



# IX International Course of Massive Data Analysis FOR GENOMICS



# Index

---

- Introduction
- Biological Databases
- Ensembl Biomart
- CellBase
- Hands on

# Introduction

## NAR Biological Database Collection

---

- ***Nucleic Acid Research*** compilation currently lists 1512 online databases!
  - <http://www.oxfordjournals.org/nar/database/c>
- Features:
  - Many different databases for each category, which should I use?
  - No standards: different IDs, methods, servers, formats, ...
  - Lack of international initiatives, many local and small databases
  - Different gene IDs, more than 50
  - *In vivo* vs *in silico* databases

# Introduction

## Data repositories

---

- Data in biology is open and available for all scientific community
- Microarrays repositories:
  - GEO <http://www.ncbi.nlm.nih.gov/geo>
  - ArrayExpress <http://www.ebi.ac.uk/arrayexpress>
- NGS repositories:
  - SRA <http://www.ncbi.nlm.nih.gov/sra>
  - ENA <http://www.ebi.ac.uk/ena/about/about>
  - 1000Genomes <http://www.1000genomes.org>
  - ENCODE <http://encodeproject.org>

# Introduction

## Sequence databases

### Genome Reference Consortium (GRC)

<http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/index.shtml>

The **GRC** is a collaborative effort and only works with input from the larger scientific community.

We strive to work closely with external groups to gather all relevant data.

The **GRC** is now working to create **assemblies** that better represent this **diversity** and provide more robust substrates for genome analysis.

**Genome Reference Consortium**

[GRC Home](#) | [Data](#) | [Help](#) | [Report an Issue](#) | [Contact Us](#) | [Credits](#) | [Curators Only](#)

[Human Overview](#) | [Human Issues under Review](#) | [Human Assembly Data](#) | [Report a problem](#)

### Human Genome Overview

Information concerning the continuing improvement of the human genome.

◀ Regions containing alternate-loci  
■ Regions containing fix patches  
● Regions containing novel patches

An ideogram representation of the latest human assembly, GRCh37.p5 (not showing unplaced or unlocalized sequences).

The GRC is working hard to provide the best possible reference assembly for human. We do this by both generating multiple representations ( [alternate loci](#) ) for regions that are too complex to be represented by a single path. Additionally, we are releasing regional fixes known as [patches](#) . This allows users who are interested in a specific locus to get an improved representation without affecting users who need chromosome coordinate stability.

**Getting Data**  
GRCh37 (Latest Major Release): [FTP](#)  
GRCh37 patch release 5 (Latest Minor Release): [FTP](#)  
Information on regions under review: [FTP](#)

**Next assembly update**  
The next assembly update (patch release 6) will be a minor update (only patches) and will happen in Sep 2011

# Introduction

## Sequence databases


### Mission:



- To provide freely available **data** and bioinformatics **services** to all facets of the scientific community in ways that promote scientific progress.
- To contribute to the advancement of biology through basic investigator-driven **research** in bioinformatics.
- To provide advanced bioinformatics **training** to scientists at all levels, from PhD students to independent investigators.
- To help **disseminate** cutting-edge technologies to industry.

Funded by EMBL

## European Bioinformatics Institute (EBI)

<http://www.ebi.ac.uk/>

EMBL-EBI  European Bioinformatics Institute

Databases Tools Research Training Industry About Us Help Site Index  

Explore the EBI:

[FIND](#)

Examples: [ROA1\\_HUMAN](#), [tpi1](#), [Sulston...](#) [Help](#) [Feedback](#)

### Data Resources and Tools

- [ENA](#)
- [UniProt](#)
- [ArrayExpress](#)
- [Ensembl](#)
- [InterPro](#)
- [PDBe](#)
- [Genomes](#)
- [Nucleotide Sequences](#)
- [Protein Sequences](#)
- [Macromolecular Structures](#)
- [Small Molecules](#)
- [Gene Expression](#)
- [Protein Expression](#)
- [Molecular Interactions](#)
- [Reactions & Pathways](#)
- [Protein Families](#)
- [Enzymes](#)
- [Literature](#)
- [Taxonomy](#)
- [Ontologies](#)
- [Patent Resources](#)
- [Sequence Similarity & Analysis](#)
- [Pattern & Motif Searches](#)
- [Structure Analysis](#)
- [Text Mining](#)
- [Downloads](#)
- [Web Services](#)

# Introduction

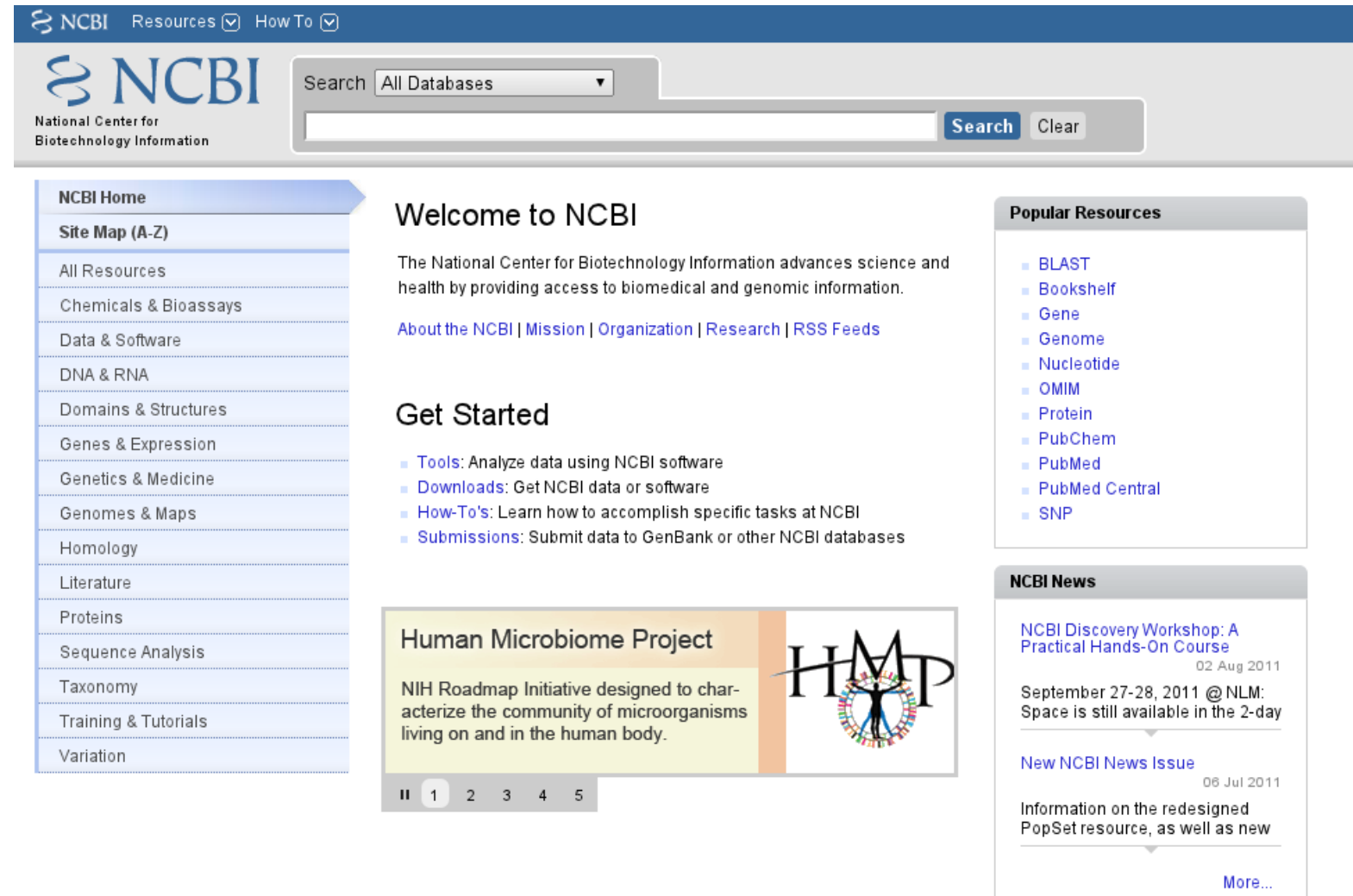
## Sequence databases

### National Center for Biotechnology Information (NCBI)

Set of **tools** and **databases** for genomic and biomedical studies and analysis.

Financed by the USA.  
Is the American competence of the EBI in both objectives and resources.

<http://www.ncbi.nlm.nih.gov/guide/>



The screenshot shows the NCBI homepage with a navigation menu on the left, a search bar at the top, and several content sections. The navigation menu includes: NCBI Home, Site Map (A-Z), All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation. The main content area features a 'Welcome to NCBI' message, a 'Get Started' section with links to Tools, Downloads, How-To's, and Submissions, and a 'Human Microbiome Project' banner. On the right, there are 'Popular Resources' and 'NCBI News' sections.

**NCBI Home**

- Site Map (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

**Welcome to NCBI**

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

**Get Started**

- Tools:** Analyze data using NCBI software
- Downloads:** Get NCBI data or software
- How-To's:** Learn how to accomplish specific tasks at NCBI
- Submissions:** Submit data to GenBank or other NCBI databases

**Human Microbiome Project**

NIH Roadmap Initiative designed to characterize the community of microorganisms living on and in the human body.

**Popular Resources**

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

**NCBI News**

[NCBI Discovery Workshop: A Practical Hands-On Course](#)  
02 Aug 2011  
September 27-28, 2011 @ NLM:  
Space is still available in the 2-day

[New NCBI News Issue](#)  
06 Jul 2011  
Information on the redesigned PopSet resource, as well as new

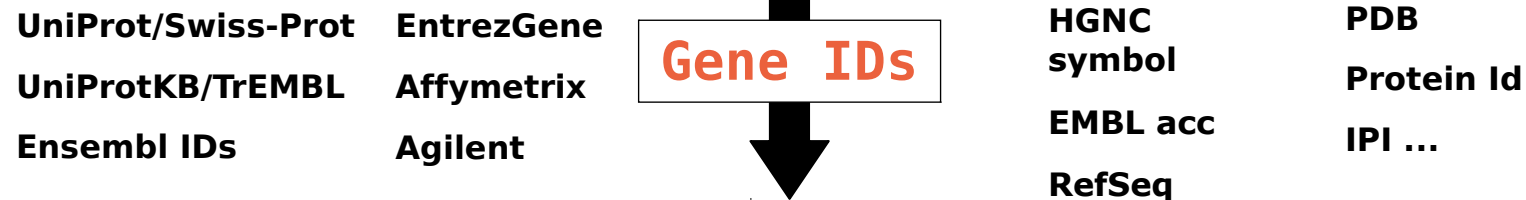
[More...](#)

# Introduction

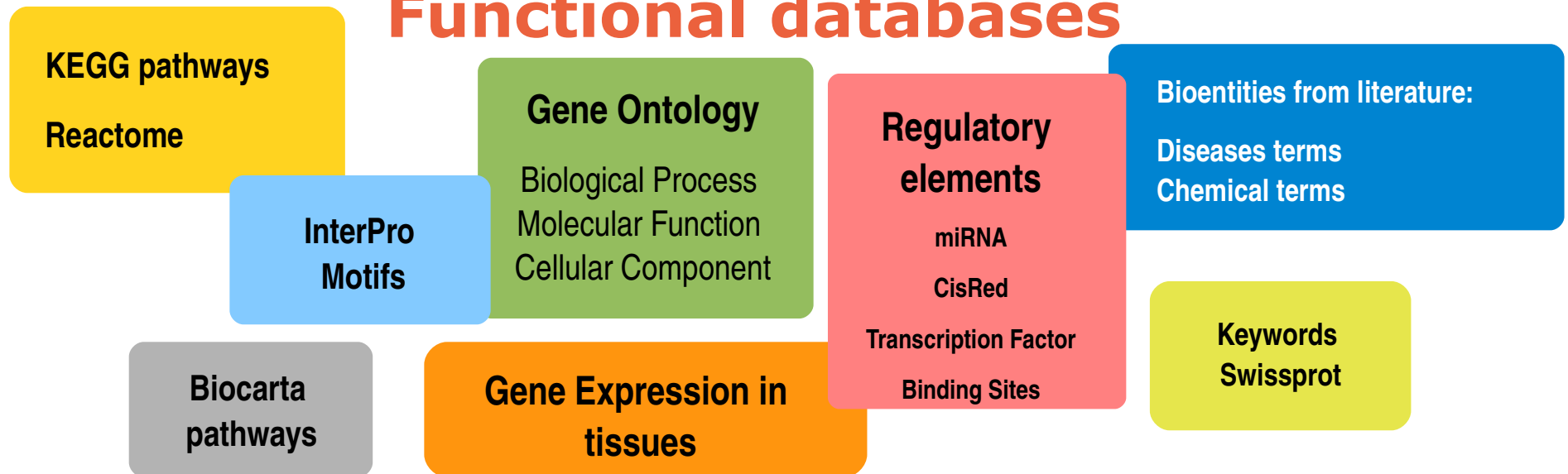
## Functional Annotation

### Overview

Some of the biological databases contain *Functional Information* of genes and sequences



## Functional databases





# Introduction

## Functional Annotation

### Gene Ontology (GO terms)

- The *Gene Ontology* project provides a **controlled vocabulary** to describe gene and gene product attributes in any organism.
- Latest version has **38137** terms (March, 2013)
- The controlled vocabularies of terms are structured.

<http://www.geneontology.org/>

# Introduction

## Functional Annotation

The three categories of GO:

### Molecular Function

The tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*

### Biological Process

Broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions

### Cellular Component

Subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*

Gene Ontology (GO terms)

- [-] [GO:0003673 : Gene Ontology \(65883\)](#)
- [-] [GO:0008150 : biological process \(44405\)](#)
  - [+] [GO:0007610 : behavior \(357\)](#)
    - [GO:0000004 : biological process unknown \(7877\)](#)
  - [-] [GO:0009987 : cellular process \(32672\)](#)
    - [+] [GO:0007154 : cell communication \(5384\)](#)
    - [+] [GO:0008219 : cell death \(744\)](#)
    - [+] [GO:0030154 : cell differentiation \(464\)](#)
    - [+] [GO:0008151 : cell growth and/or maintenance \(28802\)](#)
    - [+] [GO:0006928 : cell motility \(911\)](#)
    - [+] [GO:0006944 : membrane fusion \(257\)](#)
  - [+] [GO:0016265 : death \(793\)](#)
  - [+] [GO:0007275 : development \(4615\)](#)
  - [+] [GO:0008371 : obsolete \(1581\)](#)
  - [+] [GO:0007582 : physiological processes \(31124\)](#)
  - [+] [GO:0016032 : viral life cycle \(115\)](#)
- [+] [GO:0005575 : cellular component \(32869\)](#)
- [+] [GO:0003674 : molecular function \(53910\)](#)

# Introduction

## Functional Annotation

Gene Ontology (GO terms)

- **AmiGO** provides a web interface to search and browse the ontology and annotation data

<http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>

- **QuickGO** (EBI) provides also a web interface

<http://www.ebi.ac.uk/ego>

# Introduction

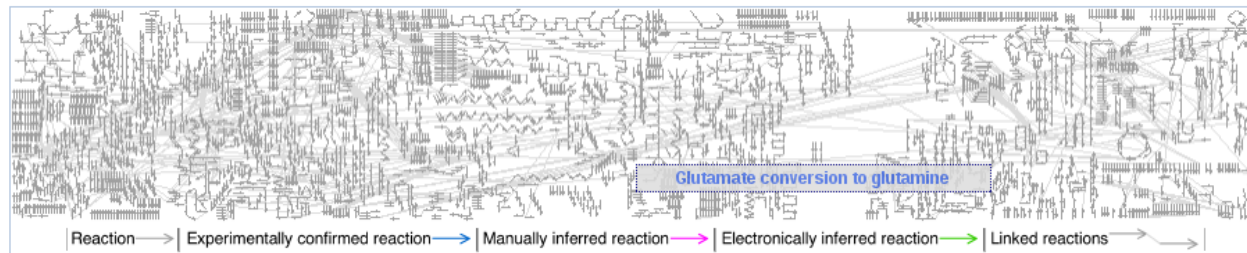
## Functional Annotation

It is a free, online, open-source, **curated pathway database** encompassing many areas of human biology. Information is authored by expert biological researchers.

<http://www.reactome.org/>

### Reactome - a curated knowledgebase of biological pathways

The data displayed is for **Homo sapiens** Use the menu to change the species. Check  for cross-species comparison.



Apoptosis	Axon guidance	Biological oxidations	Botulinum neurotoxicity
Cell junction organization	Cell Cycle Checkpoints	Cell Cycle, Mitotic	DNA Repair
DNA Replication	Diabetes pathways	Electron Transport Chain	Gap junction trafficking and regulation
Gene Expression	Hemostasis	HIV Infection	Influenza Infection
Integration of energy metabolism	Integrin cell surface interactions	Metabolism of lipids and lipoproteins	Membrane Trafficking
Metabolism of amino acids and derivatives	Metabolism of carbohydrates	Metabolism of nitric oxide	Metabolism of nucleotides
Metabolism of polyamines	Metabolism of porphyrins	Metabolism of proteins	Metabolism of RNA
Metabolism of vitamins and cofactors	Muscle contraction	mRNA Processing	Myogenesis
Pyruvate metabolism and Citric Acid (TCA) cycle	Regulation of beta-cell development	Regulatory RNA pathways	Signaling by BMP
Signaling by EGFR	Signaling by FGFR	Signaling by GPCR	Signaling by PDGF
Signaling in Immune system	Signaling by Insulin receptor	Signaling by NGF	Signaling by Notch
Opioid Signalling	Signaling by Rho GTPases	Signaling by TGF beta	Signaling by VEGF
Signaling by Wnt	<b>Synaptic Transmission</b>	Telomere Maintenance	Transcription
Transmembrane transport of small molecules			

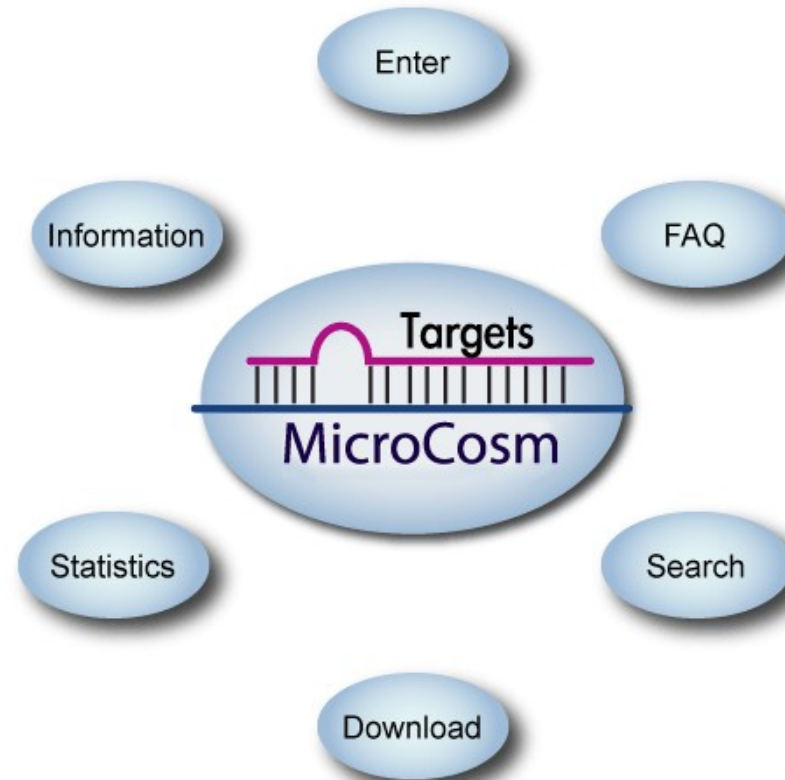
# Introduction

## Functional Annotation

microRNAs



- Involved in gene **regulation**.
- Last version has **16,772** entries (Release 17, April 2011).
- The **target database** contains computationally predicted targets for microRNAs across many species.



<http://www.mirbase.org/>

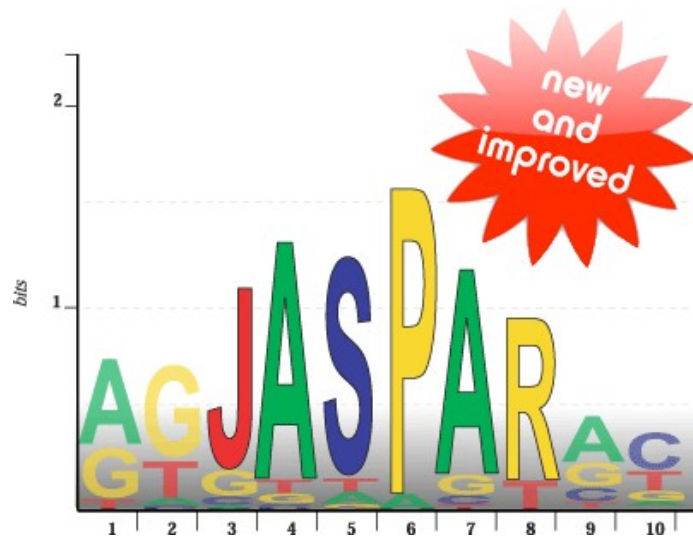
<http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/>

# Introduction

## Functional Annotation

### Jaspar Transcription Factor Binding Sites

- The **JASPAR** database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes.
- The prime difference to similar resources (**TRANSFAC**, etc) consist of the open data access, non-redundancy and quality.



<http://jaspar.genereg.net/>

# Introduction

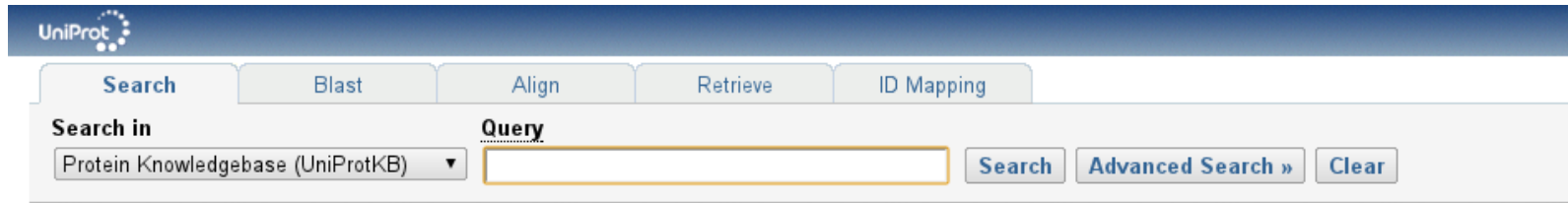
## Protein databases

### UniProtKB/Swiss-Prot

contains 531473 sequence entries

UniProt, protein sequence and information

<http://www.uniprot.org/>



### WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### What we provide

UniProtKB	Protein knowledgebase, consists of two sections: <ul style="list-style-type: none"><li>★ Swiss-Prot, which is manually annotated and reviewed.</li><li>★ TrEMBL, which is automatically annotated and is <b>not</b> reviewed.</li></ul> Includes <a href="#">Complete Proteome Sets</a> .
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	<a href="#">Literature citations</a> , <a href="#">taxonomy</a> , <a href="#">keywords</a> and <a href="#">more</a> .

### NEWS

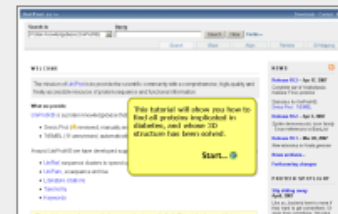
#### UniProt release 2011\_08 - Jul 27, 2011

UniProt collaboration with IMEx for the annotation of protein interactions to MIMx standard

- > [Statistics for UniProtKB: Swiss-Prot · TrEMBL](#)
- > [Forthcoming changes](#)
- > [News archives](#)

[Follow @uniprot](#)

### SITE TOUR



# Introduction

## Protein databases

### *InterPro*, protein annotation database

A centralized database of **protein families, domains, repeats and sites** in which identifiable features found in known proteins can be applied to new protein sequences.

<http://www.ebi.ac.uk/interpro/>

#### Member database information

Signature Database	Version	Signatures*	Integrated Signatures**
GENE3D	3.3.0	<a href="#">2386</a>	<a href="#">1377</a>
HAMAP	021210	<a href="#">1675</a>	<a href="#">1429</a>
PANTHER	7.0	<a href="#">80933</a>	<a href="#">1777</a>
PIRSF	2.74	<a href="#">3248</a>	<a href="#">2791</a>
PRINTS	41.1	<a href="#">2050</a>	<a href="#">2009</a>
PROSITE patterns	20.66	<a href="#">1308</a>	<a href="#">1292</a>
PROSITE profiles	20.66	<a href="#">901</a>	<a href="#">877</a>
Pfam	24.0	<a href="#">11912</a>	<a href="#">11465</a>
PfamB	24.0	<a href="#">142303</a>	<a href="#">0</a>
ProDom	2006.1	<a href="#">1894</a>	<a href="#">1008</a>
SMART	6.1	<a href="#">895</a>	<a href="#">882</a>
SUPERFAMILY	1.73	<a href="#">1774</a>	<a href="#">1154</a>
TIGRFAMs	9.0	<a href="#">3808</a>	<a href="#">3796</a>

#### Contents of InterPro 31.0 (Feb 2011)

Active site	<a href="#">97</a>
Binding site	<a href="#">65</a>
Conserved site	<a href="#">615</a>
Domain	<a href="#">5936</a>
Family	<a href="#">14194</a>
PTM	<a href="#">16</a>
Repeat	<a href="#">262</a>



# Introduction

## Protein databases

*IntAct*, protein-protein interaction database

<http://www.ebi.ac.uk/intact/main.xhtml>

*IntAct* provides a freely available, open source database system and analysis tools for protein interaction data.

All interactions are derived from **literature curation** or direct user submissions and are freely available.

The screenshot shows the IntAct website interface. At the top, there is a navigation bar with the EMBL-EBI logo and a search bar. Below the navigation bar, there is a main menu with options like Databases, Tools, Research, Training, Industry, About Us, and Help. The IntAct logo is prominently displayed on the left side. The main content area includes a search bar, a navigation menu with options like Home, Search, Interactions (272325), Browse, Lists, Interaction Details, Molecule View, and Graph. A 'Search IntAct' section provides instructions on how to perform a search and lists examples of search criteria. A 'Basic Statistics' box on the right side of the page displays the following information:

- 268,920 binary interactions.
- 57,741 proteins.
- 13,802 experiments.
- 1,706 controlled vocabulary terms.

Below the statistics, there is a section for 'Dataset of the month: September' featuring a selected reaction monitoring mass spectrometry study by Bisson et al. The study is associated with PSI-MI 2.5 and PSI-MI 1.0 datasets.

# Introduction

## Variation databases

dbSNP, the repository of all the SNPs

<http://www.ncbi.nlm.nih.gov/projects/SNP/>

NCBI dbSNP Short Genetic Variations

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for [ ] Go

ANNOUNCEMENT

8/15/2011: dbSNP Build 134 Release

Please see the build announcement for more details  
(<http://www.ncbi.nlm.nih.gov/projects/SNP/docs/b>)

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

GENERAL

RSS Feed

Contact Us

Site Map

dbSNP Homepage

Announcements

dbSNP Summary

FTP Download

HUMAN VARIATION

SNP SUBMISSION

DOCUMENTATION

SEARCH

RELATED SITES

**Search by IDs on All Assemblies**

Note: **rs#** and **ss#** must be prefixed with "rs" or "ss", respectively

ID: [ ] Reference cluster ID(rs#) [ ]

Search Reset

**Submission Information**

- By Submitter
- New Submitted Batches
- Method
- Population
- Publication

August 2011

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>	134	<a href="#">37.2</a>	<a href="#">179,506,198</a>	41,365,915 (6,961,883)	<a href="#">16,880,922</a>	73,208,602	35,627,484
<a href="#">Mus musculus</a>	132	<a href="#">37.1</a>	<a href="#">26,991,031</a>	15,522,011 (6,439,098)	<a href="#">6,696,618</a>		
<a href="#">Pongo abelii</a>	132		<a href="#">10,225,850</a>	10,065,309 (0)			
<a href="#">Pongo pygmaeus</a>	127		<a href="#">7,854,083</a>	7,854,081 (0)			
<a href="#">Rattus norvegicus</a>	130	<a href="#">4.1</a>	<a href="#">6,472,989</a>	119,436 (1,605)	<a href="#">1,024,738</a>		
<a href="#">Gallus gallus</a>	131	<a href="#">2.1</a>	<a href="#">11,318,097</a>	3,504,588 (3,269,983)	<a href="#">1,452,147</a>		50
<a href="#">Glycine max</a>	127		<a href="#">6,378,350</a>	6,352,034 (234)			
<a href="#">Phoenix dactylifera</a>	133		<a href="#">3,518,029</a>	3,429,753 (0)			
<a href="#">Zea mays</a>	128		<a href="#">4,555,638</a>	4,350,627 (80)			
<a href="#">Oryza sativa</a>	128	<a href="#">4.1</a>	<a href="#">5,872,306</a>	5,359,569 (21,773)	<a href="#">1,897,895</a>		
<a href="#">Ovis aries</a>	128		<a href="#">2,899,286</a>	2,899,215 (66)			91
<a href="#">Bos taurus</a>	131	<a href="#">4.1</a>	<a href="#">4,931,454</a>	2,210,557 (13,881)	<a href="#">677,906</a>		446
<a href="#">Canis familiaris</a>	131	<a href="#">2.1</a>	<a href="#">3,527,071</a>	3,258,962 (214,713)	<a href="#">982,946</a>		17

# Introduction

## Variation databases

### *HapMap*, Human Haplotype Map

The goal is to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.

The banner features the International HapMap Project logo on the left, which includes a world map and a DNA double helix. To the right, the text "International HapMap Project" is displayed in a large, bold font. Below this, a navigation menu lists "Home", "About the Project", "Data", "Publications", and "Tutorial".

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	News
<p><a href="#">About the Project</a></p> <p><a href="#">HapMap Publications</a></p> <p><a href="#">HapMap Tutorial</a></p> <p><a href="#">HapMap Mailing List</a></p> <p><a href="#">HapMap Project Participants</a></p>	<ul style="list-style-type: none"><li>2011-06-13: <b>HapMap help desk announcement</b></li></ul> <p>There was a problem with the HapMap help desk system. In the past several weeks, emails sent to hapmap-help@ncbi.nlm.nih.gov did not reach the help desk, and thus user requests were not addressed. Please resend your email request if you sent emails to the HapMap help desk in the past several weeks. Sorry for the inconvenience.</p> <ul style="list-style-type: none"><li>2011-04-20: <b>Hapmap help desk service interruption notice</b></li></ul> <p>There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.</p> <ul style="list-style-type: none"><li>2011-02-02: <b>Haploview issues with rel 28 data</b></li></ul> <p>Recently, there are several questions about Haploview data format errors when users tried to analyze HapMap release 28 data. The current Haploview version (4.2) does not recognize the new individuals in release 28 and the software will generate an error similar to "Hapmap data format error: NA18876" when trying to open the data.</p> <p>Haploview is developed and maintained by an organization different from HapMap. Please contact Haploview help desk (haploview@broadinstitute.org) for questions specific to this software.</p> <ul style="list-style-type: none"><li>2011-01-19: <b>HapMap phase II recombination rate on GRCh37</b></li></ul> <p>The leftover of the HapMap II genetic map from human genome build b35 to GRCh37 is available. Data is <b>available for bulk download</b>.</p> <ul style="list-style-type: none"><li>2010-08-18: <b>HapMap Public Release #28</b></li></ul>
<p><b>Project Data</b></p> <p><a href="#">HapMap Genome Browser release #28 ( Phases 1, 2 &amp; 3 - merged genotypes &amp; frequencies )</a></p> <p><a href="#">HapMap3 Genome Browser release #3 ( Phase 3 - genotypes &amp; frequencies )</a></p> <p><a href="#">HapMap Genome Browser release #27 ( Phase 1, 2 &amp; 3 - merged genotypes &amp; frequencies )</a></p> <p><a href="#">HapMap3 Genome Browser release #2 ( Phase 3 - genotypes, frequencies &amp; LD )</a></p> <p><a href="#">HapMap Genome Browser release#24 ( Phase 1 &amp; 2 - full dataset )</a></p> <p><a href="#">GWAs Karyogram</a></p>	

# Introduction

## Variation databases

Mutations: OMIM, COSMIC, Mitelman, etc.

<http://www.ncbi.nlm.nih.gov/omim>

NCBI Online Mendelian Inheritance in Man

Search OMIM for [ ] Go Clear

Entrez

OMIM

- 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.

NCBI is implementing changes to help you find current content in OMIM based on resources at NCBI, and then directing you to [omim.org](http://omim.org). Please be aware that you will leave NCBI to view OMIM records. Access to full records from NCBI (e.g. web, ftp, entrez) will no longer be supported.

**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.

<http://www.sanger.ac.uk/genetics/CGP/cosmic/>

COSMIC Catalogue Of Somatic Mutations In Cancer

**What is COSMIC?**

All cancers arise as a result of the acquisition of a series of fixed DNA sequence abnormalities, mutations, many of which ultimately confer a growth advantage upon the cells in which they have occurred. There is a vast amount of information available in the published scientific literature about these changes. COSMIC is designed to store and display somatic mutation information and related details and contains information relating to human cancers. [\[more\]](#)

**News**

12th Jul 2011  
**COSMIC v54 Release**  
COSMIC v54 Release  
Five new cancer genes have received full curation of their mutation spectrum, together with seven new fusion ...

**Entry Points**

**Search**

Enter a Gene, Sample, Tissue, Pubmed Id or Mutation Description

Search ?

**Browse**

Browse by Gene  
Detailed Browse by Tissue  
Quick Browse by Tissue

**Additional Information**

Data in COSMIC is curated from [known Cancer Genes Literature](#) and [Systematic Screens](#).  
Interested in receiving COSMIC news and release information? Then sign up [\[here\]](#).

Please send all comments and suggestions to the COSMIC team at [cosmic@sanger.ac.uk](mailto:cosmic@sanger.ac.uk)

COSMIC data is freely downloadable in many formats on our FTP site:  
<ftp://ftp.sanger.ac.uk/pub/CGP/cosmic>

**Component Projects**

- Cancer Cell Line Project
- Genomics of Drug Sensitivity
- CGP Trace and Genotype Archive
- COSMIC Biomart
- CGP Resequencing Studies
- Cancer Gene Census

**Statistics**

Experiments	4531163
Tumours	619320
Mutations	177322
References	12026
Genes	19737
Fusions	6365
Structural Variants	2753
Whole Cancer Genomes	404

# Introduction

## Variation databases

<https://www.ebi.ac.uk/ega>

### The European Genome-phenome Archive



- EGA Home
- Submit to EGA
- Information
- Help
- Contact Us

#### EXPLORE THE EGA

- Browse studies
- Browse datasets
- Browse data access committees
- Browse data providers

#### USER LOGIN

#### The European Genome-phenome Archive (EGA)

The European Genome-phenome Archive (EGA) repository allows you to **explore datasets** from numerous genetic studies, supplied by a range of **data providers**. Access to datasets must be approved by the specified **Data Access Committee (DAC)**.

#### Studies

Studies are experimental investigations of a particular phenomenon or trait.



[Browse all studies](#)

Use the search box at the top of this page to find a study using a study accession number (EGASXXXXXXXXXX).

#### Datasets

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).



[Browse datasets that we hold](#)



[Browse the list of DACs](#)

#### Data Providers

Data Providers can be involved in creating studies, data submission and the designation of data access committees (DACs).



[Browse EGA Data Providers](#)

<http://www.ebi.ac.uk/dgva>

EMBL-EBI

Services

Research

Training

Industry

About us

### Database of Genomic Variants Archive



DGVA Home

#### Database of Genomic Variants archive

- DGVA Home
- Data Submission
- Data Download
- DGVA Quick tour
- Contact Us

The Database of Genomic Variants archive (DGVA) is a repository that provides archiving, accessioning and distribution of publicly available genomic structural variants, in all species.

In recent years there have been unprecedented advances in the technologies that characterise genomic variation, and it is well known that variation at the single nucleotide level is abundant across the genomes of all species. However, it is becoming clear that *genomic structural variation* - this is variation ranging from tens to millions of base pairs in size and includes insertions, deletions, inversions, translocations and locus copy number changes - accounts for more of the individual differences at the *base pair* level in humans and is likely to play a major role in disease. Two other areas of research that are becoming increasingly important in this field are discovering how genomic structural variation affects an individual's characteristics, and understanding the role that genomic structural variation has played in the evolution of species. The DGVA catalogues, stores and freely disseminates this important class of genomic variation in any species, providing a valuable resource to a large community of researchers.

Ignacio Medina  
[imedina@cipf.es](mailto:imedina@cipf.es)

Biological Databases



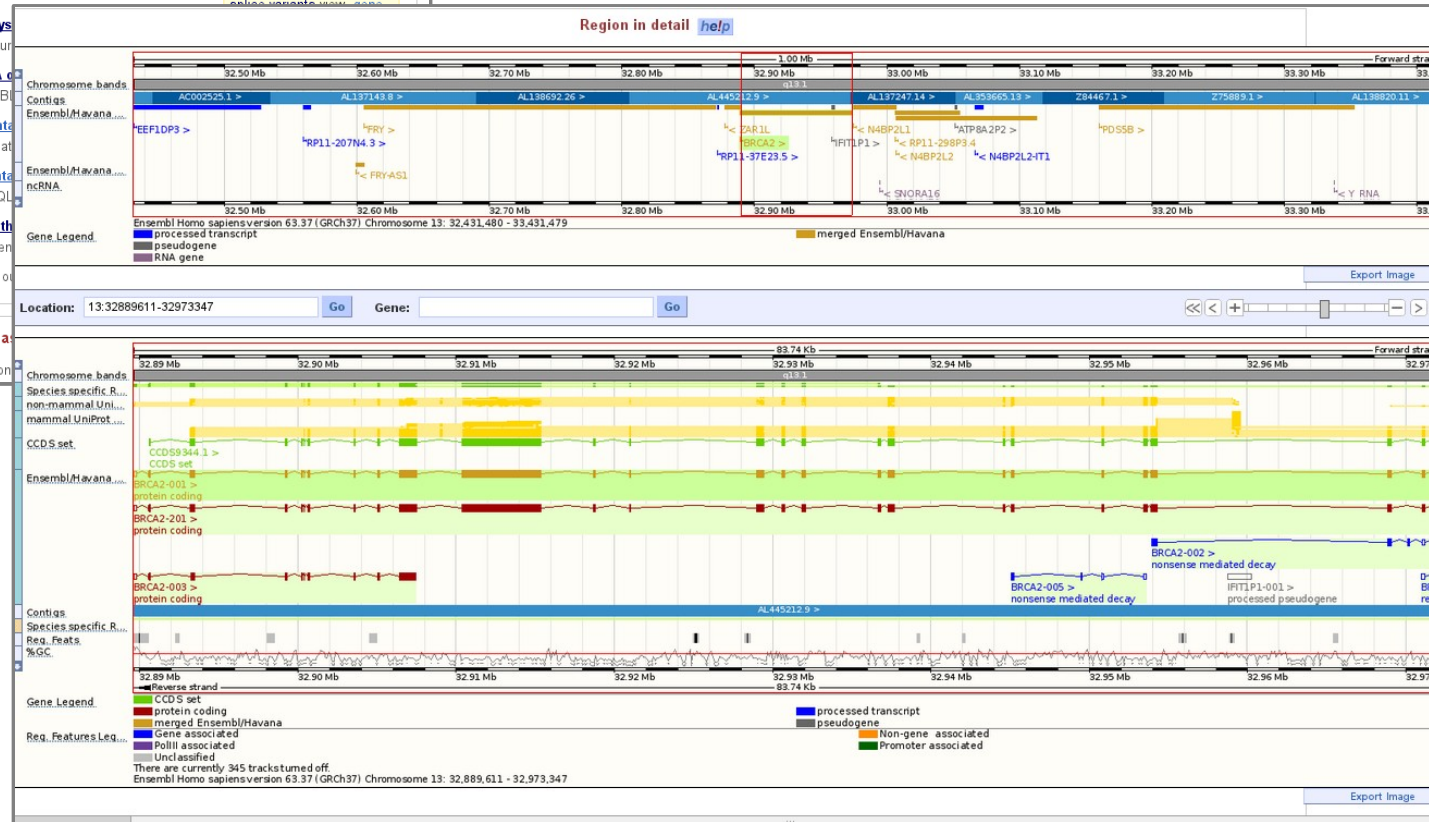
# Introduction

## Genome annotation databases and browsers

*Ensembl, the most used*

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

The screenshot shows the Ensembl website interface. At the top, there is a search bar with a dropdown menu set to 'All species' and a 'Go' button. Below the search bar, there are links for 'BLAST/BLAT', 'BioMart', 'Tools', 'Downloads', 'Help & Documentation', 'Blog', and 'Mirrors'. A navigation menu includes 'Login' and 'Register'. The main content area is divided into several sections: 'New to Ensembl?' with links for 'Learn how to use Ensembl', 'Add custom tracks', 'Upload and analysis', 'Search for a DNA sequence', 'Fetch only the data', 'Download our data', and 'Mine Ensembl with'; 'Browse a Genome' with a description of the project and links to 'Popular genomes' (Human, Mouse, Zebrafish); and 'What's New in Release' with a 'Sortable tracks' link.



Ignacio Medina  
imedina@cipf.es

Biological Databases

# Introduction

## Genome annotation databases and browsers

<http://genome.ucsc.edu/index.html>

UCSC Genome Browser

The image displays the UCSC Genome Browser interface. At the top, the header reads "UCSC Genome Bioinformatics" with a navigation menu including Genomes, Blat, Tables, Gene Sorter, PCR, VisiGene, Proteome, Session, FAQ, and Help. A sidebar on the left lists various tools like Genome Browser, ENCODE, Neandertal, Blat, Table Browser, Gene Sorter, In Silico PCR, and Genome Graphs. The main content area features an "About the UCSC Genome Bioinformatics Site" section with a welcome message and a "News" section with a Twitter icon. Below this is a detailed genomic track for Human Feb. 2009 (GRCh37/hg19) Assembly. The track shows a 29 kb region on chromosome 13 (q13.1) with various annotations including RefSeq Genes (BRCA2), Human mRNAs, Spliced ESTs, Layered HSK27Ac, DNase Clusters, Txn Factor ChIP, Mammal Cons, Multiz Alignments of 46 Vertebrates, Common SNPs (132), and RepeatMasker. The interface includes navigation controls like "move start", "move end", and "zoom in/out" buttons.

# Ensembl Biomart

## Data mining with Biomart

<http://www.ensembl.org/biomart/martview>

- Oriented to small queries
- Easy interface for users

The screenshot shows the Ensembl Biomart interface. At the top, there is a navigation bar with the Ensembl logo and links for BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. Below this is a secondary navigation bar with buttons for New, Count, Results, URL, XML, Perl, and Help. The main content area is divided into two columns. The left column contains a 'Dataset' section for 'Homo sapiens genes (GRCh37.p10)', a 'Filters' section with '[None selected]', and an 'Attributes' section with 'Ensembl Gene ID' and 'Ensembl Transcript ID'. The right column contains a 'Please select columns to be included in the output and hit 'Results' when ready' instruction. Below this are several sections of checkboxes: 'Features' (selected), 'Structures', 'Transcript Event', 'Homologs', 'Variation', and 'Sequences'. Under 'GENE:', there are two columns of checkboxes. The first column includes 'Ensembl Gene ID' (checked), 'Ensembl Transcript ID' (checked), 'Ensembl Protein ID', 'Ensembl Exon ID', 'Description', 'Chromosome Name', 'Gene Start (bp)', 'Gene End (bp)', 'Strand', 'Band', 'Transcript Start (bp)', and 'Transcript End (bp)'. The second column includes 'Associated Gene Name', 'Associated Transcript Name', 'Associated Gene DB', 'Associated Transcript DB', 'Transcript count', '% GC content', 'Gene Biotype', 'Transcript Biotype', 'Source', 'Status (gene)', and 'Status (transcript)'. At the bottom, there is an 'EXTERNAL:' section.



# CellBase

## Motivation

---

### Motivation

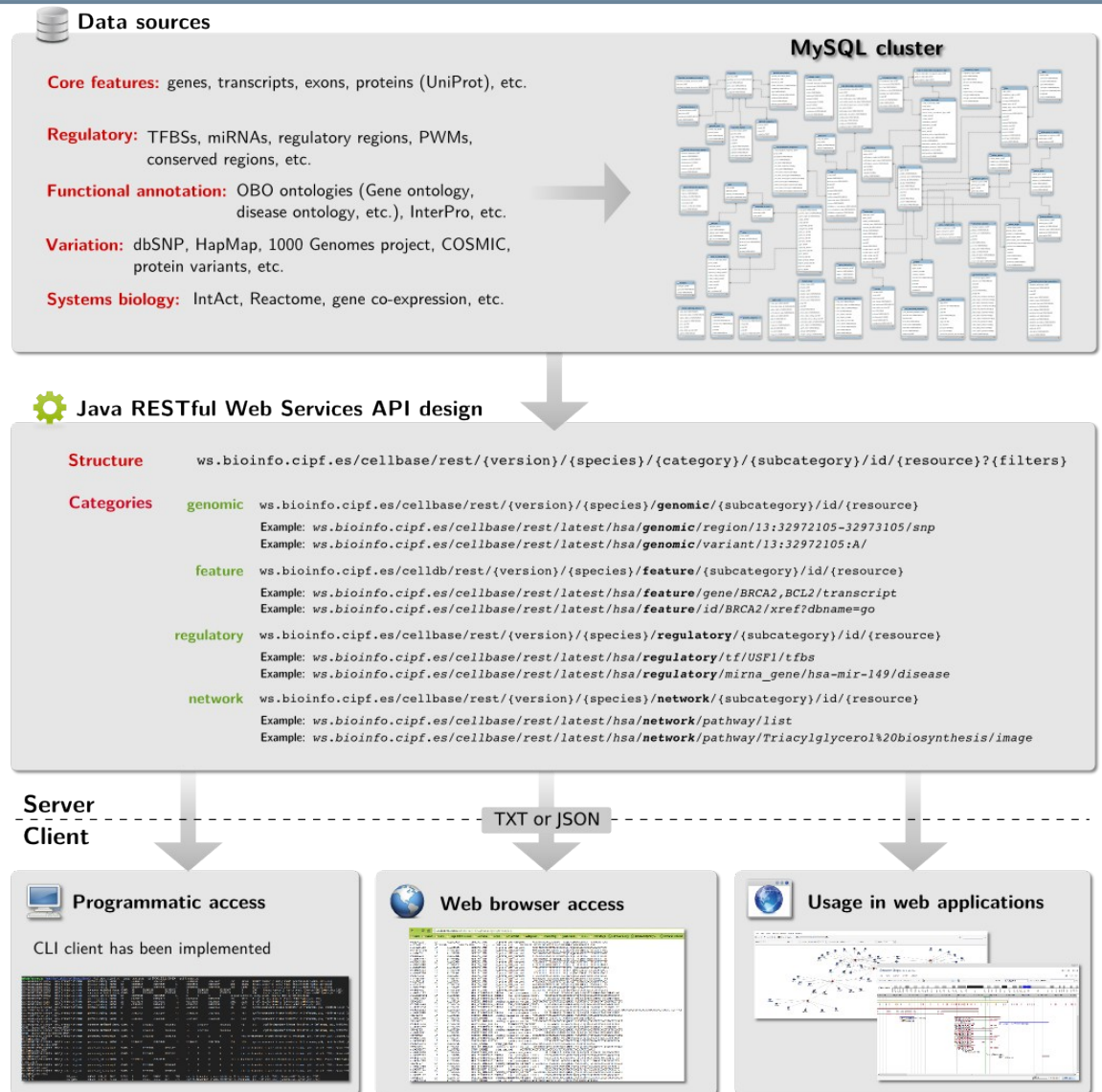
- **Exponential growth** in the number and size of biological databases and repositories. Data size can reach **hundreds of gigabytes** and involves serious problems of data access through Internet and local disks.
- **Biological information is spread out** in different databases and repositories (~1512), using different identifiers → Problem when analyzing genome-wide experiments.
- Researchers have to deal with **complex scripts** (in Perl, generally) or parse horribles XML files.

### Goals

- **Integrate and Join** the most relevant and high quality biological information in a single database.
- Facilitate **accessibility** to users.

# CellBase Overview

- A comprehensive integrative **database** and **RESTful Web Services API**.
- More than **220GB** of data and ~100 SQL tables containing the most relevant **biological information**.
- Available for **11 species**: human, mouse, rat, zebrafish, fruitfly, worm, yeast, dog, pig, mosquito and plasmodium.
- Accessible via RESTful web services and by a Perl client



# CellBase

## RESTful web services

### General Structure

`ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/{category}/{subcategory}/id/{resource}?{filters}`

<http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/mutation>

### Categories

- Genomic

Subcategories: *region*, *variant* and *position*

<http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/1:3972105-12973105/gene>

- Feature

Subcategories: *gene*, *transcript*, *exon*, *protein*, *snp* and *karyotype*

<http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs3934834/phenotype>

- Regulatory

Subcategories: *mirna\_gene*, *mirna\_mature* and *tf*

[http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna\\_gene/hsa-mir-95/disease](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna_gene/hsa-mir-95/disease)

- Network

Subcategories: *pathway*

<http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/Triacylglycerol%20biosynthesis/info>

# CellBase

## Documentation

---

Documentation site

<http://www.opencb.org/projects/cloud/doku.php?id=cellbase:overview>

Article

*Published online 12 June 2012*

*Nucleic Acids Research, 2012, Vol. 40, Web Server issue W609–W614  
doi:10.1093/nar/gks575*

### **CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources**

Marta Bleda<sup>1,2</sup>, Joaquin Tarraga<sup>1,3</sup>, Alejandro de Maria<sup>1</sup>, Francisco Salavert<sup>1,2</sup>, Luz Garcia-Alonso<sup>1</sup>, Matilde Celma<sup>4</sup>, Ainocha Martin<sup>4</sup>, Joaquin Dopazo<sup>1,2,3,\*</sup> and Ignacio Medina<sup>1,3,\*</sup>

# Accessing CellBase information

Category	Subcategory	ID	Resources
genomic	position	chr:position	gene, snp
	variant	chr:position:reference allele:other allele chr:position:other allele	snp_phenotype, mutation_phenotype, consequence_type
	region	chr:start-end	Gene, transcript, exon, snp, mutation, structural_variation, sequence, tfbs, mirna_target, cpg_island, conserved_region, regulatory
feature	gene	Any gene ID	list, info, transcript, snp, mutation, tfbs, mirna_target, protein_feature
	transcript	Ensembl transcript	info, all, gene, exon, sequence, mutation, protein_feature
	snp	rsID	Info, consequence_type, regulatory, phenotype, population_frequency, xref
	exon	Ensembl exon	Info, info, sequence, transcript
	protein	Uniprot ID and Accession	Info, feature, xref, interaction
	id	any	xref
	regulatory	tf	Gene name
mirna_gene		miRBase gene ID	Info, target, disease
mirna_mature		miRBase gene Accession	Info, gene, mirna_gene, target_gene, target, disease, annotation

# Accessing CellBase information

---

- Web services

[ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/transcript](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/transcript)

- Web client

<http://bioinfo.cipf.es/apps-beta/cellbase.html>

- Perl RESTful WS client

<http://www.opencb.org/projects/cloud/doku.php?id=cellbase:downloads>

- Download the Perl client
- Execute it!
- `./cellbase_client.pl --input-type gene --id BRCA2 --get transcript`

# Accessing CellBase information

---

## QUESTION 1:

We are interested in a particular region of the genome, 1:2201105-2319315, and we want to know if this **region contains mutations already catalogued**.

## HELP:

- Version: latest
- Species: hsa
- Category: genomic
- Subcategory: region
- ID: 1:2201105-2319315
- Resource: mutation

# Accessing CellBase information

---

## QUESTION 2:

We want to know which **microRNAs** regulate the **gene BRCA2**.

## HELP:

- Version: latest
- Species: hsa
- Category: feature
- Subcategory: gene
- ID: BRCA2
- Resource: mirna\_target



# Accessing CellBase information

---

## QUESTION 3:

We want to know the allelic and **genotypic frequencies** for a SNP, rs158691, across populations.

## HELP:

- Version: latest
- Species: hsa
- Category: feature
- Subcategory: snp
- ID: rs158691
- Resource: population\_frequency

# Accessing CellBase information

## QUESTION 4:

We need to **convert** some gene names (BRCA2, PAEP, GATA2) into Ensembl Gene **identifiers**.

## HELP:

- Version: latest
- Species: hsa
- Category: feature
- Subcategory: id
- ID: BRCA2, PAEP, GATA2
- Resource: xref
- Filter: dbname=ensembl\_gene

# Accessing CellBase information

---

## QUESTION 5:

We have obtained a **microRNA** of interest (hsa-miR-149-3p) in our analysis and we want to know if it has been **related with any disease**.

## HELP:

- Version: latest
- Species: hsa
- Category: regulatory
- Subcategory: mirna\_mature
- ID: hsa-miR-149-3p
- Resource: disease

# Accessing Ensembl BioMart information

<http://www.ensembl.org/biomart/martview>

The screenshot displays the Ensembl BioMart interface. At the top, the Ensembl logo is on the left, and navigation links for BLAST/BLAT, BioMart, Tools, Downloads, and More are in the center. On the right, there are links for Login and Register, and a search bar with the text "Search all species...". Below the navigation bar, there are buttons for New, Count, and Results, and a row of utility buttons for URL, XML, Perl, and Help. The main content area is divided into two columns. The left column contains a "Dataset" section for "Homo sapiens genes (GRCh37.p10)", a "Filters" section with "[None selected]", and an "Attributes" section listing "Ensembl Gene ID" and "Ensembl Transcript ID". Below this is another "Dataset" section with "[None Selected]". The right column is titled "Please restrict your query using criteria below" and contains a list of filter categories, each with a checkbox and a text input field: REGION, GENE, TRANSCRIPT EVENT, GENE ONTOLOGY, EXPRESSION, MULTI SPECIES COMPARISONS, PROTEIN DOMAINS, and VARIATION.

# Accessing Ensembl BioMart information

---

## QUESTION 1:

We need the complete list of **correspondences** between **genes** and their **transcripts**.

## HELP:

- Database: Ensembl Genes 70
- Dataset: Homo sapiens genes
- No filters. We want all the information!
- Attributes: Gene > Ensembl Gene ID  
Ensembl Transcript ID

# Accessing Ensembl BioMart information

## QUESTION 2:

We want to retrieve all known **somatic variants** that have been associated with a **cancer** and we just want those in **COSMIC**.

## HELP:

- Database: Ensembl Variation 70
- Dataset: Homo sapiens Somatic Short Variation (SNPs and indels)  
Filters: GENERAL VARIATION FILTERS > COSMIC
- Attributes: Sequence variation > Variation Name, Variation source, Chromosome name, Position on Chromosome (bp), Phenotype description

# Solutions to CellBase questions

## Question 1:

WS: <http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/1:2201105-2319315/mutation>

CLI: `./cellbase_client.pl --input-type region --id 1:2201105-2319315 --get mutation`

## Question 2:

WS: [http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/mirna\\_target](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/mirna_target)

CLI: `./cellbase_client.pl --input-type gene --id BRCA2 --get mirna_target`

## Question 3:

WS: [http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs158691/population\\_frequency](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs158691/population_frequency)

CLI: `./cellbase_client.pl --input-type snp --id rs158691 --get population_frequency`

## Question 4:

WS: [http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/id/BRCA2,PAEP,GATA2/xref?dbname=ensembl\\_gene](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/id/BRCA2,PAEP,GATA2/xref?dbname=ensembl_gene)

CLI: `./cellbase_client.pl --input-type id --id BRCA2,PAEP,GATA2 --get xref?dbname=ensembl_gene`

## Question 5:

WS: [http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna\\_mature/hsa-miR-149-3p/disease](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna_mature/hsa-miR-149-3p/disease)

CLI: `./cellbase_client.pl --input-type regulatory --id hsa-miR-149-3p --get disease`