IX International Course of Massive Data Analysis FOR GENOMICS

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

Ignacio Medina
imedina@cipf.es

Biological Databases

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 an 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

Ignacio Medina
imedina@cipf.es

Introduction
Sequence databases

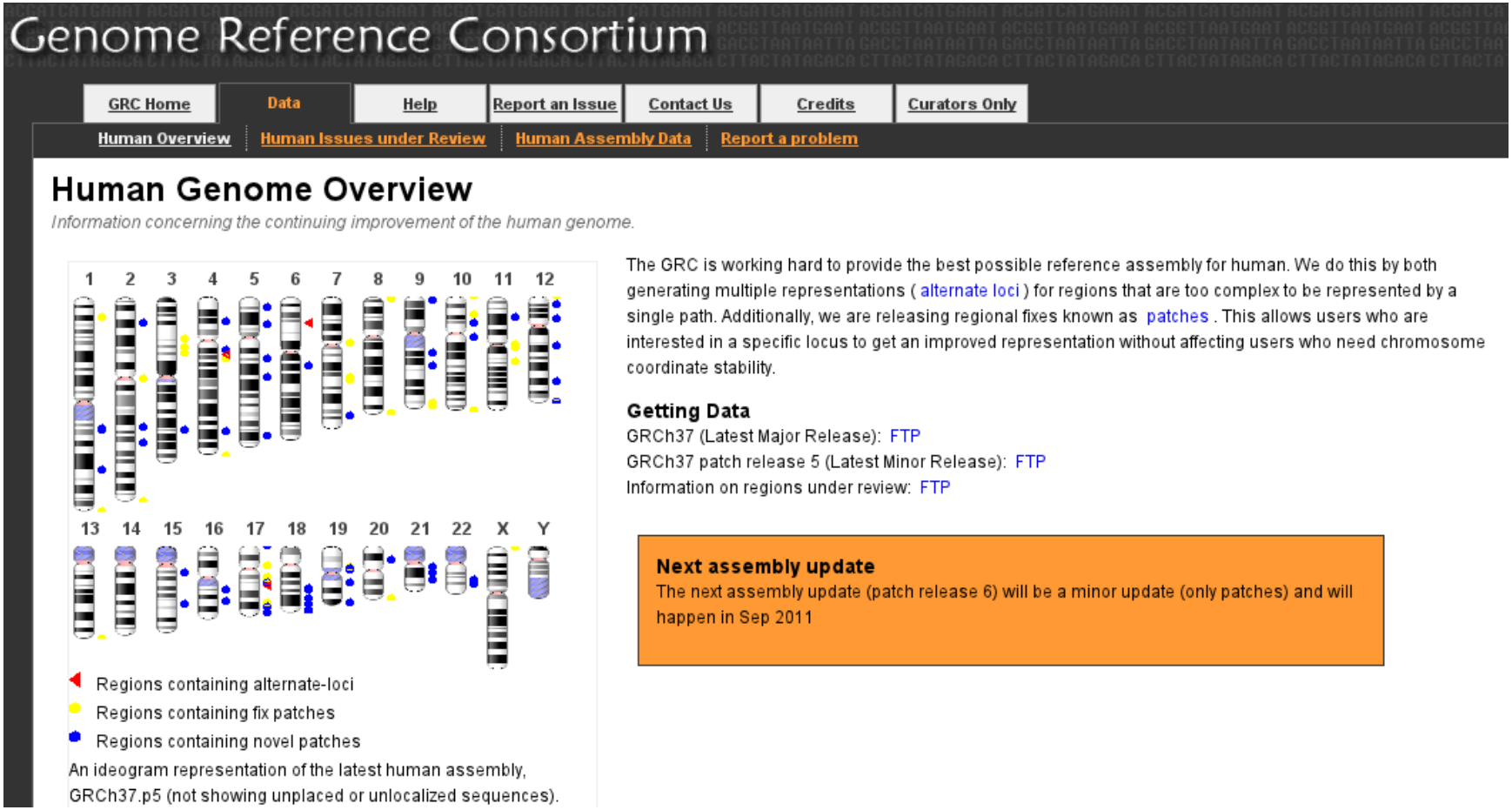
Genome Reference Consortium (GRC)

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.



Ignacio Medina
imedina@cipf.es

Introduction
Sequence databases

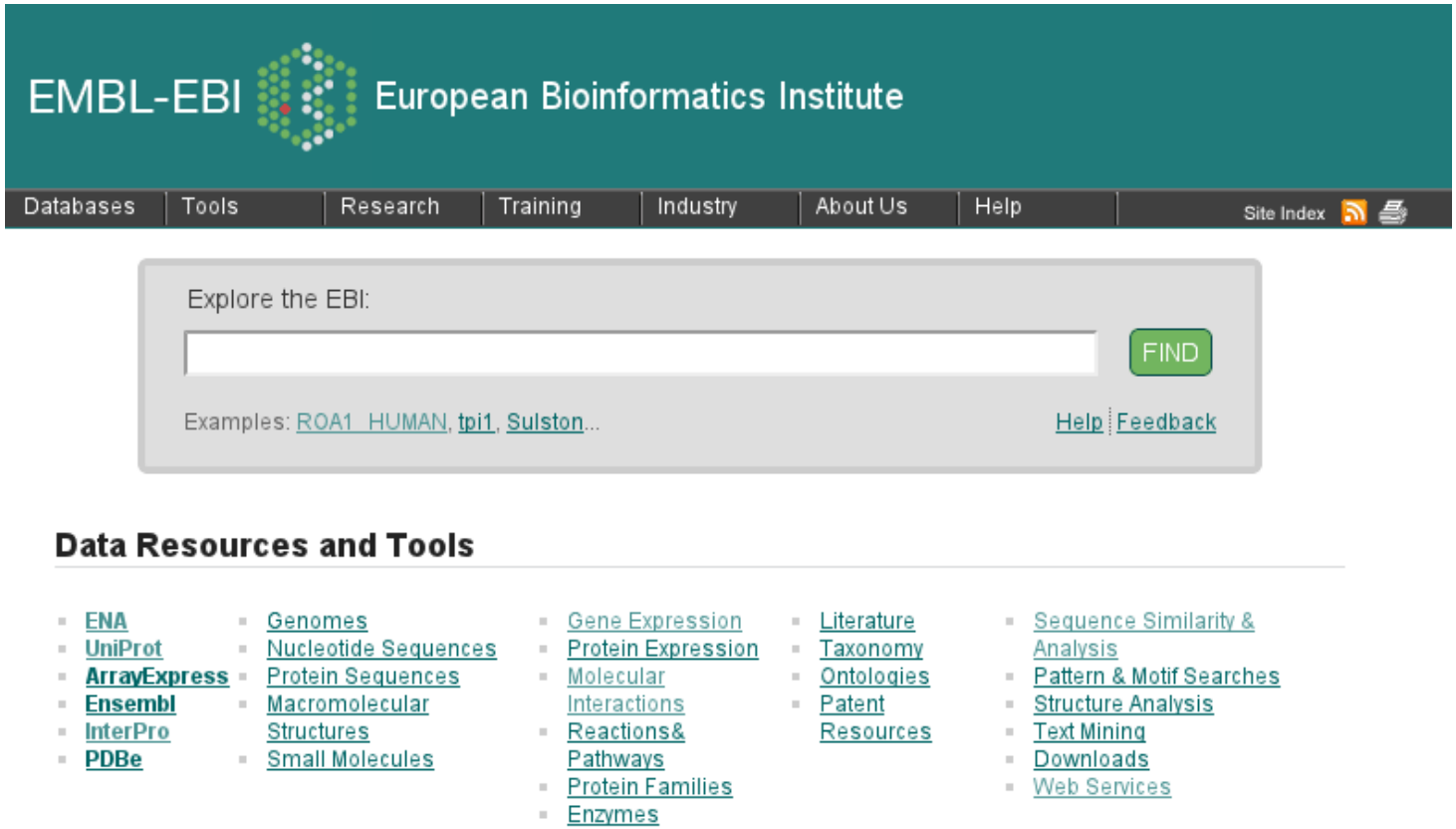
Mission:

- 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/



Ignacio Medina
imedina@cipf.es

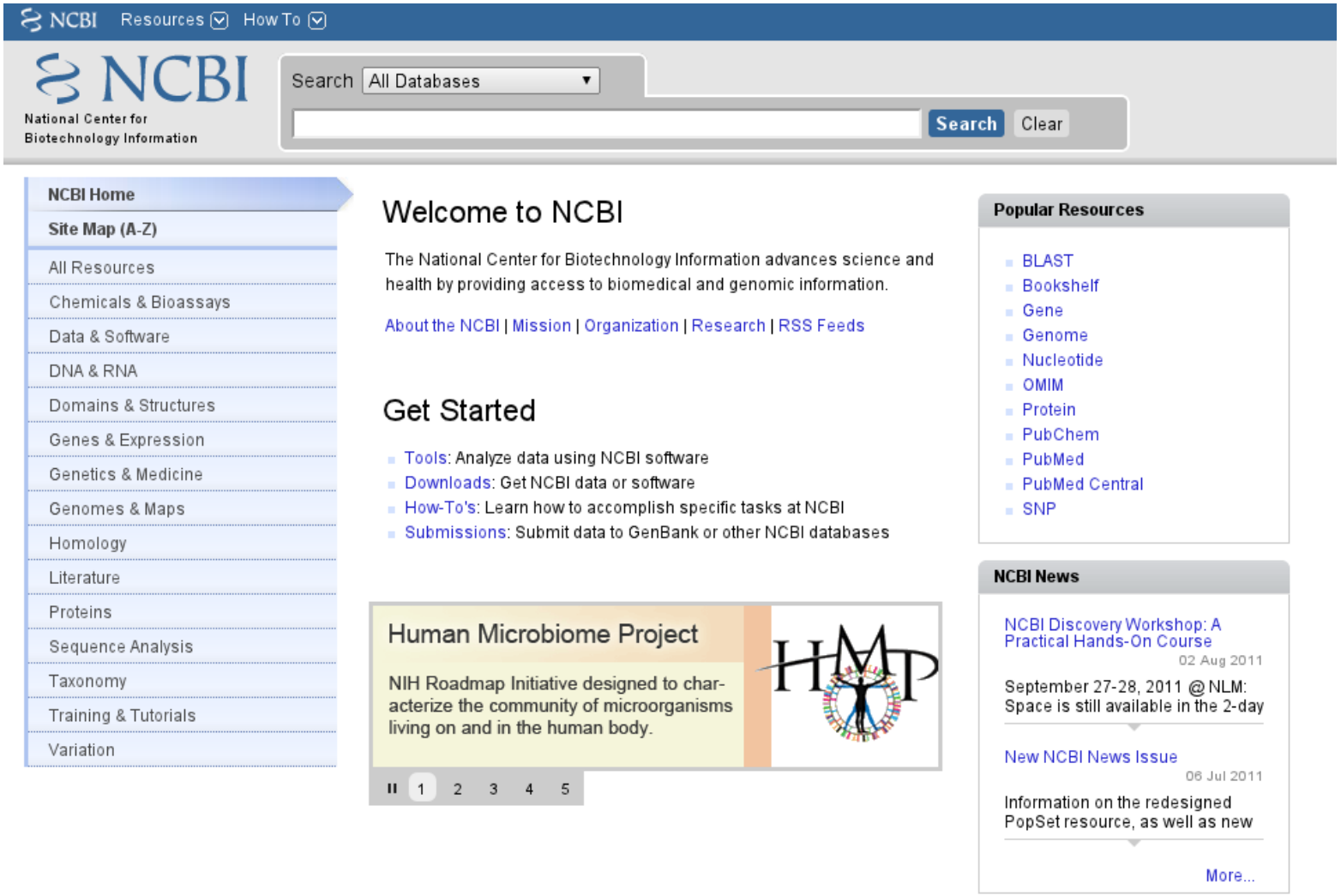
Introduction
Sequence databases

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

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.



Ignacio Medina
imedina@cipf.es

Introduction
Functional Annotation

Overview

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

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
- **InterPro Motifs**
- **Gene Ontology**
  - Biological Process
  - Molecular Function
  - Cellular Component
- **Regulatory elements**
  - miRNA
  - CisRed
  - Transcription Factor
  - Binding Sites
- **Bioentities from literature:**
  - Diseases terms
  - Chemical terms
- **Keywords Swissprot**
- **Biocarta pathways**
- **Gene Expression in tissues**

Ignacio Medina
imedina@cipf.es

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/

Ignacio Medina
imedina@cipf.es

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)

Ignacio Medina
imedina@cipf.es

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

Ignacio Medina
imedina@cipf.es

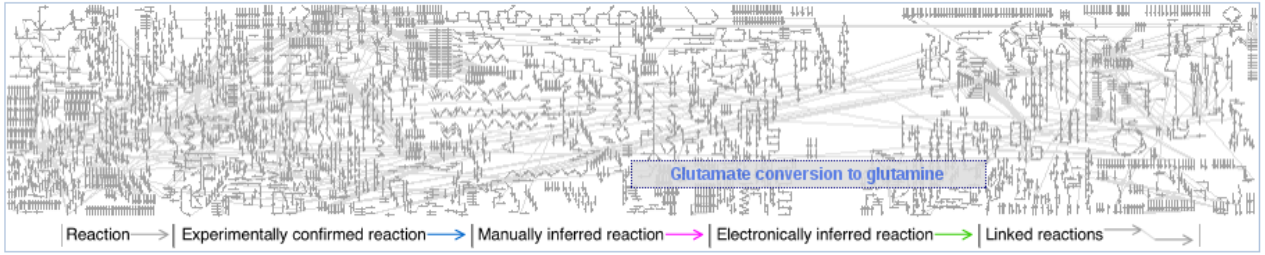
Introduction
Functional Annotation

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	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			

Ignacio Medina
imedina@cipf.es

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/

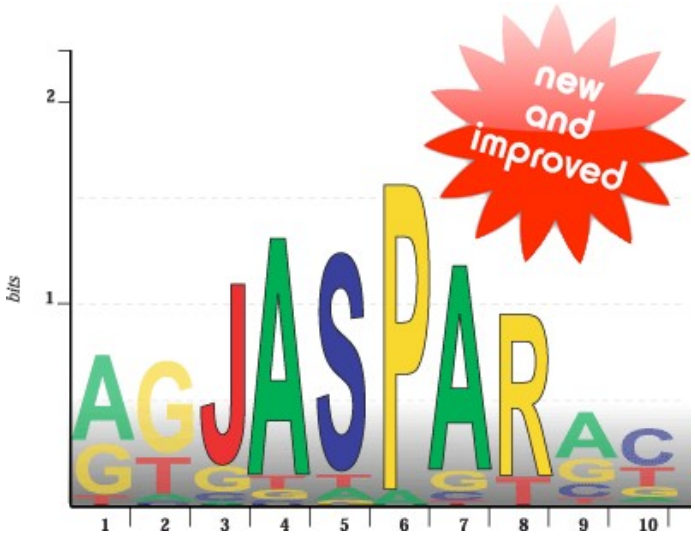
http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/

Ignacio Medina
imedina@cipf.es

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 acess, non-redundancy and quality.



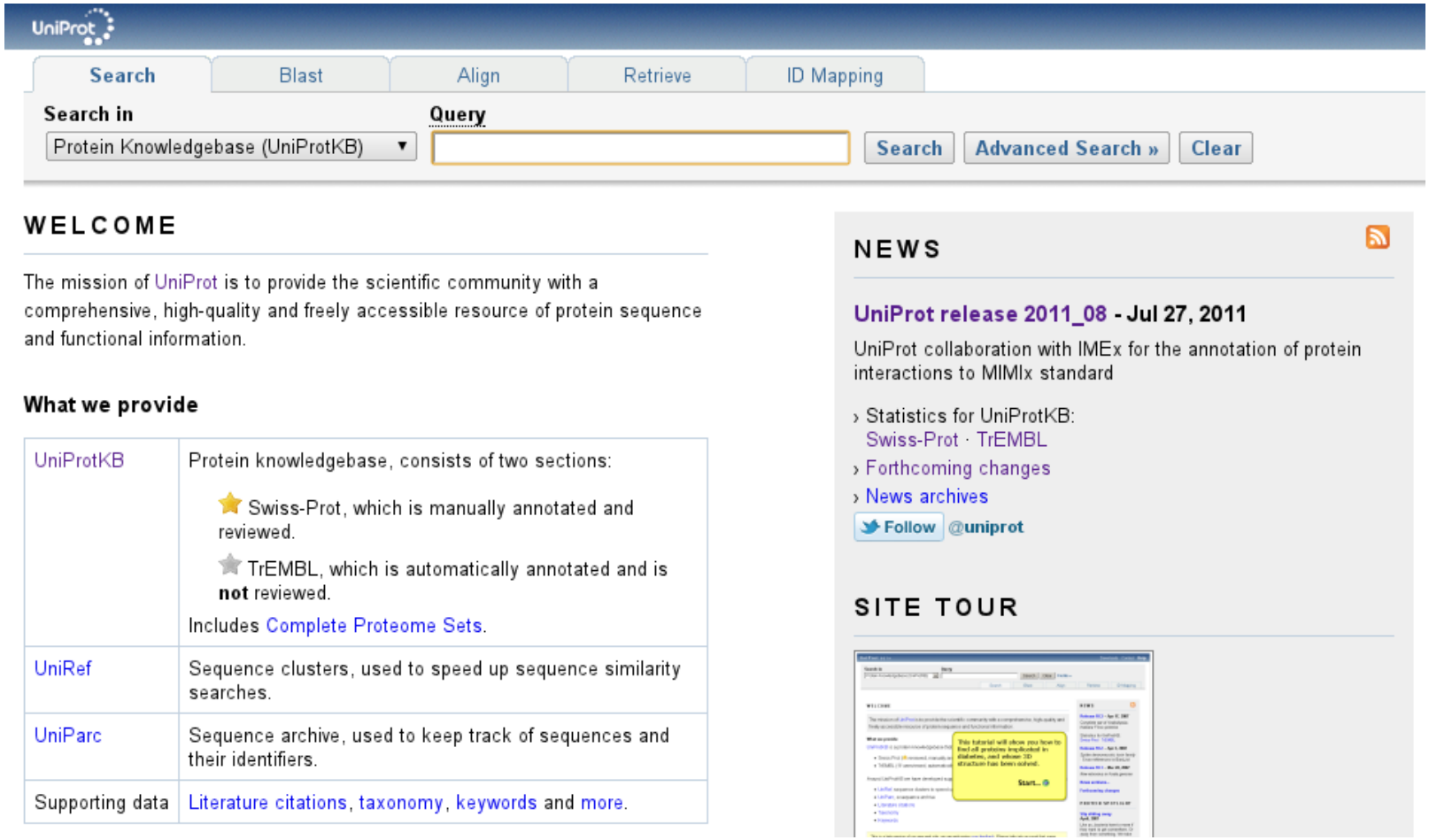
http://jaspar.genereg.net/

Ignacio Medina
imedina@cipf.es

Introduction
Protein databases

UniProt, protein sequence and information
http://www.uniprot.org/

UniProtKB/Swiss-Prot
contains 531473 sequence entries



Ignacio Medina
imedina@cipf.es

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	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

Ignacio Medina
imedina@cipf.es

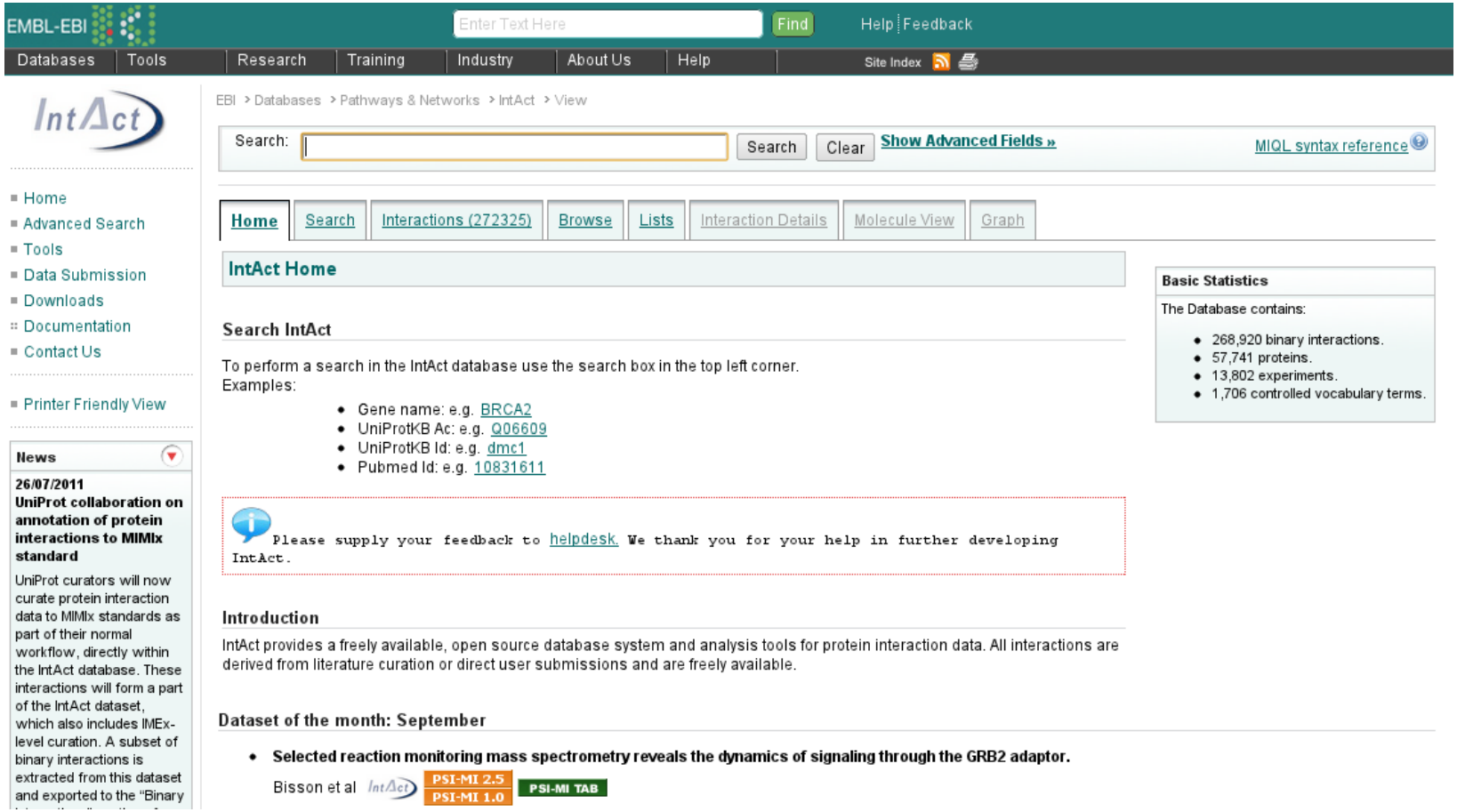
Introduction
Protein databases

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.

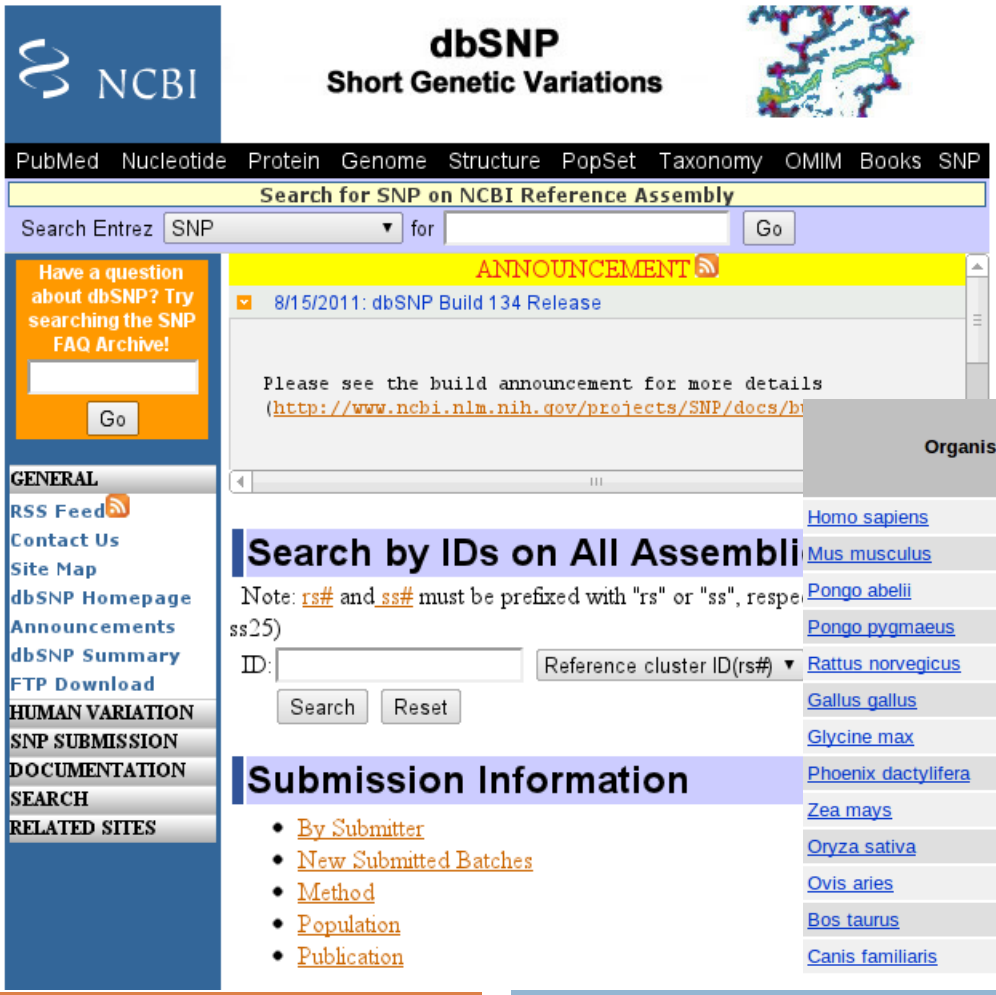


Ignacio Medina
imedina@cipf.es

Introduction
Variation databases

dbSNP, the repository of all the SNPs

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



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
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

Ignacio Medina
imedina@cipf.es

Introduction
Variation databases

HapMap, Human Haplotype Map

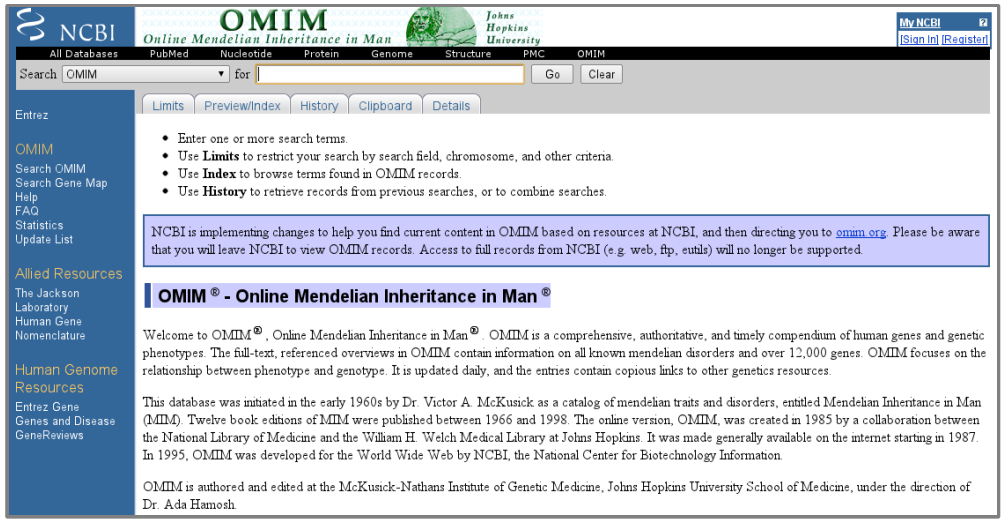
The goal is to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals.

Ignacio Medina
imedina@cipf.es

Introduction
Variation databases

Mutations: OMIM, COSMIC, Mitelman, etc.

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

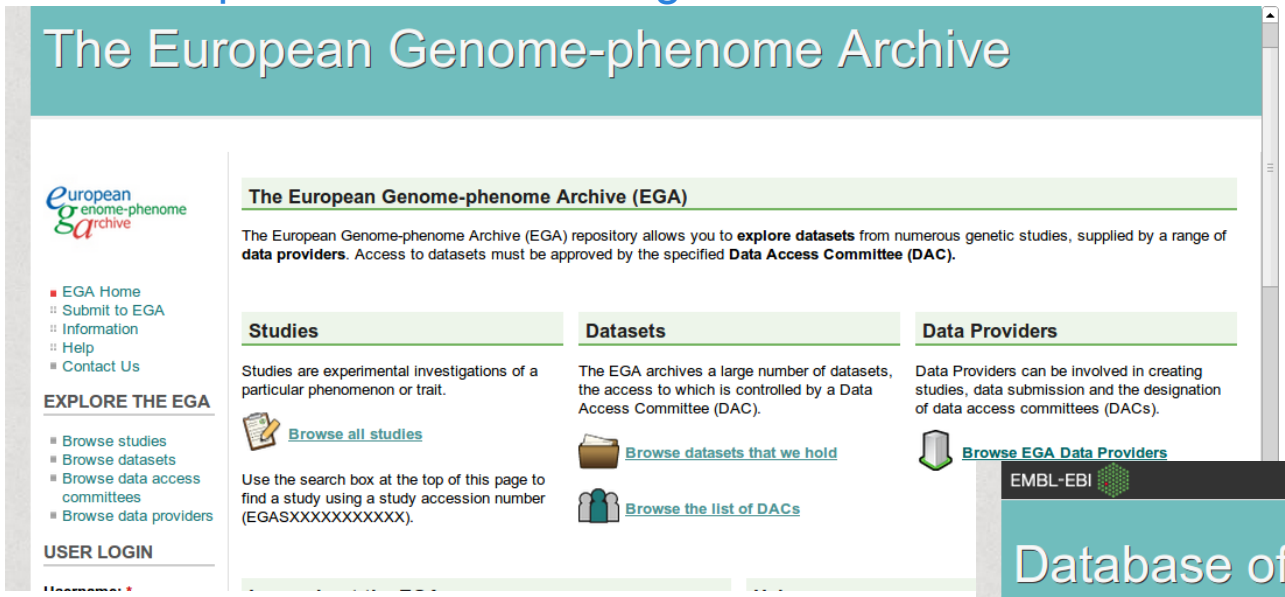


http://www.sanger.ac.uk/genetics/CGP/cosmic/

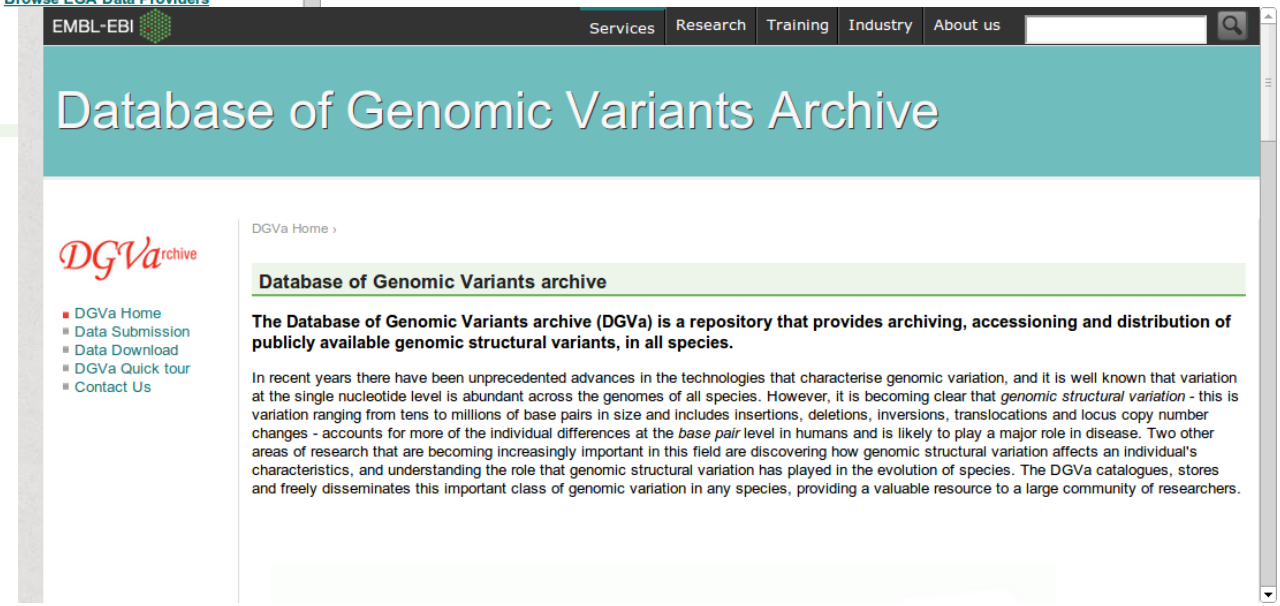
Ignacio Medina
imedina@cipf.es

Introduction
Variation databases

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



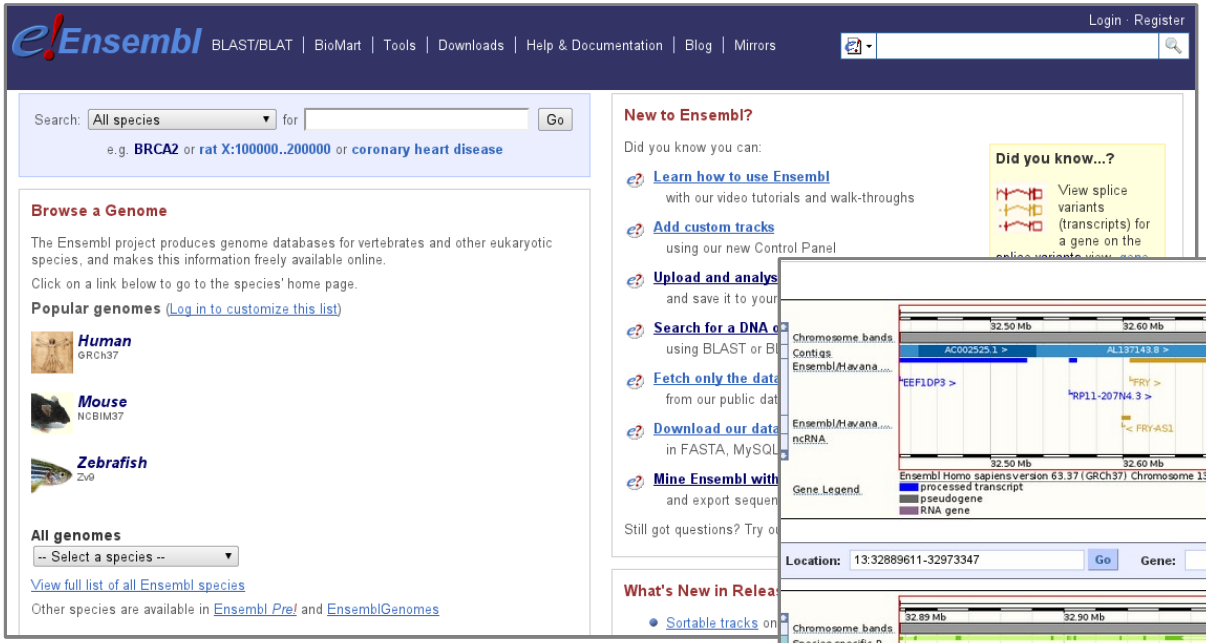
http://www.ebi.ac.uk/dgva

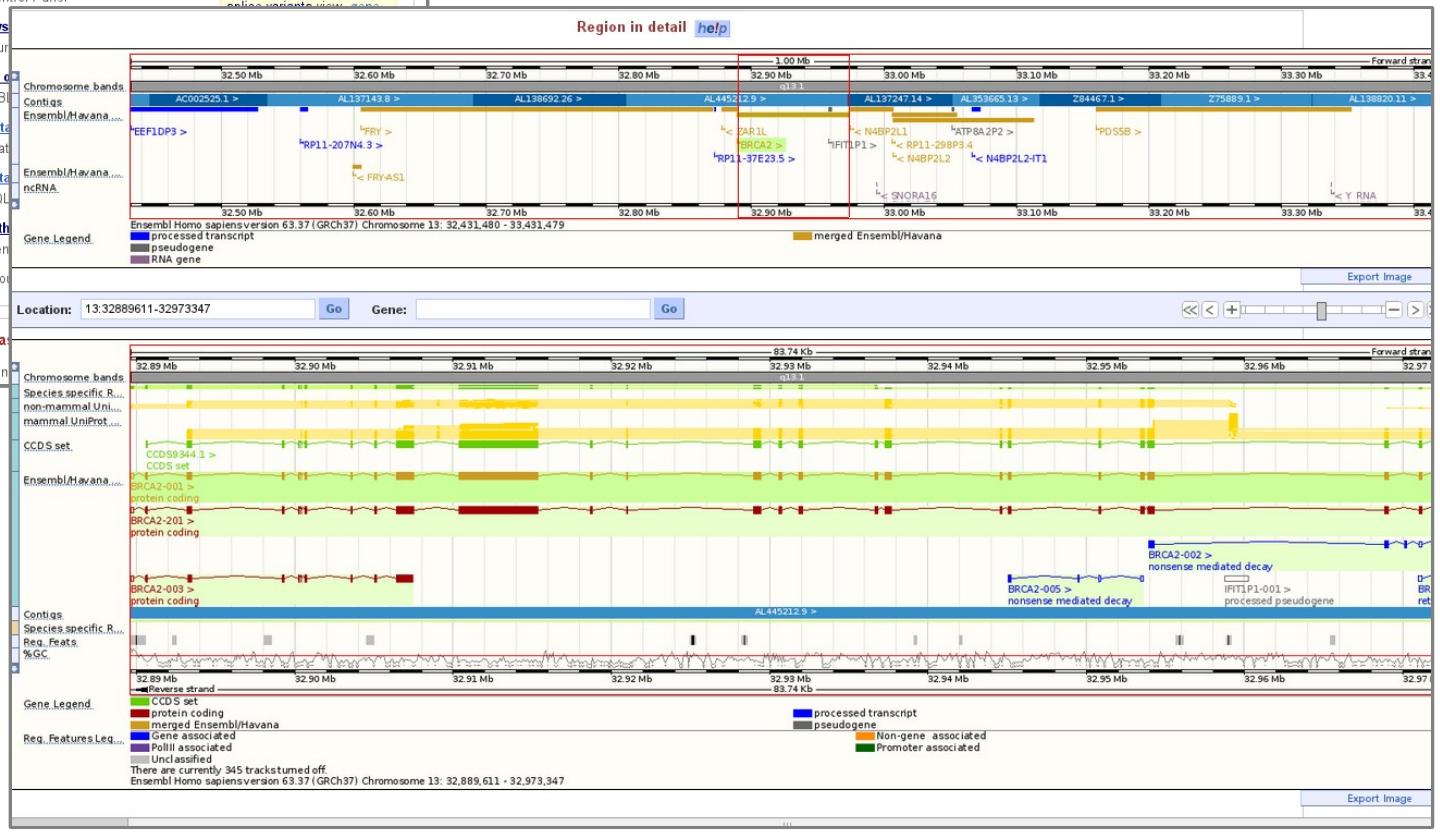


Ignacio Medina
imedina@cipf.es

Introduction

Genome annotation databases and browsers





Ensembl, the most used

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

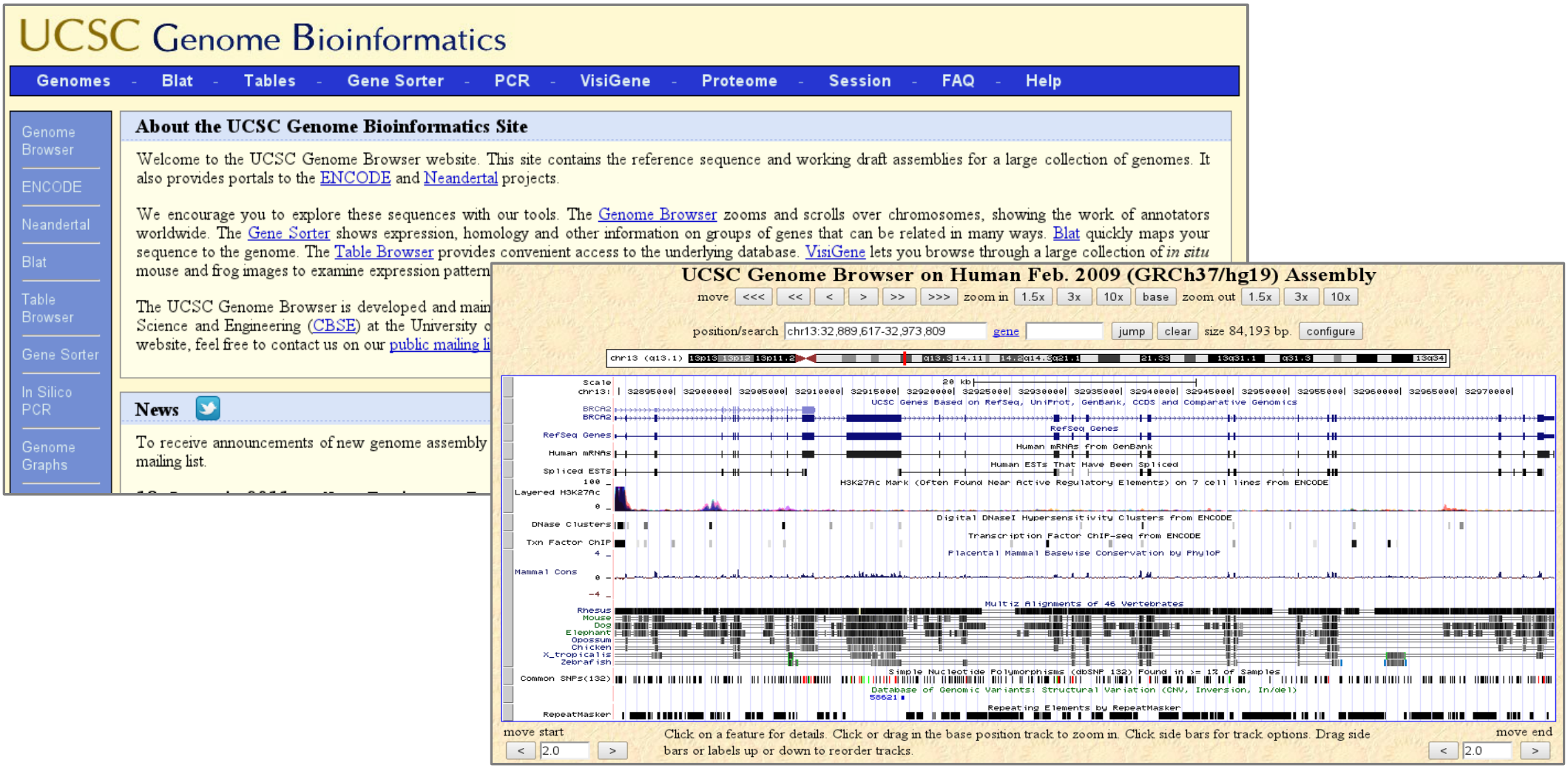
Ignacio Medina
imedina@cipf.es

Introduction

Genome annotation databases and browsers

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

UCSC Genome Browser

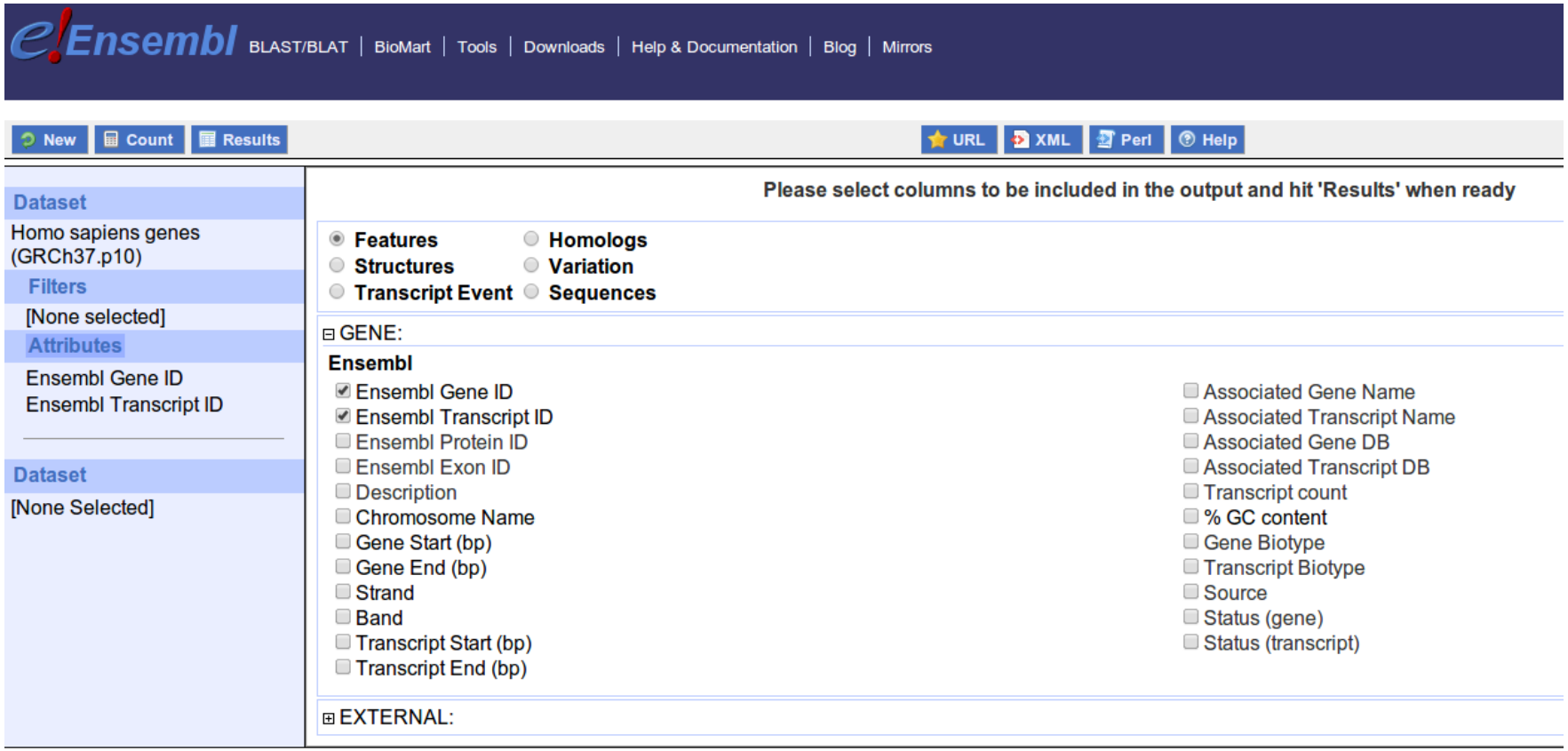


Ignacio Medina
imedina@cipf.es

Ensembl Biomart
Data mining with Biomart

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

- Oriented to small queries
- Easy interface for users



Ignacio Medina
imedina@cipf.es

CellBase
Motivation

Motivation

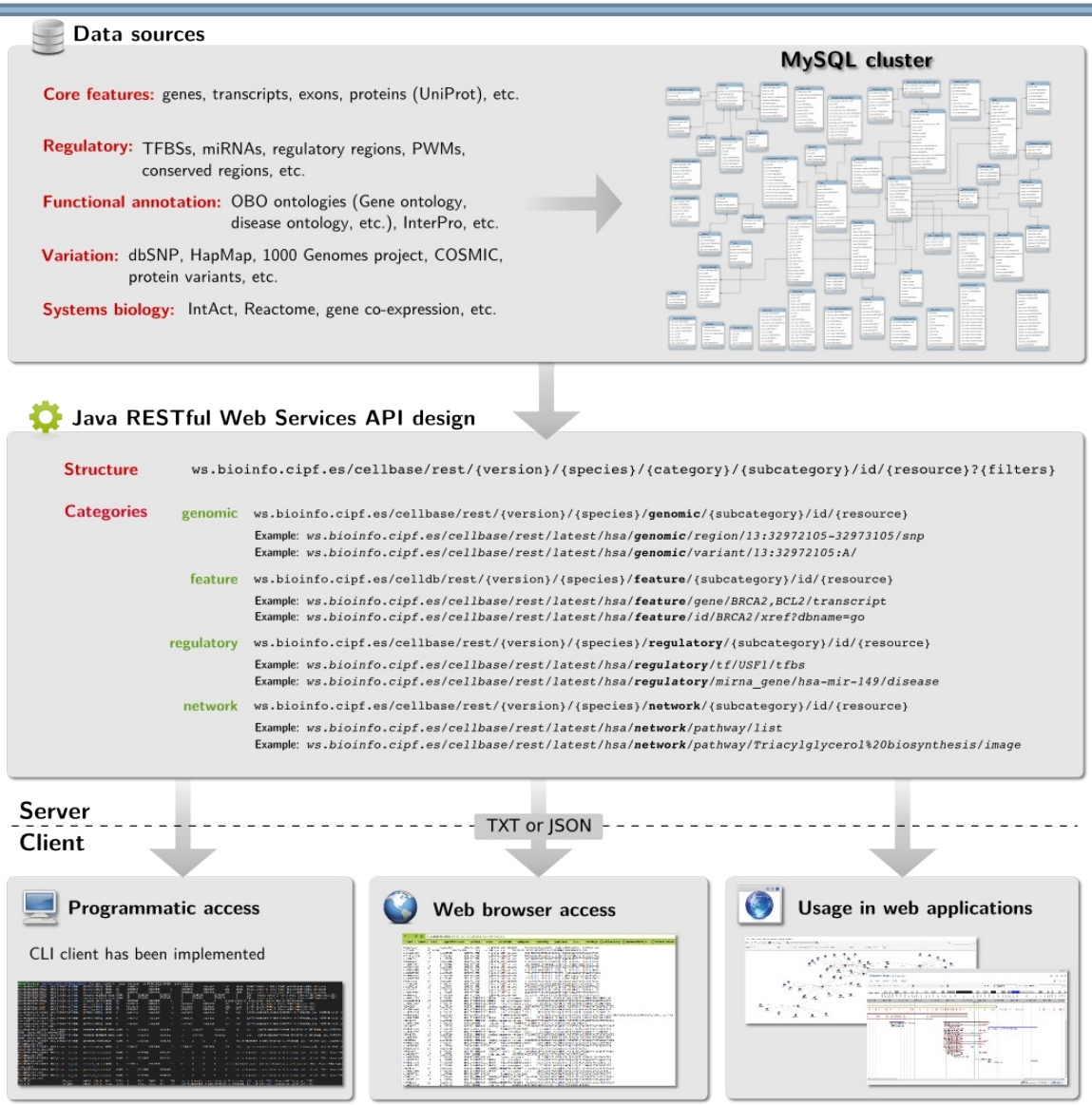
- Exponential growth in the number and size of biological databases and repositories. Data size can reach hundreds of gigabytes and involves serious problems of data access through Internet and local disks.
- Biological information is spread out in different databases and repositories (~1512), using 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



Ignacio Medina
imedina@cipf.es

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
- Network
  Subcategories: *pathway*
  http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/Triacylglycerol%20biosynthesis/info

Ignacio Medina
imedina@cipf.es

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[1,2], Joaquin Tarraga[1,3], Alejandro de Maria[1], Francisco Salavert[1,2], Luz Garcia-Alonso[1], Matilde Celma[4], Ainoha Martin[4], Joaquin Dopazo[1,2,3,*] and Ignacio Medina[1,3,*]

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	Info, tfbs, gene, target_gene, pwm, annotation
	mirna_gene	miRBase gene ID	Info, target, disease
	mirna_mature	miRBase gene Accession	Info, gene, mirna_gene, target_gene, target, disease, annotation

Marta Bleda
mbleda@cipf.es

Accessing CellBase information

- Web services

  `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`**

Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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**.

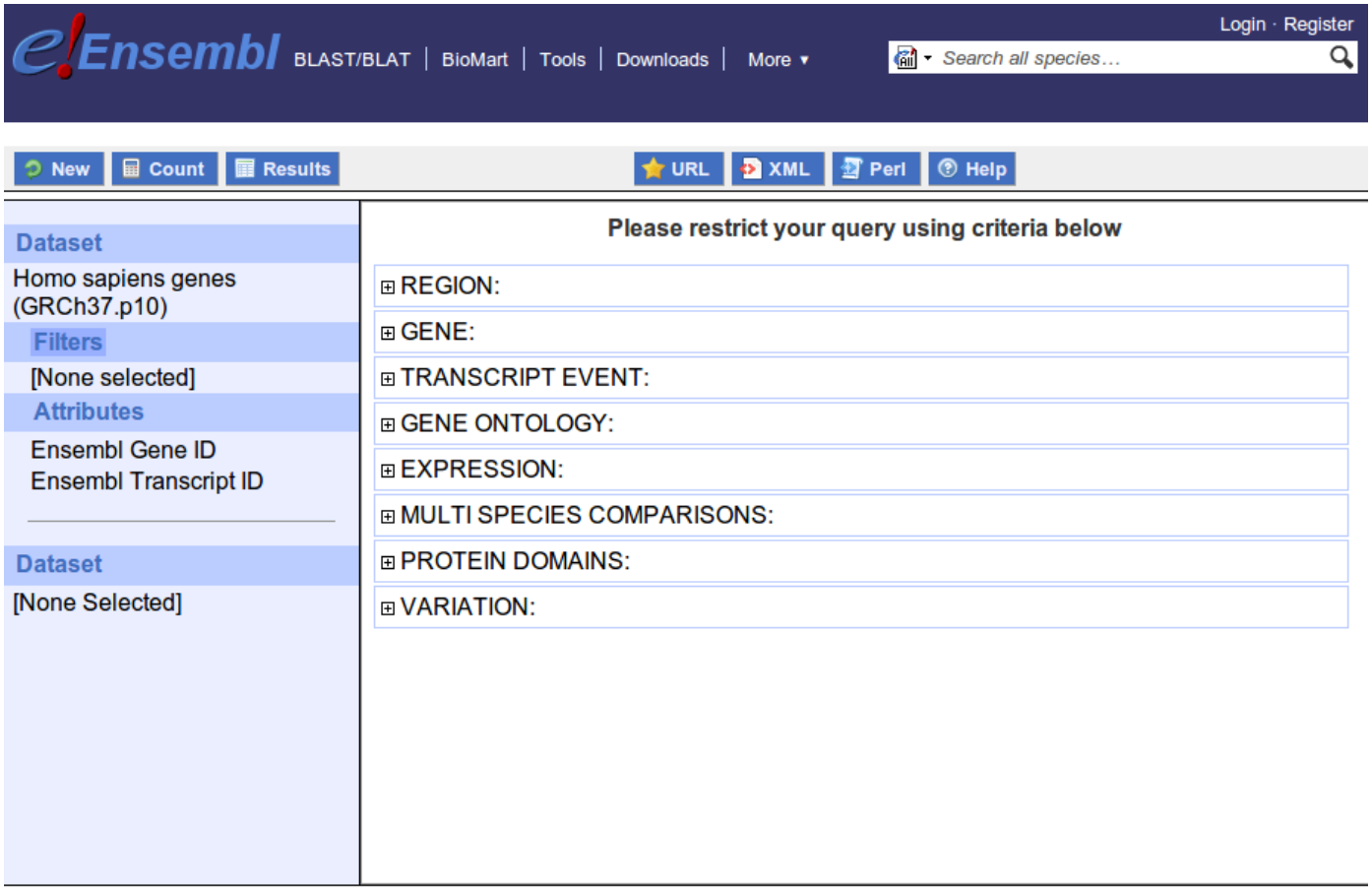
HELP:

- Version: `latest`
- Species: `hsa`
- Category: `regulatory`
- Subcategory: `mirna_mature`
- ID: `hsa-miR-149-3p`
- Resource: `disease`

Marta Bleda
mbleda@cipf.es

Accessing Ensembl BioMart information

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



Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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`

Marta Bleda
mbleda@cipf.es

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

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

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

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

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

Marta Bleda
mbleda@cipf.es