

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

**Kristel Van Steen, PhD<sup>2</sup>**

**Montefiore Institute - Systems and Modeling**

**GIGA - Bioinformatics**

**ULg**

**[kristel.vansteen@ulg.ac.be](mailto: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**

[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

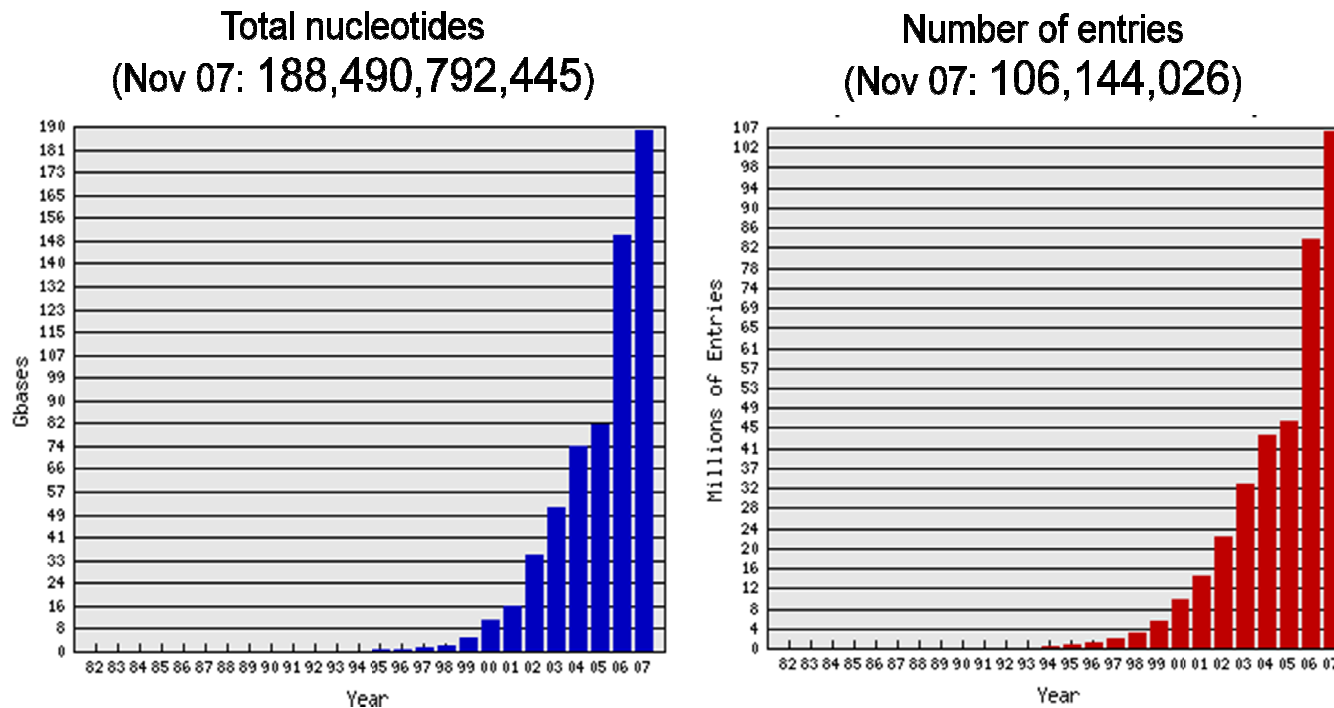
## Explosive growth of data

- 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
	<i>Mycoplasma genitalium</i>	517
	<i>Saccharomyces cerevisiae</i>	6,275
	<i>Arabidopsis thaliana</i>	~ 20,000
	<i>Caenorhabditis elegans</i>	19,099
	<i>Haemophilus influenzae</i>	1,743
	<i>Drosophila melanogaster</i>	13,601
	<i>Neisseria meningitidis</i>	2,158
	<i>Homo sapiens</i>	~ 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](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

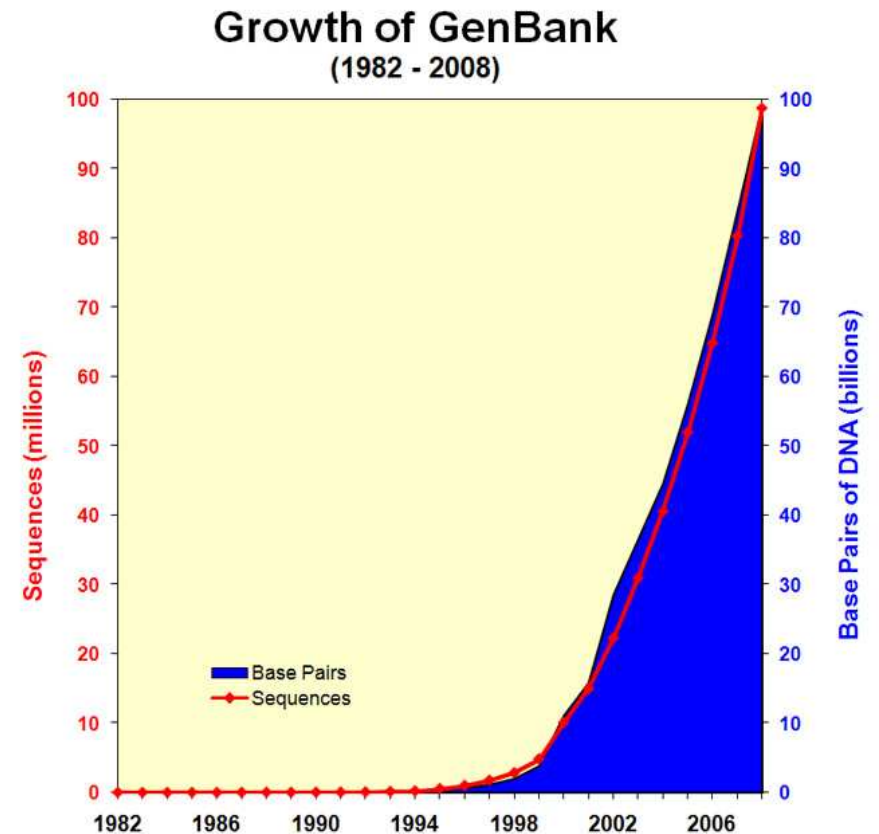
- GDB, MGD, ACeDB

- Structure Information

- PDB, NDB, CCDB/CSD

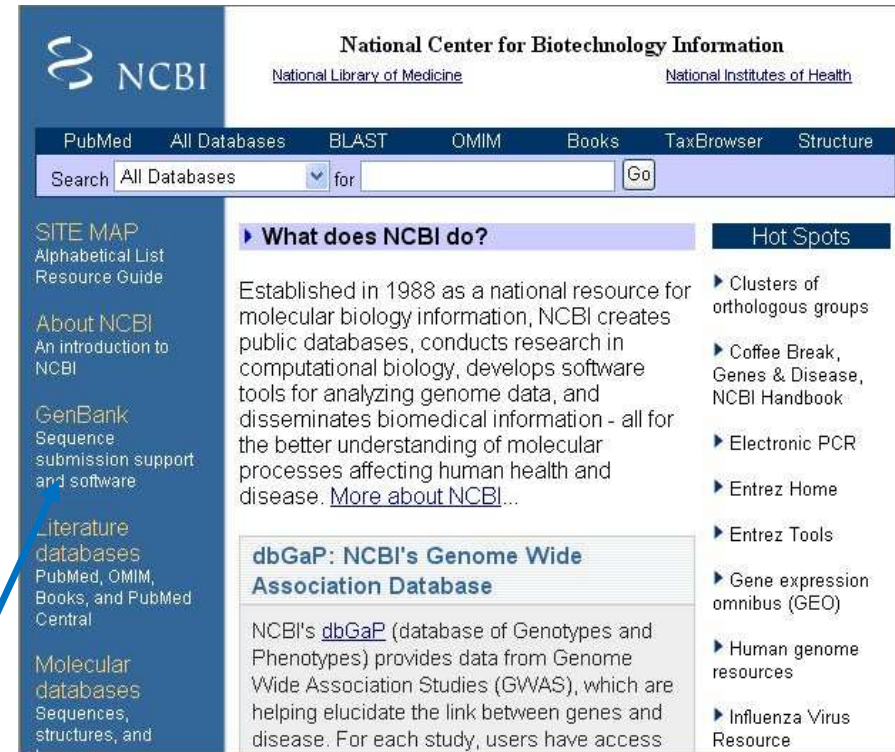
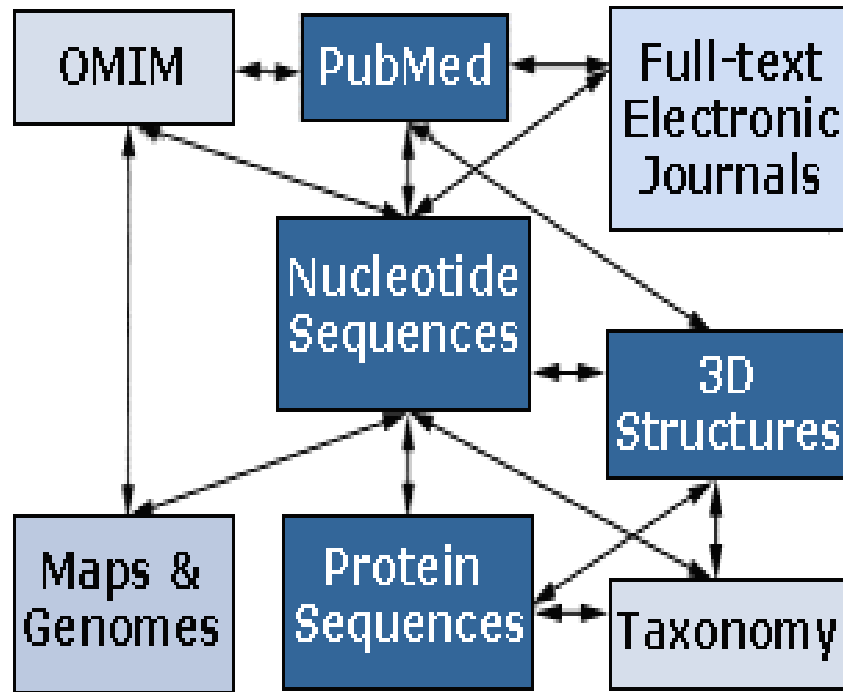
## Primary databases in detail: GenBank

- 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

**About NCBI**  
National Center for Biotechnology Information

<a href="#">About NCBI</a>	<a href="#">NCBI at a Glance</a>	<a href="#">A Science Primer</a>	<a href="#">Databases and Tools</a>
<a href="#">Human Genome Resources</a>	<a href="#">Model Organisms Guide</a>	<a href="#">Outreach and Education</a>	<a href="#">News</a>

[About NCBI Site Map](#)  
[NCBI News](#)  
[Subscribe to NCBI-Announce](#)

**NCBI at a Glance**  
**A Science Primer**  
**Databases and Tools**  
**Human Genome Resources**  
**Model Organisms Guide**  
**Outreach and Education**  
**News**

<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 TCPI-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 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 TCPI-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

NCBI Single Nucleotide Polymorphism

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for  Go

**BUILD 129**  
Have a question about dbSNP? Try searching the SNP FAQ Archive!  
  
Go

**GENERAL**  
**HUMAN VARIATION**  
Search, Annotate, Submit NEW  
Annotate and Submit Batch Data with Clinical Impact NEW  
**SNP SUBMISSION**  
**DOCUMENTATION**

dbSNP Search Options

Entrez SNP ID Numbers Submission Info Batch Locus Info Between Markers

**ANNOUNCEMENT**  
09/30/2008: SNP RefSeq mRNA Annotation Problems

Attention dbSNP user:

We discovered two problems with SNP annotation on RefSeq mRNA.  
Problem 1: A drop in the total number of SNP annotations from dbSNP build 129 onto human mRNA sequences for RefSeq

**Search by IDs on All Assemblies**

Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

Reference cluster ID(rs#)

Search Reset

**Submission Information**

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

([http://www.ncbi.nlm.nih.gov/SNP/snp\\_summary.cgi](http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi))

# NCBI SNPs

NCBI  
ENTREZ **SNP**  
Single Nucleotide Polymorphism

My NCBI  
[Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search SNP for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Click on the image below to view the connections between Entrez SNP and other databases.

NCBI  
**dbSNP BUILD 130**

Entrez SNP  
Search SNP  
Search Mouse SNP  
Common Query Filters  
Entrez Batch Query  
SNP Link Datamodel

My NCBI  
My NCBI help

Entrez SNP Help  
Searchable FAQ  
Search Fields  
Programming Utilities  
Batch Report  
Legend  
Examples  
dbSNP Home Page  
Overview

Entrez Help  
General help  
Limits  
Preview/Index

SNP

19,500  
42,500  
1,000

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

# NCBI SNPs

The screenshot shows the NCBI Entrez SNP search interface. At the top, there is a search bar with "SNP" entered and a "Go" button. Below the search bar are navigation tabs: All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. A "My NCBI" section in the top right contains links for "[Sign In]" and "[Register]".

On the left side, there is a sidebar with the NCBI logo and "dbSNP BUILD 130". Below this, there are links for "Entrez SNP", "Search SNP", "Search Mouse SNP", "Common Query Filters", "Entrez Batch Query", and "SNP Link Datamodel". Further down are "My NCBI" and "My NCBI help" links. At the bottom of the sidebar are "Entrez Help" links: "General help", "Limits", and "Preview/Index".

The main content area features a heading "Limit your search by any of the following criteria." followed by six filter panels, each with a "CLEAR" button:

- Organism:** A list of organisms with checkboxes, including Anopheles gambiae, Apis mellifera, Bison bison, Bos indicus x bos taurus, Bos taurus, Caenorhabditis elegans, Canis familiaris, Danio rerio, Gallus gallus, and Homo sapiens.
- Chromosomes:** A list of chromosomes with checkboxes, including 1, 2, 2a, 2b, 3, 4, 5, 6, 7, 8, and X.
- Chromosome Range:** Two input fields labeled "From:" and "To:" for specifying a range of chromosomes.
- Map Weight:** A list of map weights with checkboxes, including 1 and 2.
- Function Class:** A list of function classes with checkboxes, including coding nonsynonymous and nonsense.
- SNP Class:** A list of SNP classes with checkboxes, including het and in del.

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

## The “equivalent” of the US NCBI: EMBL



The image shows a screenshot of the EMBL website. At the top left is the EMBL logo, which consists of the letters 'EMBL' in a bold, sans-serif font next to a green hexagonal grid of dots with one red dot in the center. To the right of the logo is the text 'European Molecular Biology Laboratory'. Below this, there are five vertical panels, each representing a different EMBL site. Each panel has a colored header with the site name, a central image, and a white box at the bottom with the site's primary focus. The panels are: 1. EMBL-EBI Hinxton (green header, laptop image, 'European Bioinformatics Institute'); 2. EMBL Grenoble (blue header, circular building image, 'Structural Biology'); 3. EMBL Heidelberg (green header, DNA double helix image, 'Main Laboratory'); 4. EMBL Hamburg (blue header, modern building image, 'Structural Biology'); 5. EMBL Monterotondo (red header, mouse image, 'Mouse Biology'). Below the panels, there is a light blue banner with the text 'Europe's flagship laboratory for basic research in molecular biology'. At the bottom left, there is a paragraph of text: 'EMBL is at the forefront of innovation in life sciences research, technology development and transfer, and provides outstanding training and services to the scientific community in its member states. This publicly-funded non-profit institute is housed at five sites in Europe whose expertise covers the whole spectrum of molecular biology.' At the bottom right, there is a 'Please select:' label, a dropdown menu with 'More Information' selected, and a 'Show' button.

EMBL European Molecular Biology Laboratory

EMBL-EBI Hinxton  
European Bioinformatics Institute

EMBL Grenoble  
Structural Biology

EMBL Heidelberg  
Main Laboratory

EMBL Hamburg  
Structural Biology

EMBL Monterotondo  
Mouse Biology

Europe's flagship laboratory for basic research in molecular biology

EMBL is at the forefront of innovation in life sciences research, technology development and transfer, and provides outstanding training and services to the scientific community in its member states. This publicly-funded non-profit institute is housed at five sites in Europe whose expertise covers the whole spectrum of molecular biology.

Please select:  
More Information

(<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

**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.

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
<a href="#">Access</a>	<a href="#">Database queries</a> , <a href="#">Completed genomes webserver</a> , <a href="#">FTP archives</a> (EMBL release, alignments etc), <a href="#">EMBL sequence version archive (SVA)</a> , <a href="#">Browse by geography</a> .
<a href="#">Submission</a>	Primary sequence submissions, third party annotation, updates.

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

## DNA Data Bank of Japan (DDBJ)

**DDBJ**  
DNA Data Bank of Japan

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

HOME Submission How to Use Search/Analysis FTP/WebAPI Report/Statistics Contact Us [RSS](#) [Japanese](#)

▶ About DDBJ  
 ▶ How to Use  
 ▶ G and A  
 ▶ **Sequence Submission**  
 ▶ [SAKURA](#)  
 ▶ [Mass Submission](#)  
 ▶ [Data Updates](#)

▶ **Search**  
 ▶ [getentry](#)  
 ▶ [ARSA](#)  
 ▶ [TXSearch](#)  
 ▶ [BLAST](#)  
 ▶ [PSI-BLAST](#)  
 ▶ [FASTA](#)  
 ▶ [SSEARCH](#)

▶ **Phylogenetics**

**DDBJ : DNA Data Bank of Japan**  
 DDBJ (DNA Data Bank of Japan) is one of three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, through close collaboration with EBI in Europe and NCBI in USA.

**Hot Topics** [More](#)

- ▶ Aug. 28, 2009 [Change of directory structure of anonymous FTP site: Made a new directory "top"](#)
- ▶ Aug. 13, 2009 [A new high speed BLAST API was opened to the public from WABI](#)
- ▶ Aug. 11, 2009 [Redistribution of genomic sequence \(build 4\) of the cultivar Nipponbare of Japanese rice \(Oryza sativa Japonica Group\) assigned with RAP annotation](#)

**Maintenance** [More](#)

- ▶ Sep. 11, 2009 [Suspension of some parts of DDBJ activities during national holidays \(Sep.19-23\)](#)
- ▶ Aug. 04, 2009 [Request for re-submission of the queries \(Re: Homology Search and ClustalW temporary down \(Aug.3\)\)](#)
- ▶ Jun. 23, 2009 [Completion of a bug fixation: error in displaying suppressed entries by getentry](#)

**Sequence Data Submission**

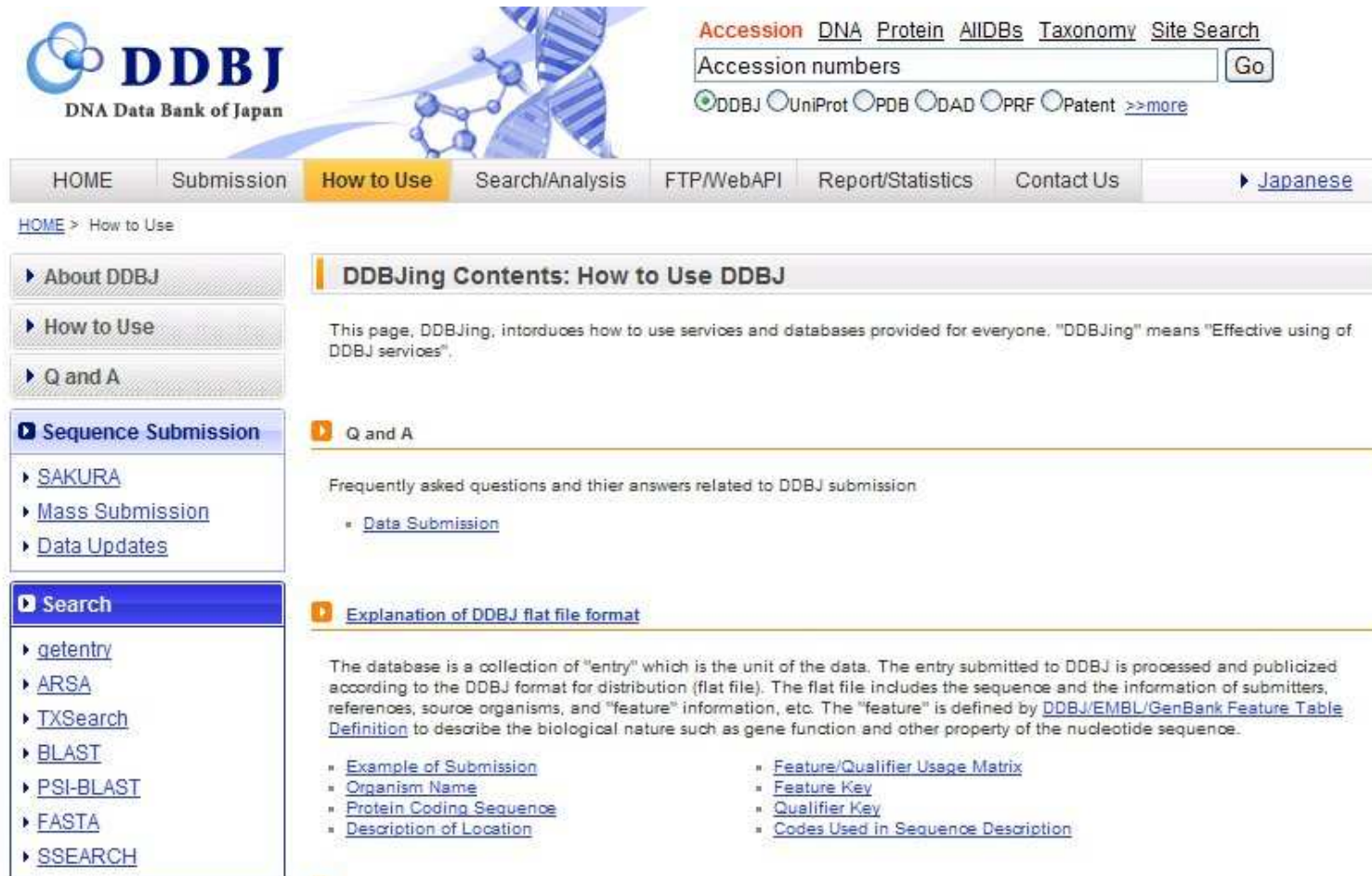
- Submit my sequences  
Orientation for the data submission.

**FTP/Web API**

- FTP ( [ftp.ddbj.nig.ac.jp](ftp://ftp.ddbj.nig.ac.jp) )  
Download data files.

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

## DNA Data Bank of Japan (DDBJ)



**DDBJ**  
DNA Data Bank of Japan

Accession DNA Protein AIIBs Taxonomy Site Search  
 Accession numbers    
 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, introduces how to use services and databases provided for everyone. "DDBJing" means "Effective using of DDBJ services".

#### Q and A

Frequently asked questions and their answers related to DDBJ submission

- [Data Submission](#)

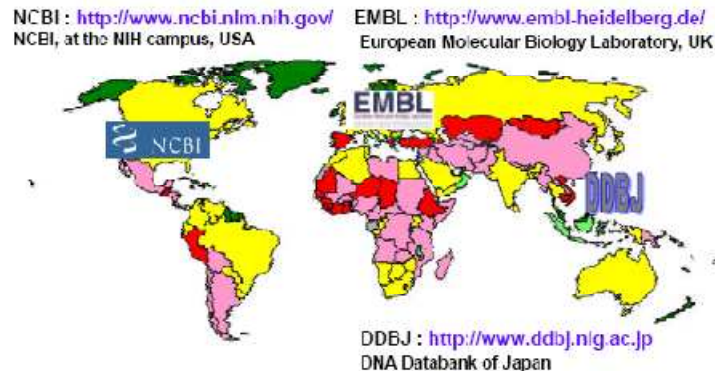
#### Explanation of DDBJ flat file format

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 Database 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

- 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 advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic

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

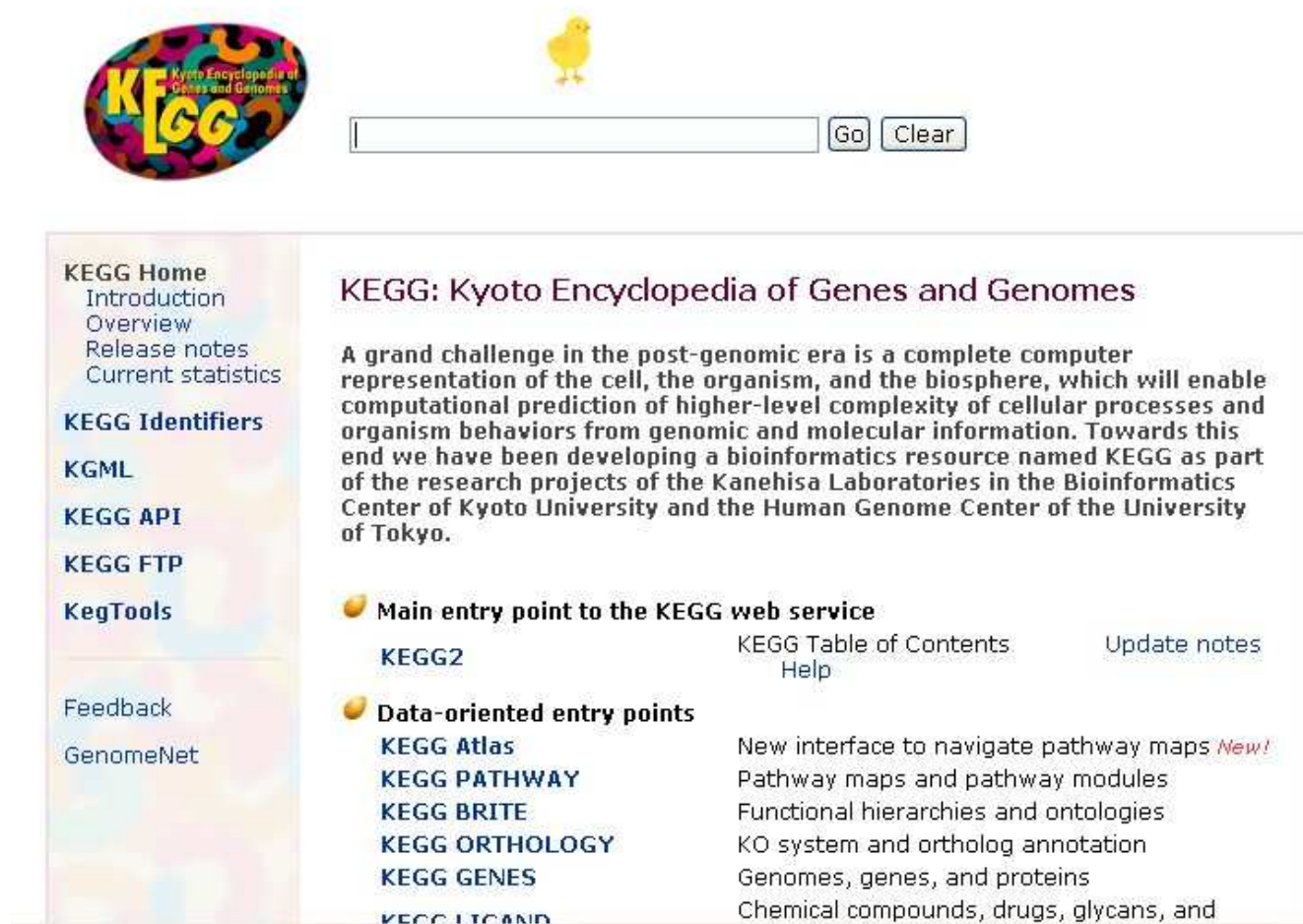
## Examples of questions that can be answered with OMIM in Entrez

- 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



**KEGG Home**  
 Introduction  
 Overview  
 Release notes  
 Current statistics

**KEGG Identifiers**

**KGML**

**KEGG API**

**KEGG FTP**

**KegTools**

Feedback  
 GenomeNet

### KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

**Main entry point to the KEGG web service**

**KEGG2**      KEGG Table of Contents      Update notes  
 Help

**Data-oriented entry points**

**KEGG Atlas**      New interface to navigate pathway maps *New!*  
**KEGG PATHWAY**      Pathway maps and pathway modules  
**KEGG BRITE**      Functional hierarchies and ontologies  
**KEGG ORTHOLOGY**      KO system and ortholog annotation  
**KEGG GENES**      Genomes, genes, and proteins  
**KEGG LIGAND**      Chemical compounds, drugs, glycans, and

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

## Secondary data bases in detail: KEGG pathways data base

KEGG PATHWAY Database  
Wiring diagrams of molecular interactions, reactions, and relations

KEGG2 ATLAS PATHWAY BRITE KO GENES SSDB LIGAND DBGET

Enter map number (Example) hsa05210

**Pathway Maps**

KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for:

- 1. Metabolism**  
Carbohydrate Energy Lipid Nucleotide Amino acid Other amino acid  
Glycan PK/NRP Cofactor/vitamin Secondary metabolite Xenobiotics Overview *New!*
- 2. Genetic Information Processing**
- 3. Environmental Information Processing**
- 4. Cellular Processes**
- 5. Human Diseases**
- 6. Drug Development**

and also on the structure relationships (KEGG drug structure maps) in:

KEGG Atlas may now be used to examine any of the KEGG pathway maps, in addition to the global metabolism map.

(<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

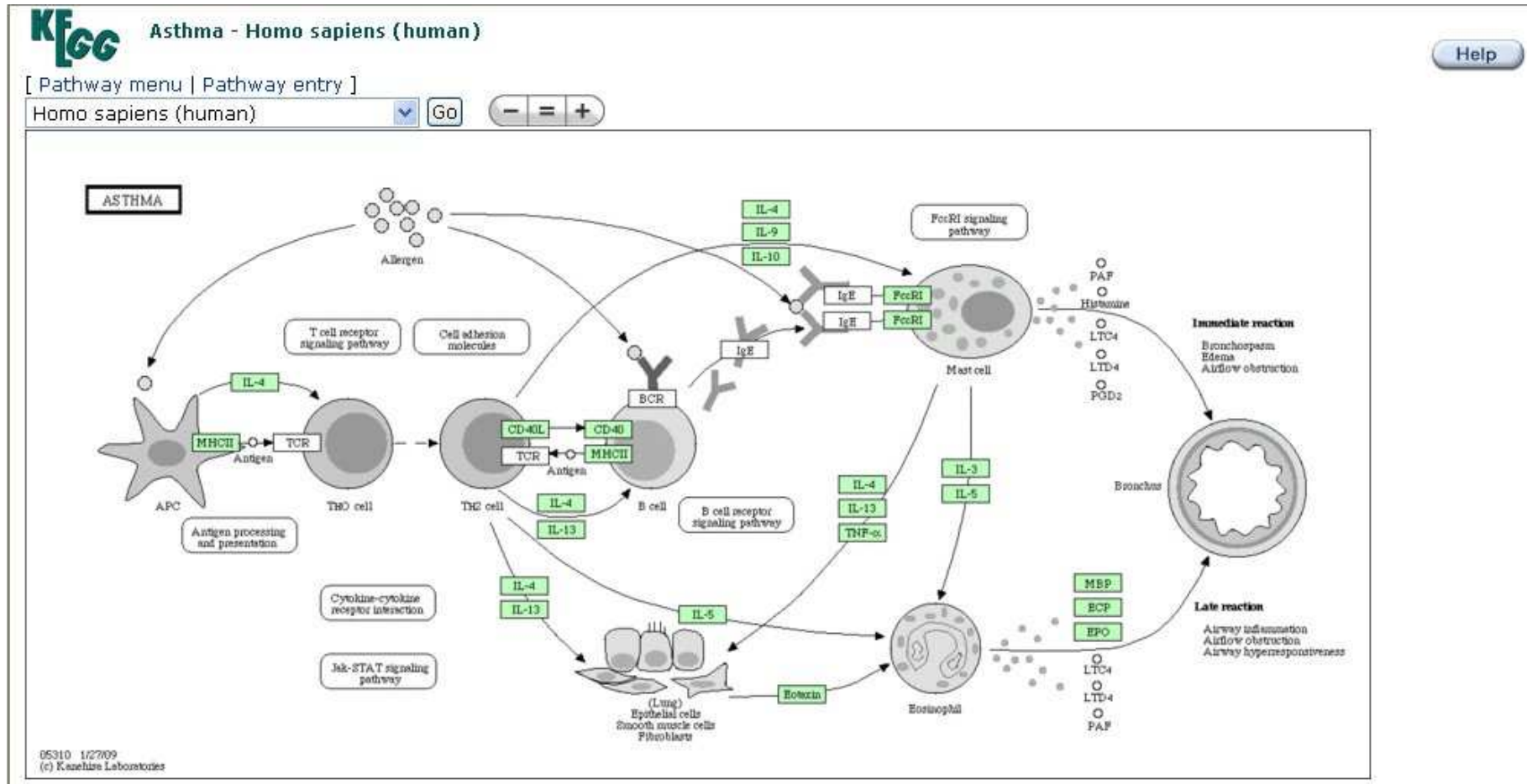
Alzheimer's disease *Revised!*  
Parkinson's disease *Revised!*  
Amyotrophic lateral sclerosis (ALS) *Revised!*  
Huntington's disease *Revised!*

### 5.4 Metabolic Disorders

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

### 5.5 Infectious Diseases

# KEGG pathway for asthma



([http://www.genome.ad.jp/kegg-bin/resize\\_map.cgi?map=hsa05310&scale=0.67](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

The screenshot displays the NCBI dbGAP website. At the top left is the NCBI logo. In the center is the dbGAP logo with the tagline "GENOTYPES and PHENOTYPES" and a graphic of a DNA helix and human figures. On the top right, there are links for "My NCBI", "Sign In", and "Register". Below the logo is a search bar containing "dbGaP" and a "for" dropdown menu, with "Go" and "Clear" buttons. A navigation bar includes tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". Below this is a "Browse dbGaP" section with "TUTORIAL" and "ABOUT dbGAP" buttons. The main content area has tabs for "By Studies", "By Diseases", and "Advanced Search". A table lists several studies with columns for Project, Study, Embargo Release, Details, Participants, and Type of Study.

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

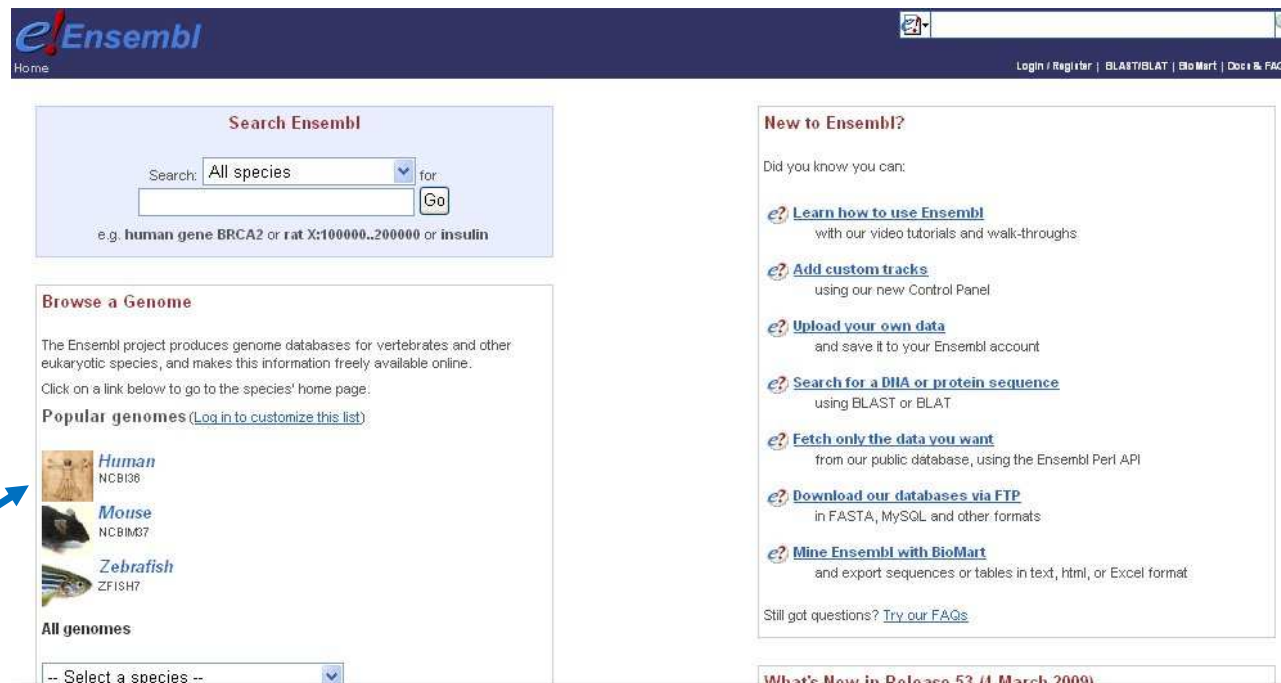
## 1.c Searching data bases

### 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

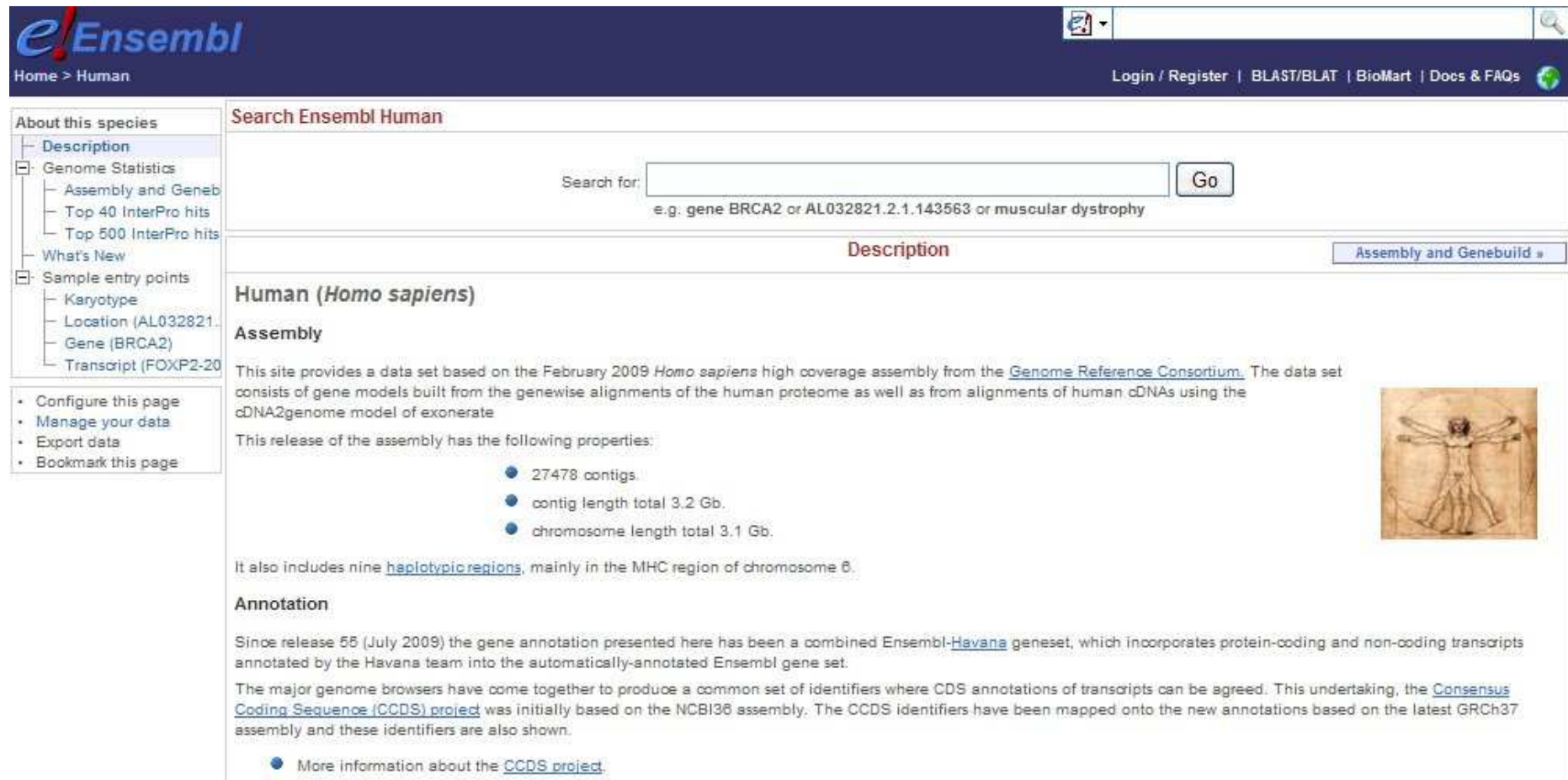


The screenshot shows the Ensembl homepage with the following elements:

- Search Ensembl:** A search box with a dropdown menu set to 'All species' and a 'Go' button. Below it, an example text reads: 'e.g. human gene BRCA2 or rat X:100000..200000 or insulin'.
- Browse a Genome:** A section with a description: 'The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page.' Below this is a link: 'Popular genomes (Log in to customize this list)'. A list of popular genomes is shown with small icons and text: 'Human NCBI36', 'Mouse NCBIM37', and 'Zebrafish ZFISH7'. Below this list is a dropdown menu labeled 'All genomes' with the text '-- Select a species --'. A blue arrow points to the 'Human' link.
- New to Ensembl?:** A section titled 'Did you know you can:' followed by several links: 'Learn how to use Ensembl' (with video tutorials), 'Add custom tracks' (using the Control Panel), 'Upload your own data' (to an Ensembl account), 'Search for a DNA or protein sequence' (using BLAST or BLAT), 'Fetch only the data you want' (using the Ensembl Perl API), 'Download our databases via FTP' (in FASTA, MySQL, etc.), and 'Mine Ensembl with BioMart' (exporting sequences or tables).
- Footer:** A link for 'What's New in Release 53 (4 March 2009)'.

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

# Ensembl genome browser portal



The screenshot displays the Ensembl genome browser interface for the Human species. At the top, the Ensembl logo is visible on the left, and navigation links for 'Login / Register', 'BLAST/BLAT', 'BioMart', and 'Docs & FAQs' are on the right. A search bar is prominently featured, with a 'Go' button and a placeholder text: 'e.g. gene BRCA2 or AL032821.2.1.143563 or muscular dystrophy'. Below the search bar, the 'Description' section is active, showing the title 'Human (*Homo sapiens*)' and the 'Assembly' section. The assembly description states it is based on the February 2009 high coverage assembly from the Genome Reference Consortium. Key properties listed include 27,478 contigs, a total contig length of 3.2 Gb, and a total chromosome length of 3.1 Gb. The 'Annotation' section mentions the Havana gene set and the CCDS project. A small image of Leonardo da Vinci's Vitruvian Man is shown on the right side of the page.

([http://www.ensembl.org/Homo\\_sapiens/Info/Index](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

**NCBI** National Center for Biotechnology Information  
National Library of Medicine National Institutes of Health

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

Search All Databases for  Go

**SITE MAP**  
Alphabetical List  
Resource Guide

**About NCBI**  
An introduction to NCBI

**GenBank**  
Sequence submission support and software

**Literature databases**  
PubMed, OMIM, Books, and PubMed Central

**Molecular databases**  
Sequences, structures, and

**What does NCBI do?**

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. [More about NCBI...](#)

**dbGaP: NCBI's Genome Wide Association Database**

NCBI's [dbGaP](#) (database of Genotypes and Phenotypes) provides data from Genome Wide Association Studies (GWAS), which are helping elucidate the link between genes and disease. For each study, users have access

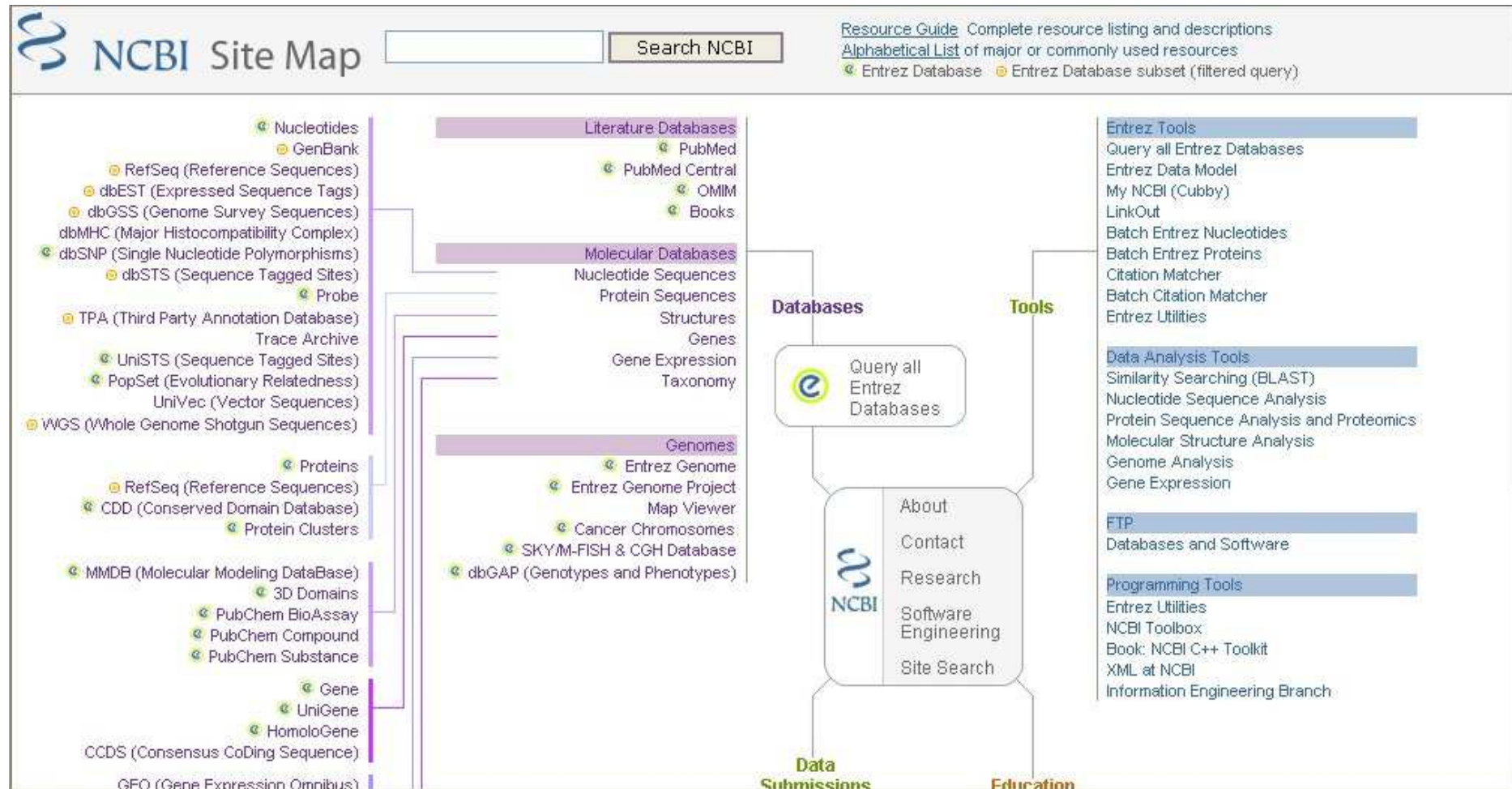
**Hot Spots**

- ▶ Clusters of orthologous groups
- ▶ Coffee Break, Genes & Disease, NCBI Handbook
- ▶ Electronic PCR
- ▶ Entrez Home
- ▶ Entrez Tools
- ▶ Gene expression omnibus (GEO)
- ▶ Human genome resources
- ▶ Influenza Virus Resource

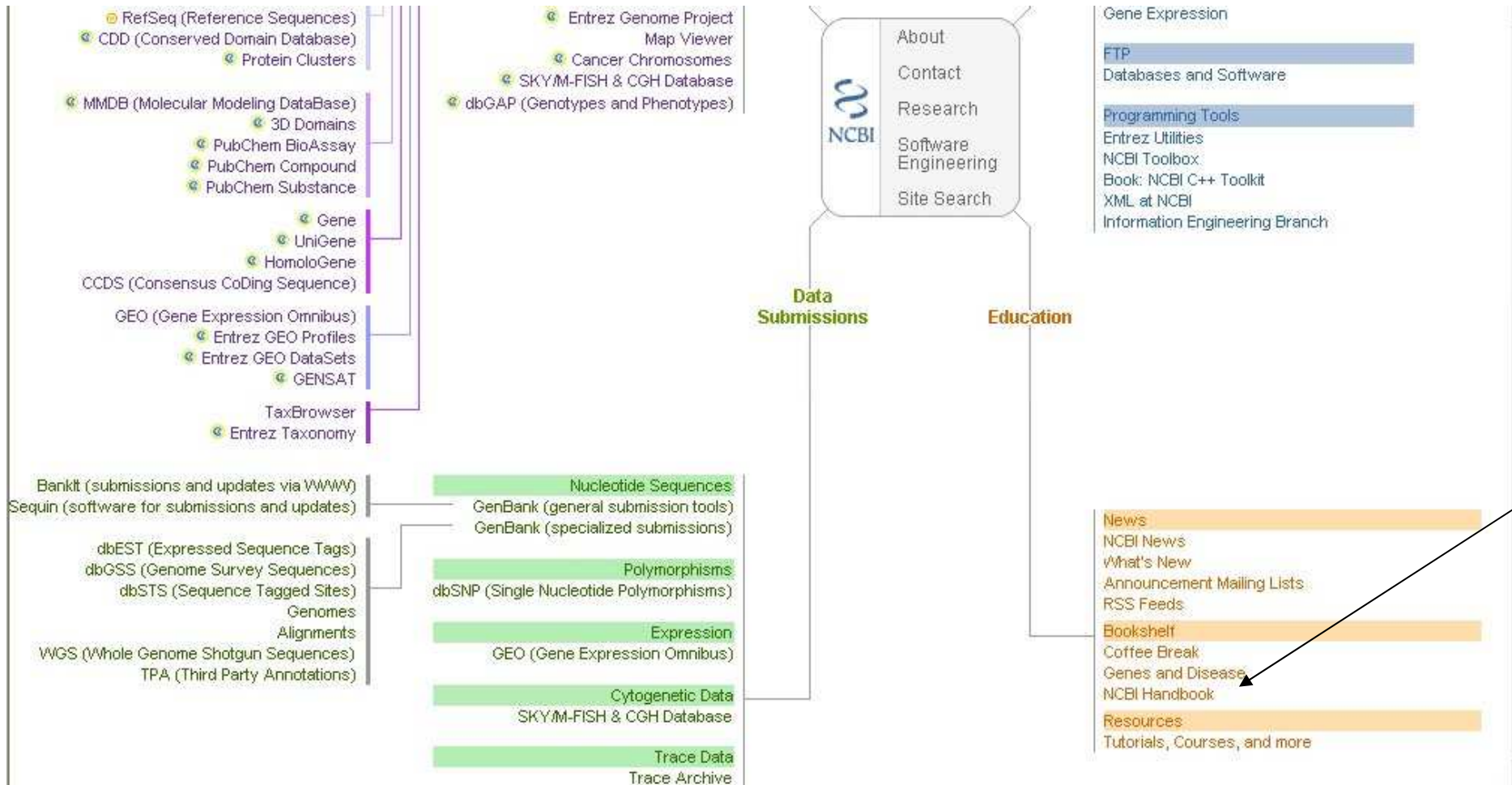
(<http://www.ncbi.nlm.nih.gov/>)



# NCBI Site Map



# NCBI Site Map (continued)



## NCBI Handbook



Navigation

→ [About this book](#)

**[Part 1. The Databases](#)**

**[Part 2. Data Flow and Processing](#)**

**[Part 3. Querying and Linking the Data](#)**

**[Part 4. User Support](#)**

**[Glossary](#)**

Search

This book  All books

PubMed

### The NCBI Handbook

Bioinformatics consists of a computational approach to biomedical information management and analysis. It is being used increasingly as a component of research within both academic and industrial settings and is becoming integrated into both undergraduate and postgraduate curricula. The new generation of biology graduates is emerging with experience in using bioinformatics resources and, in some cases, programming skills.

The National Center for Biotechnology Information (NCBI) is one of the world's premier Web sites for biomedical and bioinformatics research. Based within the National Library of Medicine at the National Institutes of Health, USA, the NCBI hosts many databases used by biomedical and research professionals. The services include PubMed, the bibliographic database; GenBank, the nucleotide sequence database; and the BLAST algorithm for sequence comparison, among many others. The NCBI Web site is visited by about 250,000 people per day.

Although each NCBI resource has online help documentation associated with it, there is no cohesive approach to describing the databases and search engines, nor any significant information on how the databases work or how they can be leveraged, for bioinformatics research on a larger scale. The NCBI Handbook is designed to address this information gap.

All of our users know how to execute a straightforward PubMed or BLAST search. However, feedback from help desk personnel and booth staff at scientific meetings suggests that people often want to know how to use our resources in a more sophisticated manner and are frequently unaware of less well-known databases that might be helpful to them. The intended audience for The NCBI Handbook is, therefore, the growing number of scientists and students who would like a more in-depth guide to NCBI resources—powerusers and aspiring powerusers.

The NCBI Handbook is focused on the relatively stable information about each resource; it is not a point-and-click user guide (this type of information can be found in the online help documents, referred to frequently but not repeated, in the Handbook). Each chapter is devoted to one service; after a brief overview on using the resource, there is an account of how the resource works, including topics such as how data are included in a database, database design, query processing, and how the different resources relate to each other. For example, the BLAST chapter briefly describes what to use BLAST for, how to use BLAST, and how to use BLAST to find related sequences.

## 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

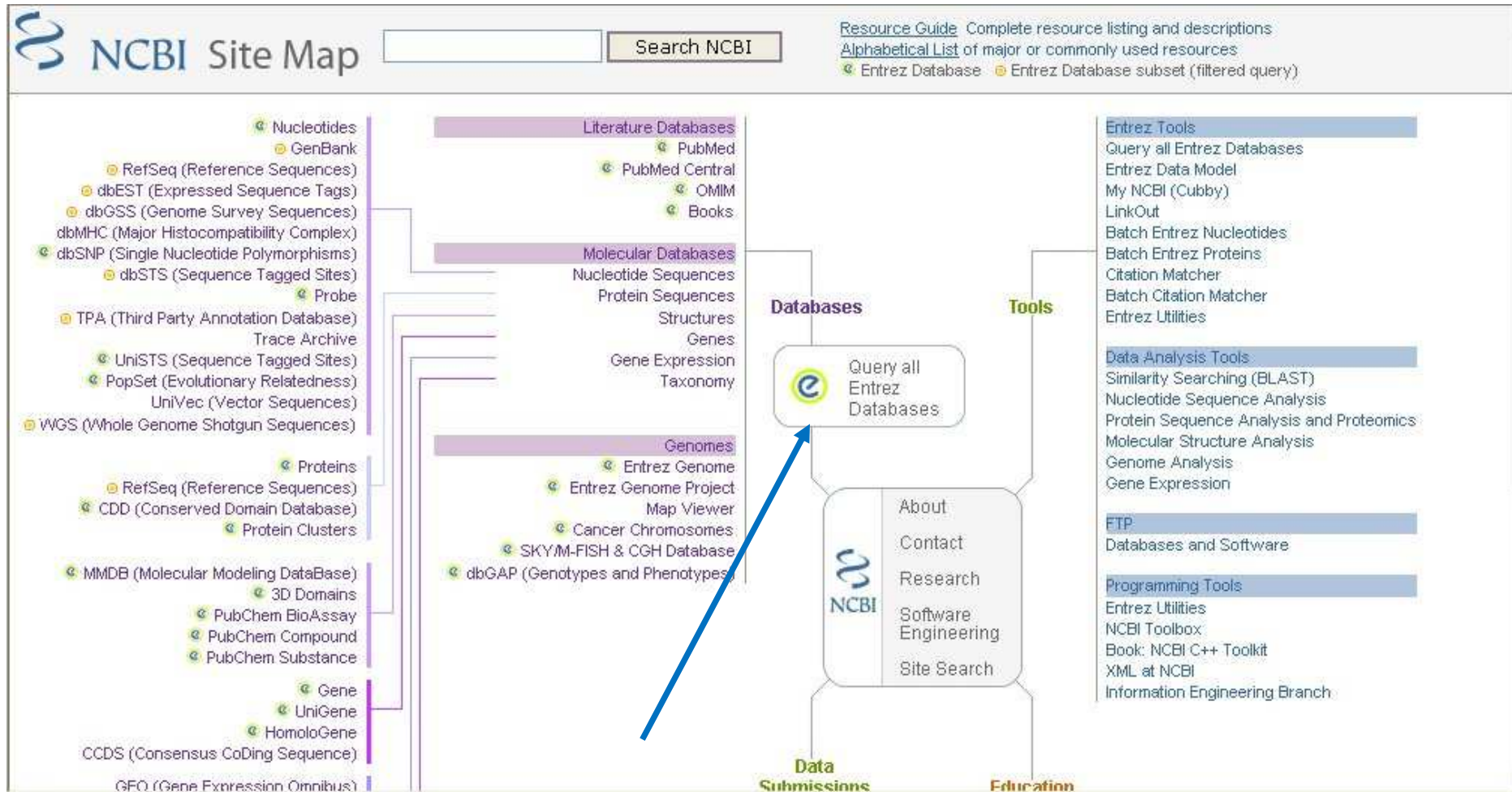
#### 22. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes

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

NCBI

Entrez, The Life Sciences Search Engine

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank Map Viewer BLAST

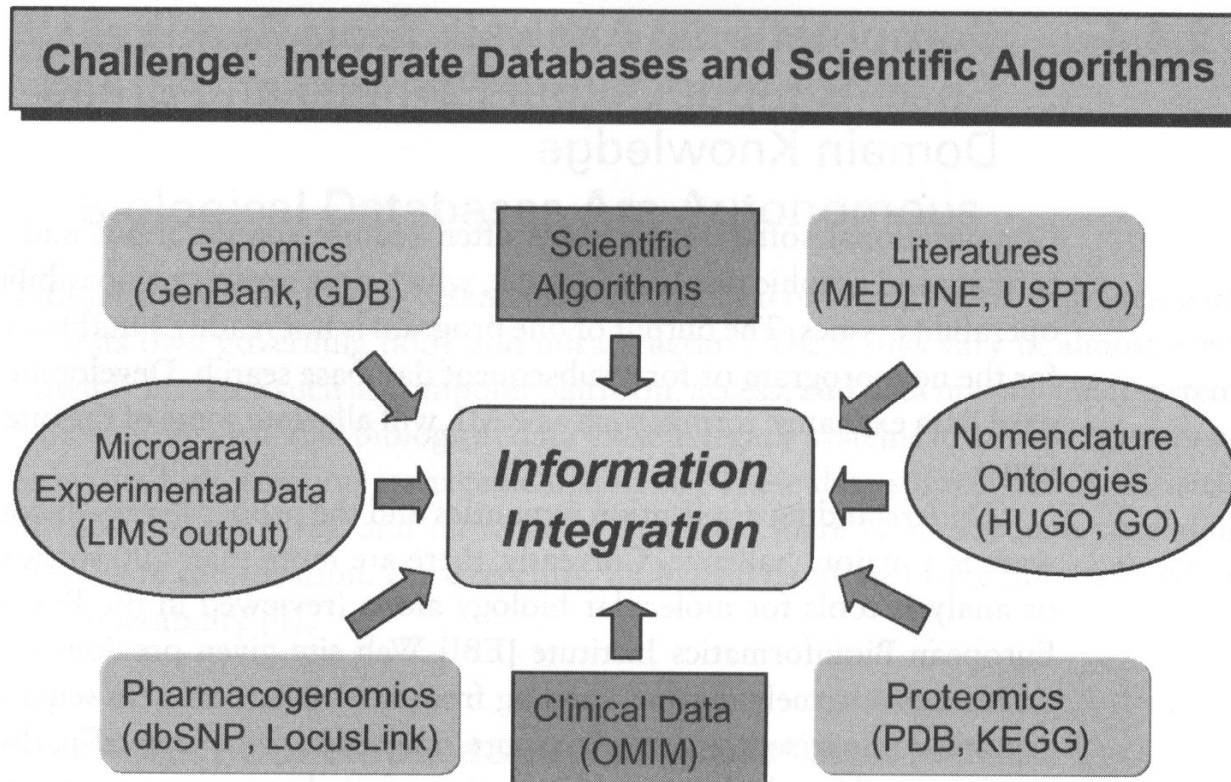
Search across databases    [Help](#)

Welcome to the Entrez cross-database search page

<b>PubMed:</b> biomedical literature citations and abstracts <input type="button" value="GO"/>	<b>Books:</b> online books <input type="button" value="GO"/>
<b>PubMed Central:</b> free, full text journal articles <input type="button" value="GO"/>	<b>OMIM:</b> online Mendelian Inheritance in Man <input type="button" value="GO"/>
<b>Site Search:</b> NCBI web and FTP sites <input type="button" value="GO"/>	<b>OMIA:</b> online Mendelian Inheritance in Animals <input type="button" value="GO"/>
<b>Nucleotide:</b> Core subset of nucleotide sequence records <input type="button" value="GO"/>	<b>dbGaP:</b> genotype and phenotype <input type="button" value="GO"/>
<b>EST:</b> Expressed Sequence Tag records <input type="button" value="GO"/>	<b>UniGene:</b> gene-oriented clusters of transcript sequences <input type="button" value="GO"/>
<b>GSS:</b> Genome Survey Sequence records <input type="button" value="GO"/>	<b>CDD:</b> conserved protein domain database <input type="button" value="GO"/>
<b>Protein:</b> sequence database <input type="button" value="GO"/>	<b>3D Domains:</b> domains from Entrez Structure <input type="button" value="GO"/>
<b>Genome:</b> whole genome sequences <input type="button" value="GO"/>	<b>UniSTS:</b> markers and mapping data <input type="button" value="GO"/>
<b>Structure:</b> three-dimensional macromolecular structures <input type="button" value="GO"/>	<b>PopSet:</b> population study data sets <input type="button" value="GO"/>

(<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](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]



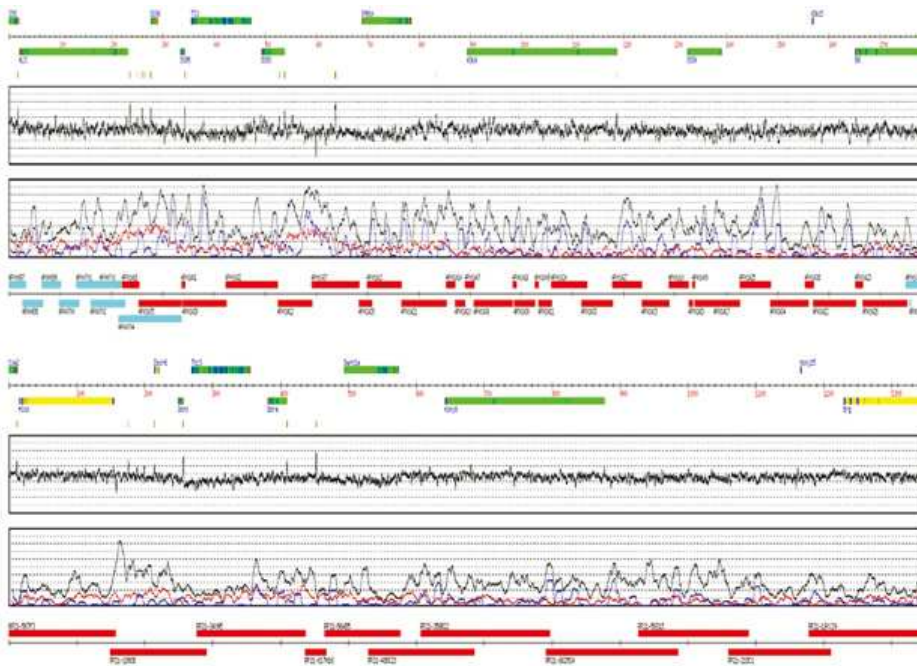
## In-class discussion document

- 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](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](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.