

# Data Analysis. Functional Profiling: FatiGO

**Joaquín Tárraga Giménez**

[jtarraga@cipf.es](mailto:jtarraga@cipf.es)

*Valencia, September 2011*

*Bioinformatics and Genomics Department  
Centro de Investigación Príncipe Felipe (CIPF)  
(Valencia, Spain)*



# Pre-genomic era

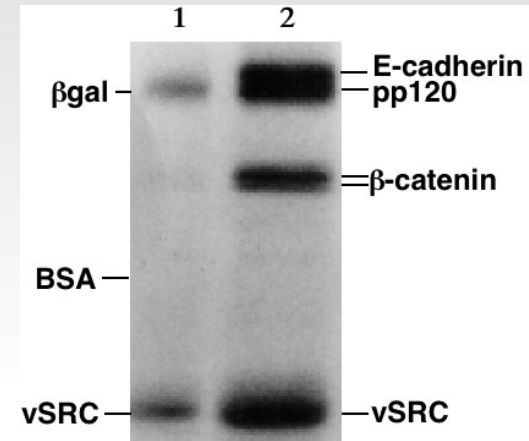
Extract as much information as possible for one single data

Easy data analysis and biological interpretation of the results

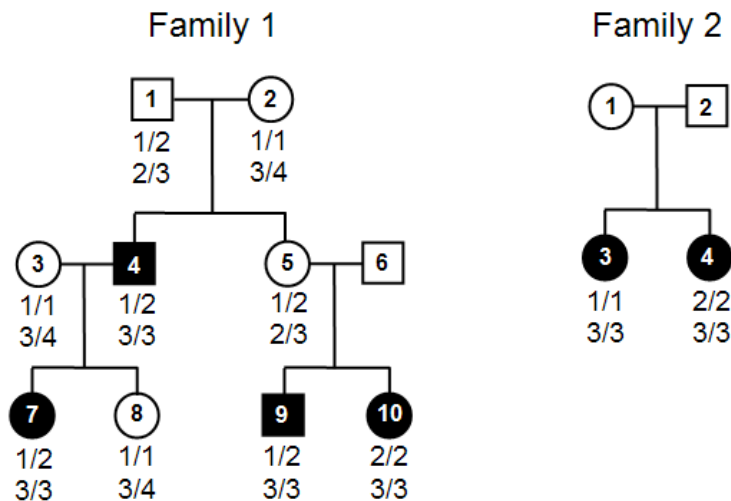
Western blot

Southern blot

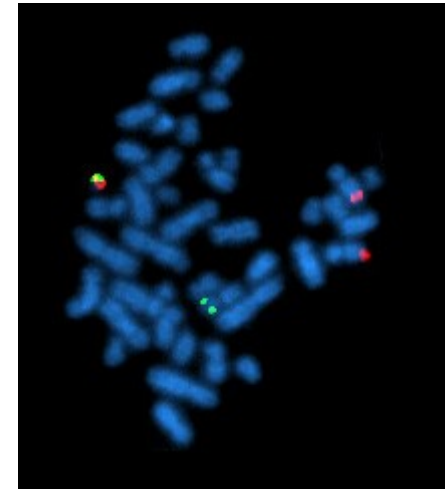
Northern blot



## Linkage studies



FISH



# Post-genomic era

## Huge amount and different type of data

In 2000 high-throughput technologies appeared



DNA sequencer: *mRNA-seq*, *Chip-seq*, *resequencing*, ...

2063 Microbial Genomes selected:

Complete - 774, Assembly - 596, Unfinished - 693

501 Eukaryotic Genome Sequencing Projects Selected:

Complete - 23, Assembly - 243, In Progress - 235

*Illumina HiSeq 2000* is the first commercially available sequencer to enable researchers to obtain ~30x coverage of two human genomes in a single run for under \$10,000



HiSeq 2000 Preliminary Performance Parameters\*

Read Length	Run Time	Output
1 x 35 bp	~1.5 days	26-35 Gb
2 x 50 bp	~4 days	75-100 Gb
2 x 100 bp	~8 days	150-200 Gb

\*Sequencing output generated with a PhIX library and cluster densities between 260,000-347,000 clusters/mm<sup>2</sup> that pass filtering on a HiSeq 2000.

**Throughput**

Up to 25 Gb per day for a 2 x 100 bp run.

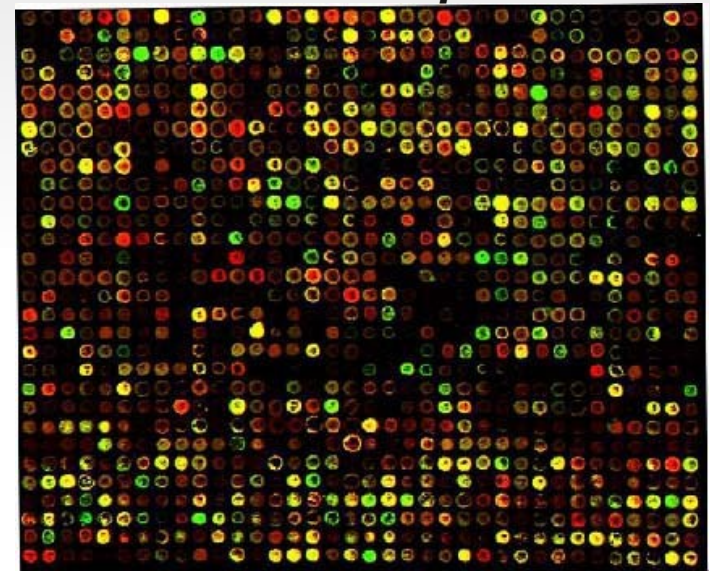
**Reads**

Up to one billion clusters passing filter, and up to two billion paired-end reads.

**Service and Support**

Illumina will ensure that your HiSeq 2000 is properly installed and qualified, and will provide ongoing maintenance and service. This industry-leading support is available in North America, Europe, and Asia.

DNA microarray: *expression*, *genotyping*, *methylation*, *exons*, *ChIP-on-chip*, ...



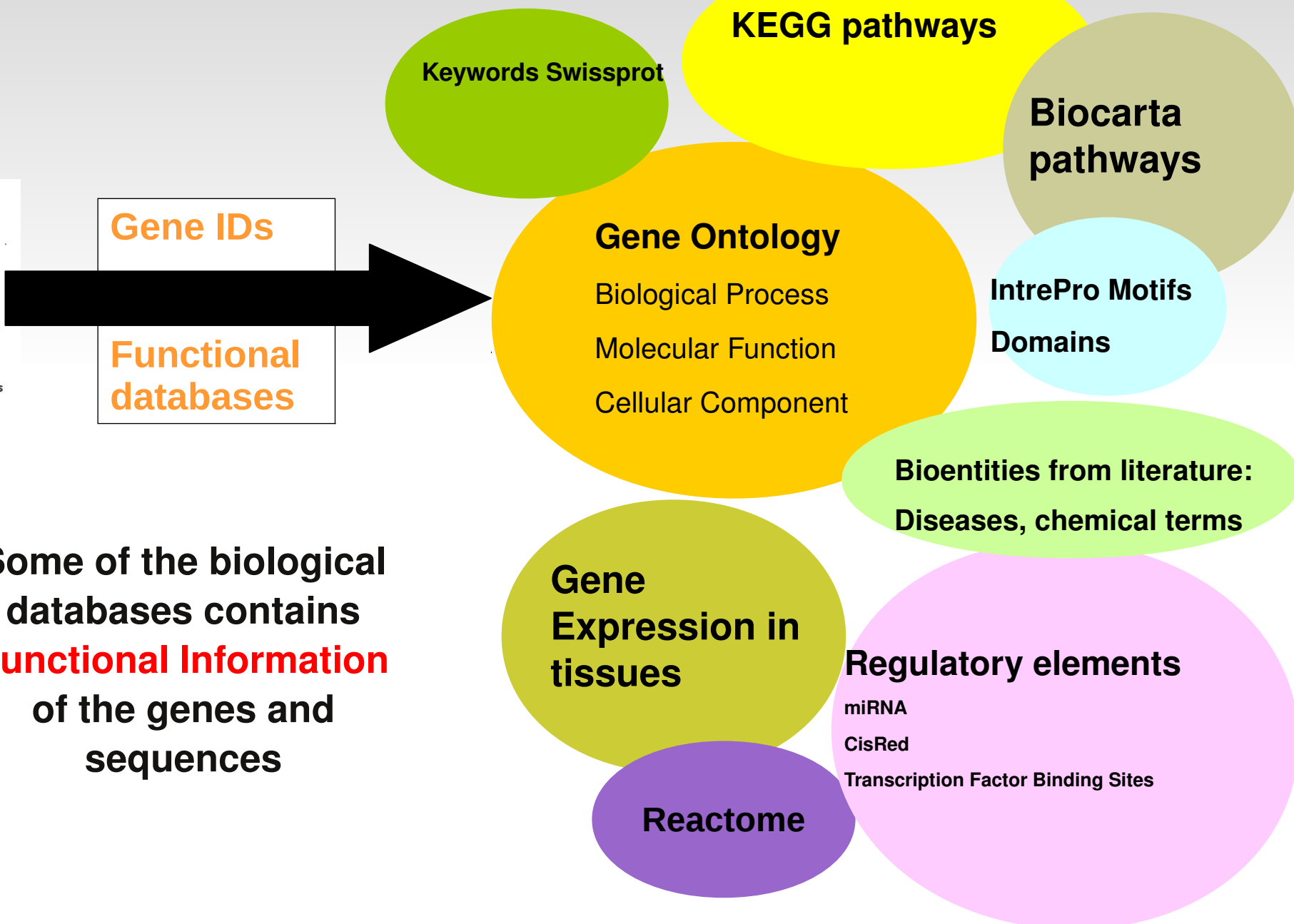
Gene Expression Omnibus

Public data

GPL Platforms	5256
GSM Samples	262357
GSE Series	10220
<b>Total</b>	<b>277833</b>

# Functional Annotation DDBB

-  **Aedes aegypti**  
home page | site map
-  **Anopheles gambiae**  
home page | site map
-  **Bos taurus**  
home page | site map
-  **Caenorhabditis elegans**  
home page | site map
-  **Canis familiaris**  
home page | site map
-  **Cavia porcellus**  
home page | site map
-  **Ciona intestinalis**  
home page | site map
-  **Ciona savignyi**  
home page | site map
-  **Danio rerio**  
home page | site map
-  **Dasypus novemcinctus**  
home page | site map
-  **Drosophila melanogaster**  
home page | site map
-  **Microcebus murinus**  
home page | site map
-  **Monodelphis domestica**  
home page | site map
-  **Mus musculus**  
home page | site map
-  **Myotis lucifugus**  
home page | site map
-  **Ochotona princeps**  
home page | site map
-  **Ornithorhynchus anatinus**  
home page | site map
-  **Oryctolagus cuniculus**  
home page | site map
-  **Oryzias latipes**  
home page | site map
-  **Otolemur garnettii**  
home page | site map
-  **Pan troglodytes**  
home page | site map
-  **Pongo pygmaeus**  
home page | site map
-  **Echinops telfairi**  
home page | site map
-  **Equus caballus**  
home page | site map
-  **Erinaceus europaeus**  
home page | site map
-  **Felis catus**  
home page | site map
-  **Gallus gallus**  
home page | site map
-  **Gasterosteus aculeatus**  
home page | site map
-  **Homo sapiens**  
home page | site map
-  **Loxodonta africana**  
home page | site map
-  **Macaca mulatta**  
home page | site map



**Gene IDs**  
**Functional databases**

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

**Gene Ontology**  
Biological Process  
Molecular Function  
Cellular Component

**Keywords Swissprot**

**KEGG pathways**

**Biocarta pathways**

**IntrePro Motifs Domains**

**Bioentities from literature: Diseases, chemical terms**

**Gene Expression in tissues**

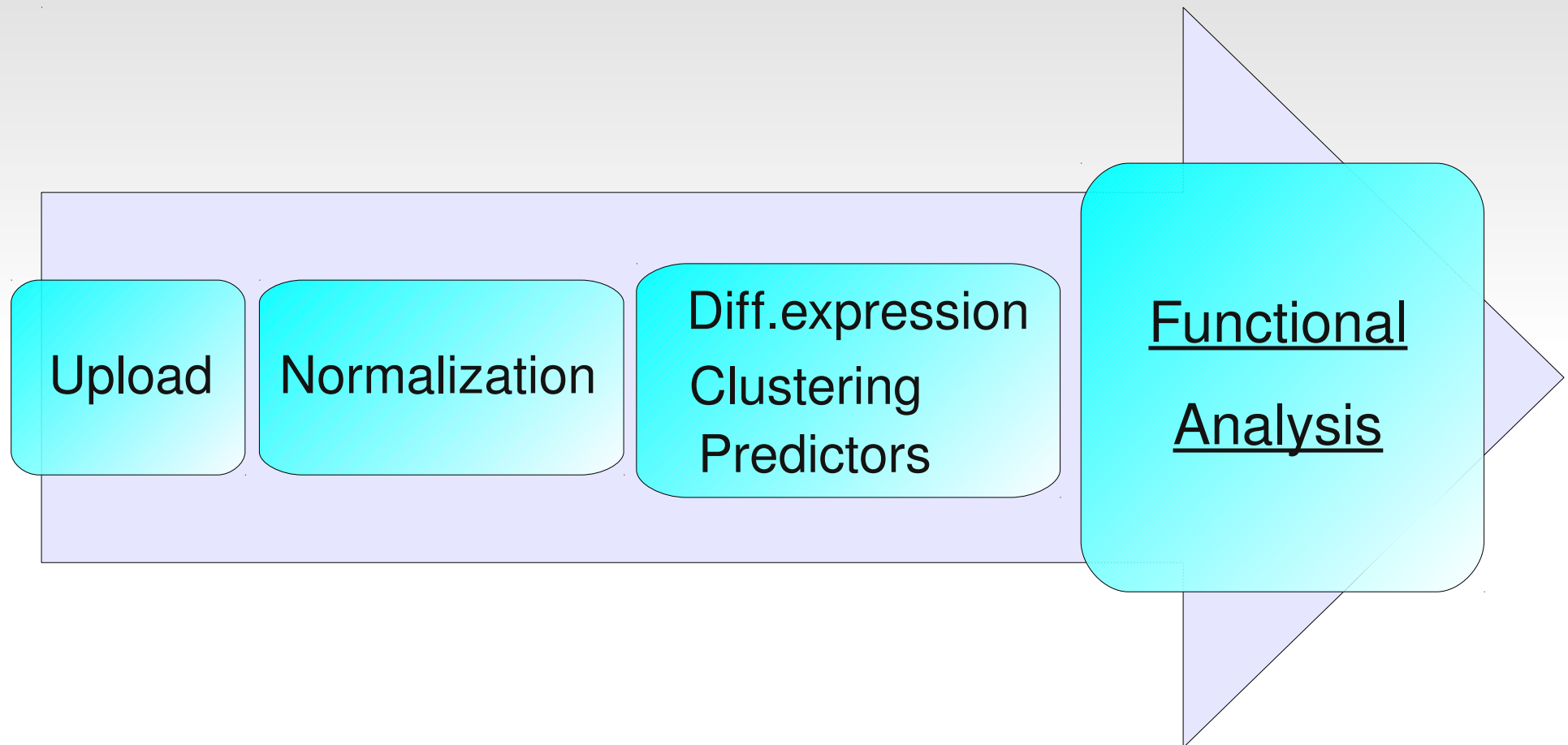
**Regulatory elements**  
miRNA  
CisRed  
Transcription Factor Binding Sites

**Reactome**

# Gene Ontology (GO)

- The objective of GO is to provide controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products.
- These terms are to be used as attributes of gene products by collaborating databases, facilitating uniform queries across them.
- The controlled vocabularies of terms are structured

# Data analysis workflow





# Functional profiling of genome-scale experiments

My data...

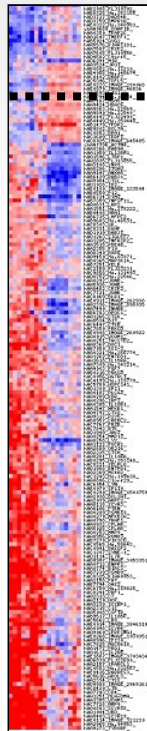
How are structured?

What are these groups?

What is this gen?

	E	F	G	H	I	J	K	L	M	N
65	578.6	*		1.4	0.26	M12481	Mouse cytoplasmic beta-actin mRNA (5_M_3 repress			
66	534.9	*		-1.6	0.22	M12481	Mouse cytoplasmic beta-actin mRNA (5_M_3 repress			
67	403.6	*		-1.5	0.15	Y61388	SGD: YEL002C Yeast S.cerevisiae WBP1 Oligosaccharyl			
68	635.2	*		-1.6	0.22	U18530	SGD: YEL018W Yeast S.cerevisiae Protein of unknown fu			
69	-567.7	*		-1.6	-0.27	M23316	SGD: YEL024C Yeast S.cerevisiae RIP1 Rieske iron-sulfur			
70	-114.5	*		-1.1	-0.03	K02207	SGD: YEL021W Yeast S.cerevisiae URA3 gene coding for			
71	-125.4	*		-1	-0.01	Cluster Incl	M16465: Calpactin I light chain /cds=(68,361) /gb=M16			
72	-1091.6	*		-1.2	-0.14	Cluster Incl	Z67748: M.musculus spermidine synthase gene /cds=(			
73	-757.2	*		-1.3	-0.17	Cluster Incl	X12973: Murine MLC1F/MLC3F gene for myosin alkali I			
74	9826.6	*		1.3	0.83	Cluster Incl	AB49035: UH-M-AH1-agw-a-06-U-01.s1 Mus musculus c			
75	-847.4	*		-1.3	-0.21	Cluster Incl	AW123542: UH-M-BH2.1-ap-f01-0-U-01.s1 Mus musculus			
76	2583.1	*		1.1	0.09	Cluster Incl	AF059832: Mus musculus proteasome alpha7/CS8 subu			
77	192.5	*		-1.2	0.05	Cluster Incl	AB006361: Mus musculus mRNA for prostaglandin D s			
78	2980.2	*		-4.4	1.63	Cluster Incl	AB006361: Mus musculus mRNA for prostaglandin D s			
79	-20.1	*		-1	0	Cluster Incl	AB011081: Mus musculus mRNA for huntingtin interact			
80	1880.9	*		-2.6	1.81	Cluster Incl	AB011081: Mus musculus mRNA for huntingtin interact			
81	753.2	*		1.2	0.1	Cluster Incl	U97170: Mus musculus protein kinase inhibitor gamma			
82	-2774.7	*		-1.9	-1.43	Cluster Incl	M36120: Keratin complex 1, acidic, gene 19 /cds=(0,12			
83	3614.4	*		-5.1	1.98	Cluster Incl	U19604: DNA ligase 1, ATP-dependent /cds=(304,3054)			
84	0	*		-0.0	0	Cluster Incl	AB51492: UH-M-BHJ-aju-4-04-U-01.s2 Mus musculus cD			
85	3310.9	*		1.2	0.24	Cluster Incl	AB025408: Mus musculus mRNA for sid47bp, complet			
86	-1291	*		-1.5	-0.42	Cluster Incl	AF059736: Mus musculus C-terminal binding protein 2			
87	-263.3	*		-1.3	-0.09	Cluster Incl	AF053454: Mus musculus tetraspan TMSF (Tspan-6)			
88	77.5	*		1.1	0.01	Cluster Incl	D45850: Hydroxysteroid 17-beta dehydrogenase 1 /cds			
89	2047.2	*		-3.3	1.1	Cluster Incl	AF039299: Mus musculus 17-beta-hydroxysteroid dehy			
90	809.9	*		-1.9	0.38	Cluster Incl	M84487: Vascular cell adhesion molecule 1 /cds=(67,2			
91	-124.3	*		-1.1	-0.03	Cluster Incl	U12884: Mus musculus C57BL/6 vascular cell adhesio			
92	-675.5	*		-1.8	-0.37	Cluster Incl	U12884: Mus musculus C57BL/6 vascular cell adhesio			
93	1465.4	*		-2.7	0.76	Cluster Incl	AJ238636: Mus musculus mRNA for nucleoside diphos			
94	638.2	*		1.1	0.1	Cluster Incl	U70475: Nuclear, factor, erythroid derived 2, like 2 /cds			
95	4969.4	*		-6.7	8.84	Cluster Incl	AF045573: Mus musculus FLLRR associated protein-			
96	148.3	*		-1.2	0.04	Cluster Incl	AB91475: u69a06.x1 Mus musculus cDNA, 3 end /cd			

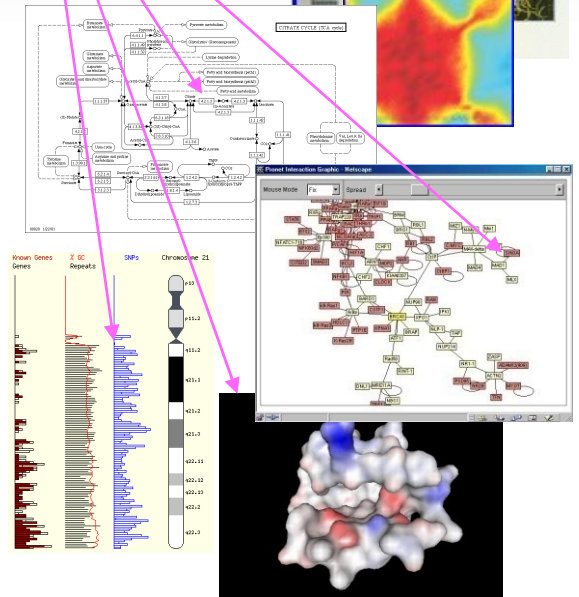
A B



Cell cycle...



I19380: Calmodulin 3 /cds=(109,558) /gb=M19380 /gi=469419  
 I4842326: UH-M-AM1-afz-b-11-0-U-01.s1 Mus musculus cDNA, 3  
 AJ242663: Mus musculus mRNA for cathepsin Z precursor (ctz  
 U12620: Diacylglycerol phosphatase 4 /cds=(117,2399) /gb=U12620 /g  
 M13444: Mouse alpha-tubulin isotype M-alpha-4 mRNA, compl  
 U11027: Mus musculus C57BL/6J Sec61 protein complex gam  
 IJ03928: Phosphofruktokinase, liver, B-type /cds=(42,2384) /gb=  
 I267745: M.musculus mRNA for phosphatase 2A catalytic subu  
 I18932: Serine/threonine kinase 6 /cds=(48,1236) /gb=U80932  
 IU47024: Maternal embryonic message 3 /cds=(137,2401) /gb=  
 AFD75136: Mus musculus Sn3-associated protein (sap3D) mR  
 M29544: Mouse carbonic anhydrase II (CAII) mRNA, 3 end /cd  
 IX74671: Neurofibromatosis 2 /cds=(576,2366) /gb=IX74671 /gi=  
 M12848: Mouse myb proto-oncogene mRNA encoding 71 kd m  
 AW125458: UH-M-BH2.2-agw-a-07-0-U-01.s1 Mus musculus cDN  
 I184903: Ribosomal protein L23 /cds=(81,501) /gb=I184903 /gi=  
 U35141: Mus musculus retinoblastoma-binding protein (mRbAp  
 I19521: Mus musculus vesicle transport protein (munc-18c) ml  
 M15268: Aminolevulinic acid synthase 2, erythroid /cds=(0,178)  
 M25149: Tissue specific transplantation antigen P91A /cds=(0,  
 I186449: Calcyclin /cds=(159,428) /gb=I186449 /gi=50271 /gi=



Analysis

Functional profiling

Links





# FatiGO

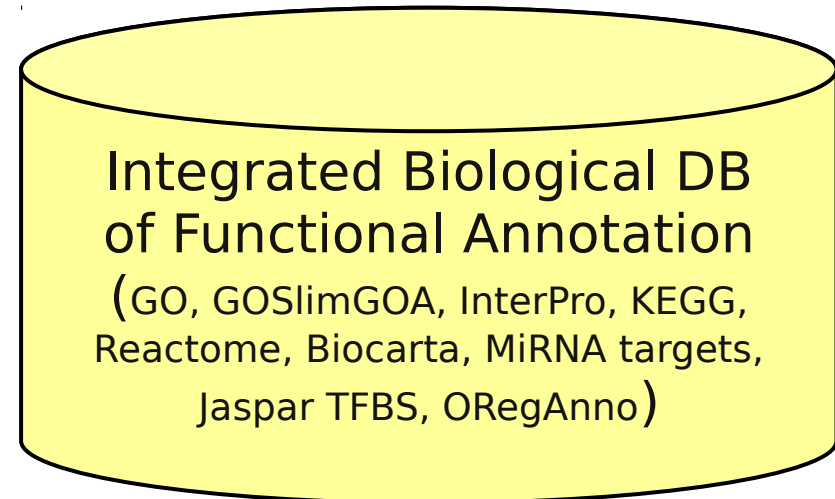
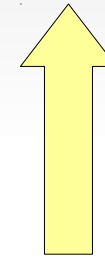
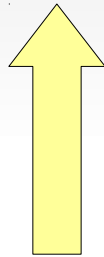
## *Questions that Functional enrichment analysis try to answer*

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

# Fatigo

## Schema

FatiGO is a web tool for:  
*statistical test, multiple test corrections, filtering ...*



# FatiGO

- A web-based tool for the functional profiling of genome-scale experiments

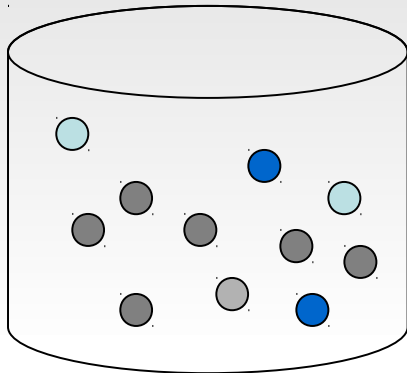
The screenshot displays the Babelomics 4 web interface. At the top, the logo 'BABELOMICS 4' is shown with the subtitle 'gene expression and functional profiling analysis suite'. Below the logo is a navigation bar with several menu items: 'Upload data', 'Processing data', 'Expression', 'Genomic', 'Functional analysis', and 'Utilities'. The 'Functional analysis' menu item is circled in red. Below the navigation bar, a status bar indicates the user 'mmarba@cipf.es' is working on a project named 'Pre-processing Agilent' with 91.30 Mb of 1.00 Gb (8.92%) used and no active sessions. A green message box contains the text: 'Welcome to the new Babelomics 4, you can still use Babelomics 3 at: <http://babelomics3>'. Below this, the 'Functional analysis' section is titled, and under 'Single enrichment analysis', the 'FatiGO' tool is listed and circled in red. A description for FatiGO reads: 'Resource to show significant over-representation of GO terms.' Below it, the 'Marmite' tool is listed with the description: 'Extracts blocks of related genes from an ordered list of genes by an associated value to the Marmite tool'.

# FatiGO features

- It allows us to compare functional annotation of:
  - **Two** list of genes
  - **One** list against the rest of genome
  - Lists of genes with user **submitted annotations**
- One statistical test for each Functional **Block** of annotation
  - Fisher's exact test
  - Multiple testing context (hundreds of annotation)
  - Filtering of annotation is convenient (the less tests the best correction)

# FatiGO test

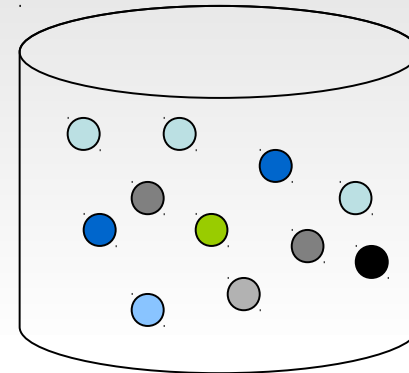
One Gene List (A)



Biosynthesis 60% ●

Are this two groups of genes carrying out different biological roles?

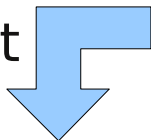
The other list (B)



Biosynthesis 20% ●



Fisher's exact test



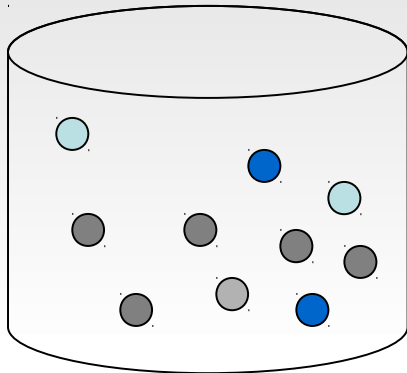
	A	B
Biosynthesis	<b>60</b>	<b>20</b>
No biosynthesis	<b>40</b>	<b>80</b>

p-value = 0.00000431

Genes in group A have significantly to do with biosynthesis

# FatiGO test

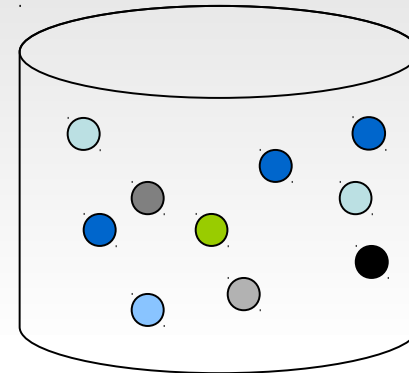
One Gene List (A)



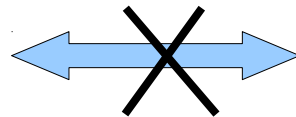
Sporulation 20% ●

Are this two groups of genes carrying out different biological roles?

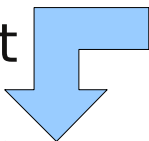
The other list (B)



Sporulation 30% ●



Fisher's exact test



	A	B
Sporulation	<b>20</b>	<b>30</b>
No sporulation	<b>80</b>	<b>70</b>

p-value = 0.964

Genes in group A DO NOT have significantly to do with sporulation.



# FatiGO form

## Define your comparison

- Id list vs Id list
- Id List vs Rest of genome
- Id List vs Rest of ids contained in your annotations (complementary list)

## Select your data

List 1 :

no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

List 2 :

no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

## Options

Fisher exact test

Remove duplicates?

## Databases

Organism

GO biological process [\[options\]](#)

# FatiGO form

Do you want to compare 2 conditions or one vs the rest of genome ?

The screenshot shows the FatiGO form with the following sections:

- Define your comparison:** Three radio button options:  Id list vs Id list,  Id List vs Rest of genome, and  Id List vs Rest of ids contained in your annotations (c...
- Select your data:** Two identical sections for List 1 and List 2. Each contains a "browse server" button, the text "no data selected.", and a link "Or go to Upload Data form: Upload [idlist]".
- Options:** Two dropdown menus: "Fisher exact test" set to "Two tailed" and "Remove duplicates?" set to "Never".
- Databases:** An "Organism" dropdown menu set to "Select an organism" and a checkbox for "GO biological process" with a link to "options".

eg. Compare 2 tissues or responder genes vs. non-responders

eg: genes that respond to one treatment against the genome

# FatiGO form

## Define your comparison

- Id list vs Id list
- Id List vs Rest of genome
- Id List vs Rest of ids contained in your annotations (c

Upload first at Data Upload

## Select your data

List 1 :  no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

List 2 :  no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

“txt” file with  
gene lists:

```
gene1  
gene2  
gene3  
...
```

Data  
selection

## Options

Fisher exact test

Remove duplicates?

## Databases

Organism

GO biological process [\[options\]](#)

# FatiGO form

## Define your comparison

- Id list vs Id list
- Id List vs Rest of genome
- Id List vs Rest of ids contained in your annotations (c

## Select your data

List 1 :

no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

List 2 :

no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

## Options

Fisher exact test

Remove duplicates?

## Databases

Organism

GO biological process [\[options\]](#)

### HINT:

- *Two tailed* for 2 lists
- *One tailed* for 1list vs rest of genome (or your annotations)

Algorithm options

### Removing duplicates:

- Choose one or other option depends on from where gene lists come from.

# FatiGO form

## Define your comparison

- Id list vs Id list
- Id List vs Rest of genome
- Id List vs Rest of ids contained in your annotations (c

## Select your data

List 1 :  no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

List 2 :  no data selected.  
Or go to Upload Data form: [Upload \[idlist\]](#)

## Options

Fisher exact test

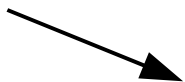
Remove duplicates?

## Databases

Organism

GO biological process [\[options\]](#)

Which type of functional information?



# FatiGO form

Which type of functional information?

**Databases**

Organism

- GO biological process [\[options\]](#)
- GO molecular function [\[options\]](#)
- GO cellular component [\[options\]](#)
- GOSlim GOA [\[options\]](#)
- Interpro [\[options\]](#)
- KEGG pathways [\[options\]](#)
- Reactome [\[options\]](#)
- Biocarta [\[options\]](#)
- miRNA targets [\[options\]](#)
- Jaspar TFBS [\[options\]](#)
- ORegAnno [\[options\]](#)

Your annotations  no data selected.  
Or go to Upload Data form: [Upload \[annotation\]](#)

**Use one or more of the given databases**

**If it is not in the databases, use your annotations option.**



# FatiGO form

Databases

Organism: Human (homo sapiens)

GO biological process [options]

GO molecular function [options]

GO cellular component [options]

**First select an organism**

## OPTIONS:

**Test all the GO or only annotated terms**

**Discard functions with too few or too many genes?**

**If you have an hypothesis, better test this first!!!!!!**

GO biological process options

GO parameters

▶ Select annotation through ontology levels

Propagate annotation to upper levels

Direct annotation

GO level must be among levels  and

Filter terms by number of annotated ids in DB

Minimum  (typically 5-20)

Maximum  (typically 500-Inf)

▶ Number of annotated ids is computed from

Genome

Your input ids

Filter terms by keywords

Keywords  (e.g. metabolism cancer)

▶ Your search must match

all keywords

any keyword

Add children of selected terms

# FatiGO form

Which type of functional information?

**Databases**

Organism

- GO biological process [\[options\]](#)
- GO molecular function [\[options\]](#)
- GO cellular component [\[options\]](#)
- GOSlim GOA [\[options\]](#)
- Interpro [\[options\]](#)
- KEGG pathways [\[options\]](#)
- Reactome [\[options\]](#)
- Biocarta [\[options\]](#)
- miRNA targets [\[options\]](#)
- Jaspar TFBS [\[options\]](#)
- ORegAnno [\[options\]](#)
- Your annotations  no data selected.  
Or go to Upload Data form: [Upload \[annotation\]](#)

**Job**

job name:

job description:

**Your annotations: useful when you work with your own annotations OR with an organism that is not in Babelomics**

**Upload first at Data Upload**

Example (your annotations):

38969_at	GO:0003677
37639_at	GO:0006306
37149_s_at	GO:0004674
37149_s_at	GO:0005525
37639_at	GO:0006306
37149_s_at	GO:0004674
...	...

# FatiGO form

**Databases**

Organism

- GO biological process [\[options\]](#)
- GO molecular function [\[options\]](#)
- GO cellular component [\[options\]](#)
- GOSlim GOA [\[options\]](#)
- Interpro [\[options\]](#)
- KEGG pathways [\[options\]](#)
- Reactome [\[options\]](#)
- Biocarta [\[options\]](#)
- miRNA targets [\[options\]](#)
- Jaspas TFBS [\[options\]](#)
- ORegAnno [\[options\]](#)

Your annotations  no data selected.  
Or go to Upload Data form: [Upload \[annoti](#)

What's  
your  
job name?

**Job**

job name:

job description:

Set up a job name  
and optionally,  
give a description.

# FatiGO

Running an example ...

# FatiGO

## Example:

Description

Molecular Apocrine Breast Cancer dataset:

- 49 Affymetrix (HG-U133A), 14,500 genes
- Human
- 3 tumor classes:  
apocrine, basal and luminal.

□

# FatiGO input

- 1) Run *differential expression* (using ANOVA test) because we are comparing 3 conditions
- 2) Send results to FatiGO:



- 3) Input FatiGO form:
  - Comparison: list vs rest of genome
  - Test: one tailed test



# FatiGO results

## Summary results:

■ Id annotations per DB :

<i>DB</i>	<i>List1</i>	<i>Genome</i>
GO biological process (levels from 3 to 9)	350 of 500 (70%) 11.26 annotations/id	11716 of 23198 (2343.2%) 5.08 annotations/id
GO molecular function (levels from 3 to 9)	344 of 500 (68.8%) 3.05 annotations/id	11370 of 23198 (2274%) 1.92 annotations/id

## Tables significant terms:

### ▼ Significant Results

■ Number of significant terms per DB :

<i>DB</i>	<i>Number of significant terms</i>
GO biological process (levels from 3 to 9)	<b>142</b>
GO molecular function (levels from 3 to 9)	<b>30</b>

# FatiGO results

## Significant results:

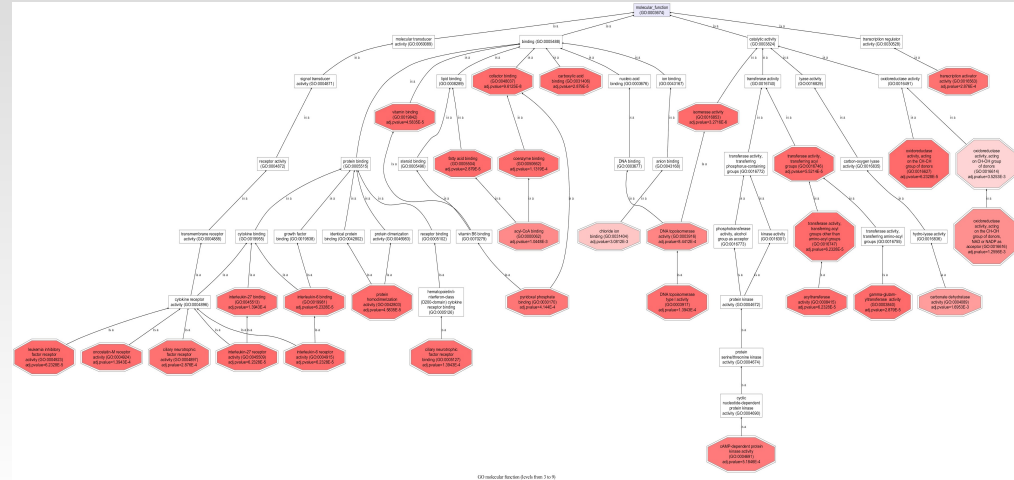
Term	Term size	Term size (in genome)	Term annotation % per list	Annotated ids	Odds ratio (log e)	pvalue	Adjusted pvalue ▲
negative regulation of apoptosis (GO:0043066)	412	403	list 1:  7.2% list 2:  1.62%	list 1: 205225_at,20979... list 2: ENSG00000001084,ENSG...	1.5495	7.006e-13	7.65e-10
negative regulation of programmed cell death (GO:0043069)	418	409	list 1:  7.2% list 2:  1.65%	list 1: 205225_at,20979... list 2: ENSG00000001084,ENSG...	1.5334	1.074e-12	7.65e-10
cellular amino acid derivative metabolic process (GO:0006575)	182	173	list 1:  4.8% list 2:  0.68%	list 1: 209604_s_at,209... list 2: ENSG00000001084,ENSG...	1.995	9.24e-13	7.65e-10
cellular amino acid and derivative metabolic process (GO:0006519)	447	447	list 1:  7.4% list 2:  1.77%	list 1: 209604_s_at,209... list 2: ENSG00000001084,ENSG...	1.491	1.7e-12	9.082e-10

↑  
**Enriched class**

↑  
**Annotated genes per GO from each list**

# FatiGO results

Graphical results:



More results:

▼ Other actions

- Open input form
- Change p value

0.1 0.05 0.01 **0.005**

# FatiGO

## Exercises

- Go to the tutorial:  
[http://bioinfo.cipf.es/babelomicstutorial/enrichment\\_analysis](http://bioinfo.cipf.es/babelomicstutorial/enrichment_analysis)
- Run worked examples
- Repeat examples modifying parameters
- Run FatiGO exercise (from the tutorial)