusters dataset contains automatically generated clusters that do not distinguish orthologs and paralogs. During manual evaluation some clusters containing paralogs can be split by curators, especially if the paralogs are known to have different functions.

History

The first complete bacterial genome of *Haemophilus influenzae* Rd KW20 sequenced and released in 1995 opened a new era in genome analysis (8). In the following year four more genomes were completed producing an unprecedented variety of protein sequences from all three major kingdoms (Archaea, Eubacteria, and Eukaryota). Comparative analysis of homologous genes has been used in evolutionary studies and functional classification since the first sequence became available, but for the first time the whole proteome of several organisms became available for comparison. New genome-scale methods were needed to provide an understanding of the true orthologous relationships of protein sequences. The protein database of Clusters of Orthologous Groups ([COGs](#)), a pioneering

¹ NCBI. ² University Mohammad 1, Morocco.

work of NCBI scientists, was the first attempt at creating a phylogenetic classification of the complete complement of proteins encoded by complete genomes (23). The COG approach is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genome are most likely to form an orthologous set. The COG project has proved to be an excellent approach for understanding microbial evolution and the diversity of protein functions encoded by their genomes. However, the major difficulty of any genomic data resource in the modern era of rapid genomic sequencing is keeping the genomic data and the annotations up-to-date.

The [RefSeq](#) project, which contains non-redundant sets of curated transcript, gene, and protein information in eukaryotes, and gene and protein information in prokaryotes, has been a very successful way to maintain and update annotated data. Given the increasing number of prokaryotic genomes being deposited, it became apparent that annotating protein families as a group was a convenient and efficient way to functionally annotate these data. The Protein Clusters database was constructed with two goals in mind: first, to routinely update RefSeq genomes with curated gene and protein information from such clusters; and second, to provide a central aggregation source for information collected from a wide variety of sources that would be useful for scientists studying protein-level or genomic-level molecular functions. In addition, curators routinely parse the scientific literature for reports of experimentally verified functions as the basis for existing or potential connections to genes/proteins, and such connections are added as annotations on each cluster. The first release of NCBI Protein Clusters in 2007 contained ~1 million proteins encoded by complete chromosomes and plasmids from three major groups: prokaryotes, bacteriophages, and the organelles (15). Since then the scope has been expanded to other taxonomic groups and proteins from draft genomes.

As of April 2013 the dataset represents more than 20 million proteins.

Data Model

Clustering is a well-known method in statistics and computer science. For a given set of entities, clusters are defined as subsets that are homogeneous and well separated. The cluster analysis should start from a definition of homogeneity and separation. Most clustering methods rely upon similarities (or distance) between entities. Protein clusters are aimed to be groups of homologous proteins. The similarity between two protein sequences is measured by maximum alignment between the sequences calculated by BLAST. There are multiple ways of defining various types of clusters that are based on criteria used to express separation or homogeneity of a cluster and separation from other clusters. NCBI Protein Clusters uses two methods for clustering, both resulting in building cliques, one based on partitioning and the other based on hierarchical aggregation.

Once clustered, each protein cluster is assigned a cluster ID and accession (letter prefix followed by digits) that is stable from version to version as long as the majority of its

proteins don't change. A protein cluster also includes certain attributes aggregated from proteins: "Gene names" (locus), "COG functional categories," "EC numbers," and "Conserved Domains." An attribute "Conserved in" defines the common taxonomical name of genomes included in the cluster. The Protein Clusters database also includes a set of "Related Clusters". Besides these attributes, each protein record in the database has "Organism name," "Protein name," "Protein accession," "Locus tag," "Length," and UniProtKB / SwissProt accession. These attributes are easily searchable within a cluster and also through the whole database.

Statistics

Proteins:	31
Conserved in:	Bacteria
Total genera:	7
Total organisms:	28
Putative Paralogs:	3
Locuses:	sacC, sacC1
COG functional category:	Carbohydrate transport and metabolism

Related Clusters

cluster	Name	Distance	Protein	Median length(aa)	Genomes
PCLA_776568	glycosyl hydrolase family 32	0.529	51	516	47
PCLA_5026923	fructan hydrolase	0.544	2	507	2
PCLA_5501389	levanase	0.569	2	446	2
PCLA_5502029	Levanase	0.578	2	509	2
PCLA_376821	levanase	0.587	21	677	21
PCLA_728286	glycosyl hydrolase family 32	0.601	19	509	13
PCLA_2508750	glycosyl hydrolase family 32	0.608	2	485	1
PCLA_2993089	levanase	0.61	3	635	3
PCLA_3254773	levanase	0.614	6	492	6
PCLA_5838420	beta-fructosidase, levanase/invertase	0.615	2	497	2

Genome Groups (clades)

Clade ID	Name	Proteins in Cluster	Total Annotated Genomes	Proteins per Genome (median)
19970	Paenibacillus	5	9	5268
19976	Paenibacillus mucilaginosus	3	3	7330
19973	Paenibacillus	2	3	6237
19975	Paenibacillus elgii	2	1	7776
21852	Paenibacillus sp. A9	2	1	4856
21853	Paenibacillus sp. PAMC 26794	1	1	5873

Protein Table

Clade ID	Organism	Protein name	Accession	Locus_tag	Length	Identical group	BLINK
22152	Actinopolyspora mortivalis DSM 44261	hypothetical protein	WP_019853821	ACTMO_06285	514	WP_019853821	◆
21263	Amphibacillus jilinensis Y1	hypothetical protein	WP_017470882	B494_02995	484	WP_017470882	◆
21267	Bacillus alcalophilus ATCC 27647	glycosyl hydrolase family protein	WP_003322074	BalcAV_07657	477	WP_003322074	◆
21936	Bacillus bataviensis LMG 21853	SacC2	WP_007084638	BABA_08076	492	WP_007084638	◆
21271	Bacillus endophyticus 2102	hypothetical protein	WP_019393615	A360_15930	531	WP_019393615	◆

Example of a bacterial cluster PCLA_5029913 glycoside hydrolase

Clustering Methods

Partitioning in Cliques

Proteins are compared by sequence similarity using BLAST all against all (E-value cutoff 10E-05; effective length of the search space is set to 5×10^8). Each BLAST score is then modified by protein length \times alignment length of the BLAST hit and the modified scores are sorted. Clusters (also known as cliques) consist of protein sets such that every member of the cluster hits every other protein member (reciprocal best hits by modified score). Cluster membership is such that for any given protein in the cluster (protein A), all the other members of the cluster will have a greater modified score to protein A than any protein outside of that cluster. During clustering, there are no cutoffs used nor strict requirements for clusters of orthologous groups, nor any check on phylogenetic distance. The initial set of uncurated clusters created in 2005 was used as a starting point for curation and has been updated periodically since then. During updates, new proteins are added to curated clusters. In the uncurated cluster set, proteins are allowed to repartition into different cluster sets, although this happens rarely and usually only in the case of smaller clusters.

Hierarchical Aggregation in Cliques

A new approach implemented for prokaryotic genomes is based on hierarchical clustering. While a hierarchical structure is conventionally represented by a dendrogram and clusters are selected as a sub-tree corresponding to a certain threshold (14, 17, 18), the hierarchical structure goes beyond simple clustering (1, 3). First, all the proteins are organized in global clusters, then links between clusters are calculated reflecting the similarity between the clusters based on several criteria.

Protein Clustering Procedure

The similarity of proteins is determined from the aggregated BLAST hits obtained by *blastp* with e-value 10^{-3} . Two proteins are considered connected if there is an aggregated BLAST hit between them satisfying criteria on hit length and score. More specifically, we require the aggregated hit lengths on each protein, $l_{ij}^{(1)}$ and $l_{ij}^{(2)}$, satisfy the inequalities $l_{ij}^{(1)} \geq \varepsilon \cdot l_i$ and $l_{ij}^{(1)} \geq \varepsilon \cdot l_j$, where l_i and l_j are lengths of proteins, and $0.5 < \varepsilon < 1$, and the aggregated BLAST score S_{ij} satisfy the inequality $S_{ij} \geq \gamma \cdot \max(S_{ii}, S_{jj})$, where S_{ii} and S_{jj} are self-scores.

The modified BLAST distance is defined as

$$d_{ij}^\alpha = 1 - \frac{S_{ij}}{\max(\chi_{ij}^{(1)} \cdot S_{ii}, \chi_{ij}^{(2)} \cdot S_{jj}, S_{ij})},$$

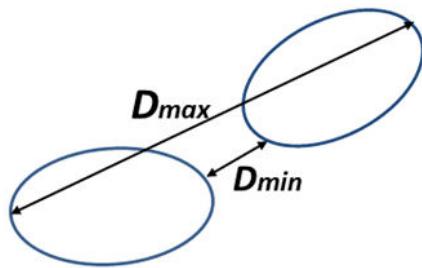


Figure 1. Minimal and maximum distance between clusters

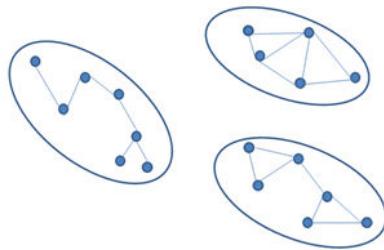


Figure 2. Disjoint sets.

where the score modifications are $\chi_{ij}^{(1)} = \max\left(\frac{l_{ij}^{(1)}}{l_i}, 1-\alpha\right)$ and $\chi_{ij}^{(2)} = \max\left(\frac{l_{ij}^{(2)}}{l_i}, 1-\alpha\right)$, and $0 < \alpha \ll 1$. Using $\alpha > 0$ allows some flexibility at the end of the sequences (the distance is reduced to $d_{ij}^0 = 1 - \frac{s_{ij}}{\max(s_{ii}, s_{jj})}$ when $\alpha = 0$). Clusters are aggregated in a hierarchical manner using the complete linkage distance, with an additional requirement that the minimum distance between clusters $d^{\min}(\Lambda, \Omega)$ should not exceed threshold δ , where $0 < \delta < 1 - \gamma$, for clusters Λ and Ω to be merged. Note that we calculate and use both $d^{\min}(\Lambda, \Omega)$ and $d^{\max}(\Lambda, \Omega)$ in our clustering procedure (see Figure 1). Because of the sparse nature of connections and applied thresholds, we build a family of trees that we consider clusters. Currently, we use the values $\varepsilon = 0.7$, $\gamma = 0.2$, $\alpha = 0.1$, and $\delta = 0.5$.

Establishing Links between Related Clusters

Each protein within a cluster should be similar to *all* other proteins in the same cluster, satisfying coverage and similarity criteria. Still, a pair of proteins in different clusters could be similar. Such clusters are designated as *related clusters* (1, 3, 12, 24). Links between related clusters are stored in *link indexes*, which are used to show neighborhoods of clusters in Entrez search.

Organization of computations. First, we eliminate redundancy and near-redundancy in the protein dataset (2, 12). Representative proteins from groups of redundant and nearly-redundant proteins are selected by the program USEARCH (5).

In order to perform clustering in parallel, the dataset is partitioned in disjoint sets (Figure 2) using a parallel implementation based on a disjoint-set forest with union-by-rank heuristics (4, 22), and then clustering is performed concurrently in partitions. When looking for disjoint sets, we only consider connections with $d_{ij}^{\alpha} \leq \delta$.

After the clustering is performed, link indexes are also calculated in parallel from the aggregated BLAST hit and protein assignment to clusters.

Dataflow

Input data are proteins from complete and draft (WGS) genomes that pass certain quality filters.

Proteins marked as incomplete in metadata (“incomplete,” “no start,” “no end,” “fragment,” etc.) are removed and only proteins that are presumed complete are analyzed. Bacterial genome clustering has a different dataflow compared to other genomes as indicated in Figure 3.

Manual Curation

One of the most important aspects of the curation process of Protein Clusters is the assignment of function that is obtained from the literature. Curated functional annotation can be propagated to all proteins within the cluster. That process improves the functional annotation of RefSeq genomes and unifies and standardizes the naming rules across various organisms and different annotation pipelines. In addition to providing functional annotation that is required for each cluster, other data are also added, such as the gene name, a detailed description about the protein, the E.C. number, and relevant publications.

Cluster Display

A protein cluster is represented by a list of protein identifiers (accessions) and the genomes that code for the proteins. Each cluster has a stable unique identifier and a functional cluster name. The cluster name is automatically calculated and followed by manual review. Each cluster provides statistics to indicate the number of proteins within that cluster and other important features including the Protein Table.

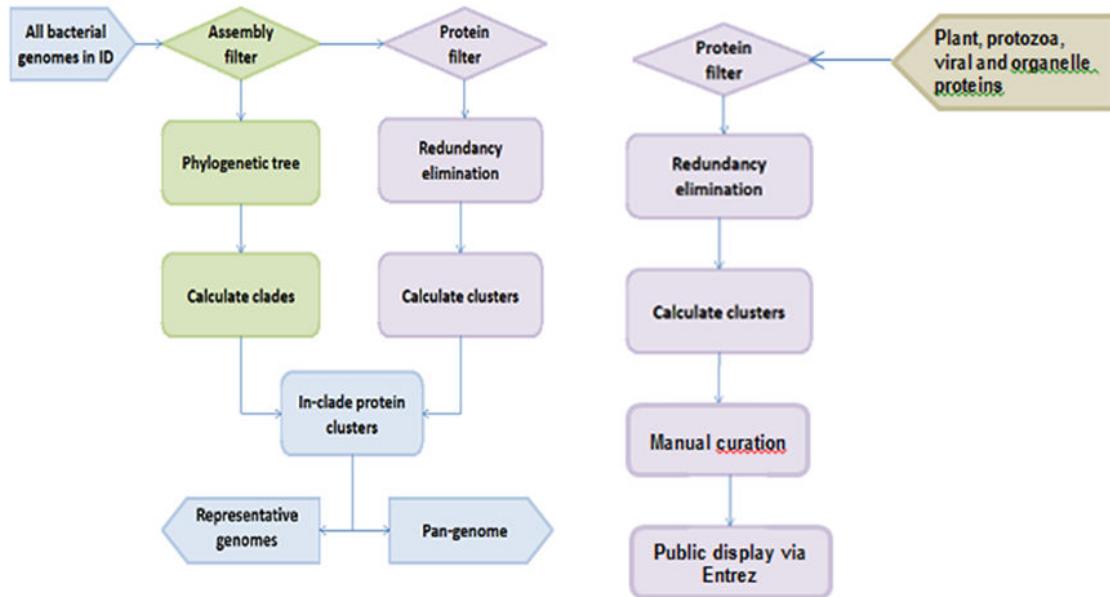


Figure 3. Dataflow for prokaryotic and eukaryotic genomes

Cluster Examples

Viruses

The ease and efficiency of nucleic acid sequencing has led to an abundance of sequence data. Because of the relatively small genome size of viruses, the influx of sequence data has been particularly large. Likewise, the ever increasing advancements and publications in virology research make it difficult for researchers to keep up with new discoveries in protein structure and function. Rapid viral evolution, combined with the relatively large number of strains and closely related species in most viral families, makes the Protein Clusters resource an ideal channel through which viral RefSeq genomes can be curated.

The *Poxviridae* is an example of a virus family with a large set of proteins having varying degrees of similarity in function, homology, and structure (13). The poxvirus RNA helicase NPH-II belongs to a family of ubiquitous ATP-dependent helicases that are required for RNA metabolism in bacteria, eukaryotes, and many viruses (6). The NPH-II family of helicases found in hepatitis C and various poxviruses have similar sequence, structure, and mechanisms of action that are essential for viral replication. The protein cluster PHA2653 includes 27 NPH-II helicase proteins from various members of the *Poxviridae*. While they share a high level of homology, evolutionary pressures have resulted in changes to both sequence and activity. Of particular interest is the fact that the poxvirus NPH-II belongs to a superfamily, SF2, of which several eukaryotic helicases that play a major role in cellular responses to viral infection also belong (19). Furthermore, the helicase core domain is a component of the dicer complex which mediates RNAi in higher

eukaryotes (9). Therefore, it stands to reason that study of the NPH-II helicases of the *Poxviridae* can serve as a model for understanding several distinct biological processes.

Frequently, several alternative names are used for viral proteins; this variation can lead to confusion for researchers and slow scientific progress. To standardize protein names, NCBI staff (viral genome curators) work closely with viral protein experts from UniProt. Such collaborations have resulted in functional naming for viral protein clusters from the family *Adenoviridae*. One of their representatives is cluster PHA3614. It combines related and highly conserved proteins from the genus *Mastadenovirus*, which presumably play an important role in host modulation (11). The old, commonly used name of proteins from the PHA3614 cluster was the E3 12.5 kDa protein. Because the size of the protein could vary in different viruses without its biological role changing, the molecular mass, as a component of protein name, was not informative and could be misleading. Therefore we proposed a new, functional name for this cluster: “putative host modulation protein E3.” All existing synonyms were included in the cluster’s functional description. Since the existence of the putative host modulation protein E3 was experimentally supported only for human adenovirus 2 and human adenovirus 5, this information was also included in the description of the cluster. These changes will be visible with the next cluster update.

Protozoa

The following is an example of the significance of publication links in a cluster of proteins as a tool to identify orthologs and paralogs.

PTZ00021 falcipain-2 ID: 2458473

Cysteine proteases identified and characterized in *Plasmodium falciparum*. Hydrolises the erythrocyte hemoglobin into its aminoacid constituents, which are used by the parasite for protein synthesis.

Statistics

Proteins:	9
Conserved in:	Plasmodium
Total genera:	1
Total organisms:	3
Putative	0
Paralogs:	
Locus:	
Structures:	6
CDDs:	PTZ00021 , smart00848:Inhibitor_I29(superfamily:cl07031) , pfam08246:Inhibitor_I29(superfamily:cl07031) , cd02248:Peptidase_C1A(superfamily:cl00298)

Filters

Protein Table

Organism	Protein name	Accession	Locus_tag	Length (aa)	BLINK
<i>Plasmodium falciparum</i> 3D7	falcipain-2B	XP_001347832	PF11_0161	482	◆
<i>Plasmodium falciparum</i> 3D7	falcipain-3	XP_001347833	PF11_0162	492	◆
<i>Plasmodium falciparum</i> 3D7	falcipain-2A	XP_001347836	PF11_0165	484	◆
<i>Plasmodium knowlesi</i> strain H	P.knowlesi ortholog of falcipain	XP_002259151	PKH_091240	477	◆
<i>Plasmodium knowlesi</i> strain H	P.knowlesi ortholog of falcipain	XP_002259152	PKH_091250	495	◆
<i>Plasmodium knowlesi</i> strain H	P.knowlesi ortholog of falcipain	XP_002259153	PKH_091260	479	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-2	XP_001615272	PVX_091405	484	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-3	XP_001615273	PVX_091410	495	◆
<i>Plasmodium vivax</i> Sal-1	vivapain-2	XP_001615274	PVX_091415	487	◆

This is a cysteine protease, originally identified and characterized in *Plasmodium falciparum*; it hydrolyses the erythrocyte hemoglobin into its amino acid constituents, which are used by the parasite for protein synthesis. In this cluster of proteins, “facipain” is present in 4 different *Plasmodium* species. Falcipain-2 differs from the falcipains in other species (i.e., vivapain -2 and -3, berghepain, etc.,) as well as within the *P. falciparum* (i.e., falcipain -3) in sequence, in the timing of expression and in the acidic environment needed for enzymatic activation, but they all appear to have the same function (20). Interestingly, the two *P. falciparum* falcipain-2 proteins in this cluster are each located in a different part of chromosome 11, although they share high amino acid sequence homology and a seemingly identical function. The differences here also appear to be in expression timing and in the level of expression. Falcipain-2A (PF11_0165) appears to be expressed earlier in the trophozoite stage and in higher amounts than falcipain-2B (PF-11_0161) (10, 20).

Also of interest is the fact that cysteine protease inhibitors have been shown to have potent anti-malarial effects. Indeed, because this family of proteases shares low sequence identity with their human counterparts, they have been given serious consideration as potential drug targets.

Plants

Although protein clustering is not specifically geared towards clustering for orthologs or paralogs, clustering does provide a view into how different proteins are related as seen in the cluster PLN03595 shown below.

Organism	Protein name	Accession	Locus_tag	Length (aa)	UniProtKB / SwissProt
Arabidopsis lyrata subsp. lyrata	hypothetical protein	XP_002668009	ARALYDRAFT_914800	1116	
Arabidopsis lyrata subsp. lyrata	phytochrome D	XP_002668148	ARALYDRAFT_915133	1165	
Arabidopsis lyrata subsp. lyrata	hypothetical protein	XP_002668441	ARALYDRAFT_493637	1112	
Arabidopsis lyrata subsp. lyrata	phytochrome B	XP_002886263	ARALYDRAFT_480851	1163	
Arabidopsis lyrata subsp. lyrata	hypothetical protein	XP_002892510	ARALYDRAFT_471053	1122	
Arabidopsis thaliana	phytochrome A	NP_011117256	AT1G09570	1014	P14712
Arabidopsis thaliana	phytochrome A	NP_172428	AT1G09570	1122	P14712
Arabidopsis thaliana	phytochrome B	NP_179469	AT2G15790	1172	P14713
Arabidopsis thaliana	phytochrome D	NP_193360	AT4G16250	1164	P42497
Arabidopsis thaliana	phytochrome E	NP_193547	AT4G18130	1112	P42498_Q56Y99
Arabidopsis thaliana	phytochrome C	NP_198433	AT5G35840	1111	P14714
Brachypodium distachyon	phytochrome B-like	XP_003558068	LOC100829838	1181	
Brachypodium distachyon	phytochrome C-like	XP_003559446	LOC100834357	1140	
Brachypodium distachyon	phytochrome A type 3-like	XP_003560548	LOC100836209	1131	
Glycine max	phytochrome A	NP_001238206	phyA	1131	
Glycine max	phytochrome B-like	XP_003533157	LOC100799831	1137	
Glycine max	phytochrome E-like	XP_003535030	LOC100808192	1120	
Glycine max	phytochrome B-like isoform 1	XP_003546314	LOC100794865	1149	
Glycine max	phytochrome E-like	XP_003546574	LOC100800339	1121	
Glycine max	phytochrome type A-like	XP_003554593	LOC100791098	1130	
Glycine max	phytochrome type A-like	XP_003555766	LOC100790763	1123	
Selaginella moellendorffii	hypothetical protein	XP_002991119	SELMODRAFT_161430	1143	
Selaginella moellendorffii	hypothetical protein	XP_002991641	SELMODRAFT_161807	1142	
Sorghum bicolor	hypothetical protein	XP_002463975	SORBIDRAFT_01g009930	1131	
Sorghum bicolor	hypothetical protein	XP_002468441	SORBIDRAFT_01g007850	1135	
Sorghum bicolor	hypothetical protein	XP_002467973	SORBIDRAFT_01g037340	1178	
Vitis vinifera	phytochrome C	XP_002268724	LOC100258014	1118	
Vitis vinifera	phytochrome E	XP_002271671	PHYE	1124	
Vitis vinifera	phytochrome B-like	XP_002278263	LOC100261882	1129	
Vitis vinifera	phytochrome A1	XP_002278610	PHYA	1124	

Organism	Protein name	Accession	Locus_tag	Length (aa)	UniProtKB / SwissProt
Medicago truncatula	Phytochrome A	XP_003591274	MTR_1g085160	1171	
Medicago truncatula	Phytochrome b1	XP_003594734	MTR_1g034040	1198	
Medicago truncatula	Phytochrome E	XP_003595571	MTR_2g049520	1122	
Oryza sativa Japonica Group	hypothetical protein	NP_001049910	Os03g0309200	1120	Q10MG9
Oryza sativa Japonica Group	hypothetical protein	NP_001051096	Os03g0719800	1128	Q10DU0
Oryza sativa Japonica Group	hypothetical protein	NP_001051296	Os03g0752100	1137	Q10CQ8
Physcomitrella patens subsp. patens	phytochrome Sc	XP_001754366	PHYPADRAFT_115388	1124	
Physcomitrella patens subsp. patens	phytochrome 5a	XP_001761145	PHYPADRAFT_200532	1123	
Physcomitrella patens subsp. patens	phytochrome 3	XP_001766035	PHYPADRAFT_185248	1123	
Physcomitrella patens subsp. patens	phytochrome 5b3	XP_001767224	PHYPADRAFT_165601	1131	
Physcomitrella patens subsp. patens	phytochrome 4	XP_001773550	PHYPADRAFT_218861	1126	
Physcomitrella patens subsp. patens	phytochrome 1	XP_001778155	PHYPADRAFT_222399	1123	
Physcomitrella patens subsp. patens	phytochrome 2	XP_001782339	PHYPADRAFT_225644	1130	
Populus trichocarpa	hypothetical protein	XP_002312330	POPTRDRAFT_832666	1142	
Populus trichocarpa	phytochrome B2	XP_002314949	POPTRDRAFT_1091155	1146	
Populus trichocarpa	phytochrome	XP_002318913	POPTRDRAFT_729311	1126	
Ricinus communis	phytochrome A, putative	XP_002512596	RCOM_1437130	1124	
Ricinus communis	phytochrome B, putative	XP_002519230	RCOM_1000590	1141	
Ricinus communis	phytochrome B, putative	XP_002519749	RCOM_0634650	1131	

PLN03595 represents a family of photoreceptors involved in the photoperiodic control of plant growth and development. This family includes diverse but structurally conserved proteins. They are expressed in different plant organs under varying light conditions. Phylogenetic analyses suggest that the phytochrome gene family is composed of four subfamilies, *PHYA*, *PHYB*, *PHYC/F*, and *PHYE*. *Arabidopsis thaliana* has an additional *PHYD* gene that originated from the *PHYB* gene after a more recent gene duplication, and there is some functional redundancy between these two. *PHYA* and its paralog *PHYC* are found in monocots as well as in dicots, but *PHYC* is missing in some dicot lineages. Rice only has three phytochrome genes: *PHYA*, *PHYB*, and *PHYC*. Monocotyledonous plants are also known to lack several members of *PHYB* subfamily. Phytochromes exhibit distinct and cooperative functions. Mutant analysis has shown that, in rice, *phyA* and *phyB* act in a highly redundant manner to control de-etiolation under continuous red light. Under continuous far-red light, *phyA* and *phyC* are involved in photoperception, but the photoperception mode of *phyC* differs between rice and *Arabidopsis* (21).

We also used proteins of the photosynthesis system as a model for clustering validation. The photosynthesis system has been chosen as it is well conserved and characterized throughout the plant kingdom. As of now, 116 clusters were identified in plants using the “photos” keyword that were annotated and curated. The number of proteins per cluster ranged from 2 to 100. These photosynthesis proteins belong to 6 or more organisms out of 23 distinct genomes. One cluster, PLN00033, contains 22 proteins belonging to 19 organisms and corresponds to the photosystem II stability/assembly factor, which is coherent with the central role this protein plays in chloroplast biogenesis and photosystem stability (25, 26, and 27). Interestingly, 10 out of these 22 proteins from 8 different organisms are annotated as hypothetical proteins.

The second most conserved cluster, PLN00037, contains 34 proteins belonging to 18 organisms and corresponds to photosystem II oxygen-evolving enhancer protein 1 (Psb O). This situation is coherent with its crucial role in photosynthesis. Here again 11

proteins are annotated as hypothetical. Generally, the most conserved proteins in the plant kingdom are known for their central role in plant growth and development. The clustering can be used to hypothesize about the most important proteins whose function is worth analyzing further. For example, PLN03089, a cluster of 65 hypothetical proteins present in both monocots and dicots, should attract more interest. Although the proteins have homology with the Glutamate-gated kainate-type ion channel receptor subunit GluR5 in *Medicago truncatula*, there is no convincing evidence of such function.

The clusters containing protein specific to a group or subgroup of plants are also very interesting to study. Examples of such clusters are the ones with proteins present in all *viridiplantae* (PLN00046: photosystem I reaction center subunit O; PLN00054: photosystem I reaction center subunit N; PLN00049: carboxyl-terminal processing protease). The corresponding proteins would be among the most important in plant photosynthesis. Some other clusters contain proteins from a specific subgroup such as algae (PLN00100).

Access

Protein Clusters are presented in NCBI's Entrez system (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters>)

The first public release of the Protein Clusters in NCBI's Entrez interface was in April 2007, and initially consisted of only prokaryotic clusters (15). The Entrez system provides a mechanism for the search, retrieval, and linkage between Protein Clusters and other NCBI databases, as well as external resources. Clusters can be searched by general text terms, and also by specific protein or gene names.

Limits and Advanced search allow clusters to be browsed by function and filtered by size and organism group. A table browser allows users to sort by the content of each column by clicking on the column header.

Search by organism name, locus tag or protein name

Hide identical proteins:

Protein Table

Clade ID	Organism	Protein name	Accession	Locus_tag	Length (aa)	Items 1 - 20 of 31 < Prev Page <input type="text" value="1"/> of	
						Identical group	
22152	Actinopolyspora mortivallis DSM 44261	hypothetical protein	WP_019853821	ACTMO_06285	514	WP_019853821	
21263	Amphibacillus jilinensis Y1	hypothetical protein	WP_017470882	B494_02995	484	WP_017470882	
21267	Bacillus alcalophilus ATCC 27647	glycosyl hydrolase family protein	WP_003322074	BalcAV_07657	477	WP_003322074	
21936	Bacillus bataviensis LMG 21833	SacC2	WP_007084638	BABA_D8076	492	WP_007084638	
21271	Bacillus endophyticus 2102	hypothetical protein	WP_019393615	A360_15930	531	WP_019393615	
21935	Bacillus nealsonii AAU1	glycosyl hydrolase family protein	WP_016205333	A499_24244	486	WP_016205333	
20034	Bacillus sp. 10403023	glycosyl hydrolase family protein	WP_010677742	B1040_010100015199	485	WP_010677742	
21943	Halobacillus sp. BAB-2008	glycoside hydrolase	WP_008633984	D479_04248	533	WP_008633984	
22245	Nocardiopsis salina YIM 90010	hypothetical protein	WP_017612606	D474_05620	515	WP_017612606	
19975	Paenibacillus elgii B69	SacC2	WP_010492966	PelgB_010100005566	489	WP_010492966	
19975	Paenibacillus elgii B69	glycoside hydrolase family protein	WP_010495814	PelgB_010100015198	492	WP_010495814	
21855	Paenibacillus ginsengihumi DSM 21568	hypothetical protein	WP_019537290	F591_24885	489	WP_019537290	
22129	Paenibacillus lactis 154	Glycosyl hydrolase family 32 domain protein	WP_007132088	PaelaDRAFT_4928	487	WP_007132088	

Protein clusters are available for download from the FTP directory (<ftp://ftp.ncbi.nih.gov/genomes/Bacteria/CLUSTERS/>) by date and by major taxonomic groups.

Related Tools

Concise Protein BLAST

The Concise Protein database contains proteins from all clusters, as well as all singletons (not clustered proteins). From the clustered set, a representative at the genus level is chosen in order to reduce the data set. Results are therefore available rapidly and the results that are returned provide a broader taxonomic range due to this data reduction.

Concise BLAST provides an option for both protein and nucleotide searches using BLASTP and BLASTX, respectively.

<http://www.ncbi.nlm.nih.gov/genomes/prokhits.cgi>

RPS-BLAST

RPS-BLAST searches against pre-calculated position-specific scoring matrices (PSSMs) created during conserved domain processing for the CD-search tool. Therefore, only protein sequences are used for this type of search. PSSMs from the curated cluster set have

been added to CDD and are also used in pre-calculated conserved domain hits available from the link menu on protein sequences and reported on each GenPept record. The curated set of PSSMs can be searched using RPS-BLAST and a protein sequence at <http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi> or the full set of PSSMs for all curated clusters is available from FTP.

ProtMap

ProtMap is a graphical gene neighborhood tool that displays clickable, linked genes upstream and downstream of the target. The tool provides useful graphical representations of the members of a particular cluster in their genome environments. All members of the cluster of interest are mapped to their genome position, and the tool displays genomic segments coding for each member of the cluster. If the genome sequence is larger than 20KB, only the relevant 10KB portion of it is shown. Users can search for the cluster of interest by using cluster access or the COG/VOG attribute of the cluster. The display is centered on protein members of the cluster. Users can select additional sets of related proteins by clicking on the corresponding colored arrows depicting a protein, or find a cluster of interest by name, protein accession, or gene locus_tag. This resource is useful in identifying paralogs as well as missing or incorrectly annotated genes.

<http://www.ncbi.nlm.nih.gov/sutils/protmap.cgi>

References

1. Ahn YY, Bagrow J, Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*. 2010 Aug 5;466(7):761–765. PubMed PMID: 20562860.
2. Cameron M, Bernstein Y, Williams HE. Clustered Sequence Representation for Fast Homology Search. *J Comput Biol*. 2007 Jun;14(5):594–614. PubMed PMID: 17683263.
3. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. *Nature*. 2008 May 1;453:98–100. PubMed PMID: 18451861.
4. Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to algorithms*. 3rd Edition, The MIT Press; 2009.
5. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010 Oct 1;26(19):2460–2461. PubMed PMID: 20709691.
6. Fairman-Williams ME, Jankowsky E. Unwinding initiation by the viral RNA helicase NPH-II. *J Mol Biol*. 2012 Feb 3;415(5):819–832. PubMed PMID: 22155080.
7. Fitch WM. Distinguishing homologous from analogous proteins. *Syst Zool*. 1970 Jun; 19:99–106. PubMed PMID: 5449325.
8. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995 Jul 28;269(5223):496–512. PubMed PMID: 7542800.

9. Gargantini PR, Serradell MC, Torri A, Lujan HD. Putative SF2 helicases of the early-branching eukaryote Giardia lamblia are involved in antigenic variation and parasite differentiation into cysts. *BMC Microbiol.* 2012 Nov 28;12:284. PubMed PMID: 23190735.
10. Goh LL, Sim TS. Homology modeling and mutagenesis analyses of Plasmodium falciparum falcipain 2A: implications for rational drug design. *Biochem Biophys Res Commun.* 2004 Oct 15;323(2):565–572. PubMed PMID: 15369788.
11. Hawkins LK, Wold WS. A 12,500 MW protein is coded by region E3 of adenovirus. *Virology.* 1992 Jun;188(2):486–494. PubMed PMID: 1585632.
12. Holm L, Sander C. Removing near-neighbor redundancy from large protein sequence collections. *Bioinformatics.* 1998 Jun;14(5):423–429. PubMed PMID: 9682055.
13. Hughes AL, Irausquin S, Friedman R. The evolutionary biology of poxviruses. *Infect Genet Evol.* 2010 Jan;10(1):50–59. PubMed PMID: 19833230.
14. Kaplan N, Sasson O, Inbar U, Friedlich M, Fromer M, Fleischer H, Portugaly E, Linial N, Linial M. ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.* 2005 Jan 1;33(Database issue):D216–D218. PubMed PMID: 15608180.
15. Klimke W, Agarwala R, Badretdin A, Chetvernin S, Ciufo S, Fedorov B, Kiryutin B, O'Neill K, Resch W, Resenchuk S, Schafer S, Tolstoy I, Tatusova T. The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 2009 Jan;37(Database issue):D216–23. PubMed PMID: 18940865.
16. Koonin EV. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* 2005;39:309–38. PubMed PMID: 16285863.
17. Krause A, Stoye J, Vingron M. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics.* 2005 Jan 22;6:6–15. PubMed PMID: 15663796.
18. Loewenstein Y, Portugaly E, Fromer M, Linial M. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics.* 2008 Jul 1;24(13):i41–49. PubMed PMID: 18586742.
19. Ranji A, Boris-Lawrie K. RNA helicases: emerging roles in viral replication and the host innate response. *RNA Biol.* 2010 Nov-Dec;7(6):775–87. PubMed PMID: 21173576.
20. Shenai BR, Sijwali PS, Singh A, Rosenthal PJ. Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of Plasmodium falciparum. *J Biol Chem.* 2000 Sep 15;275(37):29000–29010. PubMed PMID: 10887194.
21. Takano M, Inagaki N, Xie X, Yuzurihara N, Hihara F, Ishizuka T, Yano M, Nishimura M, Miyao A, Hirochika H, Shinomura T. Distinct and cooperative functions of phytochromes A, B, and C in the control of deetiolation and flowering in rice. *Plant Cell.* 2005 Dec;17(12):3311–3325. PubMed PMID: 16278346.
22. Tarjan RE. Data structures and network algorithms, CBMS 44, Society for Industrial and Applied Mathematics, Philadelphia, PA; 1983.
23. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science.* 1997 Oct 24;278(5338):631–637. PubMed PMID: 9381173.

24. Zaslavsky L, Tatusova T. Mining the NCBI influenza sequence database: adaptive grouping of BLAST results using precalculated neighbor indexing. PLoS Curr. 2009;1:RRN1124. PubMed PMID: 20029662.
25. Plücken H1. Müller B, Grohmann D, Westhoff P, Eichacker LA. The HCF136 protein is essential for assembly of the photosystem II reaction center in *Arabidopsis thaliana*. FEBS Lett. 2002 Dec 4;532(1-2):85–90. PubMed PMID: 12459468.
26. Meurer J, Plücken H, Kowallik KV, Westhoff P. A nuclear-encoded protein of prokaryotic origin is essential for the stability of photosystem II in *Arabidopsis thaliana*. *EMBO J.* 1998 Sep 15;17(18):5286-5297.
27. Peltier JB, Emanuelsson O, Kalume DE, Ytterberg J, Friso G, Rudella A, Liberles DA, Söderberg L, Roepstorff P, von Heijne G, van Wijk KJ. Central functions of the luminal and peripheral thylakoid proteome of *Arabidopsis* determined by experimentation and genome-wide prediction. *Plant Cell.* 2002 Jan;14(1):211–236. PubMed PMID: 11826309.

Small Molecules and Biological Assays

Small Molecules and Biological Activities

Rana Morris^{✉1}

Created: December 9, 2013.

Scope

The chemical and bioactivity resources at the National Center for Biotechnology Information (NCBI) are coordinated as part of the [PubChem Project](#) with data accessible as part of the PubChem Compound and Substance databases and PubChem BioAssay database, respectively. Currently the PubChem set of resources provides access to information about over [36 million chemical compounds](#) and experimental studies with results for over [700 thousand biological activity assays](#).

History

The PubChem databases were originally developed to serve as a central data repository for the NIH-sponsored [Molecular Libraries Roadmap Project](#) (Pilot phase 2004–2008). The purpose of this project was to identify novel research reagents (protein function modulators and molecular probes) and discover candidate compounds for drug development. A large set of substances were tested for biological activity in high-throughput assays. The substances were cataloged in the PubChem Substance database and the description, supporting information, and results of the studies were stored in the PubChem BioAssay database. To increase the utility of these databases, a derivative database, PubChem Compound, was created by the PubChem group to consolidate and provide a cross-referencing platform for common chemical components of the substances.

Since its inception, the scope of these resources has been dramatically expanded for use by an expanding and diverse variety of researchers, from organic chemists to molecular biologists to drug design specialists. In addition to the data from the original NIH Molecular Libraries Initiative, substance information and biological activity assays are now submitted by various organizations:

- NIH Molecular Libraries participants and other laboratories with Bioassay Screening Results
- The NIH Substance Repository
- Organizations specializing in these areas:
 - Biological Properties
 - Chemical Reactions
 - Database Vendors

¹ NCBI; Email: morrisrc@ncbi.nlm.nih.gov.

[✉] Corresponding author.

- Imaging Agents
- Journal Publishers
- Metabolic Pathways
- Patents
- Physical Properties
- Protein 3D Structures
- siRNA Reagent Providers
- Substance Vendors
- Theoretical Properties
- Toxicology Properties

Dataflow

Data Submissions to PubChem

The [PubChem Upload portal](#) enables researchers, laboratories, and organizations to submit small molecule and bioassay data to the PubChem Substance and PubChem BioAssay databases. Data can be submitted to the PubChem Substance or BioAssay databases in any of the following ways:

- PubChem Submission Wizards—for small numbers of submissions, a series of guided forms assist novice submitters in entering substance and/or assay data without requiring knowledge of detailed data specifications. After data is typed or imported, the Upload system will prepare a properly formatted file that conforms to the data specifications.
- Pre-formatted File Uploads—for small or large numbers of records, Upload accepts pre-formatted files in several formats which may be GZIP-compressed. For PubChem Substances, formats include: SDF – Chemical Structure Data File, CSV – Comma-separated Variables, and for BioAssay formats include: ASN.1 – Abstract Syntax Notation 1, XML – Extensible Markup Language, CSV – Comma-separated Variables).
- FTP Depositions of Pre-formatted Files—for large and/or frequent data uploads, depositions of pre-formatted files by FTP are recommended. Once the data are in the FTP directory, submitters can review and edit them, and commit them to PubChem using the Upload system.

While the PubChem group and the user community-at-large recommend that submitters supply as much information about their substances and bioactivity assays as possible (including such things as chemical structures of tested substances and related PubMed records), the following are the minimal data that must be submitted:

To PubChem Substance:

- Submitter information
 - Name
 - Organization

- Address
- Contact Information
- Source Classification
- Signed [Data Transfer Agreement](#)
- Substance “name”
- Source’s identifier

To PubChem BioAssay:

- Submitter information (same as PubChem Substance Submitter information)
- Assay Information
 - Descriptive Assay title
 - Source’s identifier
 - Assay Information
 - Assay type
 - List of Substances tested
 - Description
 - Protocol
 - Data Table with Defined Outcomes
 - Definitions

The PubChem Upload system can also be used for updating existing PubChem records. Existing PubChem Substance or BioAssay records can be retrieved and loaded into the Web interface for direct editing and review of the revised records. Once corrections have been made and verified, they will be committed and stored in the system.

The PubChem databases have a rolling update schedule. As new data is submitted to PubChem, it is processed for addition to both the relevant primary databases as well as to PubChem Compound (when applicable). Data is released to the website and updated on the FTP site as soon as it is ready, generally within 48 hours.

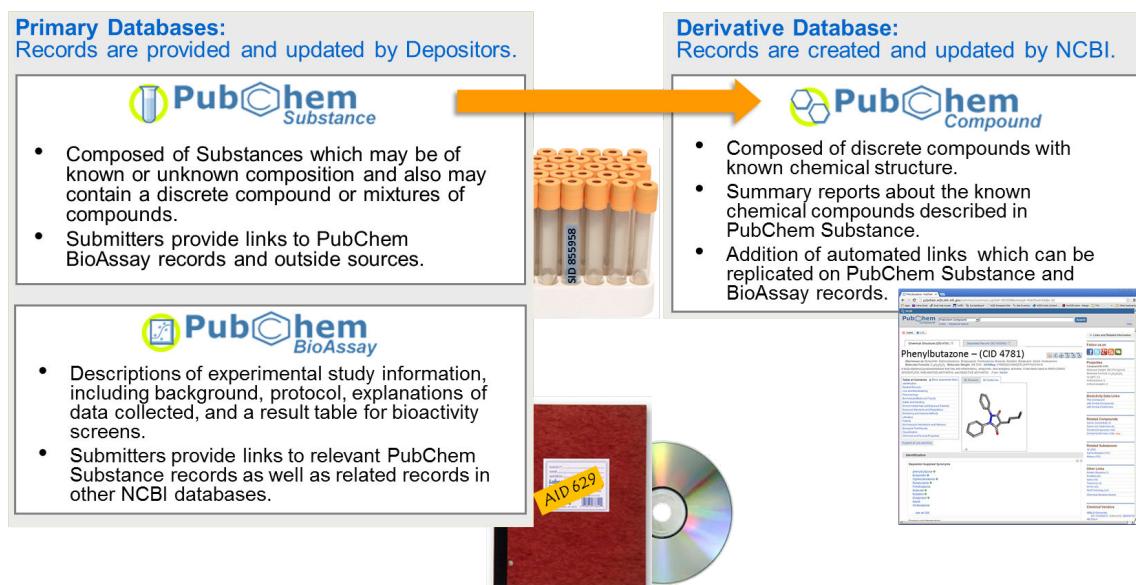
[Primary and Derivative Databases](#)

PubChem’s primary databases contain records of data as submitted to the PubChem group:

- [PubChem Substance](#): Information provided to PubChem by submitters. Substances may be of known or unknown chemical composition, and may contain a discrete compound or mixtures of compounds. Links and information in the record are provided and maintained by submitters. Please note that multiple records representing the same substances may be submitted by different laboratories or organizations.
- [PubChem BioAssay](#): Information provided to PubChem by submitters including study background, protocol, and results for experimental bioactivity screens of chemical substances that have been submitted to PubChem Substance.

The PubChem Compound database is a derivative database created by a consolidation of PubChem Substance Records with additional curated information created and maintained by the PubChem group.

- **PubChem Compound:** Composed of discrete chemicals with known molecular structure. Summary reports are generated from the known chemical compounds described in PubChem Substance and are supplemented with additional information. Links back to related PubChem Substances, PubChem BioAssays, and other NCBI and external sites are provided by the PubChem group.



For example, a Molecular Libraries Screening Center Laboratory at Emory University submitted an Estrogen Receptor-alpha coactivator high-throughput chemical binding screen. This was given the PubChem BioAssay Identifier [AID 629](#). Around the same time, a collaborating group, the Molecular Libraries Screening Center's Small Molecule Repository, submitted information about the substances in their system including 86,096 that were tested in this particular assay. One of those tested and found to be active in AID 629 was given the PubChem Substance Identifier [SID 855958](#), which was one of the many linked to the PubChem BioAssay record.

In addition to AID 629, this same substance (SID 855958) was tested in over 2,900 assays by various groups. Further information about this chemical was uploaded to the PubChem group by several other organizations including biological property databases, journal publishers, and chemical vendors as submissions to create their own PubChem Substance records. As of December 1, 2013, 284 PubChem Substance records were submitted for this same chemical. The PubChem group then consolidated all information and submitted links for this chemical and included additional calculated chemical properties and information provided by several other key sources (NLM's Medical Subject Headings, MeSH, DailyMed, and Hazardous Substances Data Bank [HSDB]) and created a reference record in the PubChem Compound database with a PubChem Compound

identifier [CID 4781](#). This record contains links back to all corresponding PubChem Substance and BioAssay records.

Access

Searching PubChem Databases Using the Web Interface

NCBI's chemical and bioactivities data can be accessed through the NCBI website and the [PubChem project homepage](#). Searching for PubChem data can be performed through the common Web Search (Entrez) mechanism. For the [PubChem Compound](#), [PubChem Substance](#) and [PubChem BioAssay](#) databases, there are highly specialized and useful Limits pages to assist in specifying key types of data. These databases also maintain Entrez Advanced pages with Advanced Search Builders and Search History tables.

For users who would like to retrieve records from a list of PubChem-related record identifiers (PubChem Compound Identifiers – CIDs, PubChem Substance Identifiers – SIDs, and BioAssay Identifiers – AIDs), [NCBI's Batch Entrez](#) permits the upload of a text file with subsequent retrieval from the selected database.

In addition to text-based searches using the Web Search (Entrez) system, the PubChem group has created a [Chemical Structure Search](#) mechanism in which two dimensional drawings or textural representations of chemical structures (SMILES – Simplified molecular-input line-entry system, SMARTS – Smiles arbitrary target specification, InChIs – International Chemical Identifiers) can be used to query the PubChem Compound or Substance databases to find records for chemicals with matching or similar structures.

Downloading PubChem Data from the Web

After performing a search, groups of retrieved records can be downloaded using the “Send to” function common to all Web Search (Entrez) displays. Individual PubChem records can be downloaded directly from the Web interface with links at the top-right-hand side of the screen. PubChem records display record download buttons for available file formats (SDF, CSV, ASN.1, XML).

The PubChem group has also created Download Facility tools which enable the quick upload of a list of PubChem record identifiers (CIDs, SIDs, or AIDs) and download of PubChem Compound or Substance information ([PubChem Structure Download Facility](#)) or PubChem BioAssay information ([BioAssay Download Facility](#)) in a number of file formats.

For users who would like to download the data in bulk, files are available on the [PubChem FTP site](#), which is mirrored on the [Aspera-plugin version of the NCBI FTP site](#).

Accessing PubChem Data Using Programming Interfaces

The PubChem group provides scripting access for users in a number of formats. As with other NCBI databases, the PubChem resources are accessible using the [NCBI Entrez Utilities API Interface \(EUtilities\)](#). In addition, the Group has developed a PubChem-specific RESTful API Interface (PUG-REST) as well as an [XML-intensive PubChem Power User Gateway \(PUG\)](#).

Related Tools

[Standardization Service](#) mimics the initial steps of PubChem's data processing pipeline and, therefore, enables users to see how submitted structures would be handled. In addition, this service will enable the conversion of chemical structures in SMILES, InChI, or SDF formats to a different format. Standardization through this service is limited to a single structure at a time.

[Identifier Exchange Service](#) enables the upload and conversion of a list of identifiers (CIDs, SIDs, or Registry IDs), InChIs or synonyms with retrieval of identifiers for identical or similar PubChem Compound or Substance records. A file or correspondence table can be downloaded.

[Data Dicer](#) is an alternative to the Web Search (Entrez) Limits page, providing guided searches and tabular retrievals for information pertaining to bioactivity outcomes for gene/protein targets or screened small molecules in bioactivity assays listed in the PubChem BioAssay database.

[Classification Browser](#) searches PubChem records annotated with hierarchies/terms and displays the distribution of these in the context of the specific ontology. Please note that this only operates on the subset of PubChem records that have been annotated with terms from the available hierarchies.

[Structure Search](#) searches for identical or similar chemical structure records using CID or SIDs, Names (SMILES, SMARTS, InChI, Synonyms, MeSH terms), Molecular Formulas, Molecular Weights, SDF file, or even hand-drawn structures. Similarity thresholds can be customized and record retrieval filters can be set using an interface similar to that of the PubChem Web Search (Entrez) Limits page or using search histories from PubChem Compound or Substance databases. The 2D structure searching strategy uses a modified-Tanimoto scoring system, while the 3D structure searching strategy is based on comparison of calculated shapes of 3D conformers with positions of the functional groups. Search results displayed sorted in decreasing order of the resulting calculated 2D- or 3D-similarity score are retrieved and displayed as a traditional PubChem Compound or Substance search results list.

[Score Matrix Service](#) enables the downloading of 2D- or 3D-similarity scoring matrix values calculated by the PubChem Structure Search.

[Structure Clustering Tool](#) assists in exploration of chemical-structure space and displays chemical structural similarity and population diversity relationships in a tree format using the Single Linkage algorithm. Calculations are based on 2D-similarity scores (modified-Tanimoto scoring) or 3D-similarity scores (conformer, functional group scoring).

[BioActivity Summary/DataTable](#) searches and retrieves biological screening results for individual or a set of chemical samples or displays the information content of PubChem BioAssay records. Tables are downloadable in a number of formats.

[BioActivity Plot Service](#) displays bioactivity assay data in customizable scatter plots or histograms.

[Structure-Activity Relationship HeatMap](#) enables interactive visualization for exploring PubChem data, displaying a chemical structure tree on a Y-axis with the bioassays clustered by relatedness on the X-axis. Activity data values are shown as colored squares in the resulting grid. The display of the grid is customizable with respect to color and zoom level and can be downloaded.

[Webpage Widgets](#) provides code for configurable PubChem data displays including record summaries as well as tables for placement in Web pages.

References

The PubChem Project

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W623-33. Epub 2009 Jun 4. doi: [10.1093/nar/gkp456..](https://doi.org/10.1093/nar/gkp456)

PubChem Substance and PubChem Compound Databases

Bolton E, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Chapter 12 IN Annual Reports in Computational Chemistry, Volume 4, American Chemical Society, Washington, DC, 2008 Apr.

PubChem BioAssay Database

Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. PubChem's BioAssay Database. Nucleic Acids Res. 2012 Jan;40(1):D400-12. (Epub 2011 Dec 2) doi: [10.1093/nar/gkr1132..](https://doi.org/10.1093/nar/gkr1132..)

NCBI PubChem BioAssay Database

Yanli Wang¹ and Stephen H Bryant¹

Created: March 14, 2014.

Scope

NCBI's PubChem BioAssay database (1-5) (<http://pubchem.ncbi.nlm.nih.gov>) is a public repository for archiving biological tests of small molecules and siRNA reagents. Small molecule bioactivity data contained in the BioAssay database consist of information generated through high-throughput screening experiments, medicinal chemistry studies, chemical biology research, as well as literature curation. In addition, the BioAssay database contains data from RNAi screens against targeted genes or complete genomes aiming to identify critical genes responsible for a biological process or disease condition. BioAssay data continue to grow rapidly and are integrated with the rest of the NCBI resources, making PubChem a widely used public information system for accelerating chemical biology research and drug development.

The mission of the PubChem resource is to deliver free and easy access to all deposited data, and to provide intuitive data analysis tools. The PubChem BioAssay database is organized as a set of relational databases deployed on Microsoft SQL servers. The infrastructure allows for seamlessly storing the submitted BioAssay records, tracking and versioning subsequent updates, and supporting data retrieval and analysis.

As a repository, PubChem constantly optimizes and develops its data submission system, answering many demands of both high and low volume depositors.

PubChem's BioAssay data is integrated into the NCBI [Entrez information retrieval system](#), thus making PubChem data searchable and accessible by Entrez queries. In addition, the [PubChem information platform](#) provides Web-based and programmatic tools for users to search, review, and download bioactivity data for a BioAssay record, a compound, a molecular target, or a publication. PubChem also provides a suite of integrated services enabling users to collect, compare, and analyze biological test results across multiple assay bioassay projects.

PubChem BioAssay Standard & Data Model

PubChem provides a flexible BioAssay data model (1,2) and database schema to accommodate bioactivity data produced by diverse experimental procedures. The data model continues to expand to support new types of information generated as experimental methodologies evolve.

¹ NCBI.

An assay record is represented by a unique PubChem BioAssay accession, or AID. A BioAssay record is organized in two parts, the assay description and the assay results, and has links to the corresponding records of the substances that were tested by the assay, which are stored in the PubChem Substance database. Updates are tracked and a BioAssay record is versioned if any part of the record gets updated.

The assay description section includes an assay title, data source, assay description, experimental protocols, tested reagent category (e.g., small molecule vs siRNA), comments, assay targets, cross references to other databases at NCBI, and assay readout descriptions.

The assay result section includes the results for all tested substances. Results reported per substance can include regular assay readout, such as IC₅₀ inhibition activity at a given test concentration. Per-substance assay data can also include annotations, including target description; comment on the individual biological test result; cross-links to other NCBI resources, such as Gene ID and PubMed ID (PMID); and URLs to the depositor's website. Assay data are provided in a tabular format, with one tested substance per row and one assay test readout or annotation per column. A substance needs not have results reported for all defined test readouts. Multiple test result field definitions may be specified per assay, each with a unique test identifier (TID), name, description, data type, data unit, and annotation for cross-references. As a result, one can report replications of a specific readout as well as one or multiple series of dose-response data points. An example BioAssay record (Dose response biochemical screening assay for inhibitors of c-Jun N-Terminal Kinase 3 (JNK3)) can be accessed at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1284>.

Biological screening data submitted to PubChem is diverse and assay specific. As such, there are no specific requirements on the presence of particular test results or assay readouts; however, PubChem requires a summary result for each tested substance or chemical sample. The summary result is two-fold: bioactivity outcome and bioactivity score. The "bioactivity outcome" partitions results and includes five categories: chemical probe, active, inactive, inconclusive, and unspecified. The "bioactivity score" facilitates the separation of highly active compounds from the inactive ones. Many biological assays employ a dose-response scheme, with a primary endpoint. PubChem requires that this key readout, denoted as an "active concentration summary," has micro-molar units, and that the experimental concentrations for the corresponding dose-response readouts (referred to as "tested concentrations," also in micro-molar units) be designated on the respective test result fields as an attribution. These specialized readouts, together with the summary results, allow PubChem users to classify and rank hits of a screening test. They also support cross links from the BioAssay record to PubChem compounds, and allow PubChem to provide tools to enable in-depth data analysis and comparison across multiple BioAssay results.

PubChem BioAssay tracks the screening stage of a high throughput screening (HTS) assay project, if multiple BioAssay records are submitted for the project. The stages of an HTS

project include: “screening,” a primary high-throughput assay where the activity outcome is based on percentage inhibition from a single dose; “confirmatory,” a low-throughput assay where the activity outcome is based on a dose-response relationship with multiple tested concentrations; “summary,” an assay summarizing information from multiple BioAssay submissions for validated chemical probes or small molecule leads ; and “other,” assays that do not fit the previous categories.

Assay targets are important information that should also be included in BioAssay records when possible. The “classical” assay model allows for the specification of assay target, either a single molecule or a complex, for the entire assay record, along with descriptions for the target molecules including gene and taxonomy information. In this model, the bioactivity outcomes provided in the entire assay dataset are solely for the specific target or target complex; for example, to describe the biological effect of the small molecules on the functionality of one enzyme.

PubChem also supports the presentation and annotation of multiple highly-related bioactivity outcomes, such as a profiling assay against a panel of molecular targets, in a single assay. Such a panel-type PubChem BioAssay record can contain multiple test readouts and respective bioactivity outcome annotations for each individual target, or similarly for each individual cell line or species defined within the “panel.” Each such target, cell line or species is regarded as a “panel component” in the data model, which can have its own “bioactivity outcome” or “active concentration” designated. An example panel assay (Kinase Inhibitor Selectivity Profiling Assay) can be accessed at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1433>.

A third bioassayBioAssay data model serves a general purpose for assays with multiple and substance- specific targets, but is primarily designed to support the representation of gene targets and test results for siRNA screenings, where each tested siRNA is aimed to suppress a specific gene target by design. This model allows one to specify a specific target and the relevant information for each individually tested sample (such as a siRNA reagent). As examples, an RNAi screening bioassayBioAssay (RNAi Global Initiative pilot viability screen of human kinase and cell cycle genes) can be accessed at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1622>; and a small molecule screening bioassayBioAssay (Experimentally measured binding affinity data derived from PDB) at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1811>.

PubChem tracks cross-references specified in a BioAssay submission, such as links to corresponding data in PubMed, Taxonomy, Gene, OMIM, or 3D structure of the target. In addition, the BioAssay data model distinguishes primary PubMed citations (references that contain experimental information directly relevant to the BioAssay record and can therefore aid the users’ interpretation and utilization of the assay data) from other PubMed citations that refer to or discuss the assay in a more general way.

In addition to providing data fields that capture essential information describing a BioAssay record, the BioAssay data model provides a flexible “categorized comment” mechanism that allows depositors to provide additional types of descriptive information,

which are not explicitly listed as allowable tags in the data specifications document (http://pubchem.ncbi.nlm.nih.gov/upload/html/tags_assay.html). For example, this mechanism allows depositors to provide information pertinent to a focused research area, to comply with recommendations on a data standard from a working group, or to meet the guidelines of data exchange and sharing as required by a research community. Such a semi-structured data model also allows PubChem to accommodate a greater diversity of information critical to multiple research communities. An example, a BioAssay containing categorized comments (A CPE Based HTS Assay for Antiviral Drug Screening Against Dengue Virus) can be accessed at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=540333>.

Tracking BioAssay Update

All data fields in a BioAssay record can be updated except the data source and the RegID (the tracking identifier provided by the depositor for the assay). An update to textual description or annotation triggers a version change for the assay description section (referred as description version). An update to a substance result is tracked by increasing the “test result” version. Duplicate tests and revisions to an existing test are both considered as test result updates. An update that involves a change to a test result definition (such as data type) or the addition or removal of test result fields triggers a major version change for the BioAssay record. For the major update, all BioAssay test results must be restated by the data depositor upon such fundamental changes. The BioAssay accession number, i.e., AID, remains unchanged upon these three types of updates. Description revision number and test result version number are associated with and counted against the major version of a BioAssay record. Whenever a major version is incremented, the description revision and test result version are reset to “1.” Only the current version of a description and corresponding test results are shown in the PubChem display system and indexed in the Entrez system by default, although all revisions are archived, tracked, and retrievable.

PubChem BioAssay Data Specification

The hierarchical data in the PubChem BioAssay archive is encoded in the data structure ASN.1 notation. All information about a single assay can be contained in a single ASN.1 or equivalent XML data object. It provides separate tagged fields for each aspect of the assay as detailed in the available specification in ASN.1 and XML Schema formats, respectively:

<ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.asn>

<ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem.xsd>

Assay Neighboring and Related BioAssays

BioAssay records in PubChem can be related to each other in multiple ways. Several types of BioAssay relationships are computed in an automated way by PubChem. “Related BioAssays by Activity Overlap” tracks Bioassay records that share one or more active compounds. It allows one to rapidly identify, and thereby avoid, promiscuous inhibitors, or to help discover more complex target-based relationships. “Related BioAssays by Protein Target Similarity” tracks BioAssay records that have the same protein targets, or related protein targets (based on sequence similarity). It allows one to group compounds tested against the same or related targets; to isolate chemical agents with distinct biological effects, such as agonists and antagonists; or to evaluate selectivity of tested compounds. “Related BioAssays by BioSystems via Protein (or Gene) Target” tracks BioAssay records targeting on common biological pathways. This relationship identifies associations between genomic scanning studies that employed RNAi, and small molecule screening discovery studies that employed gene and protein targets. It allows one to take the responsible genes identified in RNAi knockdown experiments and identify small molecule therapeutics suggested in the small molecule screening tests. “Related BioAssays by Same Publication” links together BioAssay records that are extracted from the same publication, hence allows one to relate the results for better interpretation as illustrated in the publication.

Independent of computed BioAssay neighboring, “Related BioAssays” may be specified by the assay depositor. Normally, these relationships are specified when further confirmatory or counter-screenings are performed, thus providing the means to gather all screening data produced by the same screening campaign or assay project. Typically, a “Summary” assay is defined within such a grouping; it provides an overview of how each assay is involved in the overall effort, recaps the findings, and links to the individual assays as cross-references. To better support decision making, PubChem now also clusters and links up BioAssay records submitted for the same assay projects based on such “pair-wise” cross-references. Additional BioAssay relationship may be derived in future time, such as based on disease-target associations.

Public Access, Search, and FTP site

Data from the PubChem BioAssay database can be accessed via Web tools, direct Entrez queries, the FTP site, BLAST service, as well as from other NCBI databases that have links to PubChem (for example, a PubMed record about a medicine may contain a link to the corresponding PubChem record for that medicine).

A BioAssay record can be accessed by accession (AID) through the BioAssay Summary service at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID>, where “myAID” is a valid numeric PubChem BioAssay accession (AID). This service provides access to, and allows one to download, all deposited assay information, such as assay description, protocol and assay data. (As example, AID 1284 (“Dose response biochemical screening assay for inhibitors of c-Jun N-Terminal Kinase 3 (JNK3)) can be accessed at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1284>)

```
|-- ASN  
| |-- 0000001_0001000.zip ...  
|-- CSV  
| |-- Data  
| | |-- 0000001_0001000.zip ...  
| '-- XML  
| |-- 0000001_0001000.zip ...  
|-- XML  
| |-- 0000001_0001000.zip ...  
|-- Concise  
| |-- XML  
| |-- CSV  
| |-- ASN  
|-- README
```

Figure 1. PubChem BioAssay FTP directory structure.

pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=1284.) The service also lists information about the assay target, including depositor-provided molecular information and annotations derived by PubChem about protein family classification, the corresponding gene, pathway, and homologous 3D structures. Furthermore, the BioAssay Summary service provides a central entry point to a set of data analysis tools for the bioactive compounds identified in the assay. These analysis tools can be accessed through the “BioActivity Summary,” “Structure-Activity Analysis,” and “Structure Clustering” links that appear in the “BioActive Compounds” section of the assay record. They allow one to cluster the scaffolds of the tested compounds, examine, and visualize SAR relationships, and evaluate target specificity or promiscuity properties of the tested compounds. In addition, the “Related BioAssays” section lists assays that may be related to the one under review, and links to further detailed summaries of the BioAssay relationship. Cross-references to other NCBI databases, such as PubMed, are listed under the “Links” section.

The BioAssay database is indexed in Entrez and can be directly queried by entering text into the search box found on the BioAssay home page (<http://www.ncbi.nlm.nih.gov/pcassay/>), on the PubChem home page (<http://pubchem.ncbi.nlm.nih.gov>), or at the top of many PubChem Web pages. Descriptive information content in the BioAssay database is indexed under multiple fields to facilitate general as well as specific searches for BioAssay records. A full list of indexed fields and filters are documented at the PubChem Help page (http://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_Index). For assistance with the construction of complex text queries or performing a specific search, one may use the “Limits” and “Advanced” search pages at <http://www.ncbi.nlm.nih.gov/pcassay/limits> and <http://www.ncbi.nlm.nih.gov/pcassay/advanced> respectively. Search results in Entrez are presented in tabular format where each row provides a result summary including assay title, data source, cross references, and links to the corresponding BioAssay record and assay data pages.

The BioAssay database is cross-linked to a number of other databases in Entrez, such as the PubChem Substance and Compound databases, PubMed, Entrez Protein, and Entrez Gene, and more. This makes it possible to access BioAssay even if you start your search in another database, by following the links from the record(s) you retrieve to the associated BioAssays data. In addition, the NCBI BLAST service allows one to search sequences of BioAssay targets. If any of the proteins listed on a BLAST results page were used as targets

of BioAssays, they are flagged on the BLAST results page and linked to the BioAssay records. Integrating molecular sequence information of BioAssay targets with the BLAST service provides an additional path for biologists to discover and utilize the screening results within PubChem.

PubChem BioAssay FTP (<ftp://ftp.ncbi.nlm.nih.gov/pubchem/BioAssay>) provides open access to deposited BioAssay records. PubChem updates the BioAssay FTP site with new and modified BioAssay records on a daily basis in an incremental way. One can check the time stamp for the new post or update, and check the nature and history of an update by referring to the “assay.ftpdump.history” file at the FTP site. In addition to depositor-provided BioAssay records, annotations derived by PubChem from automated computation of BioAssay relationships can also be downloaded at <ftp://ftp.ncbi.nlm.nih.gov/pubchem/BioAssay/AssayNeighbors/>.

PubChem allows one to download BioAssay records in ASN, XML, and “comma-separated values” (CSV) formats. The structure of the FTP site is organized according to the respective data formats as shown in Figure 1, e.g., ASN and XML sub-directories provide BioAssay records containing both assay description and data in ASN.1 and XML format, respectively. The CSV sub-directory provides CSV-formatted assay data and XML-formatted assay description. The “Concise” directory contains the XML/ASN/CSV sub-directories with the same structure, but provides only summary assay results including bioactivity outcome, score, and active concentration. Because of the large number of BioAssay records, bulk downloads from the FTP site are now assisted by the “zip” compression of multiple records per file with BioAssay AID ranges in the filenames, such as “0000001_0001000.zip.”

Miscellaneous information, such as related BioAssay data, protein and gene target identifier lists, and a file containing a list of records (list of AIDs) that have been updated, etc., are also provided under various sub-directories at <ftp://ftp.ncbi.nlm.nih.gov/pubchem/BioAssay>.

BioAssay Tools

To make the vast bioactivity information easily accessible to the scientific community, PubChem provides a suite of integrated services enabling users to collect, compare and analyze biological test results, identify and validate drug targets, and evaluate chemical and RNAi probes.

BioAssay records can be accessed and downloaded at <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID>, where “myAID” is a valid numeric PubChem BioAssay accession (AID). Plotting functions are provided for drawing dose-response curve and readout analysis through histogram and scatterplot. PubChem offers additional services for users to access and download summarized bioactivity information for a compound (<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?sid=mySID>, <http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myCID>), for a protein assay target (<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myGI>), and equivalently, for a gene

(<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myGeneID>). Assay descriptions and data tables can also be retrieved and downloaded through programmatic interfaces using the PubChem Power User Gateway (PUG, <http://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html>), including PUG/SOAP (http://pubchem.ncbi.nlm.nih.gov//pug_soap/pug_soap_help.html) and PUG/REST (http://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST.html) facilities.

In addition, PubChem utilizes the summary results (e.g., active vs inactive) and specialized readouts (i.e., IC₅₀) and provides Web-based tools for: 1) rapid data retrieval, analysis, integration, and comparison of biological screening results across multiple BioAssay records; 2) exploratory structure-activity analysis; 3) target selectivity examination; 4) reviewing related BioAssay records. These tools integrate chemical, target, literature, and biological activity information. They also support the navigation and in-depth data analysis that facilitates identification of bioactive compounds, study of biological profiling, and polypharmacology properties for drug or drug candidate molecules, and discovery of biological interesting targets contained within PubChem databases. A list of Web-based bioactivity analysis tools and their URLs are summarized in Table 1, which can also be accessed from the PubChem BioAssay home page at <http://pubchem.ncbi.nlm.nih.gov/assay>. The uses of these tools are described in detail in several publications (1-5).

Table 1. A list of Web-based PubChem services for the BioAssay resource

Service	Description	URL example
BioActivity Analysis Services	Home page for bioactivity data analysis services	http://pubchem.ncbi.nlm.nih.gov/assay/
BioAssay Summary	BioAssay summary page for a given AID	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID
BioAssay Data Table (concise view)	Concise data table for a given AID. The table includes SID,CID, structure, bioactivity outcome, score, and active concentration value if available	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=datatable http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=r&aid=myAID
BioAssay Data Table (complete view)	Complete data table for given AID, including all deposited test results	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=datatable http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=r&resultsummary=detail&aid=myAID
BioAssay Test Results Selection	Select/search BioAssay test results	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?q=t&aid=myAID
BioAssay Search	Search BioAssay Database with Entrez	http://www.ncbi.nlm.nih.gov/pcassay/

Table 1. continues on next page...

Table 1. continued from previous page.

Service	Description	URL example
BioAssay Search, Limits page	An interface for constructing an Entrez query	http://www.ncbi.nlm.nih.gov/pcassay/limits
BioAssay Search, Advanced Page	An interface for reviewing search history and combining search results	http://www.ncbi.nlm.nih.gov/pcassay/advanced
PubChem Upload System	Chemical structure and BioAssay submission tool	http://pubchem.ncbi.nlm.nih.gov/upload/#welcome
BioActivity Summary - Assay-centric	BioActivity Summary presented from the assay point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary - Compound-centric	BioActivity Summary presented from the compound point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary - Target-centric	BioActivity Summary presented from the target point of view	http://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi
BioActivity Summary	BioActivity information for a single SID,CID, GI, GeneID	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?sid=mySID http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?cid=myCID http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?cid=myGI http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?cid=myGeneID
Structure-Activity Analysis (SAR)	Structure-Activity relationship analysis and visualization in a heatmap-style display.	http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=heat
Related BioAssays	Related BioAssays by: Activity Overlap, Target Similarity, Deposited Annotation, Same Publication, or Common BioSystems.	http://pubchem.ncbi.nlm.nih.gov/assay/assayHeatmap.cgi?service=assayneighbor&aid=myAID
Scatter Plot/Histogram	BioAssay test results plotting functions	http://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype=2
Dose-response curve	Draw dose-response curves for confirmatory assays containing dose-response data points	http://pubchem.ncbi.nlm.nih.gov/assay/assayHeatmap.cgi?service=assaygraph&aid=523&sid=16952359http://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype=1

Table 1. continues on next page...

Table 1. continued from previous page.

Service	Description	URL example
BioAssay download	Assay data download service	http://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=myAID ftp://ftp.ncbi.nlm.nih.gov/pubchem/BioAssay/

BioAssay Submissions and Updates

PubChem Upload (<http://pubchem.ncbi.nlm.nih.gov/upload/>) supports the submission of chemical structures, BioAssay experimental results, annotations, drug targets, siRNAs, and more. The system provides an extensive set of wizards, inline help tips and guided tutorials to assist the submitter, based on their preference, to enter data and descriptive information by Web form or by file. PubChem Upload integrates convenient spreadsheet formats (CSV, Excel & OpenOffice) as well as XML-based data specifications to accommodate submitters of individual assays as well as institutional providers of data from large scale screening studies. A “Preview” facility displays incoming data in a mock record format to show how it will appear in PubChem before being released by the submitter. Such visual feedback to the submitter along with an automated suite of validation checks help insure data integrity and that everything appears as expected. Help documents and a tutorial provide an overview the PubChem Upload system and how it can be used:

The brief help document provides basic information about the PubChem Upload tool, including sample files for submitting substances and assays: http://pubchem.ncbi.nlm.nih.gov/upload/docs/upload_help.html

The complete help document includes the information provided in the brief document, plus technical details about the PubChem Upload tool and FTP submissions: http://pubchem.ncbi.nlm.nih.gov/upload/docs/upload_help_complete.html

The tutorial provides step-by-step examples of the procedure for submitting substances and/or bioassays to PubChem: <http://pubchem.ncbi.nlm.nih.gov/upload/tutorial/>

Summary

The PubChem BioAssay database is set up to serve as a public repository for bioactivity data of small molecules and RNAi. A streamlined information platform is provided at PubChem with a suite of tools allowing users to query the databases and analyze the retrieved BioAssay data. Integration with the Entrez system provides annotation services by linking small molecule modulators or effective RNAi reagents, as identified by screening experiments in the BioAssay database, to genomic and literature resources at NCBI. To meet the increasing demand from public users and from rapid growth of data volume and complexity, PubChem maintains and develops its service to the community as a public data repository by optimizing and expanding its BioAssay data model for

supporting broader types of information, by developing infrastructure to ensure database scalability, by improving the deposition system to ease information exchange, and by enhancing search, retrieval, analysis, and download tools. PubChem welcomes the community to utilize the resource, provide feedback, and to further contribute data content to the repository.

References

1. Wang Y, Suzek T, Zhang J, Wang J, He S, Cheng T, Shoemaker BA, Gindulyte A, Bryant SH. PubChem BioAssay: 2014 update. Nucleic Acids Res. 2014 Jan; 42(Database issue):D1075–82. doi: [10.1093/nar/gkt978](https://doi.org/10.1093/nar/gkt978). doiEpub 2013 Nov 5. PubMed PMID: 24198245.
2. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, et al. PubChem's BioAssay Database. Nucleic Acids Res. 2012;40(Database issue):D400-12. Epub 2011/12/06. doi: [10.1093/nar/gkr1132](https://doi.org/10.1093/nar/gkr1132). gkr1132 [pii].
2. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, et al. An overview of the PubChem BioAssay resource. Nucleic Acids Res. 2010;38(Database issue):D255-66. Epub 2009/11/26. doi: gkp965 [pii] doi: [10.1093/nar/gkp965](https://doi.org/10.1093/nar/gkp965).
3. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 2009;37(Web Server issue):W623-33.
4. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: Integrated Platform of Small Molecules and Biological Activities. Annu Rep Comput Chem. 2008;4(Chapter 12): 217-41.

Tools

The BLAST Sequence Analysis Tool

Thomas Madden, PhD^{✉1}

Created: March 15, 2013.

Scope

A sequence similarity search often provides the first information about a new DNA or protein sequence. A search allows scientists to infer the function of a sequence from similar sequences. There are many ways of performing a sequence similarity search, but probably the most popular method is the “Basic Local Alignment Search Tool” (BLAST) (1, 2). BLAST uses heuristics to produce results quickly. It also calculates an “expect value” that estimates how many matches would have occurred at a given score by chance, which can aid a user in judging how much confidence to have in an alignment.

As the name implies, BLAST performs “local” alignments. Most proteins are modular in nature, with one or more functional domains occurring within a protein. The same domains may also occur in proteins from different species. The BLAST algorithm is tuned to find these domains or shorter stretches of sequence similarity. The local alignment approach also means that an mRNA can be aligned with a piece of genomic DNA, as is frequently required in genome assembly and analysis. If instead BLAST started out by attempting to align two sequences over their entire lengths (known as a global alignment), fewer similarities would be detected, especially with respect to domains and motifs.

There are many different flavors of BLAST searches:

- The megaBLAST nucleotide-nucleotide search, optimized for very similar sequences (in the same or in closely related species), first looks for an exact match of 28 bases, and then attempts to extend that initial match into a full alignment (3, 4).
- The BLASTN nucleotide-nucleotide search looks for more distant sequences.
- BLASTP performs protein-protein sequence comparison, and its algorithm is the basis of many other types of BLAST searches such as BLASTX and TBLASTN.
- BLASTX searches a nucleotide query against a protein database, translating the query on the fly.
- TBLASTN searches a protein query against a nucleotide database, translating the database on the fly.
- PSI-BLAST first performs a BLASTP search to collect information that it then uses to produce a Position-Specific-Scoring-Matrix (PSSM). A PSSM for a query of length N is an N x 20 matrix. Each of the N columns corresponds to a letter in the

¹ NCBI; Email: madden@ncbi.nlm.nih.gov.

[✉] Corresponding author.

query, and each column contains 20 rows. Each row corresponds to a specific residue and describes the probability of related sequences having that residue at that position. PSI-BLAST can then search a database of protein sequences with this PSSM.

- RPSBLAST (Reverse-Position-Specific BLAST) can very quickly search a protein query against a database of PSSMs that were usually produced by PSI-BLAST.
- DELTA-BLAST produces a PSSM with a fast RPSBLAST search of the query, followed by searching this PSSM against a database of protein sequences (5).

A brief summary of “how BLAST works” will assist the reader in understanding the rest of this chapter. BLAST uses heuristics, which really means that it takes shortcuts to get to the proper answer faster. It is useful to break the BLAST search down into a few different phases called setup, preliminary search, and traceback. In the setup phase, BLAST reads in the query, search parameters, and database. It may first check the query for low-complexity or other repeats, and then produces a set of “words”—short, fixed-length sequences based on the query. They are used to initiate matches in the database (or “subject”) sequences. In the preliminary search, a number of steps are performed on every sequence in the database. First, the database is scanned for matches to the words, and those are used to initiate a gap-free extension. Second, gap-free extensions that achieve a certain score are used to seed a gapped extension that only calculates the score and extent and leaves to a later stage the time- and memory-consuming work of calculating insertions and deletions. Gapped extensions that achieve a specified score are saved, though lower-scoring matches may be deleted if too many matches are found. In the final traceback phase of the search, gapped extensions saved in the preliminary phase are used as seeds for a gapped extension that also calculates the insertions and deletions and may use more sensitive parameters. More details on the BLAST algorithm are provided in (1, 2).

BLAST provides a variety of ways to perform a search:

- NCBI BLAST website at <http://blast.ncbi.nlm.nih.gov> (6-8). The website requires no setup or registration, is simple to use, produces results quickly, and requires only a web browser. The BLAST website is a shared public resource, so users who send in many hundreds or thousands of searches in a short time may run afoul of [usage guidelines](#).
- BLAST URL API. The [URL API documentation](#) describes the URL parameters of the website that will not change and can be used in the future. This API also uses a shared public resource, so users should respect [usage guidelines](#).
- BLAST standalone applications. Users with specialized or proprietary data, or who require resources beyond what NCBI can provide, can use the BLAST+ applications to run BLAST in “standalone mode,” (9) on their own computers. BLAST+ in standalone mode uses data on local servers, using the user’s own data and/or using BLAST databases downloaded from NCBI. To use standalone mode, users must have sufficient computational resources for their searches. BLAST databases can be large (many gigabytes), and setting up a standalone BLAST+ environment requires

some effort. On the other hand, standalone BLAST+ may be the best option for sophisticated users. BLAST+ applications run on Mac OSX, Windows, and most flavors of LINUX/UNIX. Instructions on the use of BLAST+ can be found at <https://www.ncbi.nlm.nih.gov/books/NBK279690/>

- BLAST+ remote service. Users who need to do many searches at the NCBI, or who want to script searches, can use the remote service available with the BLAST+ applications. Instructions on the use of the remote service with the BLAST+ applications can be found at <https://www.ncbi.nlm.nih.gov/books/NBK279690/> The remote service also uses a shared public resource, so users should respect [usage guidelines](#).
- C++ Application Programming Interface (API). For very specialized applications, the NCBI C++ Toolkit offers a programmatic interface to BLAST described at https://ncbi.github.io/cxx-toolkit/pages/ch_blast The API supports both standalone and remote searches (at the NCBI).

This chapter focuses on the specifics of how BLAST works, most especially at the NCBI, and how to avoid using it incorrectly. There are links to documentation and videos about using BLAST at http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs.

History

NCBI produced the first version of BLAST around 1990, accompanied by the publication of the first BLAST paper by Altschul et al. (2). This version performed only gap-free alignments, but provided p-values that allowed users to judge the statistical significance of their result. A gapped version of BLAST appeared in 1997 (1). This release also included PSI-BLAST, which could produce a PSSM and then search the database with it. Both of these BLAST versions made use of the NCBI C toolkit. In late 2009, the NCBI started supporting a newer version of BLAST (called BLAST+) (9) based upon the NCBI C++ toolkit as a development platform. Afterwards, the C toolkit and the older BLAST packages were deprecated and users were strongly encouraged to use the BLAST+ applications for standalone needs. The NCBI BLAST website was built with the C++ toolkit and BLAST+.

Data Model

Structured Output: Flexible Results

Most users who are familiar with the BLAST report think of it as the output from BLAST, but the real picture is somewhat more complicated. A BLAST search first produces results as structured output, which permits automatic and rigorous checks for syntax errors and changes. Typical report formats such as the BLAST or GenBank report do not permit such automatic checks. BLAST output can be represented as XML or ASN.1 (Abstract Syntax Notation 1), enabling automated syntax and structure validation. For example, the structure of an XML document can be ensured by validating it against its DTD or

Schema. ASN.1, used extensively at the NCBI since 1990, is also constrained by a module definition (similar to a schema), and its binary format is very compact, making it ideal for transmission over networks.

BLAST search results are first represented as C++ objects, which can be used directly to output data formatted for further processing. Since many of the objects have ASN.1 representations, they can also be serialized to ASN.1 and written to disk or sent over a network. These serialized objects can then be used to recreate the original C++ objects at a different time or place.

There are a few advantages to this procedure. First, results that are serialized to disk can be formatted later. Second, results that are sent over a network can be formatted on another machine. Third, serialized results can be formatted multiple times in different ways using the same ASN.1 objects. Finally, since many NCBI tools produce objects corresponding to the ASN.1 modules, they have a straightforward way to exchange results.

While ASN.1 makes storage and transmission of BLAST results efficient, the XML representations of these objects provides a bridge to common XML tool chains. The choice of which to use depends on the application.

The Alignment: Data You Really Need

BLAST outputs its alignments using the [ASN.1 SeqAlign module](#). A SeqAlign indicates where an alignment starts and ends on sequences, provides the coordinates of insertions and deletions, and (for DNA) tells the strand to which the query sequence aligned. It also lists scores, expect values, and sequence identifiers for the alignment. A SeqAlign does not contain the actual query or database sequences, or other information such as the titles for any sequences, but only the sequence identifiers. In addition to the alignment data, BLAST reports require other information about the database sequence, such as the actual sequence, title, and taxonomy information. Since those data don't occur in the SeqAlign, BLAST report formatters retrieve sequence data as needed, using the sequence identifier, from the BLAST database or from Entrez.

The Rest: Data You Might Need

In addition to alignment data, BLAST also produces other outputs that characterize the search inputs and results:

The [Blast4-request](#) ASN.1 type contains the query, the search parameters, and the database. It is often referred to as a “search strategy”. Both the website and the standalone applications can produce the search strategy as an ASN.1 object, and the website can store it as a “Saved search”. The user may use the search strategy to repeat the search at a later time, with optional changes to the query or database.

The [Blast4-archive](#) ASN.1 type contains the query and search information as a search strategy, the alignment information as a SeqAlign, the location of any masks applied to

the query (for low-complexity or interspersed repeats), and the Karlin-Altschul parameters used to calculate statistics.

The search strategy and Blast4-archive types depend on the ASN.1 module defined for the BLAST+ remote service, which is the network interface of the NCBI BLAST queuing system Splitd (described below).

Dataflow

Searches at the NCBI: How It's Done

The NCBI uses a custom queuing system called Splitd to schedule searches that were submitted via the website or BLAST+ remote service (Figure 1). The Splitd system first parses the input from a user and produces an ASN.1 object with the relevant information that is stored in an MSSQL database. Some searches require a preparatory step such as retrieval of data from Entrez. In that case, a daemon requests the data from Entrez after verifying that it has not been cached by a recent search. Another example with a preparatory step is the RPSBLAST search used to produce a PSSM by DELTA-BLAST.

Once any preparatory steps have completed, the Splitd server queues the search. Queuing priority depends partly on whether the user has other searches queued: the search may be penalized (i.e., put back further in the queue) if the user has many uncompleted searches. Splitd spreads each search over multiple machines (each of which is running a backend), where each backend searches only a part of the database. The partial result produced by a backend is called a chunk. The backend performs only the setup and preliminary search (as described above), so its final product is a set of matches listing scores, extents, and database sequences identifiers for those matches. Results from the backends are forwarded to a merger, which collects all chunks in memory, merges them into a single result, and writes the complete set of preliminary results to an MSSQL database. These results are then sent to the traceback daemon, which performs the final step of producing the alignment that includes insertions and deletions, final scores, and extents. Finally, the traceback is written to the MSSQL database and the search is marked as finished. The user can then format the search in a variety of formats as shown in Figure 2.

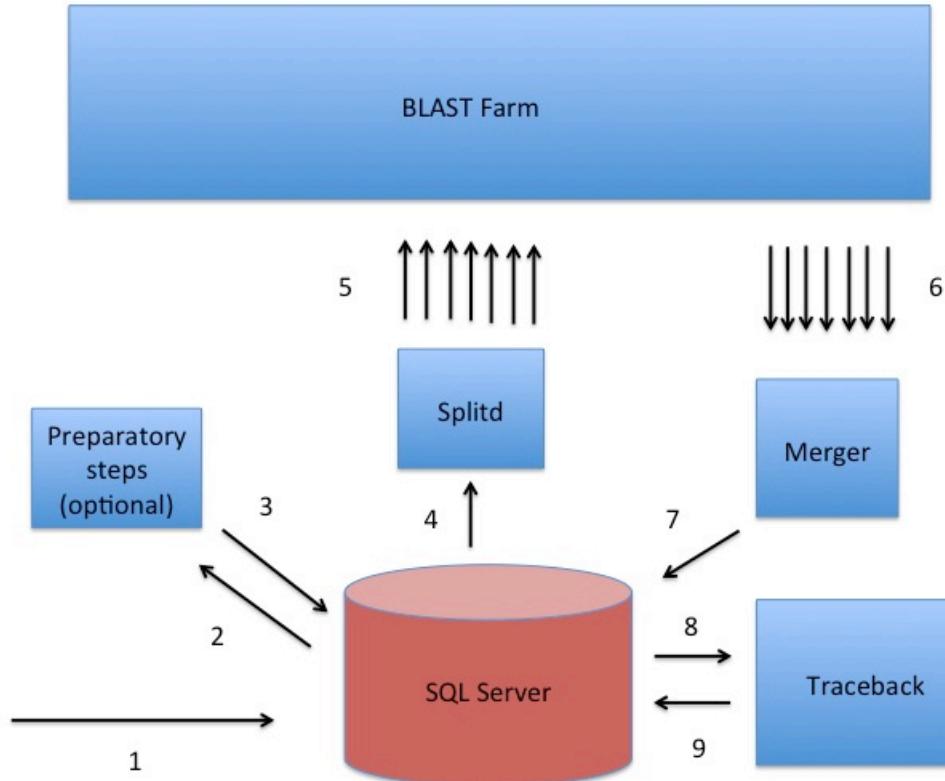
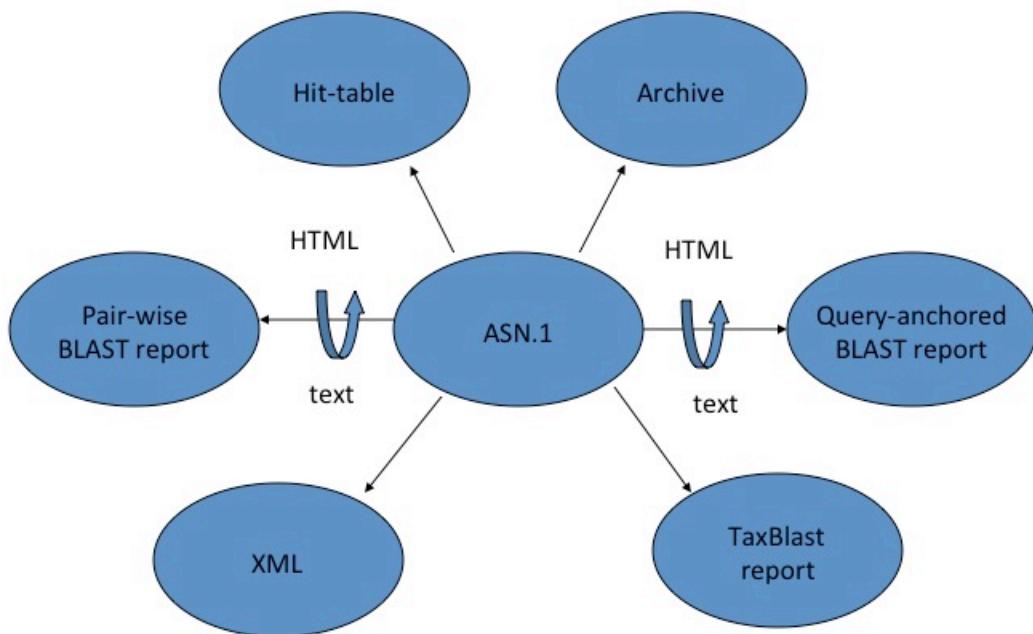


Figure 1: A sketch of the Splitd system used at the NCBI for processing BLAST requests. The numbers in the figure identify steps in the process. First, the search is inserted into the SQL database as ASN.1 [1]. Second, an optional preparatory step may retrieve data from Entrez or produce a PSSM using RPSBLAST [2, 3]. Third, Splitd pulls the search from the SQL database [4], queues it, and spreads it across several backends [5]. The search is processed on the BLAST farm. Results from backends are preliminary, because they include only the extent of alignments as well as the scores, but no insertions or deletions. Next, these results are sent to the Merger [6]. The Merger merges all pieces into one result which it stores in the SQL server [7]. The traceback then pulls the preliminary results from the SQL database and produces results that include insertions and deletions [8]. Finally, the traceback places the full results into the SQL database [9]. At this point the user may retrieve the results using the Request ID that the system issued.



2

Figure 2: Different output formats that can be generated from the Splitd ASN.1. Most reports require additional information not stored in the ASN.1, such as database sequences or taxonomy information retrieved from other sources.

BLAST Databases: Some Details to Keep in Mind

Generally, BLAST does not directly search GenBank flatfiles. Rather, sequences are transformed into BLAST databases with a special format that makes searching more efficient. The BLAST indexing processes splits and indexes the sequence records, producing several files. The “header” and the “sequence” files are the most important ones. The header file contains information such as the sequence title and taxonomy information and is used mostly during formatting of the BLAST report. The sequence file contains the sequence information and is used most heavily during the BLAST search. DNA has a small alphabet (four letters, if there are no ambiguities) so the DNA sequence file consumes a little more than one byte per four bases. A BLAST database is normally partitioned into multiple volumes, with each volume representing a contiguous subset of the database. The size of the volume can be specified when the database is created, but the NCBI has found that a volume size of about one gigabyte works well. For the Sequence Read Archive (SRA), BLAST searches the underlying SRA objects directly. This is efficient because the SRA objects group the data in a manner similar to that of BLAST.

As mentioned above, the results produced by BLAST (e.g., the SeqAlign) do not contain the database sequences, but only identifiers for them. BLAST must use the sequence identifiers to retrieve the sequence data from some other source, such as the BLAST database or Entrez. This means that an identifier must uniquely identify a sequence in the database. Furthermore, the query sequence should not have the same identifier as any sequence in the database, unless the query sequence itself is in the database. Any BLAST database or FASTA file from the NCBI website that contains GI numbers already satisfies the uniqueness criterion. Ambiguous identifiers are normally a problem only when custom databases are produced and care is not taken in assigning identifiers. The identifier for a FASTA entry is the first token (meaning the letters up to the first space) after the > sign on the definition line. The simplest case is to simply provide a unique token (e.g., 1, 2, and so on), but it is possible to construct more complicated identifiers that might, for example, describe the data source. NCBI supports a specific syntax for such parsable identifiers described at https://ncbi.github.io/cxx-toolkit/pages/ch_demo#ch_demo.T5.

The makeblastdb application produces BLAST databases from FASTA files, and its –parse_seqids flag instructs makeblastdb to expect unique parsable identifiers. If the identifiers are not parsed, then makeblastdb adds some ad hoc (internal) identifiers to the BLAST database, but this may limit the options for display or further processing of the alignments in the result. The BLAST+ user manual at <https://www.ncbi.nlm.nih.gov/books/NBK279690/> presents detailed instructions about how to use makeblastdb.

Many users are familiar with the process of producing the BLAST database from FASTA files, but this is not the process used to create most BLAST databases available at the NCBI. Rather, most BLAST databases on NCBI servers are produced directly from the central NCBI ID system (described in the chapter on Dataflow). Nucleotide databases at the NCBI can contain tens of billions of bases, so de novo indexing can consume significant resources and time. Therefore, most BLAST databases are updated incrementally, rather than being completely rewritten every time sequence data changes. The ID system also adds other available information (such as taxonomy) to the BLAST database. The NCBI makes many of these databases available as FASTA in the BLAST portion of the FTP site. These FASTA files are produced from the original BLAST databases.

BLAST Reports: Look Before You Parse

BLAST can produce a number of different reports, but it is important to understand the purpose of each report to use it effectively. The standard BLAST report consists of several sections, including a table of descriptions (sequence titles) and alignments. It is meant as a human-readable report, and is subject to change with little or no notice. The NCBI strongly discourages parsing this report, and provides other formats that are better suited for automated processes. One of the simplest formats is the tabular output, which provides basic information in an easy-to-parse form. As stated earlier, structured output allows for automatic and rigorous checks for syntax errors and changes. The standard

BLAST report and the tabular report are not structured output, so they do not permit automated checks for syntax errors and changes. Only a structured report such as XML or ASN.1 can allow for such automated checks. The NCBI makes available a special BLAST XML report that contains much of the same data (e.g., the sequences) as the standard BLAST report, but it allows formal checks for correctness. BLAST can also produce ASN.1 output.

Access

The NCBI supports a number of methods to submit BLAST searches to the Splitd system.

The most heavily used way to submit searches is via the NCBI website at blast.ncbi.nlm.nih.gov. Figure 3 presents the submission page for nucleotide-nucleotide searches. The user simply inputs a sequence identifier or raw sequence and clicks the BLAST button. If desired, they may change the database or limit their search taxonomically using autocomplete menus (Figure 4). There are a number of other ways to modify the search, but most users make minimal changes to the defaults. At this point, the Blast.cgi script parses the input, constructs an ASN.1 representation of the search, inserts the search into an MSSQL database, and returns a Request ID (RID) to the user. The Splitd system then processes the request as shown in Figure 1. Meanwhile, the user's browser periodically polls the server, checking for complete results. Once the results are complete, the page displays the report, which the user may reformat as desired. Results are usually saved for 36 hours on the server. The user may use the RID to retrieve the results.

BLAST® Basic Local Alignment Search Tool

Home Recent Results Saved Strategies Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

From
To

Or, upload file no file selected

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Human genomic + transcript Mouse genomic + transcript Others (nr etc.): Nucleotide collection (nr/nt)

Organism Optional: Enter organism name or id—completions will be suggested Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional: Enter an Entrez query to limit search

Program Selection

Optimize for: Highly similar sequences (megablast) More dissimilar sequences (discontiguous megablast) Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences) Show results in a new window

Figure 3: Nucleotide-nucleotide search page.

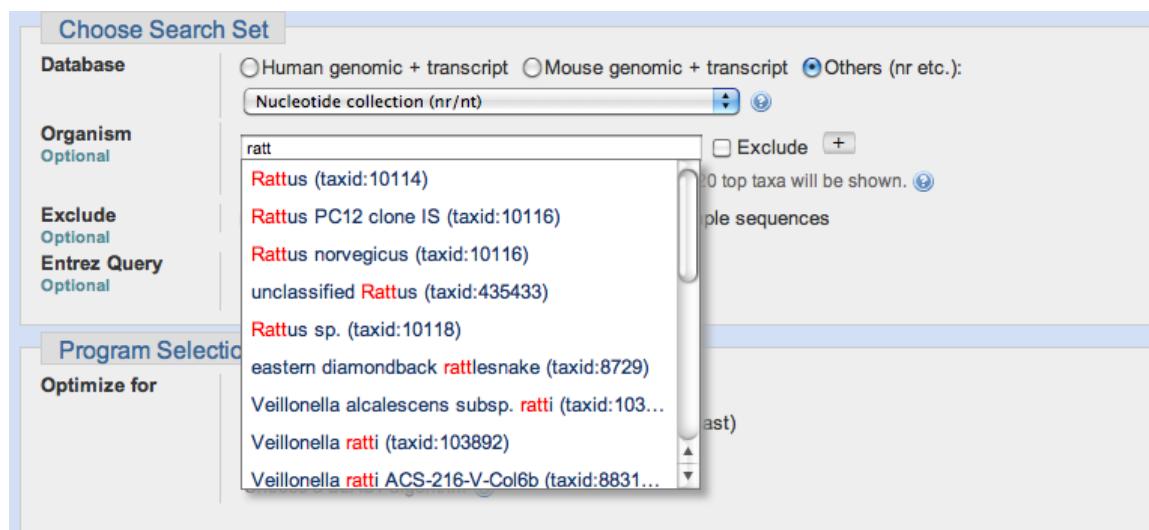


Figure 4: Detail from the nucleotide-nucleotide search page. The autocomplete menu presents suggestions as the user types. Once a user has selected an entry, an Entrez query is constructed and processed as a “preparatory” step by the Splitld system (see Figure 1 and text).

Users may also access the NCBI BLAST service through the BLAST+ remote service, which is a network service that uses ASN.1 to communicate between the client and the server. In this case, the client sends the query, parameters, and database to the server in the form of an ASN.1 request. Splitld processes the request in the manner described previously. An RID is assigned and sent back to the client. The client polls for the status of the result on a regular basis. Once the search is done, the ASN.1 results returned to the client include the alignment (as a SeqAlign) and masking information. Because the SeqAlign does not contain the database sequences, the BLAST+ remote client fetches sequence data from the NCBI as needed for formatting.

Programmers can use the NCBI “URL API” interface and the HTTP protocol to create BLAST jobs and retrieve BLAST results. The [URL API documentation](#) describes the URL parameters of Blast.cgi that will not change and can be used in the future, and explains how to interpret BLAST results programmatically.

Related Tools

There are many tools that run BLAST searches and post-process the output for specific purposes. Three tools supported by the NCBI are:

- Primer-BLAST (10). Primer-BLAST finds primers that would amplify only a specific gene. It first uses Primer3 to identify primers on a gene sequence template, and then uses BLAST to search the template against the specified databases. It extensively post-processes the BLAST output to identify primers that uniquely amplify the desired gene. Primer-blast uses the BLAST+ remote service to send the search to Splitld and to receive results. It makes full use of the ASN.1-encoded objects to post-process BLAST results and create presentations.

- IgBLAST (11). IgBLAST annotates the variable regions of an immunoglobulin sequence, which includes a variable (V), diversity (D), and a joining (J) segment. These different segments have different characteristic lengths and require different BLAST parameters. IgBLAST orchestrates the multiple BLAST searches needed, and then presents a unified report to the user. It calls either the BLAST API directly or uses the BLAST+ remote service.
- VecScreen. The VecScreen service identifies vector contamination in a query. It uses a specialized database and extensively post-processes the initial results produced by BLAST. Information about VecScreen is available at <http://www.ncbi.nlm.nih.gov/tools/vecscren/>

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402. PubMed PMID: 9254694.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403–10. PubMed PMID: 2231712.
3. Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comput Biol.* 2000;7(1-2):203–14. PubMed PMID: 10890397.
4. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. Database indexing for production MegaBLAST searches. *Bioinformatics.* 2008;24(16):1757–64. PubMed PMID: 18567917.
5. Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct.* 2012;7:12. PubMed PMID: 22510480.
6. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. NCBI BLAST: a better web interface. *Nucleic Acids Res.* 2008;36(Web Server issue):W5-9.
7. McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 2004;32(Web Server issue):W20-5.
8. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* 2006;34(Web Server issue):W6-9.
9. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421. PubMed PMID: 20003500.
10. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics.* 2012;13:134. PubMed PMID: 22708584.
11. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* 2013;41(Web Server issue):W34-40.

The Entrez Search and Retrieval System

Jim Ostell¹

Created: October 9, 2002; Updated: January 31, 2014.

Summary

Entrez is the text-based search and retrieval system used at the National Center for Biotechnology Information (NCBI) for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and many others. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information. These general concepts are the focus of this chapter. Other chapters cover the details of a specific Entrez database (e.g., PubMed) or a specific source of data (e.g., GenBank).

Entrez Design Principles

History

The first version of Entrez was distributed by NCBI in 1991 on CD-ROM. At that time, it consisted of nucleotide sequences from GenBank and Protein Data Bank (PDB); protein sequences from translated GenBank, Protein Information Resource (PIR), SWISS-PROT, PDB, and Protein Research Foundation (PRF); and associated citations and abstracts from MEDLINE (now PubMed and referred to as PubMed below). We will use this first design to illustrate the principles behind Entrez.

Entrez Nodes Represent Data

An Entrez "node" is a collection of data that is grouped together and indexed together. It is usually referred to as an Entrez database. In the first version of Entrez, there were three nodes: published articles, nucleotide sequences, and protein sequences. Each node represents specific data objects of the same type, e.g., protein sequences, which are each given a unique ID (UID) within that logical Entrez Protein's node. Records in a node may come from a single source (e.g., all published articles are from PubMed) or many sources (e.g., proteins are from translated GenBank sequences, SWISS-PROT, or PIR).

Note that the UID identifies a single, well-defined object (i.e., a particular protein sequence or PubMed citation). There may be other information about objects in nodes, such as protein names or Enzyme Commission (EC) numbers, that may be used as index terms to find the record, but these pieces of information are not the central organizing principle of the node. Each data object represents a stable, objective observation of data as

¹ NCBI.

much as possible, rather than interpretations of the data, which are subject to change or confusion over time or across disciplines. For example, barring experimental error, a particular mRNA sequence report is not likely to change over the years; however, the given name, position on the chromosome, or function of the protein product may well change as our knowledge develops. Even a published article is a stable observation. The fact that the article was published at a certain time and contained certain words will not change over time, although the importance of the article topic may change many times.

Entrez Nodes Are Intended for Linking

Another criterion for selecting a particular data type to be an Entrez node is to enable linking to other Entrez nodes in a useful and reliable way. For example, given a protein sequence, it is very useful to quickly find the nucleotide sequence that encodes it. Or given a research article, it is useful to find the sequences it describes, if any.

Links between Nodes

One way to achieve this is to put all of the information into one record. For example, many GenBank records contain pertinent article citations. However, PubMed also contains the article abstract and additional index terms (e.g., MeSH terms); furthermore, the bibliographic information is also more carefully curated than the citation in a GenBank entry. It therefore makes much more sense to search for articles in PubMed rather than in GenBank.

When a subset of articles has been retrieved from PubMed, it may be useful to link to sequence information associated with the abstracts. The article citation in the GenBank record can be used to establish the link to PubMed and, conversely, to make the reciprocal link from the PubMed article back to the GenBank record. Treating each Entrez node separately but enabling linking between related data in different nodes means that the retrieval characteristics for each node can be optimized for the characteristics and strengths of that node, whereas related data can be reached in nodes with different strengths.

This approach also means that new connections between data can be made. In the example above, the GenBank record cited the published article, but there was no link from that article in PubMed to the sequence until Entrez made the reciprocal link from PubMed. Now, when searching articles in PubMed, it is possible to find this sequence, although no PubMed records have been changed. Because of this design principle, the Entrez system is richly interconnected, although any particular association may originate from only one record in one node.

Links within Nodes

Another type of linking in Entrez is between records of the same type, often called "neighbors," in sequence and structure nodes. Most often these associations are computed at NCBI. For example, in Entrez Proteins, all of the protein sequences are "BLASTed" against each other, and the highest-scoring hits are stored as indexes within the node. This

means that each protein record has associated with it a list of highly similar sequences, or neighbors.

Again, associations that may not be present in the original records can be made. For example, a well-annotated SWISS-PROT record for a particular protein may have fields that describe other protein or GenBank records from which it was derived. At a later date, a closely related protein may appear in GenBank that will not be referenced by the SWISS-PROT record. However, if a scientist finds an article in PubMed that has a link to the new GenBank record, that person can look at the protein and then use the BLAST-computed neighbors to find the SWISS-PROT record (as well as many others), although neither the SWISS-PROT record nor the new GenBank record refers to each other anywhere.

Entrez Nodes Are Intended for Computation

There are many advantages to establishing new associations by computational methods (as in the GenBank-SWISS-PROT example above), especially for large, rapidly changing data sets such as those in biomedicine.

As computers get faster and cheaper, this type of association can be made more efficiently. As data sets get bigger, the problem remains tractable or may even improve because of better statistics. If a new algorithm or approach is found to be an improvement, it is possible to apply it over the whole data set within a practical timescale and by using a reasonable number of resources. Any associations that require human curation, such as the application of controlled vocabularies, do not scale well with rapidly growing sets of data or evolving data interpretations. Although these manual kinds of approaches certainly add value, computational approaches can often produce good results more objectively and efficiently.

Entrez Is a Discovery System

A data-retrieval system succeeds when you can retrieve the same data you put in. A discovery system is intended to let you find more information than appears in the original data. By making links between selected nodes and making computed associations within the same node, Entrez is designed to infer relationships between different data that may suggest future experiments or assist in interpretation of the available information, although it may come from different sources.

The ability to compare genotype information across a huge range of organisms is a powerful tool for molecular biologists. For example, this technique was used in the discovery of a gene associated with hereditary nonpolyposis colon cancer (HNPCC). The tumor cells from most familial cases of HNPCC had altered, short, repeated DNA sequences, suggesting that DNA replication errors had occurred during tumor development. This information caused a group of investigators to look for human homologs of the well-characterized *Escherichia coli* DNA mismatch repair enzyme, MutS (1). Mutants in a *MutS* homolog in yeast, *MSH2*, showed expansion and contraction of

dinucleotide repeats similar to the mutation found in the human tumor cells. By comparing the protein sequences between the yeast MSH2, the *E. coli* MutS, and a human gene product isolated and cloned from HNPCC colorectal tumor, the researchers could show that the amino acid sequences of all three proteins were very similar. From this, they inferred that the human gene, which they called *hMSH2*, may also play a role in repairing DNA, and that the mutation found in tumors negatively affects this function, leading to tumor development.

The researchers could connect the functional data about the yeast and bacterial genes with the genetic mapping and clinical phenotype information in humans. Entrez is designed to support this kind of process when the underlying data are available electronically. In PubMed, the research paper about the discovery of *hMSH2* (1) has links to the protein sequence, which in turn has links to "neighbors" (related sequences). There are lots of records for this protein and its relatives in many organisms, but among them are the proteins from yeast and *E. coli* that prompted the study. From those records there are links back to the PubMed abstracts of articles that reported these proteins. PubMed also has a "neighbor" function, **Related articles**, that represents other articles that contain words and phrases in common with the current record. Because phrases such as "*Escherichia coli*," "mismatch repair," and "MutS" all occur in the current article, many of the articles most related to this one describe studies on the *E. coli* mismatch repair system. These articles may not be directly linked to any sequence themselves and may not contain the words "human" or "colon cancer" but are relevant to HNPCC nonetheless, because of what the bacterial system may tell us.

Entrez Is Growing

The original three-node Entrez system has evolved over the past two decades to include approximately 40 nodes, many of which are described in detail in this handbook. Each one of these nodes is richly connected to others. Each offers unique information and unique new relationships among its members. The combination of new links and new relationships increases the chances for discovery. The addition of each new node creates different paths through the data that may lead to new connections, without more work on the old nodes.

How Entrez Works

Entrez integrates data from a large number of sources, formats, and databases into a uniform information model and retrieval system. The actual databases from which records are retrieved and on which the Entrez indexes are based have different designs, based on the type of data, and reside on different machines. These will be referred to as the "source databases." A common theme in the implementation of Entrez is that some functions are unique to each source database, whereas others are common to all Entrez databases.

A Division of Labor: Basic Principles

Some of the common routines and formats for every Entrez node include the term lists and posting files (i.e., the retrieval engine) used for Boolean queries, the links within and between nodes, and the summary format used for listing search results in which each record is called a DocSum. Generally, an Entrez query is a Boolean expression that is evaluated by the common Entrez engine and yields a list of unique ID numbers (UIDs), which identify records in an Entrez node. Given one or more UIDs, Entrez can retrieve the DocSum(s) very quickly. The links made within or between Entrez nodes from one or more UIDs is also a function across all Entrez source databases.

The software that tracks the addition of new or updated records or identifies those that should be deleted from Entrez may be unique for each source database. Each database must also have accompanying software to gather index terms, DocSums, and links from the source data and present them to the common Entrez indexer. This can be achieved by generating an XML document in a specific DTD that contains the terms, DocSums, and links. Although the common engine retrieves a DocSum(s) given a UID(s), the retrieval of a full, formatted record is directed to the source database, where software unique to that database is used to format the record correctly. All of this software is written by the NCBI group that runs the database.

This combination of database-specific software and a common set of Entrez routines and applications allows code sharing and common large-retrieval server administration but enables flexibility and simplicity for a wide variety of data sources.

Software

Although the basic principles of Entrez have remained the same for more than two decades, the software implementation has been through at least three major redesigns and many minor ones.

Currently, Entrez is written using the NCBI C++ Toolkit. The indexing fields (which for PubMed, for example, would be Title, Author, Publication Date, Journal, Abstract, and so on) and DocSum fields (which for PubMed are Author, Title, Journal, Publication Date, Volume, and Page Number) for each node are defined in a configuration file; but for performance at runtime, the configuration files are used to automatically generate base classes for each database. These are the basic pieces of information used by Entrez that can also be inherited and used by more database-specific, hand-coded features. The term indexes are based on the Indexed Sequential-Access Method (ISAM) and are in large, shared, memory-mapped files. The postings are large bitmaps, with one bit per document in the node. Depending on how sparsely populated the posting is, the bit array is adaptively compressed on disk using one of four possible schemes. Boolean operations are performed by using AND or OR postings of bit arrays into a result bit array. DocSums are small, fielded data structures stored on the same machines as the postings to support rapid retrieval.

The Web-based Entrez retrieval program is a fast cgi application that uses the NCBI proprietary XML/XSLT-based Web application framework called Portal. The Entrez Web application provides a common set of features for any Entrez database, presenting users with a uniform look and feel across the nodes. It is running on a pool of load-balanced front-end UNIX machines which connect to the back-end servers for query processing and retrieval. One aspect of this Portal framework is a set of XHTML templates that represents an HTML page. These templates allow the combination of static HTML content with dynamic content generated by Portal components written in XML/XSLT at tagged parts of the template. The Web page generated in an Entrez session contains elements from static templates as well as elements generated dynamically from Portal components, that may be common to all Entrez nodes, or unique to one or a few Entrez nodes. Again, this design supports a common core of robust, common functionality maintained by one group, with support for customizations by diverse groups within NCBI.

Boolean query processing, DocSum retrieval, and other common functions are implemented as a set of standalone servers written using NCBI C++ Toolkit and supported on a number of load-balanced "back-end" UNIX machines. Because Entrez can support session context (for example, in the use of query history, My NCBI preferences, Filters, etc.), a "history server" has been implemented on the back-end machines so that if a user is sent to machine "A" by the load balancer for their first query but to machine "B" for the second query, Entrez can quickly locate the user's query history and obtain it from machine "A." Other than that, the back-end machines are completely independent of each other and can be added and removed readily from *front-end Web application* support. Retrieval of full documents comes from a variety of source databases, depending on the node. These might be Sybase or Microsoft SQL Server relational databases of a variety of schemas or text files of various formats. Links are supported using the Sybase IQ database product for data storage and the NCBI proprietary link server called 'MegaLink' for fast retrieval.

References

1. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*. 1993 Dec 3;75(5):1027–38. PubMed PMID: 8252616.

C++ Toolkit

Denis Vakatov¹

Created: February 15, 2013; Updated: January 26, 2018.

Summary

The C++ Toolkit is a large body of C++ software that was built to support the medical literature and bioinformatics services that the National Center for Biotechnology Information (NCBI) makes available to the public. While the primary users of the Toolkit are within NCBI, the software is portable (Unix, Windows, Mac) and freely available with no restrictions on use.

Libraries and Applications

If you are a C++ developer you will find the portable nature of the libraries very useful in building cross-platform applications even if you do not have much interest in bioinformatics. Libraries such as those for CGI/Fast-CGI, networking, XmlWrapp, SQL database access, and serialization are quite general-purpose and can be used in a variety of applications outside the bioinformatics problem domain.

The Toolkit provides many general-purpose libraries, including:

- [CORELIB](#) - Provides a portable way to write C++ code and many useful facilities such as an application framework, argument processing, template utilities, threads, date/time, files, and strings.
- [CONNECT](#) - Networking and inter-process communication with IOSTREAM adaptors.
- [CGI](#) - CGI and Fast-CGI.
- [DBAPI](#) - SQL database access.
- [SERIAL](#) - Serialization using ASN.1, JSON, or XML.
- [GUI](#) - Portable wxWidgets and OpenGL based GUI and graphic libraries.
- [XmlWrapp](#) - XML parsing and handling, XSLT, XPath—this is an NCBI fork that adds some useful enhancements to the open-source [xmlwrapp](#) project.
- [JSONWRAPP](#) – same for JSON format.
- [UTIL](#) - Many generic facilities including compression, a diff API, floating point comparison, random number generation, thread pools, and UTF-8 conversion.

Libraries specific to bioinformatics are also provided, including:

- [ALGORITHM](#) - Sequence alignment algorithms.
- [BLAST](#) - An alignment engine.

¹ NCBI; Email: vakatov@ncbi.nlm.nih.gov.

- [OBJECT MANAGER](#) - Biological sequences (e.g., GenBank) retrieval and processing.

Applications are also provided, for example:

- [DATATOOL](#) - A data converter and C++ code generator for data storage classes.

The C++ Toolkit libraries and applications are in active development and are regularly built and tested on Unix, Windows, and Mac.

Typical Uses

Every day, thousands of people around the world use applications built on top of the C++ Toolkit. These include:

- [PubMed](#) - PubMed comprises more than 22 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher websites.
- [PubMed Central](#) - PMC is a free full-text archive of biomedical and life sciences journal literature at the National Institutes of Health's National Library of Medicine (NIH/NLM).
- [BLAST](#) - The Basic Local Alignment Search Tool is a tool for analyzing biological sequence information and is one of the most widely used bioinformatics programs.
- [Genome Workbench](#) - NCBI Genome Workbench is an integrated application for viewing and analyzing sequence data. With Genome Workbench, you can view data in publically available sequence databases at NCBI, and mix this data with your own private data. The Genome Workbench has an online counterpart, the [Sequence Viewer](#).

Installation

Before using the C++ Toolkit, you should ensure that your platform is supported by checking the [current release notes](#).

The first step in using the Toolkit is getting the source code—either by [downloading via FTP](#) or by [checking it out from Subversion](#).

Once you've gotten the source code, you can [configure and build the Toolkit](#) for your platform.

Because the C++ Toolkit supports a variety of platforms and compilers, the process of building the libraries involves determining the platform- and compiler-specific features as well as third-party packages. This is facilitated by running the platform-independent "configure" script on Unix and Mac, or building the "CONFIGURE" solution in C++ Visual Studio on Windows. Additional details on configuring and building can be found [online](#).

Successful builds result in immediately usable libraries and applications. Thus, downloading, configuring, and building comprise the installation process, and generally there is no need for a separate "install" step.

Public Releases

Typically, new [public versions](#) of the C++ Toolkit are released every year or two. The Release Notes are also updated when a new public version of the C++ Toolkit becomes available.

Documentation

An extensive C++ Toolkit book is available [online](#). There are also online browsers:

- [Doxygen](#)
- [LXR](#)
- [Subversion](#)
- [ASN.1](#)

Support

The software is provided on an as-is basis; however, the following mailing lists can be used:

- to receive [announcements](#) (read-only)
- for [general information](#)
- to contact [core developers](#)
- for help with the [Object Manager](#)
- to access [CVS logs](#) (read-only)
- to access [SVN logs](#) (read-only)

Related Resources

Many resources are available to C++ Toolkit users, and the best way to find them is either by using the sidebar on any page of the online [Toolkit Book](#), or from any of the following dedicated search pages:

- [Toolkit Book Search](#)
- [Source Code Search](#)
- [Combined Book and Source Code Search](#)

These pages all include a search tool and links to several source browsers, Subversion access to the source code, library symbol search tools, ASN.1 specifications, and more.

More generally, the [NCBI home page](#) has a wealth of information about NCBI resources, including the C++ Toolkit and many other tools.

For more information about the Toolkit, please see the online [NCBI C++ Toolkit Book](#). The [Introduction](#) provides a broad overview of the capabilities in the C++ Toolkit with links to other chapters that cover topics in more detail. The [Getting Started](#) chapter provides a description of how to obtain the C++ Toolkit, the layout of the source distribution tree, and how to get started. The online Toolkit book is intended to cover pretty much everything you need to know about the Toolkit. If you can't find answers to your questions there, please try one of the email lists above or email cpp-doc@ncbi.nlm.nih.gov.

LinkOut: Linking to External Resources from NCBI Databases

Y. Kathy Kwan¹

Created: November 14, 2013.

Scope

The power of linking is one of the most important developments that the World Wide Web offers to the scientific and research community. By providing a convenient and effective means for sharing ideas, linking helps scientists and scholars promote their research goals.

LinkOut is a powerful linking feature of the NCBI Entrez search and retrieval system. It is designed to provide users with links from database records to a wide variety of relevant online resources, including full-text publications, biological databases, consumer health information, and research tools. (See [Sample Links](#) for examples of LinkOut resources.) The goal of LinkOut is to facilitate access to relevant online resources beyond the Entrez system to extend, clarify, or supplement information found in the NCBI databases. By branching out to relevant resources on the Web, LinkOut expands on the theme of Entrez as an information discovery system.

LinkOut is not just a list of links to Web sites. Two unique aspects of LinkOut set it apart from linking features in other information retrieval systems.

1. Specificity—LinkOut links users to resources that are specific to the subject of an Entrez record, e.g., linking to the full-text article of a PubMed citation, not the table of contents of the journal; to a specific section on [Ginkgo Biloba](#), not just the searching interface for the [USDA/NRCS PLANTS Database](#).
2. Voluntary participation—Participation in LinkOut is free and voluntary. Links are provided by external parties that create the link format, URL, and functionality. Resources reside on the provider sites, and they determine who may access their content. It is a unique collaboration with no parallel in similar data retrieval systems where links are typically created by the retrieval systems.

History

LinkOut was developed in 1999 during the reengineering of the NCBI Entrez system. The system was designed to allow third party providers to send links to their resources to be used by the Entrez databases.

¹ NCBI.

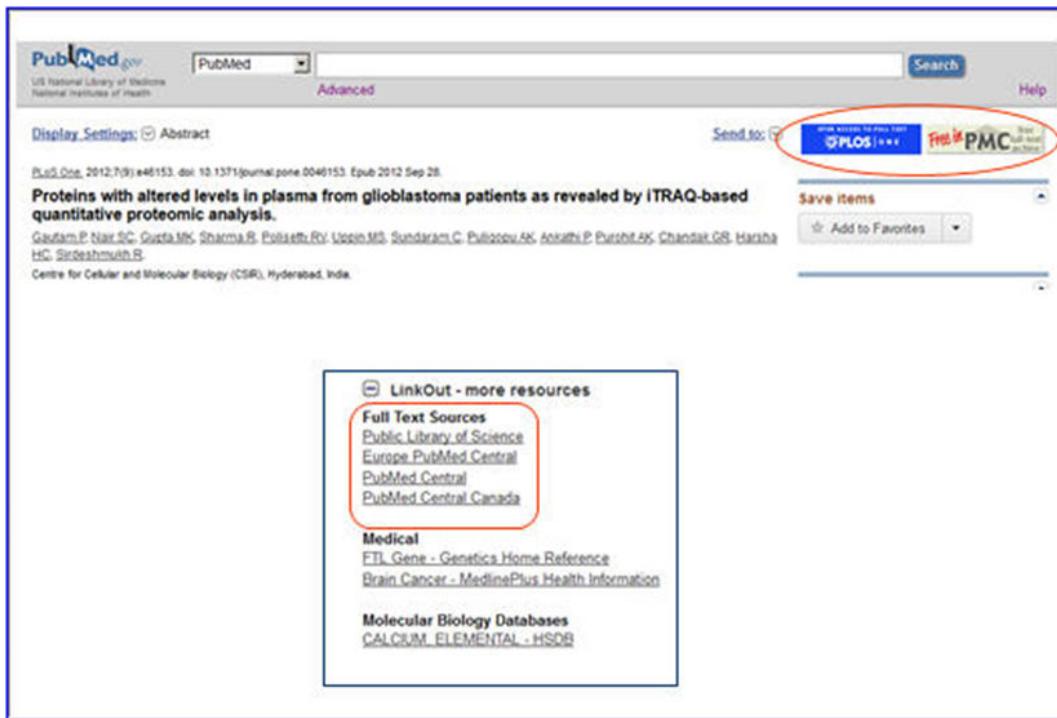


Figure 1: Full text links in PubMed

All Entrez databases can enable LinkOut. See the current list of [databases available for linking](#).

LinkOut has connected NCBI users to resources at over 3700 third party sites in more than 70 countries, for over 99 million NCBI database records. This greatly enhances the utility of the databases. Full text links in PubMed citations are an essential feature of the database, as shown in Figure 1.

Data Model

LinkOut is itself an Entrez database that holds the linking information to external resources. The separation of the Entrez database records (e.g., PubMed citations) from the external linking information (e.g., URLs to journal articles on a publisher's Web site) enables both the external link providers and NCBI to manage linking in a flexible manner. If links to external resources change, such as in the case of a Web site redesign, it will not affect the Entrez database records. Consequently, linking information can easily be updated as frequently as necessary.

The logical data unit of LinkOut is the relationship between a link and its target in the Entrez system. It is summarized in Figure 2.

Each Link Target consists of a database name and a record unique identifier (UID) that the link applies to.

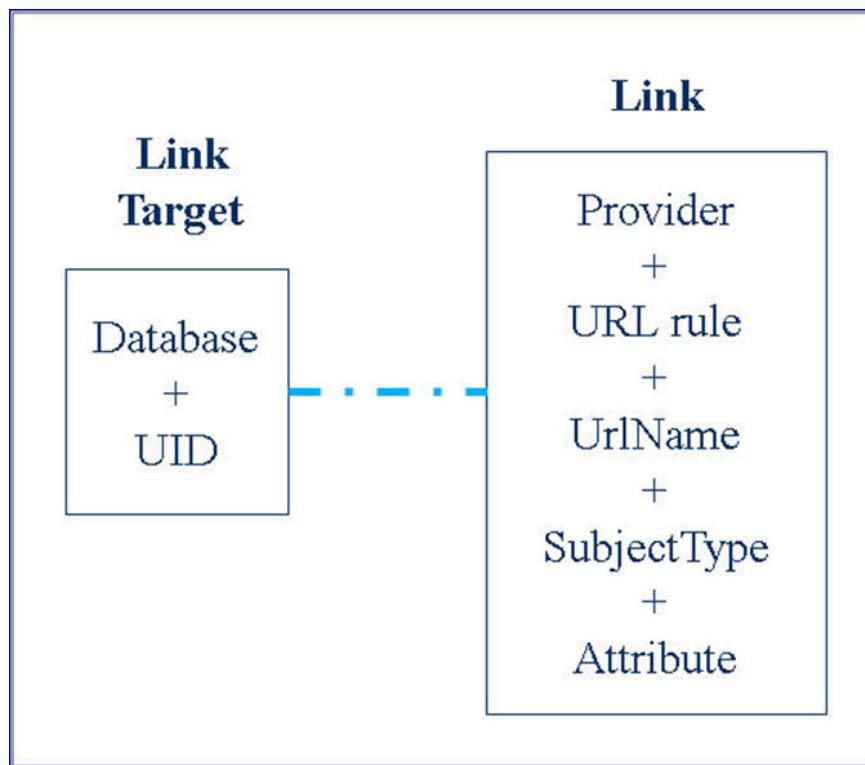


Figure 2: LinkOut Data Model

Each Link consists of:

- **Provider** that identifies the supplier of the link;
- **URL rule** that will be used to build the URL to link to the resource at the provider site;
- **UrlName** is a text string supplied by the link provider to describe the resource;
- **SubjectType** and **Attribute** are NCBI keywords to describe the resource.

LinkOut data units are retrieved by the frontend program of the target Entrez databases when the frontend is building the display of records. LinkOut filters are also built based on these data units to facilitate retrieval of LinkOut links in the target databases.

Dataflow

LinkOut dataflow is summarized in the following diagram:

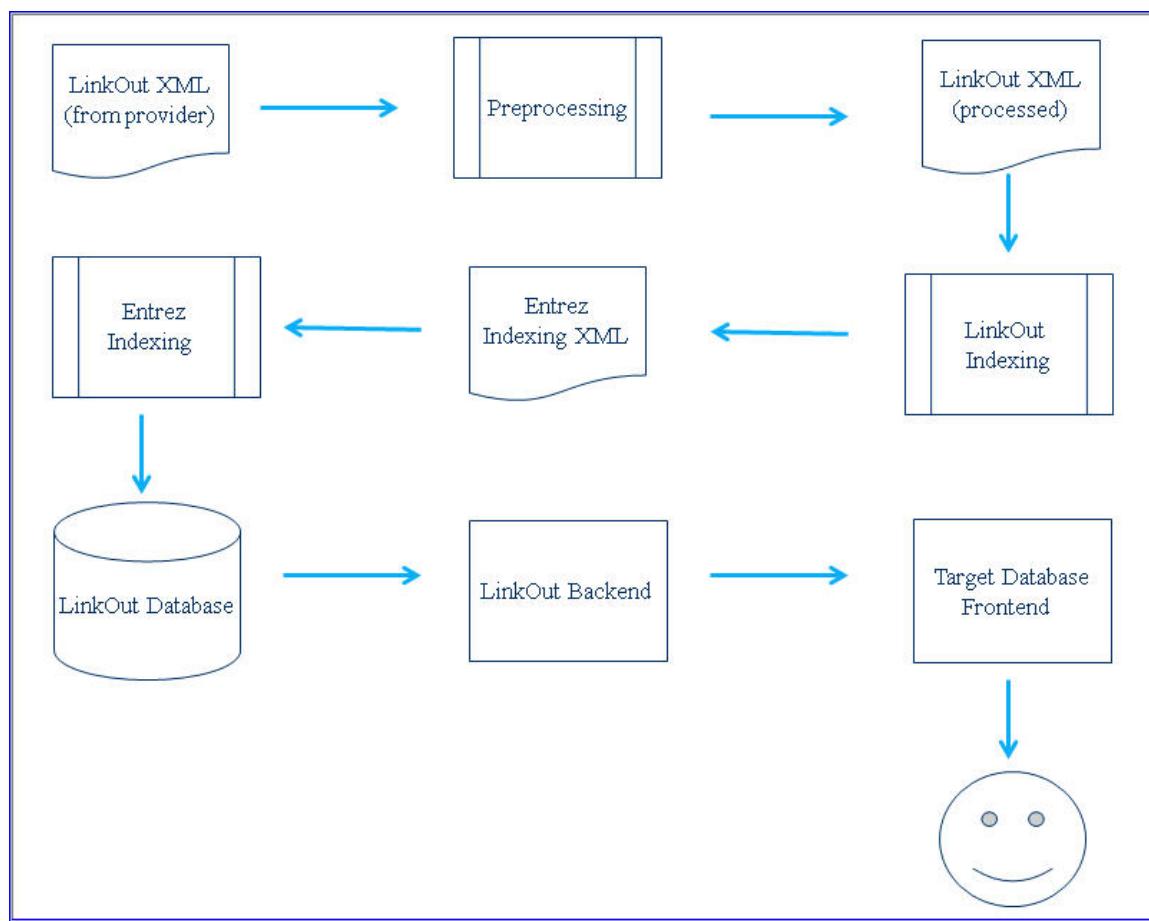


Figure 3: LinkOut Dataflow

LinkOut XML Files from Providers

LinkOut information is submitted by link providers in XML. LinkOut XML is defined by the [LinkOut Document Type Definition \(DTD\)](#).

The LinkOut DTD specifies all the elements needed to build the logical data unit for the LinkOut database. The DTD also specifies the link provider information needed for future communication with the providers regarding their links.

Two root elements are specified in the LinkOut DTD: the `<Provider>` element, which specifies information about a link provider; and the `<LinkSet>` element, which describes information about the link. Each root element is submitted to NCBI in a separate file. An identity file contains the `<Provider>` element, and a resource file contains the `<LinkSet>` element.

The identity file, `providerinfo.xml`, describes the identity of a provider, including an ID `<ProviderId>`, an abbreviated name `<NameAbbr>` assigned by NCBI, the provider's name, and other general information about the provider. There is only one `providerinfo.xml` file for each provider (see [the details of the identity file](#)).

The resource file, which contains the linking information, specifies a set of Entrez database records by UIDs or a valid query to the database, a specific rule to build the URL to an external resource, and the description of the resource using the SubjectType, Attribute, and UrlName fields. There is no standard for naming resource files, except that they must use the *.xml* extension. There may be any number of resources files associated with a ProviderId (see [the details of the resource file](#)).

Auxiliary Tools

The following tools facilitate LinkOut file submissions:

- [Library LinkOut Files Submission Utility](#) This utility was developed for libraries to generate and manage their LinkOut files. Libraries simply check off their electronic journal collections from a list of journals that participate in LinkOut, without needing to construct the LinkOut files by hand.
- [LinkOut File Validation](#) This utility is used by providers to parse their LinkOut files to ensure the accuracy of the files before submission. Besides validating the file syntax against the LinkOut DTD, this tool ensures that only allowable SubjectType and Attribute terms have been provided.

Preprocessing

LinkOut files from providers are passed through a preprocessing stage that converts them to standardized files for the LinkOut indexer.

The preprocessing stage includes three sub-processes:

1. Loading—Files submitted by providers in their FTP accounts are copied over to a staging area for the LinkOut indexer. During this sub-process, the XML files are adjusted to make sure the information in the file is valid. For example, ProviderId is a valid ID assigned by NCBI and only terms from the control lists of SubjectType and Attribute are used. The adjusted files are also validated against the LinkOut DTD to ensure XML files passed to the LinkOut indexer are with valid syntax. The loading sub-process compares the XML files in a provider's FTP account and the files in the staging area and decides which files in the FTP account should be processed. Only new and changed files in a provider's FTP account will be processed.
2. XML files generation—Holdings information entered into the Library Submission Utility are transformed into LinkOut XML files and added to the loading sub-process discussed above. Only holdings from libraries that have changed holdings information are processed.
3. Icon processing—[Icons](#) specified in the <IconUrl> element are downloaded. The icons downloaded are adjusted to make sure their size is within the allowed dimensions specified by NCBI. The processed icons are transferred to the target databases to form a part of the record display.

LinkOut Indexing

The LinkOut indexing process takes the standardized LinkOut XML files in the staging area and converts them into the standard XML suitable for Entrez indexing.

This process converts all queries specified in LinkOut file <Query> elements to UIDs of the target database. Since a <Query> in a LinkOut file can be translated to a different set of UIDs as the target database changes over time, it is necessary to process all LinkOut files in each LinkOut indexing. For example, <Query>plos one [journal]</Query> will be translated to a different set of PubMed IDs as citations for PLOS One are added to PubMed between LinkOut indexing.

LinkOut indexing is optimized to execute duplicate queries across all LinkOut files only once. In doing so, the number of queries needed to be processed has been reduced significantly and the speed of LinkOut indexing has been improved by tenfold.

Entrez Indexing

The Entrez indexing process builds the LinkOut database using the LinkOut files in the standard Entrez indexing format. The unit of the LinkOut database corresponds to the logical data unit described in the Data Model section. The LinkOut database can be queried directly with Entrez search commands internally at NCBI. It is also the data source for the LinkOut backend program. During the Entrez indexing process, LinkOut filters are generated for each target database. A LinkOut filter is a list of UIDs of a target database with a name. Filters are sent to each target database to be included during the indexing process for the database. As a result, the filters can be searchable in the target database. See the Access section for the details on LinkOut filters.

LinkOut Backend and Frontend of the Target Database

LinkOut backend handles the interaction between the LinkOut database and the frontend program of the NCBI databases. It has standardized protocols and commands that allow access to linking data in the LinkOut database more effectively and efficiently.

Each Entrez database may present LinkOut information to users differently. Generally, a frontend program accesses the LinkOut backend to find out:

1. If a specific record has any LinkOut links;
2. If there are LinkOut links, the information for each link.

The frontend program uses the information returned by the LinkOut backend and the XSLT rules set by the target database to generate the display. The frontend program is also responsible for transforming the [LinkOut entities](#) returned by the LinkOut backend to the value for a specific record to form a valid URL.

For example, in NCBI sequence databases, `&lo.pacc;` translates into the Primary Access Number of the corresponding record. In the Nucleotide database, record GI: [467096719](#), the following linking rule:

<http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=chr10.pacc>

will be translated into the following URL:

http://genome.ucsc.edu/cgi-bin/hgTracks?org=human&position=NR_102347

Access

LinkOut resources associated with a record in a [LinkOut enabled database](#) can be accessed in a variety of ways. The [Using LinkOut](#) chapter of LinkOut Help contains up-to-date information on how to access LinkOut resources.

LinkOut resources are typically presented to users by SubjectType in the LinkOut display portlet of the target database and within the [My NCBI](#) system, making it easier to browse. Attributes are used to describe the nature of a LinkOut resource, e.g., whether a resource requires a subscription for access. A short text string may be used in the UrlName element to provide an additional description for a resource. UrlName is typically used when the allowed SubjectType and Attribute terms cannot describe the resource adequately or when multiple links are available from one provider for a single record in the target database.

LinkOut Filters

To facilitate search and retrieval of LinkOut resources, there are a number of filters in the LinkOut-enabled Entrez databases. These filters, although not part of the LinkOut database, are based on the results generated in the LinkOut indexing process. A LinkOut filter is a list of UIDs for a target Entrez database. The UIDs identify records with a common property in the target database. This property is reflected in the filter name.

LinkOut filters are all prefixed with **lo**. Filters are available for all allowable SubjectType and Attribute terms, and the NameAbbr of a provider. To retrieve a set of records by a certain LinkOut property, a filter name can be entered as a search term in a database search box.

Examples of LinkOut filters include:

- **loprov**—Filter for records with links to a specific LinkOut provider. Example:
 - “*loprov*” [filter]
 - Retrieves all records with links to the journal *Public Library of Science* in PubMed.
- **loattr**—Filter for records with links of a specific LinkOut Attribute. Example:
 - “*loattrfull text online*” [filter]
 - Retrieves all records with at least one full text link in PubMed.
- **losubjt**—Filter for records with links of a specific LinkOut SubjectType. Example:
 - “*losubjorganism specific*”[filter]
 - Retrieves all records with links to resources of the SubjectType “organism specific”, i.e., resource in a database providing data specific to a particular organism or group of organisms.

- ***loall***—Filter for all records in a database with at least one LinkOut resource.

Example:

- “***loall***”[filter]

The Advanced Search Builder of each database can be used to retrieve LinkOut filters. Select **Filter**, type “**lo**”, and then click the **show index list** link to browse through the filters related to LinkOut.

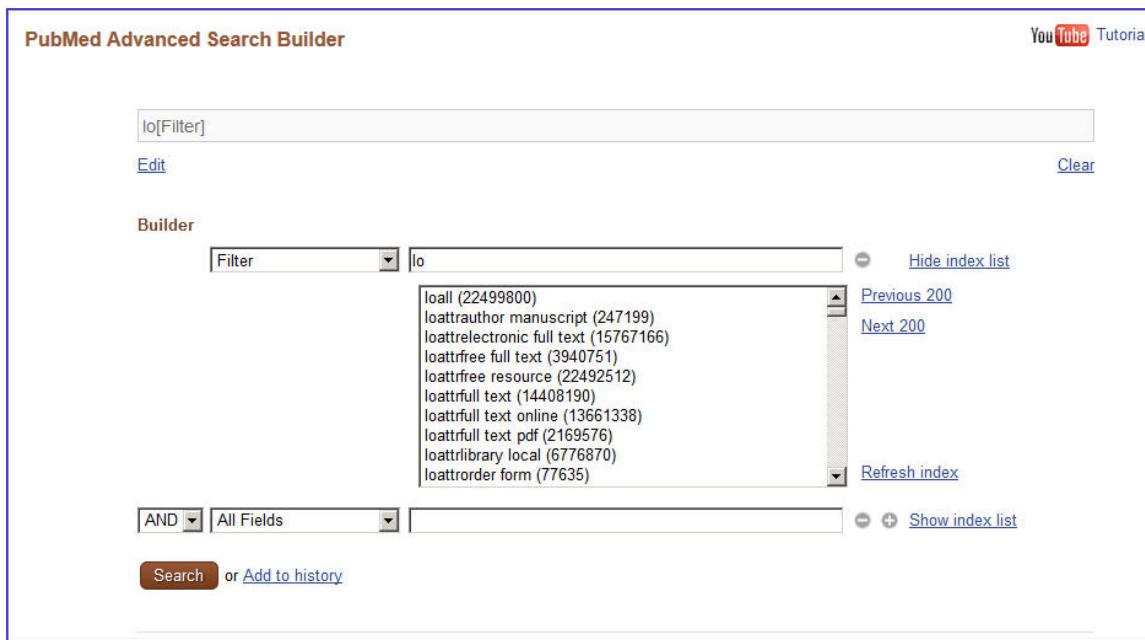


Figure 4: LinkOut Filters in Advanced Search Builder

Users can also customize LinkOut filters and icons by setting the **Filters** preferences in My NCBI.

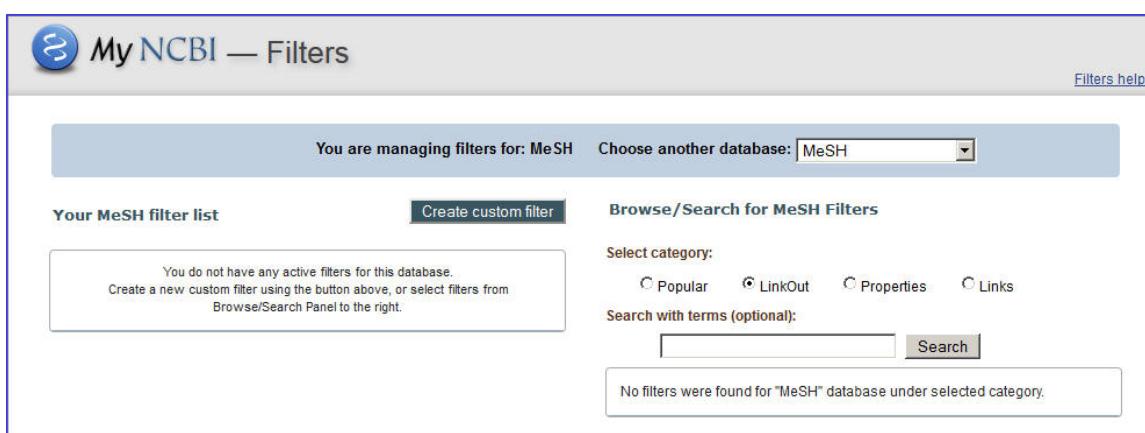


Figure 5: My NCBI Filters

Guide to LinkOut Providers

Information on LinkOut participation and file preparation is available from the manual: [LinkOut Help](#). This documentation includes specific chapters for full text providers, libraries, and providers of general resources.

Information about LinkOut is available on the [LinkOut home page](#), including FAQs, a list of existing providers, and access to various informational lists.

Users are welcome to communicate directly with the NCBI LinkOut team. Questions and comments about LinkOut can be sent to the mailing list: linkout@ncbi.nlm.nih.gov

Metadata

BioSample

Tanya Barrett, PhD¹

Created: November 14, 2013.

Scope

The BioSample database (1) stores submitter-supplied descriptive information, or metadata, about the biological materials from which data stored in NCBI's primary data archives are derived. NCBI's archives host data from diverse types of samples from any species, so the BioSample database is similarly diverse; typical examples of a BioSample include a cell line, a primary tissue biopsy, an individual organism or an environmental isolate.

The BioSample 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. This information is important for providing context to the derived data so that it may be more fully understood; it adds value, promotes re-use, and enables aggregation and integration of disparate data sets, ultimately facilitating novel insights and discoveries across a wide range of biological fields.

BioSample records are indexed and searchable. They are also reciprocally linked with the BioProjects in which they participate, as well as with derived experimental data in NCBI's primary archives including Sequence Read Archive ([SRA](#)), Gene Expression Omnibus ([GEO](#)), database of Genotypes and Phenotypes ([dbGaP](#)), as well as sections of [GenBank](#), including Expressed Sequence Tags (EST), Genome Survey Sequences (GSS), Whole Genome Shotgun (WGS), and Transcriptome Shotgun Assembly (TSA) sequences.

In summary, the BioSample database provides a dedicated environment in which to:

- Capture sample metadata in a structured way by promoting use of controlled vocabularies for sample attribute field names.
- Link sample metadata to corresponding experimental data across multiple archival databases.
- Reduce submitter burden by enabling one-time upload of a sample description, then referencing that sample as appropriate when making data deposits to other archives.
- Support cross-database queries by sample description.

¹ NCBI; Email: barrett@ncbi.nlm.nih.gov.

History

As the number and complexity of primary data archives supported by NCBI expands, a need has emerged for a shared database in which to host information about the biological samples from which those data are derived. Historically, each archive developed its own conventions for collecting sample metadata, with limited standardization of descriptions and no mechanism to indicate when the same sample was used across multiple sets of data. Furthermore, there is a growing awareness in the research community that sample metadata is essential for interpreting the data itself, and that opportunities for data re-use, aggregation, and integration increase with improved metadata.

The BioSample database was launched in 2011 to begin to help address these needs. It facilitates the capture and management of structured metadata descriptions for diverse biological samples and encourages data producers to provide a rich set of contextual metadata with their data submissions. The database was initially populated with existing sample descriptions extracted from SRA, dbGaP, EST and GSS. Over time, more NCBI archives are moving towards requiring BioSample deposit as part of data submission. As of May 2013, the database hosts almost 2 million BioSample records encompassing 18,000 species.

Data Model

The BioSample database stores descriptions of the biological materials used to generate data hosted by any of NCBI's primary data archives and, consequently, are very heterogeneous in nature. This, together with the fact that the content and granularity of metadata submitted to NCBI tends to be dependent on the context of the study, presents significant challenges in terms of procuring consistent sample descriptions from submitters.

To help address these challenges, the BioSample Submission Portal guides submitters into providing appropriate information. A number of common BioSample types are defined in the database, each comprising a package of relevant attributes with which to describe the sample. By guiding and encouraging submitters to use such attribute packages, it can be expected that the descriptions for samples deposited *via* this route will converge and become more consistent over time.

The full list and definitions of BioSample types and attributes is available for [preview and download](#). Examples include "Pathogen affecting public health," which is intended to procure information considered useful for the rapid analysis and trace back of pathogen samples, and the MiXS minimum information checklists as developed by the Genomics Standards Consortium (2) that are intended for standardizing descriptions of samples from which genomes, metagenomes, and targeted locus sequences are derived.

Attributes define the material under investigation using structured name: value pairs, for example:

tissue: liver

collection date: 31-Jan-2013

After specifying the sample type, the user is presented with a list of required and optional attribute fields to fill in, as well as the opportunity to supply any number of custom descriptive attributes. For example, if a submitter specifies that their sample is a clinical pathogen, they are required to input information about collection locality and date, host and isolation source. In addition, submitters are encouraged to provide information for additional attributes that further describe the host, disease state, etc. The values provided in some fields undergo validation to ensure proper content or format. The BioSample database is extendible in that new types and attributes can be added as new standards develop.

In addition to BioSample type (called Model in the schema) and attributes, each BioSample record also contains:

IDs: An identifier block that lists not only the BioSample accession assigned to that record, but also any other external sample identifier, such as that issued by the source database or repository.

Organism: The organism name and taxonomy identifier. The full taxonomic tree is displayed and searchable.

Title: BioSample title. A title is auto-generated if one is not supplied by the submitter.

Description: [optional] A free text field in which to store non-structured information about the sample.

Links: [optional] URL to link to relevant information on external sites.

Owner: Submitter information, including name and affiliation where available.

Dates: Information about when the record was submitted, released, and last updated.

Access: Statement about whether the record is fully public or controlled access (that is, in dbGaP).

BioSample records of interest include:

Reference BioSamples: While many samples can be considered unique and are used only once, other samples, including commercial cell lines or bacterial isolates, are used repeatedly by the research community. Major vendors, including the American Type Culture Collection (ATCC), the Coriell Institute for Medical Research , and the Leibniz Institute German Collection of Microorganisms and Cell Cultures (DSMZ), are working with us to generate official representations of commonly used and highly referenced samples. These are flagged as Reference BioSamples, so submitters who use these samples may bypass BioSample submission and simply reference relevant Reference BioSample records when depositing experimental data in any of NCBI's primary data archives. Also,

efforts are underway to map existing data from across NCBI archives to Reference BioSample records. Consequently, these Reference BioSample records serve as hubs from which users can quickly locate a multitude of diverse data sets and projects derived from a given sample.

Clinical samples: The BioSample database does not support controlled access mechanisms and thus cannot host human clinical samples that may have associated privacy concerns. Instead, clinical samples continue to be deposited in NCBI's dbGaP database. The dbGaP database then deposits abridged BioSample records that have had sensitive data attributes removed. This allows users to locate these data in BioSample, and then apply to dbGaP for access to the full descriptions as necessary.

Authenticated human cell line samples: The BioSample database hosts a growing collection of [authenticated human cell line](#) records aimed at addressing the problem of cell line misidentification (3). These records contain verified STR (short-tandem-repeat) profile information and supporting electropherogram evidence which researchers can use as a reference when checking the authenticity and purity of the cell lines from which they are publishing data.

Dataflow

Researchers typically initiate a deposit to BioSample as part of a submission to one of NCBI's primary data archives, and usually before a manuscript describing the data has been submitted to a journal for review. Researchers use their NCBI account to login and register BioSample submissions using a Web-based [Submission Portal](#) that guides them through a series of forms for input of metadata describing their samples. An XML-based submission route is also available for frequent submitters. In addition, direct data deposits to dbGaP and GEO trigger automatic creation of BioSample records.

The BioSample Submission Portal enforces provision of a minimal set of metadata via mandatory attributes for specific sample types, as well as encourages rich metadata by supporting the provision of any number of custom attributes. But ultimately, BioSample is a submitter-driven repository in that submitters are responsible for the quality and content of their deposits. Database staff respond to queries and report errors but, as with other primary data archives, submitted data are not subject to extensive curation. After passing syntactic validation, each sample is assigned a BioSample accession number which has prefix SAMN, e.g., [SAMN02048828](#). This accession number can subsequently be referenced as appropriate when submitting corresponding experimental data to the archival databases.

The BioSample records are typically released in conjunction with corresponding experimental data. At that time, the BioSample records are loaded and indexed in the [BioSample](#) database that is part of NCBI's Entrez search and retrieval system, where they may be queried and downloaded. The records are reciprocally linked to other databases where appropriate, including SRA, dbGaP, GEO, GenBank and BioProject, facilitating easy navigation to derived and related data.

Access

BioSample records may be accessed by query or by following a link from another NCBI database.

Query: Effective searches may be accomplished using the search box on the [BioSample home page](#). As with other NCBI Entrez databases, a simple free text keyword search is often sufficient to locate relevant data. However, BioSample data are indexed under several fields, meaning that users can refine their search by constructing fielded queries. Some example fielded queries are listed below and include searching by organism, attribute, or package. Users can write and execute their own search statements directly in the search boxes or use the [Advanced search](#) page to explore the indexed fields and construct multi-part fielded search statements. The [Limits](#) page may be used to restrict retrievals according to access-level, source databases, and publication dates.

Download: BioSample record content may be downloaded using the Send to: feature on the search results pages that allows download of individual or batch BioSample retrievals in text or XML formats. Furthermore, programmatic query and download functions are available using [Entrez Utilities](#).

Linking: BioSample records are reciprocally linked to related records in the archival databases. This allows users to link to, e.g., corresponding genome assembly records in the Nucleotide database, or raw sequence reads in SRA, or to navigate to the BioProject(s) in which the sample participates.

Example queries

Retrieve pathogen BioSamples released in the first quarter of 2013

`package pathogen[Properties] AND 2013/1:2013/3[Publication date]`

Retrieve BioSamples derived from bacteria of genus Shigella and for which SRA data is available:

`shigella[organism] AND biosample sra[filter]`

Retrieve BioSamples that conform to the MIGS/MIMS/MIMARKS.water package:

`package migs/mims/mimarks water[Properties]`

Retrieve BioSamples derived from mouse and for which strain and age information is available:

`(strain[Attribute Name] AND age[Attribute Name]) Mus musculus[organism]`

Retrieve BioSamples derived from fibroblast cells:

`cell type fibroblast[Attribute]`

The figure consists of two side-by-side screenshots of the NCBI BioSample interface.

Top Panel (A): BioSample Search Results

- Search Bar:** Shows the query "migs bacterial/archaeal".
- Results:** Displays 1 to 20 of 737 results, including:
 - MIGS Cultured Bacterial/Archaeal sample from Mannheimia haemolytica USDA-ARS-USMARC-185**: Sample ID: 2048830.
 - MIGS Cultured Bacterial/Archaeal sample from Mannheimia haemolytica USDA-ARS-USMARC-183**: Sample ID: 2048829.
 - MIGS Cultured Bacterial/Archaeal sample from Bibersteinia trehalosi USDA-ARS-USMARC-192**: Sample ID: 2048828; SRA: SRS419873.
- Filtering:** Options include "Send to:" (checkboxes for All, Controlled Access, EST, GSS, Public, Used by SRA), "Find related data" (Database dropdown), and "Manage Filters".

Bottom Panel (B): Full BioSample Record

- Display Settings:** Set to "Full".
- Identifiers:** BioSample: SAMN02048828; Sample name: Bibersteinia trehalosi USDA-ARS-USMARC-192; SRA: SRS419873.
- Organism:** Bibersteinia trehalosi USDA-ARS-USMARC-192, cellular organisms; Bacteria; Proteobacteria; Gammaproteobacteria; Pasteurellales; Pasteurellaceae; Bibersteinia; Bibersteinia trehalosi.
- Attributes:** A detailed table of sample characteristics, including:

Attribute	Value
environmental package	MIGS/MIMS/MIMARKS host-associated
investigation type	bacteria_archaea
biome	terrestrial biome
collection date	10-Oct-10
feature	carcass
geographic location (country and/or sea,region)	USA : Nebraska
health disease status	dead
isolation and growth condition	Not available
geographic location (latitude and longitude)	Not available
material	mucus
pathogenicity	animal
ploidy	haploid
project name	USMARC Bovine Respiratory Disease Pathogen Sequencing Project
reference for biomaterial	Not available
sample collection device	swab
specific host	Bos taurus
strain	USDA-ARS-USMARC-192
number of replicons	1
- Description:** Keywords: GSC:MixS,MIGS:3.0.
- Submission:** USDA-ARS-USMARC, Gregory Harhay, 2013-04-18.
- Related Information (E):** Links to BioProject, Nucleotide, SRA, and Taxonomy.
- Download SRA Data (F):** Provides links to SRA runs (e.g., SRS419873, SRX276683, SRX276684, SRX276685, SRX276686, SRX277383) with download statistics (e.g., 11.3G, 421.3M, 361.3M, 396.9M, 10.1G, 35.5M).

Figure 1. Screenshots of BioSample search results (top panel) and a full BioSample record (bottom panel). Users enter a query into the Search box, or use the Limits or Advanced search pages (A) and retrieve a list of matching BioSamples (B). Search results are displayed in Summary format by default, which presents the title, organism, sample type, and identifiers. Clicking a title takes the user to the full record that lists all the sample attributes, identifiers, and submitter information (C). The Send to: feature (D) allows download of individual or batch BioSample retrievals in text or XML formats. Links are provided to related records in other archives (E), in this case BioProject, Nucleotide, SRA, and Taxonomy. Where appropriate, an option to download SRA sequence data generated from that sample is provided (F).

References

- Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. BioProject

- and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* 2012;Jan40(Database issue):D57–63. PubMed PMID: 22139929.
2. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methe BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glockner FO. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol.* 2011;May29(5):415–20. PubMed PMID: 21552244.
 3. Masters JR. Cell-line authentication: End the scandal of false cell lines. *Nature.* 2012;Dec 13492(7428):186. PubMed PMID: 23235867.

BioProject

Karen Clark, PhD,¹ Kim Pruitt, PhD,¹ Tatiana Tatusova, PhD,¹ and Ilene Mizrachi, PhD¹

Created: April 28, 2013; Updated: November 11, 2013.

Scope

The BioProject database provides an organizational framework to access information about research projects with links to data that have been or will be deposited into archival databases maintained at members of the International Nucleotide Sequence Database Consortium (INSDC, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive at European Molecular Biology Laboratory (ENA), and GenBank at the National Center for Biotechnology Information (NCBI)) (1,2,3).

BioProjects describe large-scale research efforts, ranging from genome and transcriptome sequencing efforts to epigenomic analyses to genome-wide association studies (GWAS) and variation analyses. Data are submitted to NCBI or other INSDC-associated databases citing the BioProject accession, thus providing navigation between the BioProject and its datasets. Consequently, the BioProject is a robust way to access data across multiple resources and multiple submission timepoints, e.g., when there are different types of data that had been submitted to multiple databases, or sequential submissions deposited over the course of a research project.

The definition of a set of related data, a “project,” is very flexible, so using different parameters allows the creation of a complex project and various distinct sub-projects. For example, BioProject records can be established for:

- Genome sequencing and assembly
- Metagenomes
- Transcriptome sequencing and expression
- Targeted locus sequencing
- Genetic or RH Maps
- Epigenomics and functional genomics
- Phenotype or Genotype
- Variation detection

The BioProject database encompasses taxonomic diversity, from humans and animals, to plants, to prokaryotes and metagenomes. BioProjects are created for initiatives that generate a very large volume of data, data from multiple members of a consortium or collaboration, or data being submitted to multiple archival databases. BioProject

¹ NCBI; Email: kclark@ncbi.nlm.nih.gov; Email: pruitt@ncbi.nlm.nih.gov; Email: tatiana@ncbi.nlm.nih.gov; Email: mizrachi@ncbi.nlm.nih.gov.

registration is required for some database submissions including dbVar, the Sequence Read Archive (DRA/ERA/SRA), and microbial and eukaryotic genome submissions to DDBJ/ENA/GenBank. However, small datasets that have one or a few sequences, like a single viral or organellar genome, are not in scope for BioProject.

The BioProject database defines two types of projects: 1) primary submission projects, as described above, are directly associated with submitted data and may be registered by submitters of that data using the NCBI submission portal; 2) umbrella projects, which reflect a higher-level organizational structure for larger initiatives or provide an additional level of data tracking. These projects are created by request. An umbrella project groups projects that are part of a single collaborative effort but represent distinct studies that differ in methodology, sample material, or research grant. Complex research efforts may be represented with more than one layer of umbrella project such that a highest-level umbrella project is linked to one or more sub-project umbrella projects which in turn are linked to one or more Primary submission projects that describe the data in more detail.

History

The BioProject resource became public in May 2011, replacing the older NCBI Genome Project database, which had been created to organize the genome sequences in GenBank and RefSeq (4). The BioProject database was created to meet the need for an organizational database for research efforts beyond just genome sequencing, such as transcriptome and gene expression, epigenomics, and variation studies. However, because a BioProject is defined by its multiple attributes, there is flexibility for additional types of projects in the future, beyond those that were included in 2011. The new BioProject database allows more flexible grouping of projects and can collect more data elements for each project, e.g., grant information and project relevance. Projects registered in the old Genome Project database were incorporated into BioProject, and a BioProject Accession was assigned in addition to the numerical ID that was previously assigned in Genome Project.

Data Model

Primary submission projects have attributes that describe the scope, methodology, and objectives of the project. The attributes are:

- **Sample Scope** indicates the sample purity and scope. The options are monoisolate, multiisolate, multi-species, environment, synthetic, and other.
- **Material** indicates the type of material isolated from the sample. The options include genome (for a genome or metagenome), purified chromosome, transcriptome, phenotype (phenotypic descriptive data), reagent (material studied was obtained by chemical reaction, precipitation), proteome (protein or peptide data).
- **Capture** indicates the scale, or type, of information that the study is designed to generate from the sample material. The options include whole (meaning that a

specific subset was not used and which is the most common case), exome (capturing exon-specific data), and TargetedLocusLoci (specific loci such as a gene, genomic region, or barcode standard).

- **Method** is the general approach to generate the data. The options include sequence, array, and mass spectrometry.
- **Objective** is the project goals with respect to the type of data that will be generated and submitted to the data archives. Options include raw sequence reads (data to SRA); sequence, assembly and annotation (data to GenBank); expression (data to GenBank or GEO); variation (data to dbSNP or dbVAR); epigenetic markers (data to GEO or SRA) and phenotype (data to dbGaP).

The combination of attributes determines the project data type, (descriptive label), such as genome sequencing or epigenomics. Since multiple kinds of data, e.g., genome and transcriptome data, can be submitted with the same BioProject identifier, the project data type includes both values, Genome Sequencing and Transcriptome.

The BioProject also stores submitter and grant information, related publications, and links to external Web resources that are relevant for the project. Furthermore, the organism name, taxid, and infra-species identifier (strain, breed, cultivar, or isolate) of the Target of a project are currently stored in the BioProject database, and the organism name is refreshed daily by a lookup in the taxonomy database. However, by the end of 2014 the organism information will be maintained in the related BioSample database and only cached for display in the BioProject page.

The BioProject XML schema is presented on the FTP site, <ftp://ftp.ncbi.nlm.nih.gov/bioproject/>

Additional information about BioProject, including a glossary of terms, is available in the [BioProject Help document](#).

Dataflow

Primary submission records may be created through the NCBI Submission Portal via several paths: (1) interactive Web portal, <https://submit.ncbi.nlm.nih.gov/subs/bioproject/>; (2) programmatic XML-based ui-less interface; (3) as part of data submission to some resources, such as GEO or dbGaP. In addition, RefSeq processes can create primary submission BioProjects.

Umbrella projects are created by NCBI staff at the request of submitters or funding agencies. Once an umbrella project exists, submitters link to the umbrella when creating a new primary BioProject. In addition, NCBI staff can create links between an umbrella project and pre-existing sub-projects at the submitter's request.

Error-free submissions are loaded into the database and assigned a BioProject accession, which has the format of five letters plus a series of digits, e.g., PRJNA31257. BioProjects are made public immediately unless a hold-until-published (HUP) date is requested. In that case, the BioProject is released on that HUP date or when the BioProject accession or

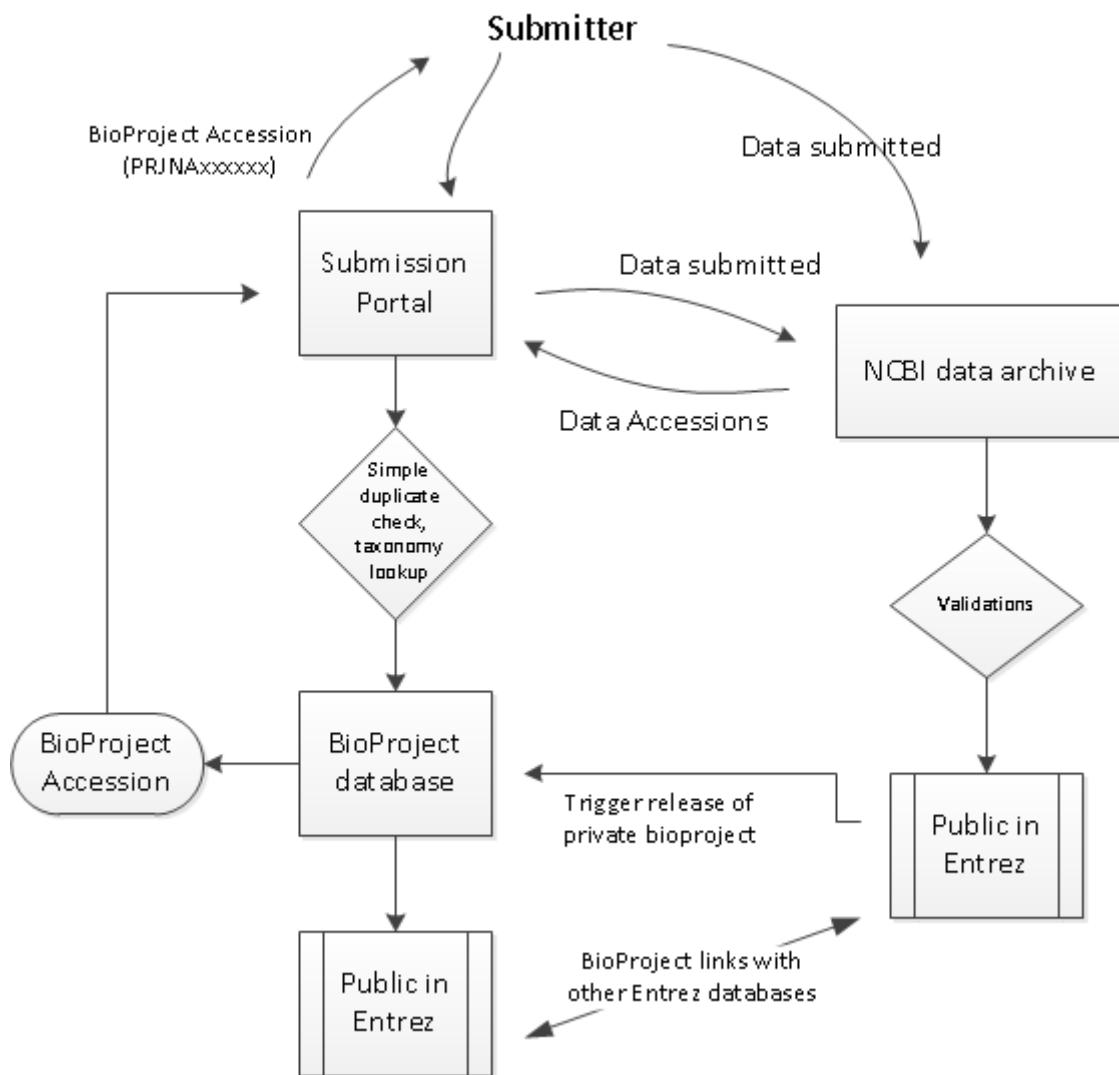


Figure 1. Workflow of BioProject submission. Projects are registered in submission portal and accession numbers are assigned for citation in related data. Related data is submitted and includes the BioProjectID accession number. Release of the data triggers release of the BioProject, if it is still confidential, and links between the BioProject and data are created in Entrez.

the linked data is cited in a publication, or when data with that BioProject accession are released, whichever of those events occurs first. Public data in NCBI archives that include a BioProject accession trigger the creation of a reciprocal link in Entrez between the data record for that archive and the BioProject (Figure 1).

Creation of a BioProject is not sufficient for publication. The data that corresponds to that BioProject also needs to be submitted to the appropriate INSDC-associated database.

Public BioProjects are exchanged with the members of the INSDC nightly.

Access

In Entrez, BioProject records may be accessed by browsing, by query, by download, or by following a link from another NCBI database.

Browsing: From the [BioProject home page](#) users can navigate to the “[By project attributes](#)” hyperlink to browse through the database content by major organism groups, project type (umbrella projects vs. primary submissions), or project data type. The table includes links to the [NCBI Taxonomy database](#) where additional information about the organism may be available and to the BioProject record.

Query: Searches can be performed in BioProject like any other Entrez database, namely by searching for an organism name, text word, or BioProject accession ([PRJNA31257](#)), or using the Advanced Search page to build a query restricted by multiple fields. Search results can be filtered by Project Type, Project Attributes, Organism, or Metagenome Groups, or by the presence or absence of associated data in one of the data archives.

Here are some representative searches:

Find BioProjects by...	Search text example(s)
A species name	<code>Escherichia coli[organism]</code>
Project data type	<code>"metagenome"[Project Data Type]</code>
Project data type and Taxonomic Class	<code>"transcriptome"[Project Data Type] AND Insecta[organism]</code>
Publication	<code>"19643200"[PMID]</code>
Submitter organization, consortium, or center	<code>JGI[Submitter Organization]</code>
Sample scope and material used	<code>"scope environment"[Properties] AND "material transcriptome"[Properties]</code>
A BioProject database identifier	<code>PRJNA33823 or PRJNA33823[bioproject] or 33823[uid] or 33823[bioproject]</code>

Download: In addition to the Entrez Web interface and the BioProject browse page, users can download the entire database and the database .xsd schema from the FTP site, <ftp://ftp.ncbi.nlm.nih.gov/bioproject/>, or use Entrez Programming Utilities ([E-utilities](#)) to programmatically access public BioProject records.

Linking: BioProject records can be found by following links from archival databases when the data cites a BioProject accession. Links may be found in several databases including SRA, Assembly, BioSample, dbVar, Gene, Genome, GEO, and Nucleotide (which includes GenBank and RefSeq nucleotide sequences).

Report formats

Citrobacter sp. KTE151

Accession: PRJNA157563 ID: 157563

Citrobacter sp. KTE151 Genome sequencing

Project Data Type: Genome sequencing; **Locus Tag Prefix:** WC7

Attributes: Scope: Monoisolate; Material: Genome; Capture: Whole; Method type: Sequencing

Relevance: Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	35
WGS master	1
Genomic DNA	12
SRA Experiments	2
Protein Sequences	4937
OTHER DATASETS	
BioSample	1
Assembly	1

See Genome Information for Citrobacter

NAVIGATE UP
This project is a component of the Studying UTI Defensins

NAVIGATE ACROSS
13 additional projects are related by organism.
236 additional projects are components of the Studying UTI Defensins.

▼ Assembly details:

Assembly	Level	WGS	BioSample	Taxonomy
GCA_000398845.1	Scaffold	ASQK00000000	SAMN00847640	1169322

▼ SRA Data Details

Parameter	Value
Data volume, Gbases	1
Data volume, Mbytes	960

Figure 2. Full Report. This Primary submission project has links to the data records of an annotated WGS genome in the Nucleotide, Protein, SRA, BioSample, and Assembly databases, and to the Genome database where information about this organism is presented. In addition, there are navigation links up to an umbrella project to which PRJNA157563 belongs, and across to other BioProjects that are related by being part of the same umbrella, or by being the same species.

Summary

The Summary view provides a concise overview of the project and includes the BioProject name (which is often the organism name), title, Taxonomy, Project data type, Attributes, the project source, submitting organization, and the BioProject accession and ID (uid). The Project name or label is linked to the full report page, shown in Figure 2.

Full Report

The Full Report display for Primary submission projects, as shown in Figure 2, includes the project name and/or title, a text description of the project (when provided), the project data type and specific project attributes, a project data section with data links, citations relevant to the project, taxonomic lineage, information about the submitting group and project funding. Navigation tools are provided near the top of the report to facilitate navigation to NCBI's taxonomically organized Genome resource, “up” to higher-

A

[Display Settings:](#)

[Send to:](#)

Studying UTI Defensins

Studying UTI defensins.

Project Type: Umbrella Comparative genomics project (**Subtype:** Comparative genomics)

Relevance: Medical

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	17705
WGS master	234
Genomic DNA	17471
SRA Experiments	430
Protein Sequences	1118448
OTHER DATASETS	
BioSample	237

SRA Data Details

Parameter	Value
Data volume, Gbases	367
Data volume, Tbytes	0.21

encompasses the following 237 sub-projects:

B

Project Type	Number of Projects		
Genome sequencing Highest level of assembly: Scaffolds or contigs SRA or Trace Total	234 2 236		
BioProject accession	Assembly level	Organism	Title
PRJNA157563	Scaffolds or contigs	Citrobacter sp. KTE151	Citrobacter sp. KTE151 (Broad Institute)
PRJNA157557	Scaffolds or contigs	Citrobacter sp. KTE30	Citrobacter sp. KTE30 (Broad Institute)
PRJNA157619	Scaffolds or contigs	Citrobacter sp. KTE32	Citrobacter sp. KTE32 (Broad Institute)
PRJNA157549	Scaffolds or contigs	Escherichia coli KTE1	Escherichia coli KTE1 (Broad Institute)
PRJNA157579	Scaffolds or contigs	Escherichia coli KTE10	Escherichia coli KTE10 (Broad Institute)
List all 236 'Genome sequencing' projects...			
Other	1		
BioProject accession	Name	Title	
PRJNA157073	UTI Defensins	UTI Defensins (Broad Institute)	

Submission:

Registration date: 20-Mar-2013

Broad Institute

Figure 3. Report page of an Umbrella BioProject. A) The data links of the sub-projects are summed in the Project Data table. B) The sub-projects are displayed, clustered by project type. The level of genome sequencing projects is included, and those projects can be sorted by that value.

level umbrella projects, or “across” to other BioProject records that are related by organism, or via a common umbrella project.

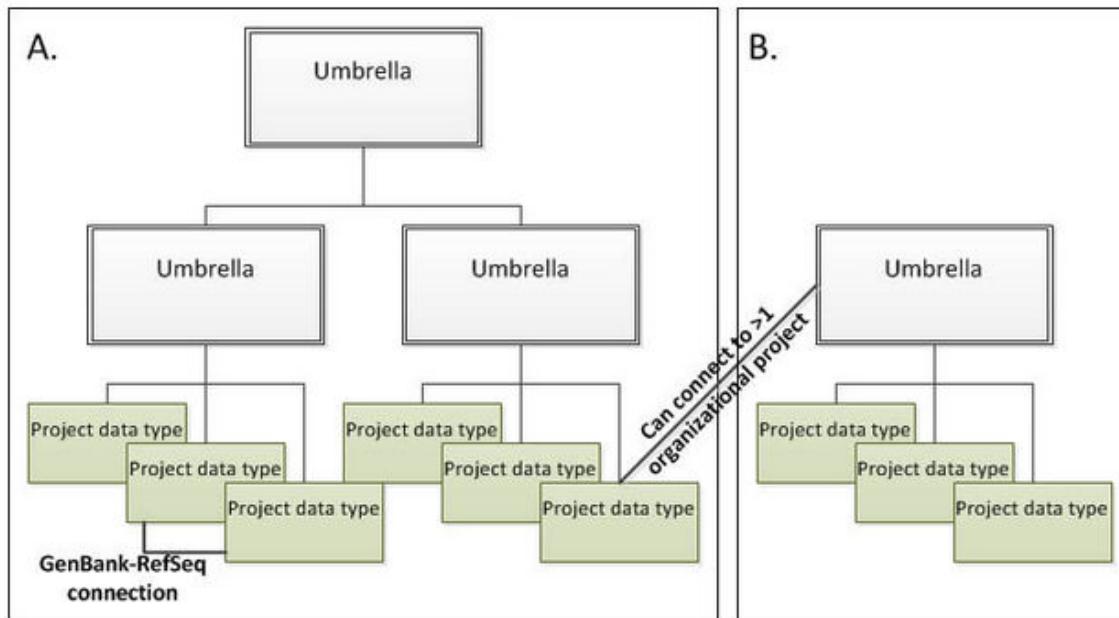


Figure 4. Schematic diagram of BioProject hierarchies. A) Large initiatives that have distinct sub-projects may have more than one level of umbrella project. For example, a top-level umbrella project groups all components of the initiative; mid-level umbrella projects reflect distinct branches of the project (such as sequencing vs. epigenetics); and several submission projects denote distinct project data types (e.g., genome sequencing, transcriptome, epigenetics, etc.). B) Other initiatives may be organized under a single umbrella project with one or many submitted projects that are connected to data. Note that a given submission project may have no connection to any hierarchical umbrella projects, or may be connected to more than one organizational layer, and there may be connections directly between submitted projects such as the indicated RefSeq to GenBank link.

When the experimental data for a BioProject is submitted to archival databases, it contains the BioProject accession that links the data to the BioProject report page. The Project Data table in the report page presents data counts from those archival databases that have links to the BioProject. Genome sequencing BioProjects also have a table that reports the genome assembly's accession number in the [Assembly](#) database, the BioSample accession, as well as the master accession number for whole genome sequencing ([WGS](#)) project, if relevant.

The page includes navigation tools to facilitate navigating to the related Genomes resource, which focuses on taxonomically organized genome sequencing projects, or to a linked umbrella project, or to “peer” projects that share a link to the same umbrella project or by shared taxonomy. If a genome assembly is represented by both an INSDC genome sequencing project, and a RefSeq genome project, then the correspondence between these projects is also indicated in the full report.

An umbrella project report page includes the relevant tabular reports listing the sub-projects that belong to that umbrella. The sub-projects may be 1) multiple primary submission projects of the same type, e.g., the HMP Reference Genome project

[PRJNA28331](#), 2) different kinds of primary submission bioprojects, e.g., [PRJNA193500](#) (Figure 3), or 3) other umbrella projects, e.g., the HMP top-most project, [PRJNA43021](#). The Data table of an umbrella project presents a sum of the data links for its grouped sub-projects, as seen for [PRJNA193500](#) (Figure 3).

Some large initiatives are represented by more than one layer of umbrella projects (see Figure 4); for instance, a top-most level may identify the largest definition of the collaboration; a second level of umbrella projects identify the primary categories of data production; and finally a third layer represents the projects that actually generate the data that is submitted. The Human Microbiome project is an example of this type of complex hierarchy where the top-most project, [PRJNA43021](#), represents the most inclusive definition of the initiative, and a secondary level (such as [PRJNA28331](#)) identifies a major sub-project to sequence multiple reference genomes each of which has a distinct project accession.

Related Tools

BioProjects may be registered with the submission portal at <https://submit.ncbi.nlm.nih.gov/subs/bioproject/>.

The submission portal is at <https://submit.ncbi.nlm.nih.gov/subs/> and is designed to be a single place where submitters can register and deposit their data for multiple NCBI archives. As of November 2013, the submission portal is operational for BioProject, BioSample, WGS genomes, TSA transcriptomes, and the Genetic Testing Registry (GTR).

Other related resources include the BioSample, Assembly, and Genome databases. BioSample and BioProject are similar as they are both entry points for aggregating and retrieving data of a single research effort or sample from various NCBI databases.

The BioProject, Genome, and Assembly databases are interconnected and can be used to access and view genome assemblies different ways. Every prokaryotic and eukaryotic genome submission has BioProject, BioSample, Assembly, and GenBank accession numbers, so users can start in any of those resources and get to the others. The BioProject and BioSample databases allow users to find related data-sets, e.g., multiple bacterial strains from a single isolation location, or the transcriptome and genome from a particular sample. The Assembly accession is assigned to the entire genome and is used to unambiguously identify the set of sequences in a particular version of a genome assembly from a particular submitter. Finally, the Genome database displays all of the genome assemblies in INSDC and RefSeq, organized by organism.

References

1. Pruitt K, Clark K, Tatusova T, Mizrachi I. BioProject Help [Internet]. Bethesda (MD): National Center for Biotechnology Information; 2011 May. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK54015>
2. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt K, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. BioProject and

- BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. 2011;40(D1):D57–63. PubMed PMID: 22139929.
3. Nakamura Y, Cochrane G, Karsch-Mizrachi I; International Nucleotide Sequence Database Collaboration. Nucleic Acids Res. 2013;41(D1):D21–24. PubMed PMID: 23180798.
 4. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Ostell J, Miller V, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Sirotnik K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2007;35(D1):D5–D12. PubMed PMID: 17170002.

Glossary

accession number — The accession number is a unique identifier assigned to a record in sequence databases such as GenBank. Several NCBI databases use the format [alphabetical prefix][series of digits]. A change in the record in some databases (e.g. GenBank) is tracked by an integer extension of the accession number, an Accession.version identifier. The initial version of a sequence has the extension “.1”. When a change is made to a sequence in a GenBank record, the version extension of the Accession.version identifier is incremented. For the sequence NM_000245.3, “.3” indicates that the record has been updated twice. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the original Accession.version identifiers.

AGP file — AGP ('A Golden Path') file is used to describe the instructions for building a contig, scaffold, or chromosome sequence. This file specifies the order, orientation, and switch points for each genomic sequence.

alignment — An alignment is a representation of the similarity between 2 nucleotide or protein sequences. In the case of protein sequences, the amino acids derived from ancestral sequences are taken into consideration in the alignment to account for conserved sequence. A pairwise alignment involves 2 sequences and a multiple alignment involves 3 or more sequences. A global alignment involves aligning the entire sequence whereas a local alignment involves aligning subsequences. The optimum alignment is determined by highest score for a given system. In a structural alignment, 3-dimensional structures of proteins under consideration are superimposed (Koonin and Galperin 2003a).

allele — One of the variant forms of a gene at a particular locus on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one). When “genes” are considered simply as segments of a nucleotide sequence, allele refers to each of the possible alternative nucleotides at a specific position in the sequence. For example, a CT polymorphism such as CCT[C/T]CCAT would have two alleles: C and T.

allele frequency — The proportion of a specific gene variant among all copies of the gene in the population.

alternate locus — A sequence that provides an alternate representation of a locus. Alternate loci are collected into additional assembly units (i.e. not in the primary assembly).

Alu — The *Alu* repeat family comprises short interspersed elements (SINES) present in multiple copies in the genomes of humans and other primates. The *Alu* sequence is approximately 300 bp in length and is found commonly in introns, 3' untranslated regions of genes, and intergenic genomic regions. They are mobile elements and are

present in the human genome in extremely high copy number. Almost one million copies of the *Alu* sequence are estimated to be present, making it the most abundant mobile element. The *Alu* sequence is so named because of the presence of a recognition site for the *Alu*I endonuclease in the middle of the *Alu* sequence. Because of the widespread occurrence of the *Alu* repeat in the genome, the *Alu* sequence is used as a universal primer for PCR in animal cell lines; it binds in both forward and reverse directions.

amino acid — Basic building block molecules of peptides and proteins. The sequence of amino acids in a protein is determined by the RNA codon sequence.

anchor sequence — Anchor sequences are molecular markers that are unique loci in genetic linkage maps of multiple species and are used in comparative genomics for cross-species mapping along co-linear genomic regions.

API

Application Programming Interface. An API is a set of routines, data structures, variables, constants and/or classes for building software applications. APIs define how software components communicate with one another. For instance, for computers running a graphical user interface, an API manages an application's windows, icons, menus, and dialog boxes.

ASN.1

Abstract Syntax Notation 1 is an international standard data-representation format used to achieve interoperability between computer platforms. It allows for the reliable exchange of data in terms of structure and content by computer and software systems of all types.

assembly

A set of chromosomes, unlocalized and unplaced (random) sequences and alternate loci used to represent an organism's genome. Assemblies are constructed from one or more assembly units. Most current assemblies are a haploid representation of an organism's genome, although some loci may be represented more than once (see alternate locus). This representation may be obtained from a single individual (e.g. chimp or mouse) or multiple individuals (e.g. human reference assembly). Except in cases of organisms which have been bred to homozygosity, the haploid assembly does not typically represent a single haplotype, but rather a mixture of haplotypes. A diploid genome assembly is a chromosome assembly that is available for both sets of an individual's chromosomes. It is anticipated that a diploid genome assembly is representing the genome of an individual. Therefore it is not anticipated that alternate loci will be defined for this assembly, although it is possible that unlocalized or unplaced sequences could be part of the assembly. An assembly is constructed from one or more assembly units.

assembly release

Release of a genome assembly. A major release is any update that changes the sequence and/or changes the chromosome coordinate system defined in the primary assembly. A

minor release is an update that does not change the coordinate system, but may add or modify information. Such events include addition of genome patches, assignment of unplaced sequences to a chromosome, or new placements for alternate loci.

assembly unit

The collection of sequences used to define discrete parts of an assembly. All assemblies must contain one assembly unit that represents the "primary assembly".

asserted position

A statement (assertion) based on experimental evidence that a variant is located at a particular position. Since asserted positions are based on experimental evidence, they cannot be seen as a conformation of the variant's position even if there are multiple claims by different submitters of a specific position for a particular variant.

NCBI does not independently verify assertions and cannot endorse their accuracy.

Note: Submissions based on asserted positions should reference a sequence accession that is part of an assembly represented in the NCBI [Assembly Resource](#). If no assembly is available however, the reference sequence can be an INSDC sequence accession. If a submission asserts a position for a variation using an accession that cannot be aligned to an assembly, the rs for that variation cannot be annotated to the human assembly, and therefore will not appear on maps or graphic representations of the assembly.

BAC

Bacterial Artificial Chromosome. The BAC cloning system is based on a bacterial plasmid vector which is capable of carrying a large segment of genomic DNA (100–300 bp) for cloning in bacteria. BACs are used in the construction of complex genomic libraries because of their high cloning efficiency and stability of the cloned DNA.

backend

With reference to web applications, the backend refers to the components (server, application, and database) not directly accessed by the user.

BAF

B-Allele Frequency

base sequence

The sequence of purines and pyrimidines in nucleic acids and polynucleotides. It is also called nucleotide sequence.

bioinformatics

Bioinformatics is an interdisciplinary field that applies computational approaches for the collection, storage, manipulation, and analysis of biological data including large datasets, to make biological discoveries or predictions. At a minimum, it encompasses computer

science, biology, genetics, genomics, statistics, mathematics and engineering to interpret biological data. It is closely related to computational biology.

BLAST

Basic Local Alignment Search Tool (Altschul et al. 1990). A sequence comparison algorithm that is used to search sequence databases for optimal local alignments to a query. See the BLAST chapter.

BLASTN

nucleotide–nucleotide BLAST. BLASTN takes nucleotide sequences and compares them against the NCBI nucleotide databases.

BLASTP

protein–protein BLAST. BLASTP takes protein sequences and compares them against the NCBI Protein databases.

BLASTX

BLASTX is an application that searches a nucleotide query against a protein database, dynamically translating the query in all six frames.

BLOB

Binary Large OObject. BLOB refers to a large piece of data, such as a bitmap. A BLOB is characterized by large field values, an unpredictable table size, and data that are formless from the perspective of a program. It is also a keyword designating the BLOB structure, which contains information about a block of data.

Boolean

This term refers to binary algebra that uses the logical operators AND, OR, XOR, and NOT; the outcomes consist of logical values (either TRUE or FALSE).

byte

In computer terms, a unit of storage that is equal to 8 bits.

CD

Conserved Domain. CD refers to a domain (a distinct functional and/or structural unit of a protein) that has been conserved during evolution. During evolution, changes at specific positions of an amino acid sequence in the protein may have occurred in a way that preserve the physico-chemical properties of the original residues, and hence the structural and/or functional properties of that region of the protein.

CDD

Conserved Domain Database. This database is a collection of sequence alignments and profiles representing protein domains conserved during molecular evolution.

cDNA

complementary DNA. A DNA sequence obtained by reverse transcription of a messenger RNA (mRNA) sequence.

CDS

Coding region, coding sequence. CDS refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon, inclusively, if complete. A partial CDS lacks part of the sequence (it may lack either or both the start and stop codons). Successful translation of a CDS results in the synthesis of a protein.

CGI

Common Gateway Interface. A mechanism that allows a Web server to run a program or script on the server and send the output to a Web browser.

chip

See DNA chip.

chromosome

The threadlike structure comprised of DNA and protein contained within the nucleus of eukaryotic cells and containing the hereditary material or genes; in prokaryotes, the circular DNA that carries the genetic information.

clinical assertion

A statement (assertion) based on experimental evidence that a variant has a clinical phenotype. Clinical assertions submitted with a variant may or may not be specific as to the nature of the associated phenotype. Since clinical assertions are based on experimental evidence, they cannot be seen as a confirmation of a clinical phenotype, even if there are multiple claims by different submitters of a specific clinical phenotype for a particular variant.

Clinical assertions can fall into one of the following categories:

- Pathogenic
- Probably Pathogenic
- Probably Non-pathogenic
- Non-pathogenic [benign]
- Affecting Drug Response
- Affecting Histocompatibility
- Unknown
- Untested
- Other

As assertion categories may change, see [ClinVar](#) for up-to-date assertion definitions.

Example: For [rs report for rs328](#), the asserted clinical significance for the cluster is clearly stated at the top of the report as well as in the “Allele” subsection.

For more information regarding clinical assertions, see the [clinvar.vcf.gz](#) section of “Human Variation Sets in VCF Format” or the [FAQ for NCBI Variation Resources](#).

Note: NCBI does not independently verify assertions and cannot endorse their accuracy. Information obtained through this resource is not a substitute for professional genetic counseling and is not intended for use as the basis of medical decision making.

CLOB

Character Large OObject

clone

A clone can be considered a self-replicating system containing a DNA fragment of interest.

cloning vector

A small DNA molecule which is capable of autonomous replication within a host cell and is used to carry a fragment of genomic DNA or cDNA to be cloned; usually a bacterial plasmid or modified bacteriophage genome.

cluster

A group that is created based on certain criteria. For example, a gene cluster may include a set of genes whose expression profiles are found to be similar according to certain criteria, or a cluster may refer to a group of clones that are related to each other by homology.

CMS

Content Management System

codon

Sequence of three nucleotides in DNA or mRNA that specifies a particular amino acid during protein synthesis; also called a triplet. Of the 64 possible codons, 3 are stop codons, which do not specify amino acids.

coding region

It is the sequence of DNA that is translated into protein and includes an initiation codon and a termination codon.

complementary DNA

See cDNA.

computational biology

Computational biology involves the development and application of data-analytical and theoretical methods, algorithms, mathematical modeling and computational simulation techniques to the understanding of biological systems and to make predictions and discoveries from biological data, including large datasets. The field has its origins in computer science, applied mathematics, statistics, biophysics, genomics, molecular biology, and many areas of biology. It is closely related to bioinformatics.

consensus sequence

A representative or most typical nucleotide or amino acid sequence in which each nucleotide or amino acid is most often found at its respective position in the group of related sequences.

conserved domains

A conserved domain of a protein is a discrete three-dimensional independently folding structure that is comprised of one or more protein sequence motifs. Protein sequence motifs are conserved amino acid sequences that are a combination of secondary structures (example, helix-loop-helix) which have been shown to be important for protein function (Koonin and Galperin 2003b).

contig

A contiguous sequence generated from determining the non-redundant path along an ordered set of component sequences. A contig should contain no gaps.

CSS

Cascading Style Sheets (CSS) specify the formatting details that control the presentation and layout of HTML and XML elements. CSS can be used for describing the formatting behavior and text decoration of simply structured XML documents but cannot display structure that varies from the structure of the source data.

Cubby

A tool of Entrez, the Cubby was used to store search strategies that could be updated as well as LinkOut preferences to specify which LinkOut providers should be displayed in PubMed, and change the default document delivery service. It has been superceded by MyNCBI.

CUI

Concept Unique Identifier

cytogenetics

A sub discipline of genetics that deals with cytological and molecular analysis of chromosomes—their cellular location, structure, function, and abnormalities.

DAC

Data Access Committee

daemon

A computer program that runs as a background process or service and is not controlled by the user.

DAR

Data Access Request

database

Store of a set of logically related data or collection of files amenable to retrieval by scripts or computer.

dataset

Permanent store of an organized collection of data, for sharing, redistribution, processing, and analysis.

dbGSS

Genome Survey Sequences Database, a division of GenBank for genome sequences.

DDBJ

DNA Data Bank of Japan, a DNA nucleotide sequence collection center and member of INSDC.

DDD

Digital Differential Display, a feature of Unigene that allows analysis of EST expression profiles.

deletion variant

Type of mutation involving the removal of a single nucleotide or segment of DNA.

deoxyribonucleic acid

See DNA.

digital differential display

See DDD.

DNA

Deoxyribonucleic acid is the chemical inside the nucleus of a cell that carries the genetic instructions for making living organisms. DNA is composed of two anti-parallel strands, each a linear polymer of nucleotides. Each nucleotide has a phosphate group linked by a phosphoester bond to a pentose (a five-carbon sugar molecule, deoxyribose), that in turn is linked to one of four organic bases, adenine, guanine, cytosine, or thymine, abbreviated

A, G, C, and T, respectively. The bases are of two types: purines, which have two rings and are slightly larger (A and G); and pyrimidines, which have only one ring (C and T). Each nucleotide is joined to the next nucleotide in the chain by a covalent phosphodiester bond between the 5' carbon of one deoxyribose group and the 3' carbon of the next. DNA is a helical molecule with the sugar–phosphate backbone on the outside and the nucleotides extending toward the central axis. There is specific base-pairing between the bases on opposite strands in such a way that A always pairs with T and G always pairs with C.

DNA chip

A DNA chip (also referred to as a DNA microarray) is an organized arrangement of DNA sequences on a solid surface in a 2-dimensional (2D) or 3D manner, either covalently or non-covalently bound to the surface. Arrays contain oligonucleotide probes or short nucleotide “known” sequences that can be used to hybridize to sequences in sample for various applications such as measuring the level of gene expression or identifying a particular mutation of interest.

DOI

Digital Object Identifier, an international standard for persistent, actionable, interoperable identifiers that can be applied to objects such as publications.

DTD

Document Type Definition. The DTD is an optional part of the prolog of an XML document that defines the rules of the document. It sets constraints for an XML document by specifying which elements are present in the document and the relationships between elements, e.g., which tags can contain other tags, the number and sequence of the tags, and attributes of the tags. The DTD helps to validate the data when the receiving application does not have a built-in description of the incoming data.

DUC

Data Use Certification

E-utilities

Structured interface to the NCBI Entrez query and database system via 9 server-side programs: EInfo (database statistics), ESearch (text searches), EPost (UID uploads), ESummary (document summary downloads), EFetch (data record downloads), ELink (Entrez links), EGQuery (global query), ESpell (spelling suggestions), ECitMatch (batch citation searching in PubMed).

EMBL — European Molecular Biology Laboratory

ENA — European Nucleotide Archive at European Molecular Biology Laboratory (EMBL)

end sequence — A sequence obtained from the unidirectional sequencing of a genomic clone insert. A set of paired end sequences can be generated if the insert is sequenced from either end.

Entrez — Entrez is a retrieval system at NCBI for searching several linked databases, such as PubMed, GenBank, and PMC. See the Entrez chapter.

epigenomics — The study of changes in the expression or repression of genes by epigenetic mechanisms such as DNA methylation or histone modification that are not a result of changes in the DNA base sequence.

eQTL — expression Quantitative Trait Loci

EST — Expressed Sequence Tag. ESTs are short (usually approximately 300–500 base pairs), single-pass sequence reads from cDNA. Typically, they are produced in large batches. They represent the genes expressed in a given tissue and/or at a given developmental stage. They are tags (some coding, others not) of expression for a given cDNA library. They are useful in identifying full-length genes and in mapping.

eukaryotic — Referring to organisms with cells having a true nucleus bounded by a nuclear membrane.

exon — Refers to the portion of a gene that encodes for a part of that gene's mRNA. A gene may comprise many exons, some of which may include only protein-coding sequence; however, an exon may also include 5' or 3' untranslated sequence. Each exon codes for a specific portion of the complete protein. In some species (including humans), a gene's exons are separated by long regions of DNA (called introns or sometimes “junk DNA”) that often have no apparent function but have been shown to encode small untranslated RNAs or regulatory information.

expressed sequence tag — See EST.

FASTA — The first widely used algorithm for similarity searching of protein and DNA sequence databases. The program looks for optimal local alignments by scanning the sequence for small matches called “words”. Initially, the scores of segments in which there are multiple word hits are calculated (“init1”). Later, the scores of several segments may be summed to generate an “initn” score. An optimized alignment that includes gaps is shown in the output as “opt”. The sensitivity and speed of the search are inversely related and controlled by the “k-tup” variable, which specifies the size of a “word” (Pearson and Lipman 1988). Also refers to a format for a nucleic acid or protein sequence.

FLAN — FLu Annotation

FlyBase — [FlyBase](#) is the primary database of genetic and genomic data for the insect family Drosophilidae.

frameshift — A mutation in which the number of nucleotides inserted or deleted from a protein coding sequence of DNA is not a multiple of 3, which results in a shift in the codon reading frame, creating an altered protein product.

frontend — With reference to web applications, the frontend refers to the interface which is directly accessible to the user through which other components such as databases and servers can be accessed.

FTP — File Transfer Protocol. A method of retrieving files over a network directly to the user's computer or to his/her home directory using a set of protocols that govern how the data are to be transported.

gap — A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acids is also penalized in the scoring of an alignment.

GB — Gigabytes; 10^9 bytes.

Gbps — Gigabits per second. Refers to the speed of data transfer.

gene expression — It is the process by which a gene is modulated for transcription to mRNA and translation to a protein. It can be measured by the levels of mRNA, protein, or observable phenotype of the cell.

gene frequency — See allele frequency.

gene trap — Gene trapping is a technique for determining gene function whereby specialized vector sequences are randomly inserted into the genome and can be used to tag genes, allowing for genetic or phenotypic approaches to study the effect of the mutation.

genetic code — The instructions in a gene that tell the cell how to make a specific protein. A, T, G, and C are the “letters” of the DNA code; they stand for the chemicals adenine, thymine, guanine, and cytosine, respectively, that make up the nucleotide bases of DNA. Each gene's code combines the four chemicals in various ways to spell out three-letter “words” that specify which amino acid is needed at every position for making a protein.

genetic recombination — Genetic recombination is the process by which DNA is broken and rejoined resulting in new arrangements, such as by crossing over of chromosomes during meiosis or by chromosomal exchange during genetic conjugation, transduction, or transformation.

genetic testing — The analysis of an individual's genome for determining the presence of a mutation, carrier status of a mutation, disease risk, or relationship to other individuals.

genome — The genome is the complete genetic material of an organism. For eukaryotic organisms, it is the DNA in all chromosomes and in mitochondria or chloroplasts; for prokaryotes, it includes the circular double-stranded DNA molecule. For viruses, it comprises DNA or RNA.

genome assembly — See assembly.

genome library — A DNA library which includes the complete sequences from the genome of an organism, i.e., introns and exons.

genomics — A field of study in genetics that applies molecular tools such as recombinant DNA technology and high-throughput sequencing, and bioinformatics approaches such as genome alignment and assembly towards the analysis of genome structure and function.

genotype — The genetic identity of an individual that does not show as outward characteristics. The genotype refers to the pair of alleles for a given region of the genome that an individual carries.

GFF — General Feature Format; it is used for the annotation of biological sequences

gnomon — Gene model prediction program. See <http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml>

haplotype — A haplotype is a set of DNA variants, SNPs, or other polymorphisms that are closely located on the same chromosome and are thus inherited together.

HapMap — Haplotype map, a now retired tool for finding genes and genetic variations that affect health and disease.

HGNC — HUGO Gene Nomenclature Committee

HGP — Human Genome Project

HIPAA — Health Insurance Portability and Accountability Act of 1996

HLA — Human Leukocyte Antigen

HMM — Hidden Markov Model

HMP — Human Microbiome Project

HomoloGene — HomoloGene is a system that automatically detects homologs, including paralogs and orthologs, among the genes of several completely sequenced eukaryotic genomes.

homology — See homologous.

homologous — The term refers to similarity attributable to descent from a common ancestor. Homologous chromosomes are members of a pair of essentially identical chromosomes, each derived from one parent. They have the same or allelic genes with genetic loci arranged in the same order. Homologous chromosomes synapse during meiosis.

HPO — Human Phenotype Ontology

HUGO — Human Genome Organization

HUP — Hold-until-published

initiation codon — The canonical AUG codon or any alternative non-canonical start codon in the messenger RNA which serves as the recognition codon for binding of N-formylmethionyl transfer RNA (tRNA^{fmet}) and addition of the first methionine during the initiation of protein translation.

INSDC — International Nucleotide Sequence Database Collaboration

insert sequence — A piece of DNA inserted into a cloning vector by recombinant DNA techniques. For genomic cloning vectors, insert sequences typically range in size from 10's of kilobases (cosmids, fosmids), to 100's of kilobases (BAC, PACs) up to ~2MB (YACs).

intergenic — Regions of the genome that lie between genes and have no known function. Intergenic DNA is mainly made up of 2 types of repeated sequences: interspersed repeats and tandemly repeated DNA.

intervening sequence — See intron.

intron — Non-coding region of the DNA that gets transcribed in the primary messenger RNA but the sequence is removed from the mature RNA transcript when exons are spliced together.

IRB — Internal Review Board

ISSN — An ISSN is an 8-digit code used to identify newspapers, journals, magazines, periodicals, and electronic and print media. See <http://www.issn.org>

JATS — Journal Article Tag Suite, a NISO DTD Standard.

JPEG — Joint Photographic Experts Group; suffix for digital image file format.

locus — The position on a chromosome where a gene is located.

long repeat — See long terminal repeat.

long terminal repeat — Retroviruses integrate a reverse-transcribed double-stranded DNA copy of their RNA genome into host DNA. The human genome contains many copies of endogenous retroviral DNA sequences integrated in the genome. The viral DNA is flanked by long terminal repeat sequences (LTR) that contain regulatory elements, signal sequences, transcription factor binding sites, and polyadenylation signals that play a role in modulating gene expression.

LTR — Long Terminal Repeat

mapping — In genomics, mapping refers to the various techniques for determining the position and relative order of markers or genes (loci) on a chromosome and relative distance between them based on recombination frequency (genetic map), the absolute position of genes and the distance between them in nucleotide base pairs (physical map), or the position of markers or genes on the chromosome based on hybridization (cytogenetic map).

MapView — MapViewer is a genome browsing tool used to view and search an organism's genome and display chromosome maps.

MathML — Mathematical Markup Language; it is used for handling mathematical equations in XML.

MD5 — MD5 (Message-Digest Algorithm 5) is a database hash function used to validate data integrity.

MedGen — [MedGen](#) is an NCBI resource providing information related to human medical genetics, such as attributes of conditions with a genetic contribution.

MeSH — Medical Subject Headings is the National Library of Medicine's controlled vocabulary thesaurus used for indexing articles in PubMed.

metagenome — A metagenome is a collective genome representative of the community of organisms, for example, microorganisms, many of which cannot be cultivated outside of their environment.

MHC — Major Histocompatibility Complex

MIAME — Minimum Information About a Microarray Experiment

microarray — See DNA chip.

My NCBI — My NCBI is a tool that retains user information and database preferences to provide customized services for many NCBI databases.

N50 — The N50 value is the size of the smallest contig (or scaffold) such that 50% of the genome is contained in contigs of size N50 or larger. It is a statistic used in genome assemblies.

NCBI — National Center for Biotechnology Information

NIHMS — National Institutes of Health Manuscript Submission system

NISO — National Information Standards Organization

NLM — National Library of Medicine

OAI — Open Archives Initiative

OMIM — [Online Mendelian Inheritance in Man](#)

open access — Open access refers to online publications, often publicly funded research output, that have no restrictions on access (e.g., no user fees) and are not subject to restrictions on use (license and copyright).

open reading frame — Sequence of DNA that begins with an initiation codon and ends with a termination codon, specifying a gene sequence.

opposite strand — It is the DNA strand facing the template strand.

ORF — Open Reading Frame

ortholog — Orthologous genes are found in different species and arise from a single genomic locus in a common ancestor. Orthologs may not have a similar function.

overlapping gene — A gene that shares part or all of its nucleotide sequence with another gene which may include regulatory elements or intron sequence.

p-value — The p-value or probability value is a measure of statistical significance. If the null hypothesis is assumed to be true, the p-value indicates the probability of observing a particular outcome or result. The lower the p-value, the lower the probability of a result occurring by chance and therefore, greater the significance. A p-value of 0.05 indicates that there is a 5% probability of a chance outcome. For historical reasons, a p-value of 0.05 has been used as a cutoff for significance. Values less than 0.05 are considered significant. If the p value is 0.001 or less (less than a 0.1% probability that the results occurred by chance), the result is seen as highly significant.

PAC — P1-derived artificial chromosome; cloning system for large DNA inserts (100-300 kb) of genomic DNA which is based on a combination of the BAC cloning system and bacteriophage P1 vector.

pairwise alignment — Alignment of the two protein or nucleic acid sequences to determine regions of similarity that may reveal structural or functional relationships.

PDB — Protein Data Bank

paralog — Paralogs are homologous genes within a species that arose from a duplication event.

PDF — Portable Document Format

pedigree — In genetics, a pedigree is a multigenerational family hierarchy tree with symbol convention used to depict inheritance of normal or disease traits by individuals in the tree.

PheGenI — [Phenotype Genotype Integrator](#), a resource for phenotypes compiled by merging NHGRI genome-wide association study (GWAS) catalog data with data from NCBI databases including Gene, dbGaP, OMIM, GTEx and dbSNP.

phenotype — The observable characteristics or features of a living organism.

phylogenetic tree — An evolutionary tree for organismal species or cellular macromolecules (example tRNA) that is built using inheritance or molecular sequence information.

PIR — [Protein Information Resource](#). A free resource of protein databases and analysis tools.

PI — Principal Investigator

PLink — PLink is a free, open-source whole genome association analysis toolset.

PMCI — PubMed Central International

PMID — PubMed Identifier; a unique identifier for each PubMed record.

pNIHMS — Portable NIHMS

Polyadenylation signal — Polyadenylation is a post-transcriptional modification to the 3' end of eukaryotic mRNA involving the addition of a sequence of ~250 adenosine nucleotides in a template-independent manner, and occurs about 30 base-pairs downstream from a short signal sequence in the primary transcript, typically AAUAAA.

Polyprotein — Precursor protein that is enzymatically processed to form the mature protein.

ProtEST — A view in the Unigene browser for comparing proteins to the EST cDNA sequences.

Pseudogene — An altered copy of the original gene, which may either arise from reverse transcription of mRNA followed by integration of the double-stranded cDNA into the chromosome at a break event (processed pseudogene) or by a gene duplication event (unprocessed pseudogene). For a long time, pseudogenes were thought to be non-functional transcriptionally and translationally, but new roles are emerging for pseudogenes as regulatory modulators.

PSI-BLAST — Position-Specific Iterative BLAST (PSI-BLAST) is an iterative search using the protein BLAST algorithm in which the amino acid frequency at each position determination built after the initial search is then used in subsequent searches. The process may be repeated, if desired, with new sequences found in each cycle used to refine the profile (Altschul et al. 1997).

PSSM — Position-Specific Scoring Matrix; a type of scoring matrix used in protein BLAST searches providing position-specific amino acid substitution scores for each position in a protein multiple-sequence alignment.

public access — See open access.

Pubreader — A viewer for reading books and journal articles at NCBI on tablet devices.

QA — See quality assurance.

QC — Quality Control

quality assessment — An assessment of quality is a part of quality assurance.

quality assurance — Standardized process designed and undertaken to avoid mistakes or errors in a released product.

quality control — A set of procedures that are performed to ensure that a product meets a specified standard.

query — Query refers to the term used in the search.

query translation — The full search expression including MeSH expansion and automatic term mapping, shown in the details box in Entrez search results.

reading frame — See open reading frame.

recombinant — Referring to the product of recombination, either DNA or gene or protein.

recombination — Recombination results from crossing-over events involving exchange of DNA sequences between structurally similar chromosomes during meiosis in the diploid cell. It is a process whereby new gene combinations are formed in the progeny. It can also be performed enzymatically on DNA *in vitro*.

reference sequence — See RefSeq.

RefSeq — RefSeq, NCBI's Reference Sequence project, is a non-redundant, annotated set of sequences that serve as reference standards. They are derived from the INSDC databases and include chromosomes, complete genomes (plasmids, organelles, viruses, archaea, bacteria, and eukaryotes), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.

RefSeqGene — A subset of RefSeq, RefSeqGene defines genomic sequences to be used as reference standards for well-characterized genes. It provides more stable gene-specific genomic sequence for each gene including upstream and downstream flanking regions, and versioning information for conversion of coordinates in case of updates.

refSNP — In dbSNP, variant information that results from aggregation of submission data by location on the genome and type of variation is assigned a refSNP (rs) identifier. The rs identifier is used as a reference for that variant location, but does not indicate the explicit sequence change at a location.

RepeatMasker — RepeatMasker is a program that analyzes a query sequence for repeat sequences, creating an output showing the annotation of the repeats as well as a modified query sequence that masks the annotated repeats.

RH map — A Radiation Hybrid (RH) map is obtained by fusing irradiated human cells with rodent cells to create hybrids. The radiation causes chromosomal breakage in a dose-dependent manner. Following fusion with the rodent cell line, the chromosomal fragments get integrated into the rodent chromosomes. The collection of hybrid cells forms a panel and can be used for mapping.

RPSBLAST — Reverse-Position-Specific BLAST (RPSBLAST) is a tool that is used to search a protein query against a database of PSSMs that were usually produced by PSI-BLAST.

scaffold — A scaffold is an ordered and oriented set of contigs. It can contain gaps, but there is typically some evidence to support the contig order, orientation, and gap size estimates.

Schema — Representation

SEF — Serials Extract File, required for Medline indexing.

sequence masking — In determining similarities between homologous sequences, it is sometimes necessary to exclude non-specific or non-homologous similarities. This is done using standard techniques by which sequences which are of low complexity or short-period tandem repeat sequences are masked.

sequence pair — Two aligned component sequences used in the generation of a contig.

sequence tagged site — See STS

SGD — Saccharomyces Genome Database. A database for the molecular biology and genetics of *Saccharomyces cerevisiae*, also known as baker's yeast.

SGML — Standard Generalized Markup Language. The international standard for specifying the structure and content of electronic documents. SGML is used for the markup of data in a way that is self-describing. SGML is not a language but a way of defining languages that are developed along its general principles. A subset of SGML called XML is more widely used for the markup of data. HTML (Hypertext Markup Language) is based on SGML and uses some of its concepts to provide a universal markup language for the display of information and the linking of different pieces of that information.

similarity score — Quantitative measure of the similarity between two sequences.

single nucleotide polymorphism — See SNP

small RNA — Refer to snRNA? [CHECKING WITH thibaudf]

small nuclear RNA — See snRNA

snRNA — snRNAs (small nuclear RNAs) are small non-coding RNAs that are localized to the nucleus and that play a role in splicing of introns from premature transcripts.

SNOMED-CT — Systematized Nomenclature of Medicine–Clinical Terms. Vocabulary of clinical terminology, maintained by [International Health Terminology Standards Development Organisation](#) (IHTSDO).

SNP — Single Nucleotide Polymorphism. Common, but minute, variations that occur in human DNA at a frequency of 1 every 1,000 bases. An SNP is a single base-pair site within the genome at which more than one of the four possible base pairs is commonly found in natural populations. Over 10 million SNP sites have been identified and mapped on the sequence of the genome, providing the densest possible map of genetic differences. SNP is pronounced “snip”.

source file — A source file refers to the original file or the file provided by the submitter.

Spidey — A software program used to align cDNA (spliced mRNA sequences) to genomic sequences (Wheelan et al. 2001).

splice form — See splice site

splice signal — See splice site.

splice site — Refers to the location of the exon-intron junctions in a pre-mRNA (i.e., the primary transcript that must undergo additional processing to become a mature RNA for translation into a protein). Splice sites can be determined by comparing the sequence of genomic DNA with that of the cDNA sequence. In mRNA, introns (non-protein coding regions) are removed by the splicing machinery; however, exons can also be removed. Depending on which exons (or parts of exons) are removed, different proteins can be made from the same initial RNA or gene. Different proteins created in this way are “splice variants” or “splice forms” or “alternatively spliced”.

Splign — A software program comprising a set of algorithms for computing cDNA-to-Genome alignments (Kapustin et al. 2008).

Splitd — NCBI queuing system for processing BLAST requests.

SQL — Structured Query Language; a programming language for managing and processing data in relational databases.

ssRNA — single-stranded RNA; viruses such as Ebola virus and Marburg virus of the Family Filoviridae and Dengue virus and Yellow fever virus of the Family Flaviviridae have a single stranded RNA genome.

structural variation — Genomic structural variation includes insertions, deletions, duplications, inversions, or chromosomal translocations longer than 50 bp. These variants can occur in coding or noncoding DNA and they can be inherited or arise sporadically in the germline or somatic cells. Some of these variants may be benign, with or without phenotypic manifestations whereas others result in disease, for example, [22q11.2 Deletion Syndrome](#).

structured output — Results written to a file in a way that they conform to an external schema or set of rules, permitting validation and machine-readability.

STS — Sequence Tagged Site. A short DNA segment that occurs only once in the human genome, the exact location and order of bases of which are known. Because each is unique, STSs are helpful for chromosome placement of mapping and sequencing data from many different laboratories. STSs serve as landmarks on the physical map of the human genome.

style sheet — A style sheet is a mechanism to separate content from presentation and contains information necessary for formatting the text. For example, Cascading Style Sheets (CSS) are used to format HTML pages.

Swiss-Prot — Swiss-Prot is a curated protein sequence database which is a part of the UniProt knowledgebase, UniProtKB. It provides a high level of annotation (such as the description of protein function, nomenclature and taxonomy, structure, domains, sequence, post-translational modifications, variants, publications, etc.) and integration with other databases.

switch point — The switch point is the base at which a contig sequence stops being generated from one component sequence and switches to being generated from the next component sequence. There must be at least one switch point between adjacent component sequences in a contig.

tagserver — The TagServer is a database in PMC that is used for storing metadata about the journal article based on information mined from the article, such as Gene and protein names and identifiers.

TBLASTN — TBLASTN is an application which searches a protein query against a nucleotide database, dynamically translating the database.

term mapping — Untagged terms that are entered in the search box are matched against a series of translation tables (example, for Medical Subject Headings [MeSH], journals, author names etc.) and indexes in a defined order until a match is found. Once the match is found, the mapping process is complete.

termination codon — Canonical or non-canonical codon at which the ribosome is released from the RNA and translation of protein synthesis ends.

termination signal — See termination codon.

tiling path — An ordered list or map that defines a set of overlapping clones that covers a chromosome or other extended segment of DNA.

TPF — Tiling Path Format. A table format used to specify the set of clones that will provide the best possible sequence coverage for a particular chromosome, the order of the clones along the chromosome, and the location of any gaps in the clone tiling path. Also used to refer to a file (Tiling Path File) in which the minimal tiling path of clones covering a chromosome is specified in Tiling Path Format.

traceback — The traceback is the process which converts the High Scoring Segment Pairs (HSPs) and generates alignments in BLAST.

transcriptome — The transcriptome refers to the full set of transcripts in a cell assembled by a method called RNA-seq in which RNA from cells is collected, sampled, and sequenced. It includes alternative splice variants, variants created by alternative transcription initiation and alternative transcription termination, and noncoding RNA genes.

transfer RNA — See tRNA.

translation initiation site — See translation start site.

translation start site — The position within an mRNA at which synthesis of a protein begins. The translation start site is usually an AUG codon, but occasionally, GUG or CUG codons are used to initiate protein synthesis.

translation stop signal — See termination codon.

tree viewer — Graphical representation of a tree hierarchy showing the root, branches, and leaves (nodes).

tRNA — Transfer RNA (tRNA) is a small RNA molecule that plays a role in protein synthesis. Typically tRNAs have highly conserved sequences and four-armed cloverleaf secondary structures formed by base pairing within the tRNA resulting in the formation of hairpin loops. Exceptions to this cloverleaf structure occur in mitochondrial and nematode tRNAs. There are two parts to the tRNA that are involved in protein synthesis: the anticodon that recognizes and binds to the complementary codon in the mRNA transcript; and the amino-acid binding site. The aminoacyl-tRNA synthetase are involved in pairing the amino acid with its cognate tRNA. Amino acids are transferred from the amino-acyl tRNA to the growing peptide chain via the formation of a peptide bond.

tRNAscan-SE — tRNAscan-SE is a program for identifying tRNA genes in DNA sequence.

UID — Identifier for a public record (e.g., publication or sequence) in the NCBI Entrez system.

UMLS — Unified Medical Language System

UniGene — UniGene is a computational system for analyzing the transcriptome, expression of transcripts, and the libraries from which they were derived, allowing evaluation of expression pattern by various parameters such as tissue or health status.

UniProt — Universal Protein Resource (UniProt) is a resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR).

URI — Uniform Resource Identifier (URI)

UTR — Untranslated region (UTR) is the region of mRNA at the 5' or 3' end that does not code for protein and typically includes regulatory sequence.

WAM — Weight Array Method (WAM), a method for sequence analysis (Zhang and Mar3 1993).

WGS — Whole Genome Shotgun; refers to sequencing approach in which the whole genome is fragmented and sequenced using a shotgun approach, and then the sequence of the fragments are reassembled for genome assembly. [CHECK – IMPROVE]

WMM — Weight Matrix Method (WMM), [CHECK, NEED REFERENCE]

XHTML — Extensible Hypertext Markup Language (XHTML)

XML — Extensible Markup Language

XQuery — XML Query; a query language for structured data such as XML.

XSL — Extensible Stylesheet Language

YAC — Yeast artificial chromosomes (YACs) were developed for cloning large fragments of genomic DNA into yeast. YACs can carry large, megabase-size inserts of genomic DNA. BACs or PACs have advantages over YACs. Cloned DNA in the YAC system is more difficult to manipulate and is often chimeric. Also, recombination events in yeast can lead to deletions or rearrangements of the insert DNA, thus YACs are less stable than BACs or PACs.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990 Oct 5;215(3):403–10. PubMed PMID: 2231712.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997 Sep 1;25(17):3389–402. Review. PubMed PMID: 9254694.
- Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, Chen HC, Agarwala R, McLaren WM, Ritchie GR, Albracht D, Kremitzki M, Rock S, Kotkiewicz H, Kremitzki C, Wollam A, Trani L, Fulton L, Fulton R, Matthews L, Whitehead S, Chow W, Torrance J, Dunn M, Harden G, Threadgold G, Wood J, Collins J, Heath P, Griffiths G, Pelan S, Grafham D, Eichler EE, Weinstock G, Mardis ER, Wilson RK, Howe K, Flliceck P, Hubbard T. Modernizing reference genome assemblies. *PLoS Biol.* 2011 Jul;9(7):e1001091. doi: [10.1371/journal.pbio.1001091](https://doi.org/10.1371/journal.pbio.1001091). PubMed PMID: 21750661.
- Genome Reference Consortium (GRC) Assembly Terminology. Available from <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/info/definitions.shtml> [Accessed October 12, 2017].
- Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol Direct.* 2008 May 21;3:20. doi: [10.1186/1745-6150-3-20](https://doi.org/10.1186/1745-6150-3-20). PubMed PMID: 18495041.
- Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003a. Chapter 4, Principles and Methods of Sequence Analysis. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK20261/>

Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003b. Chapter 3, Information Sources for Genomics. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK20256/>

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988 Apr;85(8):2444–8. PubMed PMID: 3162770.

Wheelan SJ, Church DM, Ostell JM. Spidey: a tool for mRNA-to-genomic alignments. Genome Res. 2001 Nov;11(11):1952–7. PubMed PMID: 11691860.

Zhang MQ, Marr TG. A weight array method for splicing signal analysis. Comput Appl Biosci. 1993 Oct;9(5):499–509. PubMed PMID: 8293321.