Data Analysis.
Functional Profiling: FatiGO

Joaquín Tárraga Giménez
jtarraga@cipf.es
Valencia, September 2011

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

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

er ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED
DE ENFERMEDADES RARAS

Pre-genomic era

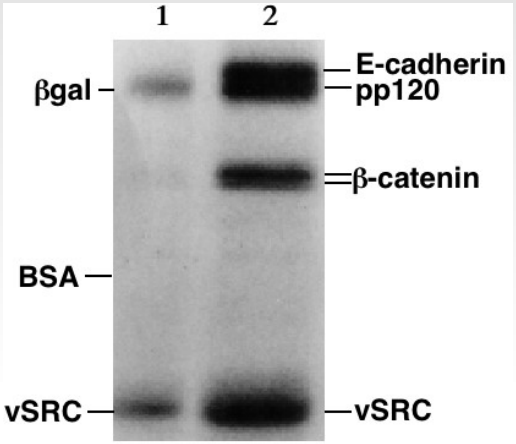
Extract as much information as possible for one single data

Easy data analysis and biological interpretation of the results

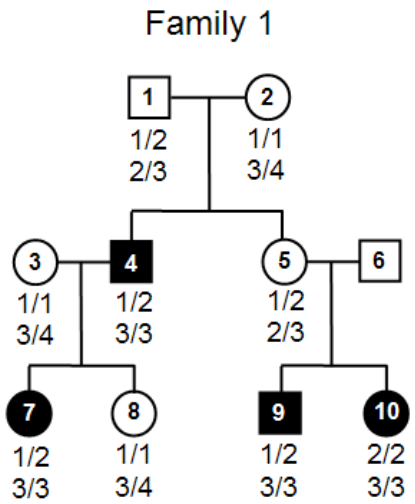
Western blot

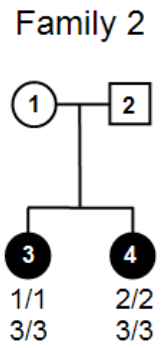
Southern blot

Northern blot

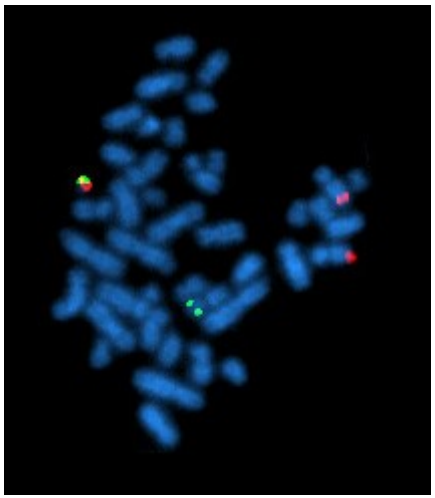


Linkage studies





FISH



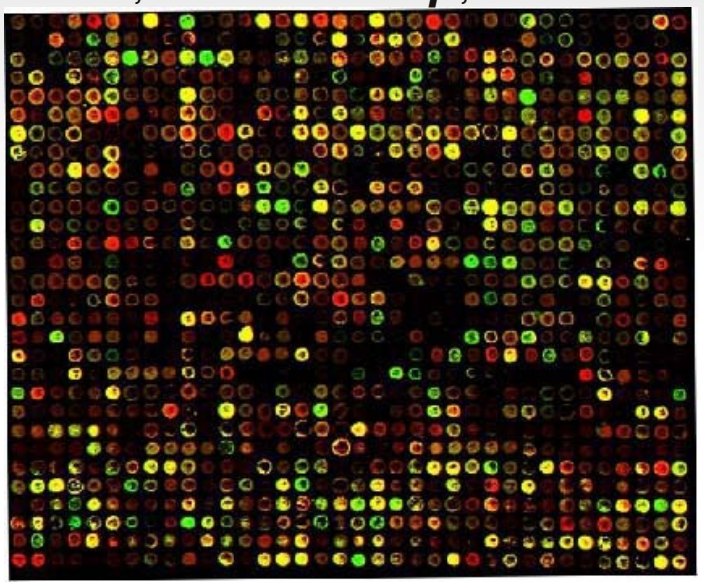
Post-genomic era

Huge amount and different type of data

In 2000 high-throughput technologies appeared

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

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



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/mm² 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.

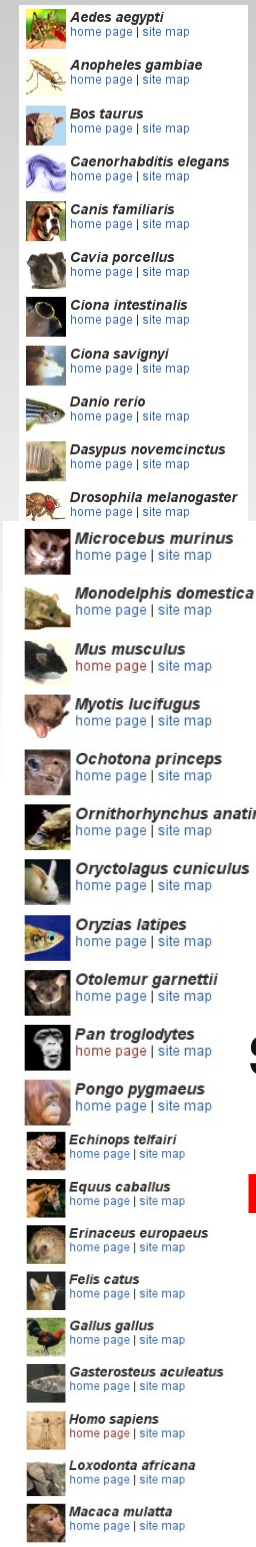
Service and Support
Illumina will ensure that your HiSeq 2000 is properly installed and qualified, and will provide ongoing maintenance and service. This industry-leading support is available in North America, Europe, and Asia.

GEO Gene Expression Omnibus

Public data

GPL Platforms	5256
GSM Samples	262357
GSE Series	10220
Total	277833

Functional Annotation DDBB



Gene IDs

Functional databases

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

Keywords Swissprot

KEGG pathways

Biocarta pathways

Gene Ontology

Biological Process

Molecular Function

Cellular Component

IntrePro Motifs

Domains

Bioentities from literature:

Diseases, chemical terms

Gene Expression in tissues

Regulatory elements

miRNA

CisRed

Transcription Factor Binding Sites

Reactome

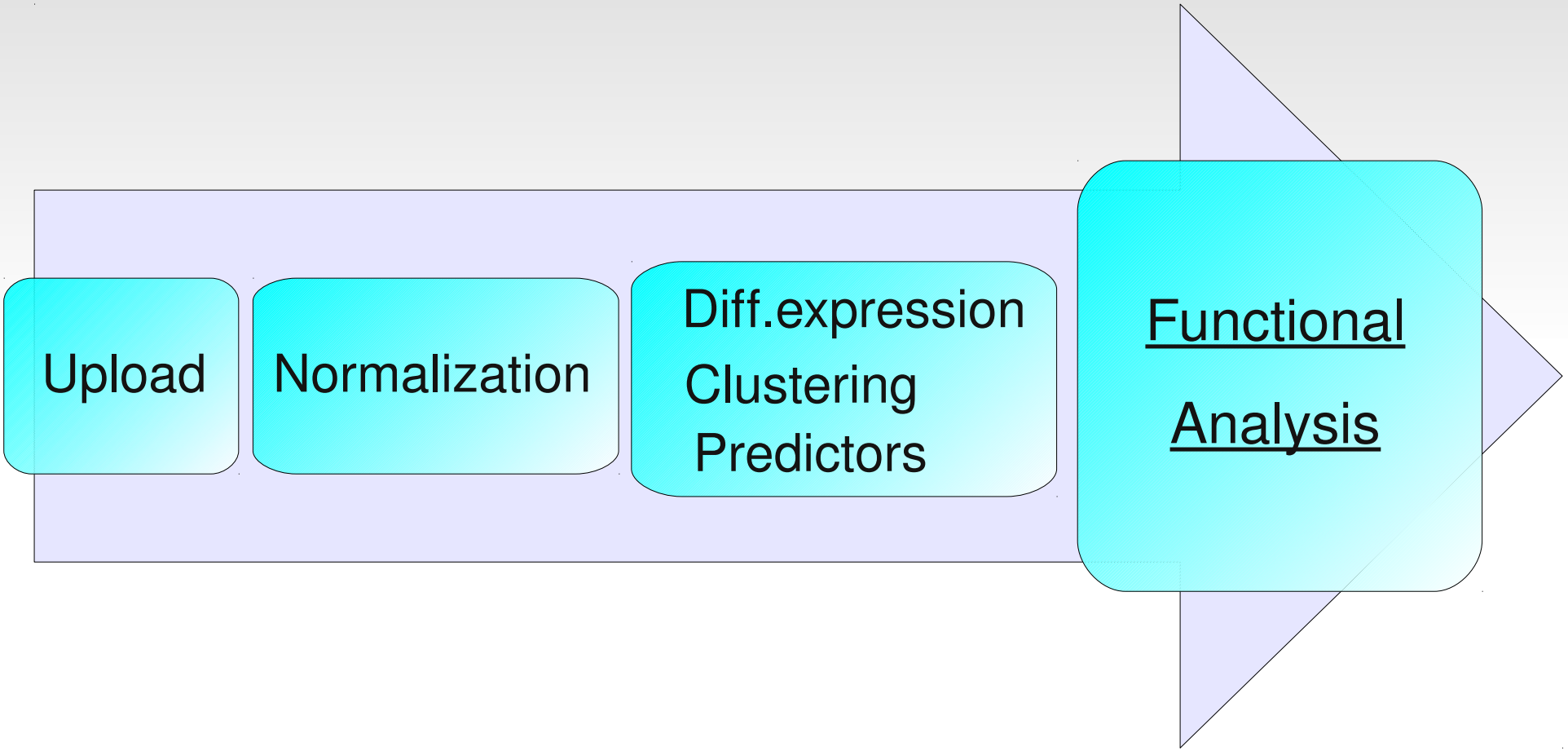
Gene Ontoloty (GO)

- 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



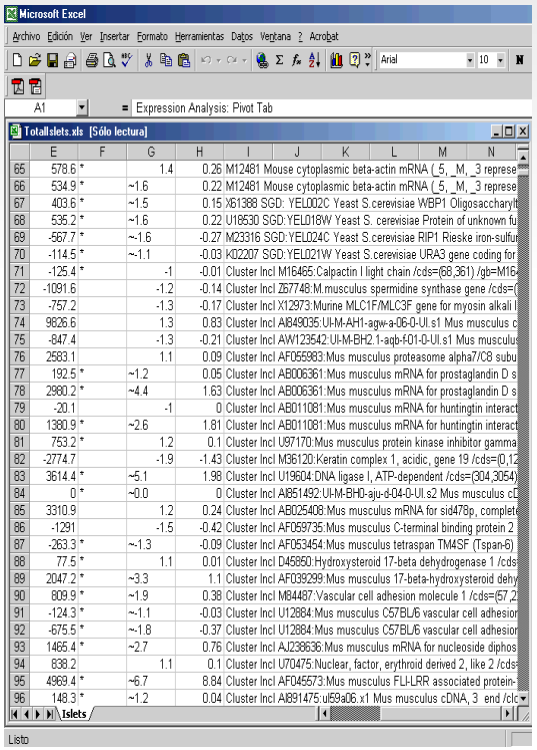
Functional profiling of genome-scale experiements

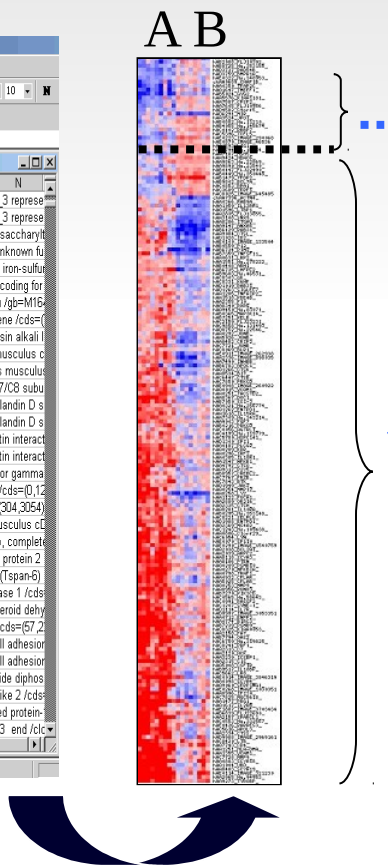
My data...

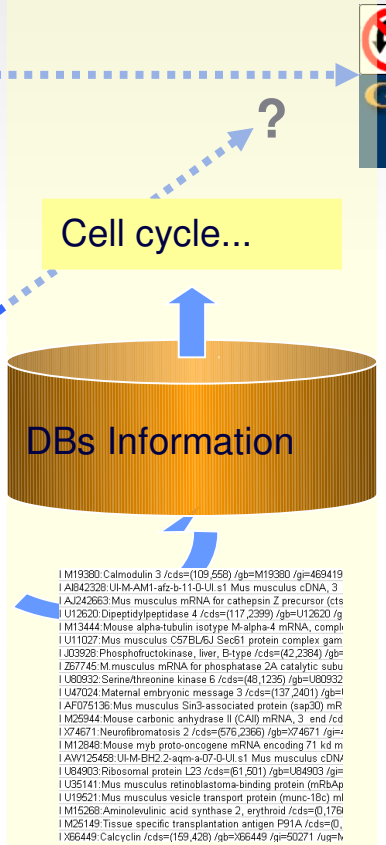
How are structured?

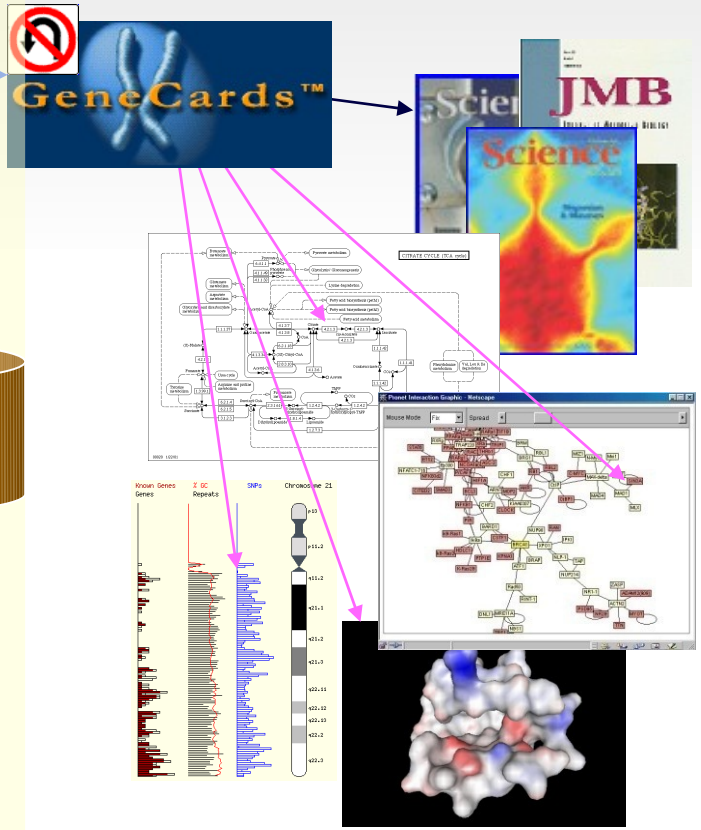
What are these groups?

What is this gen?









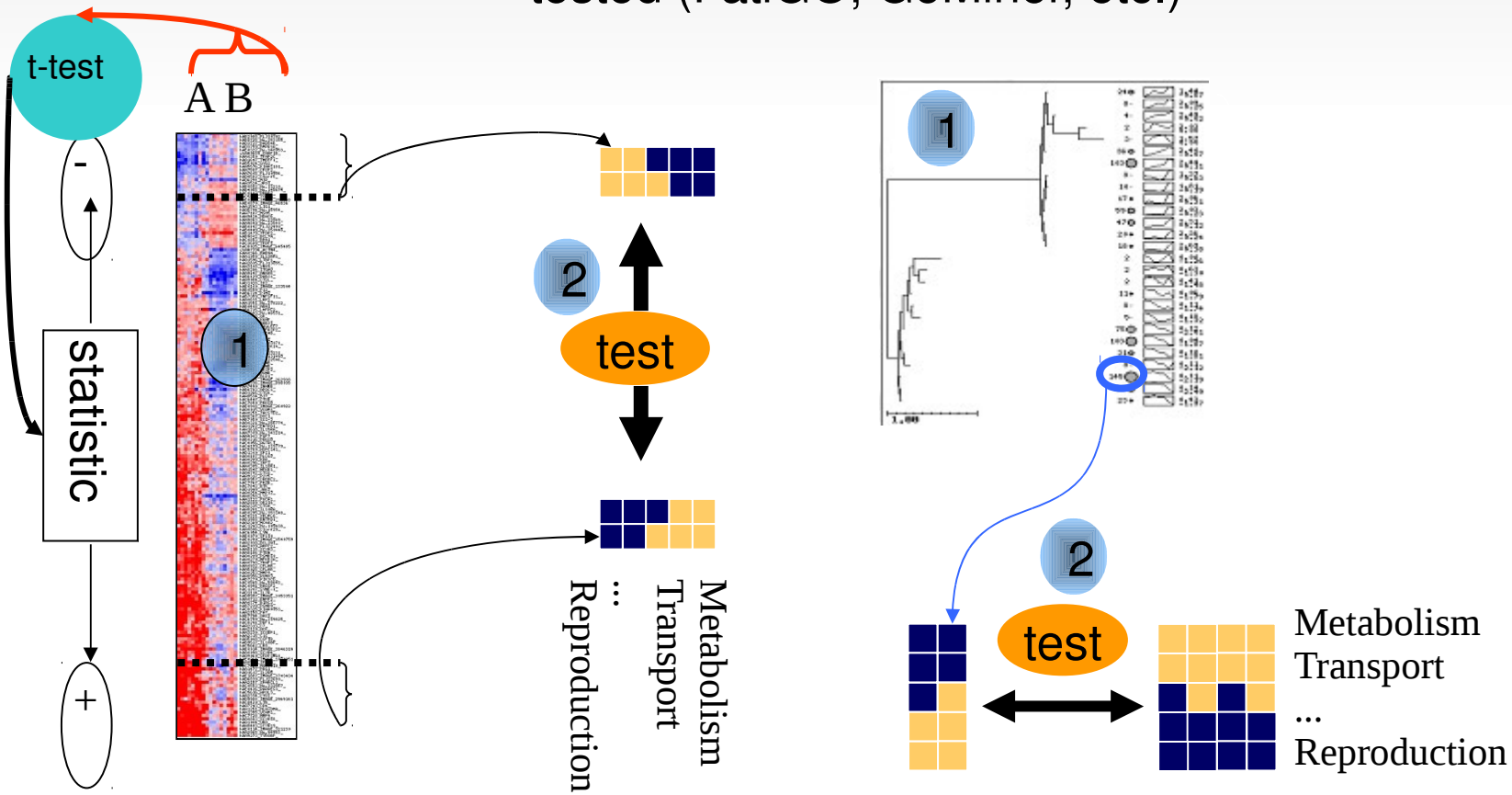
Analysis

Functional profiling

Links

Two-steps functional analysis

1. Genes are selected based on their experimental values and...
2. Enrichment in functional terms is tested (FatiGO, GoMiner, etc.)

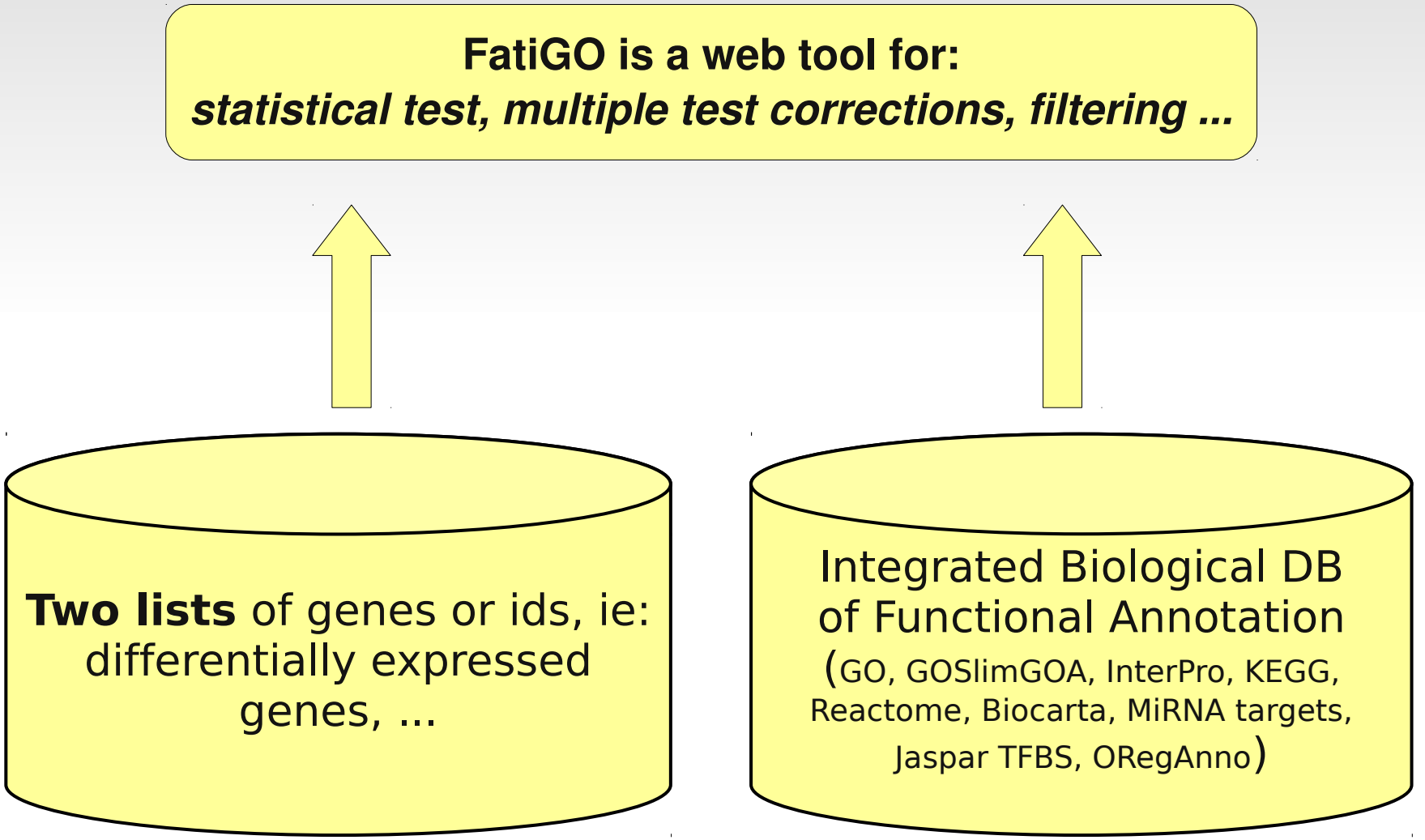


FatiGO

Questions that Functional enrichment analysis try to answer

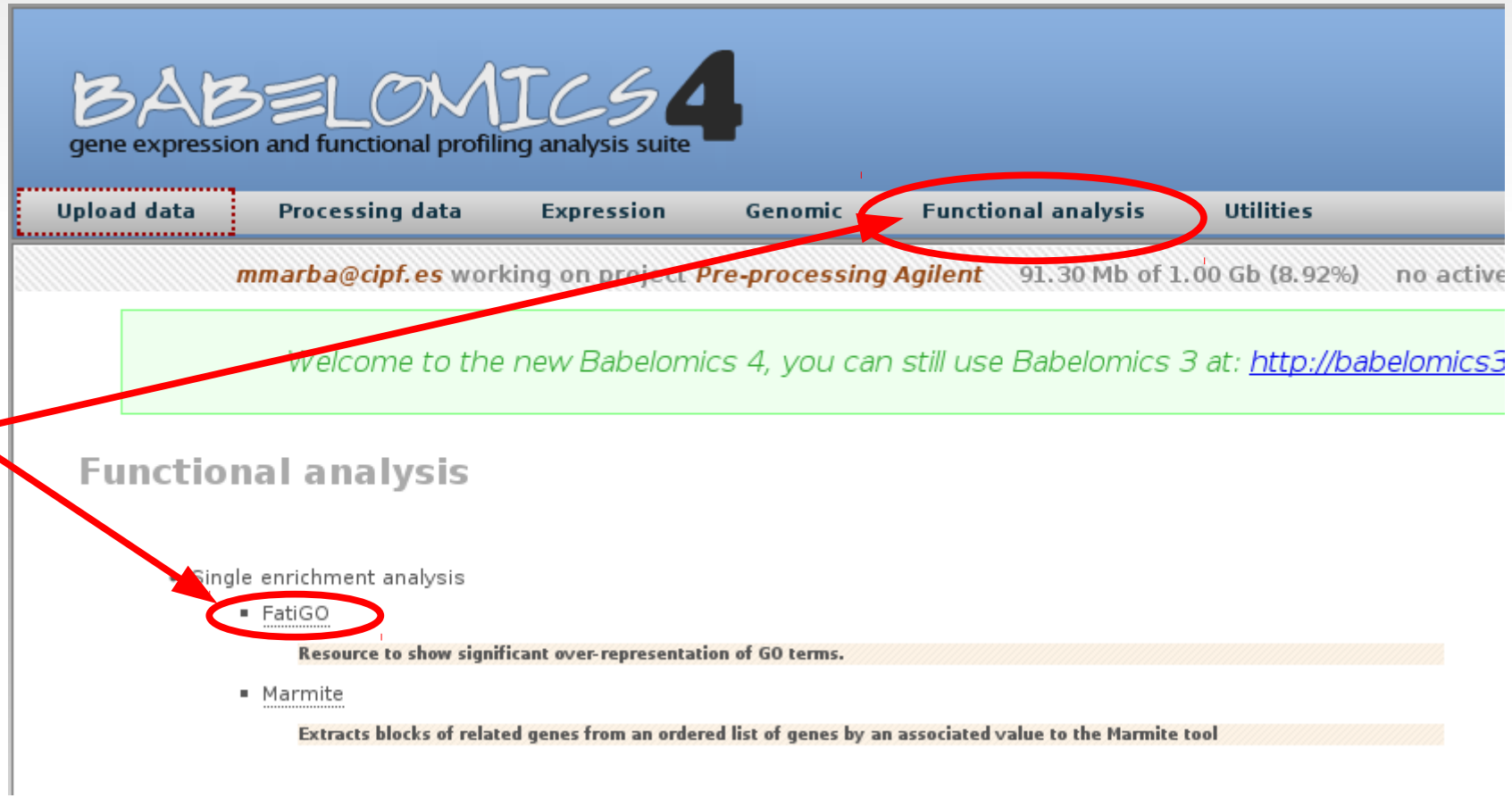
- Is there any significant functional enrichment in my gene list?
- Are these genes involved in same pathways?
- Are they sharing a specific microRNA regulation?
- Are they involved in the same disease?

Fatigo
Schema



FatiGO

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

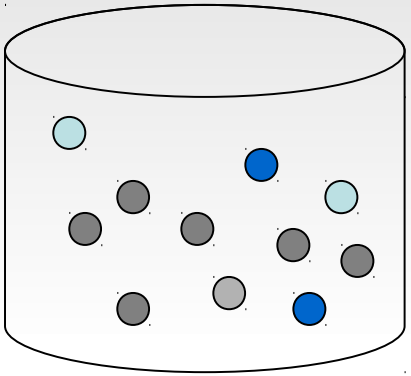


FatiGO features

- 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

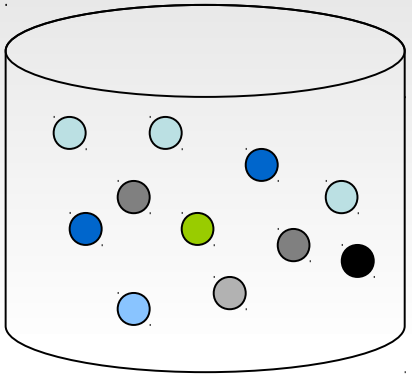
One Gene List (A)



Biosynthesis 60%

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

The other list (B)



Biosynthesis 20%

Fisher's exact test

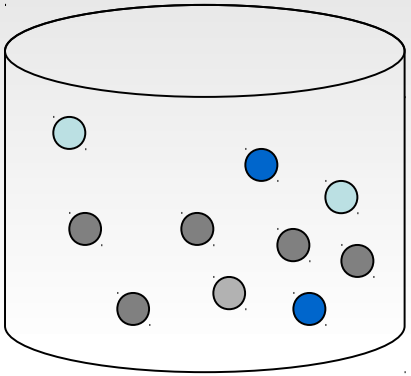
	A	B
Biosynthesis	**60**	**20**
No biosynthesis	**40**	**80**

p-value = 0.00000431

Genes in group A have significantly to do with biosynthesis

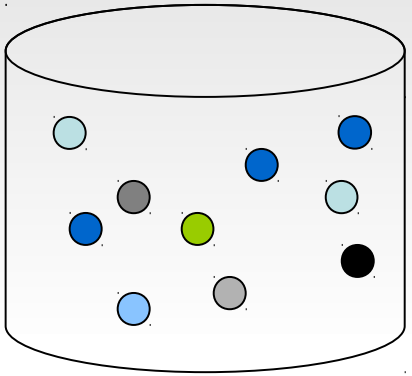
FatiGO test

One Gene List (A)

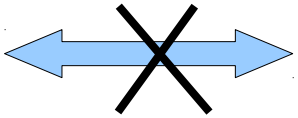


Are this two groups of genes carrying out different biological roles?

The other list (B)



Sporulation 20%



Sporulation 30%

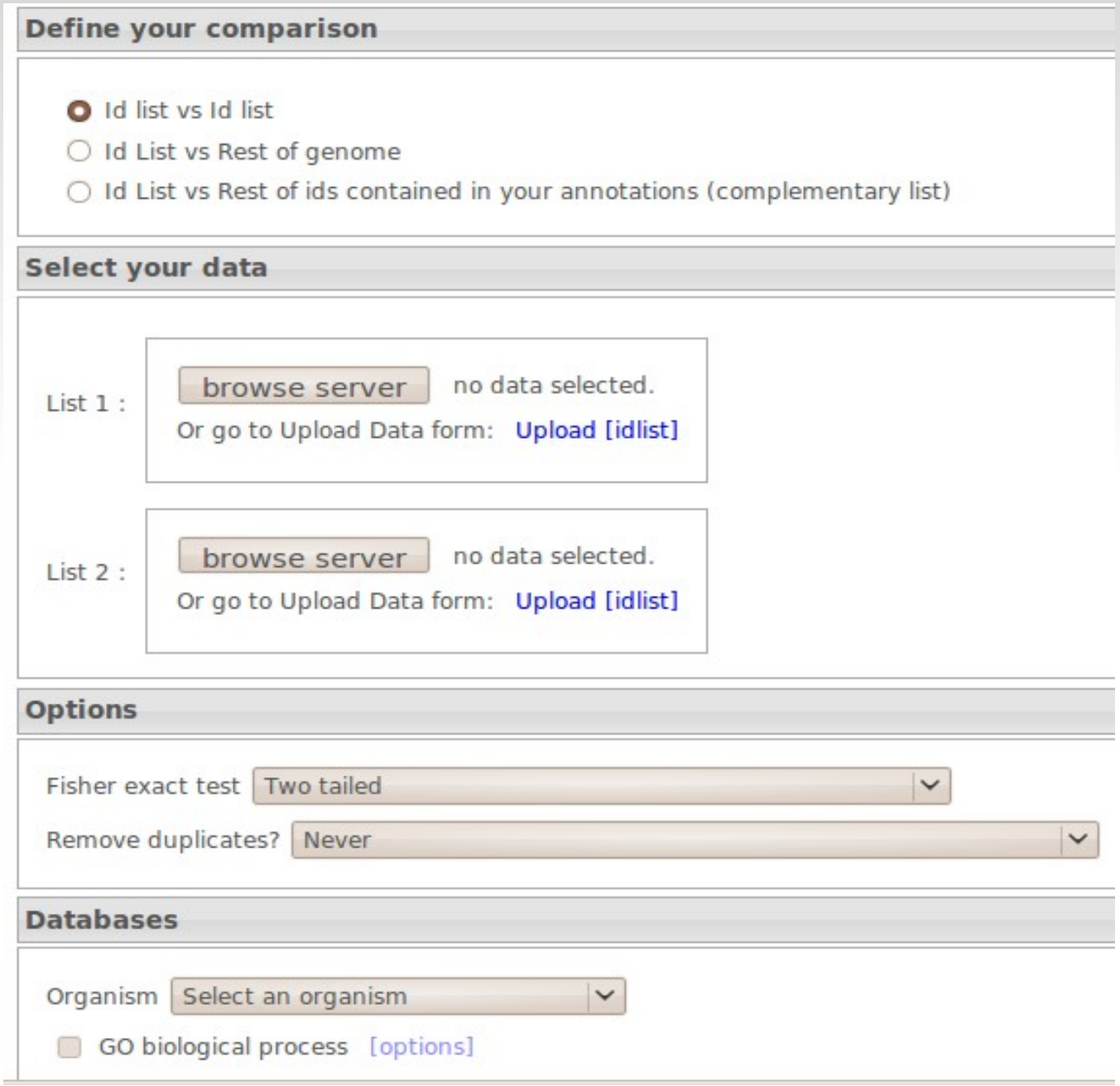
Fisher's exact test

	A	B
Sporulation	**20**	**30**
No sporulation	**80**	**70**

p-value = 0.964

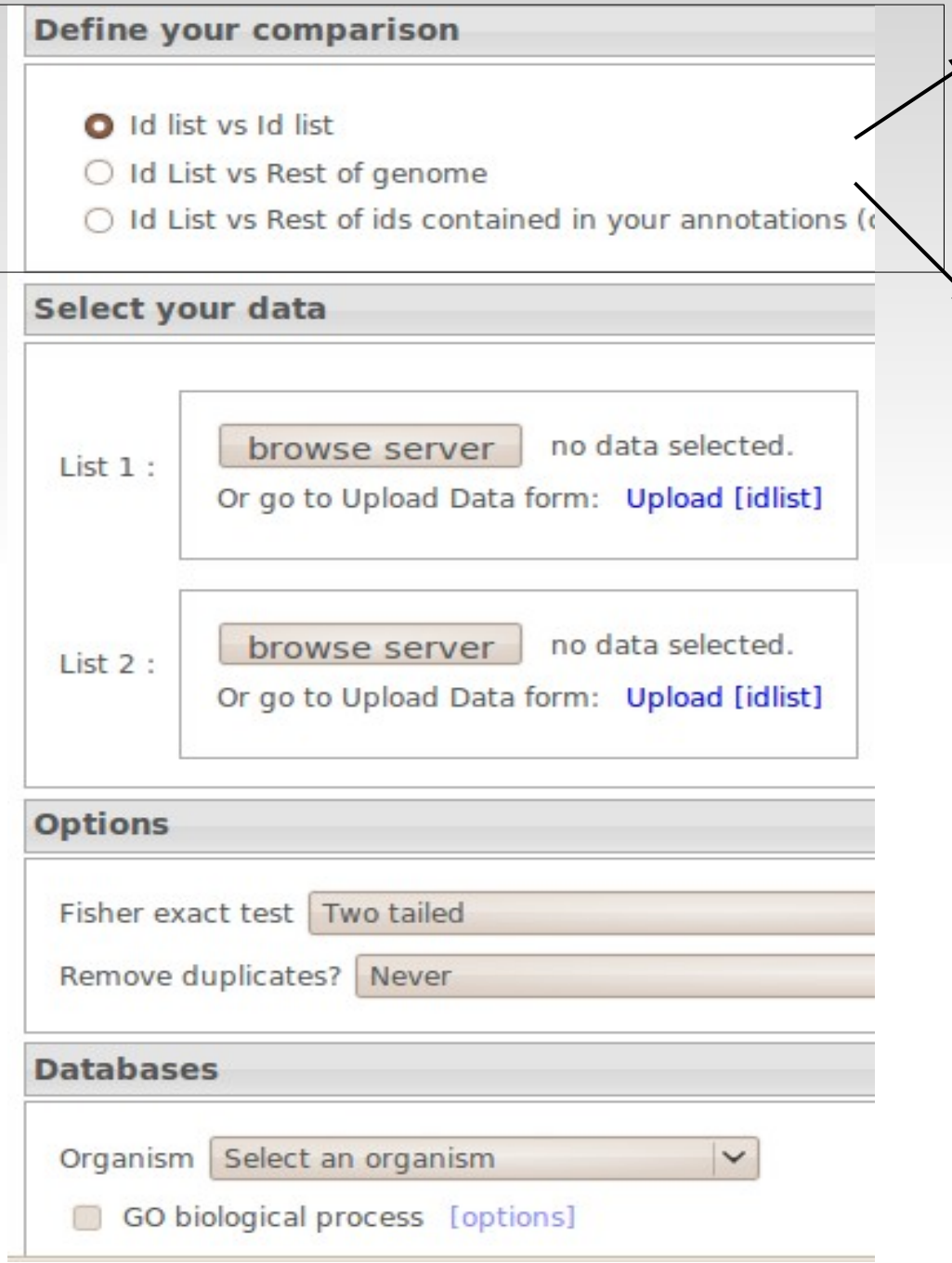
Genes in group A DO NOT have significantly to do with sporulation.

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 :
browse server
no data selected.
Or go to Upload Data form: Upload [idlist]
List 2 :
browse server
no data selected.
Or go to Upload Data form: Upload [idlist]
Options
Fisher exact test
Two tailed
Remove duplicates?
Never
Databases
Organism
Select an organism
GO biological process [options]

FatiGO form

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

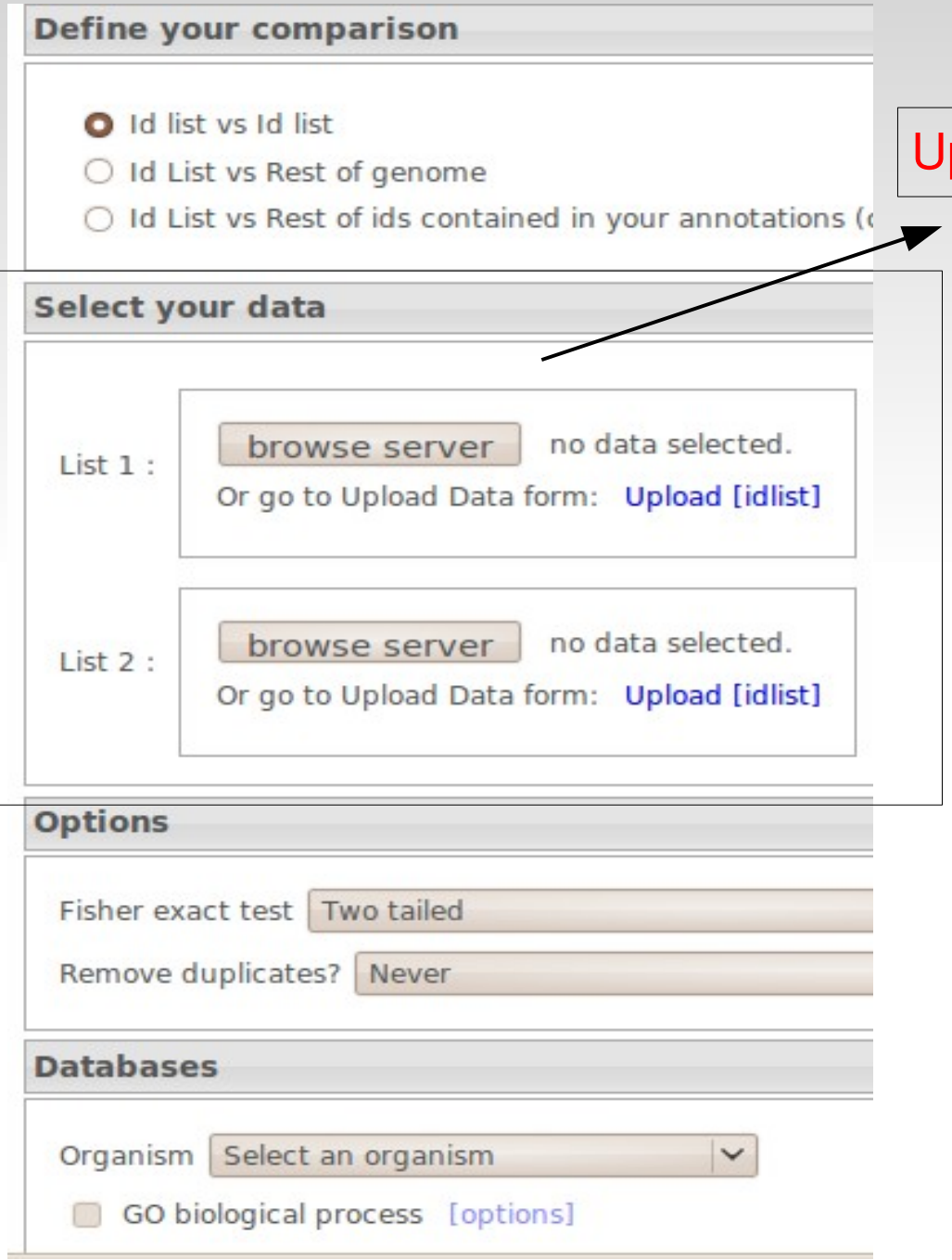


eg. Compare 2tissues or responder genes vs. non-responders

eg: genes that respond to one treatment against the genome

FatiGO form

Data selection

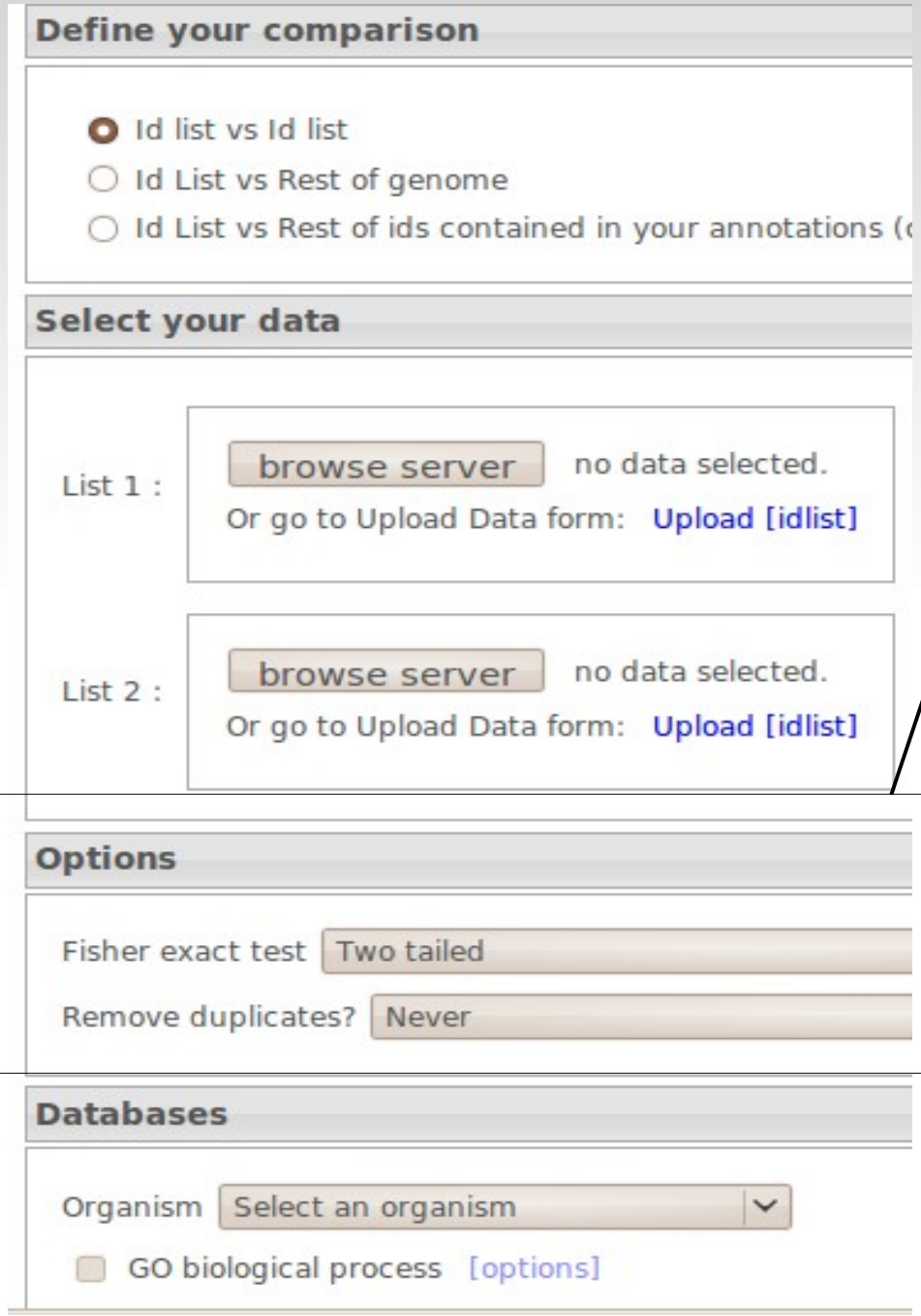


Upload first at Data Upload

"txt" file with gene lists:

gene1
gene2
gene3
...

FatiGO form



Algorithm options

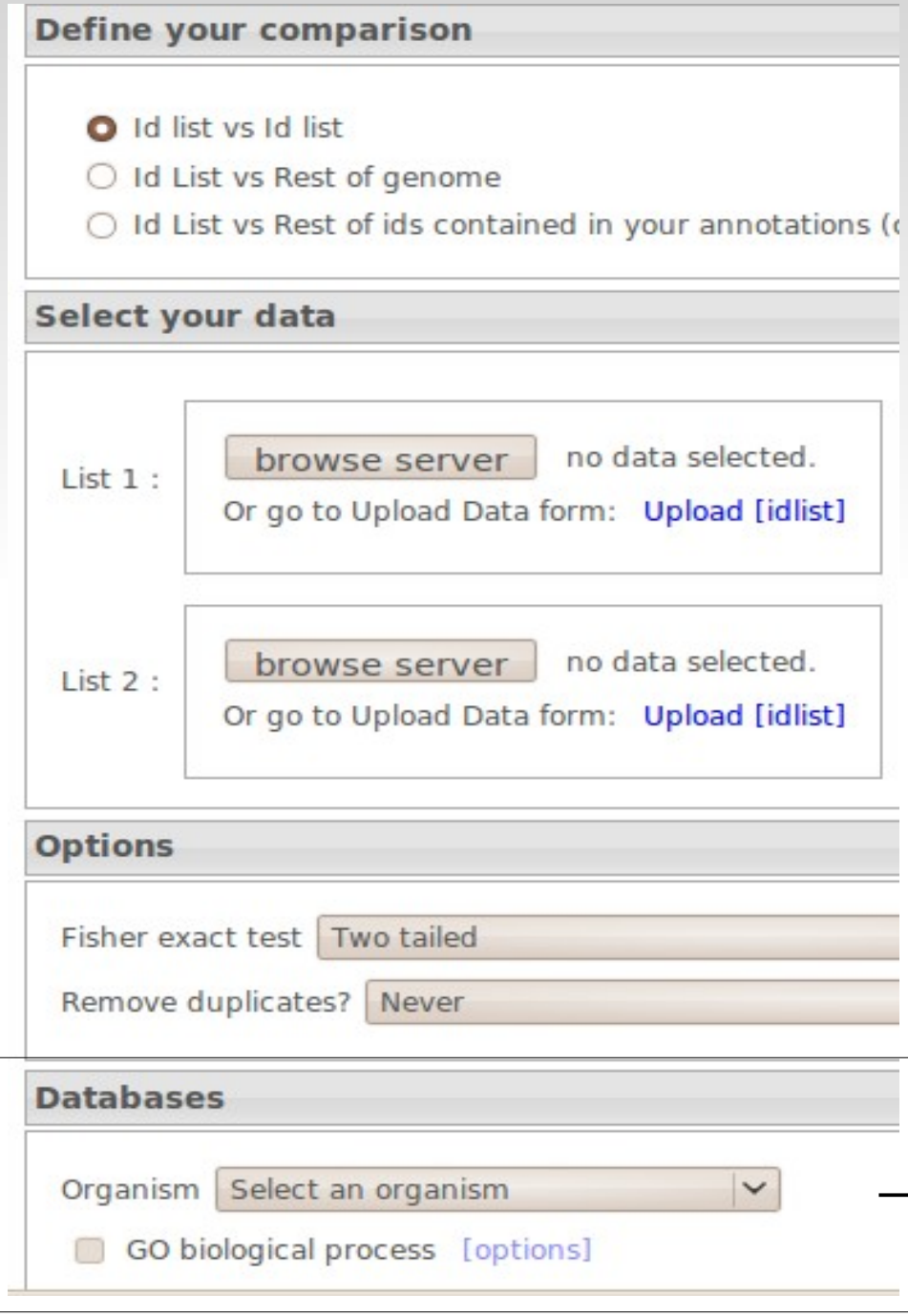
HINT:

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

Removing duplicates:

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

FatiGO form

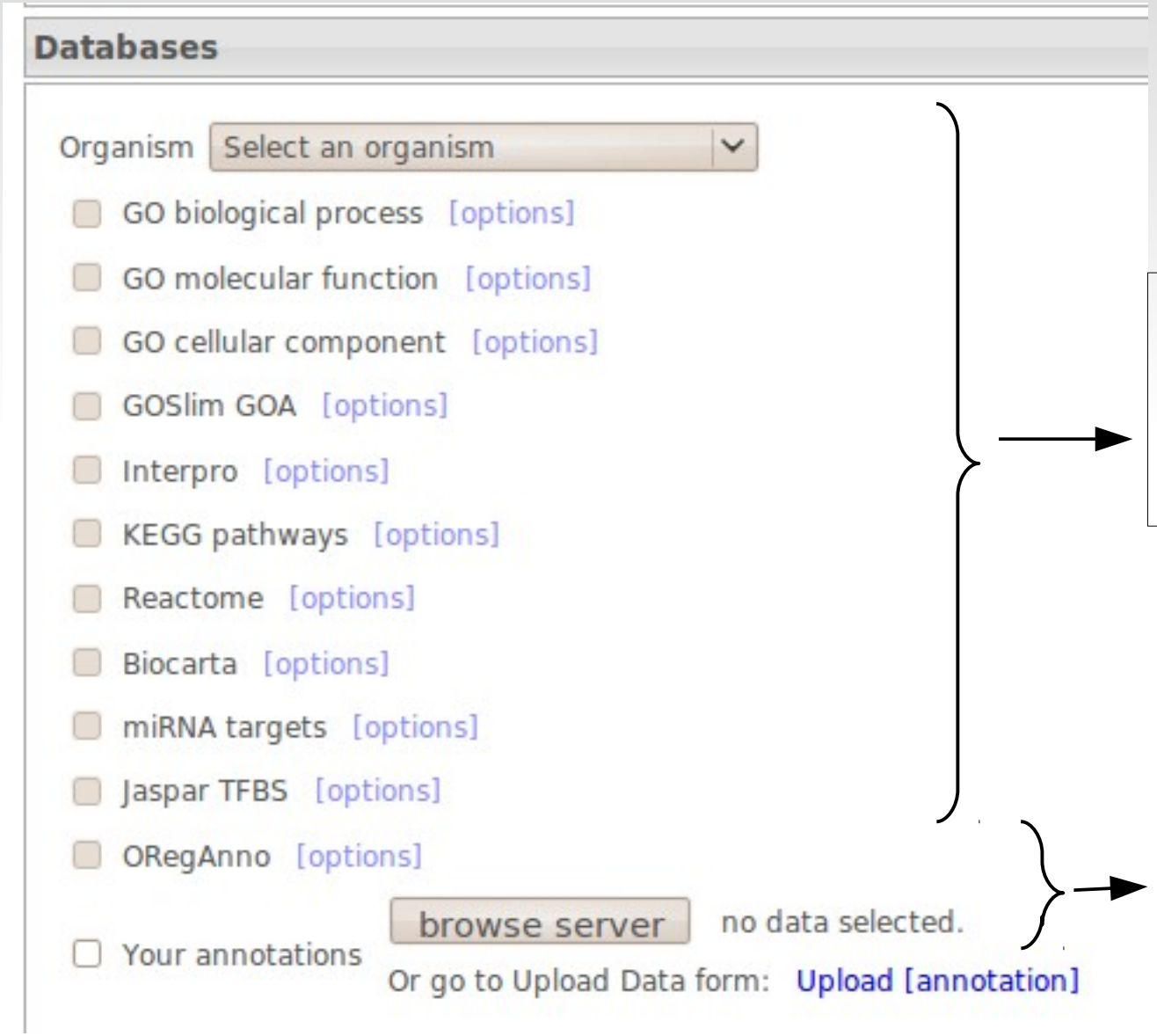


Which type of functional information?

?

FatiGO form

Which type of functional information?



Use one or more of the given databases

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

FatiGO form

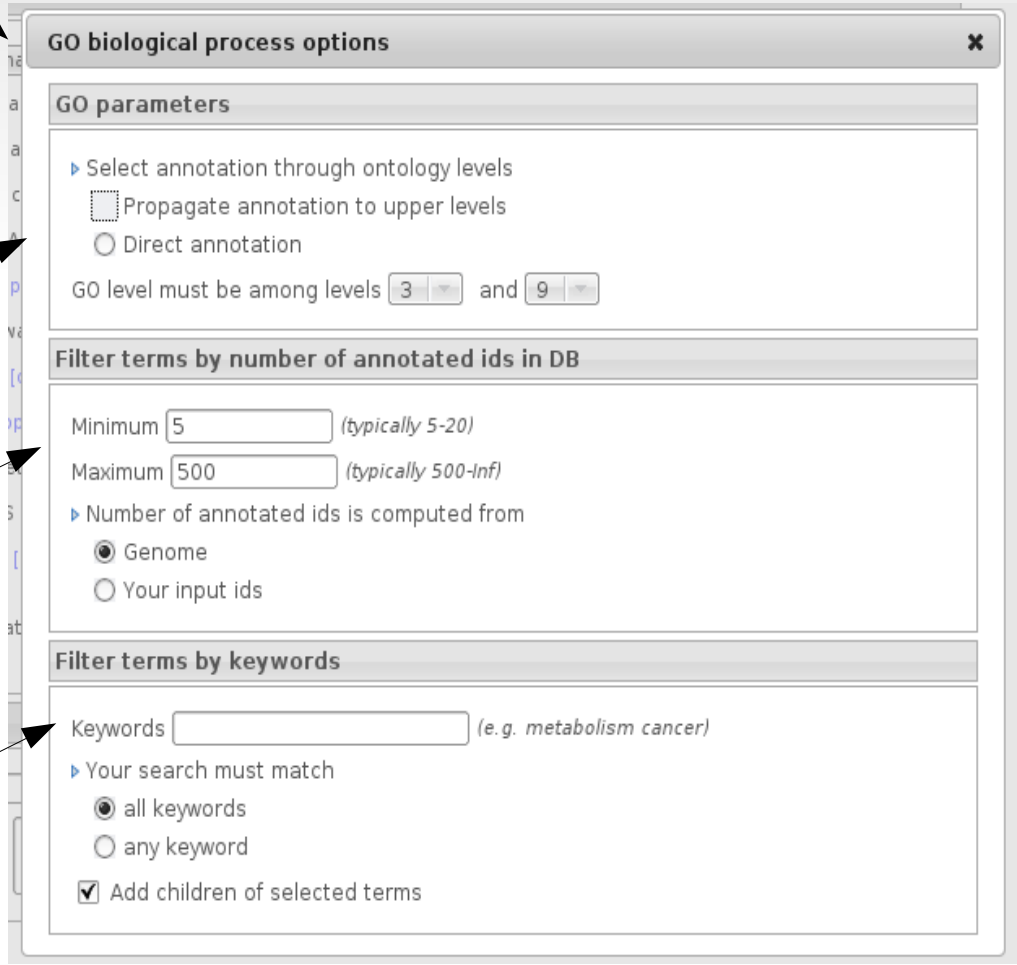
First select an organism

OPTIONS:

Test all the GO or only annotated terms

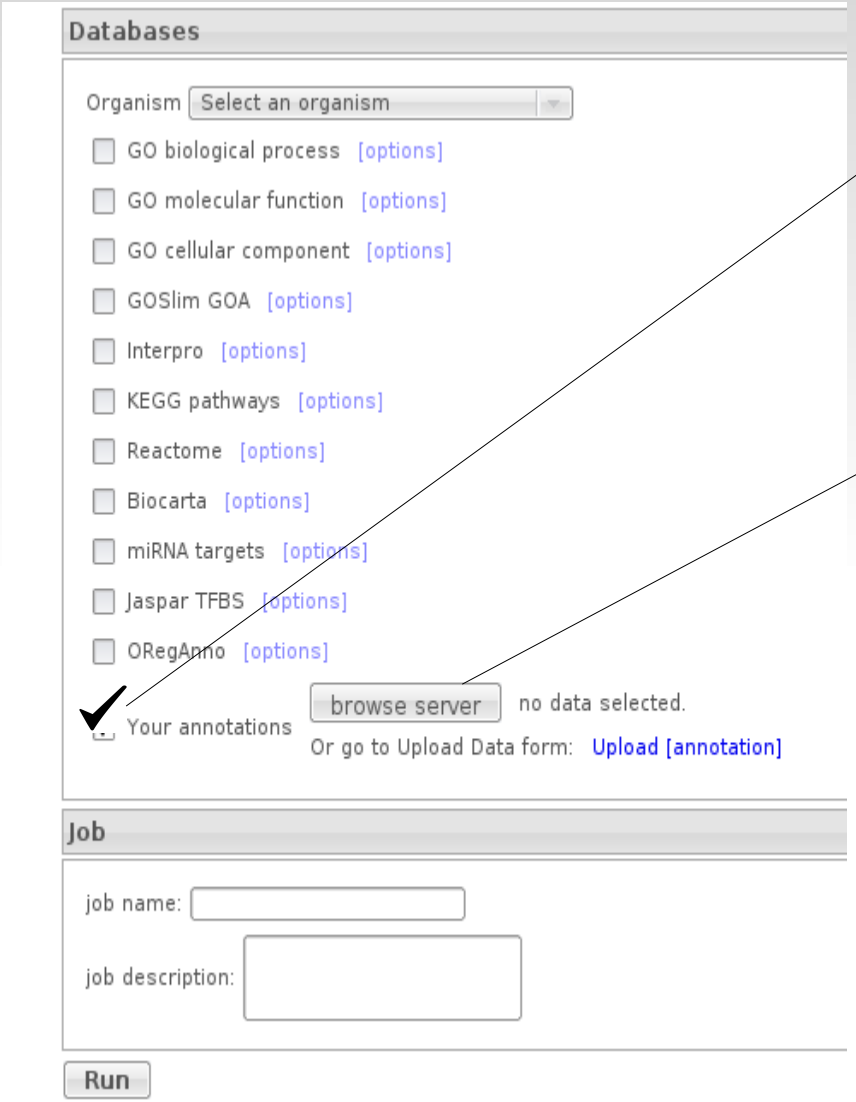
Discard functions with too few or too many genes?

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



FatiGO form

Which type of functional information?



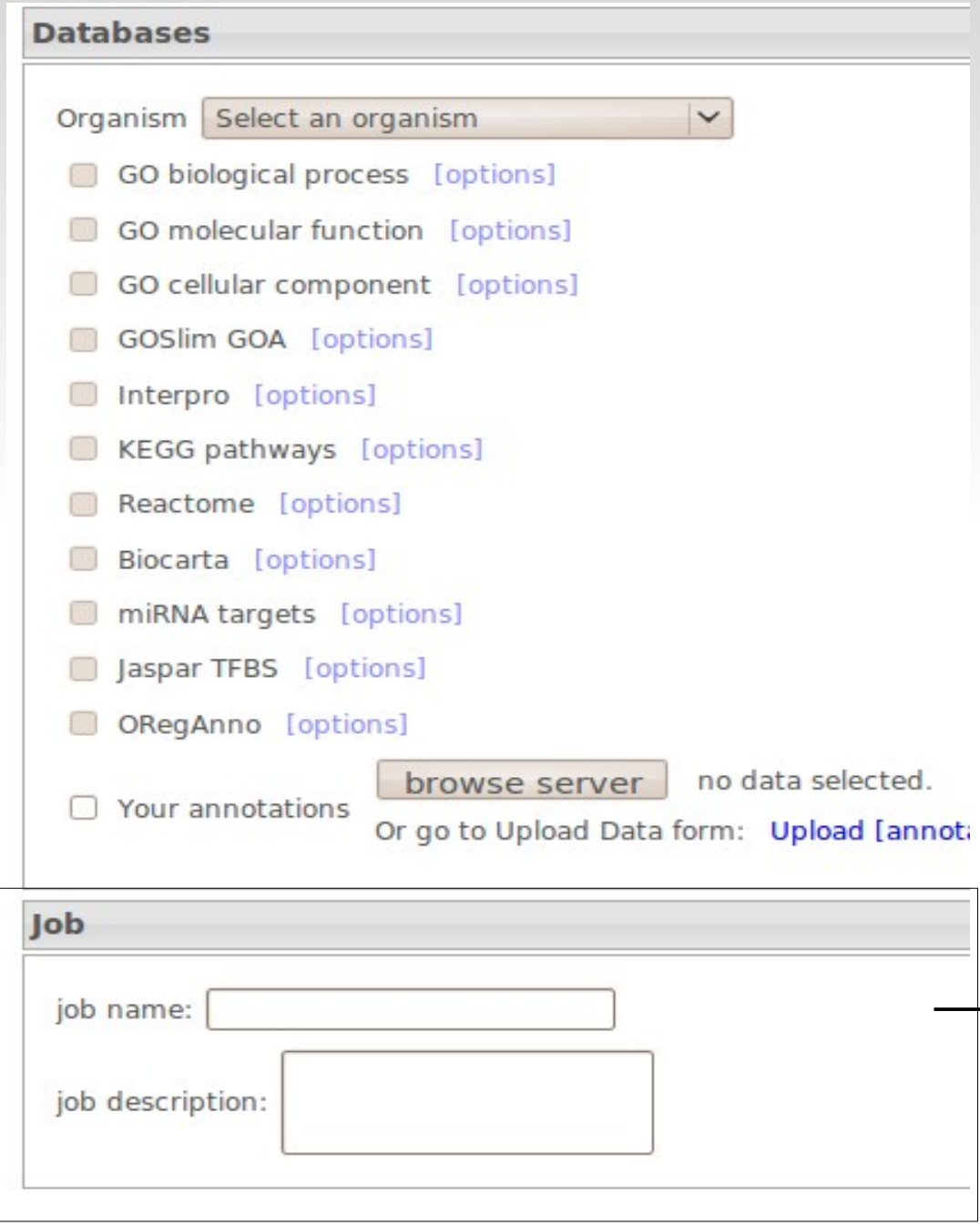
<u>Your annotations</u>: useful when you work with your own annotations OR with an organism that is not in Babelomics

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



What's your job name?

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 3conditions
- 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 :

DB	List1	Genome
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 :

DB	Number of significant terms
GO biological process (levels from 3 to 9)	142
GO molecular function (levels from 3 to 9)	30

FatiGO results

Significant results:

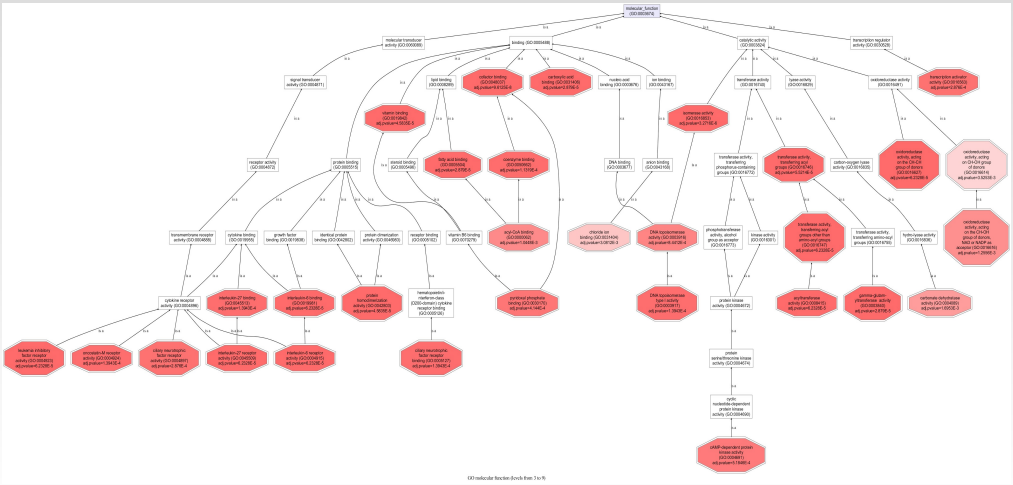
Term	Term size	Term size (in genome)	Term annotation % per list	Annotated ids	Odds ratio (log e)	pvalue	Adjusted pvalue
negative regulation of apoptosis (GO:0043066)	412	403	list 1: 7.2% list 2: 1.62%	list 1: 205225_at,20979... list 2: ENSG00000001084,ENSG...	1.5495	7.006e-13	7.65e-10
negative regulation 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 par 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
- Run worked examples
- Repeat examples modifying parameters
- Run FatiGO exercise (from the tutorial)