BABELOMICS

Functional Enrichment Analysis: FatiGO & Fatiscan

Martina Marbà
mmarba@cipf.es

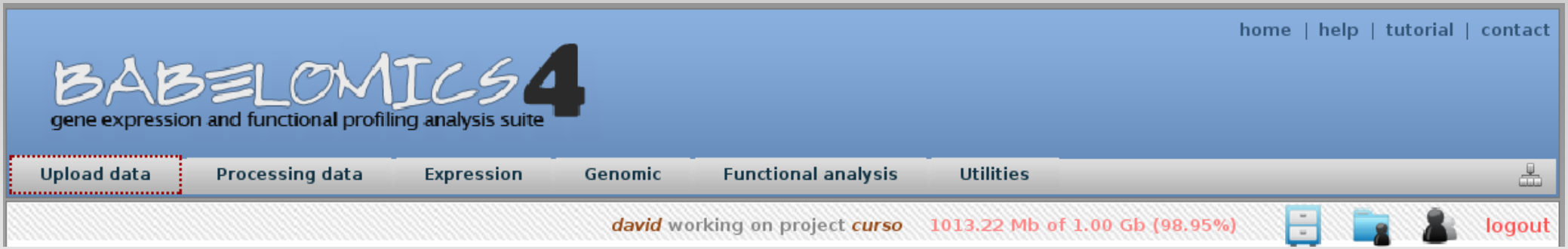
Valencia, 25th March 2011

Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

INB

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

er ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS



BABELOMICS

A systems biology web resource for the functional interpretation of genome-scale experiments.

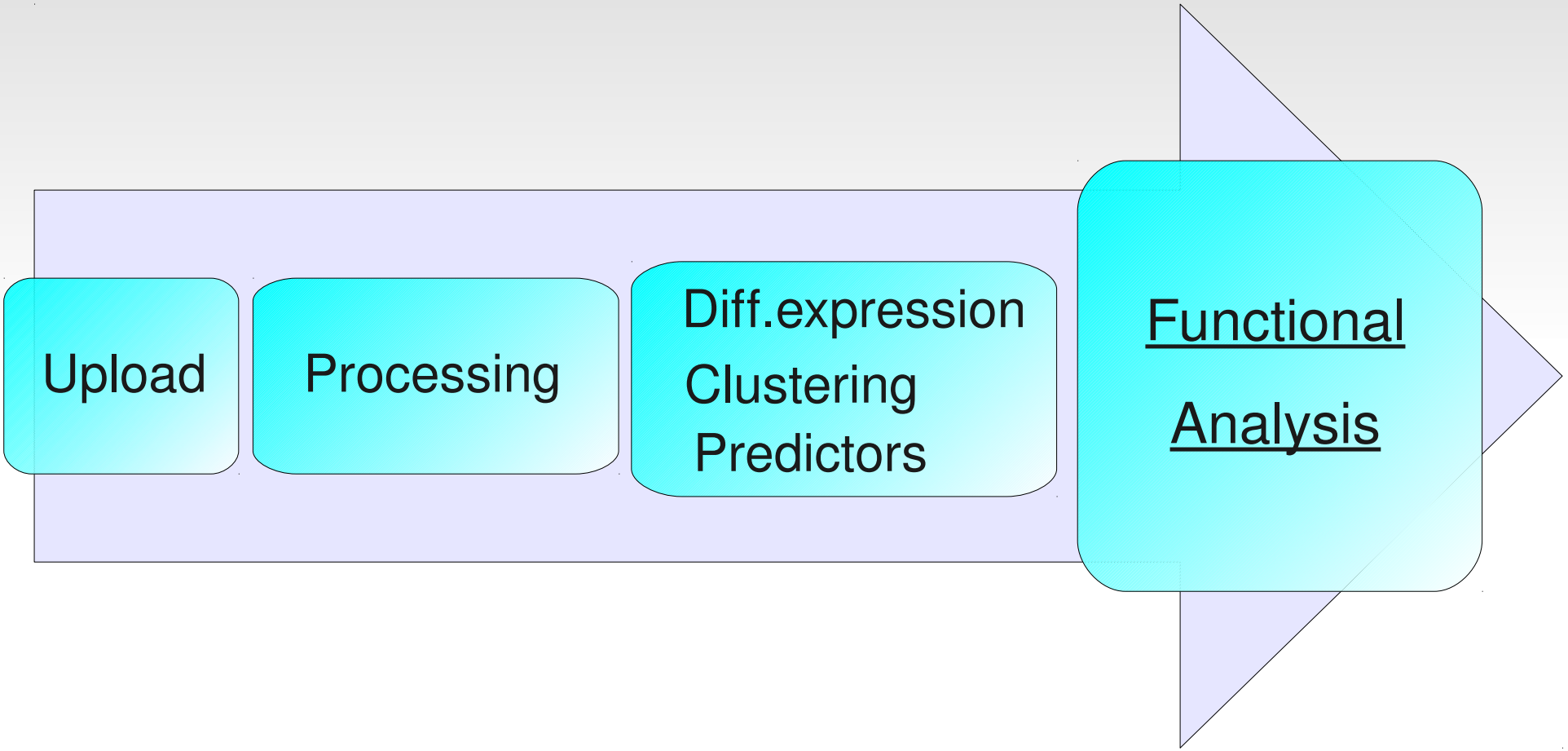
http://www.babelomics.org

Questions

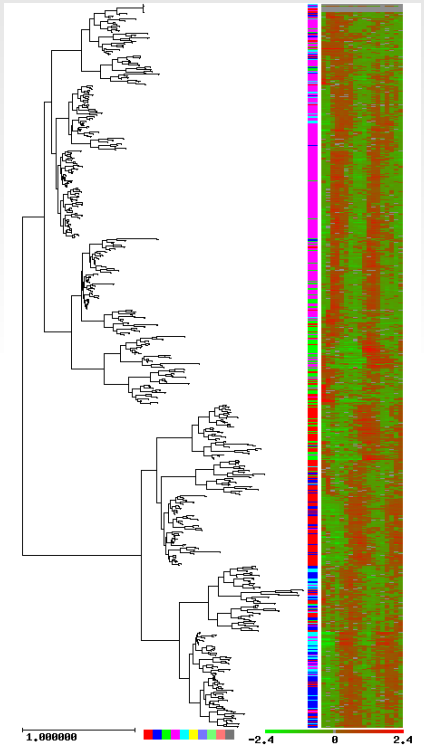
Questions that Functional enrichment analysis try to answer

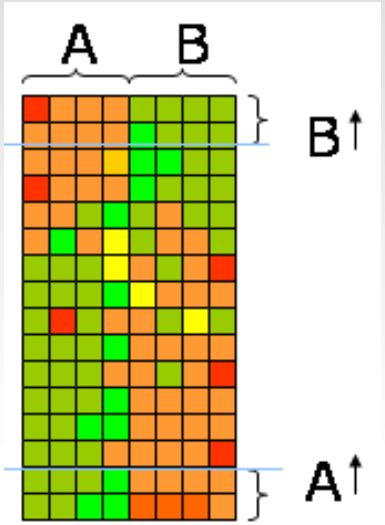
- Is there any significant functional enrichment in my gene list?
- Are these genes involved in same pathways?
- What biological processes differentiate a healthy control from a diseased case?
- Do these genes share a specific microRNA regulation?
- Are they involved in the same disease?

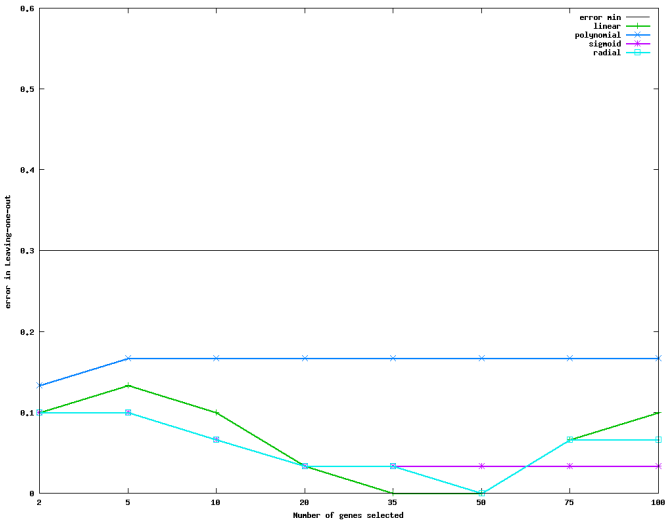
Data analysis workflow

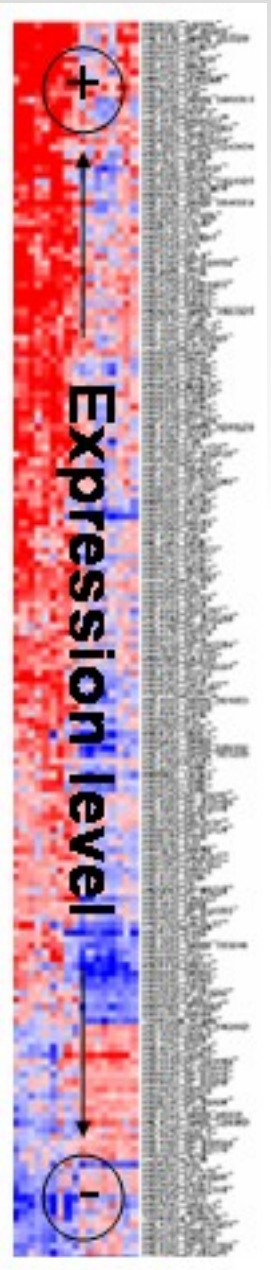


Genome-scale experiment output

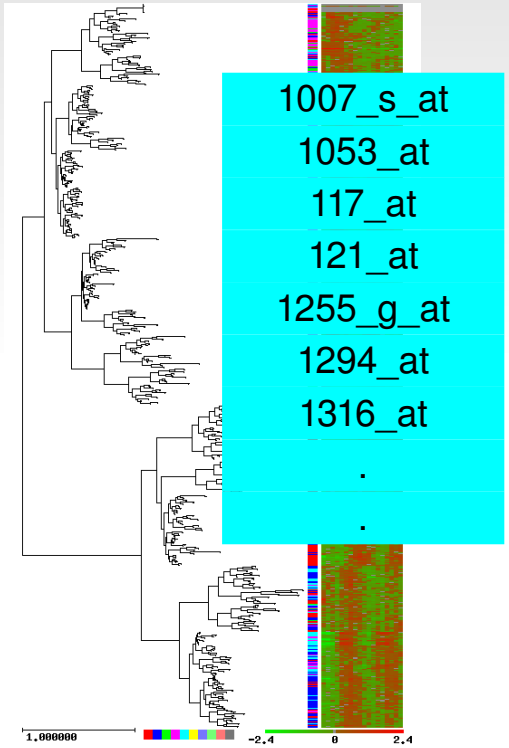




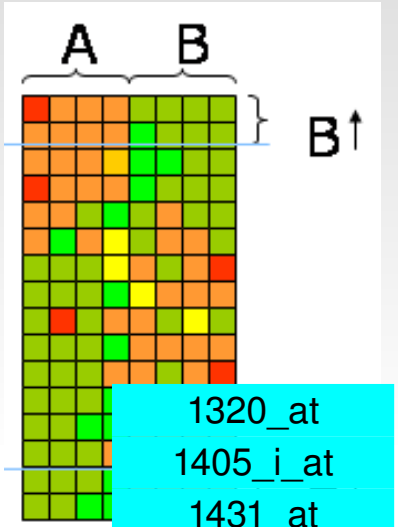




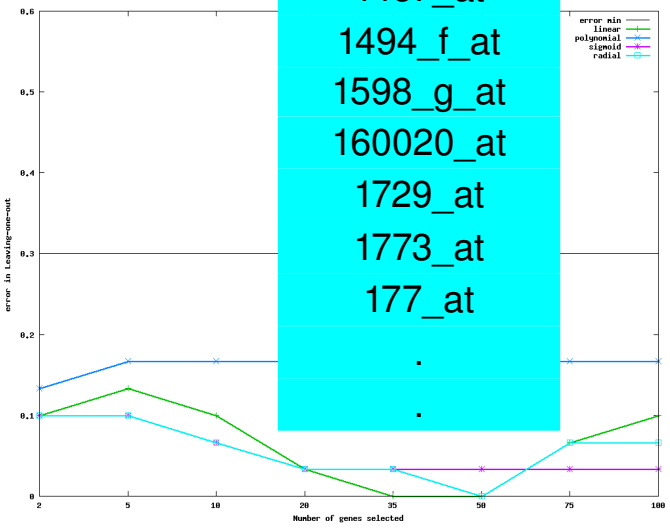
Genome-scale experiment output



1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

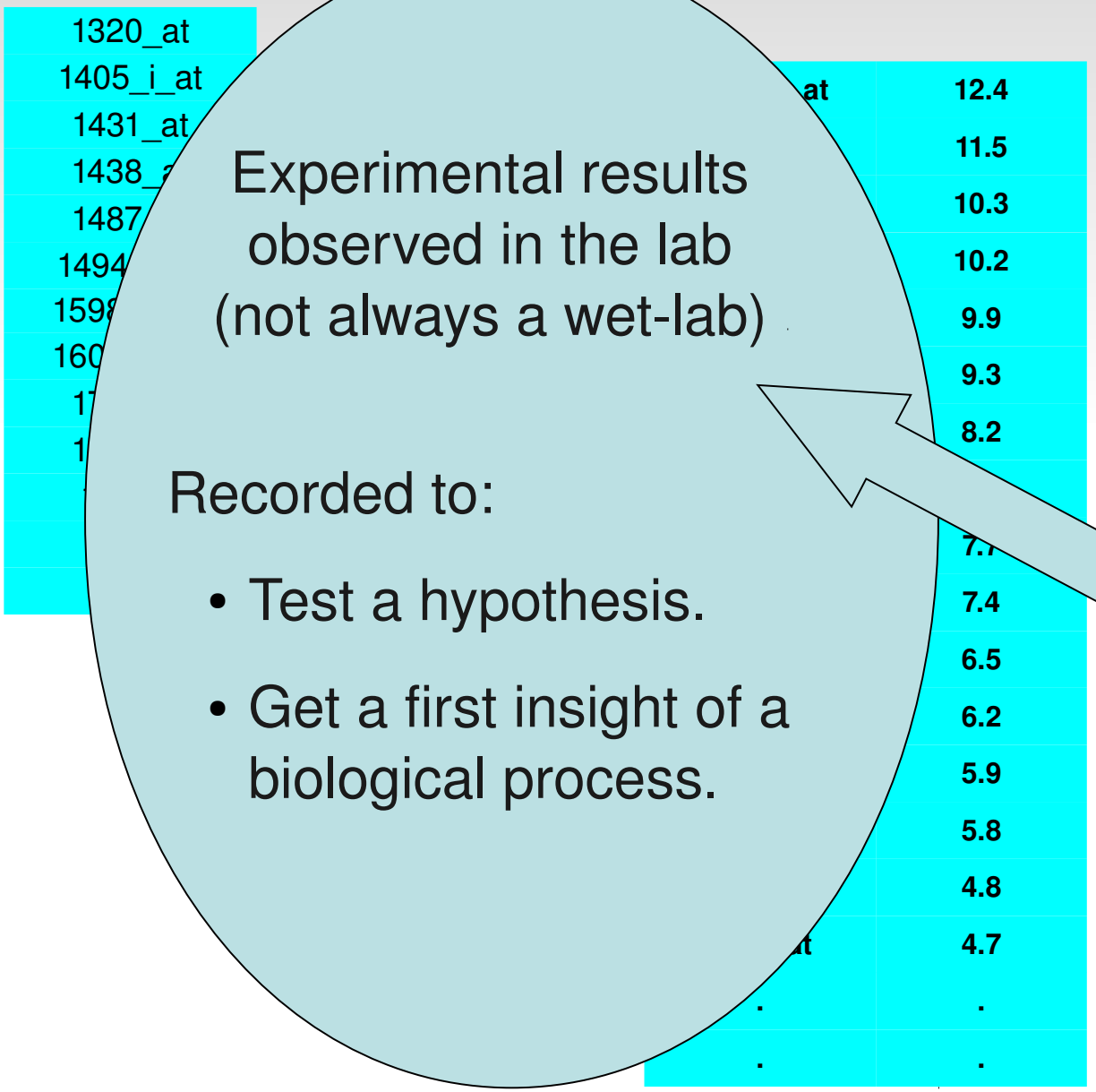


1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.



1007_s_at	12.4
1053_at	11.5
117_at	10.3
121_at	10.2
1255_g_at	9.9
1294_at	9.3
1316_at	8.2
1320_at	8.1
1405_i_at	7.7
1431_at	7.4
1438_at	6.5
1487_at	6.2
1494_f_at	5.9
1598_g_at	5.8
160020_at	4.8
1729_at	4.7
.	.
.	.

Functional interpretation

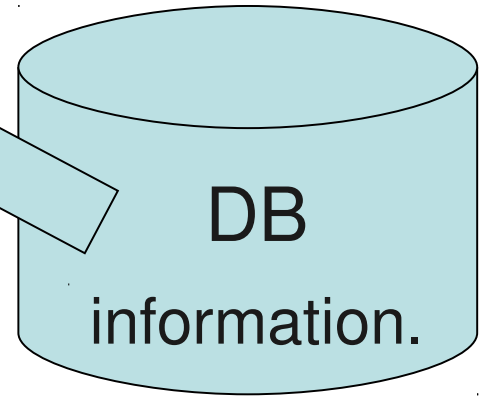


Experimental results observed in the lab (not always a wet-lab)

Recorded to:

- Test a hypothesis.
- Get a first insight of a biological process.

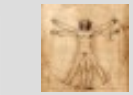
To “interpret” experimental results is to use **current knowledge** to rearrange them in a meaningful way.



Already tested and stored

Babelomics Databases

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









Homo Sapiens | *Mus musculus* | *Rattus Norv.* | *Gallus gallus* | *Danio-rerio* | *Drosophila melanogaster* | *C. elegans* | *Sacchar. cerevisae* | *Arabidopsis thaliana*

Gene IDs

UniProt/Swiss-Prot	RefSeq	HGNC symbol	PDB
UniProtKB/TrEMBL	Agilent	EMBL acc	Protein Id
Ensembl IDs	Affymetrix	Entrez Gene	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

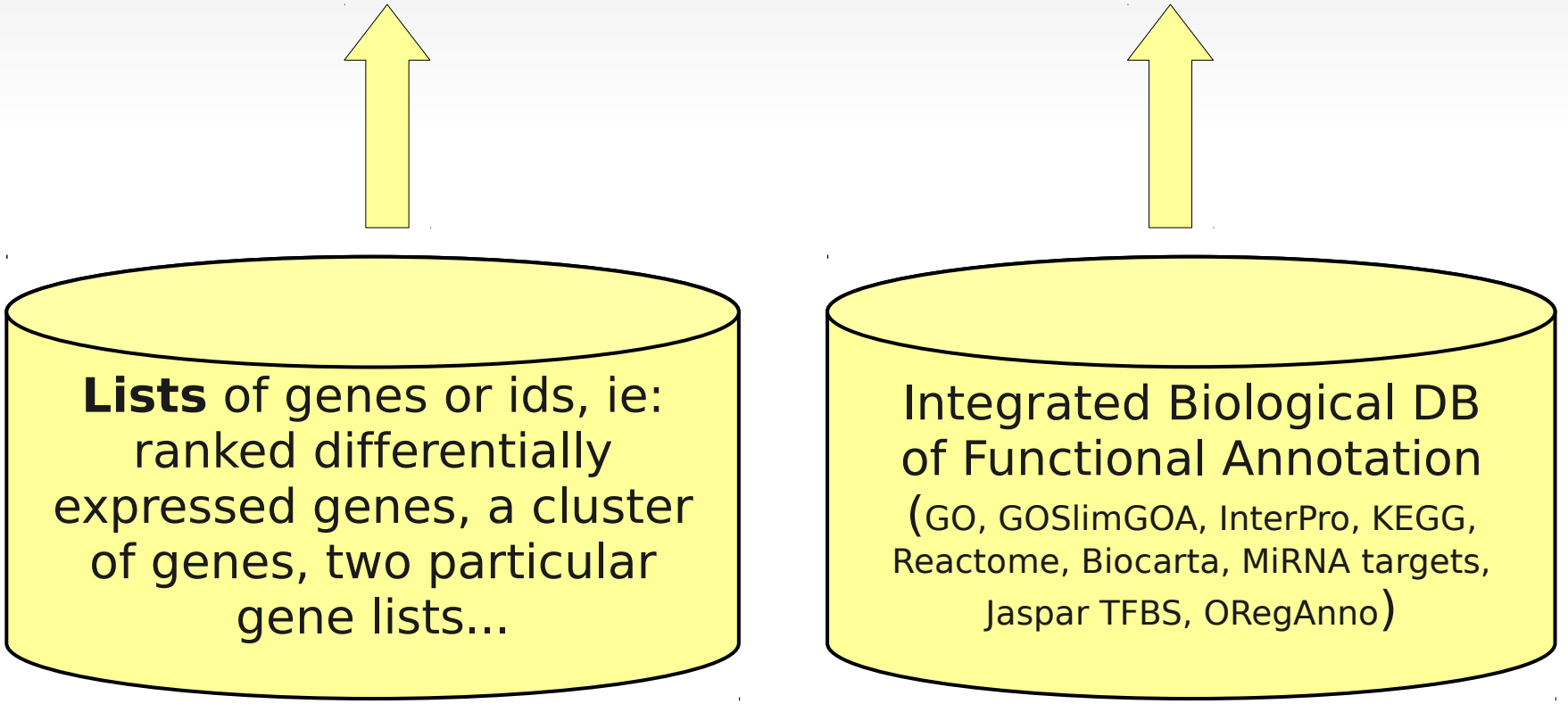
InterPro Motifs

Gene Expression in tissues

Bioentities from literature:
Diseases terms
Chemical terms

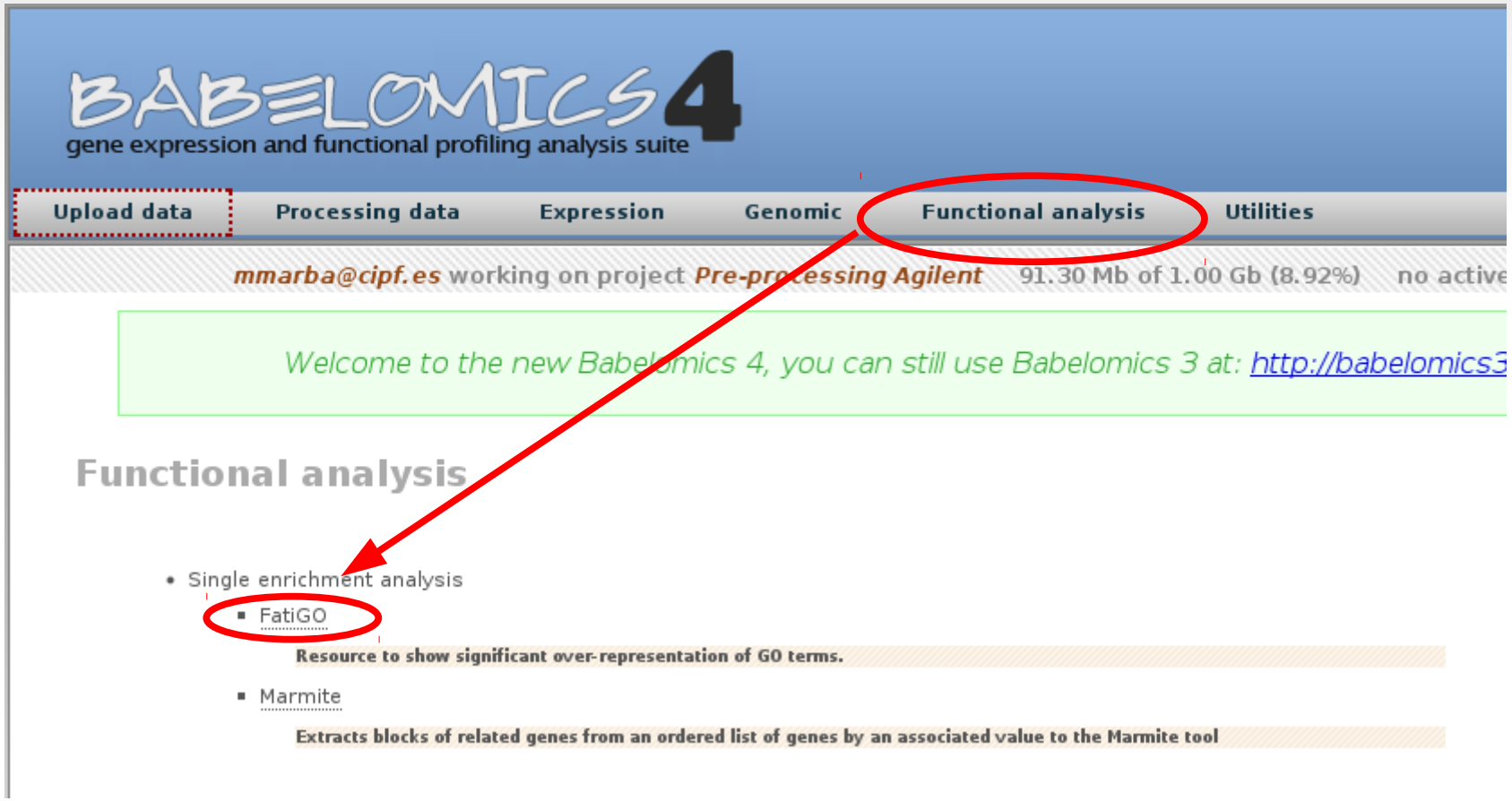
Functional interpretation
FatiGO, Fatiscan

FatiGO and Fatiscan are web tools for:
statistical test, multiple test corrections, filtering ...



FatiGO

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



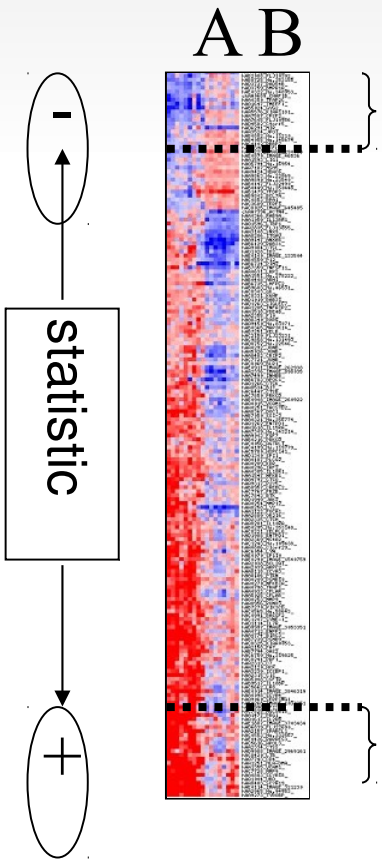
FatiGO features

- 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)

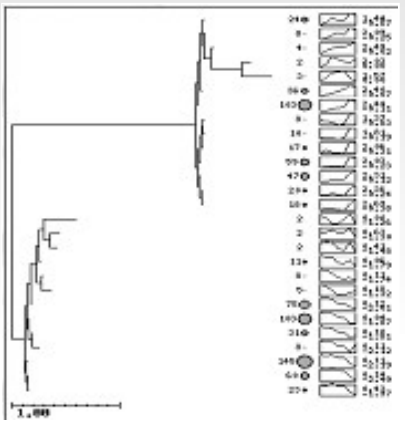
FatiGO features

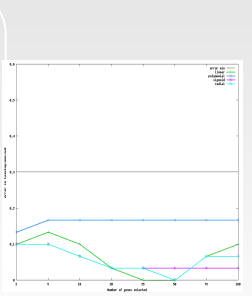
- It allows us to compare functional annotation of:
  - **Two** list of genes
  - **One** list against the rest of genome
  - Lists of genes with user **submitted annotations**
- One statistical test for each Functional **Block** of annotation
  - Fisher's exact test
  - Multiple testing context (hundreds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

Simple enrichment analysis



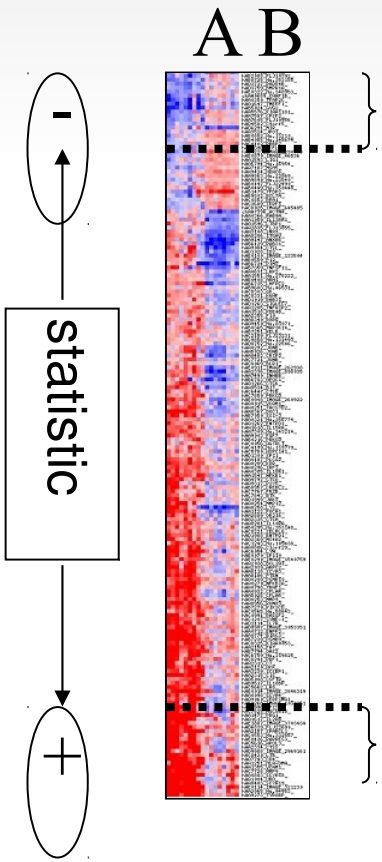
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.



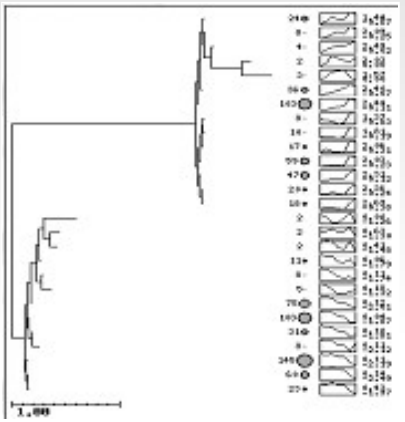


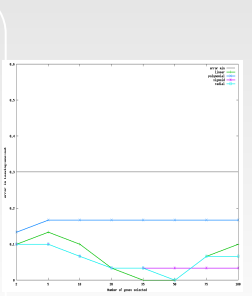
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

Simple enrichment analysis



1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.





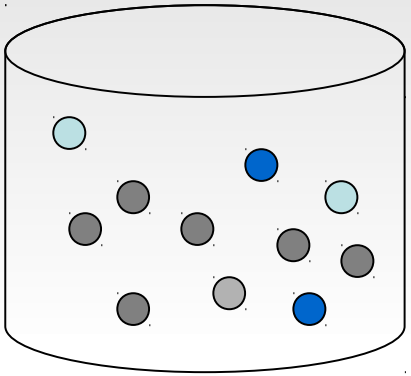
GO

4/7 ~ 2/11

1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

FatiGO test

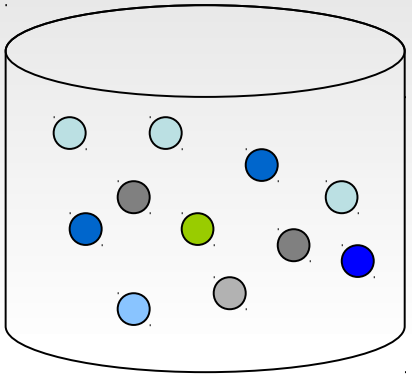
One Gene List (A)



Biosynthesis 60%

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

The other list (B)



Biosynthesis 20%

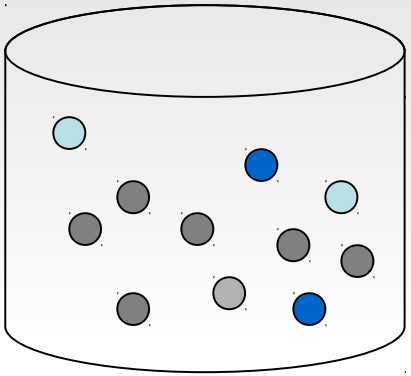
Genes in group A have significantly to do with **biosynthesis.**

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

p.val = 5.318e-09

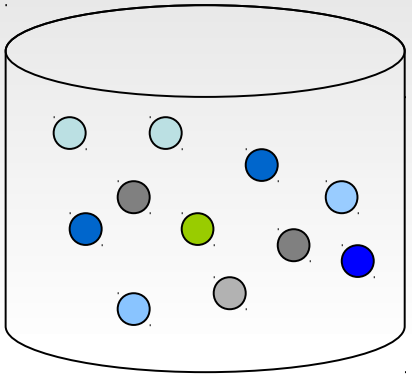
FatiGO test

One Gene List (A)

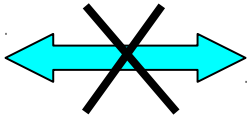


Are this two groups of genes carrying out different biological roles?

The other list (B)



Sporulation 20%



Sporulation 30%

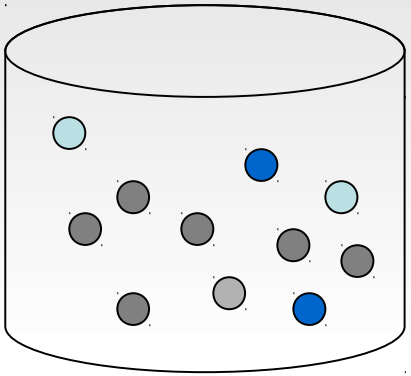
Genes in group A have significantly to do with biosynthesis, **but not with sporulation**.

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

p.val = 0.964

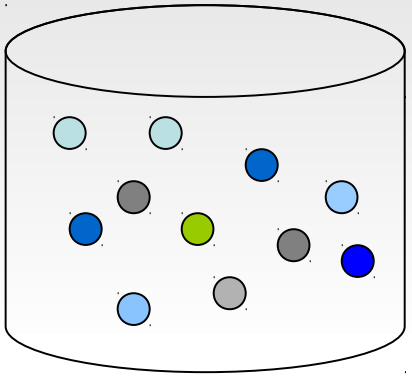
FatiGO test

One Gene List (A)



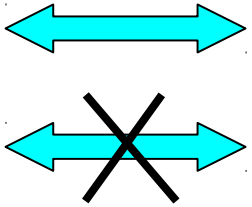
Are this two groups of genes carrying out different biological roles?

The other list (B)



Biosynthesis 60%

Sporulation 20%



Biosynthesis 20%

Sporulation 30%

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

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

We do this for each term (GO, miRNA, Interpro , …)!!!

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

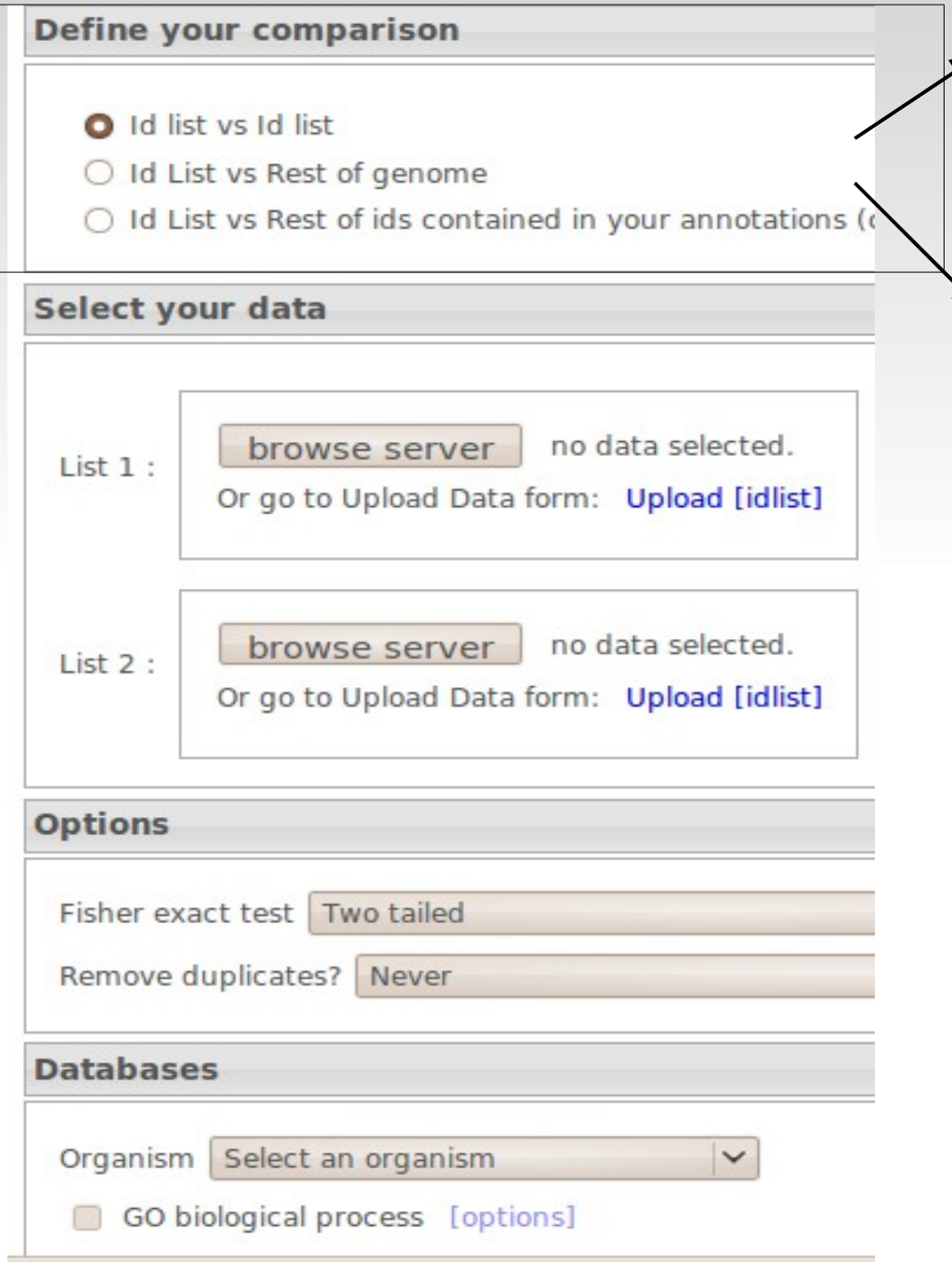
Databases

Organism: Select an organism

[ ] GO biological process [options]

FatiGO form

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

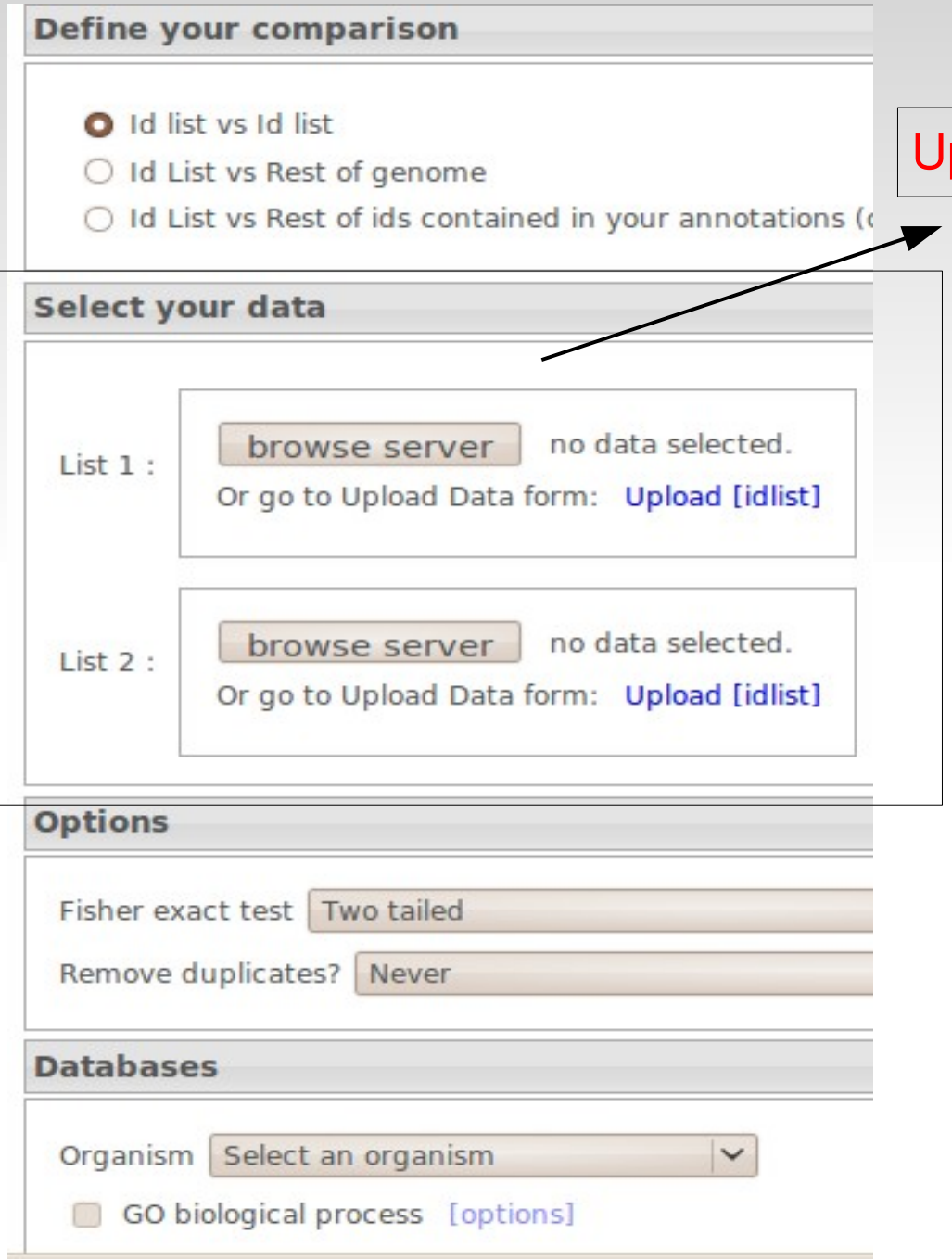


eg. Compare 2tissues or responder genes vs. non-responders

eg: genes that respond to one treatment against the genome

FatiGO form

Data selection

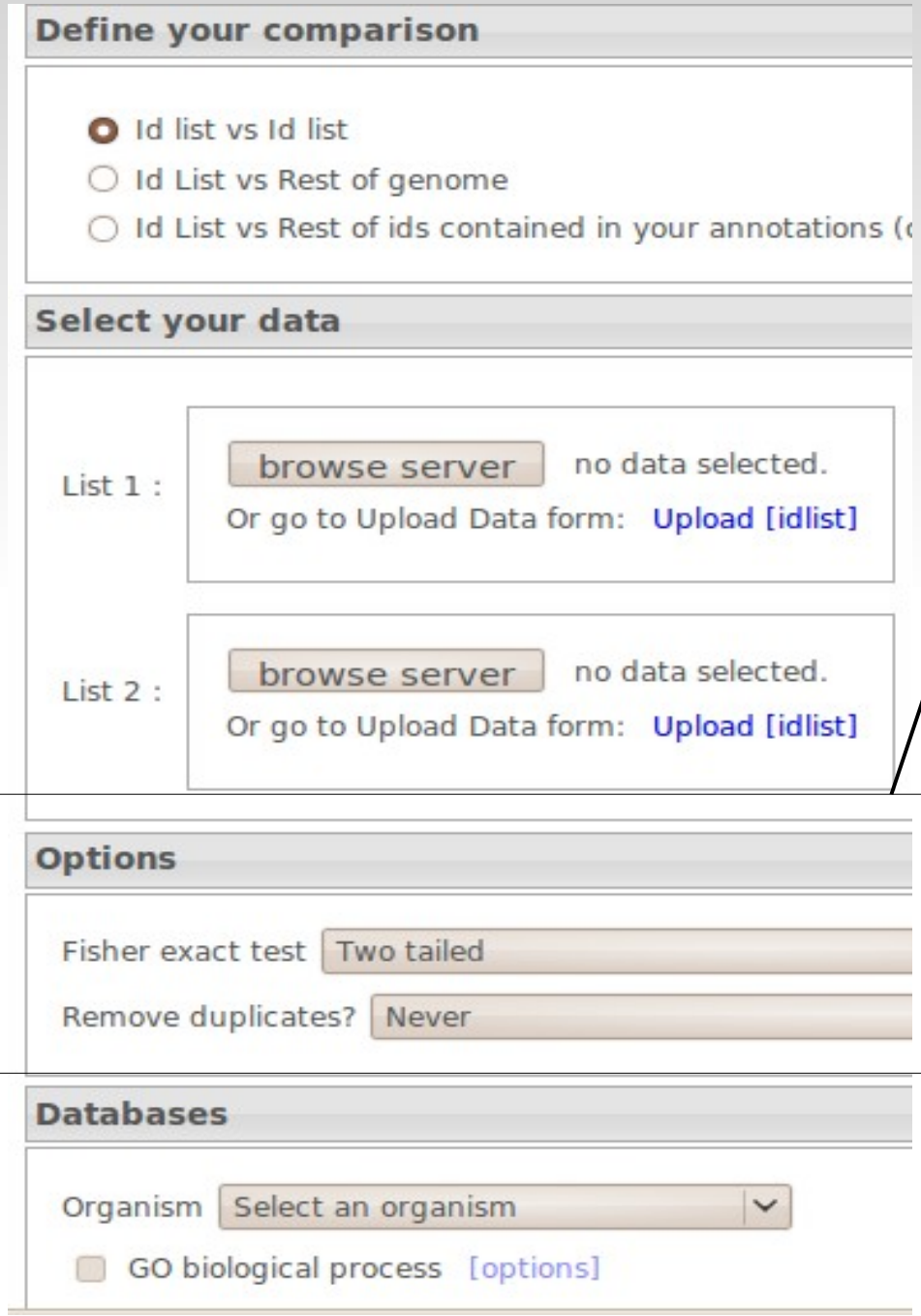


Upload first at Data Upload

"txt" file with gene lists:

gene1
gene2
gene3
...

FatiGO form



Algorithm options

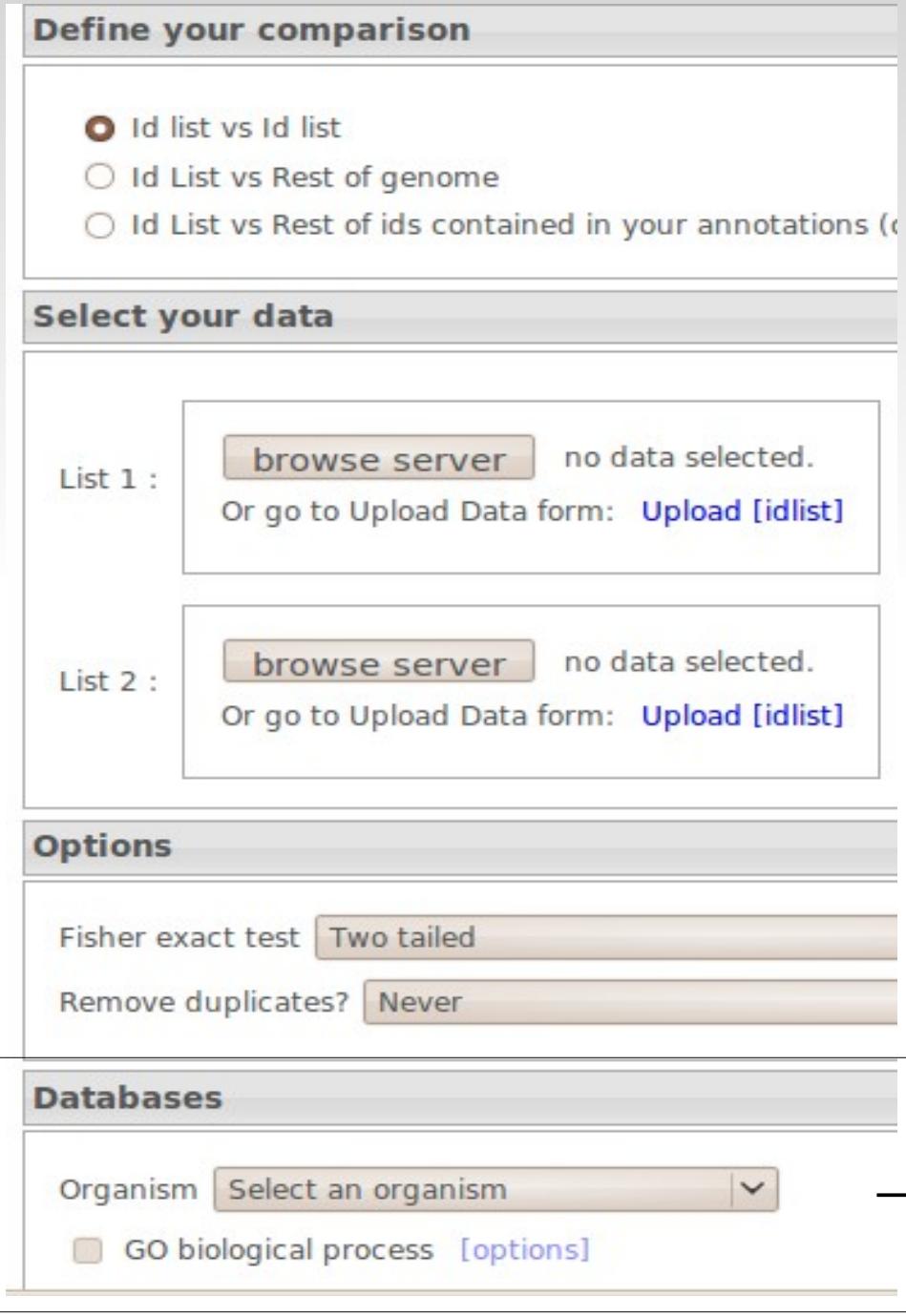
HINT:

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

Removing duplicates:

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

FatiGO form

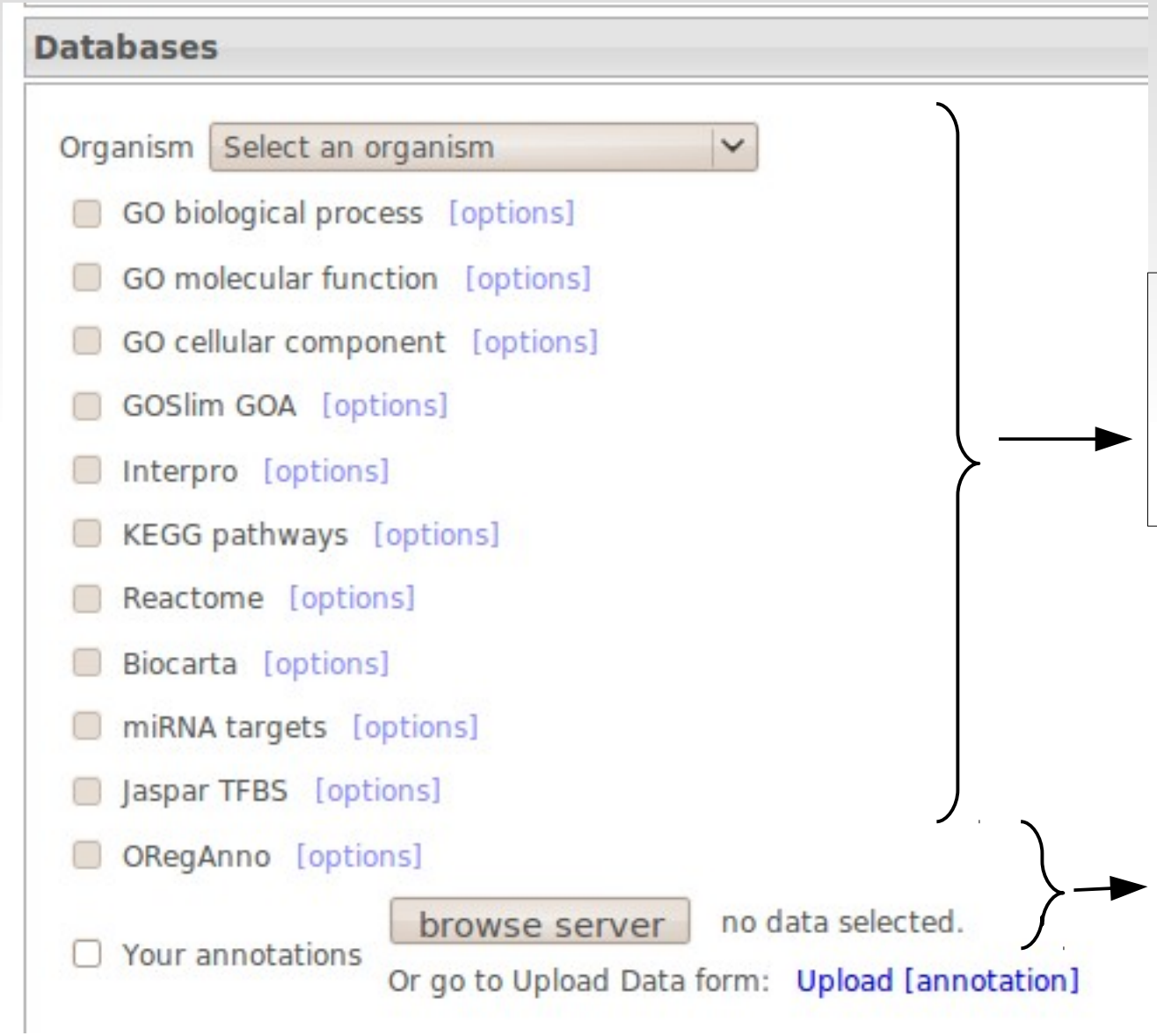


Which type of functional information?

?

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

A.

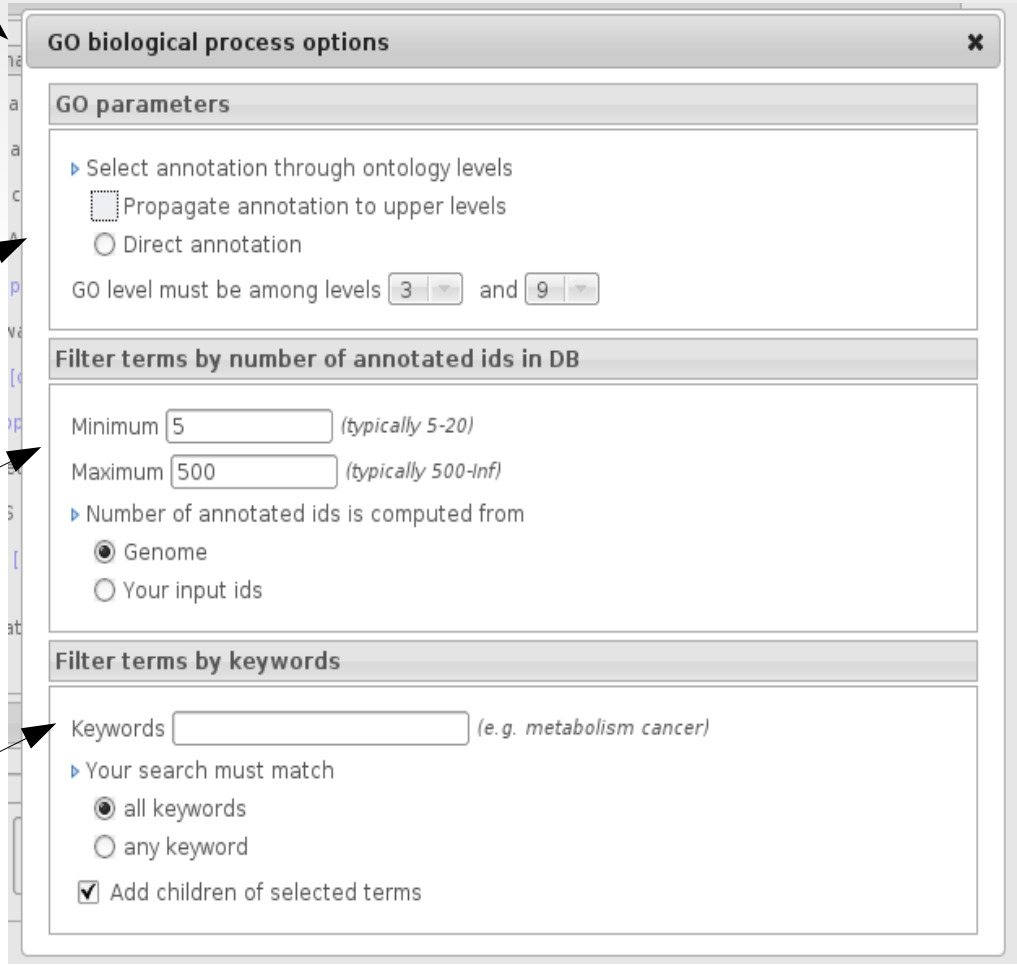
First select an organism

OPTIONS:

Test all the GO or only annotated terms

Discard functions with too few or too many genes?

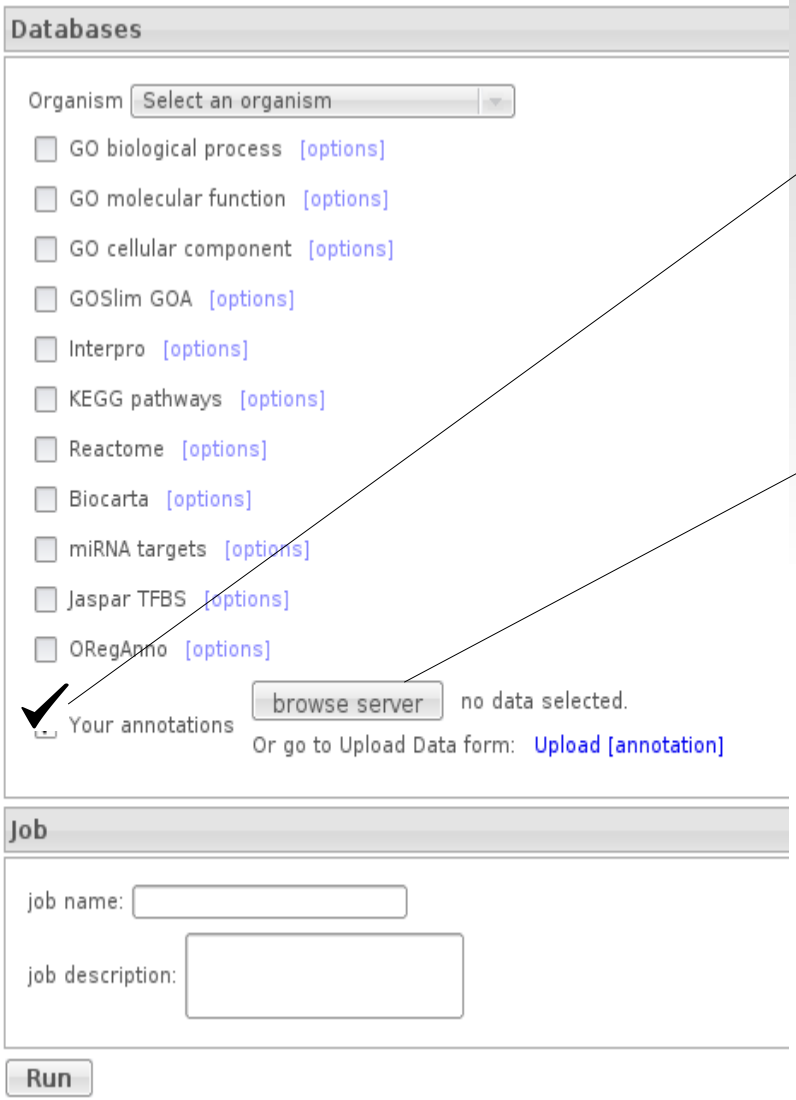
If you have an hypothesis, better test this first!!!!!!



FatiGO form

B.

Which type of functional information?



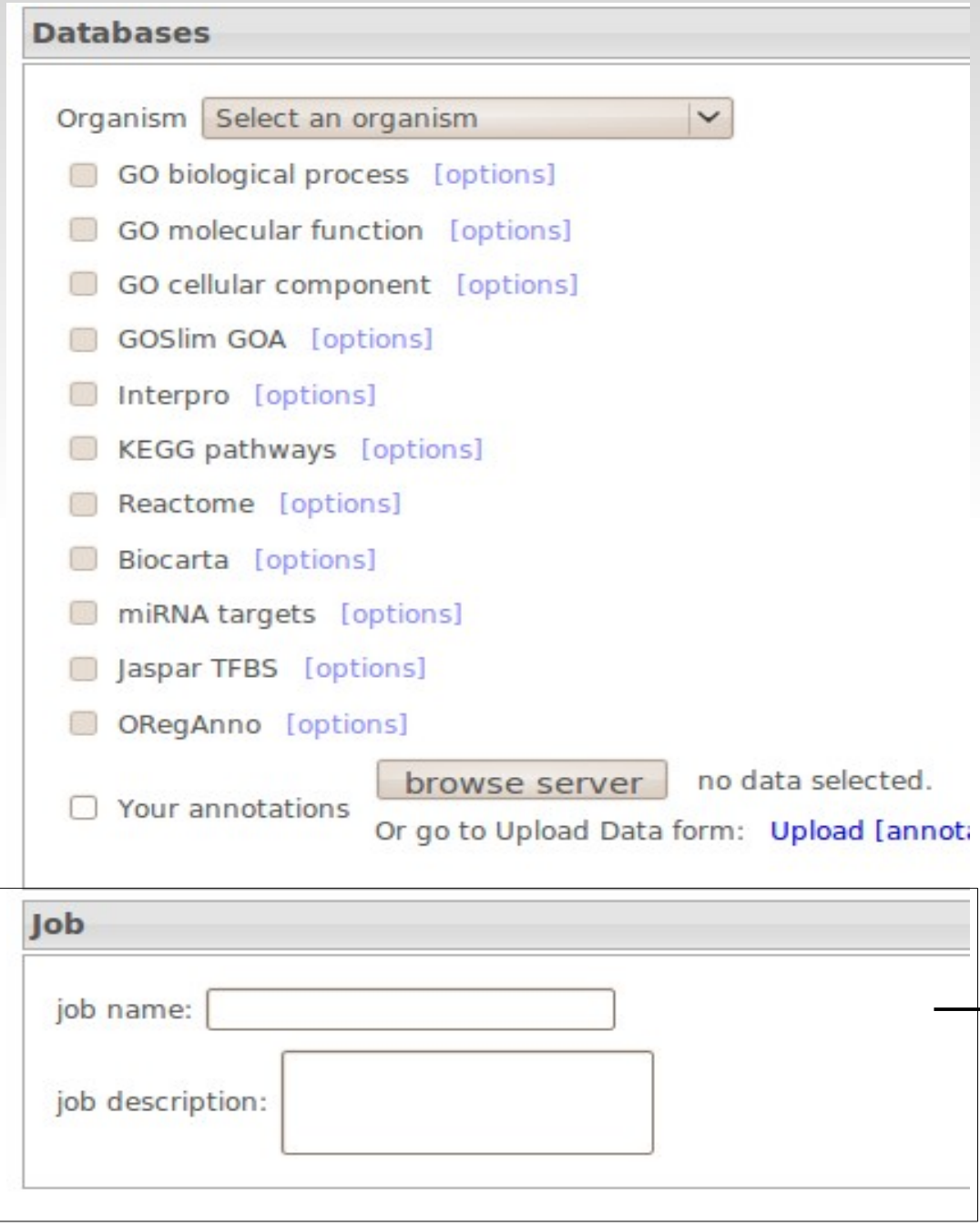
<u>Your annotations</u>: useful when you work with your own annotations OR with an organism that is not in Babelomics

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 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:

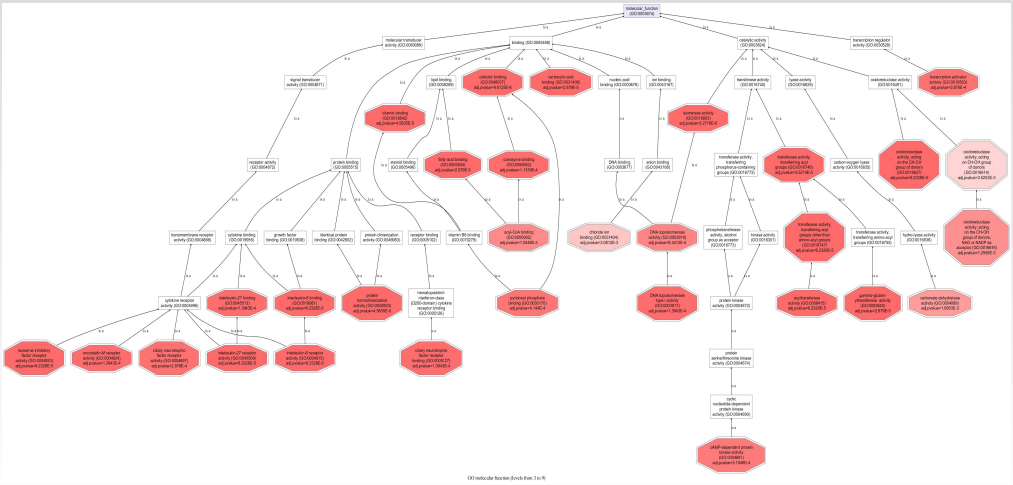
Term	Term size	Term size (in genome)	Term annotation % per list	Annotated ids	Odds ratio (log e)	pvalue	Adjusted pvalue
negative regulation of apoptosis (GO:0043066)	412	403	list 1: 7.2% list 2: 1.62%	list 1: 205225_at,20979... list 2: ENSG00000001084,ENSG...	1.5495	7.006e-13	7.65e-10
negative regulation 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:



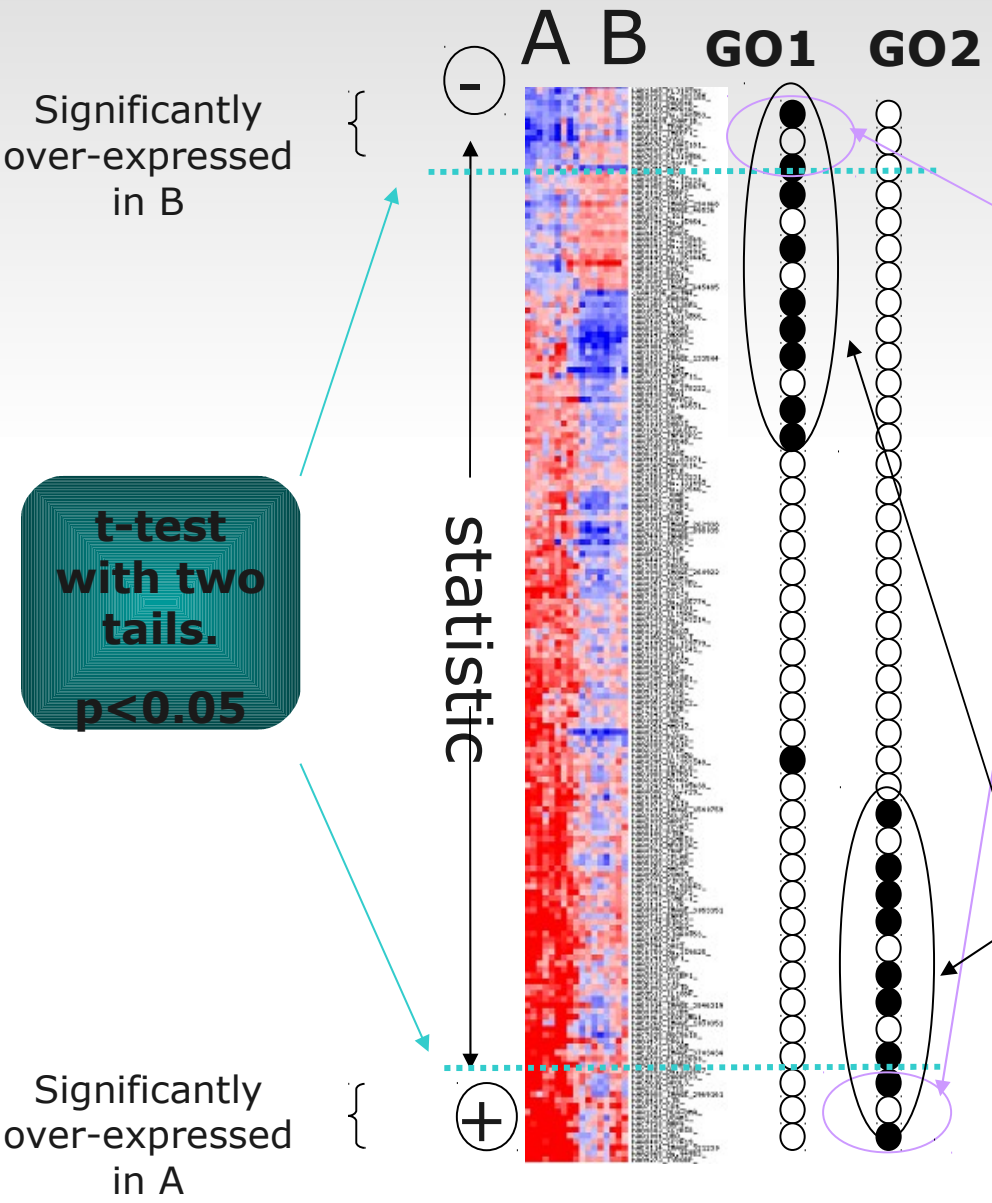
More results:

Other actions

- Open input form
- Change p value

0.1 *0.05* *0.01* *0.005*

FatiGO approach may not be very powerful

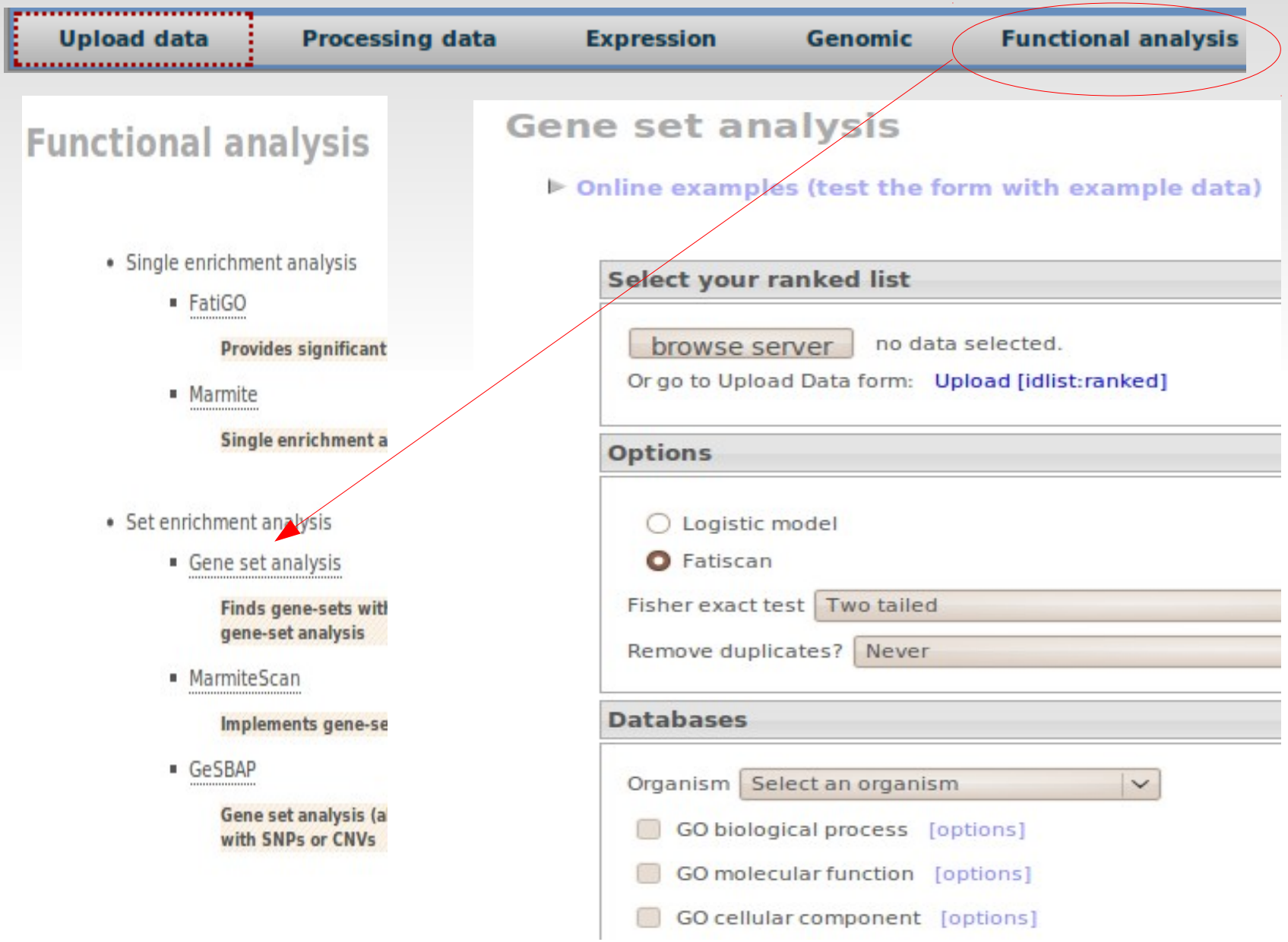


Very few genes selected to arrive to a significant conclusion on GO1 and GO2

If a threshold based on the experimental values is applied, and the resulting selection of genes compared for enrichment of a functional term, this might not be found

Functional Classes expressed as blocks in A and B

FatiScan


Upload data
Processing data
Expression
Genomic
Functional analysis
Functional analysis
Single enrichment analysis
FatiGO
Provides significant
Marmite
Single enrichment a
Set enrichment analysis
Gene set analysis
Finds gene-sets with
gene-set analysis
MarmiteScan
Implements gene-se
GeSBAP
Gene set analysis (a
with SNPs or CNVs
Gene set analysis
Online examples (test the form with example data)
Select your ranked list
browse server
no data selected.
Or go to Upload Data form:
Upload [idlist:ranked]
Options
Logistic model
Fatiscan
Fisher exact test
Two tailed
Remove duplicates?
Never
Databases
Organism
Select an organism
GO biological process
[options]
GO molecular function
[options]
GO cellular component
[options]

Fatiscan features

- Interpret a **ranked list of genes.**
- There is not need for choosing a cut-off. All information is included.

- 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)

Fatiscan features

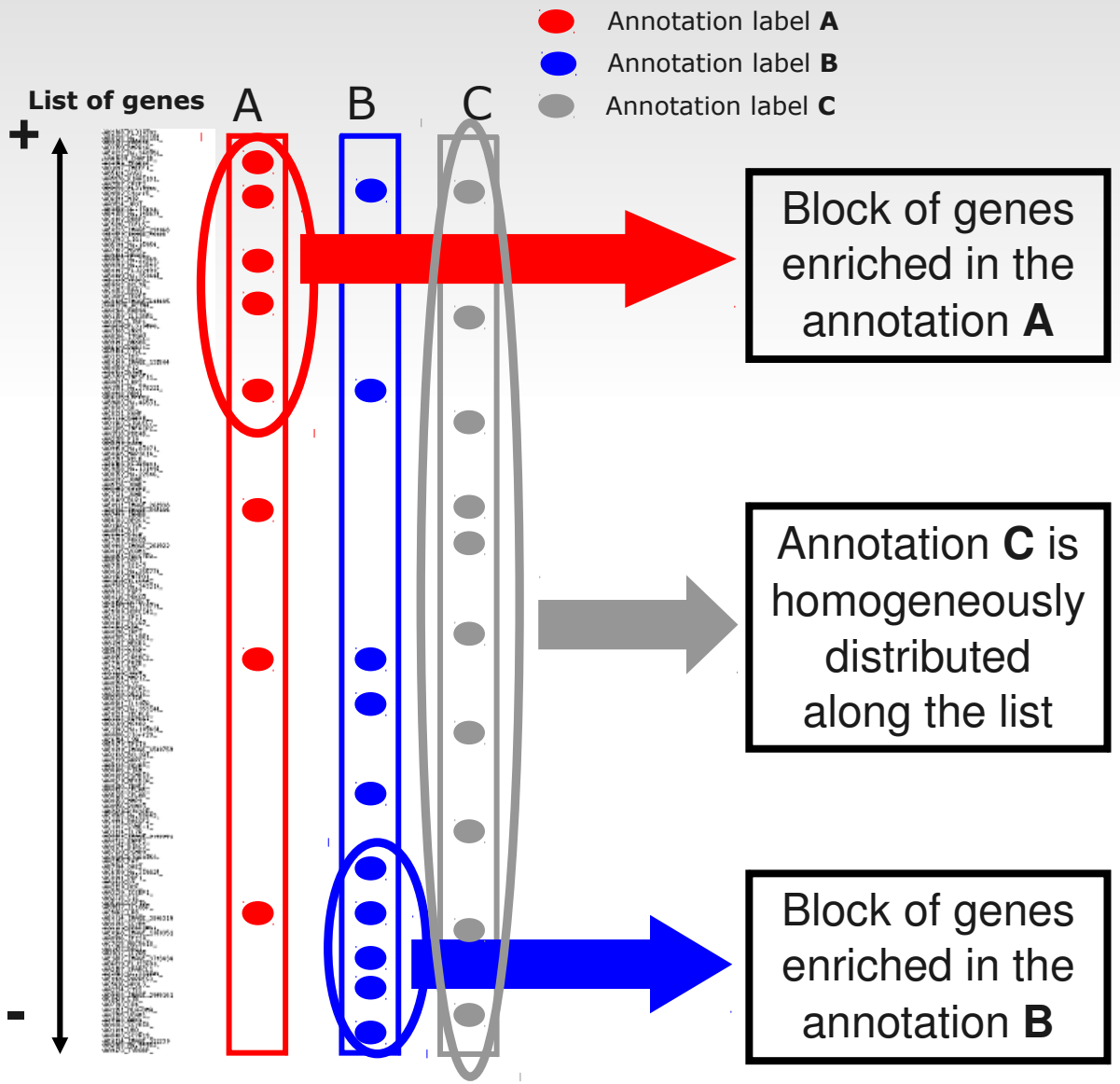
- Interpret a **ranked list of genes.**
- There is not need for choosing a cut-off. All information is included.

- 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)

FatiScan

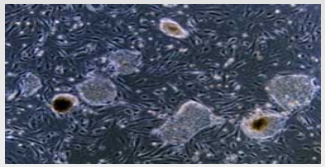
Testing along an ordered list

- Index ranking genes according to some biological aspect under study.
- Database that stores gene class membership information.
- **FatiScan** searches over the whole ordered list, trying to find runs of functionally related genes.

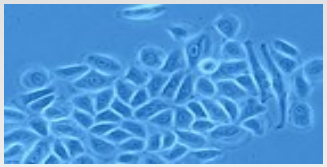


FatiScan Example – two classes

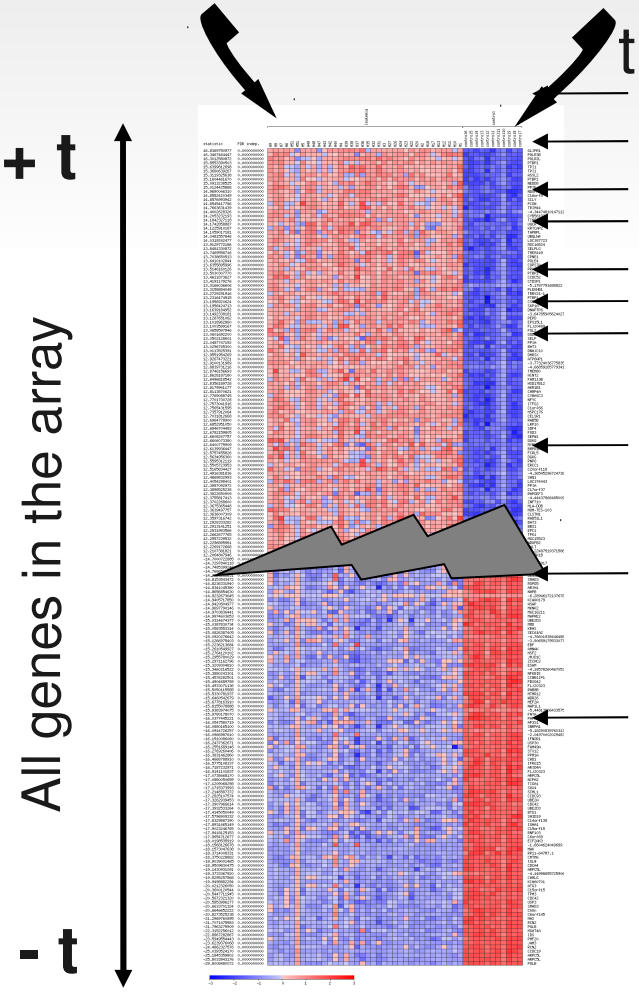
Tumor



Control



t ~ Tumor mean expression – Control mean expression

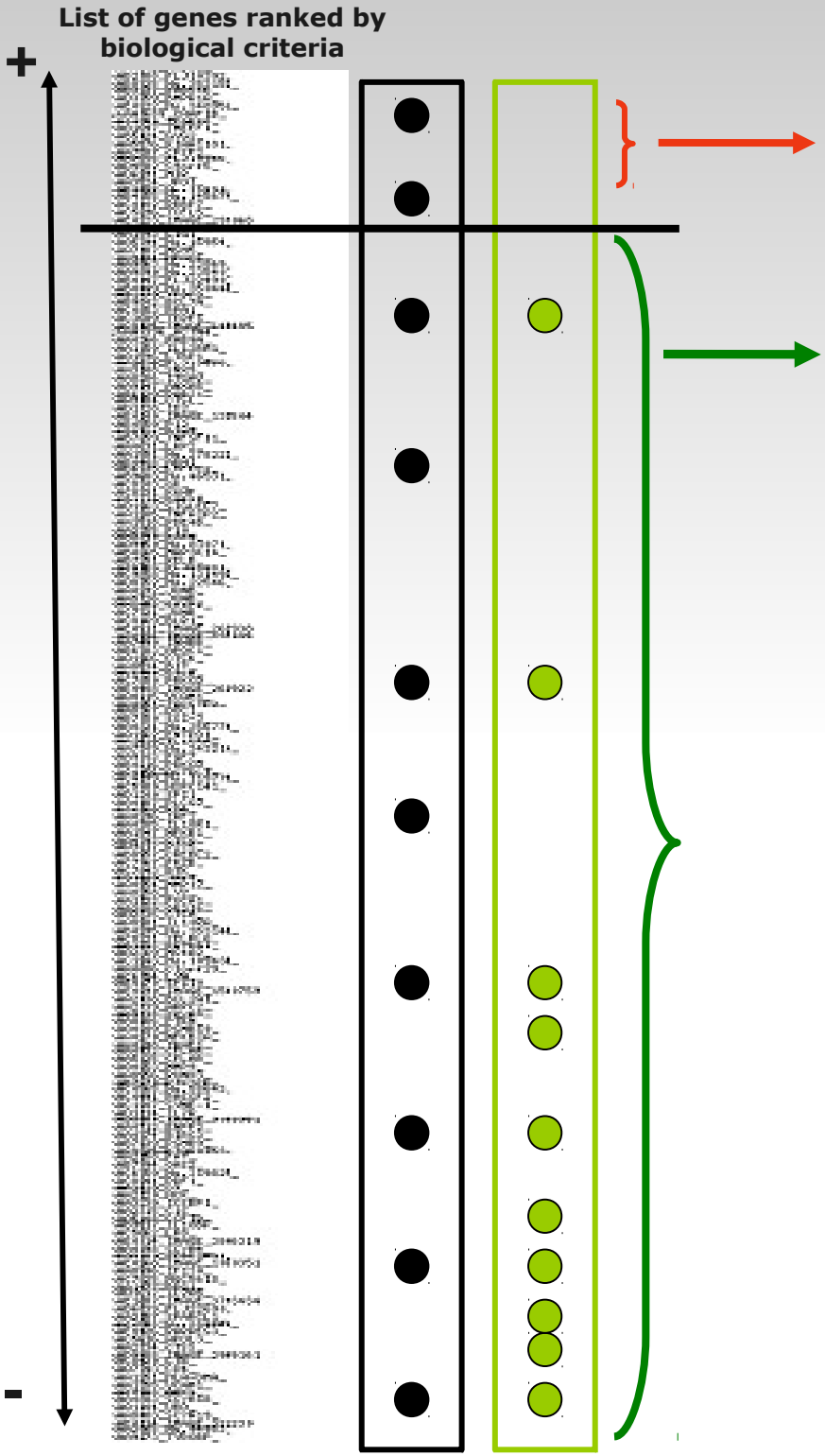


Proliferation

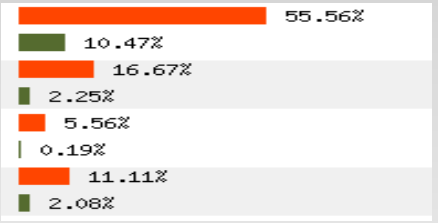
Is more associated with the genes on the top of the list

Is more associated with the genes that show higher expression in Tumors

Scanning test using partitions

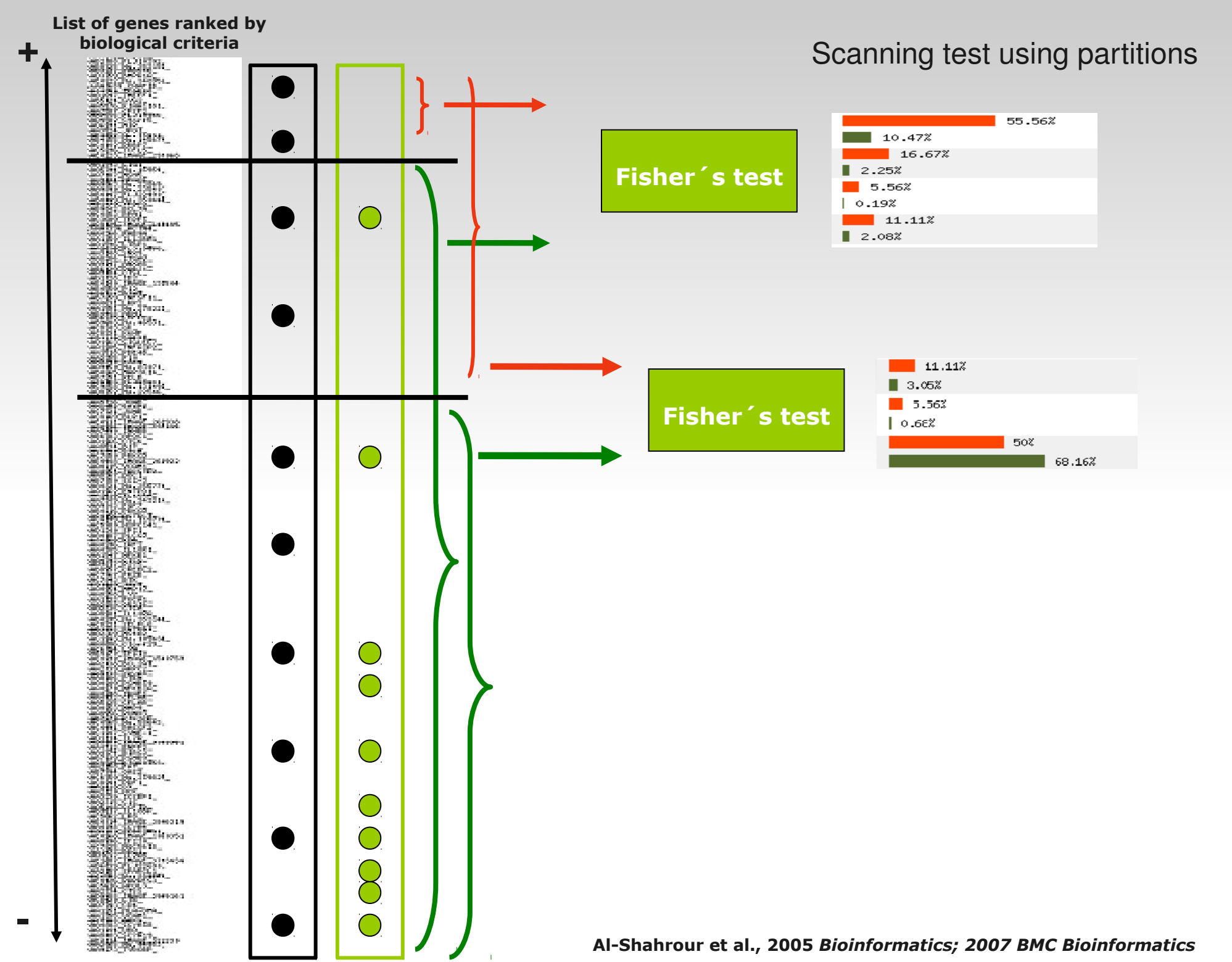


Fisher´s test



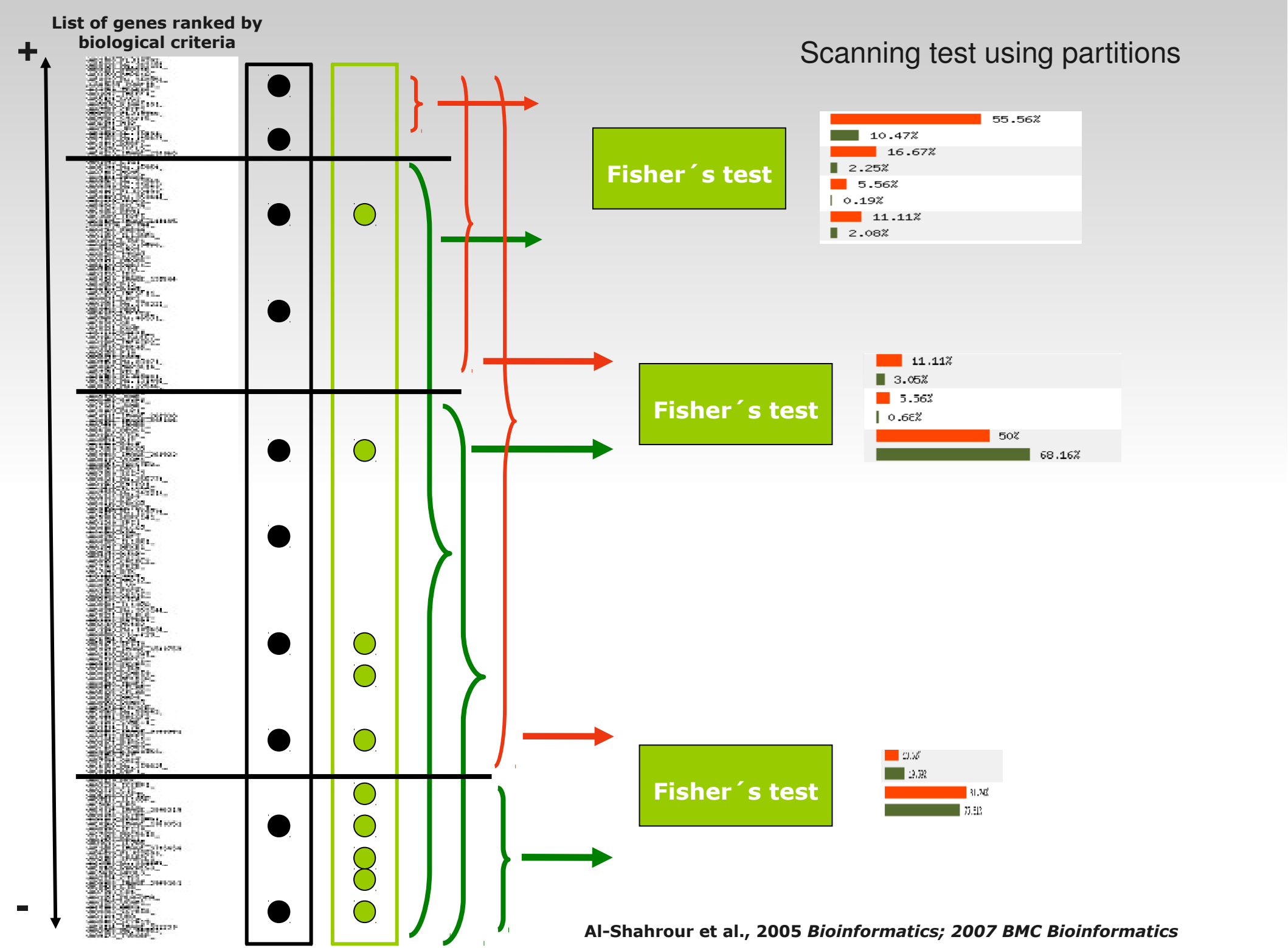
Al-Shahrour et al., 2005 *Bioinformatics; 2007 BMC Bioinformatics*

Scanning test using partitions



Al-Shahrour et al., 2005 ***Bioinformatics; 2007 BMC Bioinformatics***

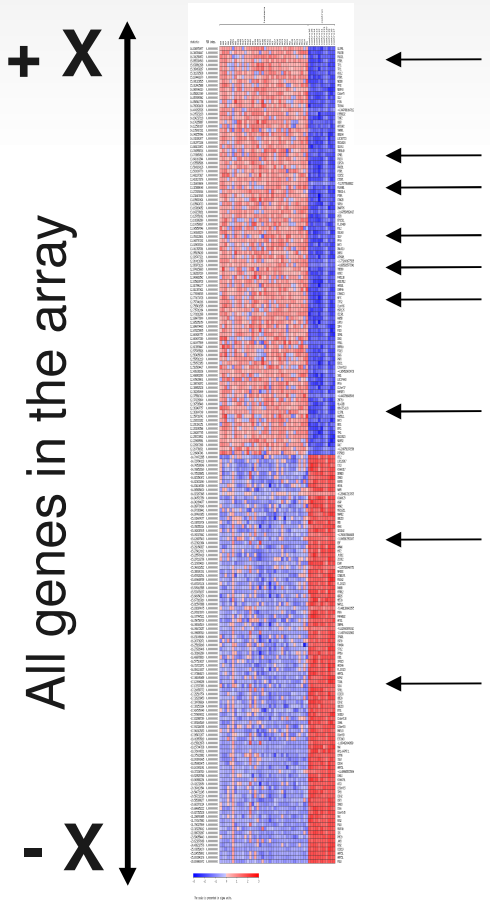
Scanning test using partitions



Al-Shahrour et al., 2005 *Bioinformatics; 2007 BMC Bioinformatics*

New test

- Not using partitions
- But logistic regression model



$$\ln\left(\frac{P(g\in F)}{P(g\notin F)}\right) = K + \alpha X$$

alpha > 0 : **increasing** X increases the probability of the gen to be annotated

alpha < 0 : **decreasing** X increases the probability of the gen to be annotated

Babelomics Tools

***MARMITE*:** Finds differential distributions of bioentities extracted from **PubMed** between two groups of genes.

MarmiteScan: Use chemical and disease-related information to detect related blocks of genes in a gene list with associated values.

Annotation is not 0, 1 may take any value.

Number of articles in which a gene is associated to a:

- Chemical product
- Disease associated
- Drug
- Gene
- Symptom

Take Home Message

- The unit of information over which we test is shifted from genes to functional blocks
- We do one statistical test for each block (Multiple testing)
- All genes in the block are treated equally
- Genes independently may not show a strong pattern of association but the block coordinately does