# Babelomics

## Introduction to Functional Annotation Databases

Valencia, March 2011

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

# Functional Annotation DDBB

## Introduction

- Last years has been an exponential increase in the number of annotation 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/**

# Functional Annotation DDBB

## Functional Databases

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* | *Saccharomyces 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

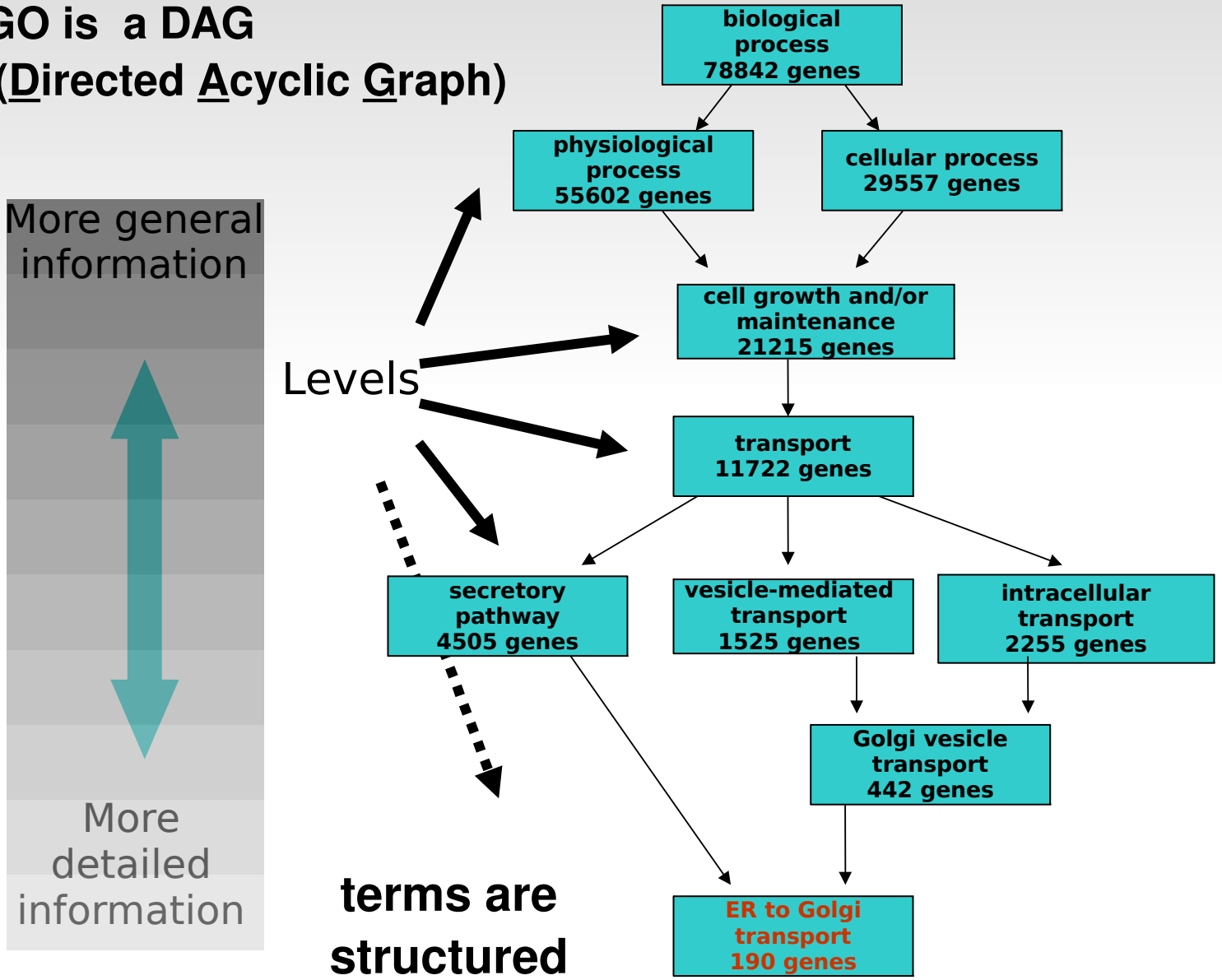
#### Cellular Component

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

# Functional Annotation DDBB

## Gene Ontology (GO terms)

**GO is a DAG**
**(Directed Acyclic Graph)**




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

## InterPro

- 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 |

# 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

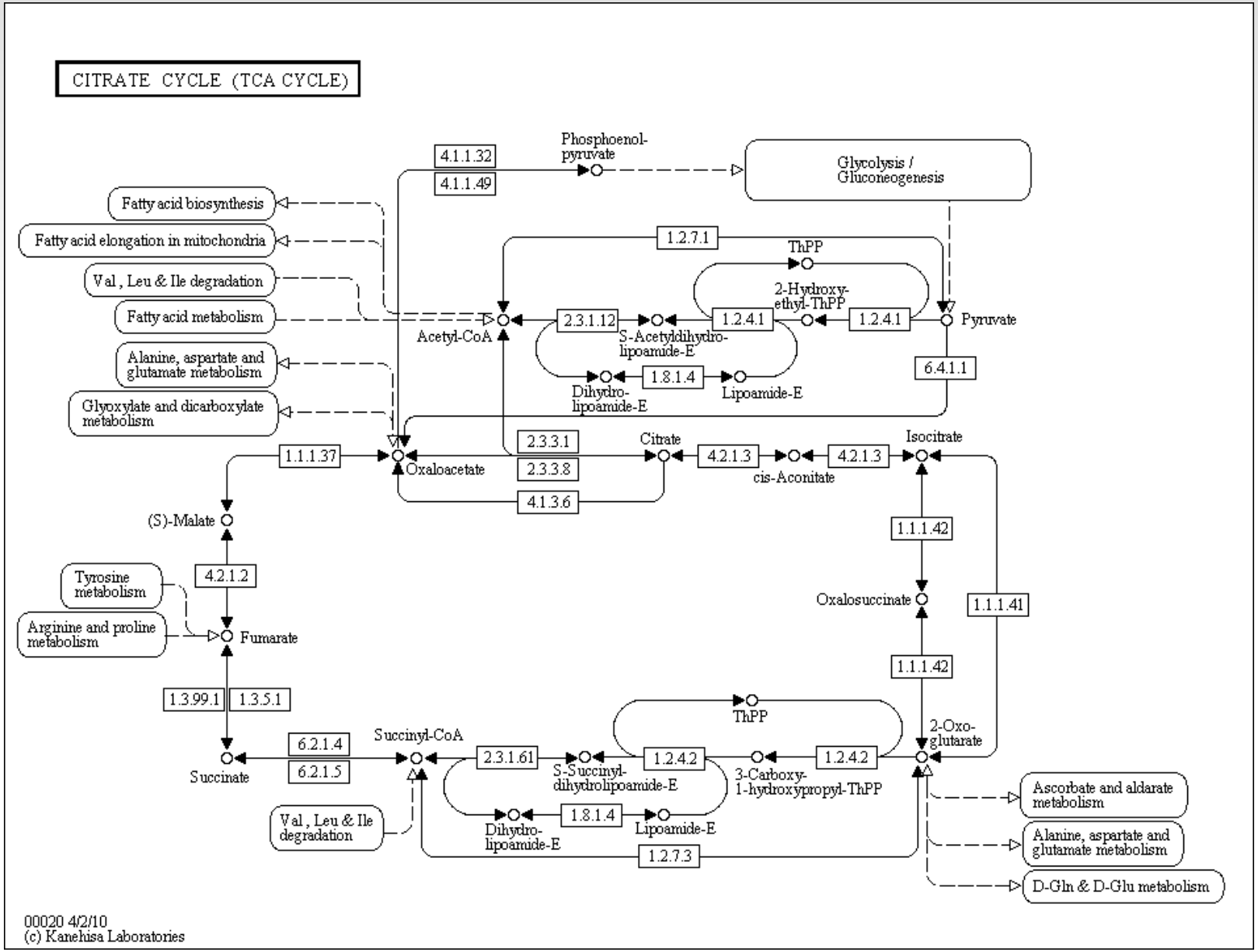
**KEGG pathways**


| **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



# Functional Annotation DDBB

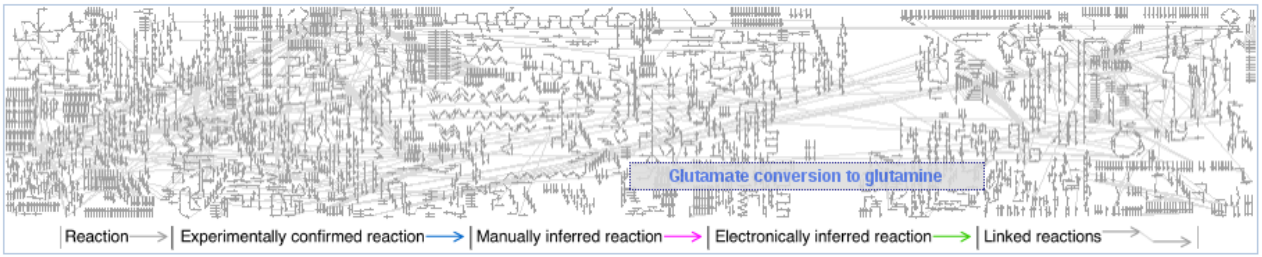
## Reactome

- 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**

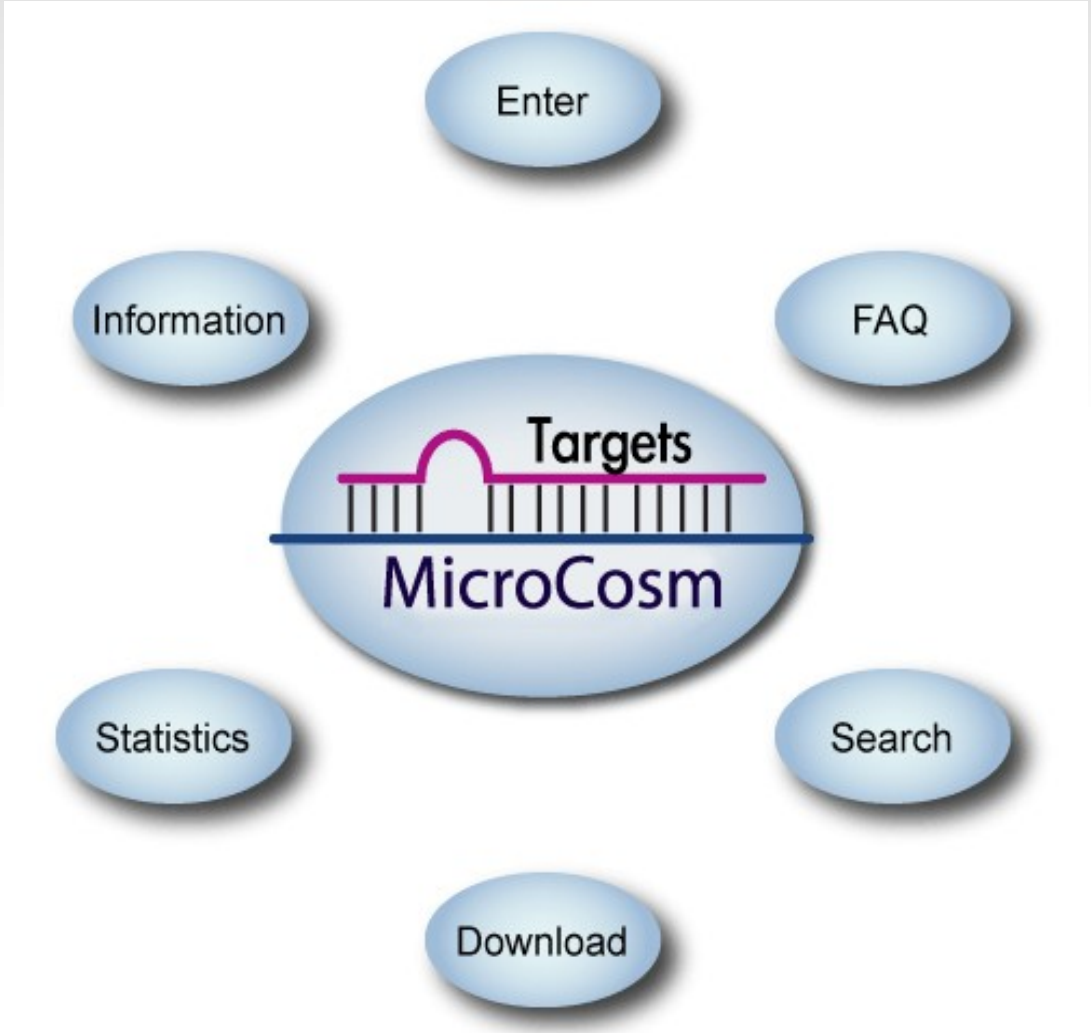
The data displayed is for Homo sapiens. Use the menu to change the species. Check for cross-species comparison.




| | | | |
|---|---|---|---|
| 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://microrna.sanger.ac.uk/**

# 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.

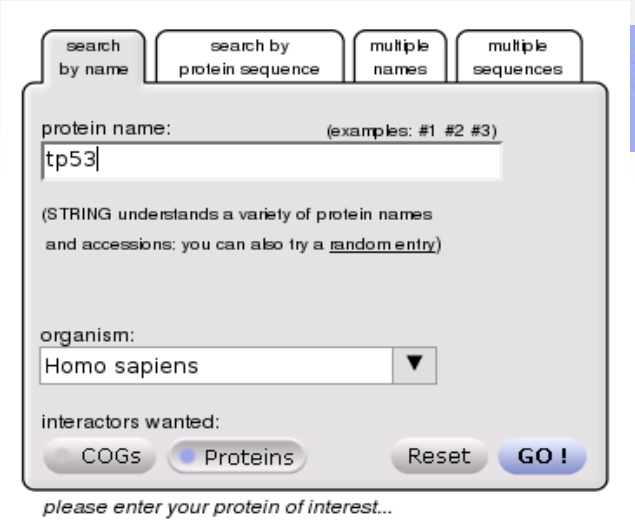
http://www.oreganno.org/oregano/

# Functional Annotation DDBB

## String

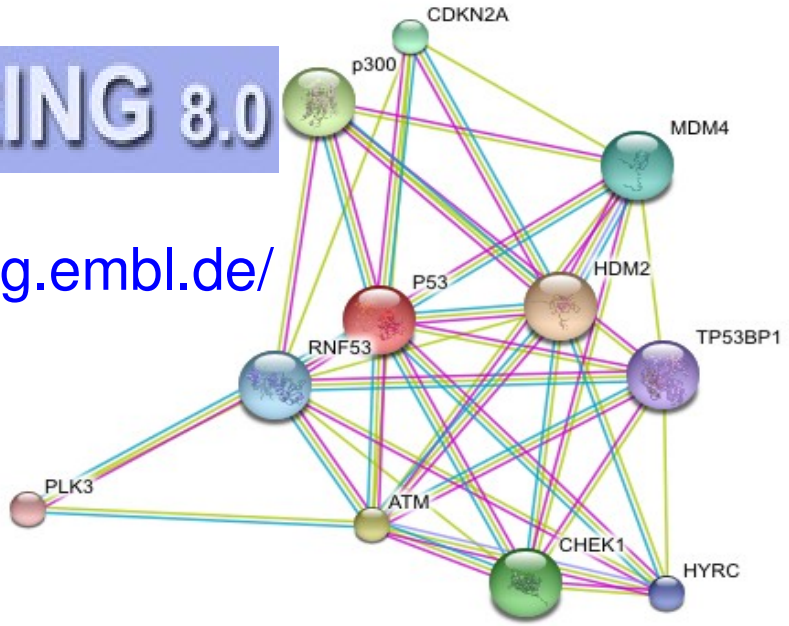
Database for known and predicted protein-protein interactions (direct and indirect associations)
Cover four sources of annotations: Genomic association (prokaryotes), high-throughput experiments (e.g. y2h), conserved co-expression, previous knowledge (text-mining).

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




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 |

# Functional Annotation DDBB

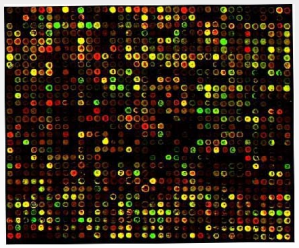
## Practical exercise

- About *BCL2, BRCA1, ATM* and *P53*(*TP53*):
  - try to find the biological process and cellular components (*GO terms*), do they share some GO terms? Is that significant?
  - Are they target of the same microRNA (*mirbase*)?
  - What about protein functional domains (*interpro*)?
  - Are they regulated by the same conserved motifs (*ORegAnno*)?
  - Are they involved in a common disease or pathway (*kegg, reactome*)?
  - ...

# Functional Annotation DDBB

## From GEPAS to Babelomics

Microarray Data



Next generation of High throughput Sequencing

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

Data Analysis

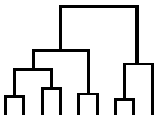

Differential expression

Genes differentially expressed

Predictors

Predicting genes

List of genes (*ie*: 120 genes)



Clustering

Genes with same expression patterns


CGH array

Genes from a deleted or duplicated region

Functional Profiling

# Functional Annotation DDBB

## Babelomics try to answer these questions

- Is there any significant functional enrichment in my gene list?
- Are these genes involved in the same pathways?
- Are they sharing a specific microRNA regulation?
- Are they involved in the same disease?
- ...