Introduction to

Biological Databases

Valencia, September 2011

Marta Bleda
Ignacio Medina (*Nacho*)
imedina@cipf.es
http://bioinfo.cipf.es/imedina
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
Valencia, Spain

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

Index

- Introduction
- Sequence Databases
- Functional Annotation Databases
- Protein Databases
- Variation Databases
- Genome Databases and Browsers
- Exercises

Introduction

- Last years has been an exponential increase in the number of biological databases and in their content.
- ***Nucleic Acids Research* online Molecular Biology Database Collection** is a public repository that lists principal ***biological databases***
- Updated every year. The Nov-2010 update includes **1330** databases !!

http://www3.oup.co.uk/nar/database/c/

Introduction

- These databases contain:
  - Data and results from experiments with microarrays, *NGS*, ...
  - Genes, transcripts and *EST* sequences
  - DNA variation and frequencies (*SNP*, mutations, indels, …)
  - Protein sequences, structures and variations
  - Functional information about what a gene/protein is doing in the cell
  - User interface to search, navigate and explore the genomes

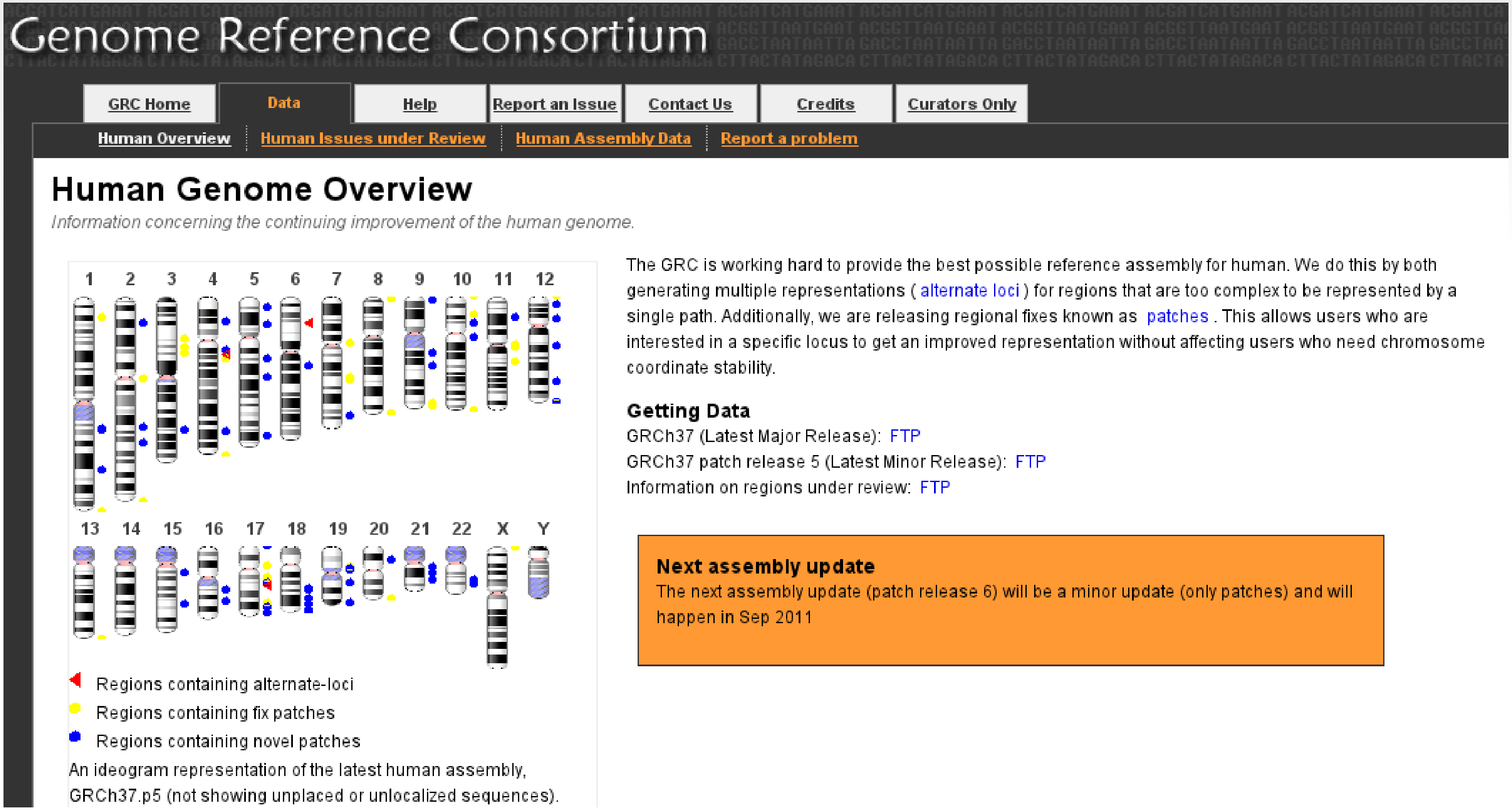
Sequence Databases
Genome Reference Consortium (*GRC*)

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

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



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

Sequence Databases

European Bioinformatics Institute (*EBI*)

Mission:

- To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress

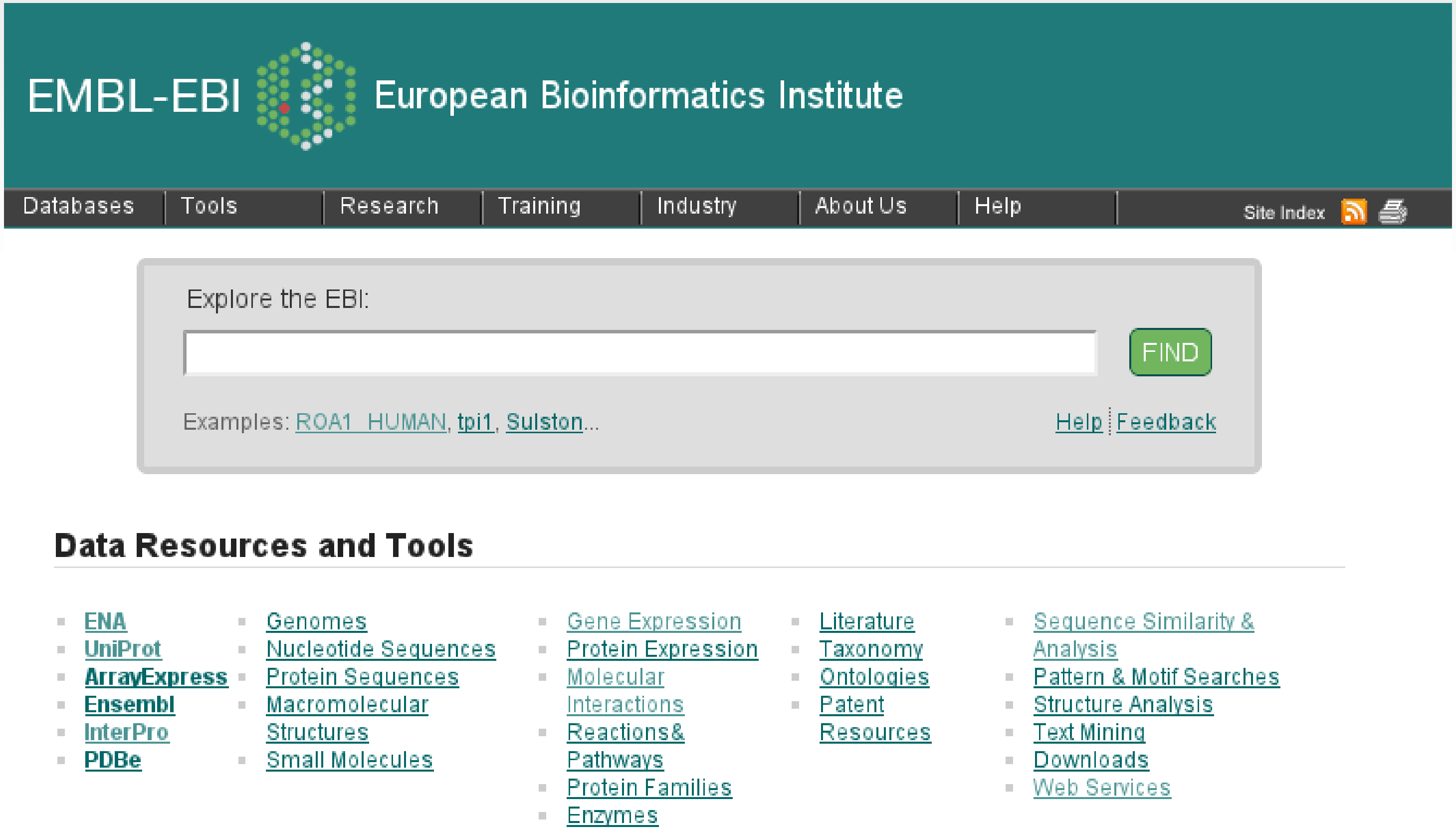
- To contribute to the advancement of biology through basic investigator-driven research in bioinformatics

- To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators

- To help disseminate cutting-edge technologies to industry

Financiado por el EMBL, por tanto con dinero Europeo

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



Explore the EBI:

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

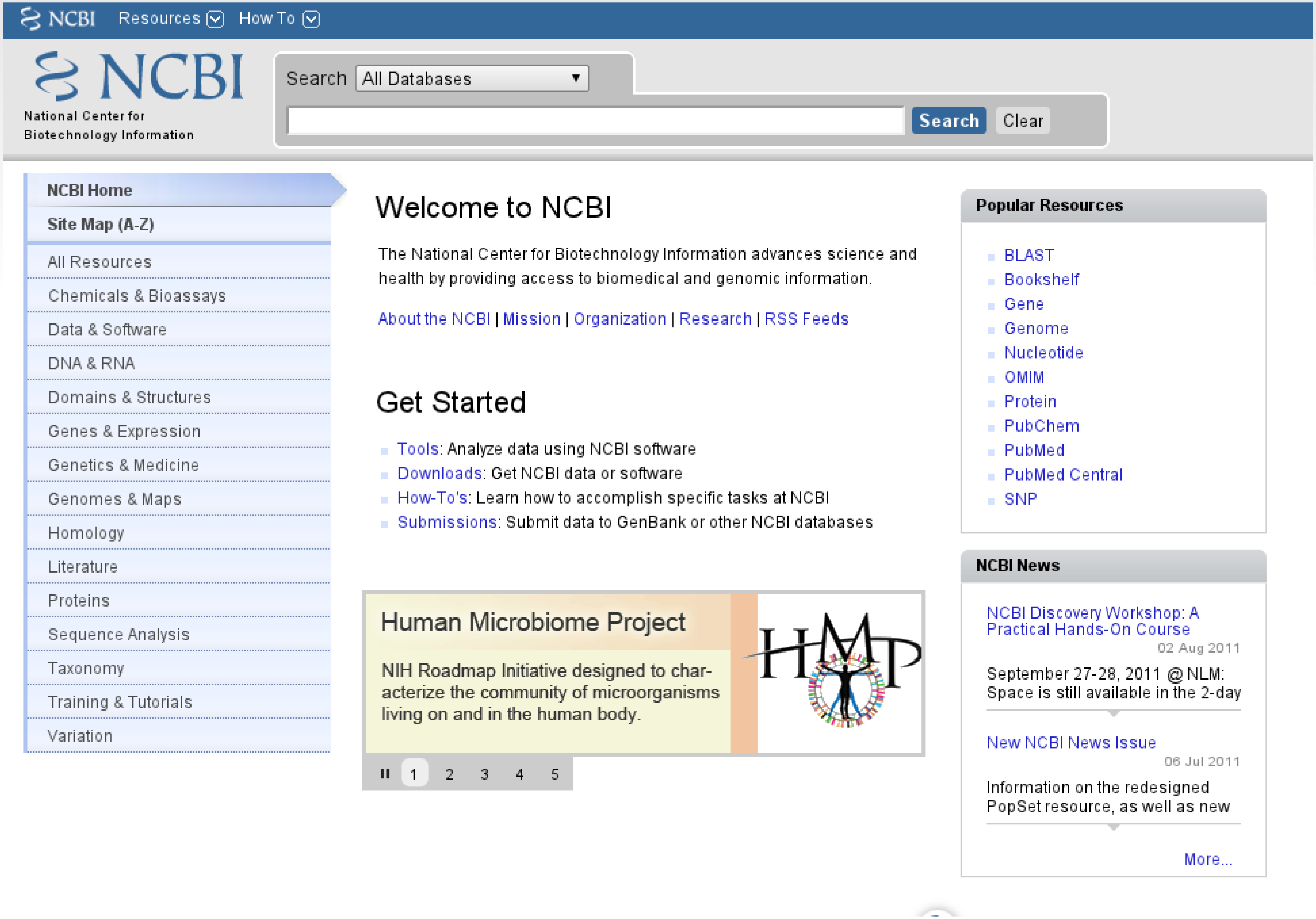
Sequence Databases

Nat. Center for Biotech. Information (*NCBI*)

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

Conjunto de herramientas y bases de datos para el estudio y análisis genómico y biomédico

Financiado por USA, en cierta forma compite con el EBI en objetivos y recursos



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.

NCBI News

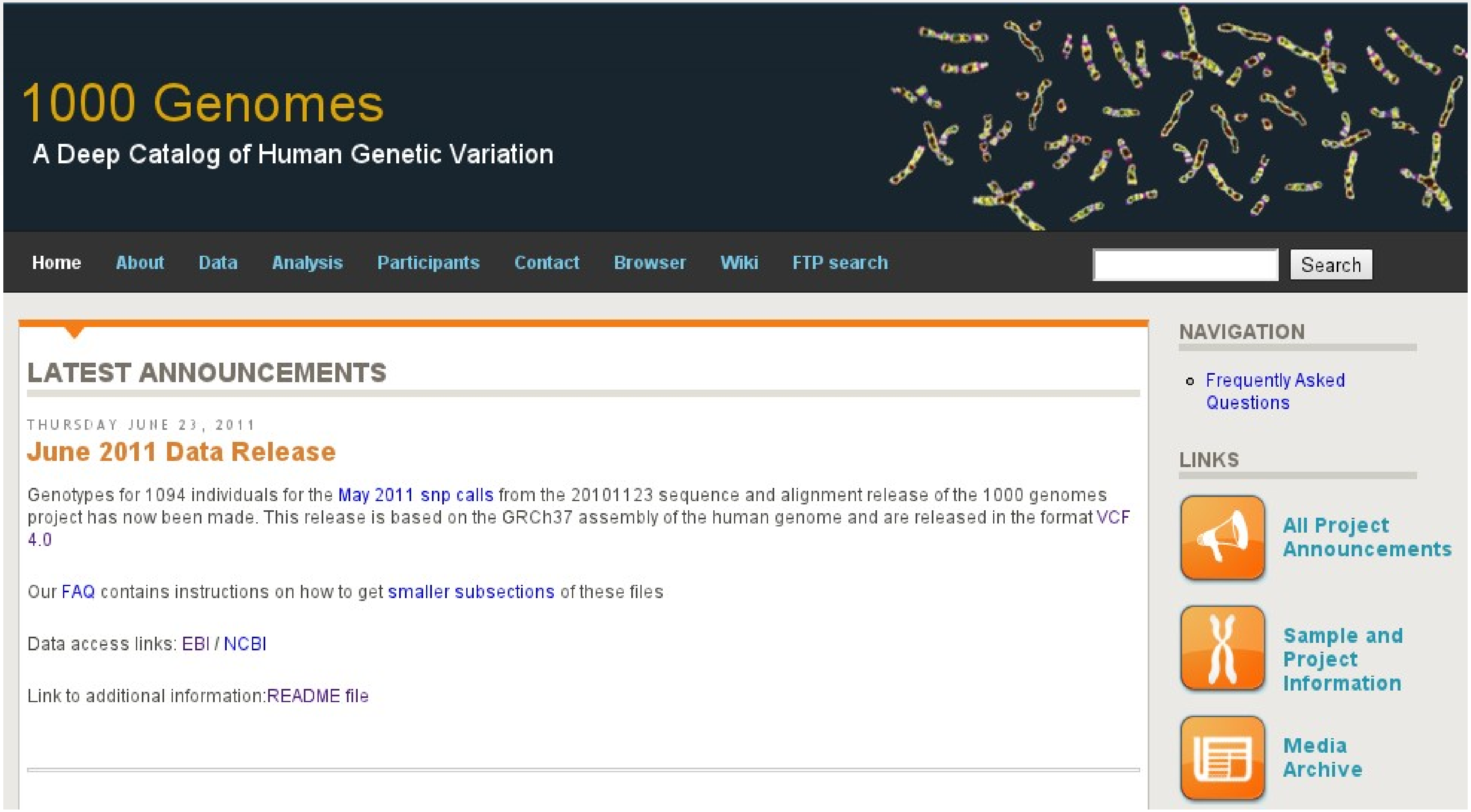
NCBI Discovery Workshop: A Practical Hands-On Course
02 Aug 2011
September 27-28, 2011 @ NLM: Space is still available in the 2-day

New NCBI News Issue
06 Jul 2011
Information on the redesigned PopSet resource, as well as new

Sequence Databases
1000 Genomes project

1000 Genomes Project is the first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation

http://www.1000genomes.org/


1000 Genomes
A Deep Catalog of Human Genetic Variation
Home
About
Data
Analysis
Participants
Contact
Browser
Wiki
FTP search
Search
NAVIGATION
Frequently Asked Questions
LINKS
All Project Announcements
Sample and Project Information
Media Archive

LATEST ANNOUNCEMENTS

THURSDAY JUNE 23, 2011

June 2011 Data Release

Genotypes for 1094 individuals for the May 2011 snp calls from the 20101123 sequence and alignment release of the 1000 genomes project has now been made. This release is based on the GRCh37 assembly of the human genome and are released in the format VCF 4.0

Our FAQ contains instructions on how to get smaller subsections of these files

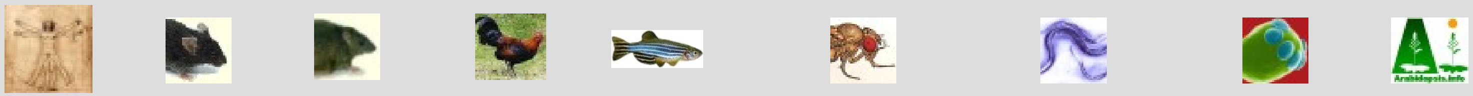
Data access links: EBI / NCBI

Link to additional information:README file

Functional Annotation DDBB

Overview

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



Homo sapiens *Mus musculus* *Rattus norvegicus* *Gallus gallus* *Danio rerio* *Drosophila melanogaster* *Caenorhabditis elegans* *Saccharmoyces cerevisae* *Arabidopsis thaliana*

UniProt/Swiss-Prot

UniProtKB/TrEMBL

Ensembl IDs

EntrezGene

Affymetrix

Agilent

Gene IDs

HGNC symbol

EMBL acc

RefSeq

PDB

Protein Id

IPI….

Functional databases

KEGG pathways

Reactome

Gene Ontology

Biological Process
Molecular Function
Cellular Component

Regulatory elements

miRNA

CisRed

Transcription Factor Binding Sites

Biocarta pathways

Keywords Swissprot

InterPro Motifs

Gene Expression in tissues

Bioentities from literature:

Diseases terms
Chemical terms

Functional Annotation DDBB

Gene Ontology (GO terms)

- The *Gene Ontology* project provides a ***controlled vocabulary*** to describe gene and gene product attributes in any organism
- Lastest version has **33808** terms (March, *2011*)
- The controlled vocabularies of terms are structured

http://www.geneontology.org/

Functional Annotation DDBB

Gene Ontology (GO terms)

The three categories of GO

Molecular Function

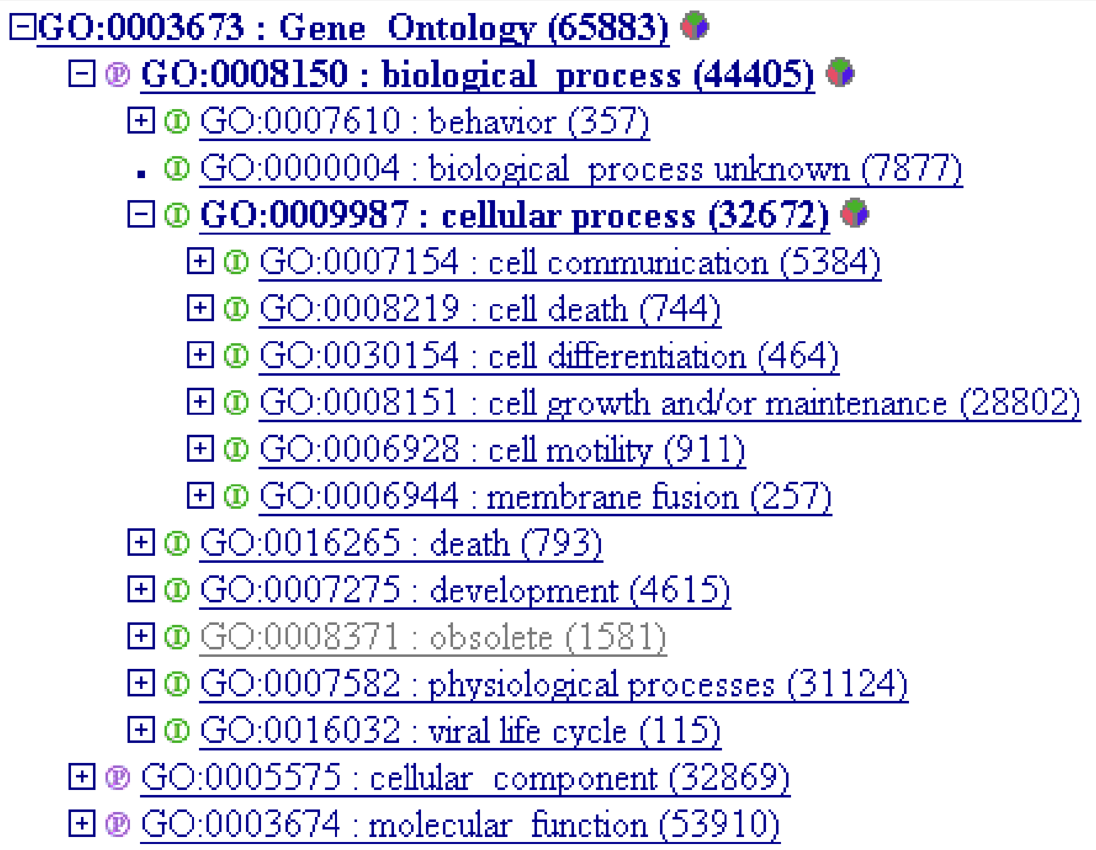
the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*

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*

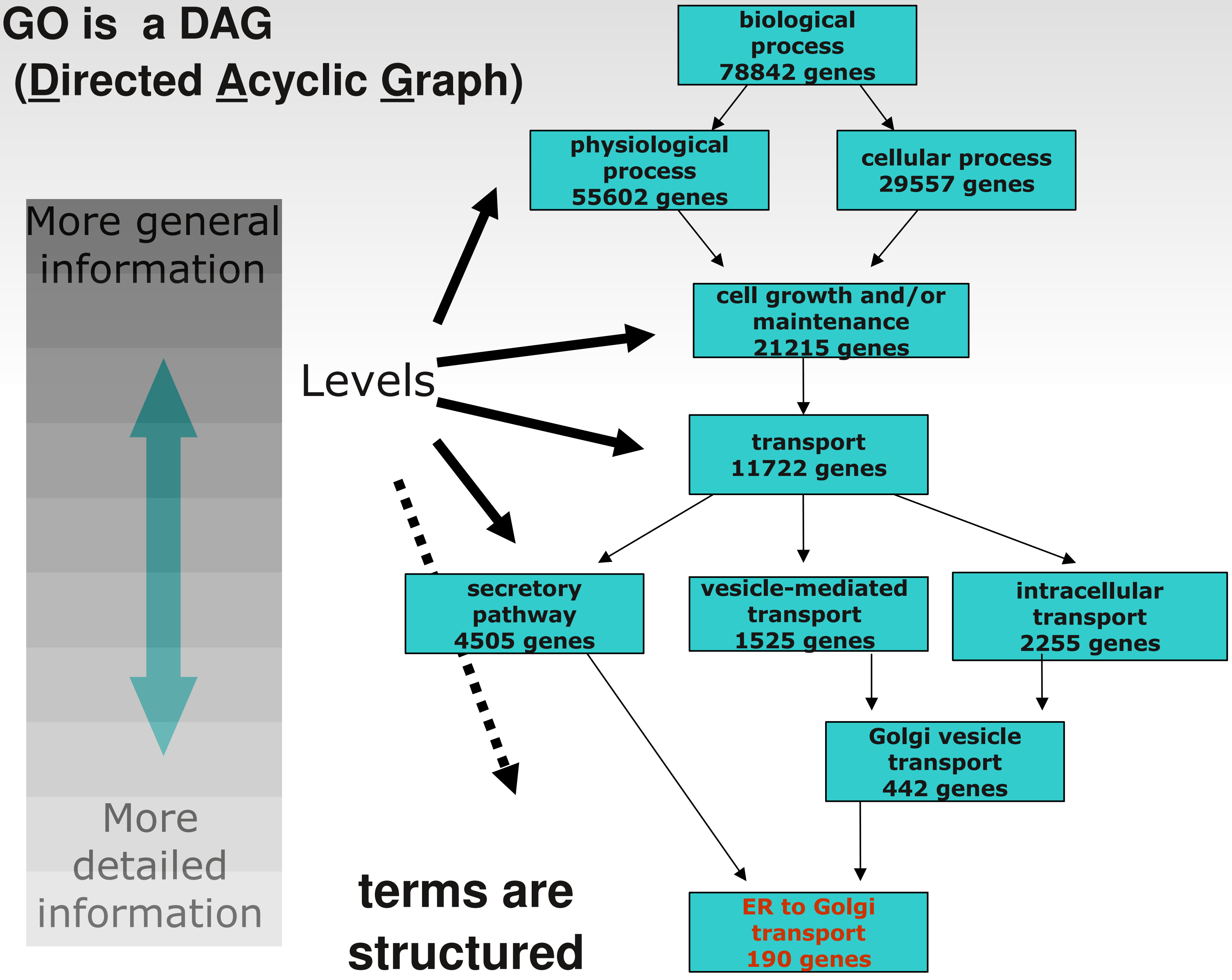


- ⊟ **GO:0003673 : Gene Ontology (65883)**
  - ⊟ Ⓟ **GO:0008150 : biological process (44405)**
    - ⊞ Ⓘ GO:0007610 : behavior (357)
    - ▪ Ⓘ GO:0000004 : biological process unknown (7877)
    - ⊟ Ⓘ **GO:0009987 : cellular process (32672)**
      - ⊞ Ⓘ GO:0007154 : cell communication (5384)
      - ⊞ Ⓘ 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)

Functional Annotation DDBB

Gene Ontology (GO terms)

GO is a DAG
(Directed Acyclic Graph)


More general information
More detailed information
Levels
biological process 78842 genes
physiological process 55602 genes
cellular process 29557 genes
cell growth and/or maintenance 21215 genes
transport 11722 genes
secretory pathway 4505 genes
vesicle-mediated transport 1525 genes
intracellular transport 2255 genes
Golgi vesicle transport 442 genes
ER to Golgi transport 190 genes

terms are structured

Annotations are given to the most specific (low) level.

True path rule:
Annotation at a term implies annotation to all its parent terms

Annotation is given with an Evidence Code:
EXP: inferred from Experiment
IDA: inferred by direct assay
TAS: traceable author statement
ISS: inferred by sequence similarity
IEA: electronic annotation

Functional Annotation DDBB

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

Functional Annotation DDBB

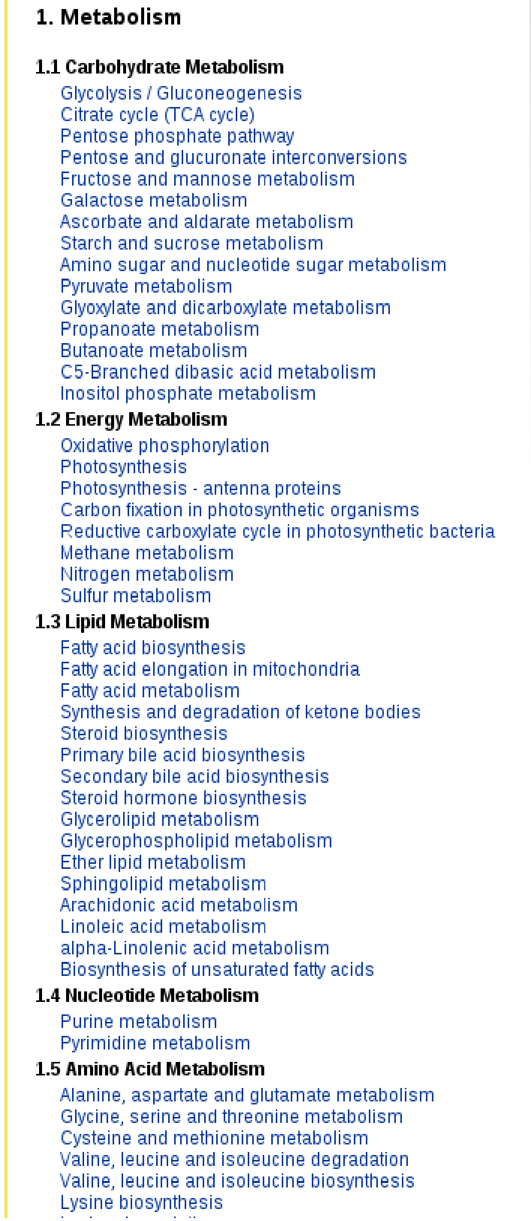
GO Slim

- ***GO slims*** are cut-down versions of the GO ontologies *containing a* ***subset*** of the terms in the whole GO. They give a broad overview of the ontology content without the detail of the specific fine grained terms

http://www.geneontology.org/GO.slims.shtml

Functional Annotation DDBB

Kyoto Encyclopedia of Genes and Genomes (KEGG)



1. Metabolism

1.1 Carbohydrate Metabolism
- Glycolysis / Gluconeogenesis
- Citrate cycle (TCA cycle)
- Pentose phosphate pathway
- Pentose and glucuronate interconversions
- Fructose and mannose metabolism
- Galactose metabolism
- Ascorbate and aldarate metabolism
- Starch and sucrose metabolism
- Amino sugar and nucleotide sugar metabolism
- Pyruvate metabolism
- Glyoxylate and dicarboxylate metabolism
- Propanoate metabolism
- Butanoate metabolism
- C5-Branched dibasic acid metabolism
- Inositol phosphate metabolism

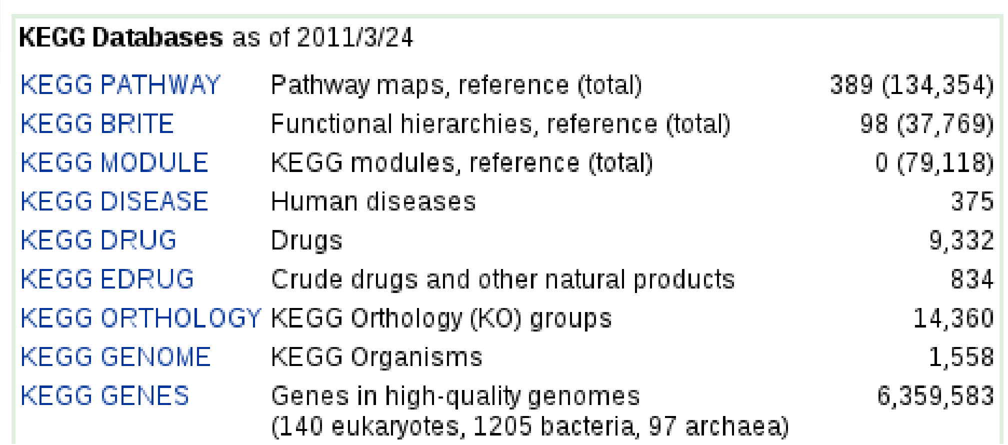
1.2 Energy Metabolism
- Oxidative phosphorylation
- Photosynthesis
- Photosynthesis - antenna proteins
- Carbon fixation in photosynthetic organisms
- Reductive carboxylate cycle in photosynthetic bacteria
- Methane metabolism
- Nitrogen metabolism
- Sulfur metabolism

1.3 Lipid Metabolism
- Fatty acid biosynthesis
- Fatty acid elongation in mitochondria
- Fatty acid metabolism
- Synthesis and degradation of ketone bodies
- Steroid biosynthesis
- Primary bile acid biosynthesis
- Secondary bile acid biosynthesis
- Steroid hormone biosynthesis
- Glycerolipid metabolism
- Glycerophospholipid metabolism
- Ether lipid metabolism
- Sphingolipid metabolism
- Arachidonic acid metabolism
- Linoleic acid metabolism
- alpha-Linolenic acid metabolism
- Biosynthesis of unsaturated fatty acids

1.4 Nucleotide Metabolism
- Purine metabolism
- Pyrimidine metabolism

1.5 Amino Acid Metabolism
- Alanine, aspartate and glutamate metabolism
- Glycine, serine and threonine metabolism
- Cysteine and methionine metabolism
- Valine, leucine and isoleucine degradation
- Valine, leucine and isoleucine biosynthesis
- Lysine biosynthesis

KEGG pathways



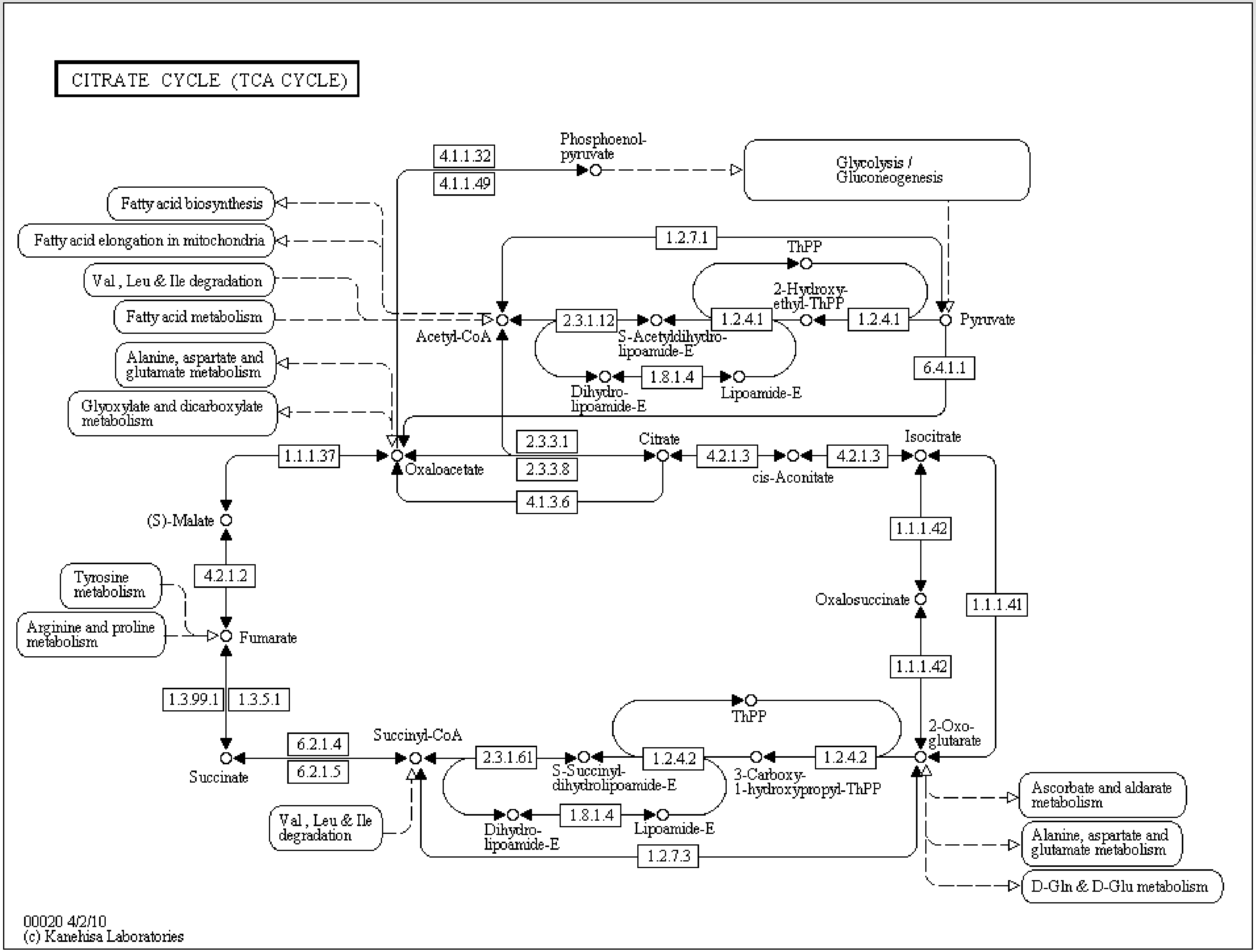
KEGG Databases as of 2011/3/24

KEGG PATHWAY	Pathway maps, reference (total)	389 (134,354)
KEGG BRITE	Functional hierarchies, reference (total)	98 (37,769)
KEGG MODULE	KEGG modules, reference (total)	0 (79,118)
KEGG DISEASE	Human diseases	375
KEGG DRUG	Drugs	9,332
KEGG EDRUG	Crude drugs and other natural products	834
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	14,360
KEGG GENOME	KEGG Organisms	1,558
KEGG GENES	Genes in high-quality genomes (140 eukaryotes, 1205 bacteria, 97 archaea)	6,359,583

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

Functional Annotation DDBB

KEGG



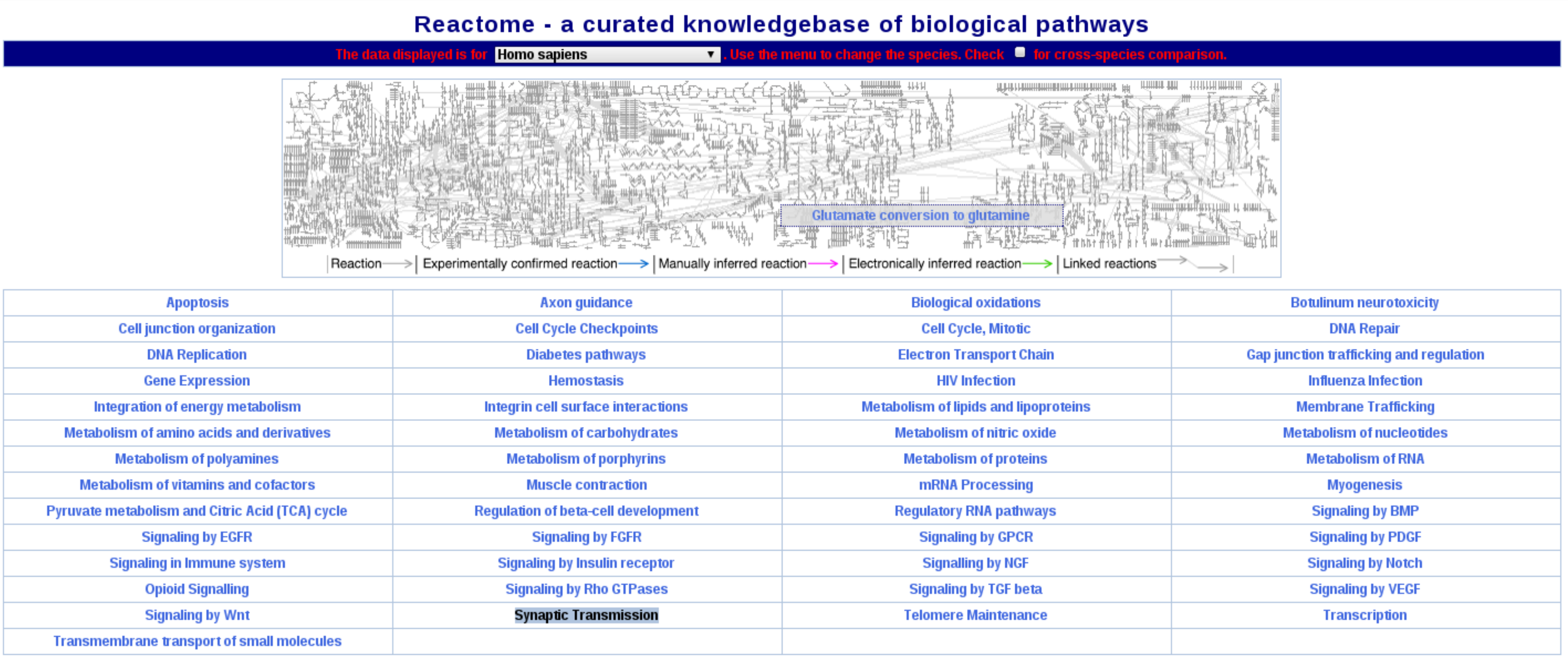
00020 4/2/10
(c) Kanehisa Laboratories

Functional Annotation DDBB

Reactome

- 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

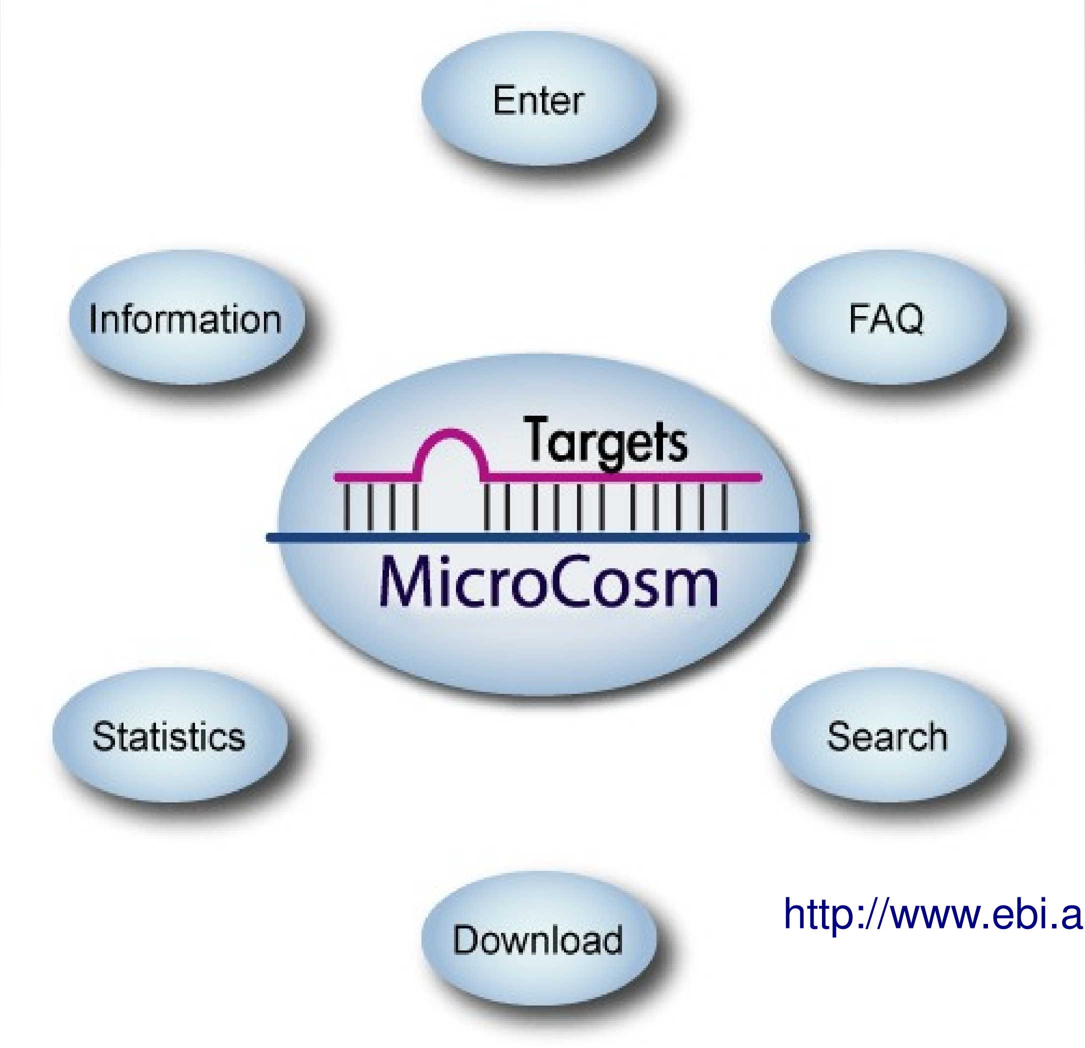
The data displayed is for Homo sapiens ▾ . Use the menu to change the species. Check □ for cross-species comparison.

Glutamate conversion to glutamine
Reaction
Experimentally confirmed reaction
Manually inferred reaction
Electronically inferred reaction
Linked reactions

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	Signalling by NGF	Signaling by Notch
Opioid Signalling	Signaling by Rho GTPases	Signaling by TGF beta	Signaling by VEGF
Signaling by Wnt	Synaptic Transmission	Telomere Maintenance	Transcription
Transmembrane transport of small molecules			

Functional Annotation DDBB

MicroRNA



- Involved in gene regulation
- Last versions has 15172 entries (Release 16, Sept 2010)
- 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/

Functional Annotation DDBB

Jaspar TFBS

- 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 acess, non-redundancy and quality

http://jaspar.genereg.net/

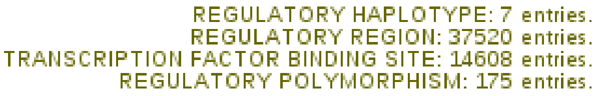
Functional Annotation DDBB

ORegAnno

- It's an open database for the curation of known regulatory elements from scientific literature (*TFBS*)
- Annotation is collected from users worldwide for various biological assays

ORegAnno

open regulatory annotation database


REGULATORY HAPLOTYPE: 7 entries.
REGULATORY REGION: 37520 entries.
TRANSCRIPTION FACTOR BINDING SITE: 14608 entries.
REGULATORY POLYMORPHISM: 175 entries.

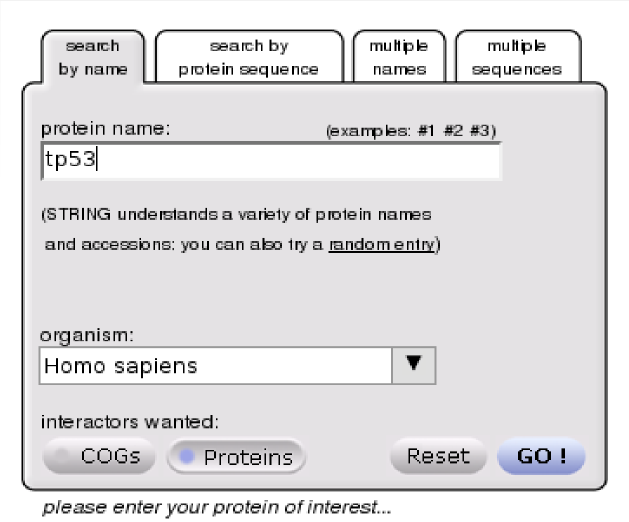
http://www.oreganno.org/oregano/

Functional Annotation DDBB

String

Database for known and predicted protein-protein interactions (direct and indirect associations)
Cover four sources of annotations: Genomic association (prokaryotes), high-throughput experiments (e.g. y2h), conserved co-expression, previous knowledge (text-mining).

A combined score is calculated for every association based on benchmarks of the different types of associations against a common reference set.



protein name: (examples: #1 #2 #3)

tp53

(STRING understands a variety of protein names and accessions; you can also try a random entry)

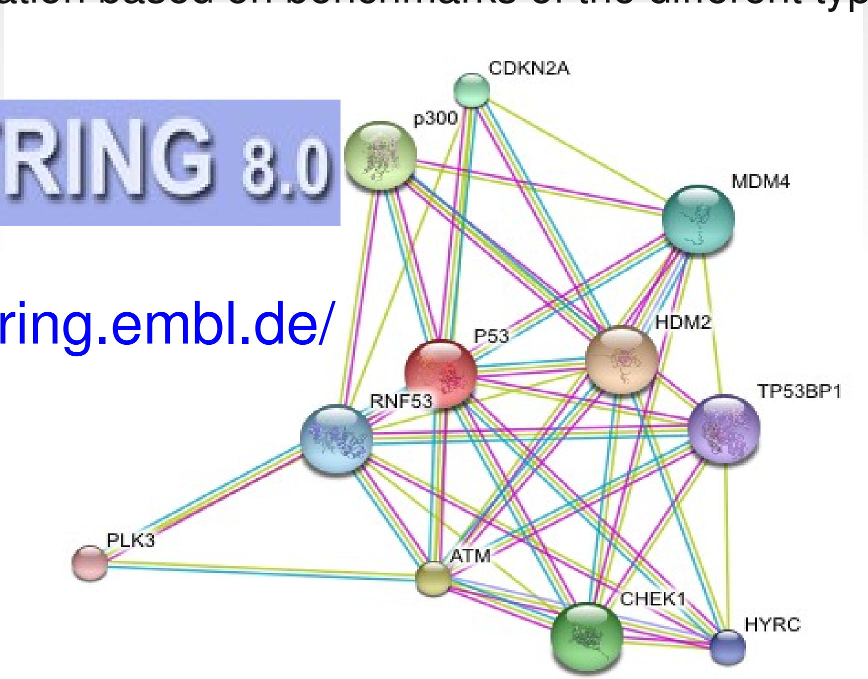
organism: Homo sapiens ▼

interactors wanted: ○ COGs ● Proteins

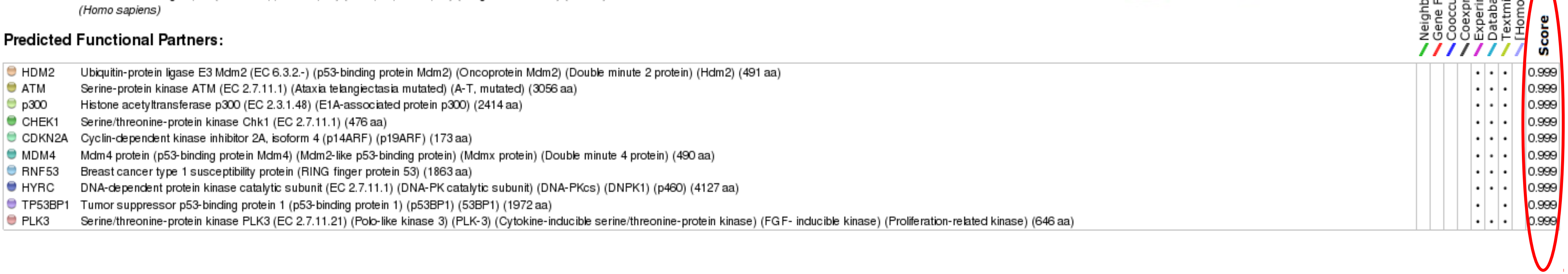
please enter your protein of interest...

STRING 8.0

http://string.embl.de/



Combined score (all sources)



Your Input:

● P53 Cellular tumor antigen p53 (Tumor suppressor p53) (Phosphoprotein p53) (Antigen NY-CO-13) (393 aa) *(Homo sapiens)*

Predicted Functional Partners:

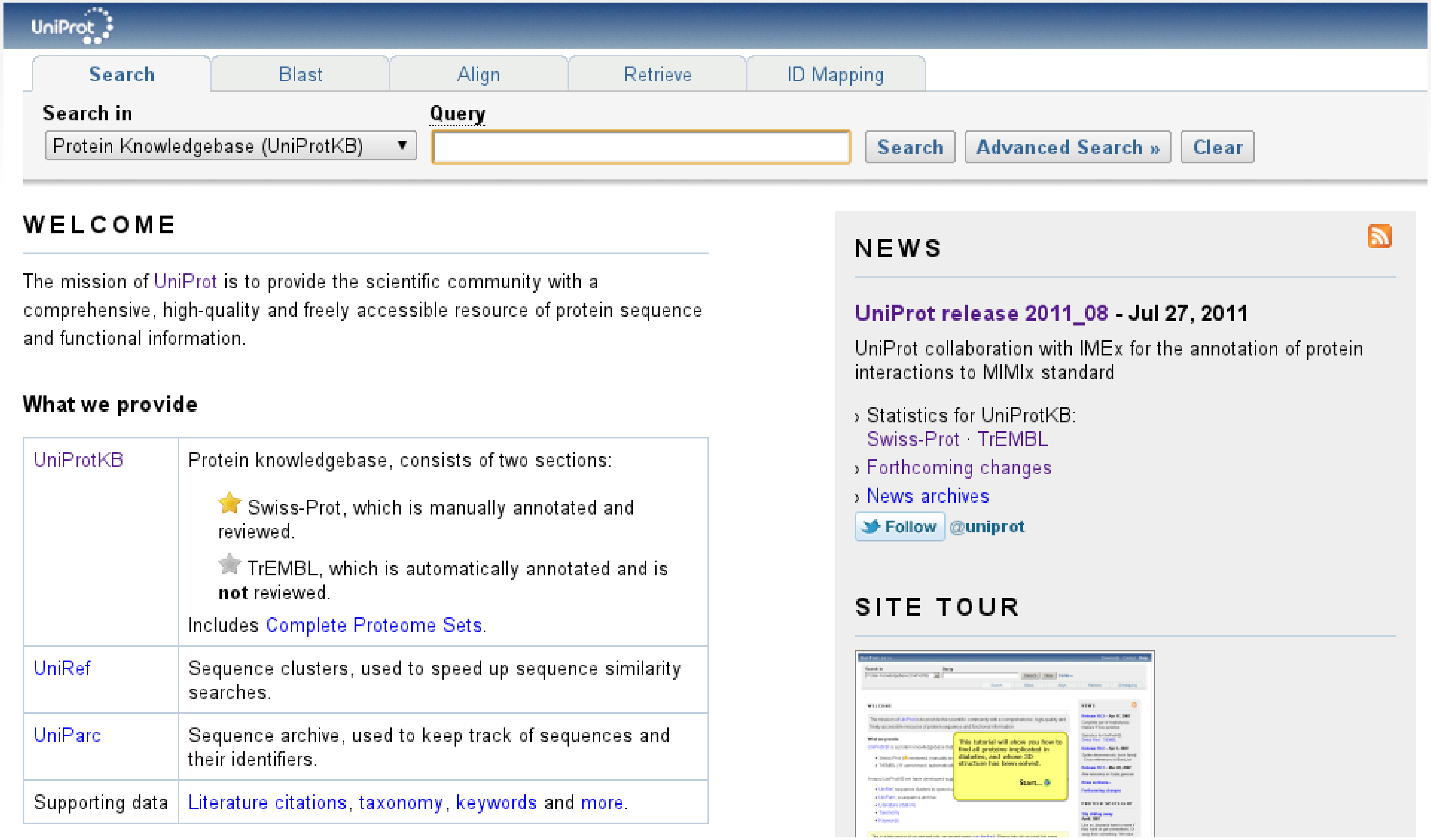
		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	**Score**
● HDM2	Ubiquitin-protein ligase E3 Mdm2 (EC 6.3.2.-) (p53-binding protein Mdm2) (Oncoprotein Mdm2) (Double minute 2 protein) (Hdm2) (491 aa)					•	•	•		0.999
● ATM	Serine-protein kinase ATM (EC 2.7.11.1) (Ataxia telangiectasia mutated) (A-T, mutated) (3056 aa)					•	•	•		0.999
● p300	Histone acetyltransferase p300 (EC 2.3.1.48) (E1A-associated protein p300) (2414 aa)					•	•	•		0.999
● CHEK1	Serine/threonine-protein kinase Chk1 (EC 2.7.11.1) (476 aa)					•	•	•		0.999
● CDKN2A	Cyclin-dependent kinase inhibitor 2A, isoform 4 (p14ARF) (p19ARF) (173 aa)					•	•	•		0.999
● MDM4	Mdm4 protein (p53-binding protein Mdm4) (Mdm2-like p53-binding protein) (Mdmx protein) (Double minute 4 protein) (490 aa)					•	•	•		0.999
● RNF53	Breast cancer type 1 susceptibility protein (RING finger protein 53) (1863 aa)					•	•	•		0.999
● HYRC	DNA-dependent protein kinase catalytic subunit (EC 2.7.11.1) (DNA-PK catalytic subunit) (DNA-PKcs) (DNPK1) (p460) (4127 aa)					•	•	•		0.999
● TP53BP1	Tumor suppressor p53-binding protein 1 (p53-binding protein 1) (p53BP1) (53BP1) (1972 aa)					•		•		0.999
● PLK3	Serine/threonine-protein kinase PLK3 (EC 2.7.11.21) (Polo-like kinase 3) (PLK-3) (Cytokine-inducible serine/threonine-protein kinase) (FGF- inducible kinase) (Proliferation-related kinase) (646 aa)					•	•	•		0.999

Protein Databases

UniProt, protein sequence and information

UniProtKB/Swiss-Prot contains
531473 sequence entries

http://www.uniprot.org/



UniProt
Search | Blast | Align | Retrieve | ID Mapping
Search in
Protein Knowledgebase (UniProtKB) ▾
Query
Search | Advanced Search » | Clear

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: ★ Swiss-Prot, which is manually annotated and reviewed. ★ TrEMBL, which is automatically annotated and is **not** reviewed. Includes Complete Proteome Sets.
UniRef	Sequence clusters, used to speed up sequence similarity searches.
UniParc	Sequence archive, used to keep track of sequences and their identifiers.
Supporting data	Literature citations, taxonomy, keywords and more.

NEWS

UniProt release 2011_08 - Jul 27, 2011

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

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL

› Forthcoming changes

› News archives

Follow @uniprot

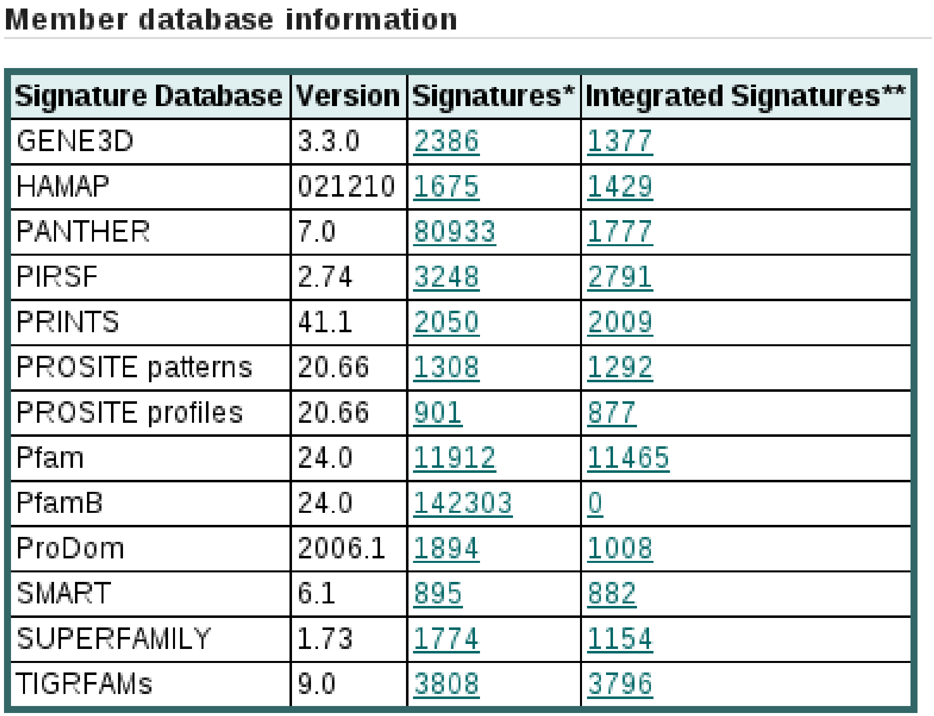
SITE TOUR

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

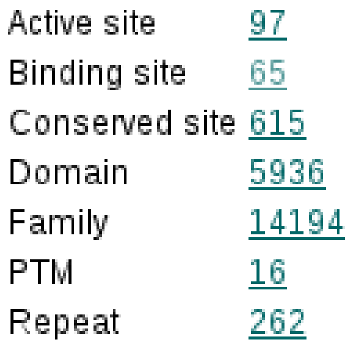
http://www.ebi.ac.uk/interpro/



Member database information

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

Contents of InterPro 31.0 (Feb 2011)



Active site	97
Binding site	65
Conserved site	615
Domain	5936
Family	14194
PTM	16
Repeat	262

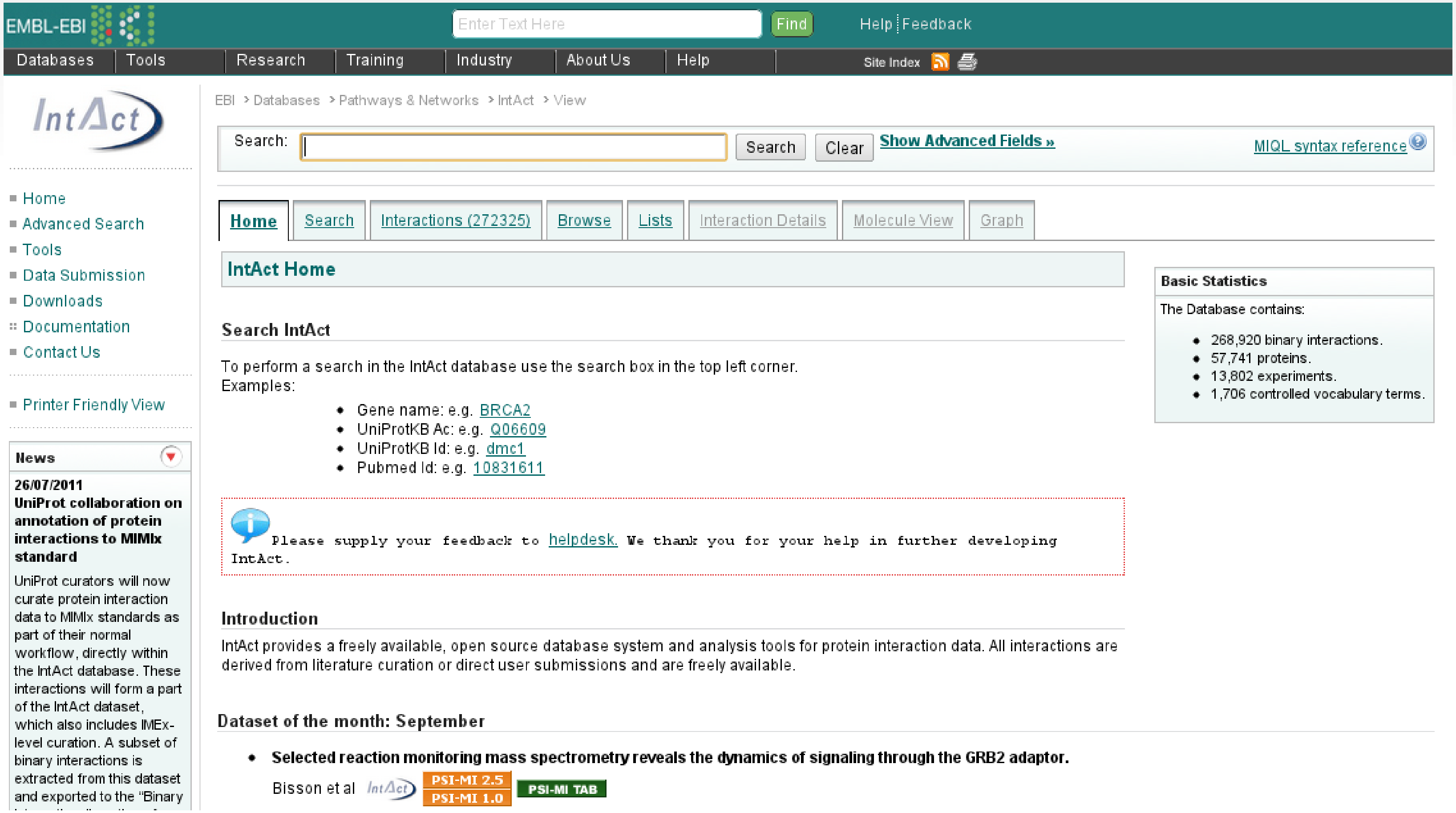
Protein Databases

IntAct, protein-protein interaction database

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

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



News

26/07/2011
UniProt collaboration on annotation of protein interactions to MIMIx standard

UniProt curators will now curate protein interaction data to MIMIx standards as part of their normal workflow, directly within the IntAct database. These interactions will form a part of the IntAct dataset, which also includes IMEx-level curation. A subset of binary interactions is extracted from this dataset and exported to the "Binary

IntAct Home

Search IntAct

To perform a search in the IntAct database use the search box in the top left corner.
Examples:

- Gene name: e.g. BRCA2
- UniProtKB Ac: e.g. Q06609
- UniProtKB Id: e.g. dmc1
- Pubmed Id: e.g. 10831611

Please supply your feedback to helpdesk. We thank you for your help in further developing IntAct.

Introduction

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.

Dataset of the month: September

- **Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor.**
  Bisson et al

Basic Statistics

The Database contains:

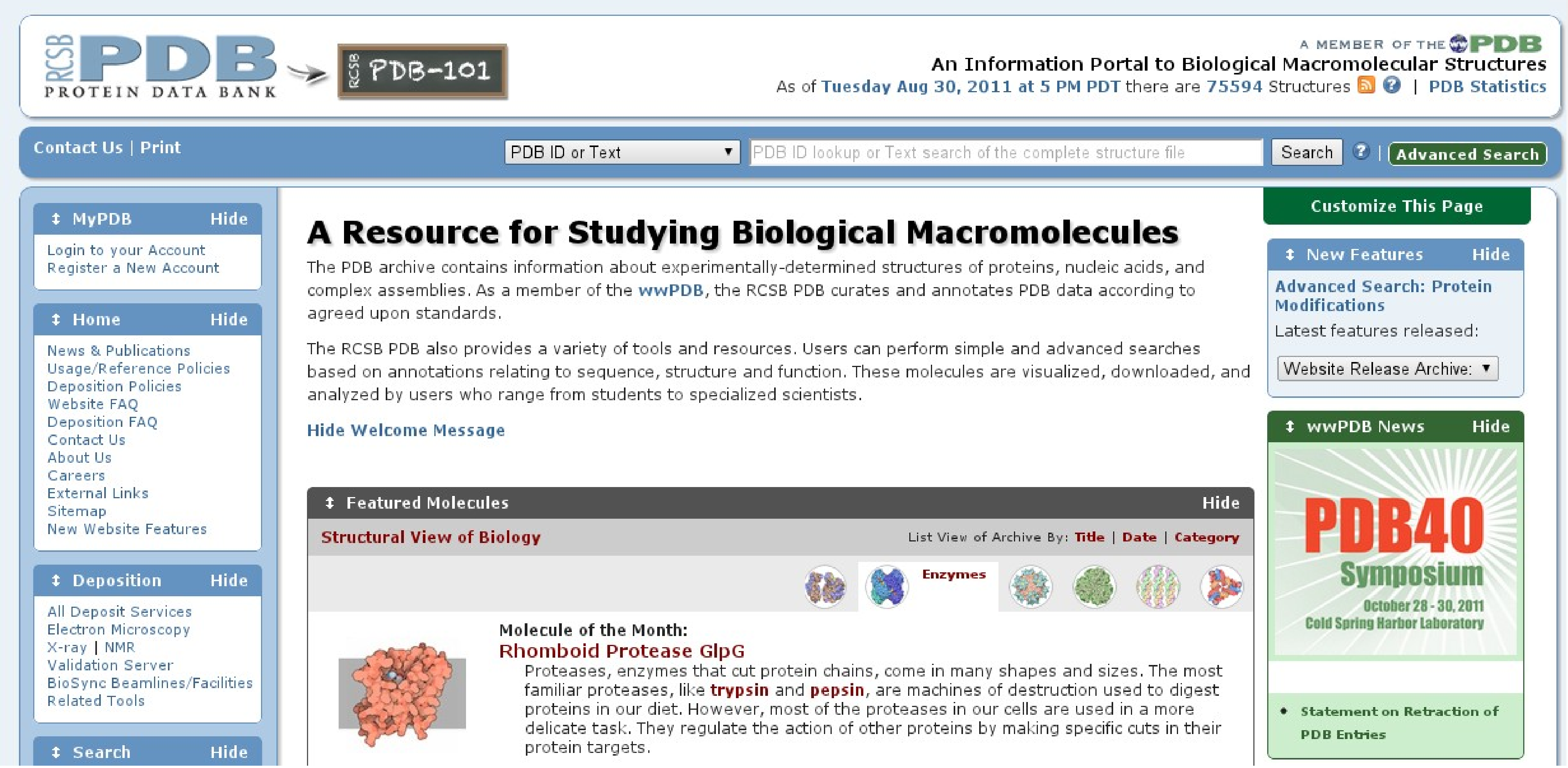
- 268,920 binary interactions.
- 57,741 proteins.
- 13,802 experiments.
- 1,706 controlled vocabulary terms.

Protein Databases

Protein Data Bank (*PDB*)

The ***PDB*** archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies

http://www.rcsb.org/pdb/home/home.do



RCSB PDB PROTEIN DATA BANK → RCSB PDB-101

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Aug 30, 2011 at 5 PM PDT there are 75594 Structures | PDB Statistics

Contact Us | Print

PDB ID or Text ▾ PDB ID lookup or Text search of the complete structure file Search ? | Advanced Search

MyPDB Hide
- Login to your Account
- Register a New Account

Home Hide
- News & Publications
- Usage/Reference Policies
- Deposition Policies
- Website FAQ
- Deposition FAQ
- Contact Us
- About Us
- Careers
- External Links
- Sitemap
- New Website Features

Deposition Hide
- All Deposit Services
- Electron Microscopy
- X-ray | NMR
- Validation Server
- BioSync Beamlines/Facilities
- Related Tools

Search Hide

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Hide Welcome Message

Featured Molecules Hide

Structural View of Biology — List View of Archive By: Title | Date | Category

Enzymes

Molecule of the Month:

Rhomboid Protease GlpG

Proteases, enzymes that cut protein chains, come in many shapes and sizes. The most familiar proteases, like trypsin and pepsin, are machines of destruction used to digest proteins in our diet. However, most of the proteases in our cells are used in a more delicate task. They regulate the action of other proteins by making specific cuts in their protein targets.

Customize This Page

New Features Hide

Advanced Search: Protein Modifications

Latest features released:

Website Release Archive: ▾

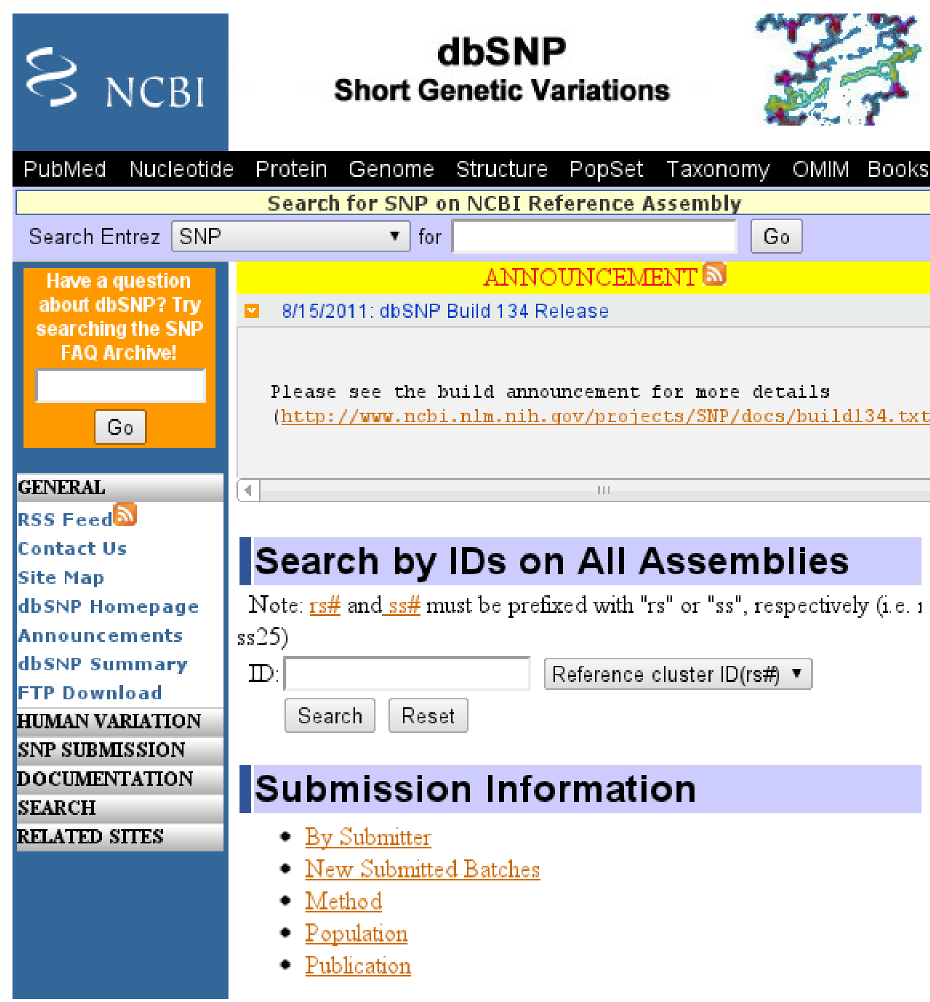
wwPDB News Hide

PDB40 Symposium October 28 - 30, 2011 Cold Spring Harbor Laboratory

- Statement on Retraction of PDB Entries

Variation Databases

dbSNP, the repository of all the *SNPs*



8/15/2011: dbSNP Build 134 Release

Please see the build announcement for more details (http://www.ncbi.nlm.nih.gov/projects/SNP/docs/build134.txt)

Search by IDs on All Assemblies

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

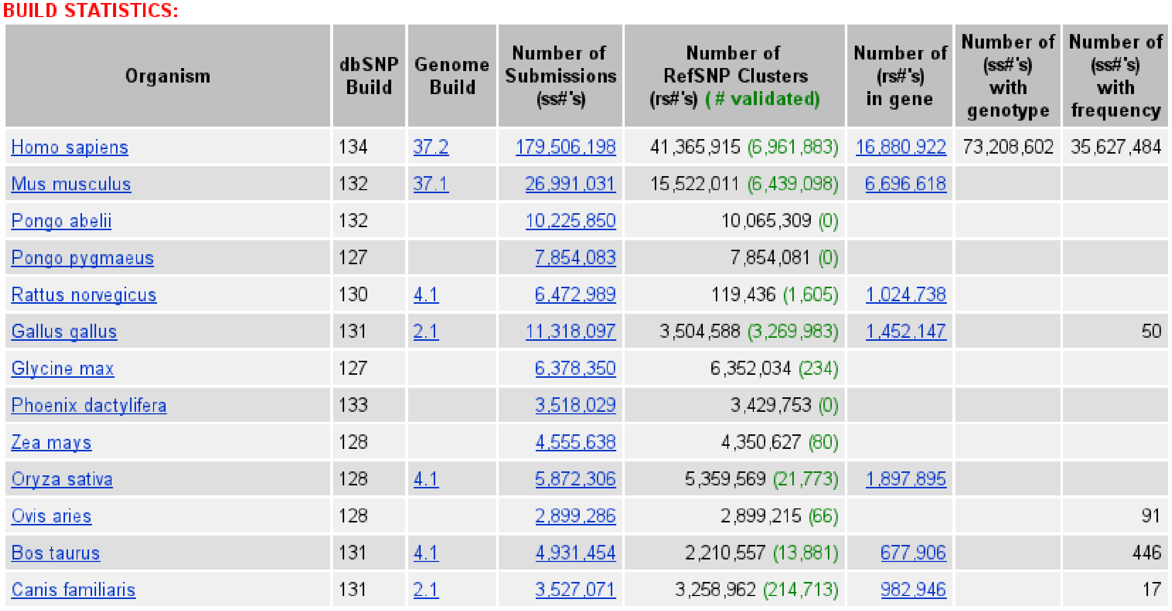
ID: Reference cluster ID(rs#)

Search Reset

Submission Information

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

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



BUILD STATISTICS:

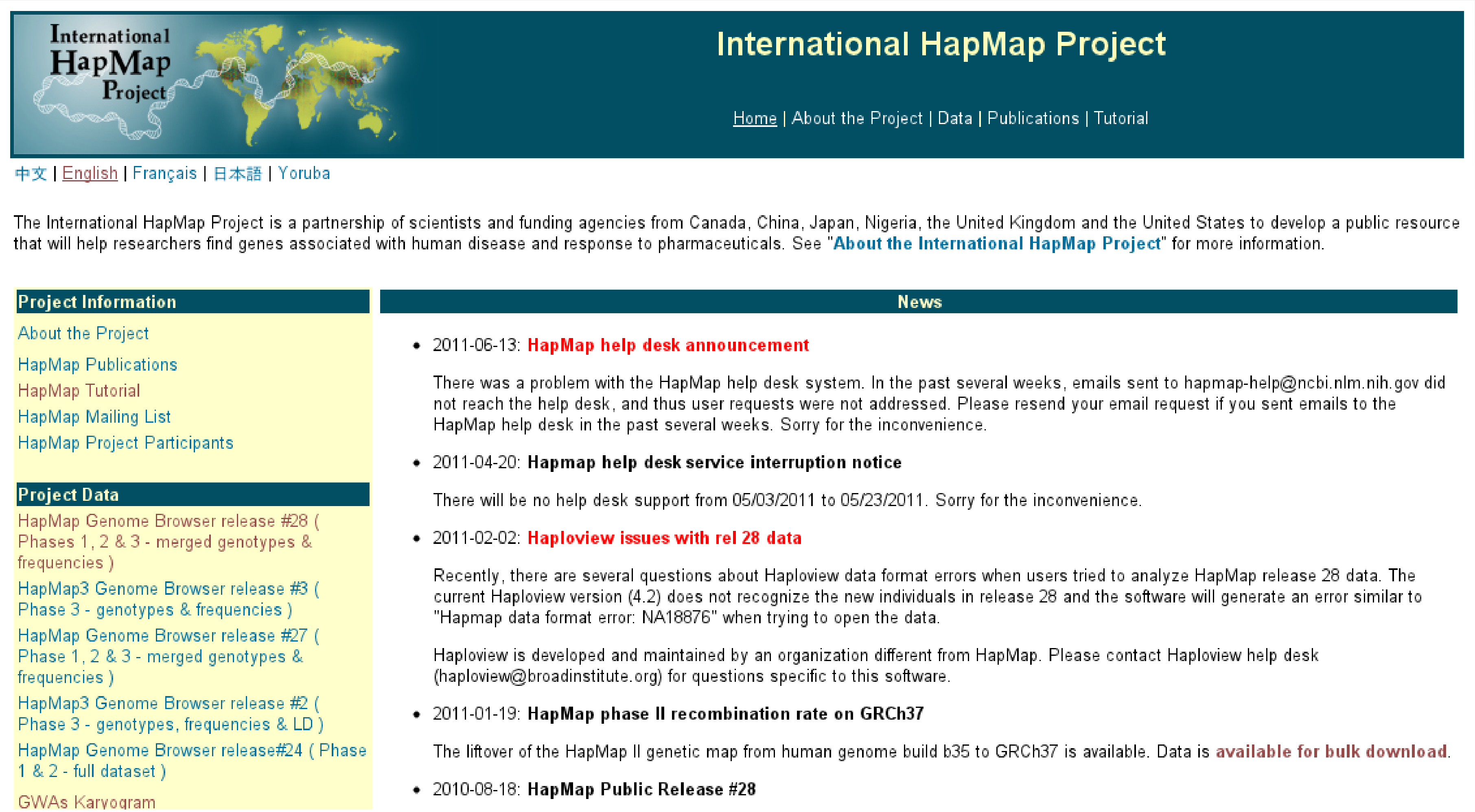
Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	134	37.2	179,506,198	41,365,915 (6,961,883)	16,880,922	73,208,602	35,627,484
Mus musculus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

Variation Databases

HapMap, human Haplotype Map

To develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals

http://hapmap.ncbi.nlm.nih.gov/



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

Project Data

HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)

HapMap3 Genome Browser release #3 (Phase 3 - genotypes & frequencies)

HapMap Genome Browser release #27 (Phase 1, 2 & 3 - merged genotypes & frequencies)

HapMap3 Genome Browser release #2 (Phase 3 - genotypes, frequencies & LD)

HapMap Genome Browser release#24 (Phase 1 & 2 - full dataset)

GWAs Karyogram

News

- 2011-06-13: **HapMap help desk announcement**

  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.

- 2011-04-20: **Hapmap help desk service interruption notice**

  There will be no help desk support from 05/03/2011 to 05/23/2011. Sorry for the inconvenience.

- 2011-02-02: **Haploview issues with rel 28 data**

  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.

  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.

- 2011-01-19: **HapMap phase II recombination rate on GRCh37**

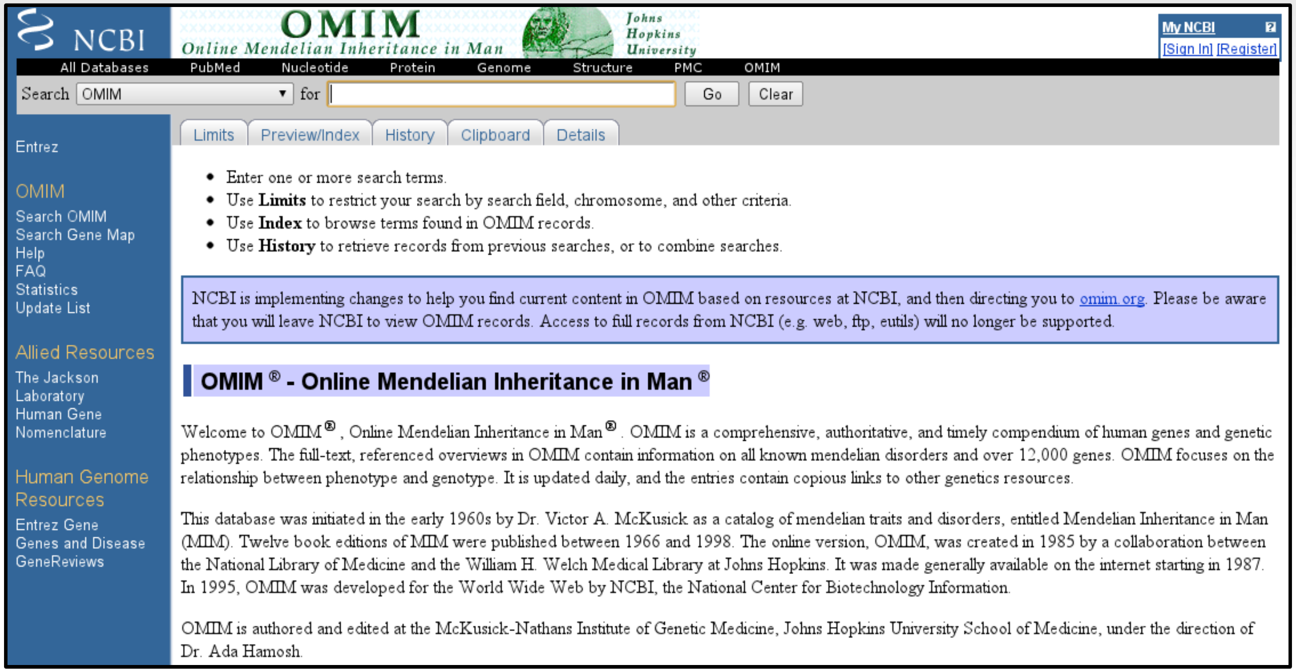
  The liftover of the HapMap II genetic map from human genome build b35 to GRCh37 is available. Data is **available for bulk download**.

- 2010-08-18: **HapMap Public Release #28**

Variation Databases

Mutations: OMIM, COSMIC, Mitelman, ...

http://www.ncbi.nlm.nih.gov/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. Please be aware that you will leave NCBI to view OMIM records. Access to full records from NCBI (e.g. web, ftp, eutils) will no longer be supported.

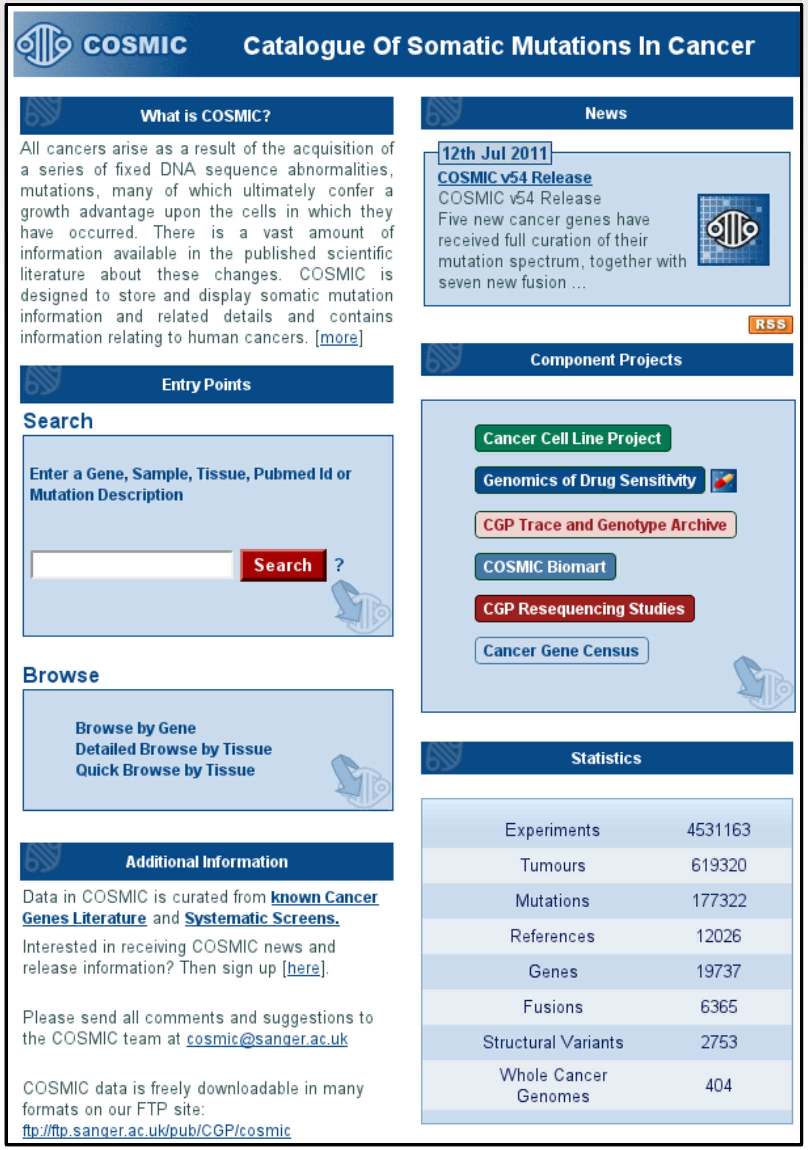
OMIM ® - Online Mendelian Inheritance in Man ®

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/



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

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

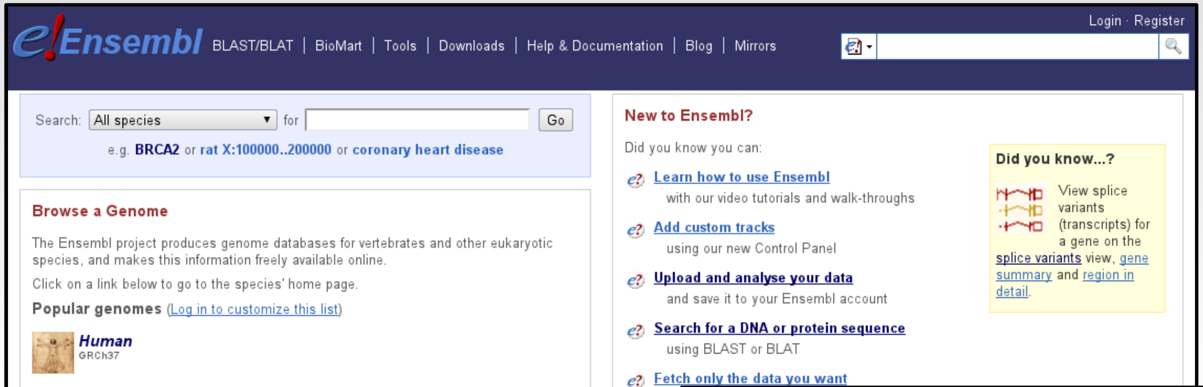
COSMIC data is freely downloadable in many formats on our FTP site: ftp://ftp.sanger.ac.uk/pub/CGP/cosmic

Statistics

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

Genome DDBB and Browsers

Ensembl, the most used and reliable



e!Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login · Register

Search: All species for Go

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

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.

Popular genomes (Log in to customize this list)

- *Human* GRCh37
- *Mouse* NCBIM37
- *Zebrafish* Zv9

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl *Pre!* and EnsemblGenomes

New to Ensembl?

Did you know you can:

- Learn how to use Ensembl
  with our video tutorials and walk-throughs
- Add custom tracks
  using our new Control Panel
- Upload and analyse your data
  and save it to your Ensembl account
- Search for a DNA or protein sequence
  using BLAST or BLAT
- Fetch only the data you want
  fro
- Dow
  in
- Min
  an

Still got

Did you know...?

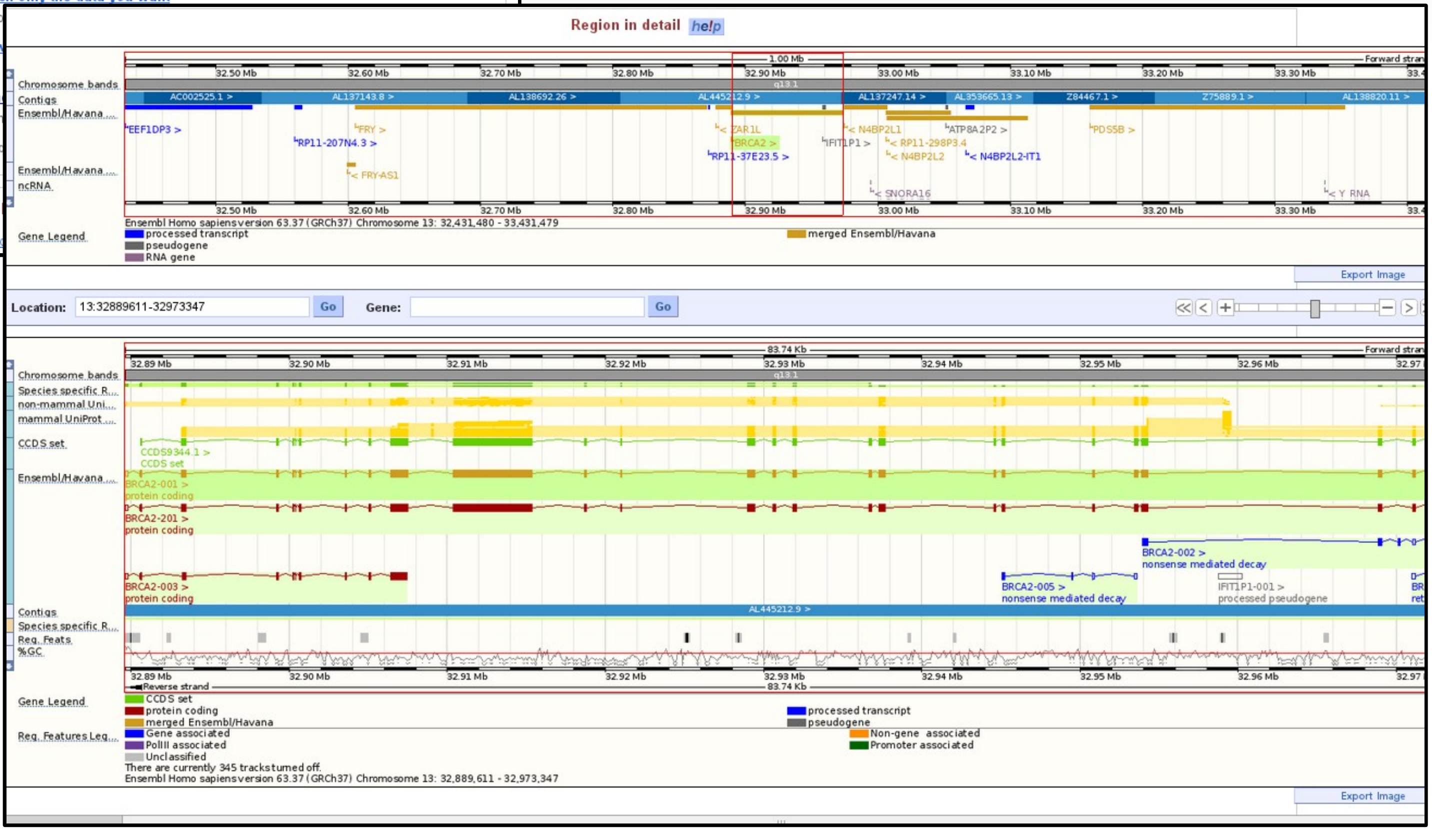
View splice variants (transcripts) for a gene on the splice variants view, gene summary and region in detail.

What's

- S

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

The ***Ensembl*** project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online



Region in detail he!p

Ensembl Homo sapiens version 63.37 (GRCh37) Chromosome 13: 32,431,480 - 33,431,479

- 🟦 processed transcript
- ⬛ pseudogene
- 🟪 RNA gene
- 🟧 merged Ensembl/Havana

Export Image

Location: 13:32889611-32973347 Go **Gene:** Go

- 🟩 CCDS set
- 🟥 protein coding
- 🟧 merged Ensembl/Havana
- 🟦 processed transcript
- ⬛ pseudogene
- 🟦 Gene associated
- 🟪 PolIII associated
- ⬜ Unclassified
- 🟧 Non-gene associated
- 🟩 Promoter associated

There are currently 345 tracks turned off.

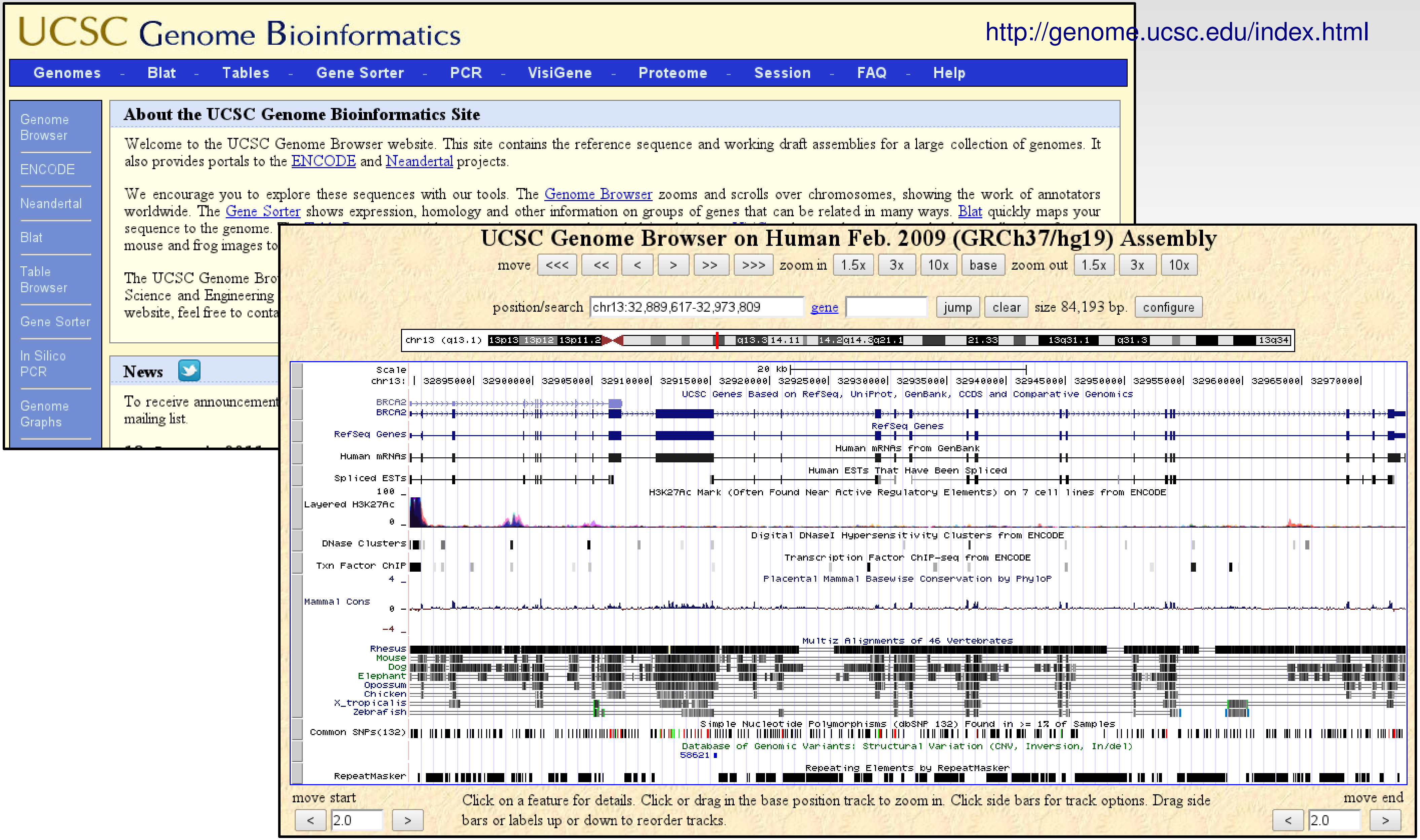
Ensembl Homo sapiens version 63.37 (GRCh37) Chromosome 13: 32,889,611 - 32,973,347

Export Image

Genome DDBB and Browsers

UCSC

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



UCSC Genome Bioinformatics

Genomes - Blat - Tables - Gene Sorter - PCR - VisiGene - Proteome - Session - FAQ - Help

About the UCSC Genome Bioinformatics Site

Welcome to the UCSC Genome Browser website. This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides portals to the ENCODE and Neandertal projects.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome.

mouse and frog images to

The UCSC Genome Bro

Science and Engineering

website, feel free to conta

News

To receive announcement

mailing list.

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x
position/search
gene
jump
clear
configure

position/search chr13:32,889,617-32,973,809 size 84,193 bp.

chr13 (q13.1)
13p13
13p12
13p11.2
q13.3
14.11
14.2
q14.3
q21.1
21.33
13q31.1
q31.3
13q34
Scale
20 kb
chr13:
32895000
32900000
32905000
32910000
32915000
32920000
32925000
32930000
32935000
32940000
32945000
32950000
32955000
32960000
32965000
32970000
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
BRCA2
BRCA2
RefSeq Genes
RefSeq Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Placental Mammal Basewise Conservation by PhyloP
4
Mammal Cons
0
-4
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 132) Found in >= 1% of Samples
Common SNPs(132)
Database of Genomic Variants: Structural Variation (CNV, Inversion, In/del)
58621
Repeating Elements by RepeatMasker
RepeatMasker
move start
< 2.0 >
move end
< 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks.

Ejercicios
Introducción

- Estamos interesados en estudiar un gen llamado ***BCL2****.*
- Utilizando las bases de datos explicadas anteriormente vamos a buscar información acerca de:
  - Secuencia génica
  - Información funcional y reguladora
  - Variaciones conocidas
  - Proteinas

Ejercicios

Información sobre la secuencia génica de *BCL2*

- Desde la página de Ensembl (http://www.ensembl.org/) intenta responder a las siguientes preguntas:
  - Indica la localización del gen y en qué cadena se encuentra
  - ¿Para cuántos tránscritos codifica?
  - ¿Y para cuantas proteínas?
  - Encuentra su secuencia de DNA
  - Indica el número de exones que contiene el gen

Ejercicios

Información funcional y reguladora de *BCL2*

- Ayúdate de GO (http://www.geneontology.org/) y encuentra los términos “biological process” y “cellular components” (GO terms) relacionados con el gen.

- Utiliza MicroCosm (http://www.ebi.ac.uk/enright-srv/microcosm/) para determinar si existe algún microRNA que regule a este gen.

- Dirígete a OregAnno (http://www.oreganno.org/). ¿Existe algún factor de transcripción conocido que regule a este gen?

- Utiliza KEGG (http://www.genome.jp/kegg/) y Reactome (http://www.reactome.org/) para determinar en qué rutas (pathways) podemos encontrar este gen involucrado.

Ejercicios

Información sobre variaciones en *BCL2*

- Consulta en dbSNP (http://www.ncbi.nlm.nih.gov/snp/) el número de SNPs localizados en la secuencia de nuestro gen.
- En OMIM (http://www.ncbi.nlm.nih.gov/omim/) podemos encontrar información médica relacionada con mutaciones en genes. ¿Existe alguna enfermedad relacionada con nuestro gen?
- UniProtKB (http://www.uniprot.org/) contiene anotaciones sobre las consecencias observadas al mutar determinados aminoácidos en las secuencias proteicas. Observa qué tipo de consecuencias pueden tener estas mutaciones sobre la proteína.

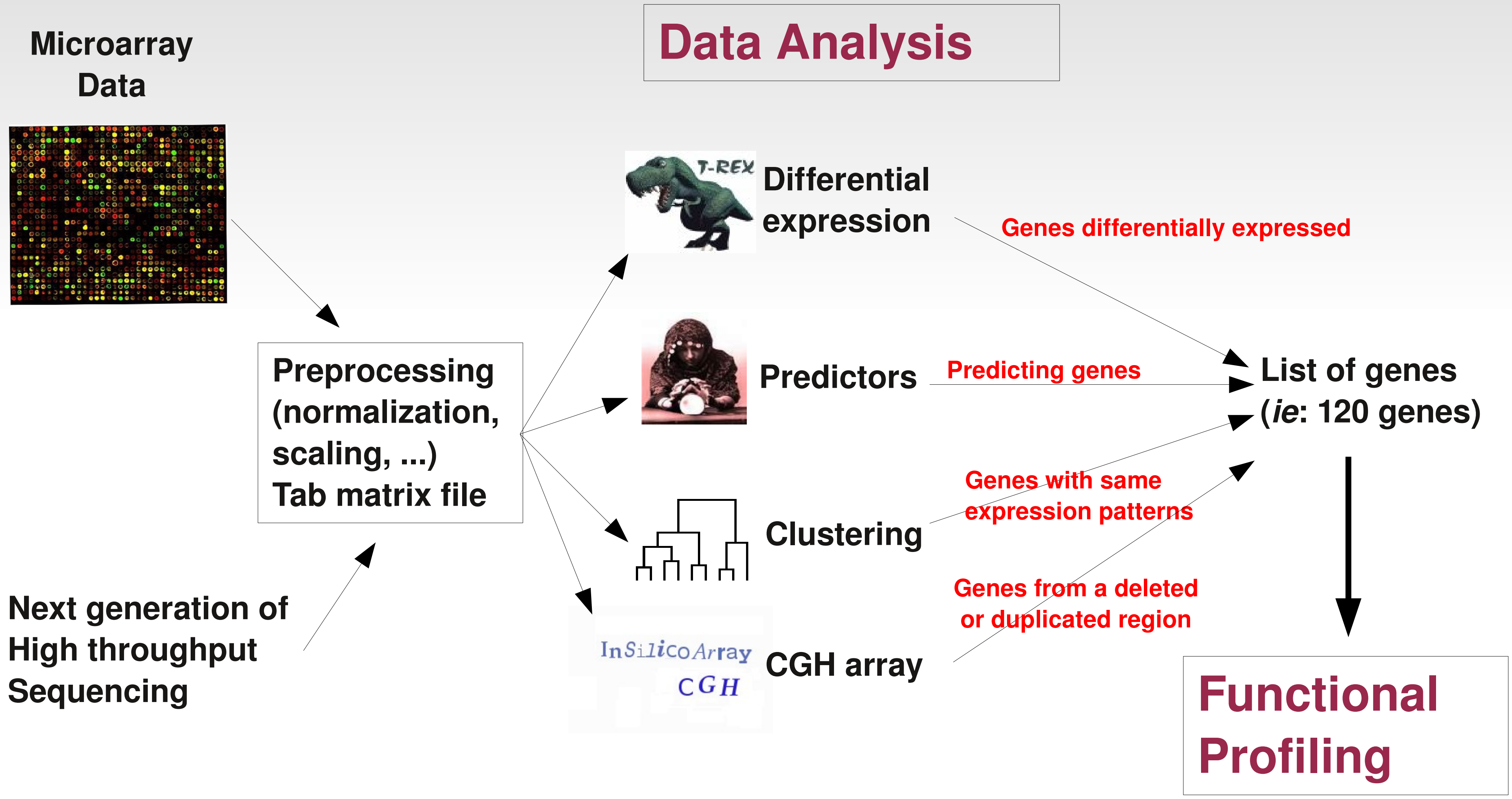
Ejercicios

Información sobre las proteínas producidas por *BCL2*

- Obtén la siguiente inforamción de la proteína BCL2:
  - Secuencia (*UniProt*)
  - ¿Qué dominos proteicos funcionales tienen (*interpro*)?
  - Estructura 3D (*PDB*)

Functional Annotation DDBB

From GEPAS to Babelomics



Data Analysis

Microarray Data

Preprocessing (normalization, scaling, ...) Tab matrix file

Next generation of High throughput Sequencing

Differential expression

Predictors

Clustering

CGH array

Genes differentially expressed

Predicting genes

Genes with same expression patterns

Genes from a deleted or duplicated region

List of genes (*ie*: 120 genes)

Functional Profiling