Babelomics

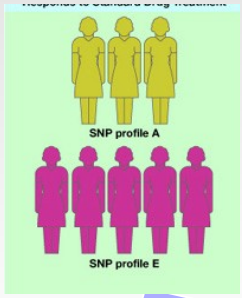
Experimental Data Analysis

Jun 2014 in Cambridge

David Montaner
dmontaner@cipf.es
http://bioinfo.cipf.es/dmontaner
Bioinformatics and Genomics Department
Centro de Investigacion Principe Felipe (CIPF)
(Valencia, Spain)

Genetic Research

Genes in the DNA...



…code for proteins...

…produces the final phenotype

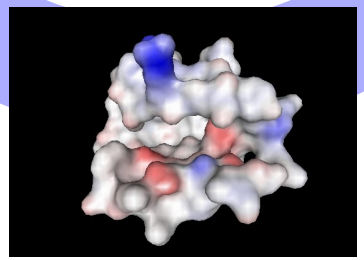
>protein kinase

acctgttgatggcgacagggactgtatgctgatctatgctgatgcatgcatgctgactactgatgtgggggctattgacttgatgtctatc....

From genotype to phenotype.

…whose structure accounts for function...

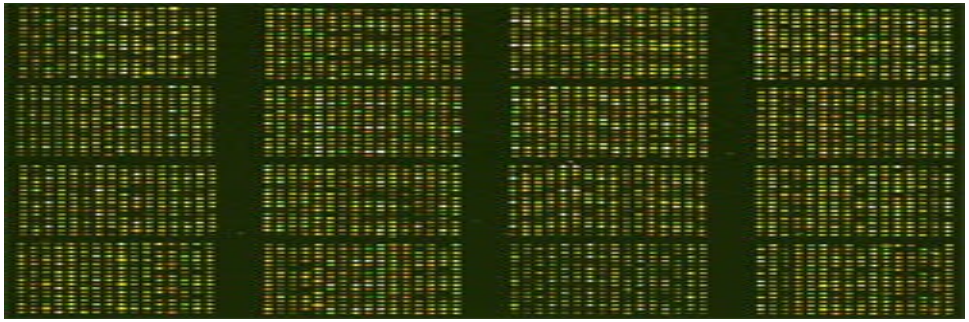
…plus the environment...

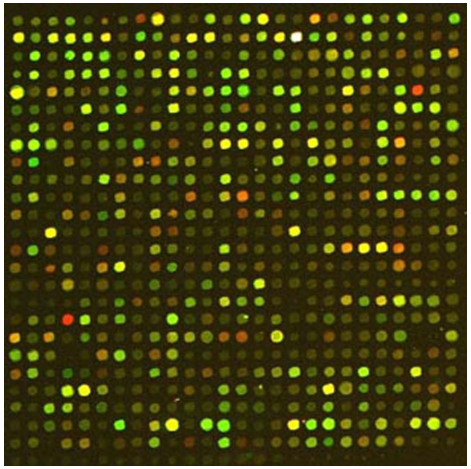


Data is information

High Throughput Technologies

- 1988 arrayed DNAs were used
- 1991 oligonucleotides are synthesized on a glass slide through photolithography
- 1995 DNA Microarrays
- 1997 Genome wide Yeast Microarray
- 2005 First next-generation sequencing system



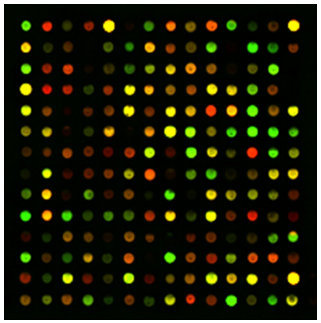


Nature: Milestones DNA Technologies

High Throughput Technologies

The road of excess leads to the palace of wisdom.

William Blake
Proverbs of Hell
(1790–1793)



Next Generation Sequencing
SOLID 6Gbp per round

Genes in the DNA...

>protein kinase

acctgttgatggcgacagggactgtatgctgatc
tatgctgatgcatgcatgctgactactgatgtggg
ggctattgacttgatgtctatc....

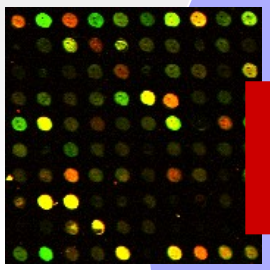
…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype...

…when expressed in the proper moment and place...

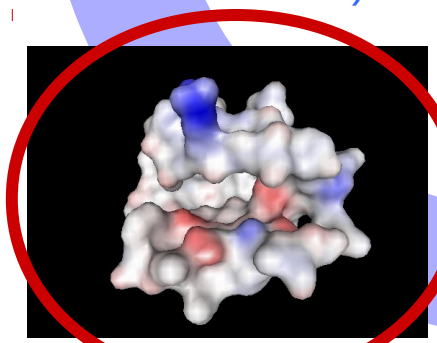
A typical tissue is expressing among 5000 and 10000 genes



Data ≠ information

(in the functional post-genomic scenario)

…code for proteins...

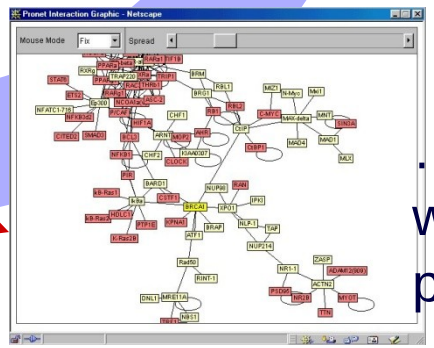


That undergo post-translational modifications, somatic recombination...

100K-500K proteins

…whose structures account for function...

…conforming complex interaction networks...



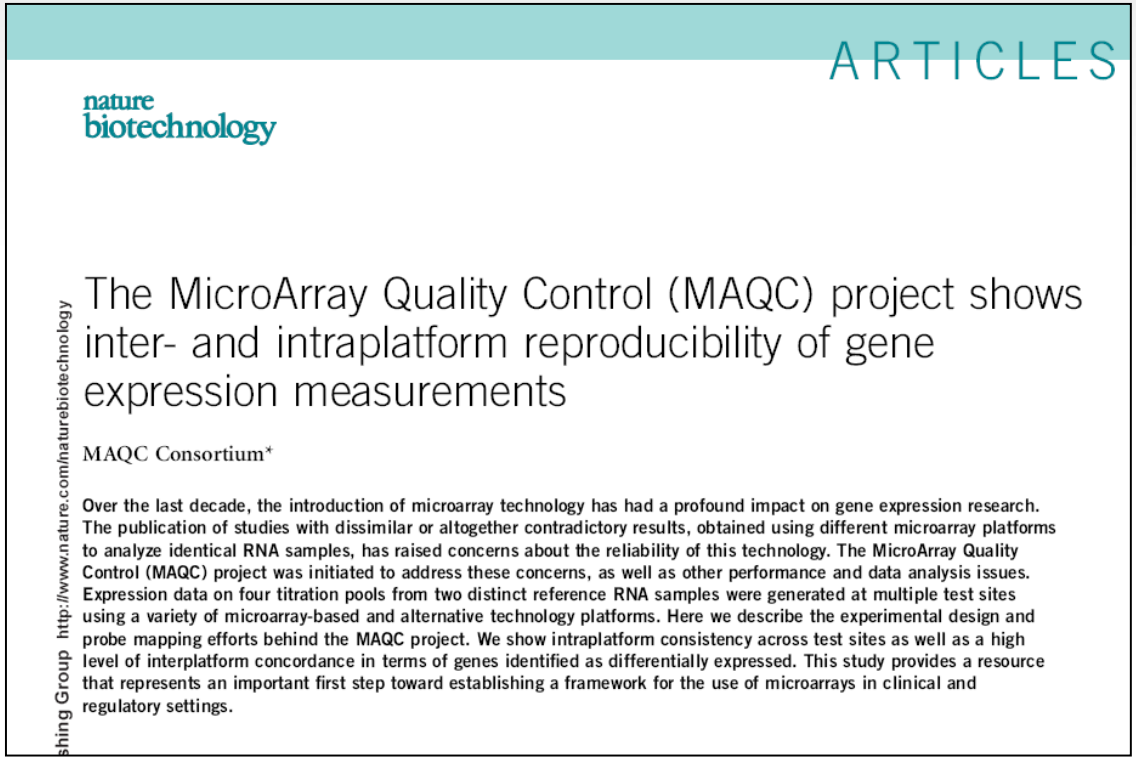
…in cooperation with other proteins…

Each protein has an average of 8 interactions

Bioinformatics

- 2003 GEPAS
- Bioinformatics department open at CIPF
- 2006 BABELOMICS
- 2008 Blast2GO
- 2010 GEPAS and some characteristics of Blast2GO are included into Babelomics.

Microarrays arrive to an acceptable level of reproducibility



ARTICLES

nature biotechnology

The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements

MAQC Consortium*

Over the last decade, the introduction of microarray technology has had a profound impact on gene expression research. The publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms to analyze identical RNA samples, has raised concerns about the reliability of this technology. The MicroArray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues. Expression data on four titration pools from two distinct reference RNA samples were generated at multiple test sites using a variety of microarray-based and alternative technology platforms. Here we describe the experimental design and probe mapping efforts behind the MAQC project. We show intraplatform consistency across test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed. This study provides a resource that represents an important first step toward establishing a framework for the use of microarrays in clinical and regulatory settings.

shing Group http://www.nature.com/naturebiotechnology

FDA approves the first predictor based on microarrays

FDA Clears Breast Cancer Specific Molecular Prognostic Test - Microsoft Internet Explorer proporcionado por CNIO

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://www.fda.gov/bbs/topics/NEWS/2007/NEW01555.html Ir Vínculos

FDA U.S. Food and Drug Administration U.S. Department of Health and Human Services

FDA Home Page | Search FDA Site | FDA A-Z Index | Contact FDA

FDA News

FOR IMMEDIATE RELEASE
P07-13
February 6, 2007

Media Inquiries:
Karen Riley, 301-827-6242
Consumer Inquiries:
888-INFO-FDA

FDA Clears Breast Cancer Specific Molecular Prognostic Test

The U.S. Food and Drug Administration (FDA) today cleared for marketing a test that determines the likelihood of breast cancer returning within five to 10 years after a woman's initial cancer. It is the first cleared molecular test that profiles genetic activity.

The MammaPrint test uses the latest in molecular technology to predict whether existing cancer will metastasize (spread to other parts of a patient's body). The test relies on microarray analysis, a powerful tool for simultaneously studying the patterns of behavior of large numbers of genes in biological specimens.

The recurrence of cancer is partly dependent on the activation and suppression of certain genes located in the tumor. Prognostic tests like the MammaPrint can measure the activity of these genes, and thus help physicians understand their patients' odds of the cancer spreading.

MammaPrint was developed by Agendia, a laboratory located in Amsterdam, Netherlands, where the product has been on the market since 2005.

"Clearance of the MammaPrint test marks a step forward in the initiative to bring molecular-based medicine into current practice," said Andrew C. von Eschenbach, M.D., Commissioner of Food and Drugs. "MammaPrint results will provide patients and physicians with more information about the prospects for the outcome of the disease. This information will support treatment decisions.

Agendia compared the genetic profiles of a large number of women suffering from breast cancer and identified a set of 70 genes whose activity confers information about the likelihood of tumor recurrence. The MammaPrint test measures the level of activity of each of these genes in a sample of a woman's surgically removed breast cancer tumor, then uses a specific formula, known as an algorithm, to produce a score that determines whether the patient is deemed low risk or high risk for spread of the cancer to another site. The result may help a doctor in planning appropriate follow-up for a patient when used with other clinical information and laboratory tests.

The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device. Several months ago, FDA issued a draft guidance document concerning the need for these complex molecular tests to meet pre-market review and post-market device requirements even when the tests are developed and used by a single laboratory. Although FDA regulates diagnostic tests sold to laboratories, hospitals and physicians, it uses discretion when regulating tests developed and performed by single laboratories.

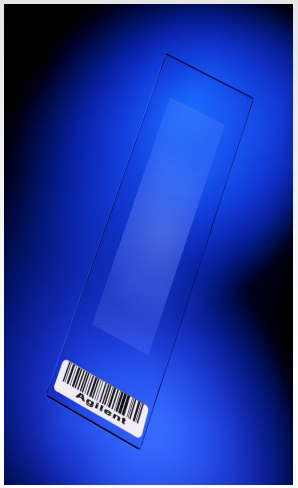
On February 8, FDA will hold a public meeting to discuss its draft guidance document describing its regulatory approach to this type of test.

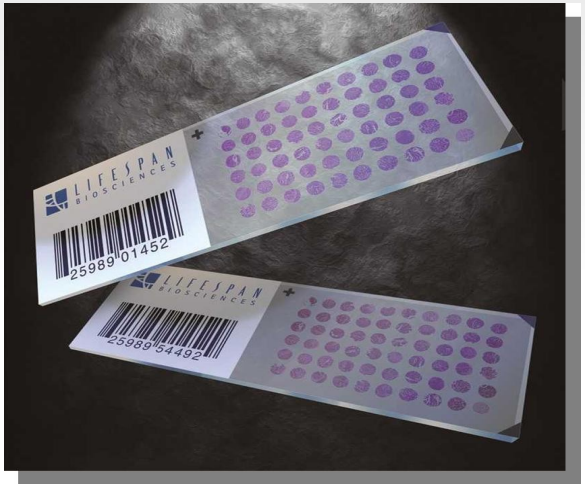
"There have been rapid advances in microarrays and other pioneering diagnostics, and a corresponding increase in the use and impact of these complex tests. This

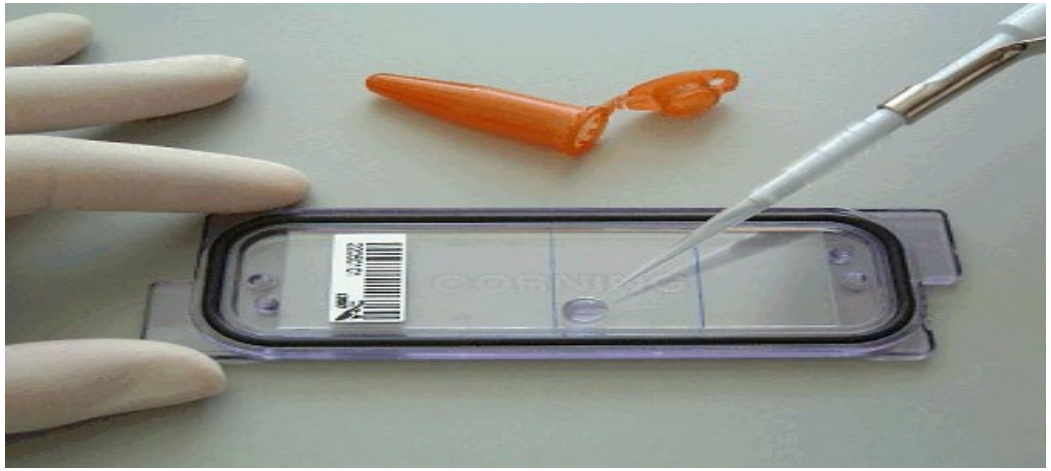
Listo Internet

February 2007: MamaPrint

DNA Microarrays: the paradigm of a post-genomic technique



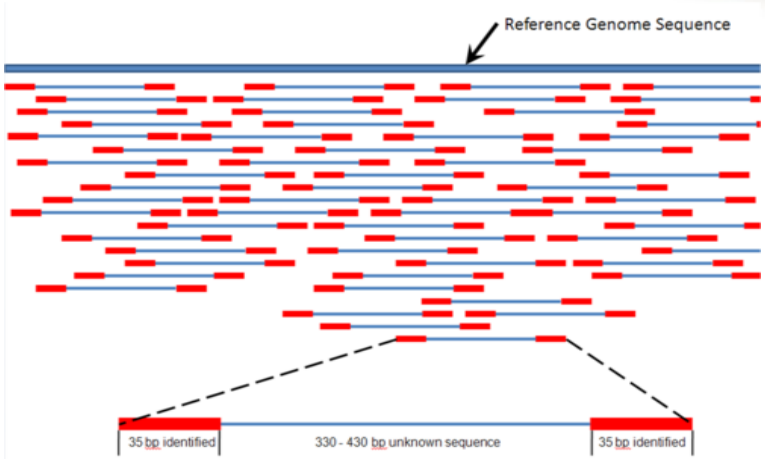


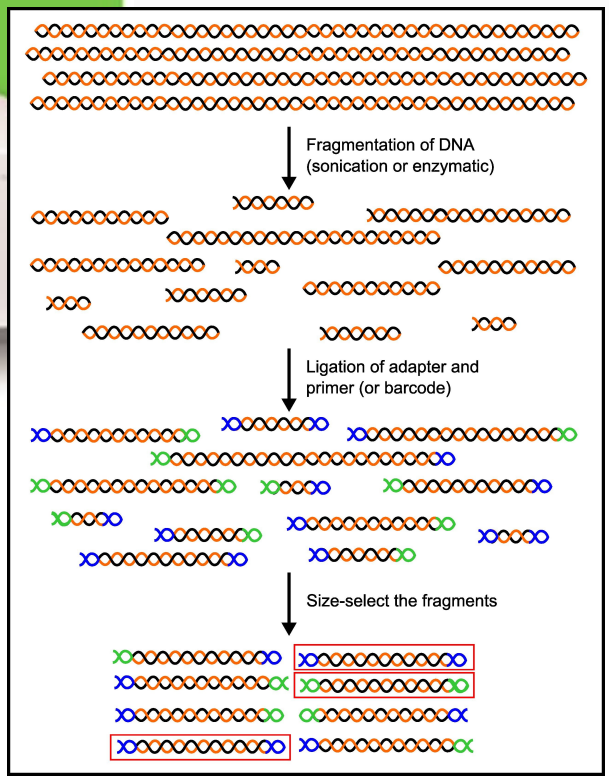


gene1	10.23	9.98	10.41	10.55	10.65	9.69
gene2	10.51	9.74	10.65	10.63	10.43	10.35
gene3	9.89	10.02	9.89	11.03	10.21	10.77
gene4	10.25	10.83	8.94	10.16	10.49	10.46
gene...	.....	.....	.....	.....	.....	.....

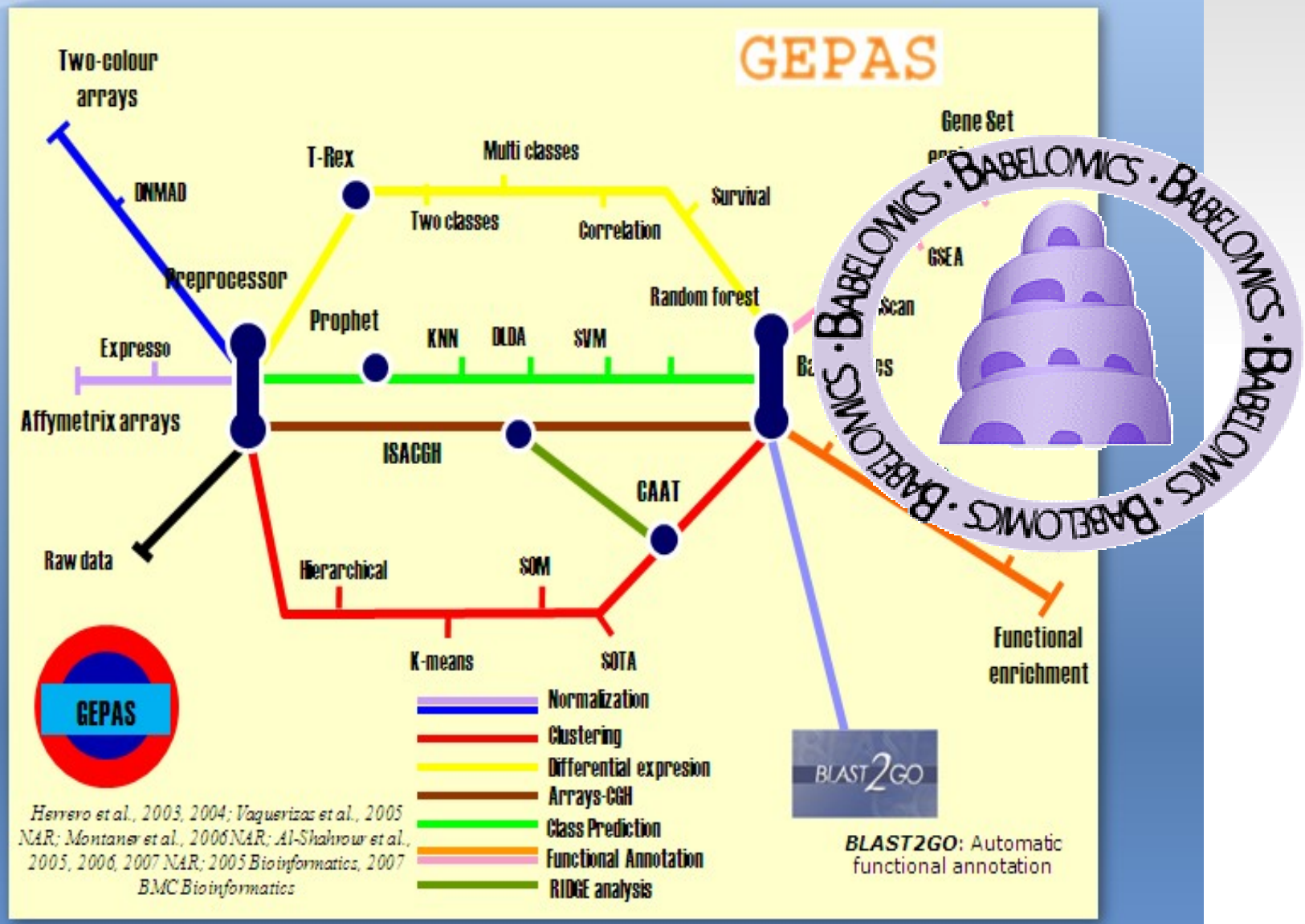
NGS


Next Generation Sequencing (NGS)
SOLiD v4
HiSeq2000
GAIIx
GS-FLX


Reference Genome Sequence
35 bp identified
330 - 430 bp unknown sequence
35 bp identified


Fragmentation of DNA
(sonication or enzymatic)
Ligation of adapter and
primer (or barcode)
Size-select the fragments

Functional interpretation



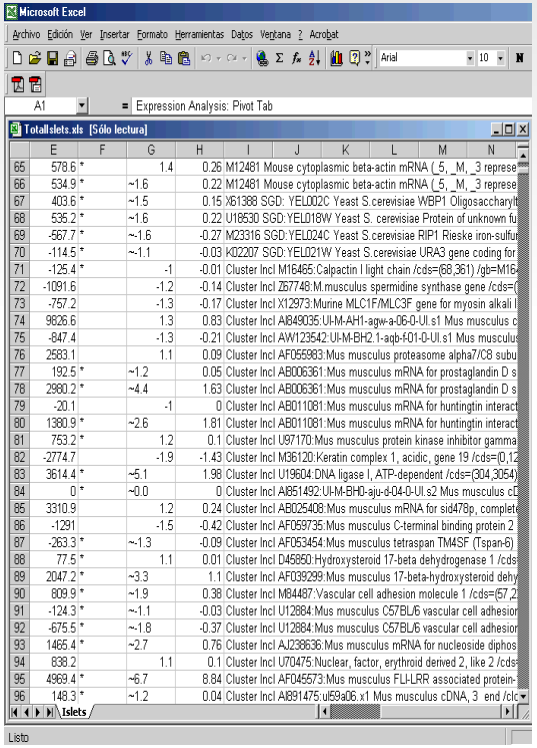
Functional profiling of genome-scale experiments in the post-genomic era

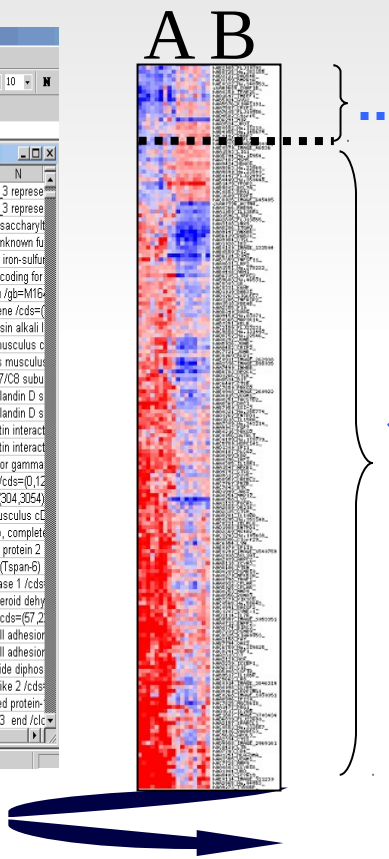
My data...

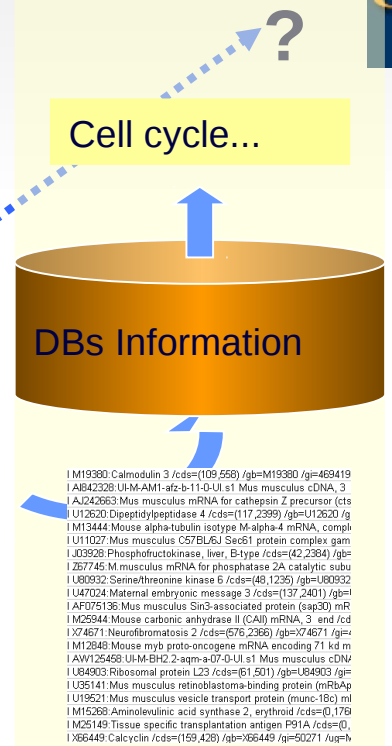
How are structured?

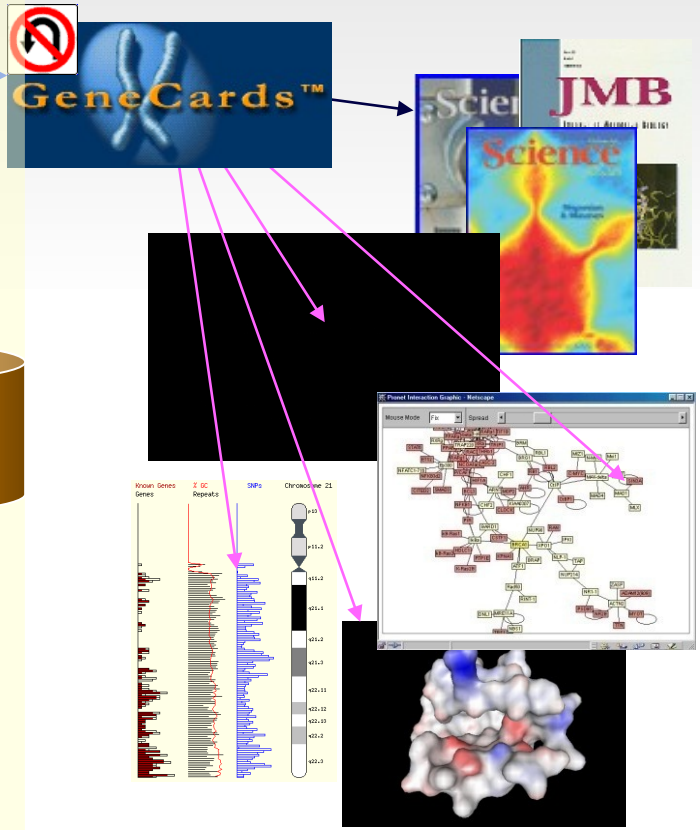
What are these groups?

What is this gen?



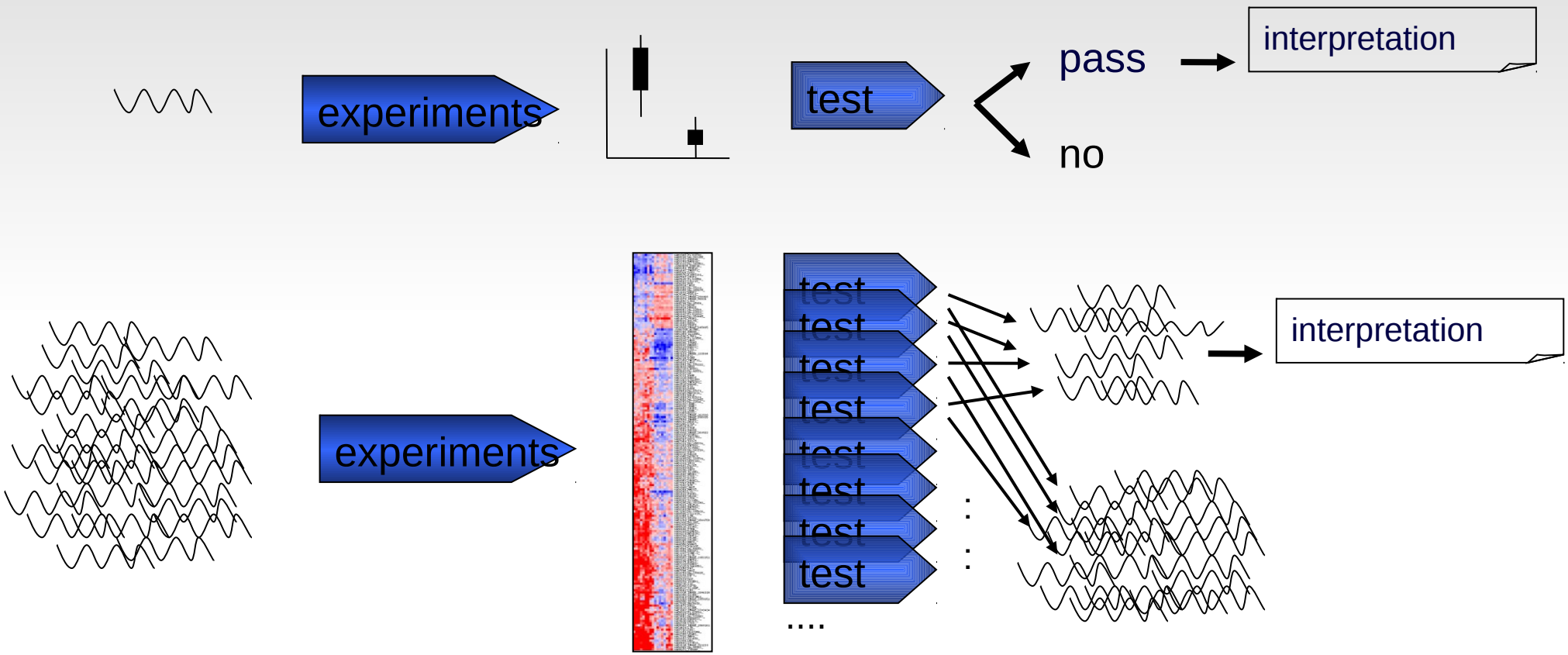






Analysis *Functional profiling* *Links*

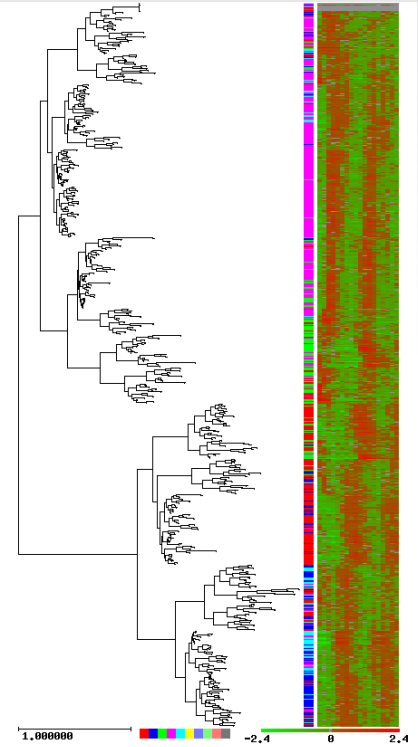
Functional enrichment approach reproduces pre-genomics paradigms

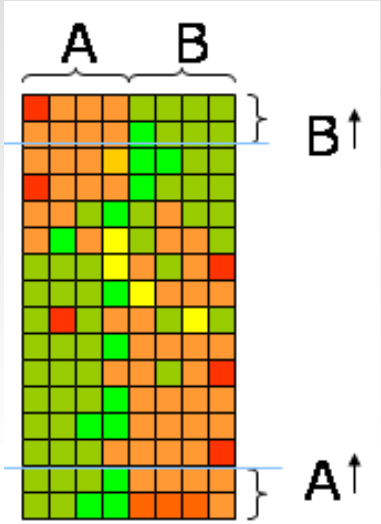


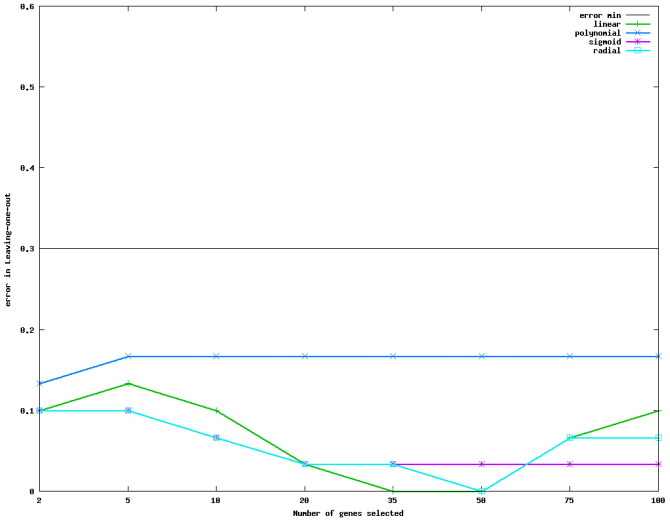
Context and cooperation between genes is not ignored

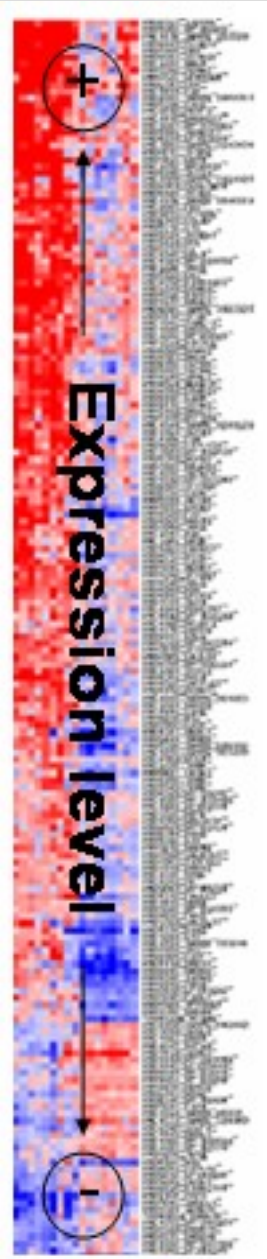
The unit of interest in the study is shifted from gene to function

Genome-scale experiment output

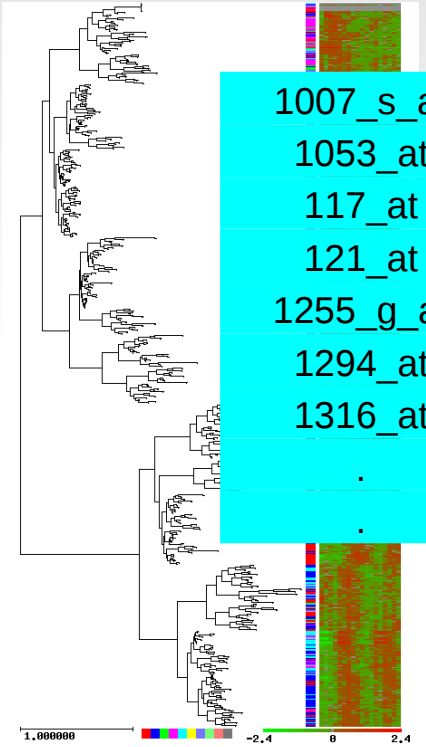




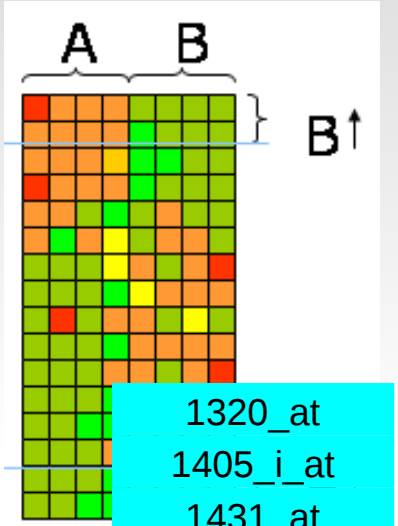




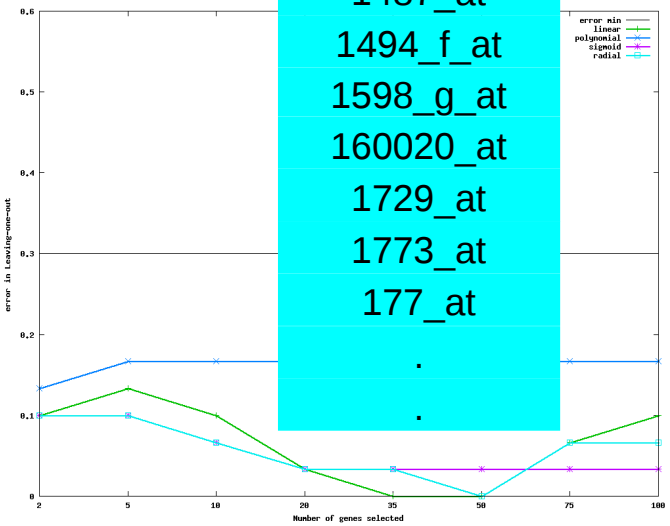
Genome-scale experiment output

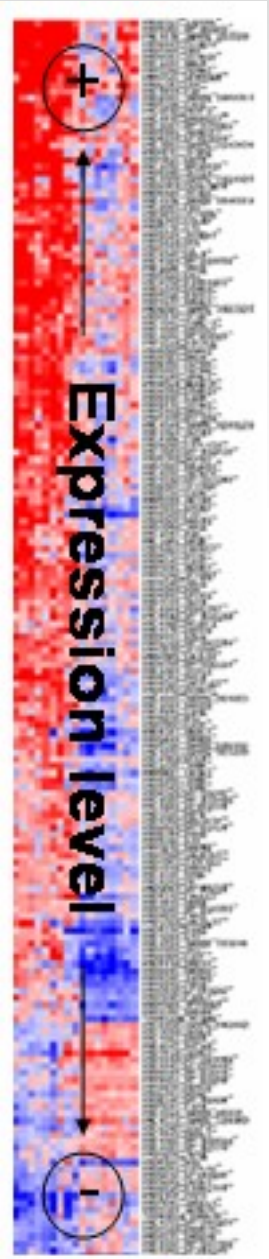


1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

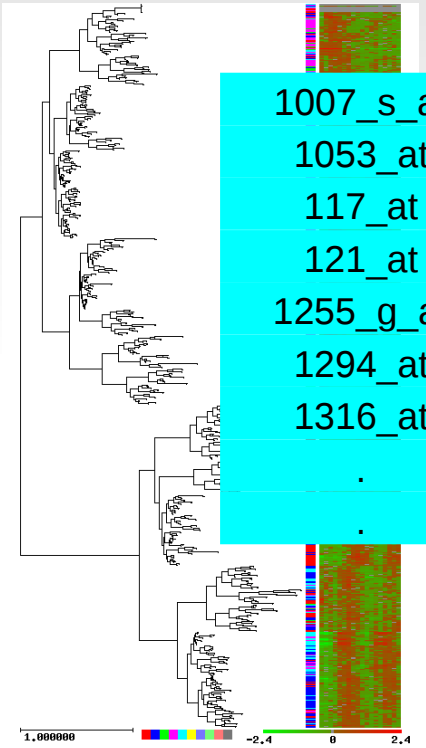


1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

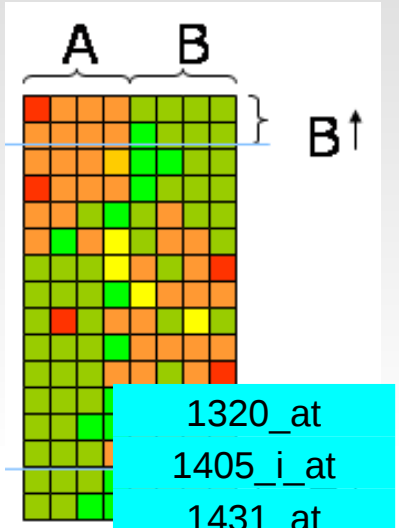




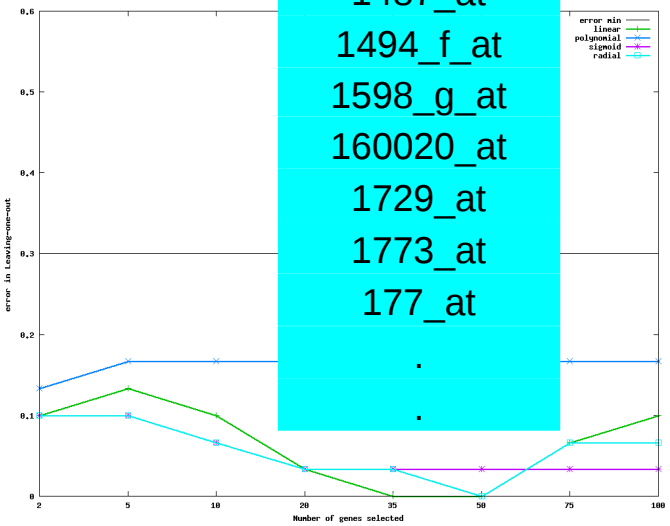
Genome-scale experiment output



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

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

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
.
.

Functional interpretation

1007_s_at
1053_at
117_at
121_at
1255_g_
1294_
1316
.

1
1
14
14
1494
1598_g
160020_
1729_at
1773_at
177_at
.
.

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

Recorded to:

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

t	12.4
	11.5
	10.3
	10.2
	9.9
	9.3
	8.2
	8.1
	7.7
	7.4
	6.5
	6.2
	5.9
	5.8
t	4.8
_at	4.7
.	.
.	.

Functional interpretation

1007_s_at
1053_at
117_at
121_at
1255_g_
1294_
1316
.

1
1
1
14
1494
1598_g
160020_
1729_at
1773_at
177_at
.
.

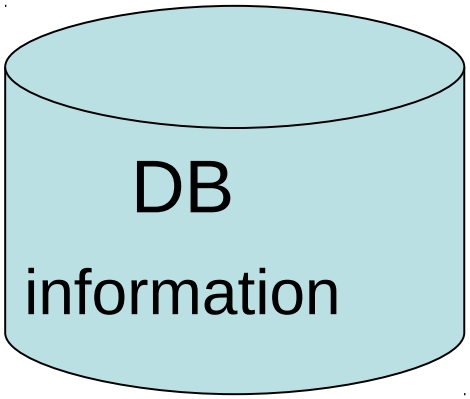
t	12.4
	11.5
	10.3
	10.2
	9.9
	9.3
	8.2
	8.1
	7.7
	7.4
	6.5
	6.2
	5.9
	5.8
t	4.8
_at	4.7
.	.
.	.

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

Recorded to:

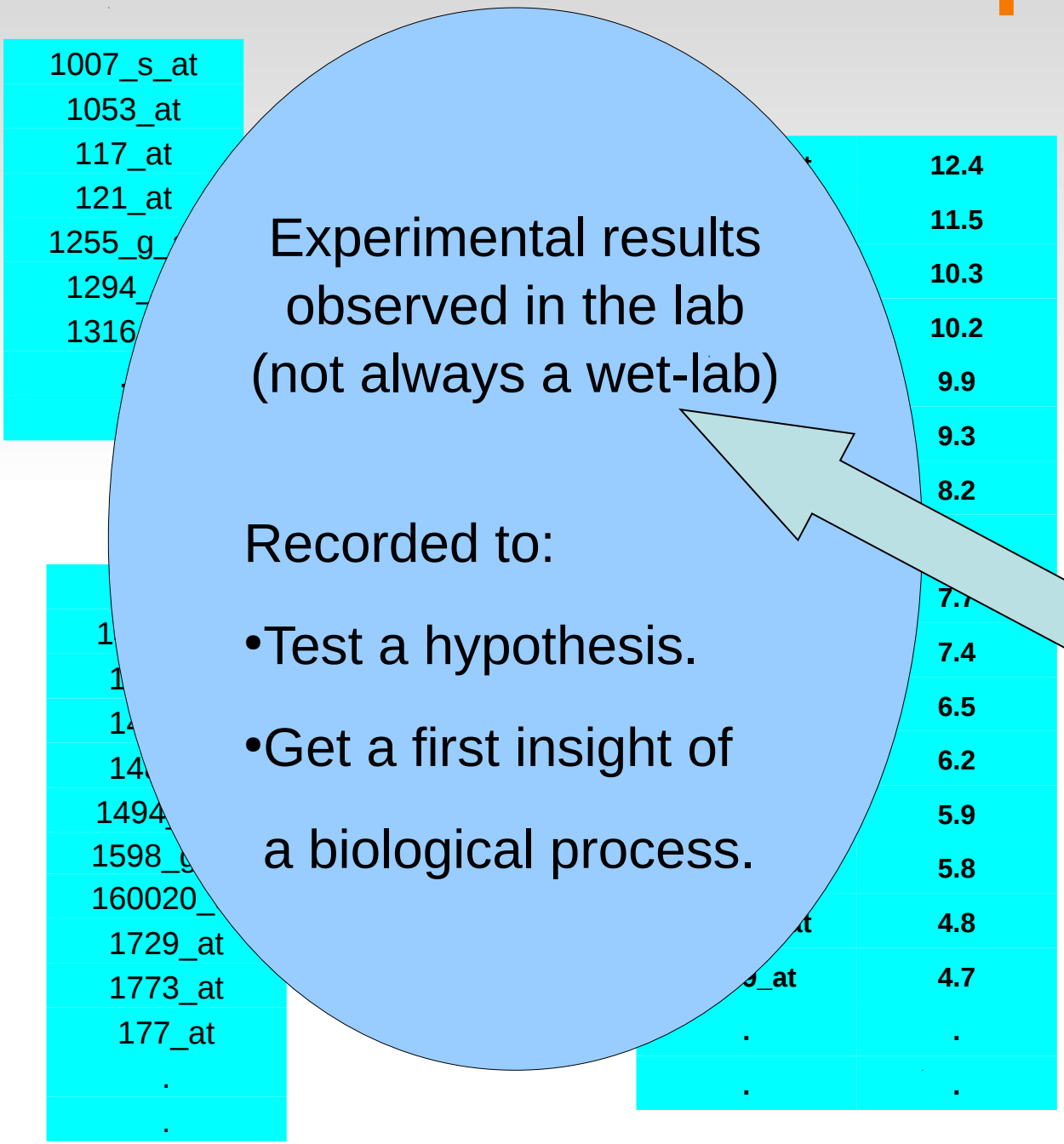
- Test a hypothesis.
- Get a first insight of a biological process.

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



Already tested and stored

Functional interpretation

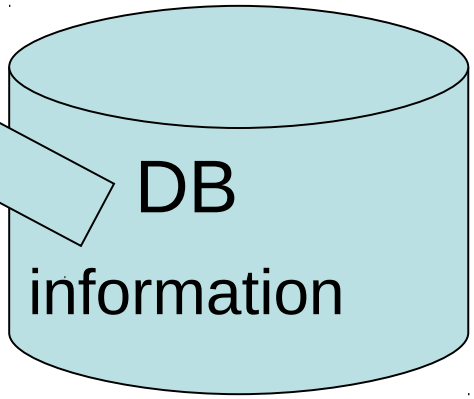


Experimental results observed in the lab (not always a wet-lab)

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

UniProt/Swiss-Prot	EntrezGene	HGNC symbol	PDB
UniProtKB/TrEMBL	Affymetrix	EMBL acc	Protein Id
Ensembl IDs	Agilent	RefSeq	IPI....

Gene IDs

Functional databases

KEGG pathways

Reactome

Gene Ontology

ess Molecular Function Cell

Regulatory elements

miRNA

CisRed

Transcription Factor

Binding Sites

Keywords Swissprot

Biocarta pathwa

Gene Expression in tissues

InterPro Moti

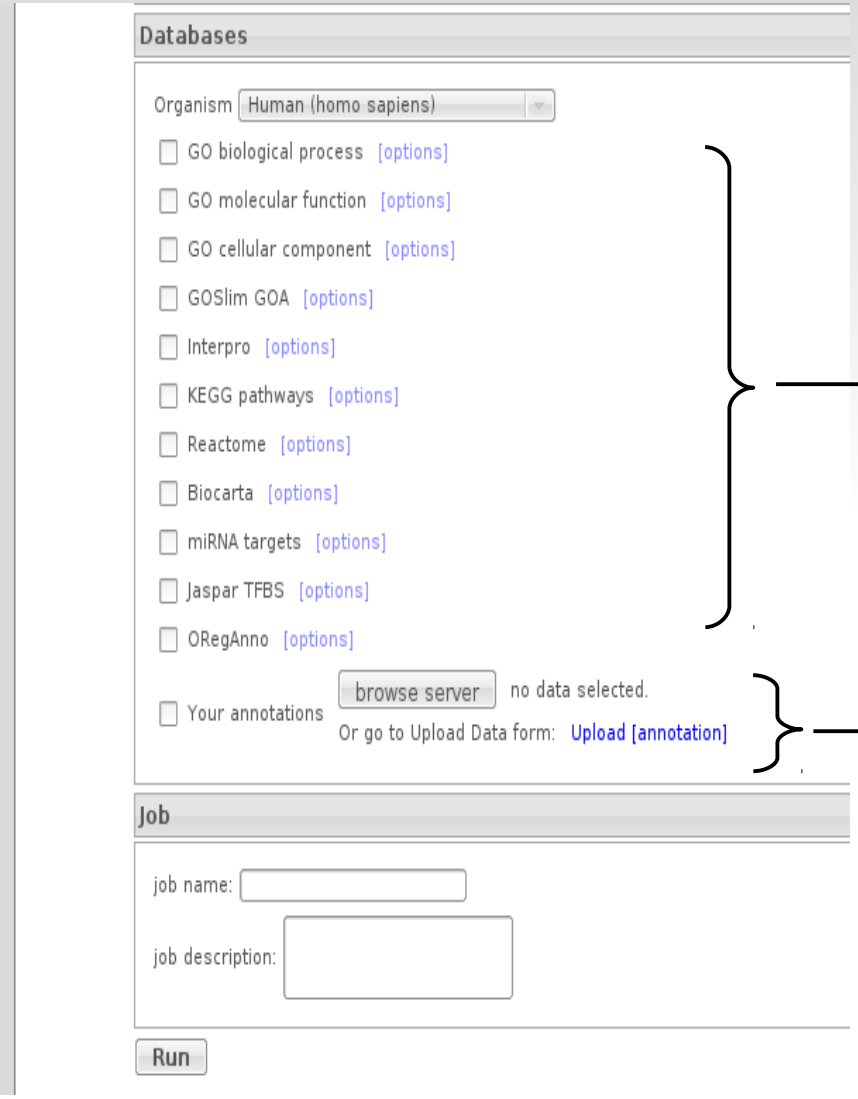
Bioentities from literature:

ses terms

Chemical terms

Babelomics Databases

Which type of functional information?

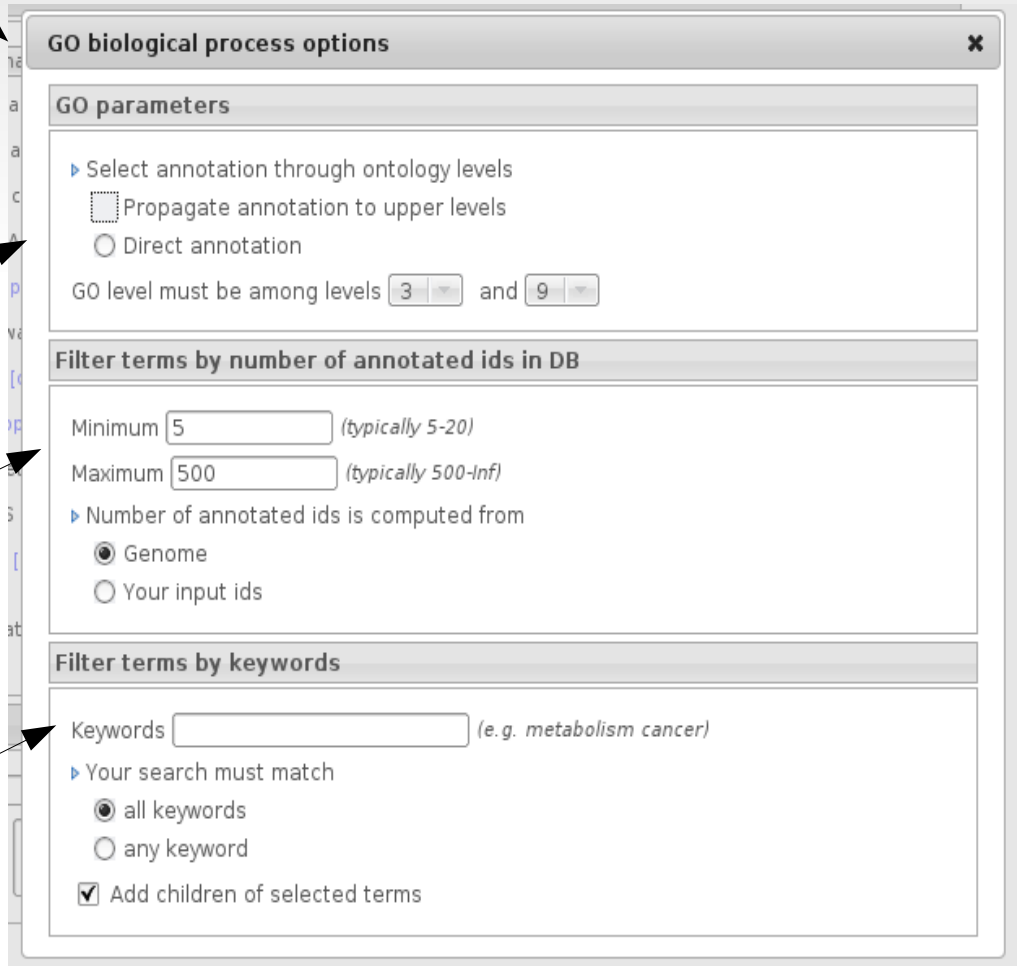


Use one or more of the given databases

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

Babelomics Databases

First select an organism



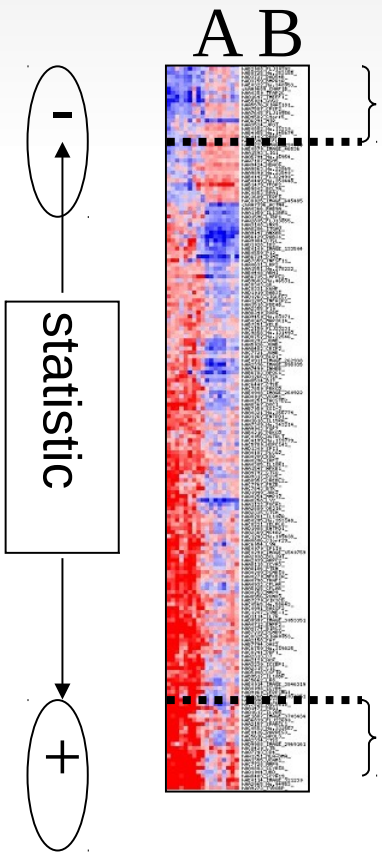
<u>OPTIONS:</u>

Test all the GO or only annotated terms

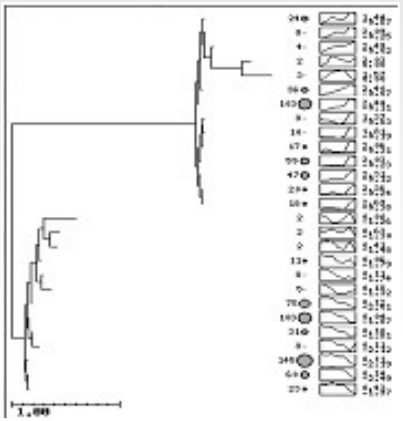
Discard functions with too few or too many genes?

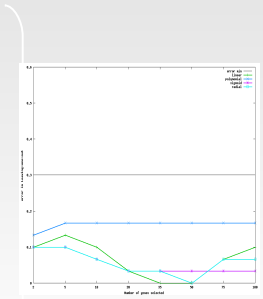
If you have an hypothesis, better test this first!!!!!!

Simple enrichment analysis



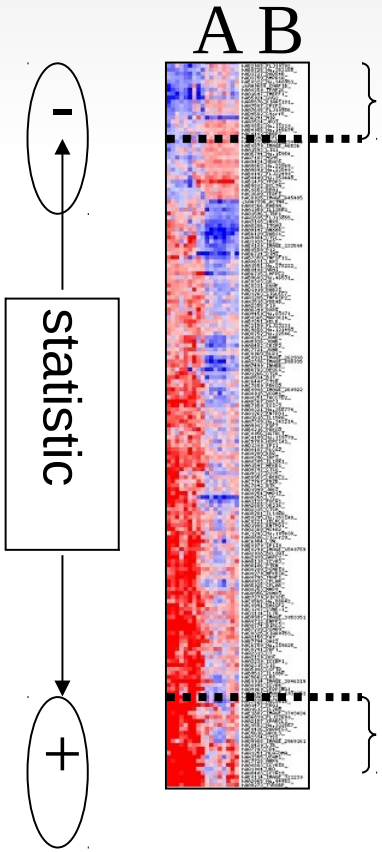
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.



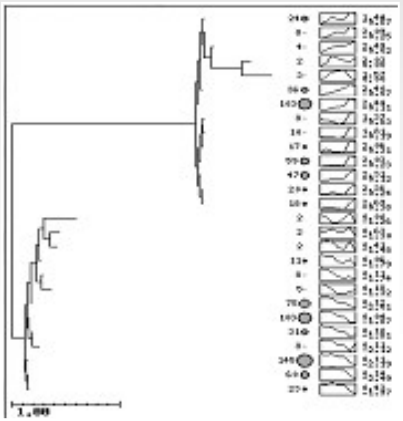


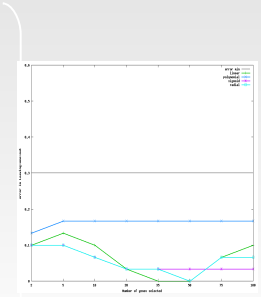
1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

Simple enrichment analysis



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





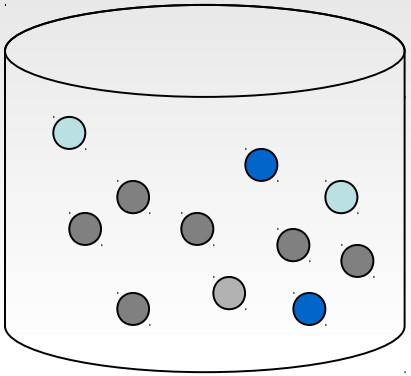
GO

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

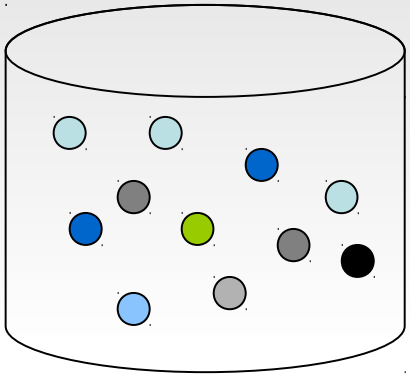
One Gene List (A)



Biosynthesis 60%

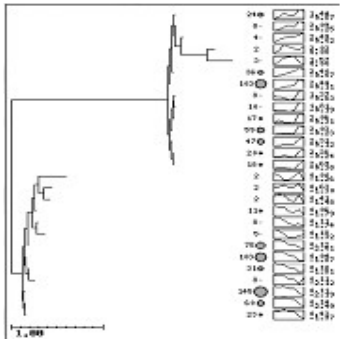
Sporulation 20%

The other list (B)



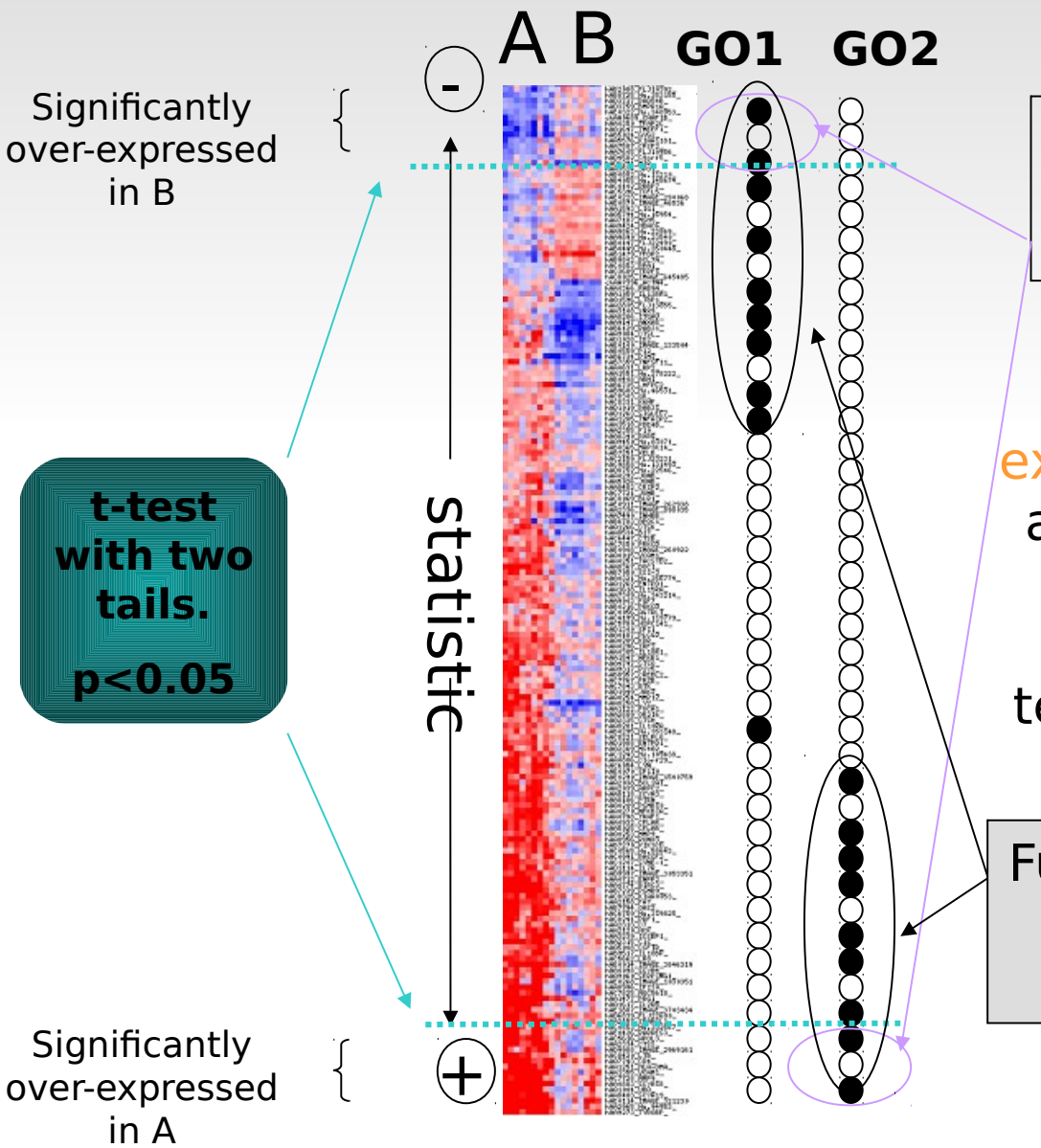
Biosynthesis 20%

Sporulation 20%



	A	B
Biosynthesis	**6**	**2**
No biosynthesis	**4**	**8**

FatiGO approach may not be very powerful



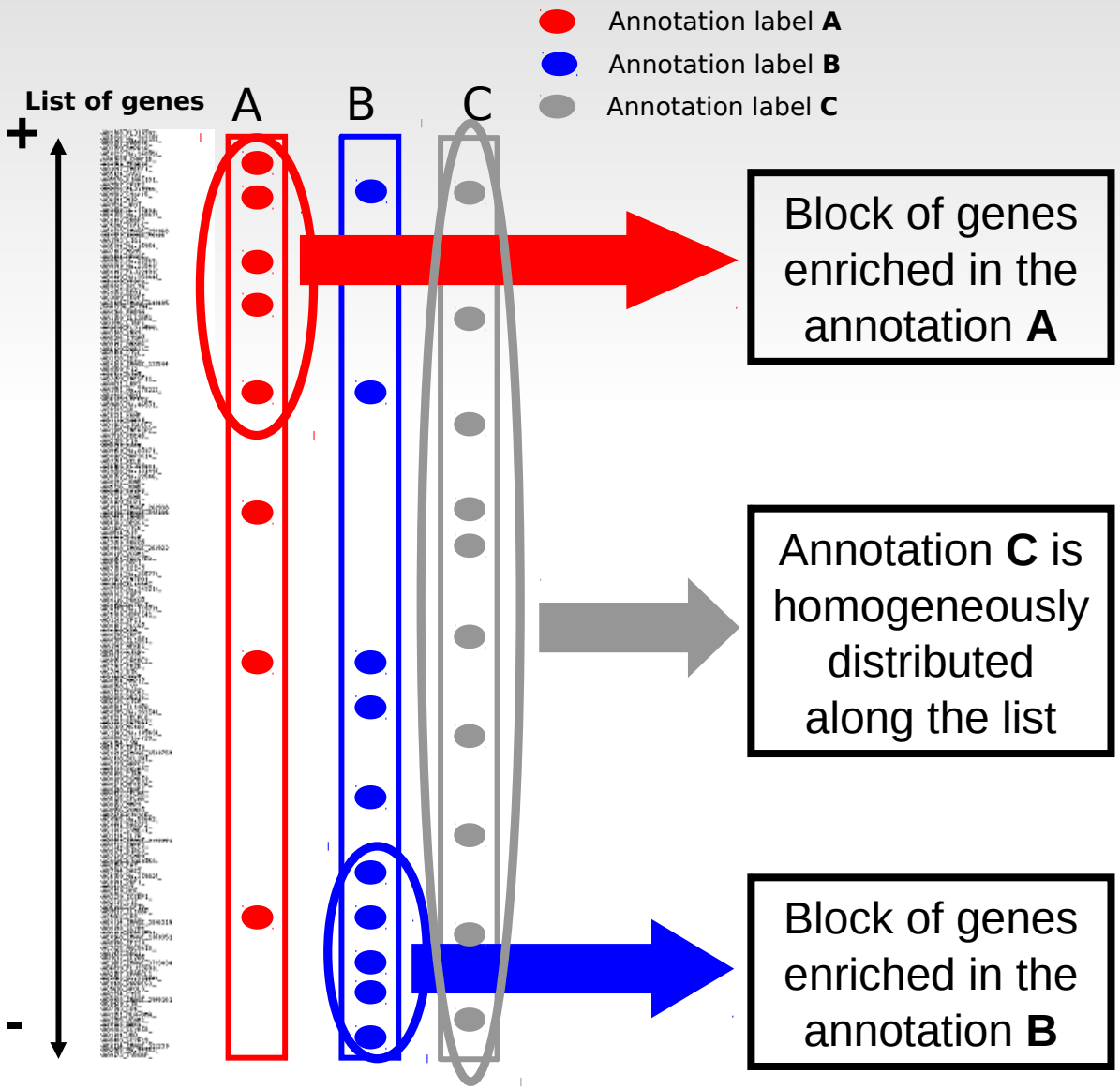
Very few genes selected to arrive to a significant conclusion on GO1 and GO2

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

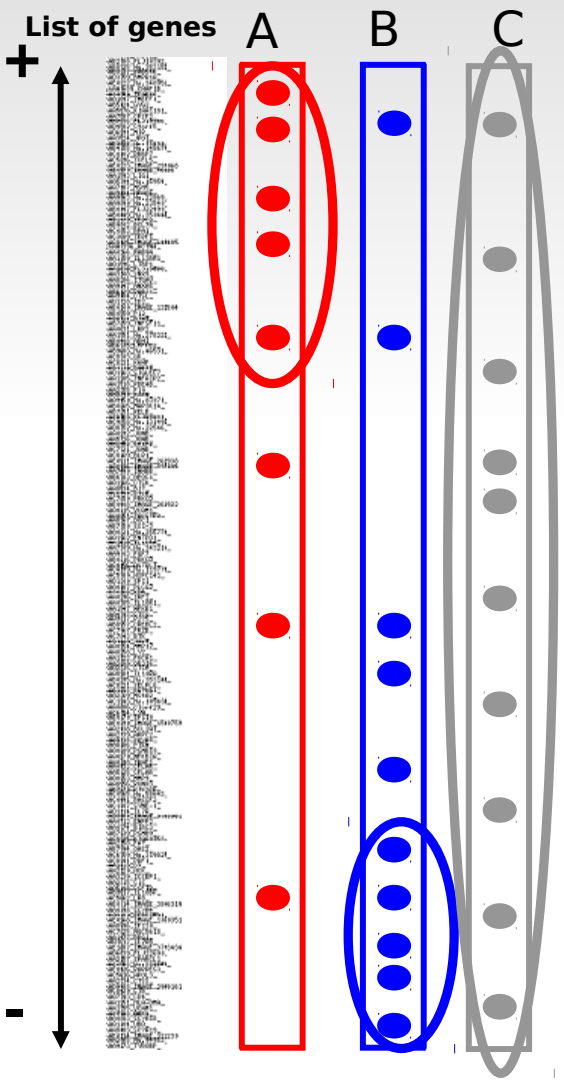
Functional Classes expressed as blocks in A and B

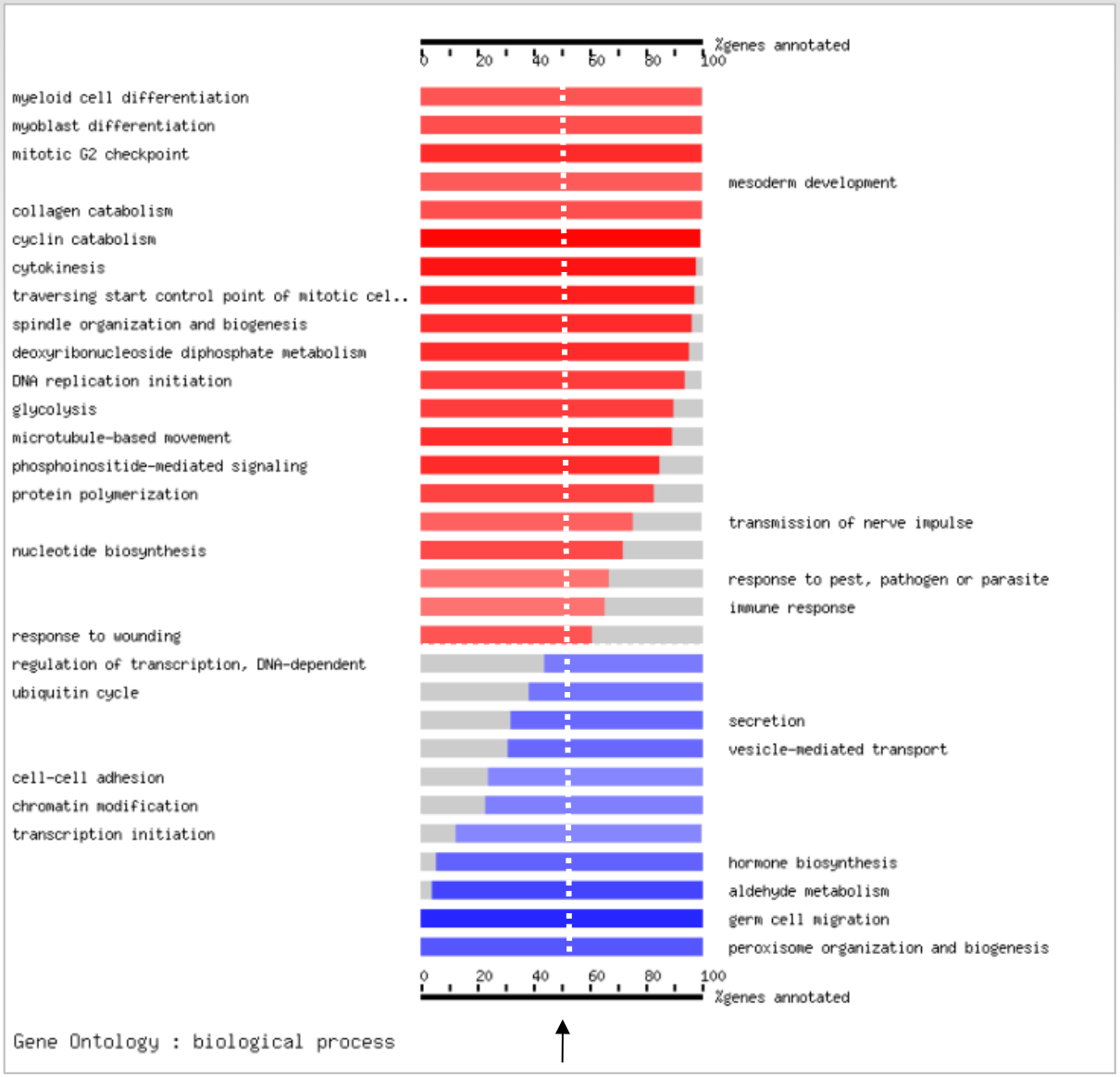
FatiScan, testing along an ordered list

- 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.



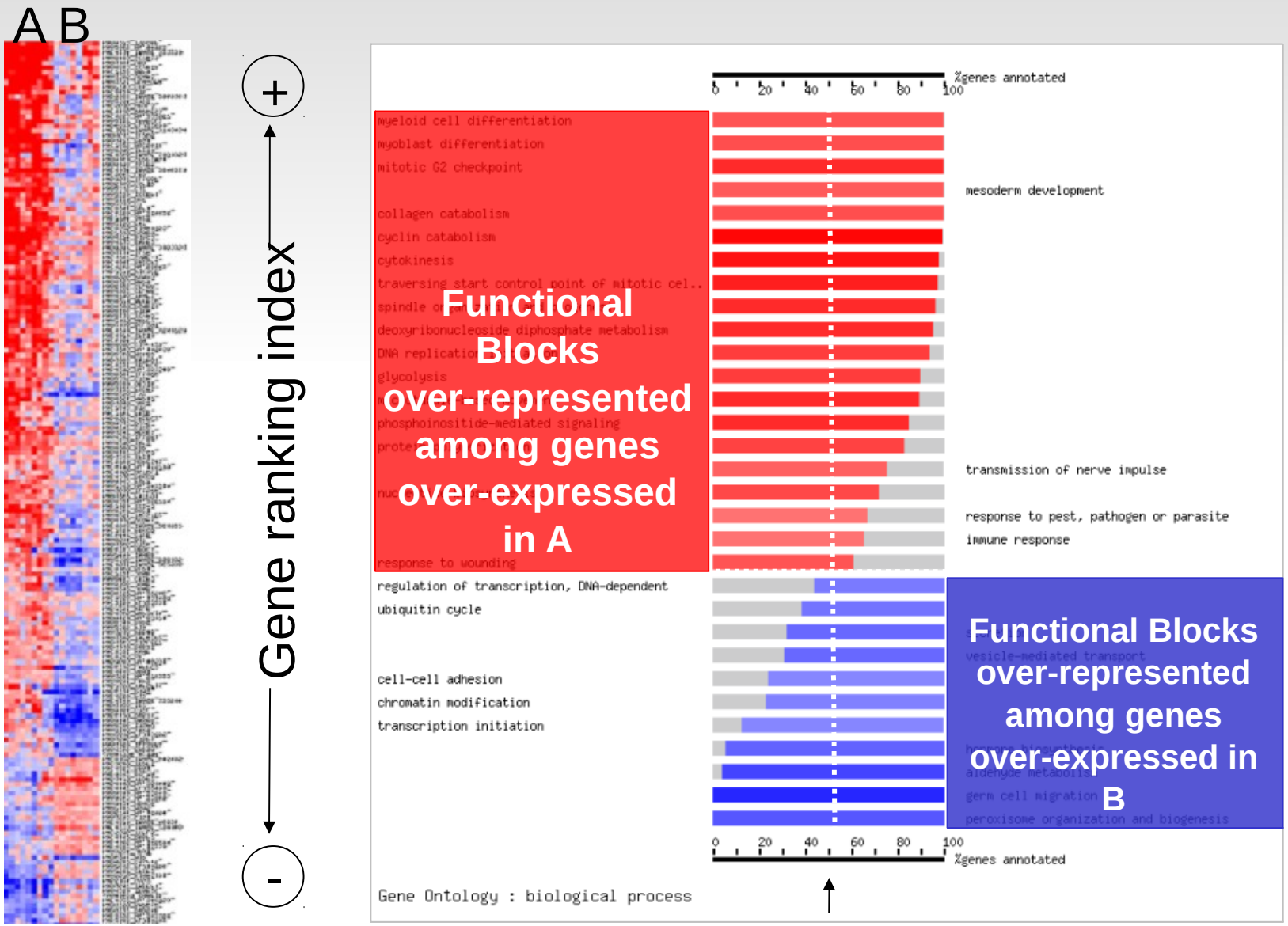
FatiScan results





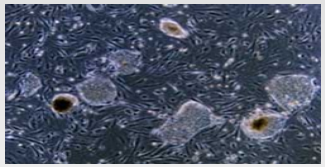
Babelomics

Fatiscan results

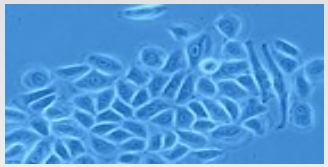


FatiScan Example – two classes

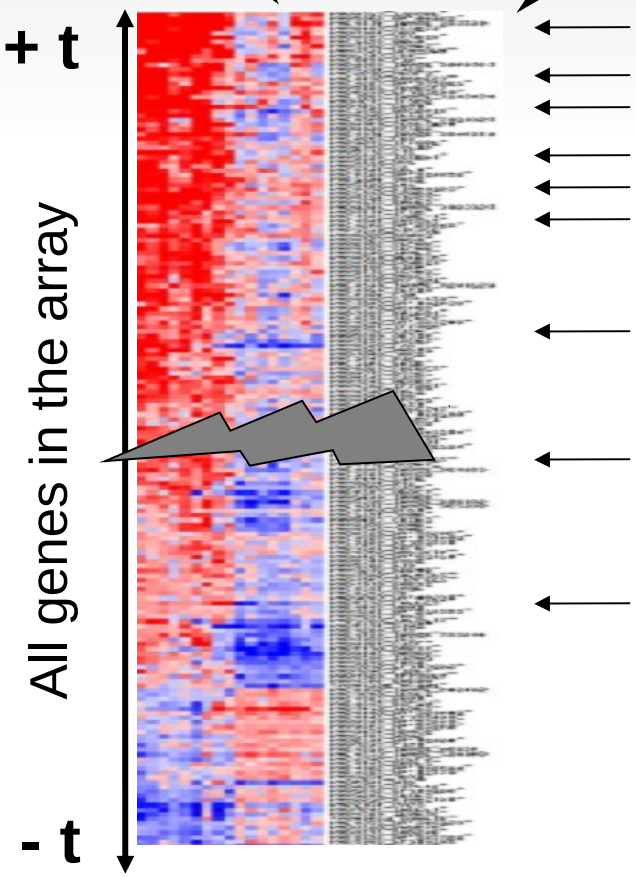
Tumor



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

FatiScan Example - Survival Analysis

- Cromer et all. **Identification of genes associated with tumorigenesis and metastatic potential of hypopharyngeal cancer by microarray analysis**. *Oncogene* 2004, **23**(14) : 2484-2498.
- 34 hypopharyngeal cancer samples taken from patients undergoing surgery.
- Analyzed using Affymetrix HG-U95A microarrays (~12650 distinct transcription features).
- Disease free survival time after intervention was recorded

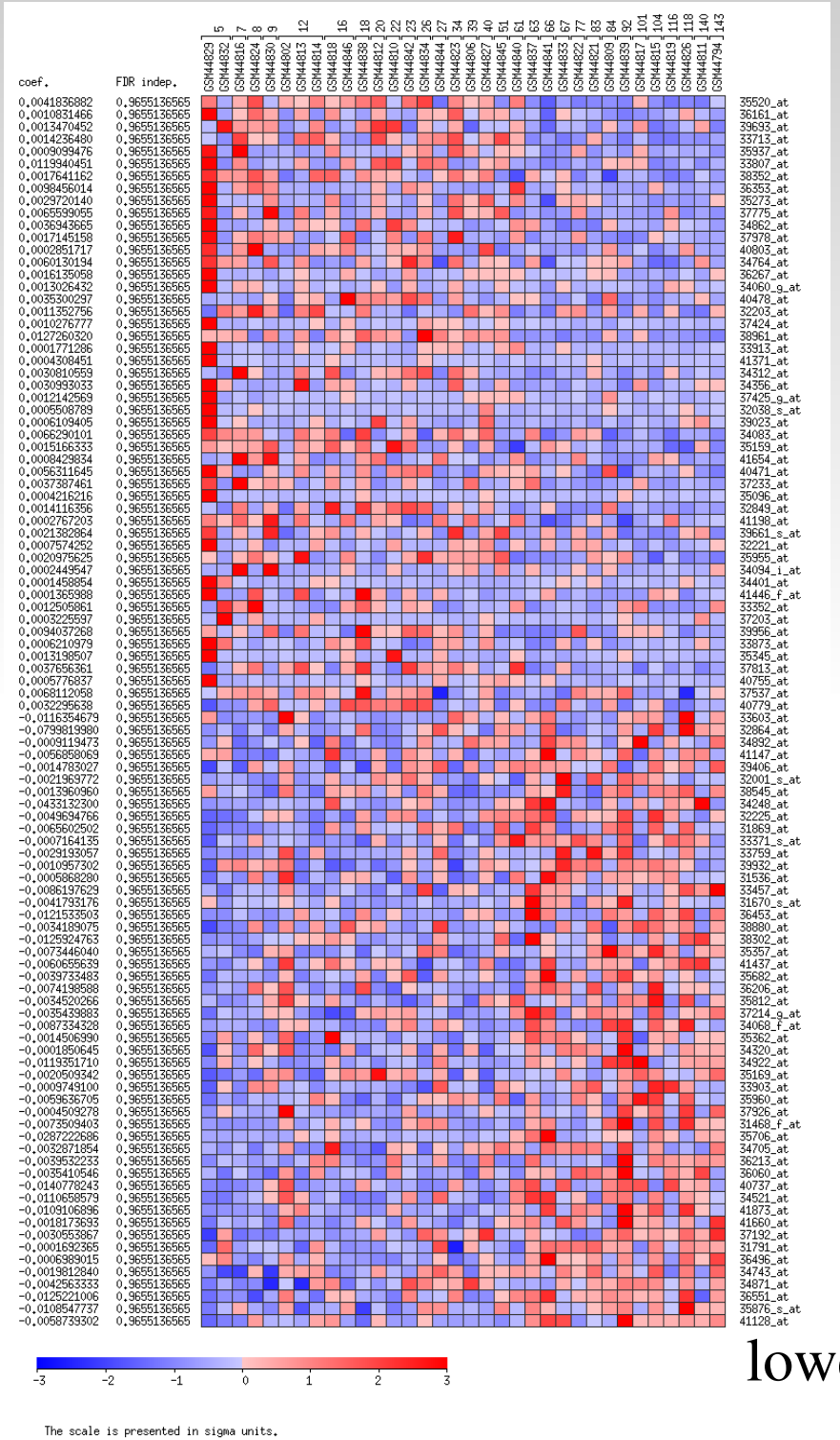
Cox proportional hazards model

$$h(t) = h_0 (t) * \exp (\beta * \text{gene expression})$$

Gene Ontology: biological process

Hazard increased with expression **+ β**

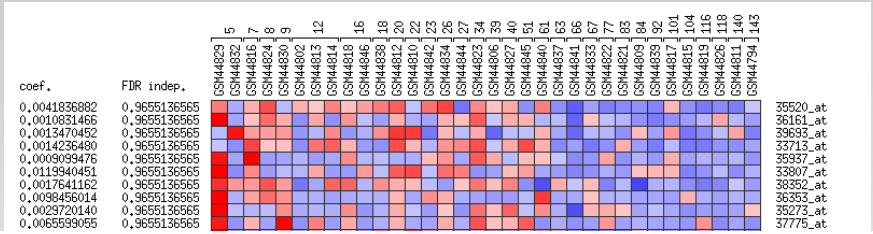
Hazard decreased with expression **- β**

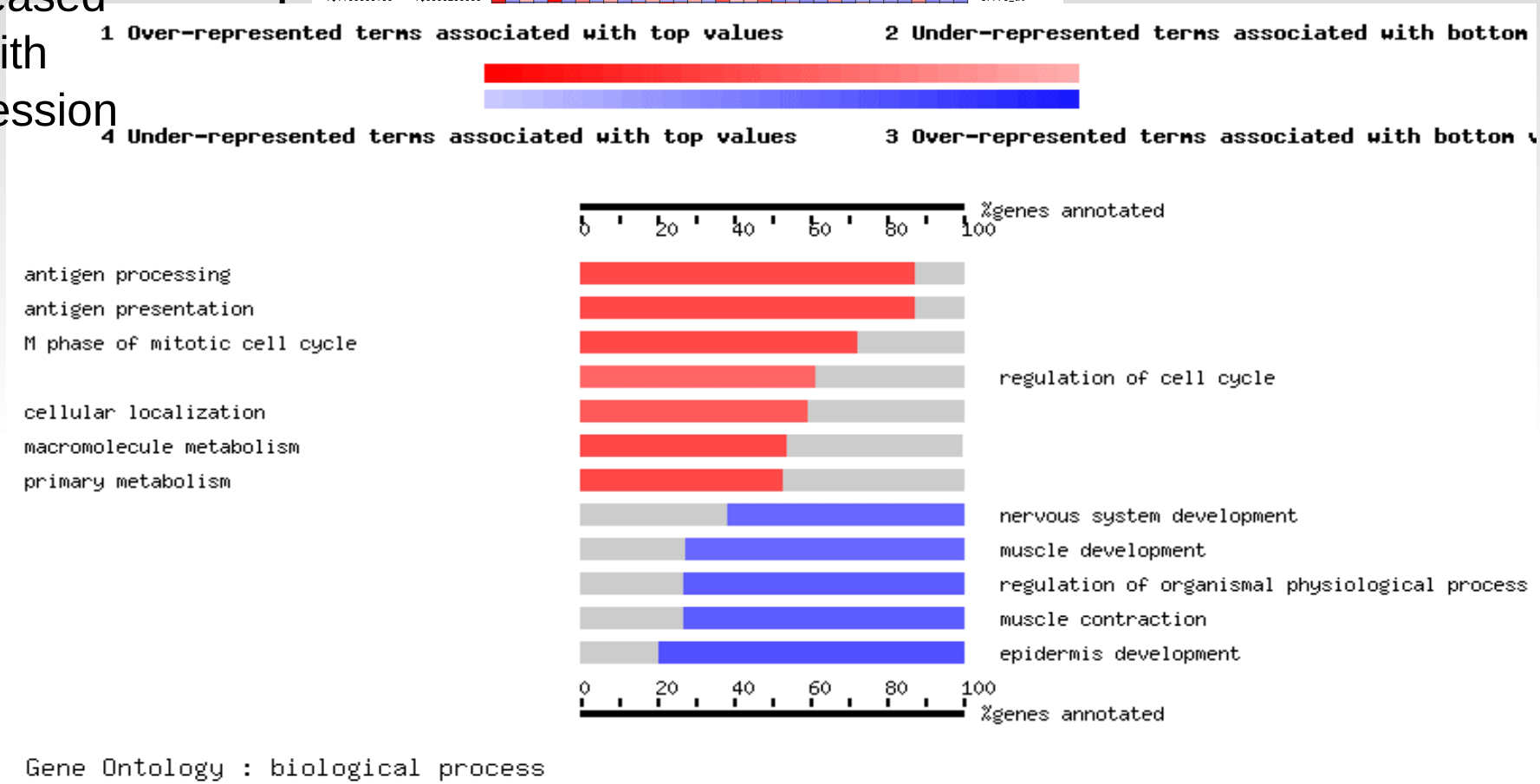


lowest p-value = 0.96

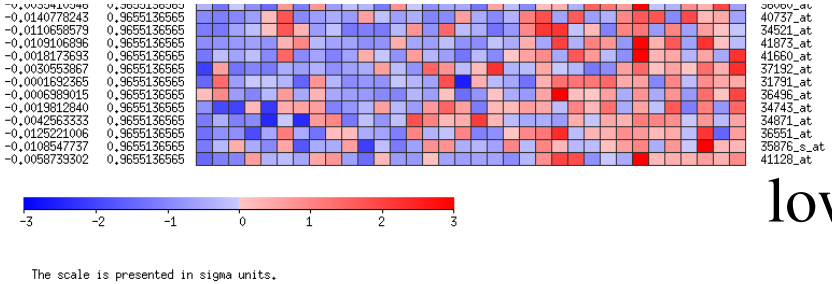
Gene Ontology: biological process

Hazard increased with expression **+ β**



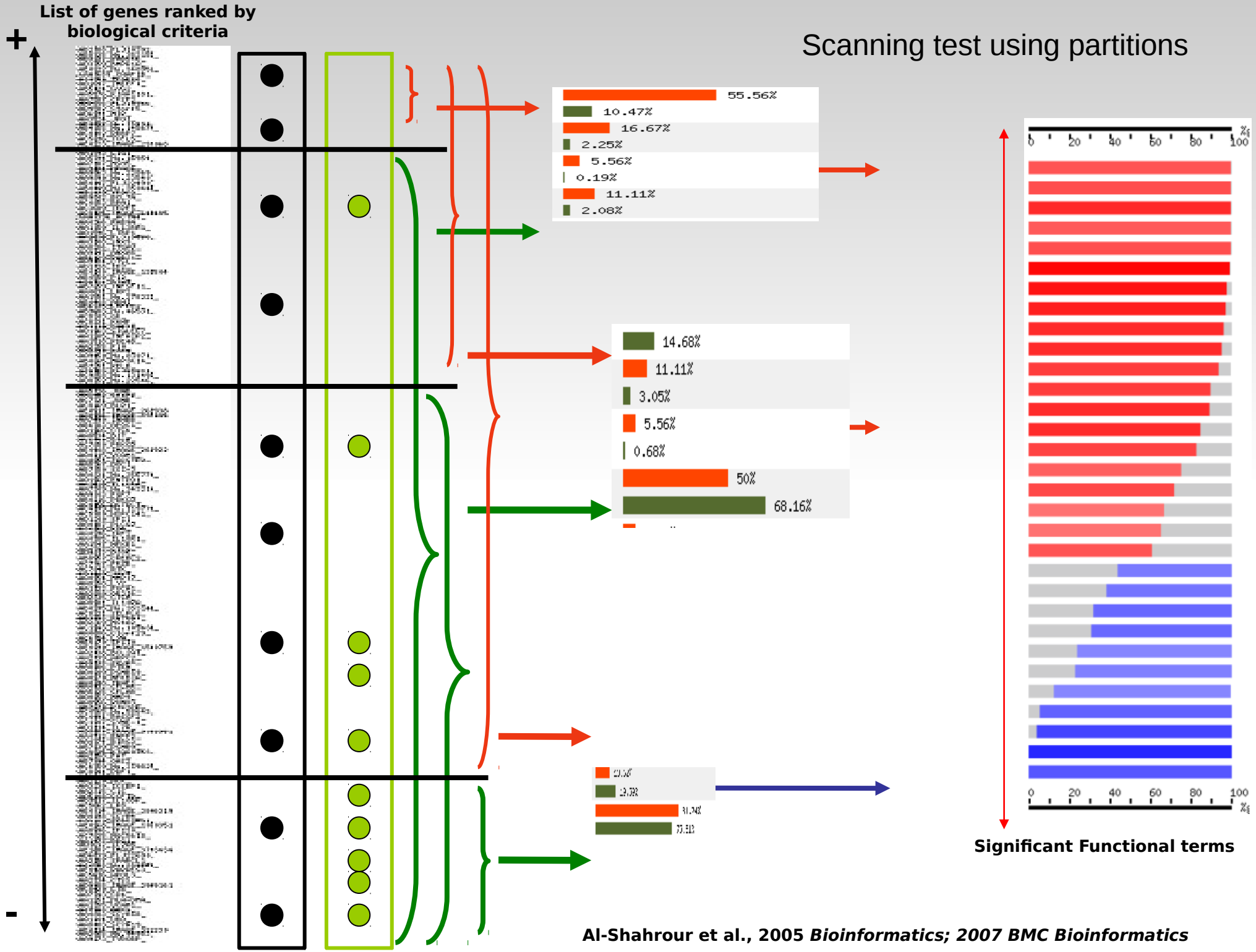


Hazard decreased with expression **- β**



lowest p-value = 0.96

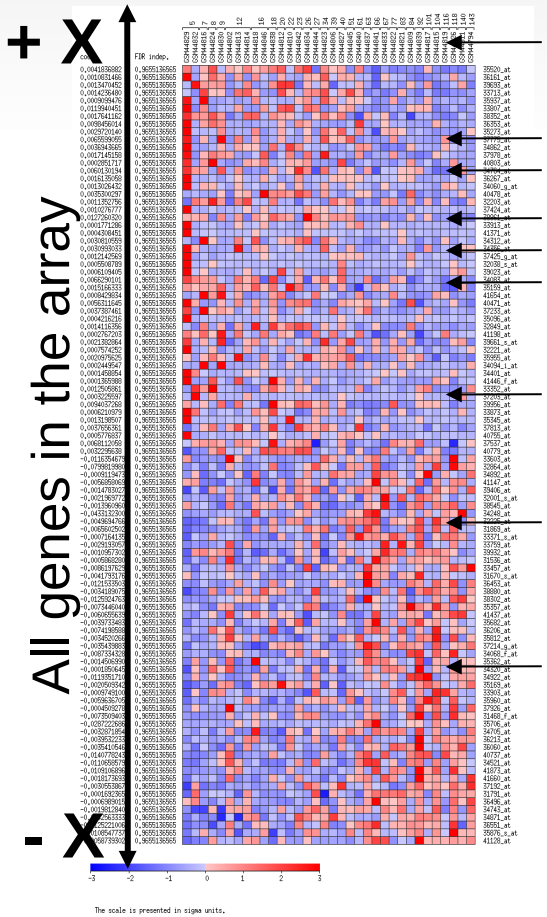
Scanning test using partitions



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

Logistic test

- Not using partitions
- But logistic regression model



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

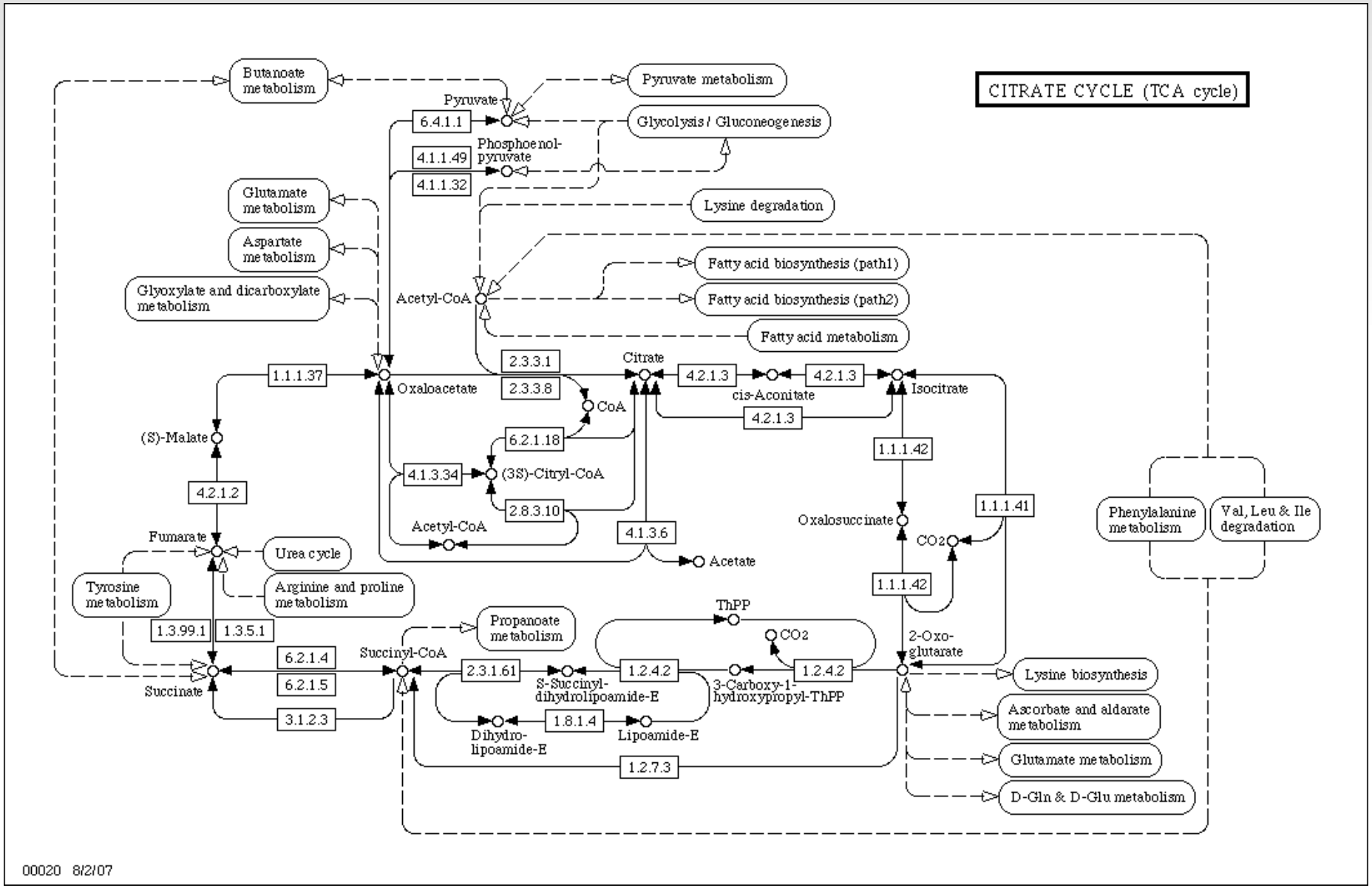
alpha > 0 : **increasing** X increases the probability of the gen to be annotated

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

Remarks

- The unit of information over which we test is shifted from genes to functional blocks (multiple testing again)
- We do one statistical test for each block
- All genes in the block are treated equally

Network modeling



Remarks

- The unit of information over which we test is shifted from genes to functional blocks (multiple testing again)
- We do one statistical test for each block
- All genes in the block are treated equally
- Annotation information is 0, 1

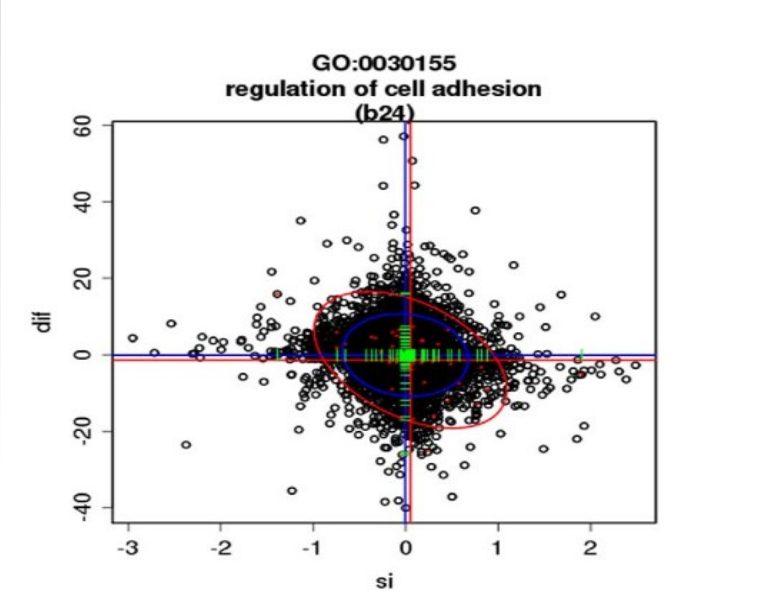
Remarks

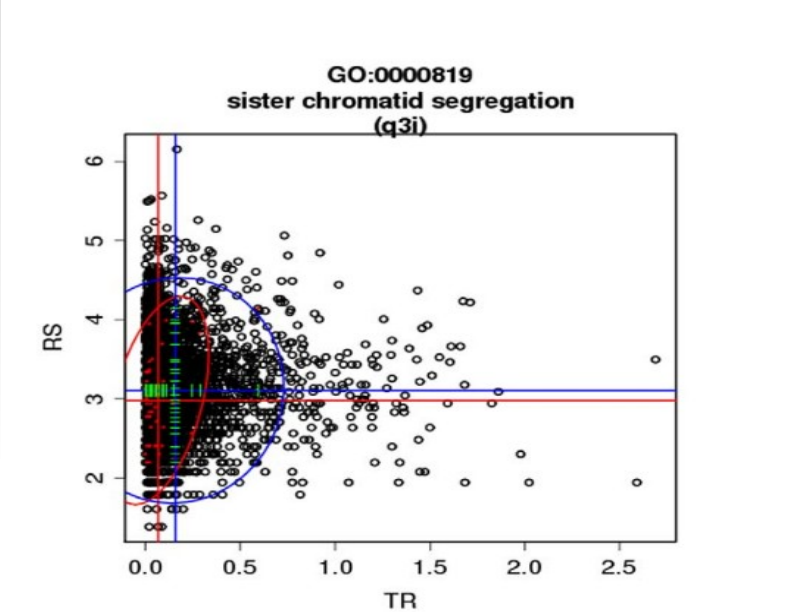
- The unit of information over which we test is shifted from genes to functional blocks (multiple testing again)
- We do one statistical test for each block
- All genes in the block are treated equally
- Annotation information is 0, 1
- Genes independently may not show a strong pattern of association but the block coordinately does.

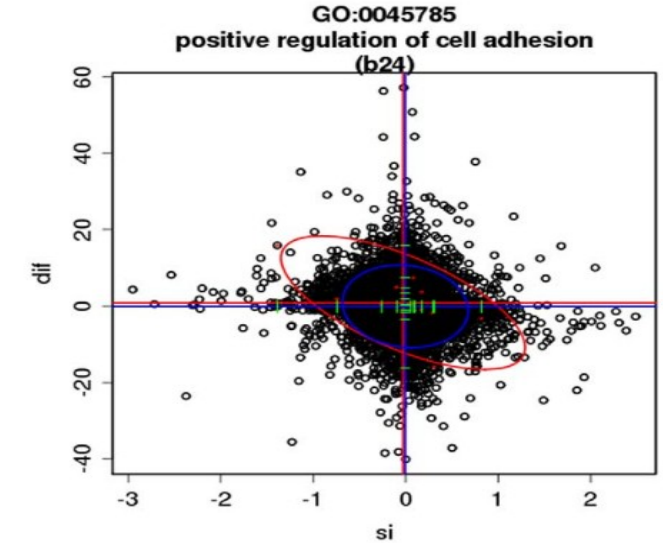
Remarks

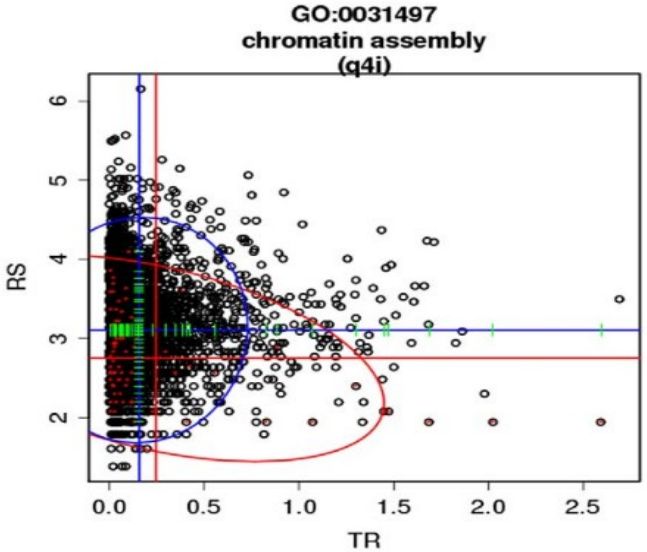
- The unit of information over which we test is shifted from genes to functional blocks (multiple testing again)
- We do one statistical test for each block
- All genes in the block are treated equally
- Annotation information is 0, 1
- Genes independently may not show a strong pattern of association but the block coordinately does.
- Only ranking genes according to a unique condition

Multidimensional Analysis





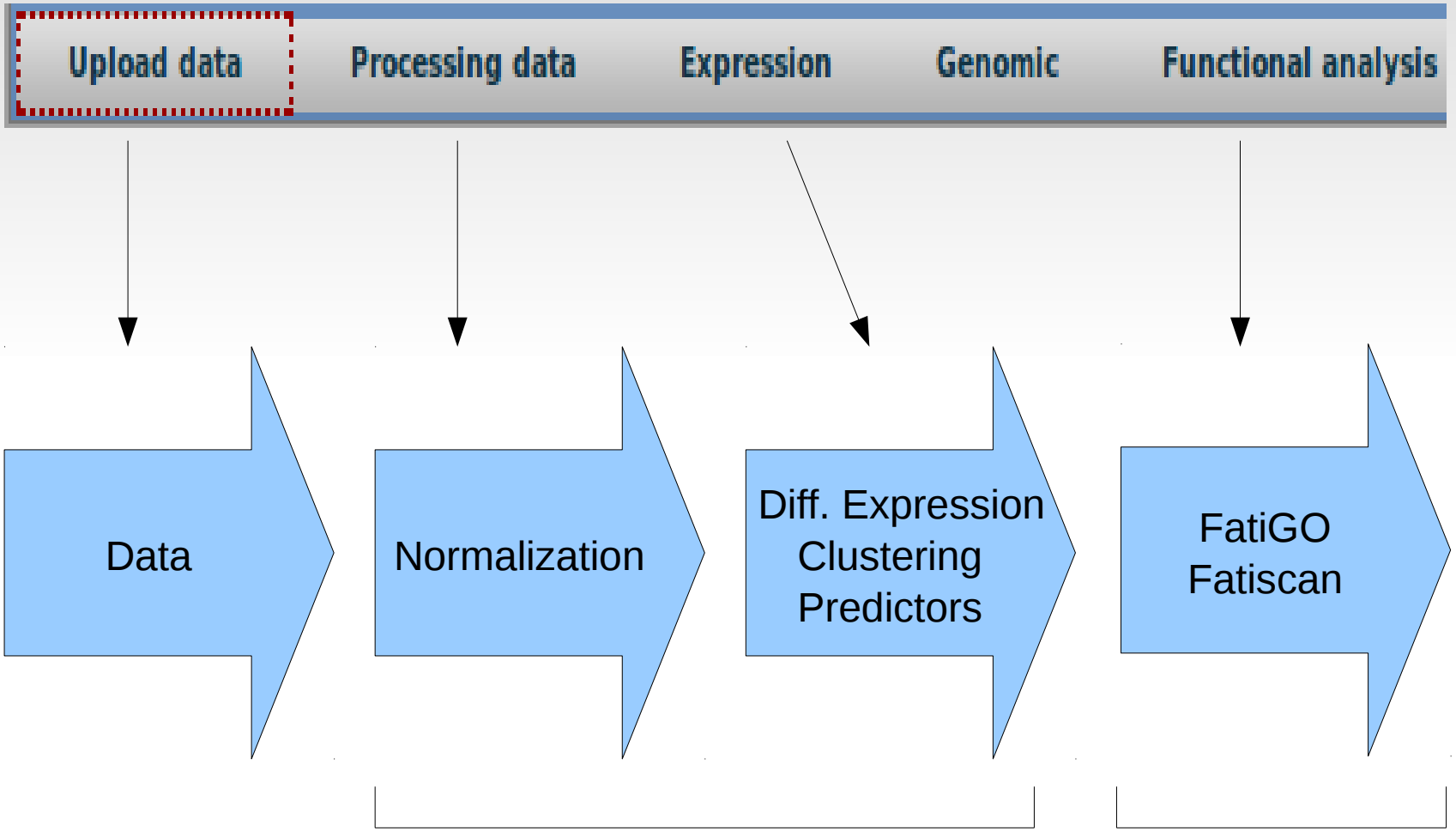




Gene Set Methods - 2 general Approaches

- Competitive Hypothesis – Babelomics
  - Each functional block is compared to the remaining genes of the genome (or the second list).
  - Independent of the test used to derive the ranking.
- Self Contained Hypothesis – Goeman 2004
  - Checks that the block it self is differentially expressed, correlated with phenotype, associated to survival...
  - Has to be developed with the test creating the ranking of the genes.

Modules / Analysis progress



Gene Expression ANALYSIS

Functional ANALYSIS

Thank You

www.babelomics.org