B I O I N F O R M A T I C S

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

SUPPLEMENTARY CHAPTER: DATA BASES AND MINING

1 What is a biological data base?

1.a Introduction

1.b Types of data bases

1.c Searching data bases

1 What is a biological data base

1.a Introduction

- Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community.
- The completion of a "working draft" of the human genome -an important milestone in the Human Genome Project - was announced in June 2000 at a press conference at the White House and was published in the February 15, 2001 issue of the journal Nature.

The Human Genome Project

genomics.energy.gov Human Genome Project Information • Genomics:GTL • DOE Microbial Genomics • home

Human Genome Project Information

About the HGP | Ethical / Legal Issues | Medicine | Education | Gene Gateway | Research Archive

Sequence Databases | Landmark Papers | Sequence Insights | Related Projects

Landmark HGP Papers

Basic Information
- FAQs
- Glossary
- Acronyms
- Links
- Genetics 101
- Publications
- Meetings Calendar
- Media Guide

About the Project
- What is it?
- Goals
- Landmark Papers
- Sequence Databases
- Timeline

General Human Genome Project Papers

- Mapping and sequencing of structural variation from eight human genomes, *Nature*, May 1, 2008
- Identification and Analysis of Function Elements in 1% of the Human Genome by the ENCODE Pilot Project, *Nature*, June 14 2007
- Finishing the euchromatic sequence of the human genome, *Nature*, Oct. 21, 2004
- Human genome: Quality assessment of the human genome sequence. *Nature*, **429**, 365-368 (27 May 2004)
- Building on the DNA revolution, April 11, 2003 entire issue of *Science* with insights from the completion of the HGP finished sequence
- Double helix at 50, April 24, 2003 entire issue of *Nature* with insights from the completion of the HGP finished sequence
- The human genome Feb.16, 2001, entire issue of *Science* with insights from the completion of the HGP and Celera working draft
- The human genome Feb.15, 2001, entire issue of *Nature* with insights from the completion of the HGP working draft

Spin-offs of the Human Genome Project

International HapMap Project

International HapMap Project

Home | About the Project | Data | Publications | Tutorial

中文 | English | Français | 日本語 | Yoruba

The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "**About the International HapMap Project**" for more information.

Project Information

About the Project
HapMap Publications
HapMap Tutorial
HapMap Mailing List
HapMap Project Participants
HapMap Mirror Site in Japan

Project Data

HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
HapMap Genome Browser (Phase 1 & 2

News

- 2009-02-09: **HapMap3 Phased Haplotypes available**

  Phased haplotypes for consensus HapMap3 release 2 data has been phased for autosomes are now **available for bulk download**.

- 2009-02-06: **HapMap Public Release #27 (merged II+III)**

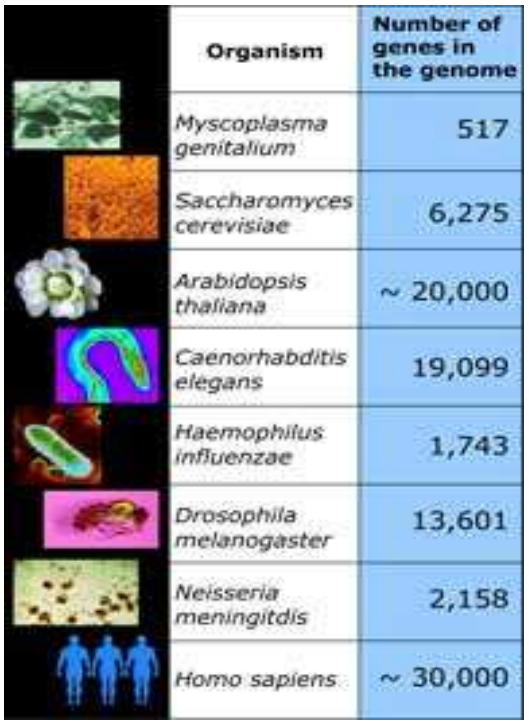
  Genotypes and frequency data for the three phases of the project (I+II: rel #24 and III: release #2), were combined in NCBI build 36 (dbSNP b126) coordinates. Data is **available for downloading** and also **available for browsing**. Click here to read the latest **release notes**.

- 2009-01-07: **HapMap Phase 3 draft 2 release available for download**

  Genotypes and frequency data for phase 3 (NCBI build 36, dbSNP b126) of the HapMap are **available for bulk download**. This dataset will subsequently be merged with phase I+II data, and once merged, the complete dataset

Explosive growth of data

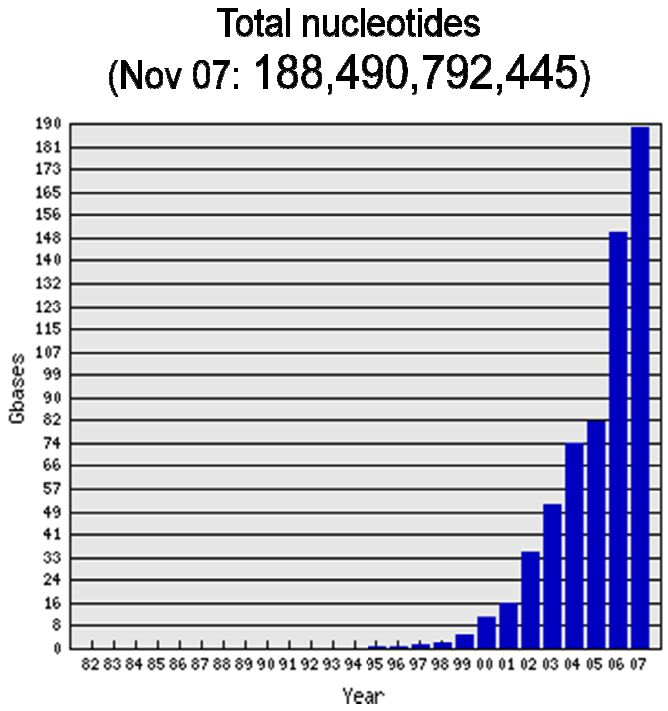
- In particular, advances in biotechnology and sequencing techniques lead to accumulation of biological data:
  - 100's of mammalian genomes
  - SNP chips of 500,000 and above
  - Organism-wide gene expression profiles
  - Proteome snapshots characterizing translation products across time and tissues
  - Modeling of cellular processes and pathways

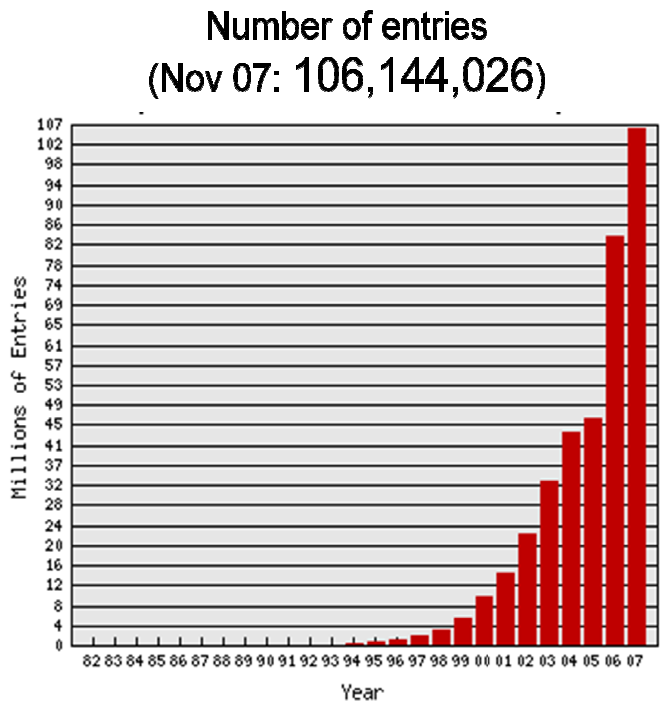


Organism	Number of genes in the genome
Myscoplasma genitalium	517
Saccharomyces cerevisiae	6,275
Arabidopsis thaliana	~ 20,000
Caenorhabditis elegans	19,099
Haemophilus influenzae	1,743
Drosophila melanogaster	13,601
Neisseria meningitdis	2,158
Homo sapiens	~ 30,000

(UIC Bioinformatics Group)

EMBL data base growth





- This has led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.

What is a biological data base?

- *Biological data bases* are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses.
- They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics.
- Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures

What is a biological data base?

Type of databases	Information they contain
Bibliographic databases	Literature
Taxonomic databases	Classification
Nucleic acid databases	DNA information
Genomic databases	Gene level information
Protein databases	Protein information
Protein families, domains and functional sites	Classification of proteins and identifying domains
Enzymes/ metabolic pathways	Metabolic pathways

- A simple database might be a single file containing many records, each of which includes a overlapping “format” of information.

.

Desired properties of data bases

For researchers to benefit from the data stored in a database, two additional requirements must be met:

- easy access to the information
- a method for extracting only that information needed to answer a specific biological question

- Data must be in certain format for the programs to recognize them.
- Every database can have its own format, but some data elements are essential for every database:
  - Unique identifier or accession code
  - Name of depositor
  - Literature reference
  - Deposition date
  - The real data

Biological data bases: some statistics

- More than 1000 different databases
  - 968 databases reported in *The Molecular Biology Database Collection: 2007 update* by Galperin, Nucleic Acids Research, 2007, Vol. 35, Database issue D3-D4
  - Metabase: database of biological databases, http://biodatabase.org/index.php/Main_Page
- Database sizes: <100kB to >100GB (EMBL >500GB)
  - DNA: >100GB
  - Protein: 1GB
  - 3D structure: 5GB
- Update (adding new data) frequency: daily to annually
- Freely accessible (as a rule)

1.b Types of data bases

Primary data bases

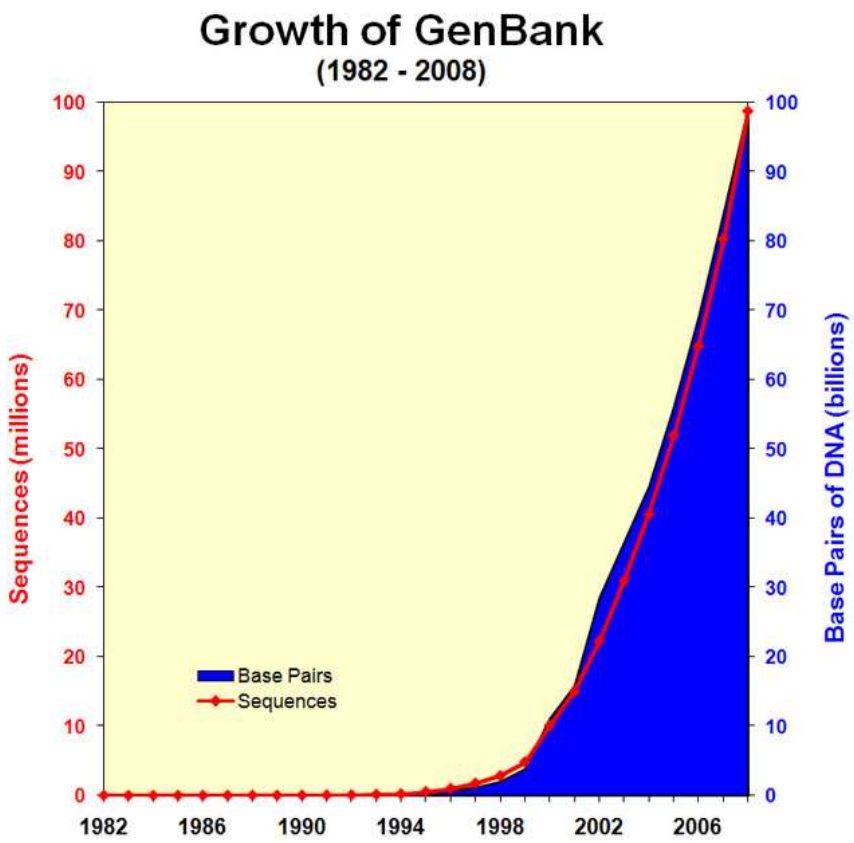
- Real experimental data
- Biomolecular sequences or structures and associated annotation information:
  - organism,
  - function,
  - mutation linked to disease,
  - functional/structural patterns,
  - bibliographic, etc

Examples of primary data bases

- Sequence Information
  - DNA: EMBL nucleotide sequence data base, Genbank, DDBJ
  - Protein: SwissProt, TREMBL, PIR, OWL
- Genome Information
  - GDB, MGD, ACeDB
- Structure Information
  - PDB, NDB, CCDB/CSD

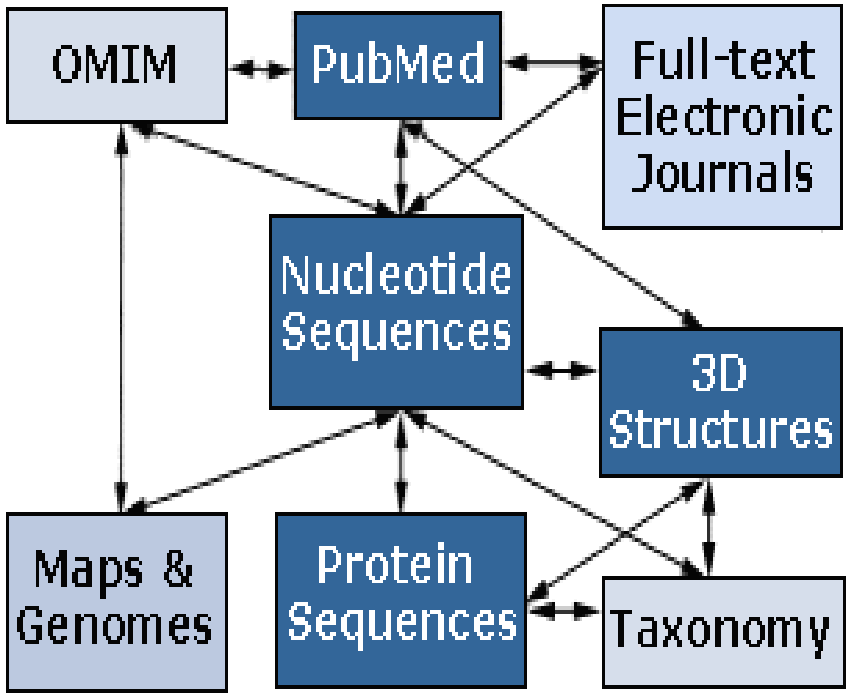
Primary databases in detail: GenBank

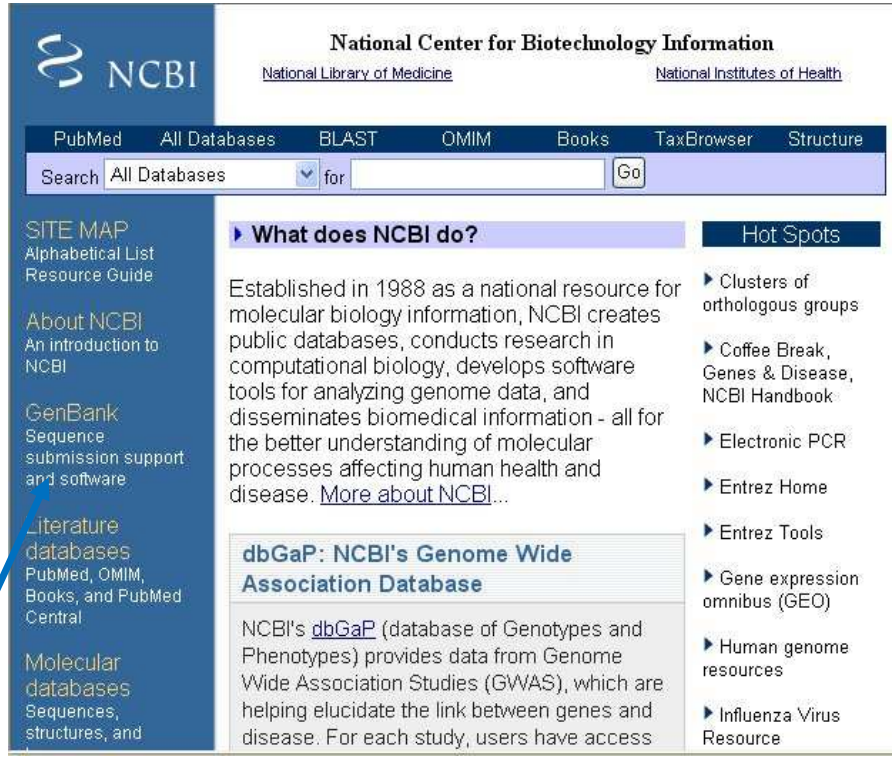
- GenBank is the NIH genetic sequence database
- Genbank is an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2008 Jan; 36(Database issue):D25-30).
- It is connected to other data bases available at NCBI (National Center for Biotechnology Information).



(http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)

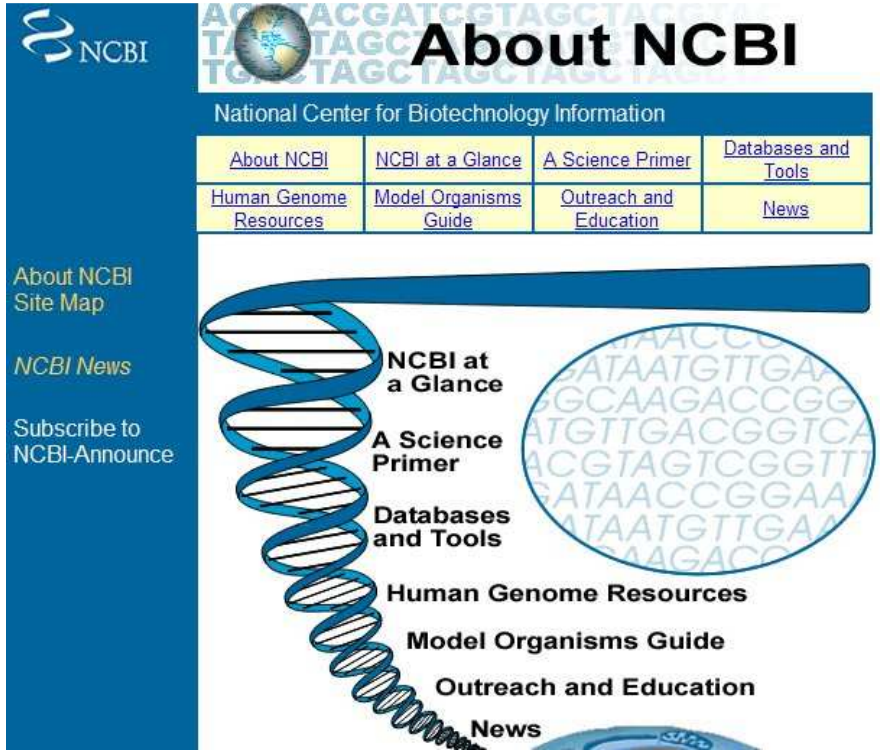
NCBI





(http://www.ncbi.nlm.nih.gov/)

NCBI



http://www.ncbi.nlm.nih.gov/About/

- Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease.

GenBank

NCBI

GenBank Overview

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search Entrez for Go

NCBI Home

NCBI Site Map

Submit to GenBank

Submit an update

Search GenBank

GenBank and RefSeq: a comparison

BLAST

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2008 Jan;36(Database issue):D25-30). There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a *Saccharomyces cerevisiae* gene.

In The News: Platypus Genome

Explore Platypus Genome resources.

- Platypus Genome Project
- Platypus Taxonomic and Sequence Resources
- Platypus Genome Resource Guide
- Duck-Billed Platypus Genome Sequence Published (NIH Press Release)

(http://www.ncbi.nlm.nih.gov/Genbank/index.html)

GenBank sample record

NCBI **Sample GenBank Record**

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
```

NCBI Resource Guide

NCBI **NCBI Resource Guide**

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

*Each link in this **Resource Guide** leads to a **brief description of the resource** on this page, then to the resource itself. A graphical **Site Map** and an **Alphabetical Quicklinks Table** provide direct links to resources and bypass the descriptions.*

RESOURCES BY CATEGORY

About NCBI
programs and services, contact information, NCBI handbook, **news** (what's new, NCBI News, announcements e-mail lists, RSS feeds), exhibit schedule, postdoctoral fellowships, organizational structure, resource statistics, site search

GenBank
overview, submit sequences, submit genomes, sample record, GenBank divisions, statistics, release notes, international collaboration, FTP GenBank

Molecular Databases
nucleotides, proteins, structures, genes, gene expression, taxonomy

Literature Databases
PubMed, PubMedCentral, Journals, OMIM, Books, Citation Matcher

ALPHABETICAL INDEX

with links to resource descriptions

*(To bypass descriptions, use the **Alphabetical Quicklinks Table**.)*

About NCBI	GenBank sample record	Plant Genomes
Announcements	Genes	Protein Sequences
ASN.1	Genes and Disease	PubChem
BankIt	Genomes (data, projects, submissions)	PubMed
BLAST	GENSAT	PubMed Central
BLink	GEO	RefSeq
Books	Glossaries	Research at NCBI
Cancer Chromosomes	Handbook	Retroviruses
CCDS	HIV Interactions	SAGEmap
CDART	HTGs	Science Primer
CDD	HomoloGene	Seminars

(http://www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide.html)

GenBank sample record information

Sample Record - detailed description of each field in a GenBank record.
Includes, for example, information about accession number formats, sequence identifiers (GI number and accession.version), a listing of GenBank divisions, and more. Describes some commonly annotated biological features, such as CDS, and provides links to documents that list and define the complete set of biological features that can be annotated on sequence records. Includes a link to a sequence revision history tool that can be used to track changes that have occurred to the sequence data in a record. Also lists the Entrez search field(s) that can be used to search each part of a sequence record.

GenBank Divisions - summary of GenBank divisions, including abbreviations, full spellings, information about what the GenBank divisions are, and what they are *not*. (This information is part of the GenBank sample record, described above.)

Access GenBank - through Entrez Nucleotides. Search by accession number, author name, organism, gene/protein name, and a variety of other text terms. Additional information about Entrez is below. Use BLAST for sequence similarity searches against GenBank and other databases. An option to download the GenBank full release and updates via FTP is also available.

Growth Statistics (graph) - see also Release Notes sections 2.2.6 (per division statistics), 2.2.7 (per organism statistics), 2.2.8 (growth of GenBank). For statistics on other NCBI databases, please see the page that summarizes sources of Statistics for NCBI Resources.

GenBank Release Notes - A document that accompanies each full release (described in "What is GenBank?", above) of the GenBank database. The release notes describe the format and content of the flat files that comprise the release. They also include notices of recent and upcoming changes, information about GenBank divisions, growth statistics, citing GenBank, and more.

- **Current Release Notes**
- **Past Release Notes**

Genetic Codes - synopsis of 17 genetic codes; used to ensure correct translation of coding sequences in GenBank records.

GenBank Bionet Newsgroup - A moderated list that includes announcements of new GenBank releases, recent and upcoming changes, and discussion among subscribers. For information on how to subscribe by e-mail, see the NCBI Announcements Email Lists page.

(http://www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide.html#SampleRecord)

GenBank sample record information

NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy Structure

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
```

(http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html)

GenBank sample record information

LOCUS

The LOCUS field contains a number of different data elements, including locus name, sequence length, molecule type, GenBank division, and modification date. Each element is described below.

- **Locus Name**

The locus name in this example is SCU49845.

The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries, the last character was one of a series of sequential integers. (See GenBank release notes section 3.4.4 for more info.)

However, the 10 characters in the locus name are no longer sufficient to represent the amount of information originally intended to be contained in the locus name. The only rule now applied in assigning a locus name is that it must be unique. For example, for GenBank records that have 6-character accessions (e.g., U12345), the locus name is usually the first letter of the genus and species names, followed by the accession number. For 8-character character accessions (e.g., AF123456), the locus name is just the accession number.

The RefSeq database of reference sequences assigns formal locus names to each record, based on gene symbol. RefSeq is separate from the GenBank database, but contains cross-references to corresponding GenBank records.

(http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#LocusB)

Statistics at NCBI

NCBI **Statistics for NCBI Resources**

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI Home

Site Map
Resource Guide
Alphabetical List

About NCBI
general and contact information

GenBank
submit your sequence, general information

Molecular Databases
nucleotides, proteins, structures and taxonomy

Literature

- Database Statistics
  - **General tips** for obtaining **Entrez database statistics**
  - Additional statistics web pages for specific databases:
    - Consensus CDS (CCDS) Database
    - dbEST
    - dbGSS
    - dbSNP
    - **GenBank**
    - Gene database
    - Gene Expression Omnibus (GEO)
    - OMIM
    - RefSeq
    - Taxonomy
- Genome Statistics
  - Entrez Genome (database statistics)
  - Statistics for Individual Prokaryotic and Viral Genomes
  - Statistics for Individual Eukaryotic Genomes
- Usage Statistics
  - PubMed Usage

(http://www.ncbi.nlm.nih.gov/Sitemap/Summary/statistics.html#GenBankStats)

Primary databases in detail: dbSNP

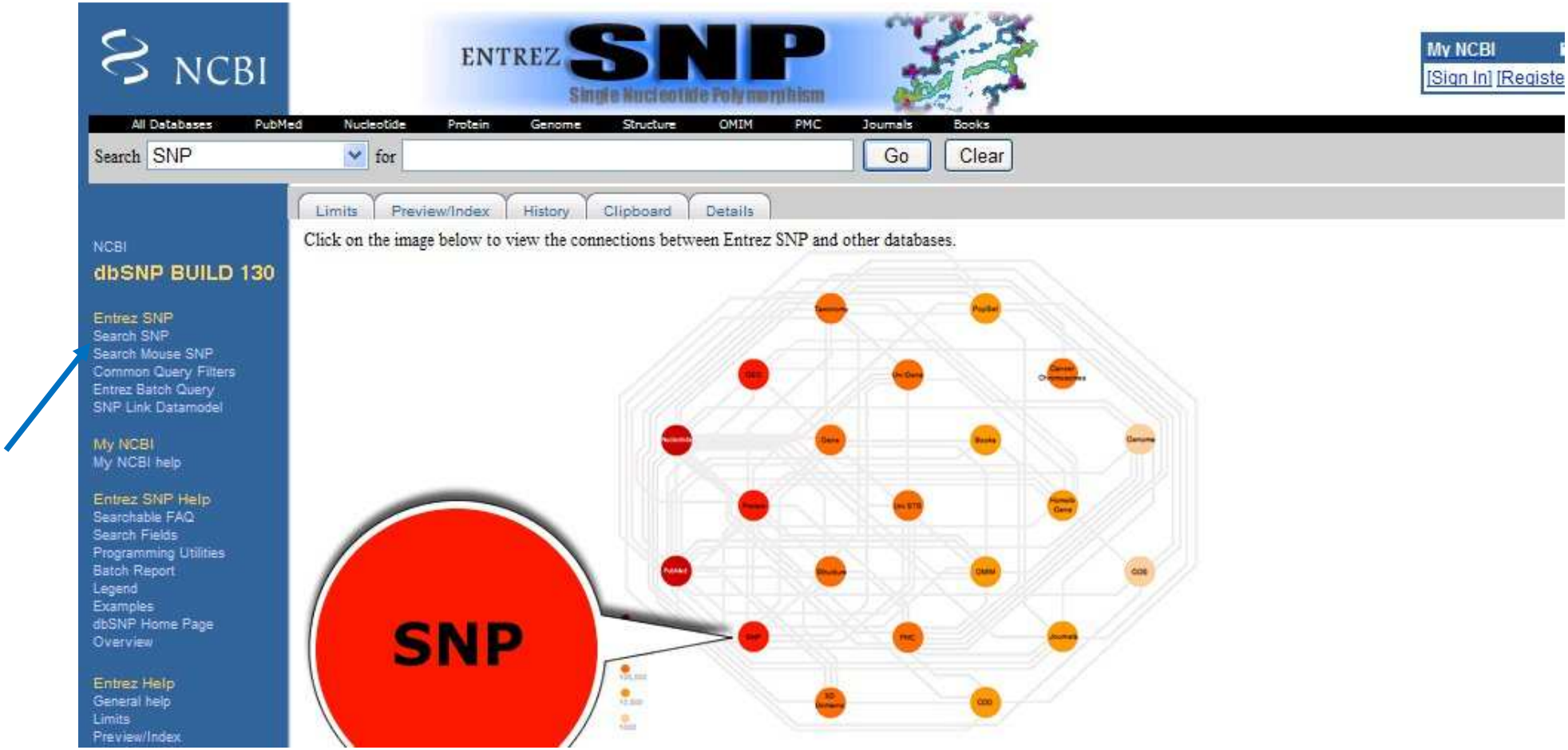
(http://www.ncbi.nlm.nih.gov/projects/SNP/)

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	129	36.3	55,949,131	14,708,752 (6,573,789)	6,136,008		784,257
Mus musculus	128	37.1	18,645,060	14,380,528 (6,447,366)	5,878,592	11,225,458	
Gallus gallus	128	2.1	3,641,959	3,293,383 (3,280,002)	1,452,147		
Oryza sativa	128	4.1	5,872,081	5,418,373 (22,057)			
Canis familiaris	126	2.1	3,526,996	3,301,322 (217,525)	982,946		17
Pan troglodytes	127	2.1	1,544,900	1,543,208 (112,654)	527,665	1,544,895	2
Bos taurus	128	3.1	2,233,086	2,223,033 (14,371)	577,507	10,202	277
Monodelphis domestica	128	2.1	1,196,103	1,194,131 (0)	287,496		
Anopheles gambiae	128	2.2	1,368,906	1,131,534 (0)			
Apis mellifera	128	4.1	1,118,192	1,117,049 (16)	69,462		
Danio rerio	128	2.1	700,855	662,322 (3,091)	305,414	2,298	
Felis catus	127	1.1	327,037	327,037 (0)			
Plasmodium falciparum	127		185,071	185,071 (47)		199	
Rattus norvegicus	126	4.1	47,711	43,628 (1,605)	18,881		

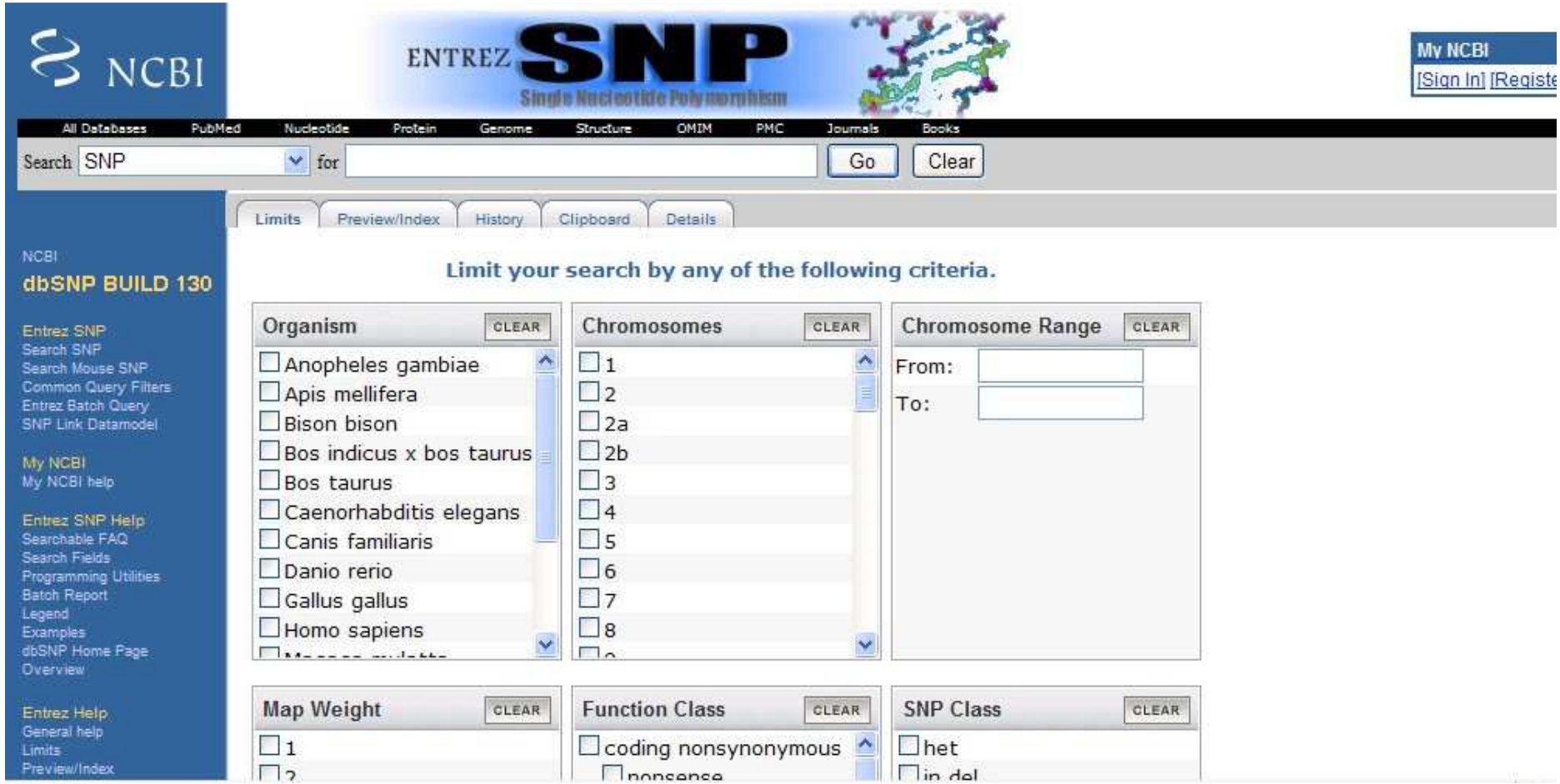
(http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi)

NCBI SNPs



(http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp&cmd=search&term=)

NCBI SNPs



(http://www.ncbi.nlm.nih.gov/snp/limits)

The “equivalent” of the US NCBI: EMBL

(http://www.embl.org/)

Primary data bases in detail: EMBL nucleotide sequence data base

EMBL-EBI | EB-eye Search | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback

Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

- EMBL-Bank Home
- Access
- Documentation
- News
- Submission
- Publications
- People
- Contact

EMBL Fetch

Fetch an EMBL record by id

Go

Hands-on Training

30th April - 1st May 2009: Short Read Bioinformatics hands-on EBI training course...more

EBI > Databases > EMBL-Bank

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

EMBL NUCLEOTIDE SEQUENCE DATABASE

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 99, March 2009), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucleic Acids Research 2009 37: D19-D25. provides further information and details.

The EMBL nucleotide sequence database is part of the The Protein and Nucleotide Database Group (**PANDA**). This is jointly headed by Dr. Rolf Apweiler and Dr. Ewan Birney, with Dr. Birney taking responsibility for Nucleotides.

Link	Explanation
Access	Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography.
Submission	Primary sequence submissions, third party annotation, updates.

(http://www.ebi.ac.uk/embl/index.html)

DNA Data Bank of Japan (DDBJ)

(http://www.ddbj.nig.ac.jp/)

DNA Data Bank of Japan (DDBJ)

DDBJ
DNA Data Bank of Japan

Accession DNA Protein AllDBs Taxonomy Site Search
Accession numbers Go
DDBJ UniProt PDB DAD PRF Patent >>more

HOME | Submission | How to Use | Search/Analysis | FTP/WebAPI | Report/Statistics | Contact Us | Japanese

HOME > How to Use

- About DDBJ
- How to Use
- Q and A

Sequence Submission
- SAKURA
- Mass Submission
- Data Updates

Search
- getentry
- ARSA
- TXSearch
- BLAST
- PSI-BLAST
- FASTA
- SSEARCH

DDBJing Contents: How to Use DDBJ

This page, DDBJing, intorduces how to use services and databases provided for everyone. "DDBJing" means "Effective using of DDBJ services".

Q and A

Frequently asked questions and thier answers related to DDBJ submission

- Data Submission

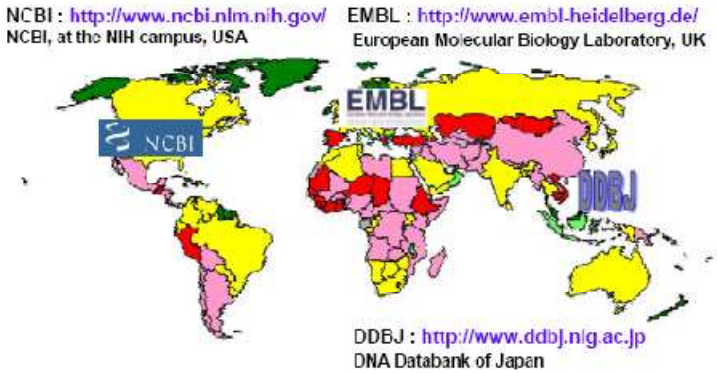
Explanation of DDBJ flat file format

The database is a collection of "entry" which is the unit of the data. The entry submitted to DDBJ is processed and publicized according to the DDBJ format for distribution (flat file). The flat file includes the sequence and the information of submitters, references, source organisms, and "feature" information, etc. The "feature" is defined by DDBJ/EMBL/GenBank Feature Table Definition to describe the biological nature such as gene function and other property of the nucleotide sequence.

- Example of Submission
- Organism Name
- Protein Coding Sequence
- Description of Location
- Feature/Qualifier Usage Matrix
- Feature Key
- Qualifier Key
- Codes Used in Sequence Description

(http://www.ddbj.nig.ac.jp/ddbjingtop-e.html)

The International Sequence Data base Collaboration



- These three databases have collaborated since 1982. Each database collects and processes new sequence data and relevant biological information from scientists in their region

- These databases automatically update each other with the new sequences collected from each region, every 24 hours. The result is that they contain exactly the same information, except for any sequences that have been added in the last 24 hours.
- This is an important consideration in your choice of database. If you need accurate and up to date information, you must search an up to date database.

(S Star slide: Ping)

Secondary data bases

- Derived information/ curated or procesed
- Fruits of analyses of sequences in the primary sources:
  - patterns,
  - blocks,
  - profiles etc.

  which represent the most conserved features of multiple alignments

Examples of secondary data bases

- Sequence-related Information
  - ProSite, Enzyme, REBase
- Genome-related Information
  - OMIM, TransFac
- Structure-related Information
  - DSSP, HSSP, FSSP, PDBFinder
- Pathway Information
  - KEGG, Pathways

Secondary data bases in detail: OMIM

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

My NCBI
[Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | OMIM

Search OMIM for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Entrez

OMIM
Search OMIM
Search Gene Map
Search Morbid Map

Help
OMIM Help
How to Link

FAQ
Numbering System
Symbols
How to Print
Citing OMIM
Download

OMIM Facts
Statistics
Update Log
Restrictions on Use

Allied Resources

- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria.
- Use **Index** to browse terms found in OMIM records.
- Use **History** to retrieve records from previous searches, or to combine searches.

OMIM ® - Online Mendelian Inheritance in Man ®

Welcome to OMIM ®, Online Mendelian Inheritance in Man ®. OMIM is a comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes. The full-text, referenced overviews in OMIM contain information on all known mendelian disorders and over 12,000 genes. OMIM focuses on the relationship between phenotype and genotype. It is updated daily, and the entries contain copious links to other genetics resources.

This database was initiated in the early 1960s by Dr. Victor A. McKusick as a catalog of mendelian traits and disorders, entitled Mendelian Inheritance in Man (MIM). Twelve book editions of MIM were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library of Medicine and the William H. Welch Medical Library at Johns Hopkins. It was made generally available on the internet starting in 1987. In 1995, OMIM was developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

OMIM is authored and edited at the McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, under the direction of Dr. Ada Hamosh.

NLM's Profiles in Science -- The McKusick Papers More...

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by

(http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)

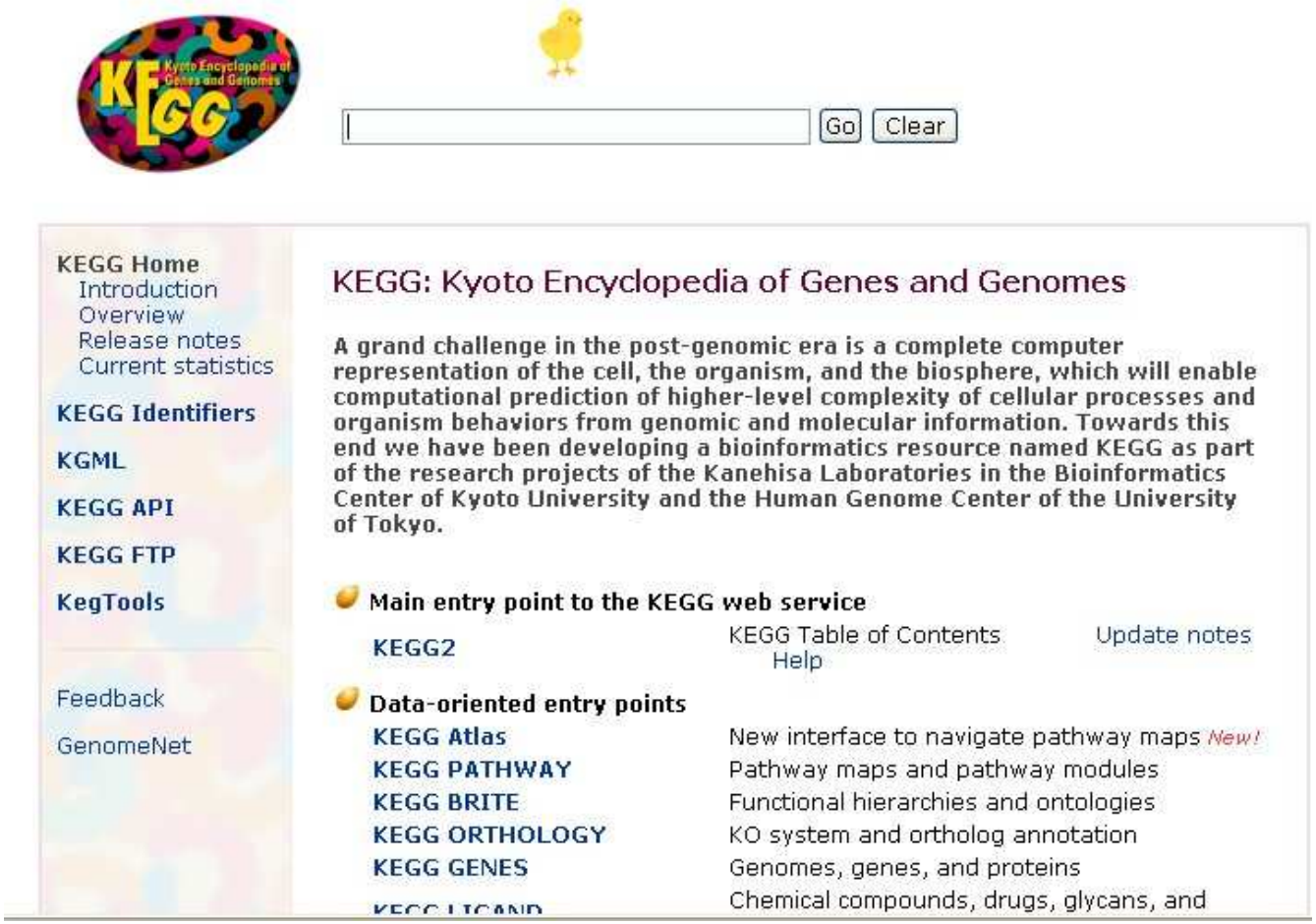
Examples of questions that can be answered with OMIM in Entrez

- What human genes are related to hypertension? Which of those genes are on chromosome 17? (strategy)
- List the OMIM entries that describe genes on chromosome 10. (strategy)
- List the OMIM entries that contain information about allelic variants. (strategy)
- Retrieve the OMIM record for the cystic fibrosis transmembrane conductance regulator (CFTR), and link to related protein sequence records via Entrez. (strategy)
- Find the OMIM record for the p53 tumor protein, and link out to related information in Entrez Gene and the p53 Mutation Database (strategy)

The "strategy" links lead to the Sample Searches section in the document

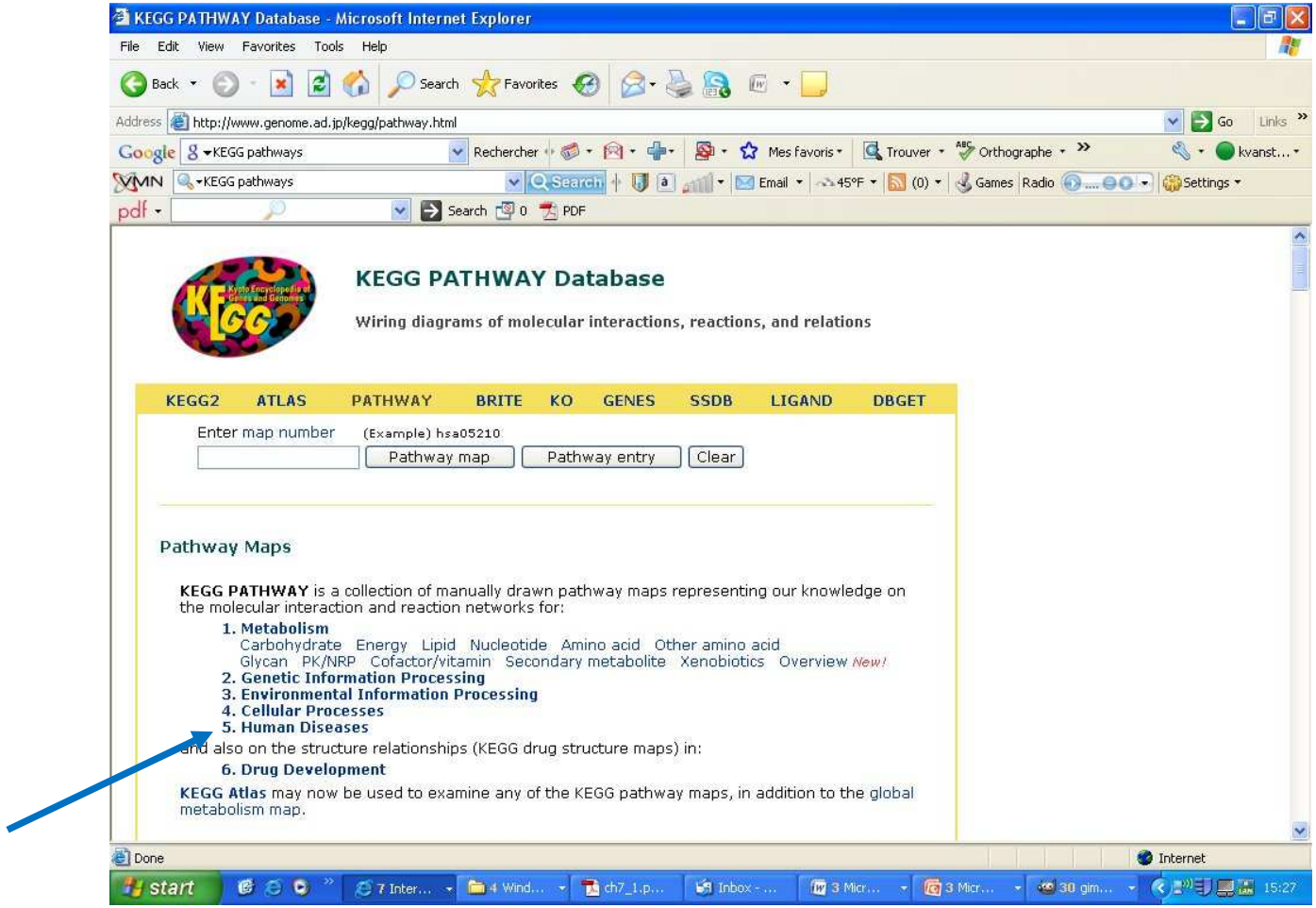
(http://www.ncbi.nlm.nih.gov/Omim/omimhelp.html#MainFeatures)

Secondary data bases in detail: KEGG portal



(http://www.genome.jp/kegg/)

Secondary data bases in detail: KEGG pathways data base



(http://www.genome.ad.jp/kegg/pathway.html)

5. Human Diseases

5.1 Cancers

Pathways in cancer (overview)
Colorectal cancer
Pancreatic cancer
Glioma
Thyroid cancer
Acute myeloid leukemia
Chronic myeloid leukemia
Basal cell carcinoma
Melanoma
Renal cell carcinoma
Bladder cancer
Prostate cancer
Endometrial cancer
Small cell lung cancer
Non-small cell lung cancer

KEGG DISEASE

Human diseases
ICD-10 disease classification

Pathways in cancer

5.2 Immune Disorders

Asthma
Systemic lupus erythematosus
Autoimmune thyroid disease
Allograft rejection
Graft-versus-host disease
Primary immunodeficiency

5.3 Neurodegenerative Diseases

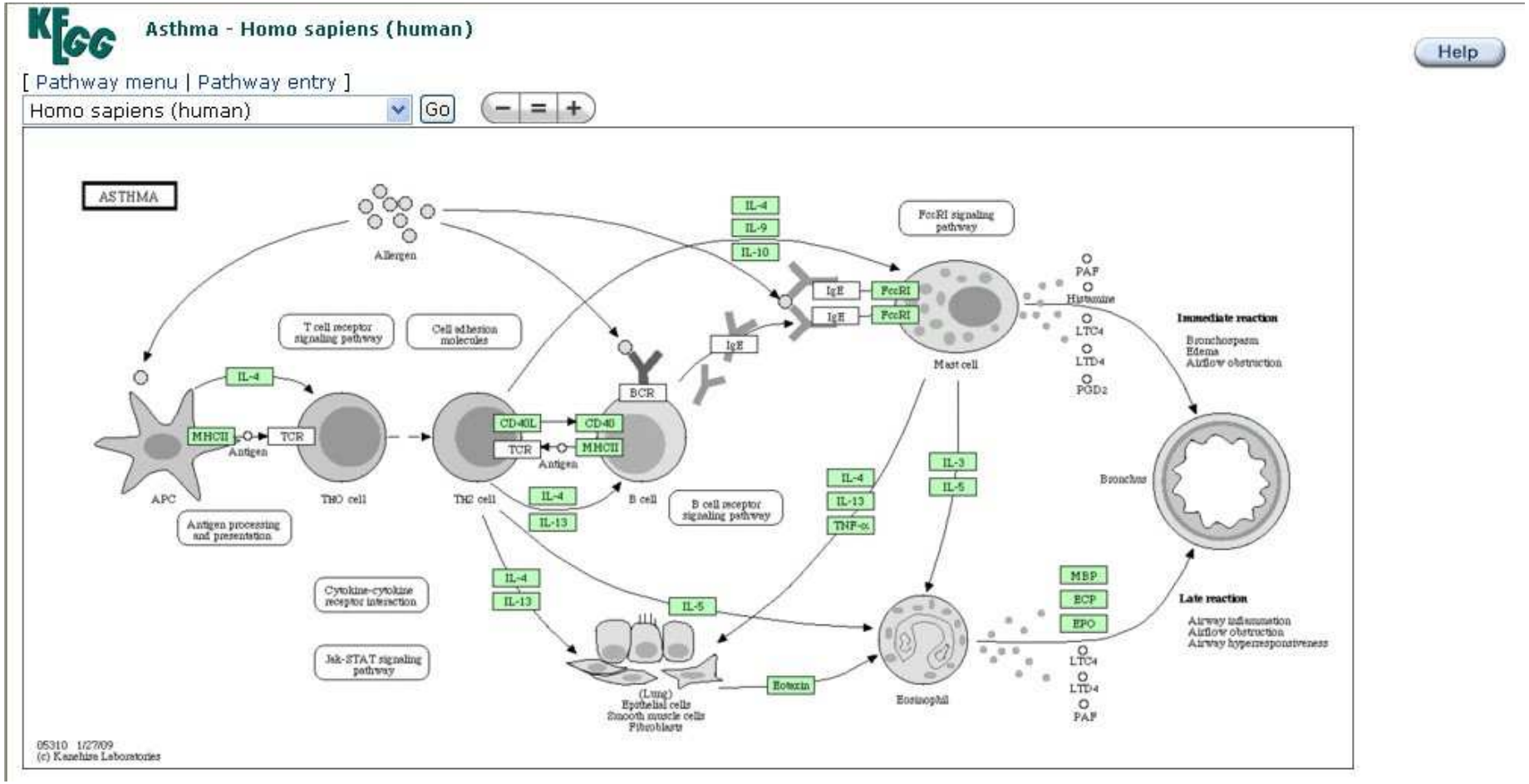
Alzheimer's disease *Revised!*
Parkinson's disease *Revised!*
Amyotrophic lateral sclerosis (ALS) *Revised!*
Huntington's disease *Revised!*

5.4 Metabolic Disorders

Type I diabetes mellitus
Type II diabetes mellitus
Maturity onset diabetes of the young

5.5 Infectious Diseases

KEGGpathway for asthma



(http://www.genome.ad.jp/kegg-bin/resize_map.cgi?map=hsa05310&scale=0.67)

Secondary data bases in detail: NCBI dbGaP

dbGaP Overview

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype. Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The advent of high-throughput, cost-effective methods for genotyping and sequencing has provided powerful tools that allow for the generation of the massive amount of genotypic data required to make these analyses possible.

dbGaP provides two levels of access - open and controlled - in order to allow broad release of non-sensitive data, while providing oversight and investigator accountability for sensitive data sets involving personal health information. Summaries of studies and the contents of measured variables as well as original study document text are generally available to the public, while access to individual-level data including phenotypic data tables and genotypes require varying levels of authorization.

View Certificate of Confidentiality

The data in dbGaP will be pre-competitive, and will not be protected by intellectual property patents. Investigators who agree to the terms of dbGaP data use may not restrict other investigators' use of primary dbGaP data by filing intellectual property patents on it. However, the use of primary data from dbGaP to develop commercial products and tests to meet public health needs is encouraged.

Submission Policy

Submitters who are not Federally-funded and affiliated with an NIH IC will need to work with an NIH DAC so that proposed submission can be reviewed for consistency with appropriate policies to protect the privacy of research participants and confidentiality of their data. Submissions to dbGaP will not be accepted without assurance that the submitting institution approves the submission and has verified that the data submission is consistent with all applicable laws and regulations, as well as institutional policies. Submitters must also identify any limits on research uses of the data that are specifically set by individual research participants, e.g., through their informed consent.

Data Content and Organization

(http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html)

NCBI as portal to dbGAP

NCBI

dbGaP GENOTYPES and PHENOTYPES

My NCBI [Sign In] [Register]

Search dbGaP for Go Clear

Limits Preview/Index History Clipboard Details

NCBI

dbGaP
About dbGaP
Browse dbGaP
Authorized Access
Email Alerts
dbGaP Tutorial
Security Procedures
FTP Download
Publications
Contact dbGaP
dbGaP Alert NEW

Other Services

MeSH Browser
Clinical Queries

Browse dbGaP TUTORIAL ABOUT dbGAP

By Studies | By Diseases | Advanced Search

Project	Study	Embargo Release	Details	Participants	Type of Study
CIDR	CIDR: Genome Wide Association Study in Familial Parkinson Disease (PD)	Feb 13, 2009	V D A	1991	Case-Control
CIDR	CIDR: Collaborative Study on the Genetics of Alcoholism (COGA)	Oct 06, 2009	V D A	1945	Case-Control
COG	Genome-Wide Association Study of Neuroblastoma	Dec 18, 2008	V D A	1032	Case-Control
GAIN	Genotyping the 270 HapMap samples for GAIN by Broad		V D A	-	Parent-Offspring Trios
GAIN	Search for Susceptibility Genes for Diabetic Nephropathy in Type 1 Diabetes (GoKinD study participants), GAIN	Jul 09, 2008	V D A	1825	Case-Control

(http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap)

Tertiary data bases

- Tertiary sources consist of information which is a distillation and collection of primary and secondary sources.
- These include:
  - structure databases
  - flatfile databases

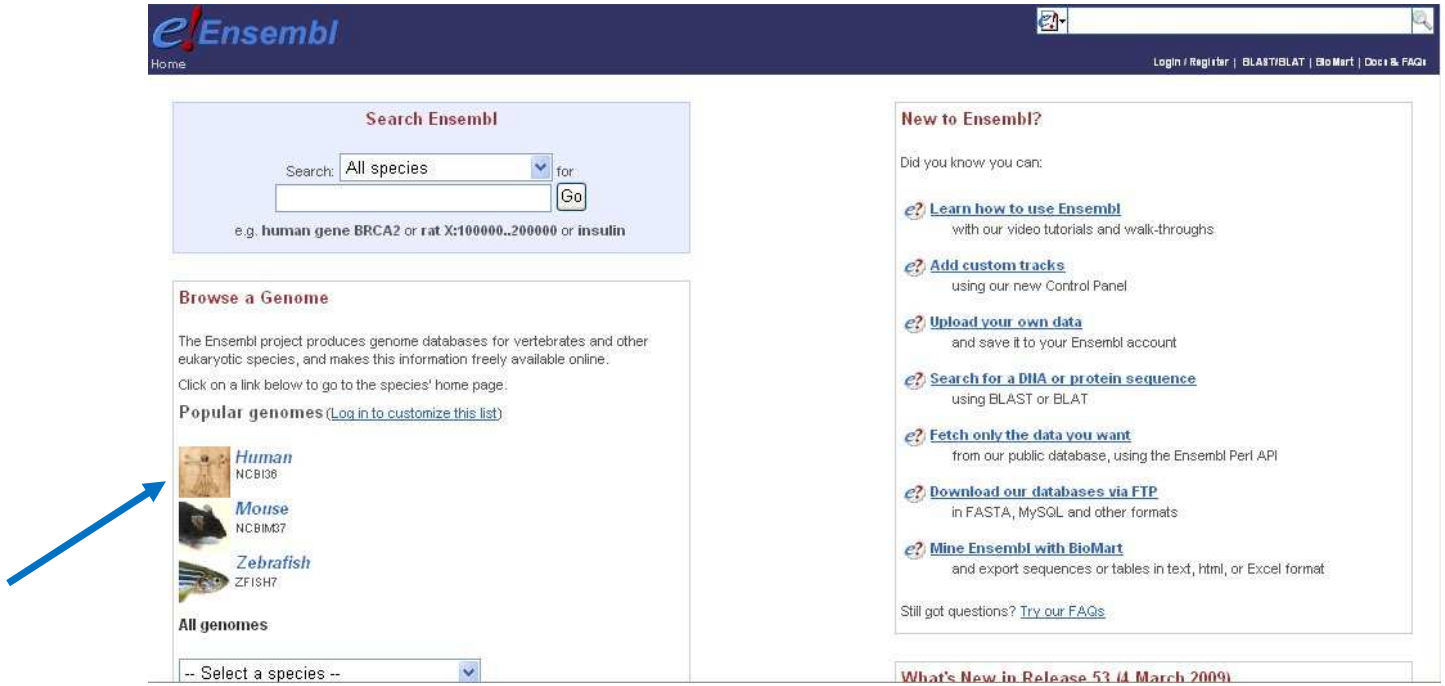
1.c Searching data bases

Where the h… is the d… thing?

- Start looking in some of the big systems (EMBL, NCBI, KEGG, etc).
- Read their help pages.
- Use their data.
- Follow their hyperlinks.

Ensembl genome browser portal

- Ensembl is a joint project between EMBL-EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes



(http://www.ensembl.org/index.html)

Ensembl genome browser portal

Ensembl

Home > Human

Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

About this species
- Description
- Genome Statistics
  - Assembly and Geneb
  - Top 40 InterPro hits
  - Top 500 InterPro hits
- What's New
- Sample entry points
  - Karyotype
  - Location (AL032821.
  - Gene (BRCA2)
  - Transcript (FOXP2-20

- Configure this page
- Manage your data
- Export data
- Bookmark this page

Search Ensembl Human

Search for: Go

e.g. gene BRCA2 or AL032821.2.1.143563 or muscular dystrophy

Description

Assembly and Genebuild »

Human (*Homo sapiens*)

Assembly

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly from the Genome Reference Consortium. The data set consists of gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

It also includes nine haplotypic regions, mainly in the MHC region of chromosome 6.

Annotation

Since release 55 (July 2009) the gene annotation presented here has been a combined Ensembl-Havana geneset, which incorporates protein-coding and non-coding transcripts annotated by the Havana team into the automatically-annotated Ensembl gene set.

The major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed. This undertaking, the Consensus Coding Sequence (CCDS) project was initially based on the NCBI36 assembly. The CCDS identifiers have been mapped onto the new annotations based on the latest GRCh37 assembly and these identifiers are also shown.

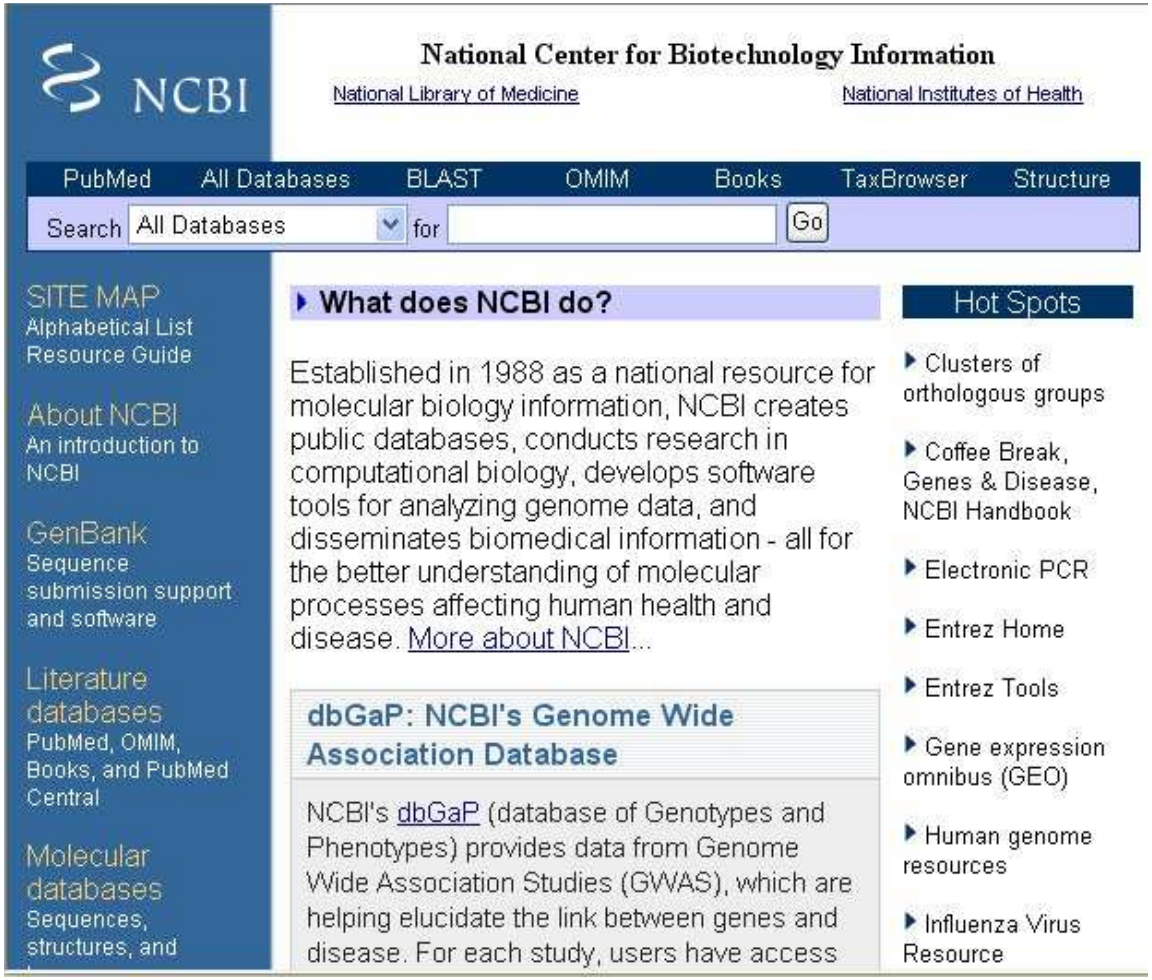
- More information about the CCDS project.

(http://www.ensembl.org/Homo_sapiens/Info/Index)

Contigs

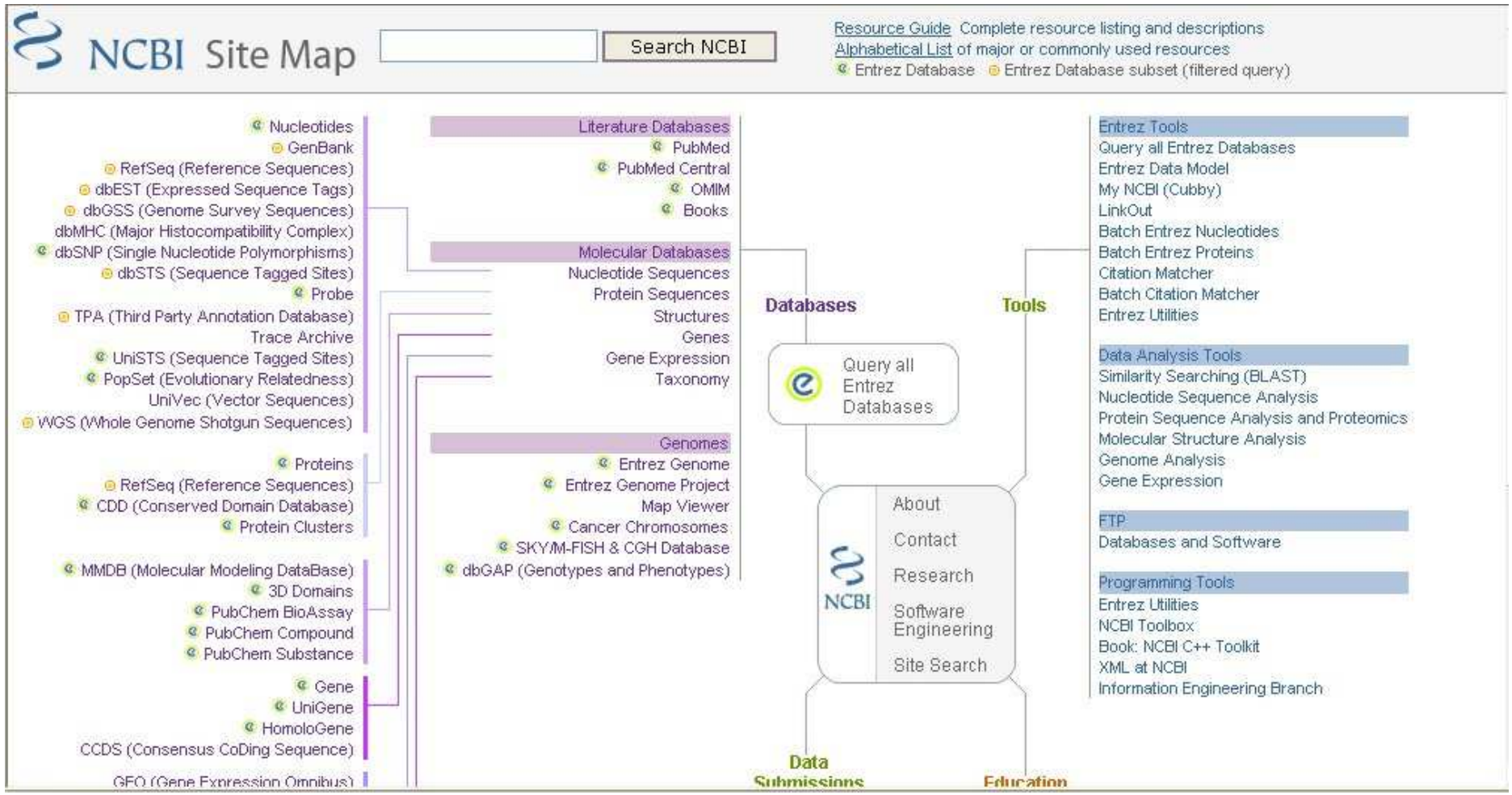
- In order to make it easier to talk about our data gained by the shotgun method of sequencing, researchers have invented the word "contig".
- A contig is a set of gel readings that are related to one another by overlap of their sequences.
- All gel readings belong to one and only one contig, and each contig contains at least one gel reading.
- The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig

Entrez genome browser portal

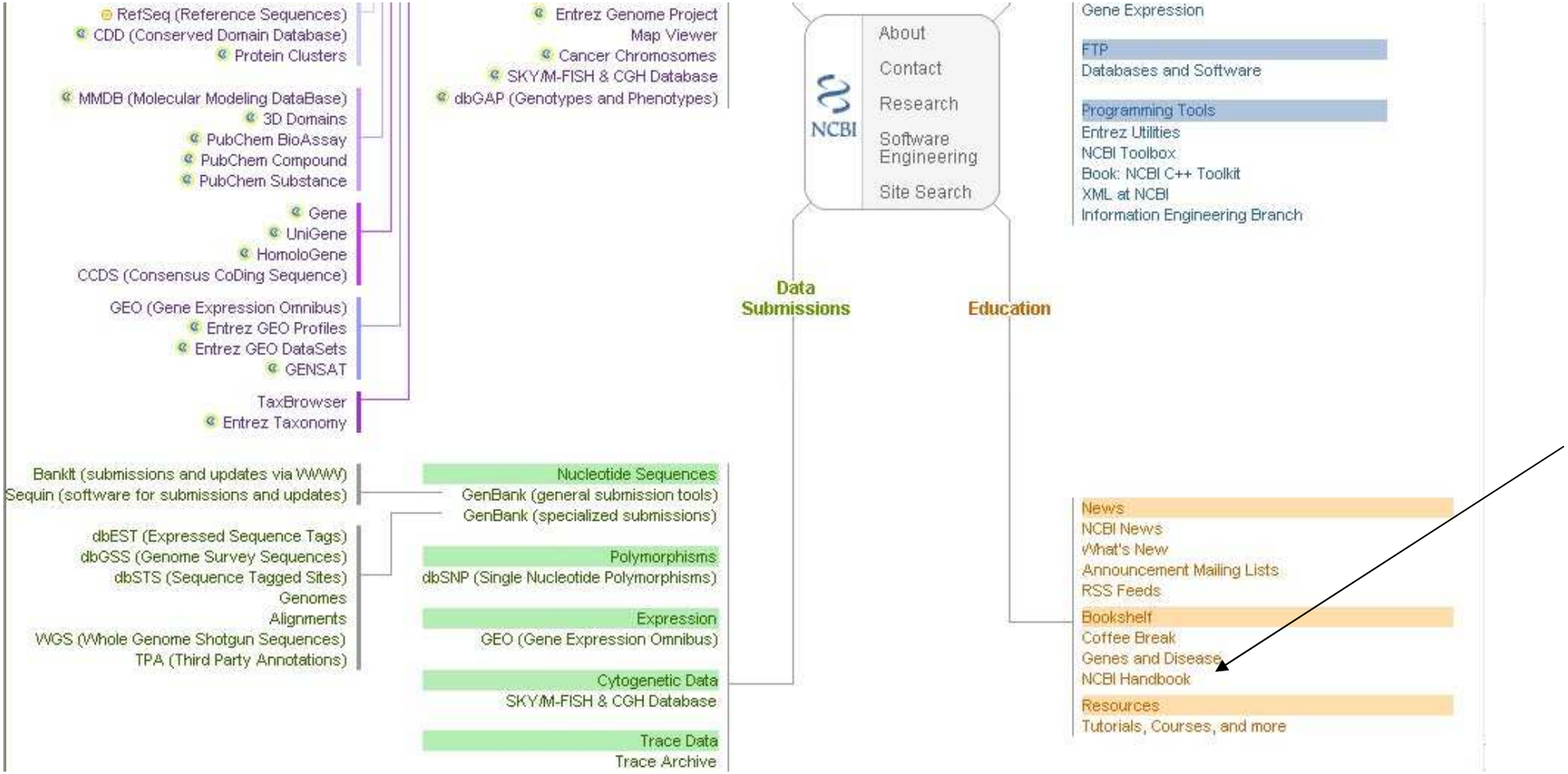


(http://www.ncbi.nlm.nih.gov/)

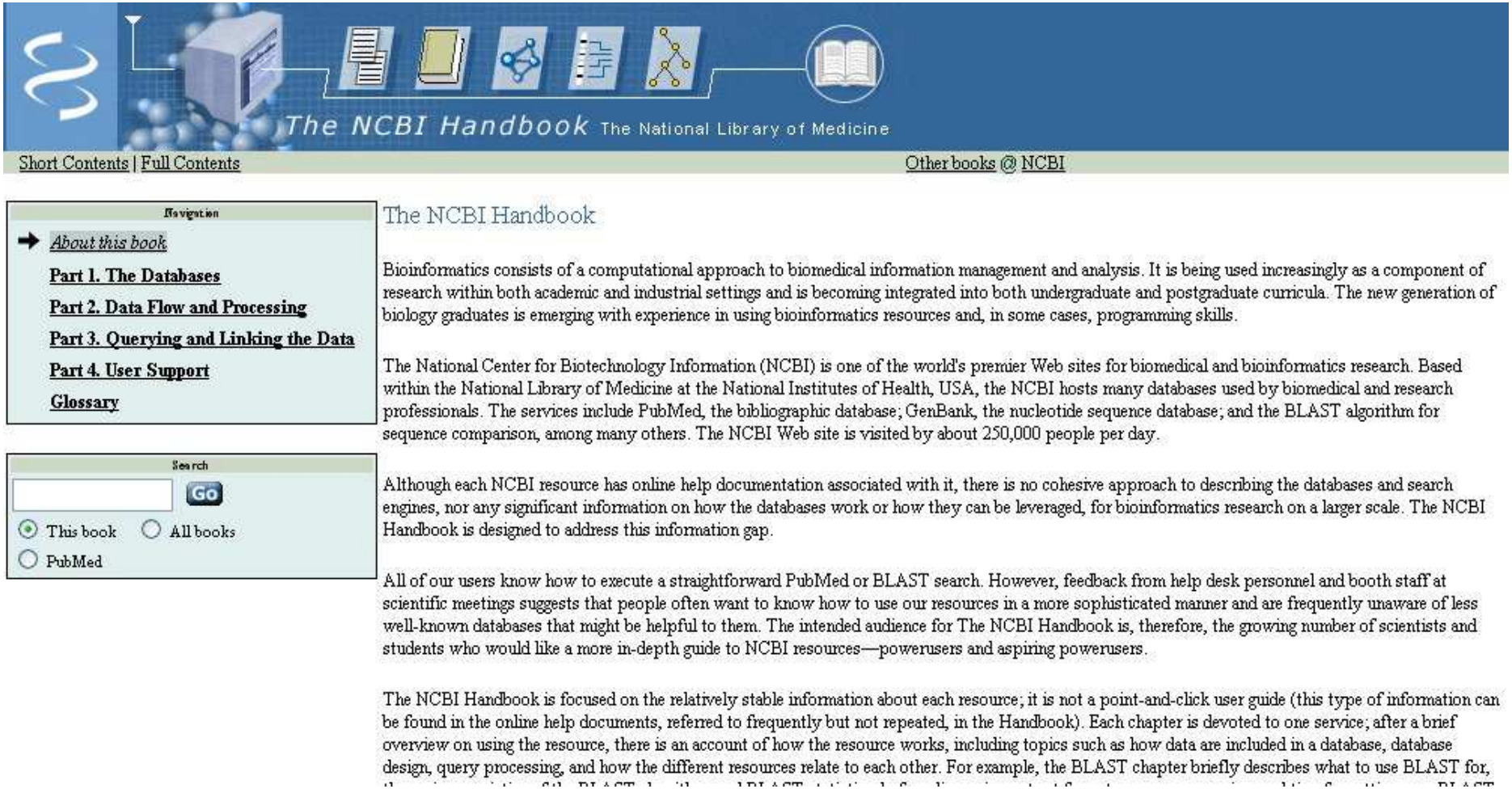
NCBI Site Map



NCBI Site Map (continued)



NCBI Handbook



NCBI Handbook snapshot

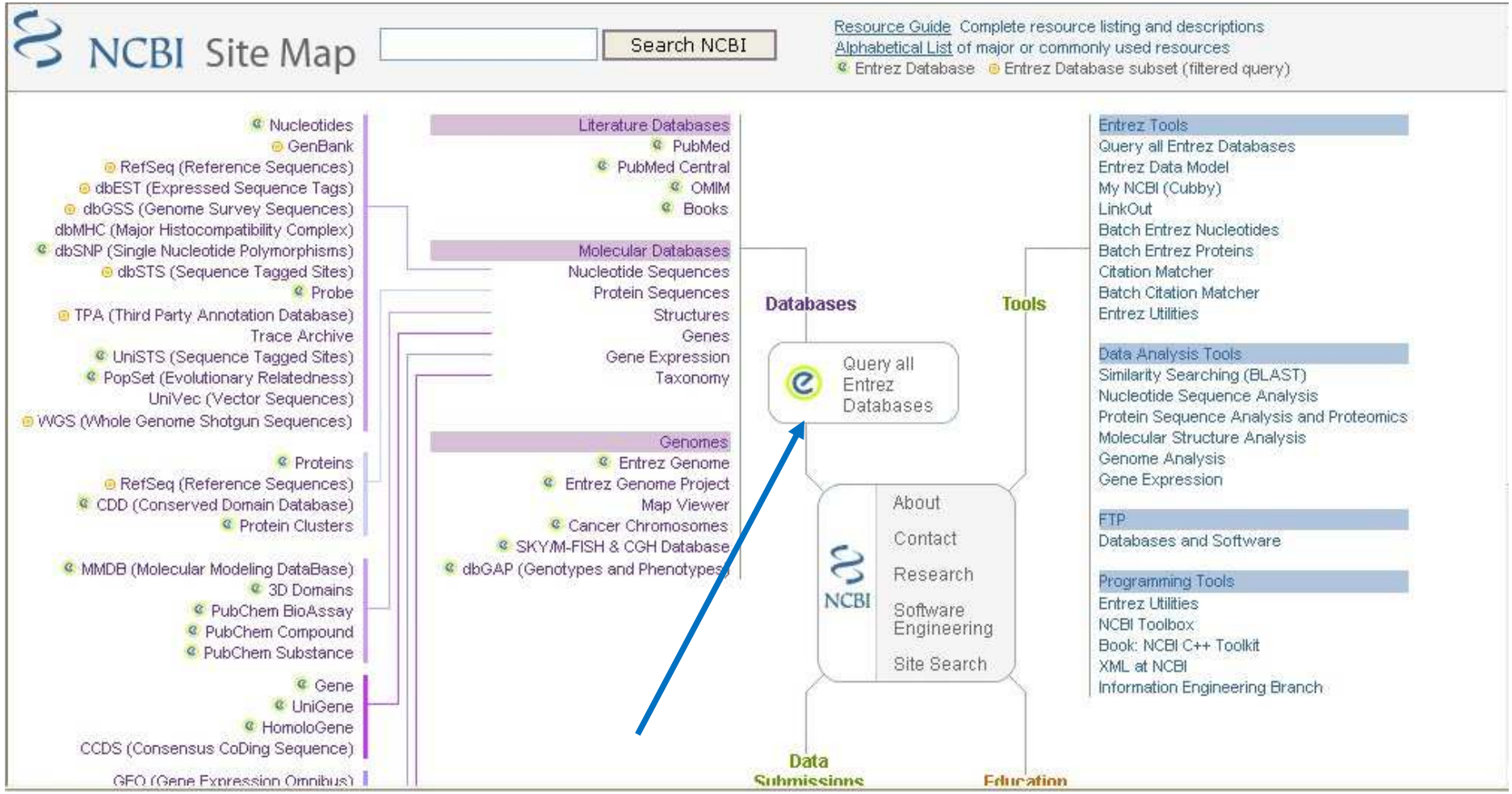
Paul Kitts.
Created: October 9, 2002, Updated: August 13, 2003

Part 3. Querying and Linking the Data

15. The Entrez Search and Retrieval System
Jim Ostell.
Created: October 9, 2002, Updated: August 13, 2003
16. The BLAST Sequence Analysis Tool
Tom Madden.
Created: October 9, 2002, Updated: August 13, 2003
17. LinkOut: Linking to External Resources from Entrez Databases
Kathy Kwan.
Created: October 9, 2002, Updated: August 13, 2003
18. The Reference Sequence (RefSeq) Project
Kim Pruitt, Tatiana Tatusova, and Donna Maglott.
Created: October 09, 2002, Updated: January 3, 2007
19. Entrez Gene: A Directory of Genes
Donna Maglott, Kim Pruitt, and Tatiana Tatusova.
Created: March 3, 2005
20. Using the Map Viewer to Explore Genomes
Susan M. Dombrowski and Donna Maglott.
Created: October 9, 2002, Updated: August 13, 2003
21. UniGene: A Unified View of the Transcriptome
Joan U. Pontius, Lukas Wagner, and Gregory D. Schuler.
Created: October 9, 2002, Updated: August 13, 2003
22. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes
Eugene V. Koonin.
Created: October 9, 2002, Updated: August 13, 2003

Part 4. User Support

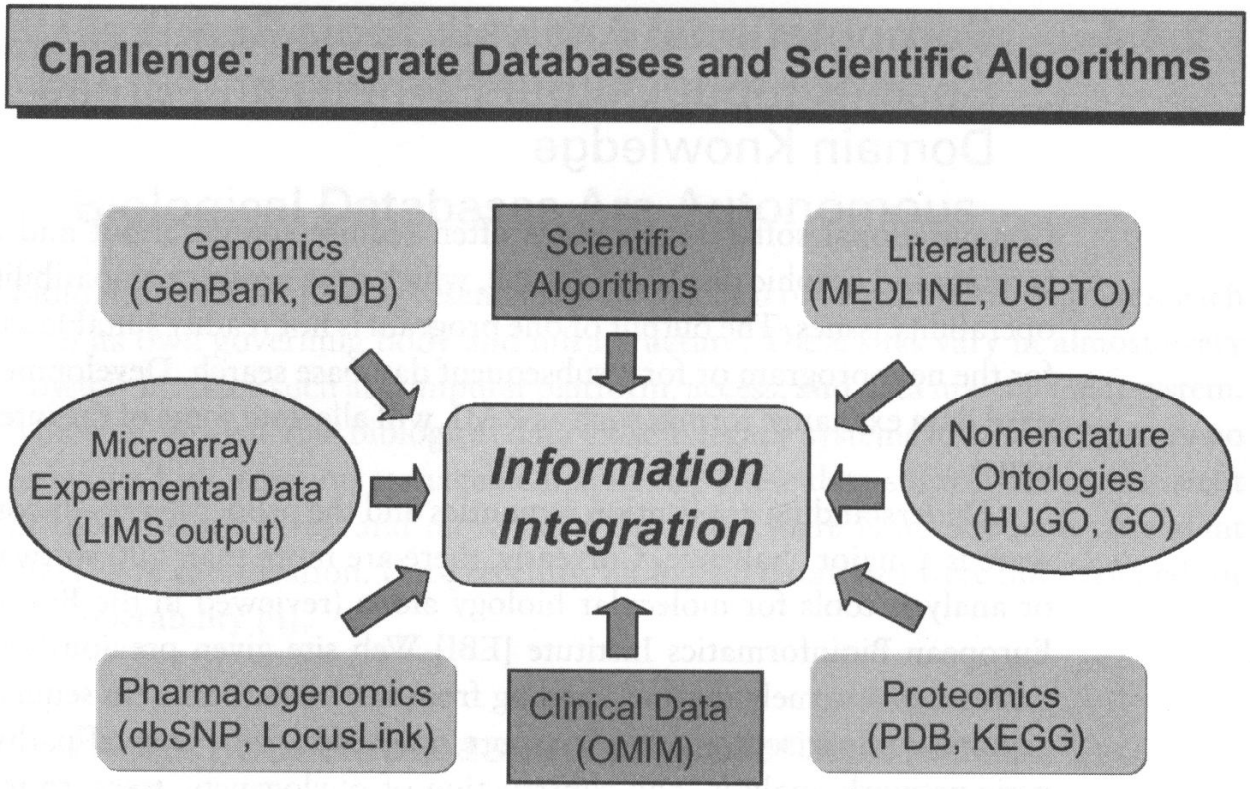
NCBI Site Map



Entrez: An integrated database search and retrieval system

(http://www.ncbi.nlm.nih.gov/sites/gquery)

Information integration is essential: data aggregation from several databases



(Bioinformatics: Managing Scientific Data)

References:

- Deonier et al. *Computational Genome Analysis*, 2005, Springer. (Chapter 10)
- Hahne et al. *Bioconductor Case Studies*, 2008, Springer (Chapter 9,10)
- URLs:
  - http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf

Background reading:

- Roos 2001. Bioinformatics – trying to swim in a sea of data. *Science*, 16 (291):1260-1261.
- Philippi et al 2006. Addressing the problems with life-science databases for traditional uses and systems biology. Nature Reviews Genetics – Perspectives 7: 482-.
- Alfred 2001. Mining the bibliome. Nature Reviews Genetics – Highlights 2: 401.
- Eglen 2009. A quick guide to teaching R programming to computational biology students. PLoS computational biology 8: e1000482.
- HT_BioC_manual: http://htseq.ucr.edu/ (part of R BioConductor Manual)
- Jain et al. 2000. Data clustering: a review. ACM Computing Surveys. 31 (3), September 1999. [Sections 1-4, 5.1,5.2,5.4]

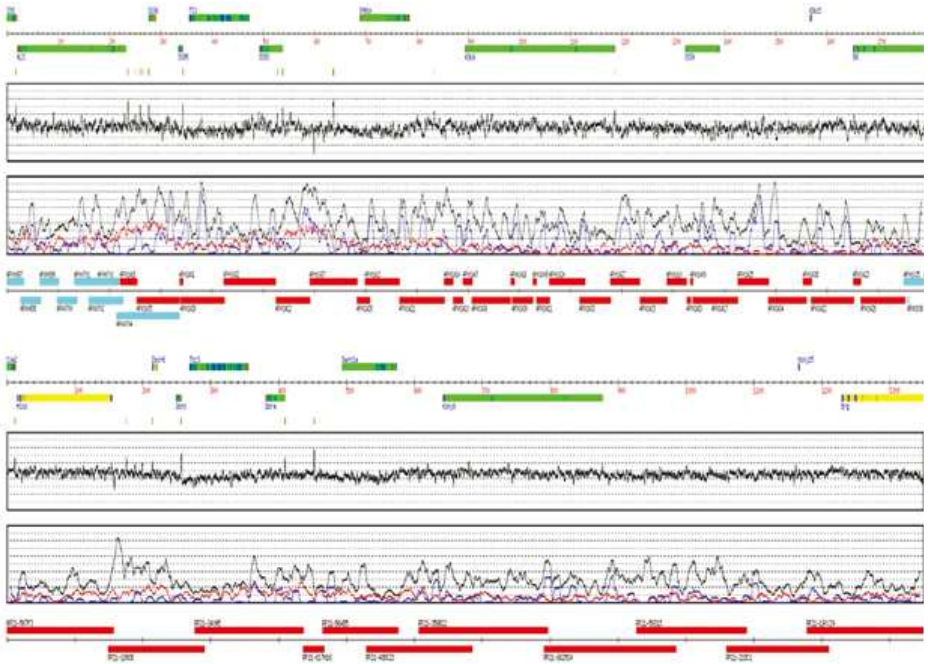
In-class discussion document

- Mailman et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. Nature Genetics 39(10): 1181-.
- Flintoft 2005. From genotype to phenotype: a shortcut through the library. Nature Reviews Genetics 6: 1.

Questions: In class reading_3.pdf

Preparatory Reading:

- Facts about Human Genome Sequencing: http://www.ornl.gov/sci/techresources/Human_Genome/faq/seqfacts.shtml
- Insights learned from the human DNA sequence http://www.ornl.gov/sci/techresources/Human_Genome/project/journals/insights.shtml



(Nature, May 18, 2000 issue)

Human chromosome 21 is the causative chromosome of Down's syndrome, which is the most frequent neonatal disorder. Sequencing chromosome 21 has revealed the existence of 11 genes within the essential region of Down's syndrome (upper panel). It is supposed that the overexpressions of these genes are related to the symptoms of Down's syndrome, such as mental retardation. In addition, we determined the sequence in the corresponding region of the mouse genome (bottom panel) and conducted a comparative study. Although 10 genes were well conserved in the mouse genome, a gene designated DSCR9 was found only in the human genome.