

# Microarray data analysis

## Introduction

Department of Bioinformatics, Centro de Investigación  
Príncipe Felipe, and  
Functional genomics node, INB, Spain.

<http://www.gepas.org>.

<http://www.babelomics.org>

<http://bioinfo.cipf.es>



**ciberer**



PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

# Background

*Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.*

**Sydney Brenner, 1980**

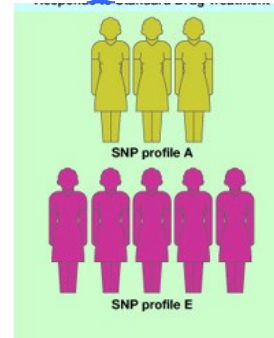


The introduction and popularisation of high-throughput techniques has drastically changed the way in which biological problems **can** be addressed and hypotheses can be tested.

But not necessarily the way in which we really address or test them...

# The pre-genomics paradigm

Genes in the  
DNA...



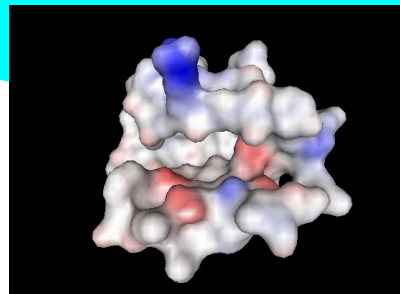
...code for  
proteins...

```
>protein kinase  
acctgttgatggcgacagggactgtatgctgac  
tatgctgatgcacatgctgactactgatgtggg  
ggctattgactgatgtctatc....
```

From genotype to  
phenotype.

...produces the final  
phenotype

...whose structure  
accounts for function...



...plus the  
environment...

Now: 22240 (NCBI build 35 12/04)

50-70% display alternative splicing

25%-60% unknown

Transfrags

>protein kinase

```
acctgttgatggcagaggactgtatctgac  
tatgtgatgcacatgctgactactgatgtggg  
ggctattgacttgatctatc....
```

Genes in the  
DNA...



...which can be different  
because of the variability.

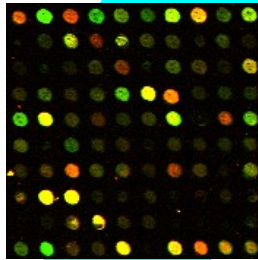
10 million  
SNPs



...whose final  
effect  
configures the  
phenotype...

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

A typical tissue is  
expressing among  
5000 and 10000  
genes



From genotype  
to phenotype.

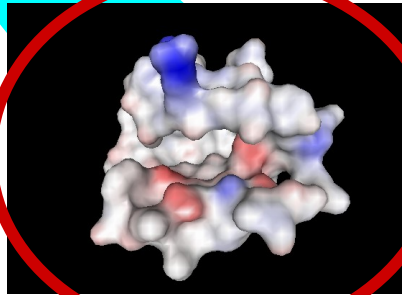
(in the functional post-genomics  
scenario)

...code for  
proteins...

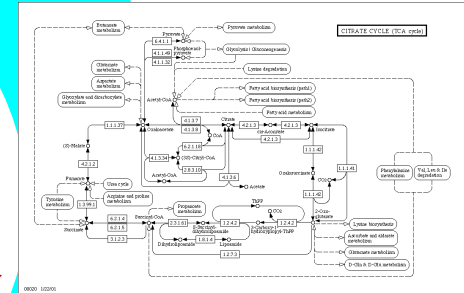
That undergo post-  
translational  
modifications, somatic  
recombination...

100K-500K proteins

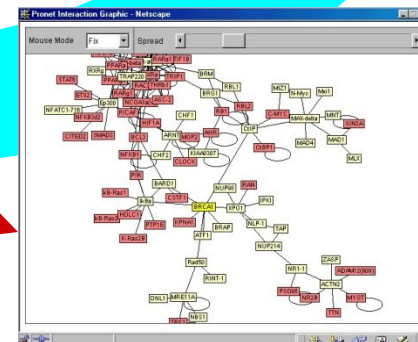
...whose structures  
account for function...



...conforming complex  
interaction networks...

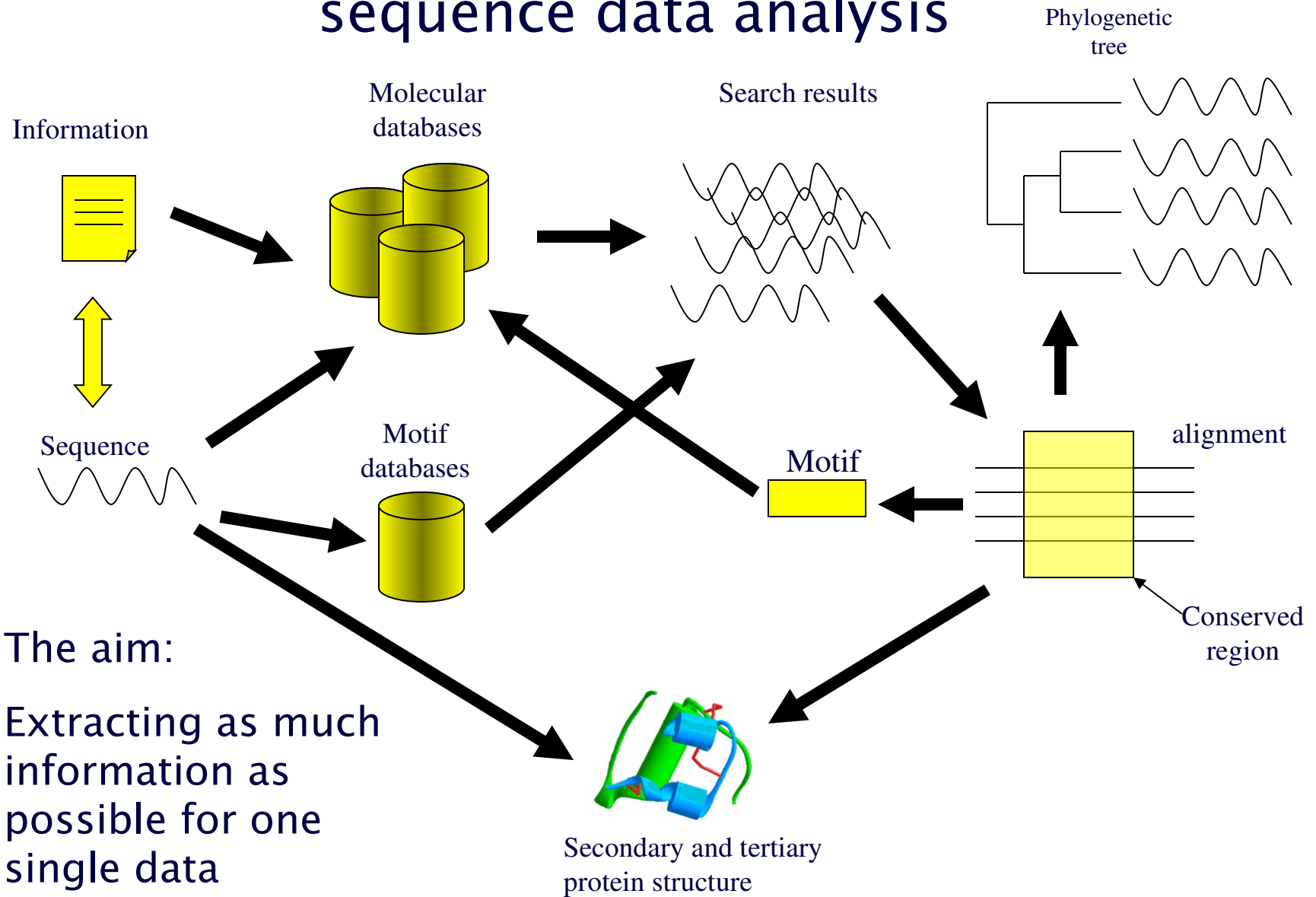


...in  
cooperation  
with other  
proteins...

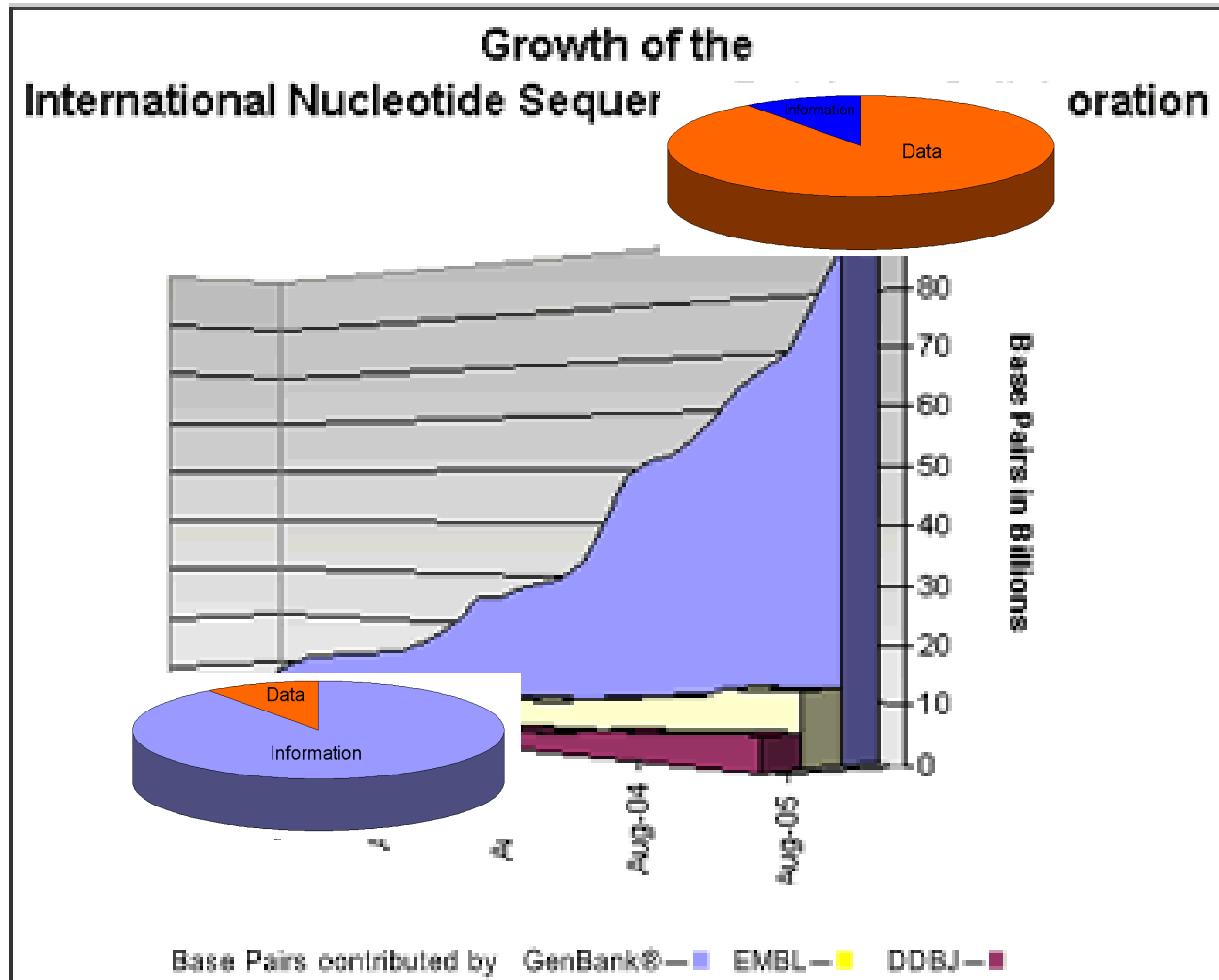


Each protein has an average  
of 8 interactions

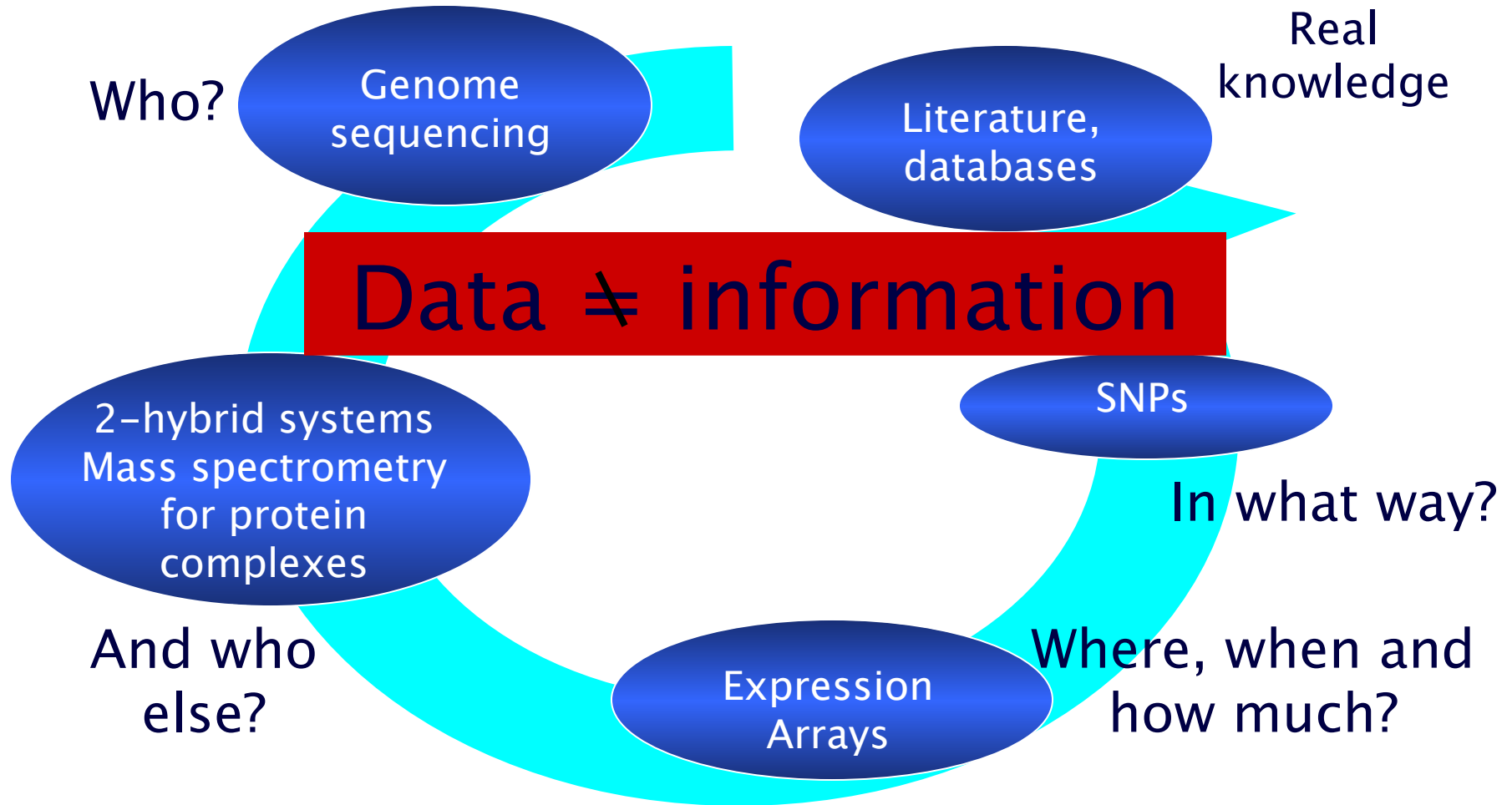
# Bioinformatics tools for pre-genomic sequence data analysis



# Post-genomic vision



# Post-genomic vision



Our capacity of producing data surpasses our capacity of analysing data

Guilty by association

# Genome wide data and a note of caution:

Risks of the “guilty by association” concept.

Genome-wide technologies allows us to produce vast amounts of data.

But... dealing with many data (omic data) increase the occurrence of spurious associations due to chance

Hypothesis  $\longrightarrow$  Experiment  $\longrightarrow$  test

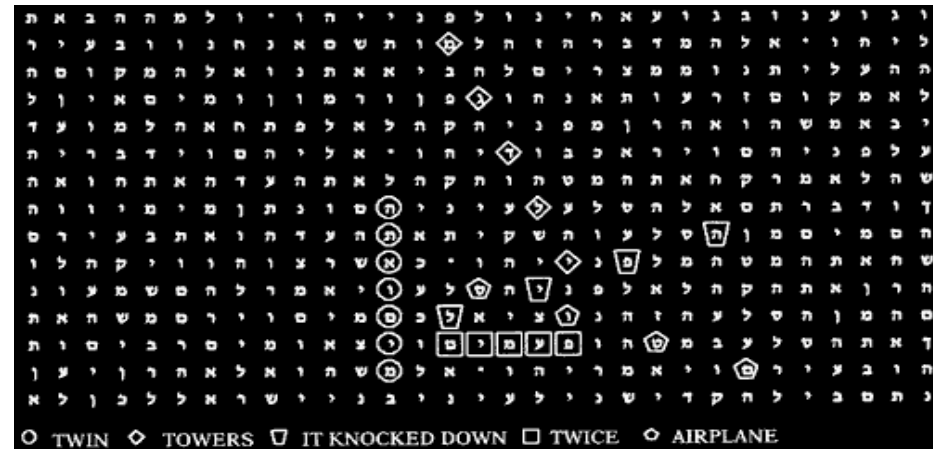
Is gene A involved in process B?

Experiment  $\longrightarrow$  (sometimes) test  $\longrightarrow$  Hypothesis

Is there any gene (or set of genes) involved in any process?

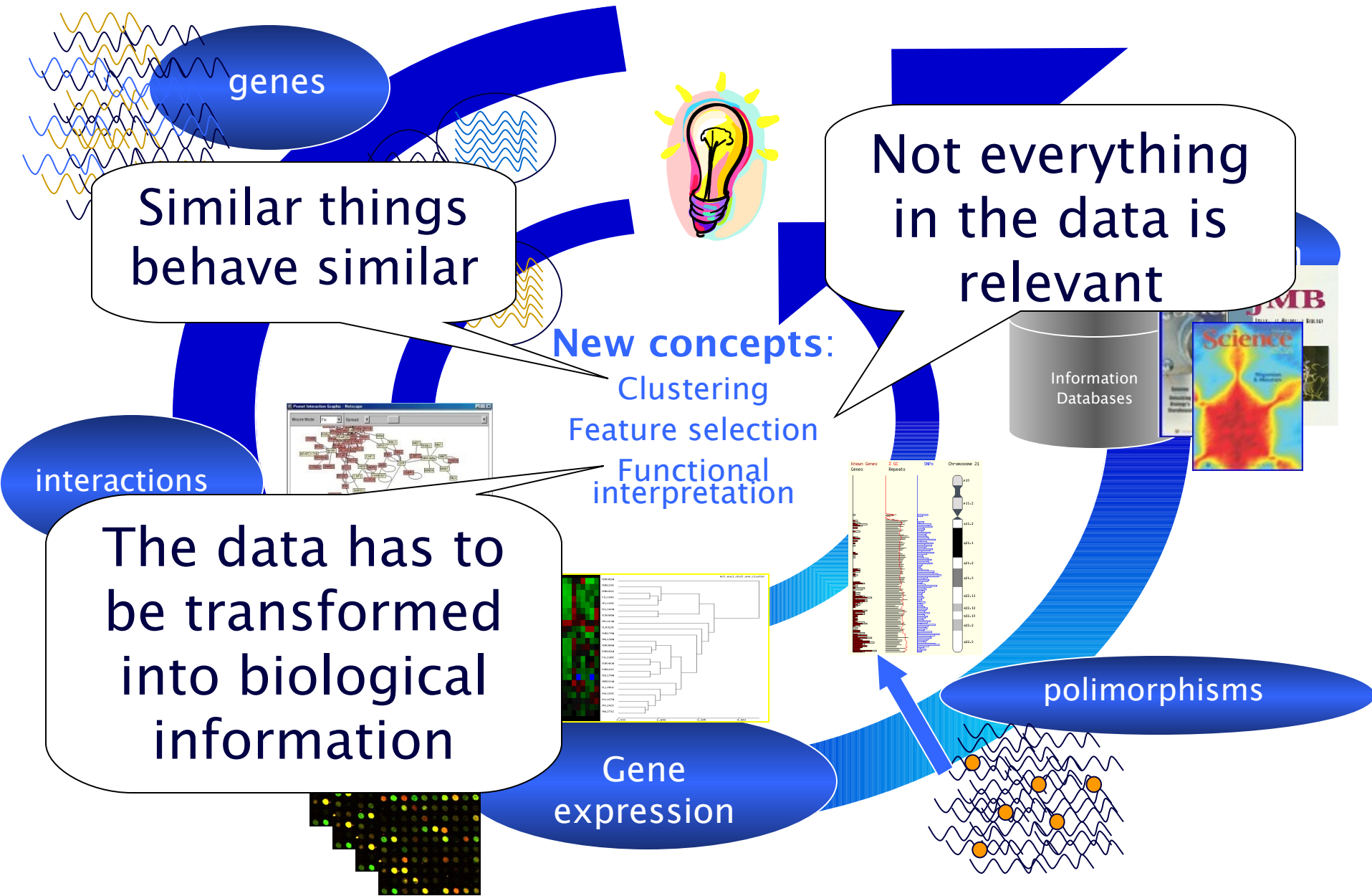
Sure, but... Is it real? (many hypotheses are rejected while this one is accepted *a posteriori*: numerology)

The test is dependent on the hypothesis and not *vice versa*





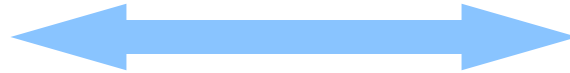
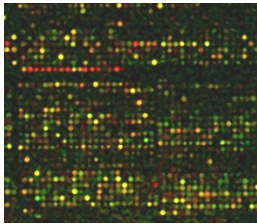
# Post-genomic vision: whole system picture



# Gene expression profiling.

## Historic perspective

Differences at phenotype level are the visible cause of differences at molecular level which, in many cases, can be detected by measuring the levels of gene expression. The same holds for different experiments, treatments, strains, etc.



- **Classification of phenotypes / experiments.** Can I distinguish among classes (either known or unknown), values of variables, etc. using molecular gene expression data? (**sensitivity**)
- **Selection of differentially expressed genes** among the phenotypes / experiments. Did I select the relevant genes, all the relevant genes and nothing but the relevant genes? (**specificity**)
- **Biological roles the genes are carrying out in the cell.** What general biological roles are really represented in the set of relevant genes? (**interpretation**)

# Microarrays arrive to an acceptable level of reproducibility

nature  
biotechnology

OCTOBER 2006  
www.nature.com/nbt/journal/v24/n10s

Produced with support from



United States  
Environmental Protection  
Agency



Agilent Technologies

The MicroArray Quality Control Consortium

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 Ir

Dirección <http://www.fda.gov/bbs/topics/NEWS/2007/NEW01555.html>

---



[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.

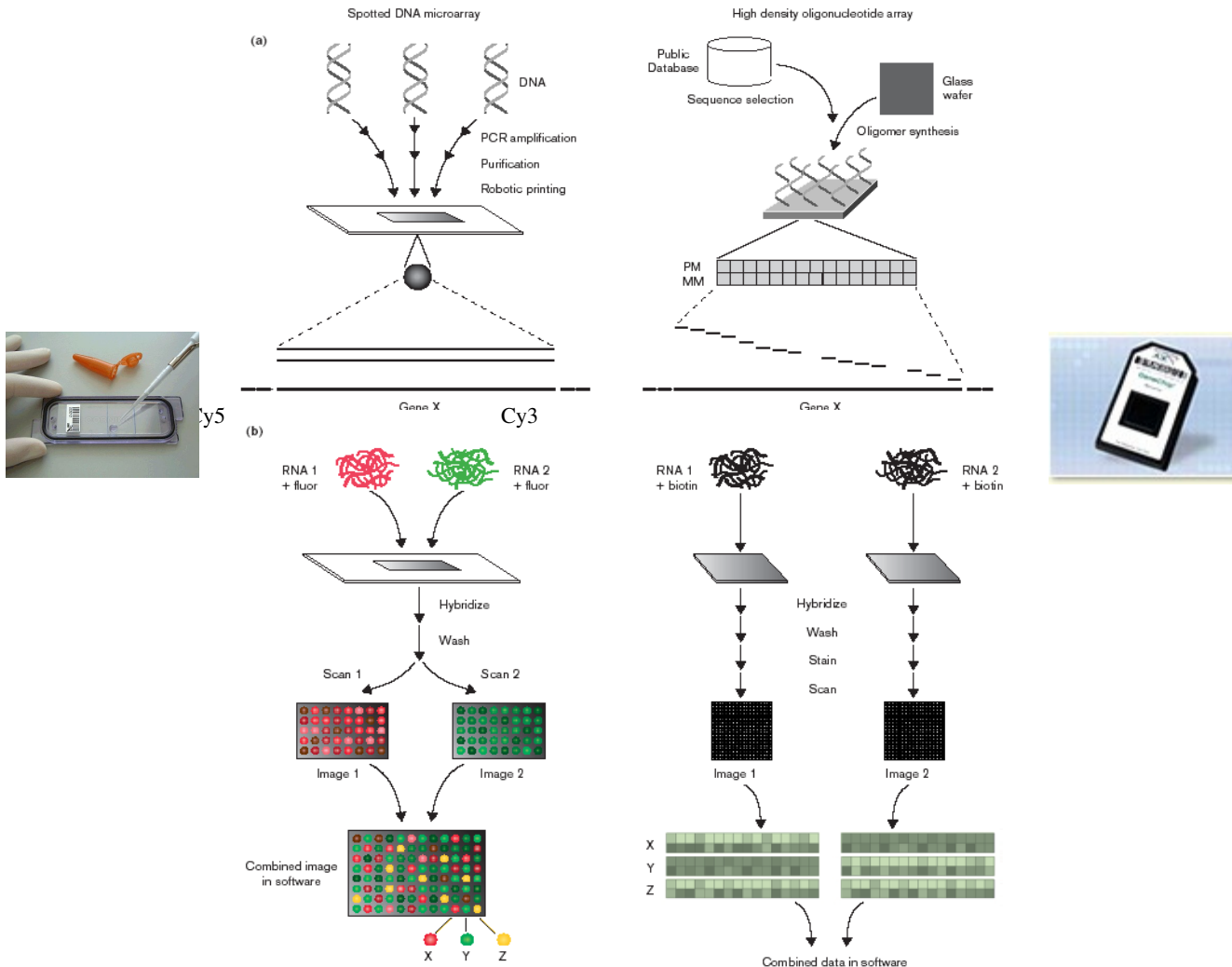
The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device. Several months ago, FDA issued a draft guidance document concerning the need for these complex molecular tests to meet pre-market review and post-market device requirements even when the tests are developed 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

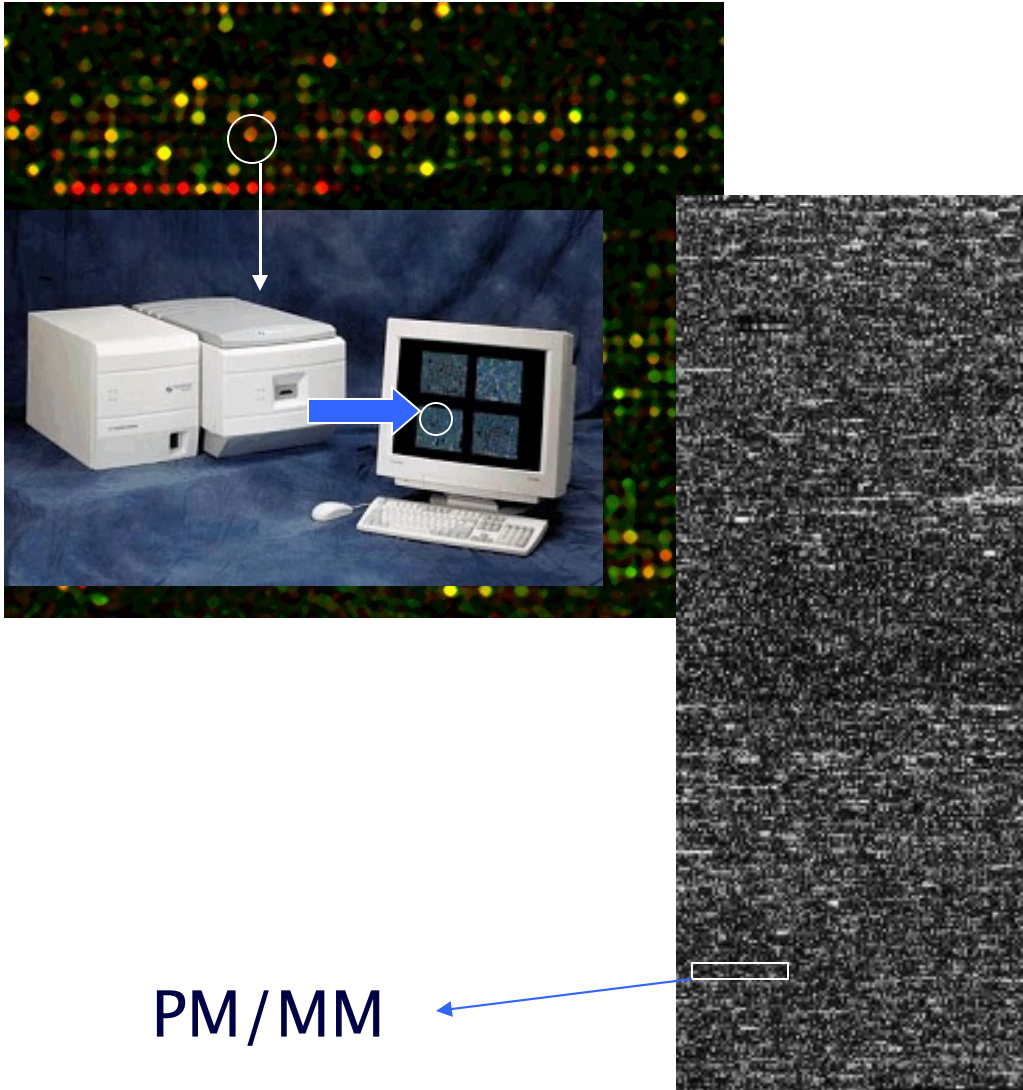
Listo Internet

# DNA microarrays: the paradigm of a post-genomic technique





# Transforming images into numbers



## Two-color

Test sample labeled red (**Cy5**)

Reference sample labeled green (**Cy3**)

Red : gene overexpressed in test sample

Green : gene underexpressed in test sample

**Yellow** – equally expressed

**red/green** – ratio of expression

## One color

**Intensity** of a gene using the probes

## Affymetrix

**Intensity** of a gene using the probes  
PM and in MM

Scanners generate a graphic file.

Software analyzes the file: GenePix Pro  
(by Axon Instruments, Inc.) or Imagene  
(By Biodiscovery, Inc.)

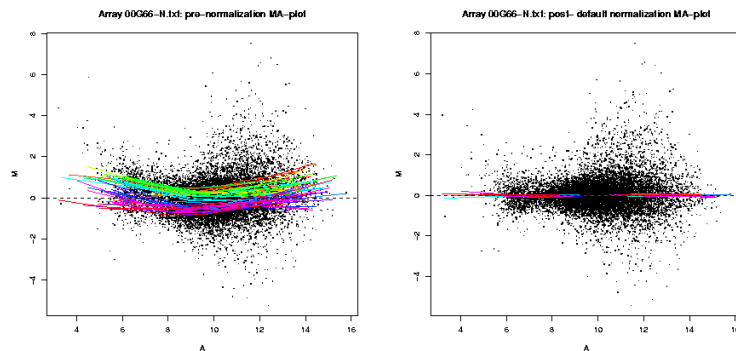
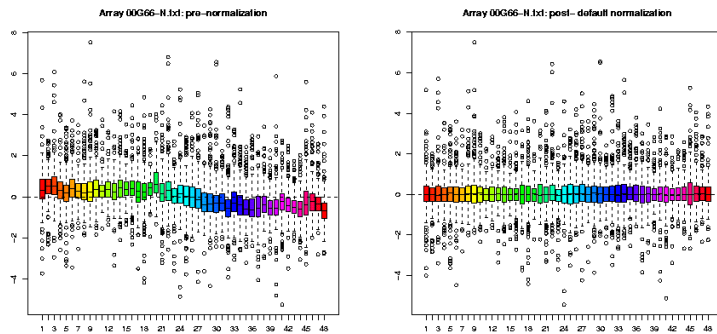
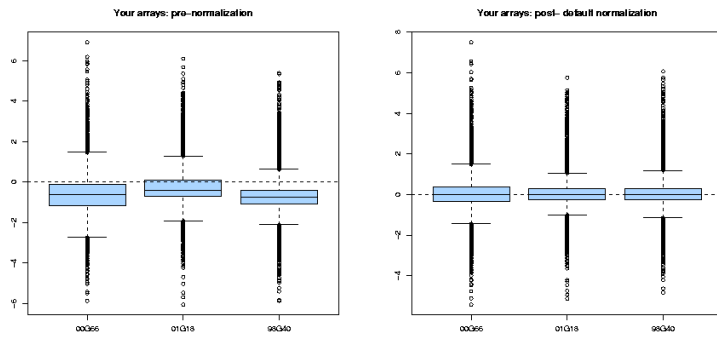
There are free systems too: TIGR  
Spotfinder, ScanAlyze, etc

# Normalisation

There are many sources of error that can affect and seriously bias the interpretation of the results.

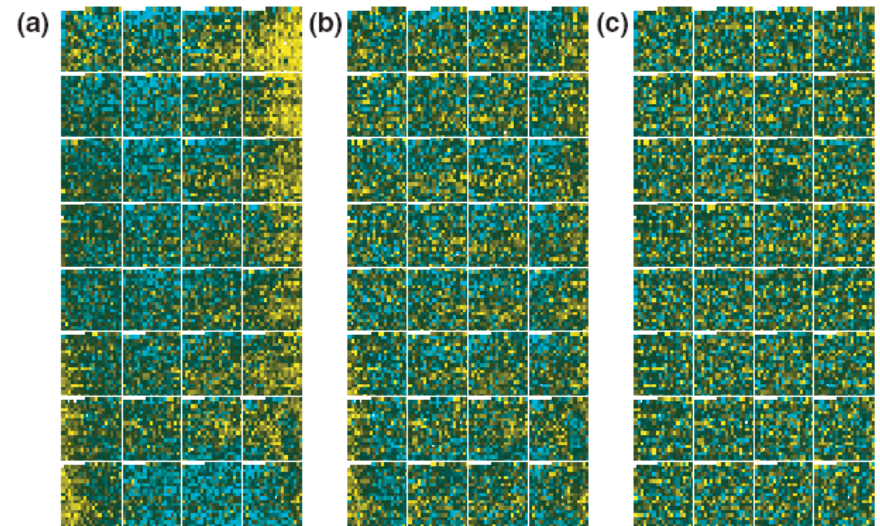
Differences in the efficiency of labelling, the hybridisation, local effects, etc.

Normalisation is a necessary step before proceeding with the analysis

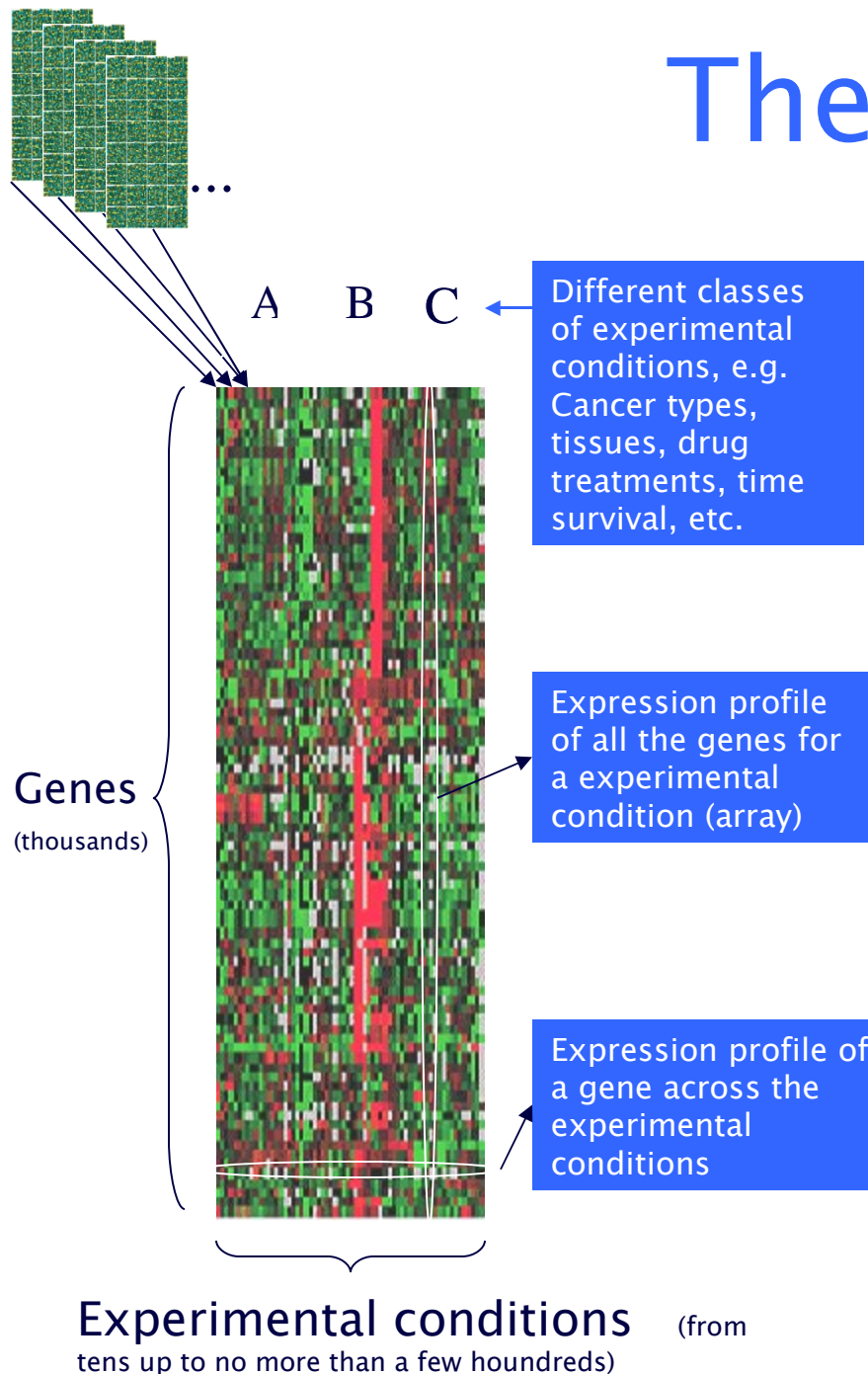


Before (left) and after (right) normalisation. A) BoxPlots, B) BoxPlots of subarrays and C) MA plots (ratio versus intensity)

(a) After normalization by average (b) after print-tip lowess normalization (c) after normalisation taking into account spatial effects



# The data



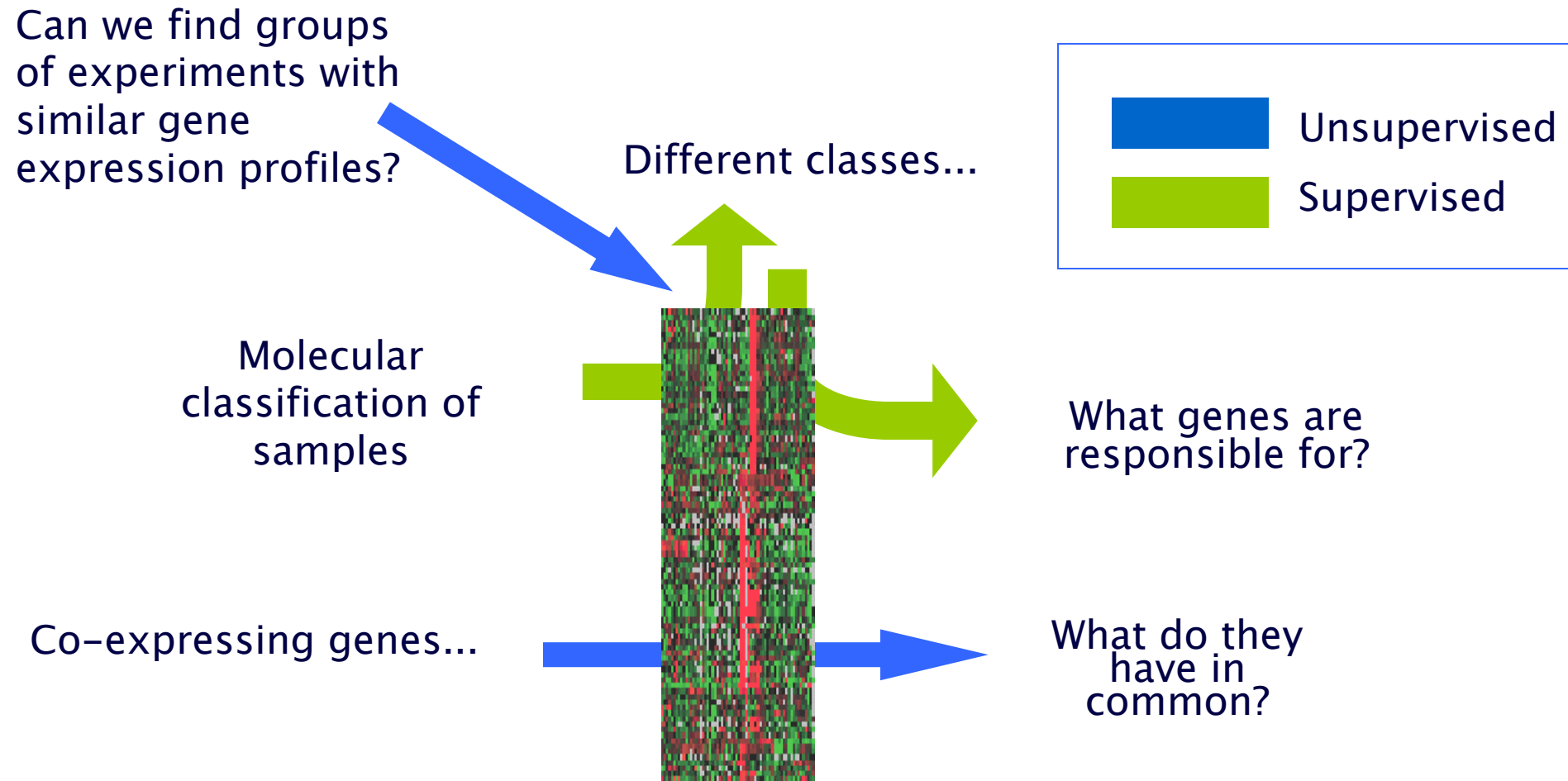
## Characteristics of the data:

- We NEVER deal with individual arrays, we deal with collections of arrays obtained for a given experimental design
- Most of the genes are not informative with respect to the trait we are studying (account for unrelated physiological conditions, etc.)
- Number of variables (genes) is several orders of magnitude larger than the number of experiments
- Low signal to noise ratio



# Studies must be hypothesis driven.

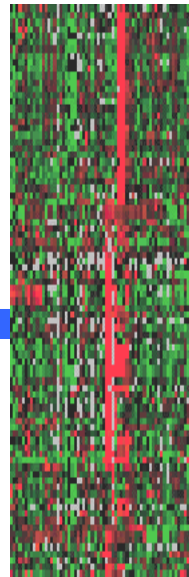
What is our aim? Class discovery? sample classification? gene selection? ...



# Unsupervised problem: class discovery

Our interest is in discovering clusters of items (genes or experiments) which we do not know beforehand

Can we find groups of experiments with similar gene expression profiles?



Co-expressing genes...



- What genes co-express?
- How many different expression patterns do we have?
- What do they have in common?
- Etc.

# Unsupervised clustering methods:

## Method + distance: produce groups of items based on its global similarity

Non hierarchical

hierarchical

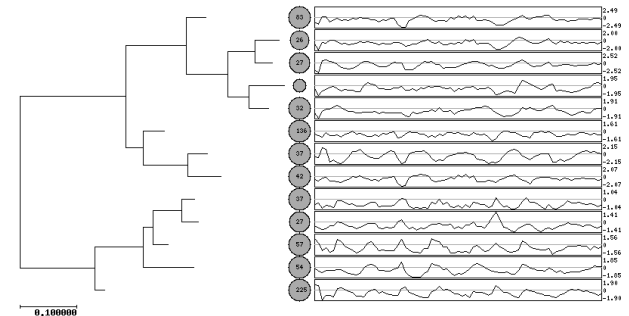
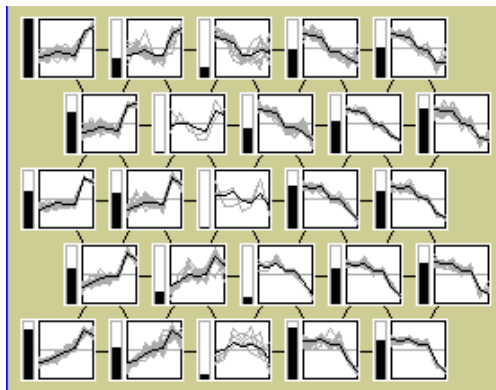
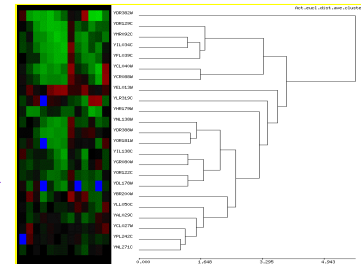
K-means, PCA

SOM

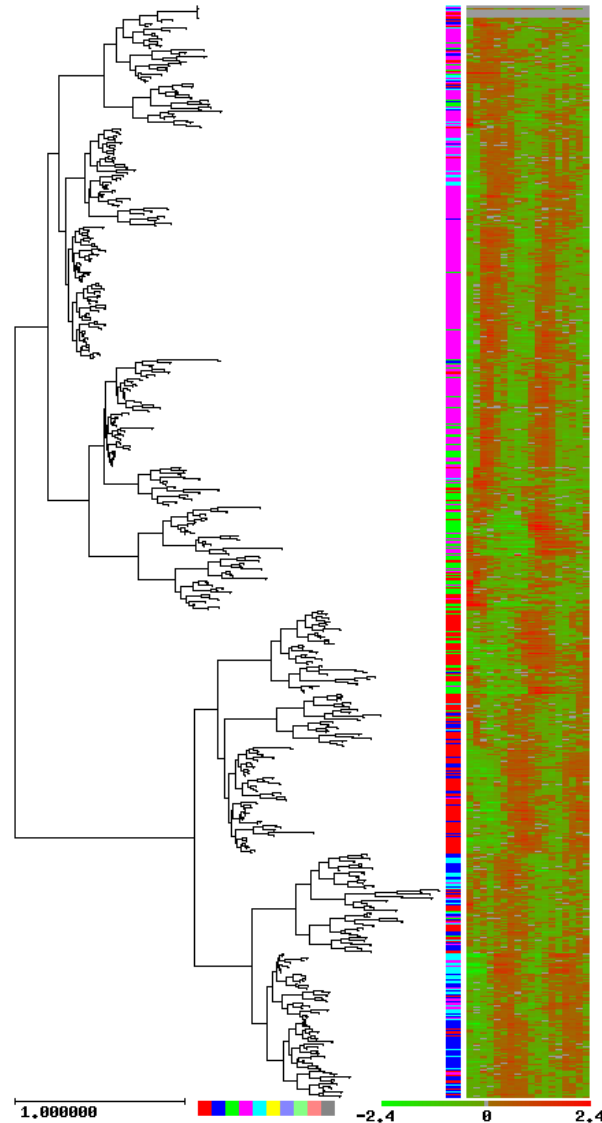
UPGMA

SOTA

Different  
levels of  
information



# An unsupervised problem: clustering of genes.



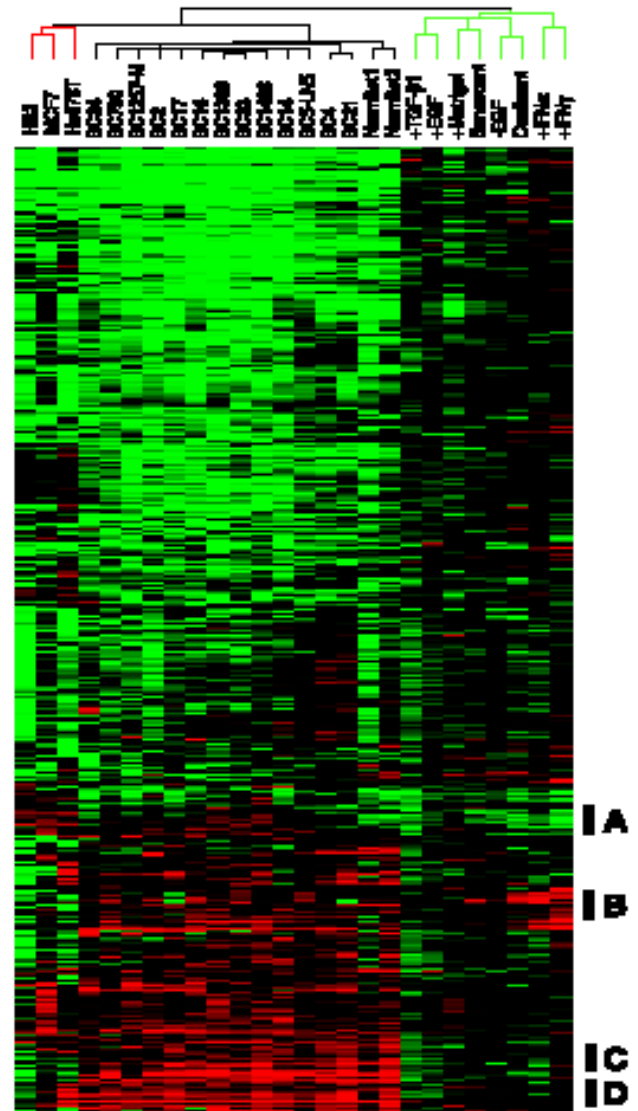
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

# Clustering of experiments: The rationale

If enough genes have their expression levels altered in the different experiments, we might be able of finding these classes by comparing gene expression profiles.

## Distinctive gene expression patterns in human mammary epithelial cells and breast cancers

Overview of the combined *in vitro* and breast tissue specimen cluster diagram. A scaled-down representation of the 1,247-gene cluster diagram. The black bars show the positions of the clusters discussed in the text: (A) proliferation-associated, (B) IFNregulated, (C) B lymphocytes, and (D) stromal cells.



Perou et al., PNAS 96 (1999)

# Clustering of experiments: The problems

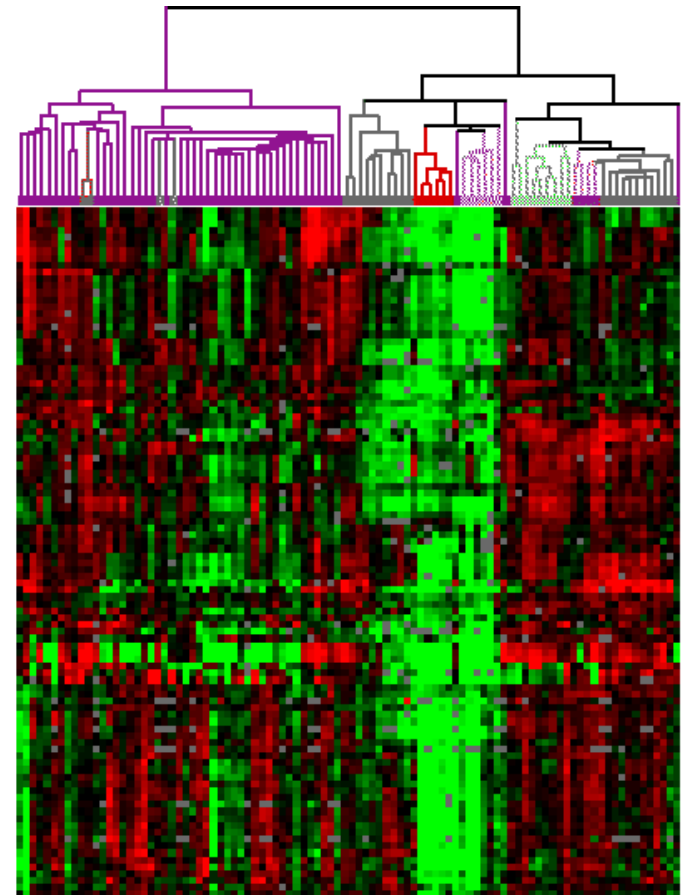
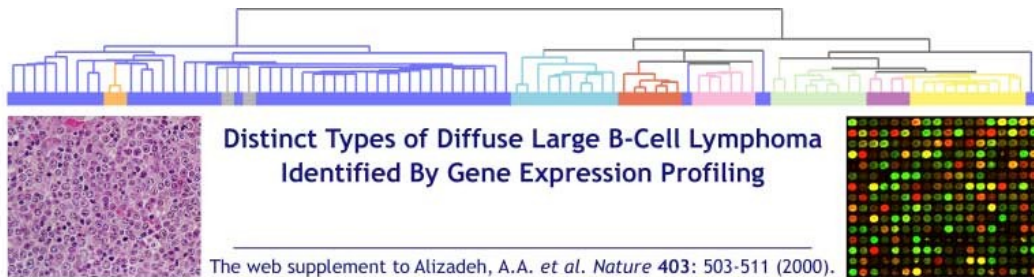
Any gene (regardless its relevance for the classification) has the same weight in the comparison.

If relevant genes are not in overwhelming majority we will find:

Noise

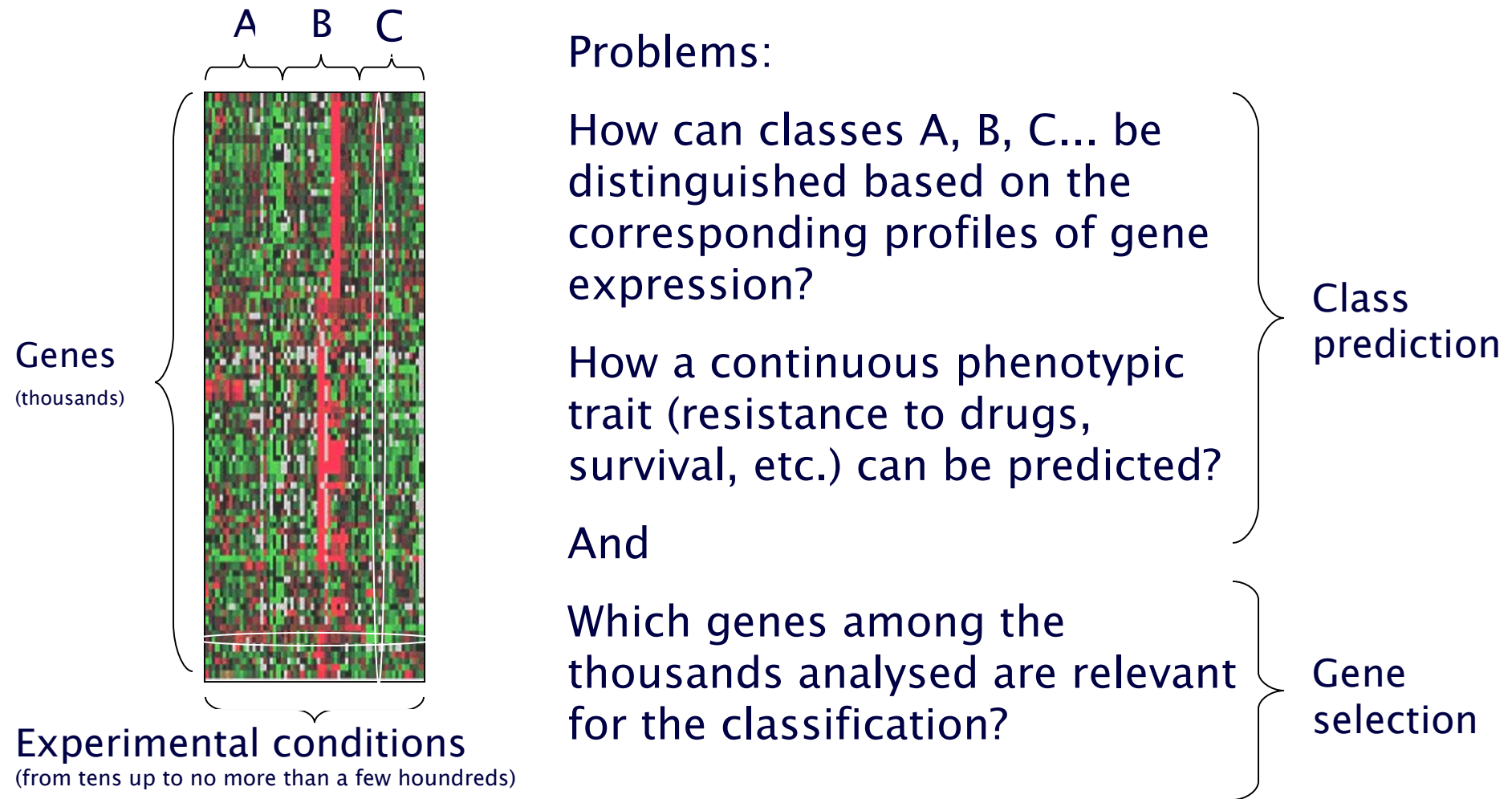
and/or

irrelevant trends



# Supervised problems: Class prediction and gene selection, based on gene expression profiles

Information on classes (defined on criteria external to the gene expression measurements) is used.



# Gene selection.

The simplest way: univariant gene-by-gene. Other multivariant approaches can be used

- **Two classes**

- T-test
  - Bayes
  - Data-adaptive
  - Clear

- **Multiclass**

- Anova
  - Clear

- **Continuous variable (e.g. level of a metabolite)**

- Pearson
  - Spearmam
  - Regression

- **Survival**

- Cox model

The screenshot shows the T-Rex form within the GEPA web interface. The browser window title is "GEPAS - T-Rex : form - Microsoft Internet Explorer". The address bar shows "http://t-rex.bioinfo.cipf.es/cgi-bin/t-rex.cgi". The page header includes navigation links: "normalization | preprocessing | clustering | supervised classification | differential expressions | functional annotation | cgh arrays | viewers". The main heading is "Tools > Differential expression > T-Rex". Below this is a "T-Rex : form" section with tabs for "two classes", "multi classes", "correlation", and "survival". The "two classes" tab is selected. The form contains several input fields and radio buttons: "Expression data" (with an "Examiner..." button), "Class labels" (with an "Examiner..." button), "Test" (with radio buttons for "t-test", "Bayes", "Data adaptive", and "CLEAR test"), "Significance level" (set to 0.05), "Image appearance" (with radio buttons for "yes" and "no"), "Standardize" (with radio buttons for "yes" and "no"), "Rows" (set to 100), "Scale" (set to -3/+3), "Project name (optional)", and "E-mail (optional)". At the bottom of the form are "Reset" and "Submit" buttons, with a "Run" button next to "Submit". Below the form is a "References:" section with three citations. The footer of the page provides contact information for the Centro de Investigación Príncipe Felipe, CIPF.

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección <http://t-rex.bioinfo.cipf.es/cgi-bin/t-rex.cgi>

Links [Ensembl Genome Browser](#) [NCBI HomePage](#) [Google Scholar](#) [Bioinformatics - Manuscript Central \[TM\]](#) Norton Internet Security

**GEPAS**  
Gene Expression Pattern Analysis Suite v3.0  
Bioinformatics Department - CIPF

normalization | preprocessing | clustering | supervised classification | differential expressions | functional annotation | cgh arrays | viewers

Tools > Differential expression > T-Rex

**T-Rex : form**

two classes multi classes correlation survival

Expression data Examiner...

Class labels Examiner...

Test

t-test Bayes Data adaptive CLEAR test

Significance level 0.05

Image appearance Standardize yes no Rows 100 Scale -3/+3

Project name (optional)

E-mail (optional)

Reset reset

Submit Run

References:

Vaquerizas J.M., Conde L., Yankilevich P., Cabezon A., Minguez P., Diaz-Urriarte R., Al-Shahrour F., Herrero J. & Dopazo J. (2005). [Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data](#). *Nucleic Acids Research* 33 (Web Server issue): W616-W620.

Herrero J., Vaquerizas J.M., Al-Shahrour F., Conde L., Mateos Á., Santoyo J., Diaz-Urriarte R. & Dopazo J. (2004). [New challenges in gene expression data analysis and the extended GEPAS](#). *Nucleic Acids Research* 32 (Web Server issue): W485-W491.

Herrero J., Al-Shahrour F., Diaz-Urriarte R., Mateos Á., Vaquerizas J.M., Santoyo J. & Dopazo J. (2003). [GEPAS, a web-based resource for microarray gene expression data analysis](#). *Nucleic Acids Research* 31(13): 3461-3467

Centro de Investigación Príncipe Felipe, CIPF - Avda. Autopista del Saler, 16 - 46013 Valencia - Spain - +34 96 328 96 80

Internet

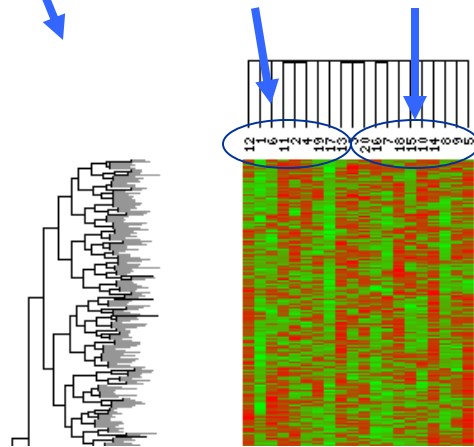
The T-rex tool



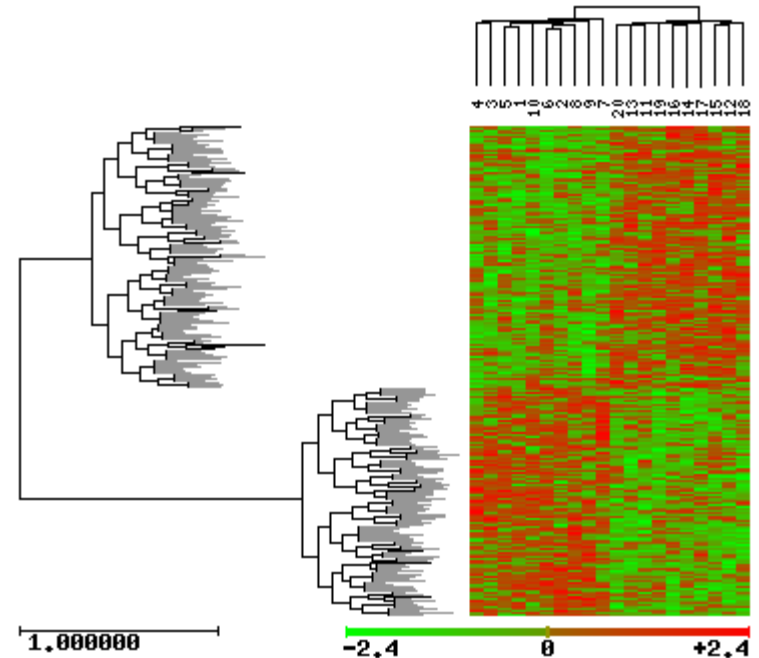
# A simple problem: gene selection for class discrimination

~15,000 genes

Case(10)/control(10)



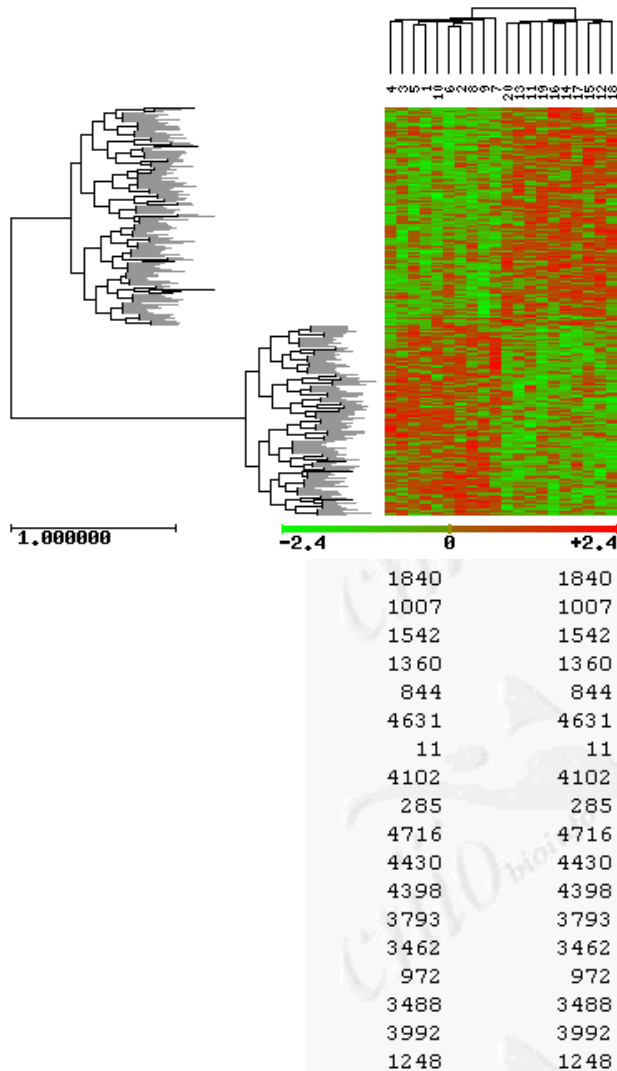
thebest - [04/10/2003 18:57:43 GMT]



Genes differentially expressed  
among classes (t-test), with p-  
value < 0.05

# Sorry... the data was a collection of random numbers labelled for two classes

thebest - [04/10/2003 18:57:43 GHT]



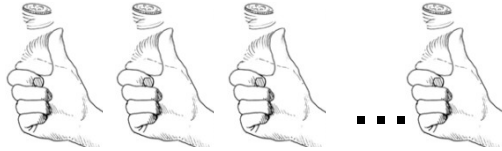
So... Why do we find good p-values?

unadj.p	adj_p	FDR_indep	FDR_dep	obs_stat
0.00019998	0.152685	0.49995	1	5.47044
0.00019998	0.746225	0.49995	1	4.49902
0.0009999	0.983002	0.861025	1	4.01726
0.00149985	0.986401	0.861025	1	3.99374
0.00129987	0.9959	0.861025	1	3.86046
0.00169983	0.9996	0.861025	1	3.7251
0.00169983	0.9996	0.861025	1	3.66628
				62427
				60596
				58109
				52935
				43721
				41937
				41428
				.4025
				40212
				37412
0.00539946	1	0.8888	1	3.36813
0.00219978	1	0.861025	1	3.35909
0.0029997	1	0.861025	1	3.35235
0.00439956	1	0.8888	1	3.28286
0.00669933	1	0.8888	1	3.2427
0.00559944	1	0.8888	1	3.23225
0.00279972	1	0.861025	1	3.22175
0.00429957	1	0.8888	1	3.19595
0.0039996	1	0.8888	1	3.19547
0.0069993	1	0.8888	1	3.12957
0.00849915	1	0.8888	1	3.0987
0.00779922	1	0.8888	1	3.09834

You were not interested *a priori* in the first (whatever), best discriminant, gene.

Adjusted p-values must be used!

# On the problem of multiple testing



= 10 heads.  $P=0.5^{10}=0.00098$

Take one coin, flip it 10 times. Got 10 heads? Use it for betting



10 heads !!!

⋮



1000 coins

$$P = 1 - (1 - 0.5^{10})^{1000} = 0.62$$

It is not the same getting 10 heads with **my** coin than getting 10 heads in **one among** 1000 coins

Will you still use this coin for betting?

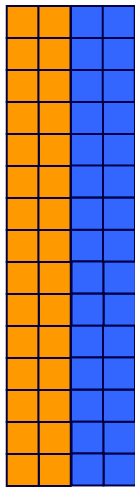
[illegible]

# Of predictors and molecular signatures

What is a predictor?

Intuitive notion:

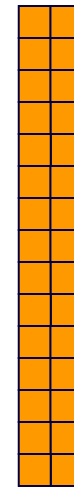
A B X



Is X, A  
or B?



$\text{Diff (B, X)} = 2$



$\text{Diff (A, X)} = 13$

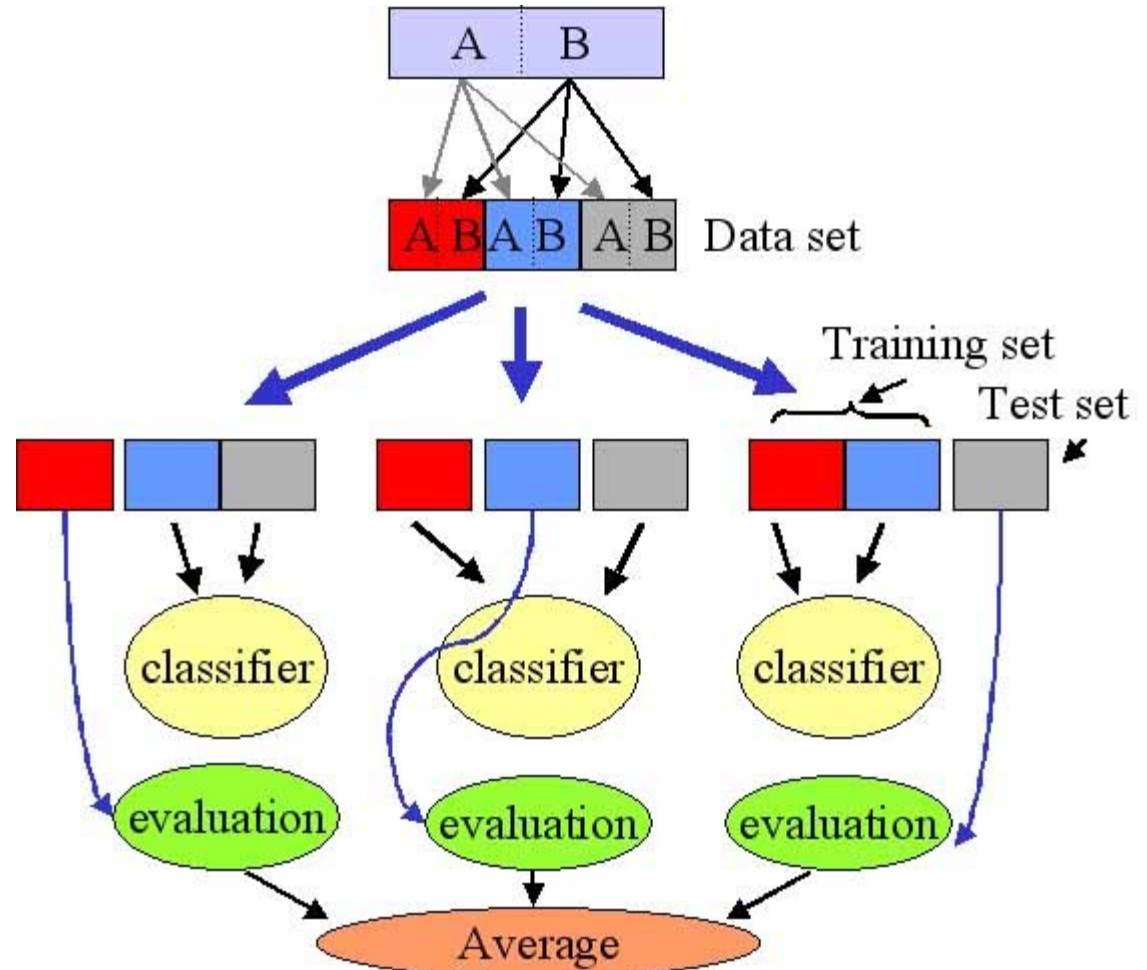
Most probably X belongs to class B

Algorithms: DLDA, KNN, SVM, random forests, PAM, etc.

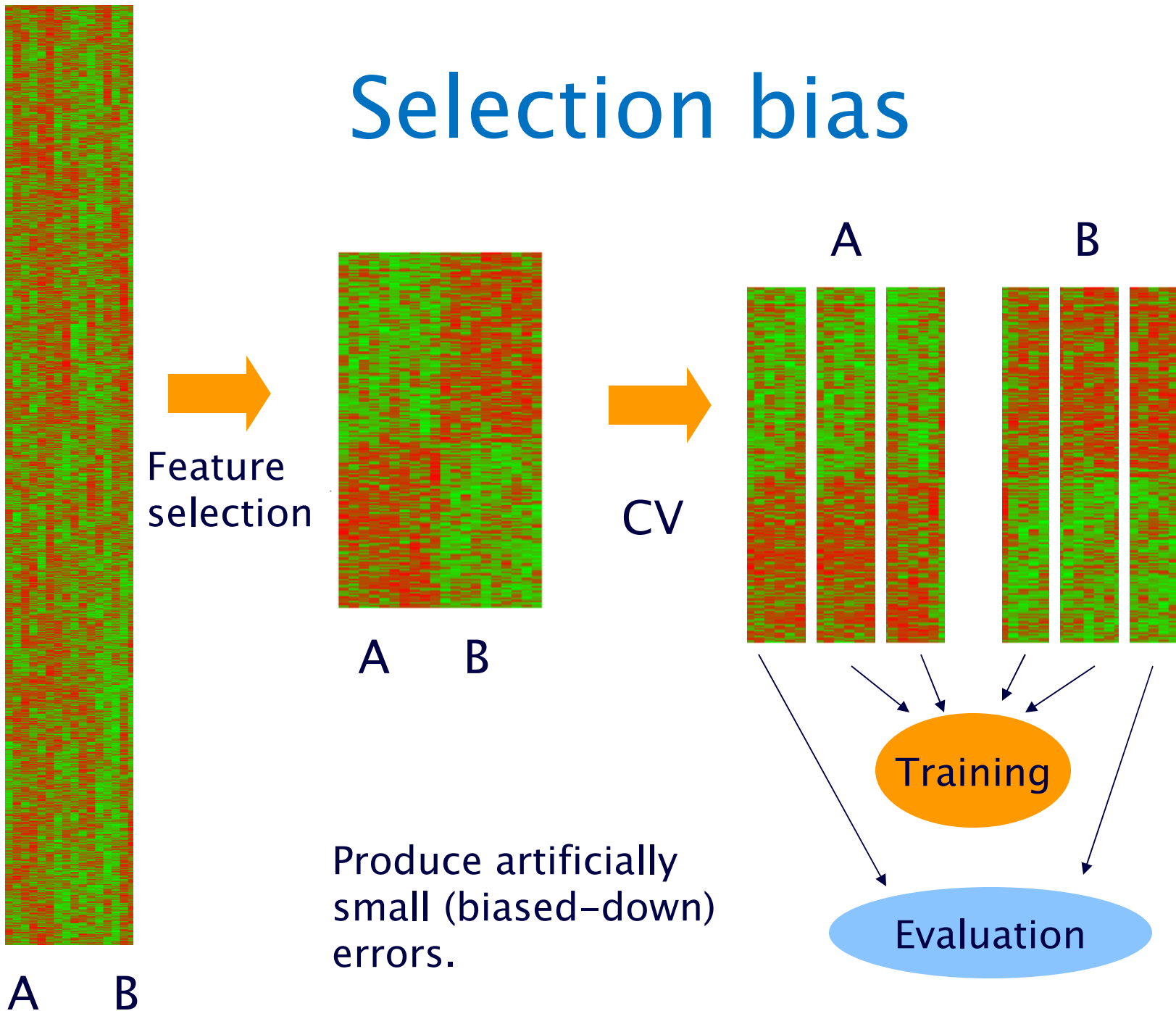
# Cross-validation

The efficiency of a classifier can be estimated through a process of cross-validation.

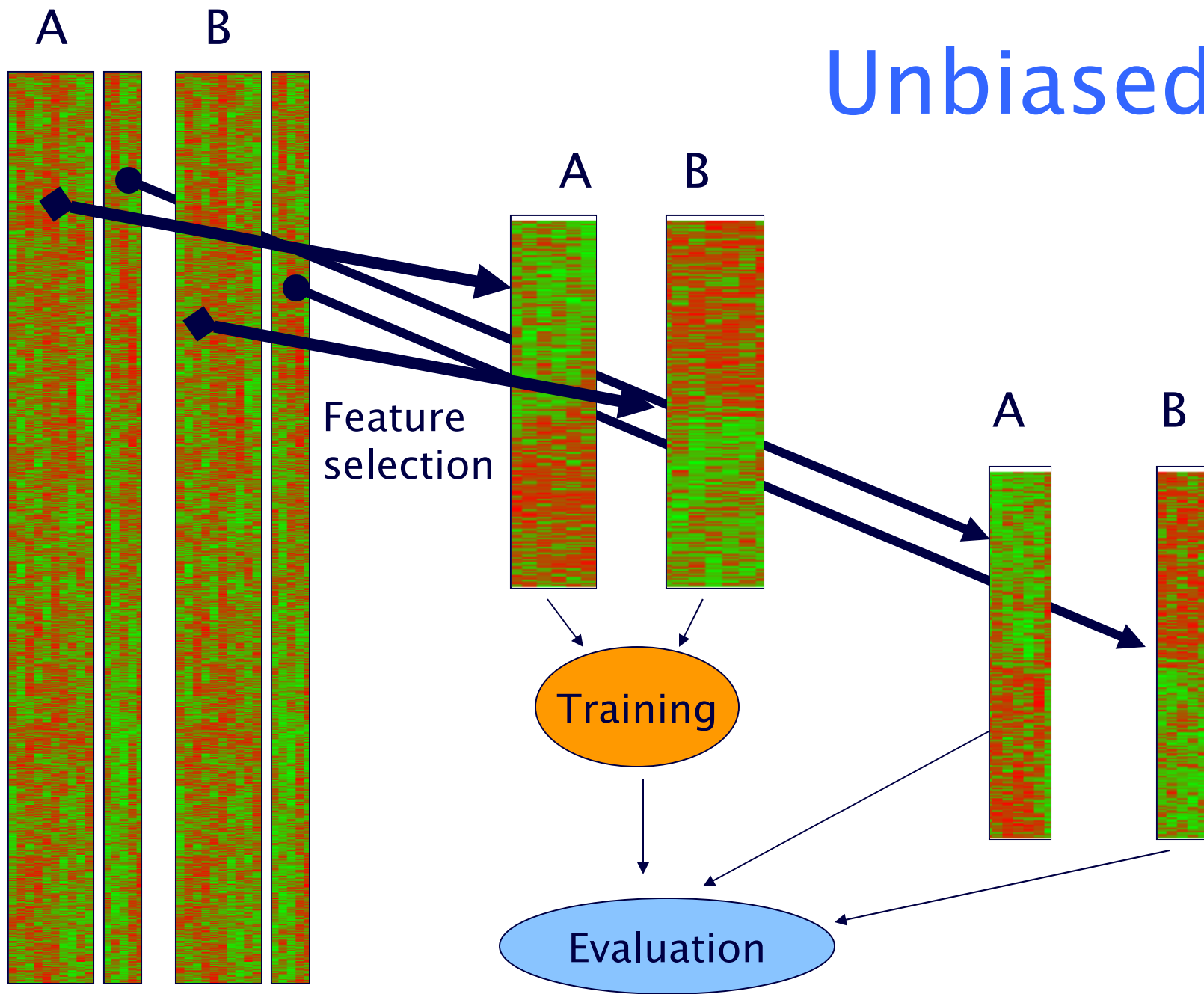
Typical are three-fold, ten-fold and leave-one-out (LOO), in case of few samples for the training



# Selection bias

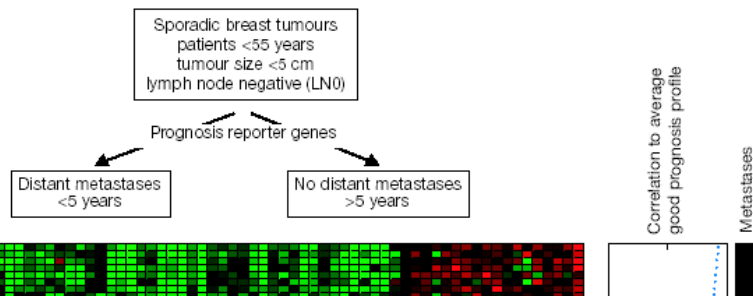


# Unbiased CV





# Predictor of clinical outcome in breast cancer



Genes are arranged to their correlation with the prognostic groups

The screenshot shows the Agendia website as viewed in a Microsoft Internet Explorer browser. The browser's address bar displays "http://www.agendia.com/". The website features the Agendia logo at the top left. The main content area has a blue background with the text "The Future is now" and "The future is now" in a stylized font. Below this, it says "diagnostics by genomics" and "Now". A large, stylized "Life" logo is visible in the background. A blue arrow points from the text "Pronostic classifier with optimal accuracy" to the website. The website also includes a section titled "Tumor profiling to improve cancer treatment" and a paragraph about Agendia's expertise in gene expression analysis.

Pronostic classifier with optimal accuracy

*van't Veer et al., Nature, 2002*

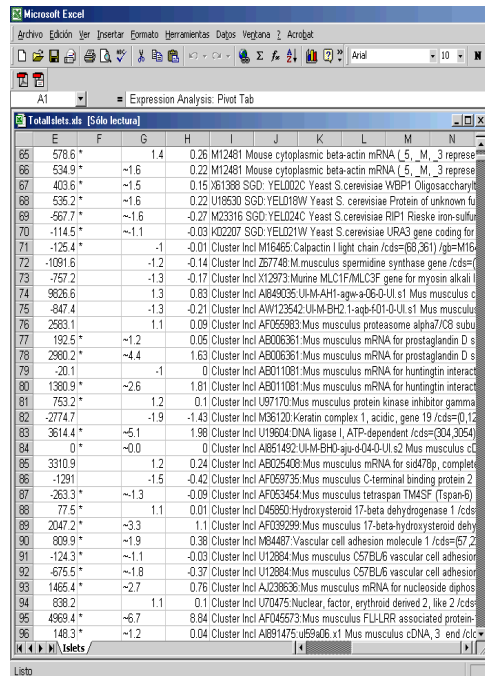
# Functional profiling of genome-scale experiments in the post-genomic era

My data...

How are structured?

What are these groups?

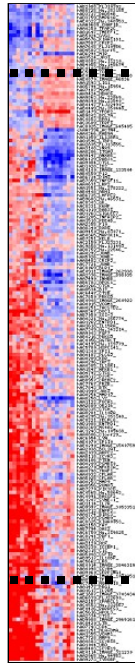
What is this gen?



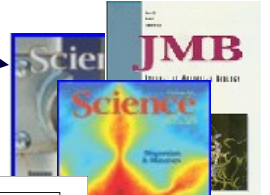
Microsoft Excel - Expression Analysis: Pivot Tab

	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)	represe			
66	534.9*	~1.6		0.22	M12481	Mouse cytoplasmic beta-actin mRNA (5, M, 3)	represe			
67	403.6*	~1.5		0.15	X61366	SIGD: YELU02C	Yeast S. cerevisiae WBP1 Oligosaccharyl			
68	535.2*	~1.6		0.22	U18530	SIGD: YELD18W	Yeast S. cerevisiae Protein of unknown fu			
69	-567.7*	~1.6		-0.27	M23316	SIGD: YELD24C	Yeast S. cerevisiae RIP1 Rieske iron-sulfur			
70	-114.5*	~1.1		-0.03	K02207	SIGD: YELD21W	Yeast S. cerevisiae URA3 gene coding for			
71	-125.4*		-1	-0.01	Cluster Incl M16465	Calpactin I light chain /cds=(88,361)	/g=M164			
72	-1091.6		-1.2	-0.14	Cluster Incl ZB746	M. musculus spermidine synthase gene /cds=(				
73	-757.2		-1.3	-0.17	Cluster Incl X12673	Mouse MLC1/FMLC3P gene for myosin alkali				
74	9626.6		1.3	0.83	Cluster Incl AB49035	U1-M-BH2-1-agg-01-0-UI s1	Mus musculus c			
75	-847.4		-1.3	-0.21	Cluster Incl AW123542	U1-M-BH2-1-agg-01-0-UI s1	Mus musculus			
76	2583.1		1.1	0.09	Cluster Incl AF059893	Mus musculus proteasome alpha7/08 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	1380.9*	~2.6		1.81	Cluster Incl AB011081	Mus musculus mRNA for huntingtin interact				
81	753.2*		1.2	0.1	Cluster Incl UB9170	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	U1-M-BH2-1-agg-01-0-UI s2	Mus musculus c			
85	3310.9		1.2	0.24	Cluster Incl AB025408	Mus musculus mRNA for sid478p, complet				
86	-1291	~1.5	-0.42	-0.42	Cluster Incl AF059735	Mus musculus C-terminal binding protein 2				
87	-263.3*	~1.3		-0.09	Cluster Incl AF053454	Mus musculus tetraspan TM4SF (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 AF039289	Mus musculus 17-beta-hydroxysteroid de				
90	809.9*	~1.9		0.38	Cluster Incl M64487	Vascular cell adhesion molecule 1 /cds=(67,2				
91	-124.3*	~1.1		-0.03	Cluster Incl U12884	Mus musculus C57BL/6 vascular cell adhesion				
92	-675.5*	~1.8		-0.37	Cluster Incl U12884	Mus musculus C57BL/6 vascular cell adhesion				
93	1465.4*	~2.7		0.76	Cluster Incl A123836	Mus musculus mRNA for nucleoside diphos				
94	838.2		1.1	0.1	Cluster Incl U70475	Nuclear, factor, erythroid derived 2, like 2 /cds=				
95	4369.4*	~6.7		8.84	Cluster Incl AF045673	Mus musculus FLJ-LRR associated protein-				
96	148.0*	~1.2		0.04	Cluster Incl AB91475	ucb9a06.x1	Mus musculus cDNA, 3' end /cds=			

A B

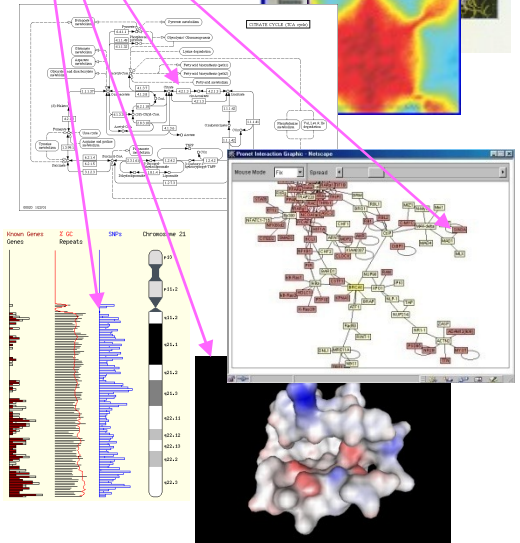


GeneCards™



Cell cycle...

DBs Information



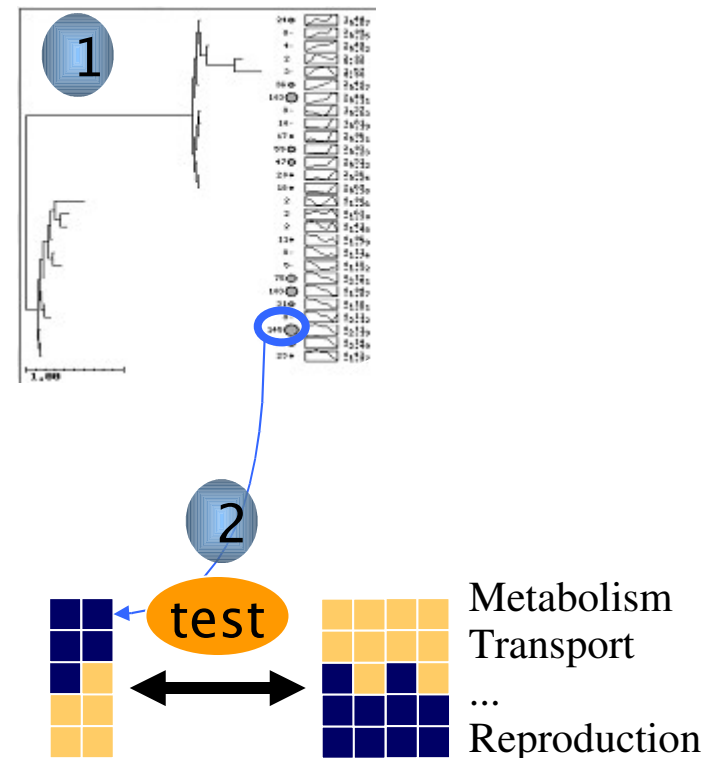
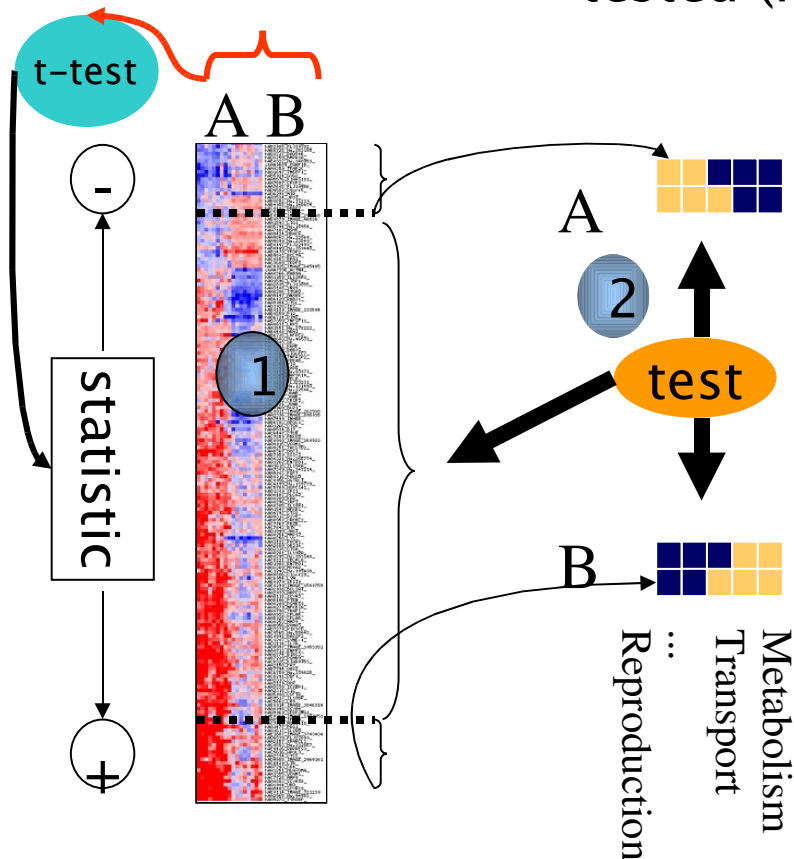
Analysis

Functional profiling

Links

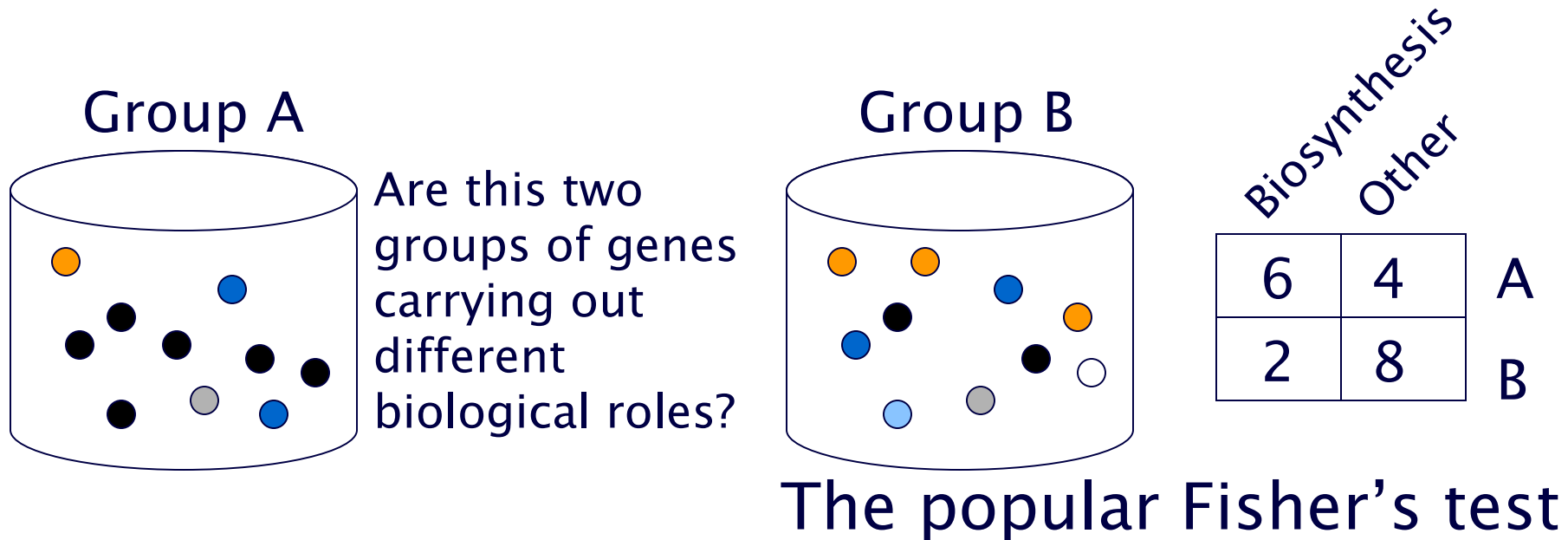
# Two-steps functional interpretation

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



# Testing two GO terms

(remember, we have to test thousands)

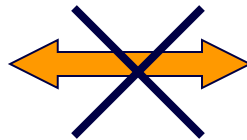


Biosynthesis 60% ●



Biosynthesis 20% ●

Sporulation 20% ●



Sporulation 20% ●

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

# GO terms found in sets of 50 genes

GO	Definition	p-value	Adjusted p-value
GO:0006790	sulfur metabolism	0.0595683	1
GO:0042592	homeostasis	0.0157944	0.300094
GO:0016265	death	0.116317	1
GO:0050874	organismal physiological process	0.151987	1
GO:0008152	metabolism	0.129865	1
GO:0019058	viral infectious cycle	0.016503	0.181353
GO:0019059	initiation of viral infection	0.0123062	0.459417
GO:0009056	catabolism	0.0276032	1
GO:0006766	vitamin metabolism	0.00875837	0.604328
GO:0007155	cell adhesion	0.122953	1

Each row corresponds to a random selection of 50 genes from the *E. coli* genome, compared with respect to the rest of the genome.

GO terms in blue (p-value < 0.05 in individual test) have asymmetrical distributions by chance (see adjusted p-values).

# How to test significant differences in the distribution of biological terms between groups of genes?

FatiGO: GO-driven data analysis

Provides a statistical framework able to deal with multiple-testing hypothesis

The left screenshot shows the Gene Ontology homepage in Microsoft Internet Explorer. The URL is <http://www.geneontology.org/>. The page features a navigation menu on the left, a main heading "Gene Ontology Home", and a "Popular Links" section. In the footer, under "GO website", the link "Tools for using GO" is circled in red. An orange arrow points from this link to the right screenshot.

The right screenshot shows the "Tools for Gene Expression Analysis" page in Microsoft Internet Explorer. The URL is <http://www.geneontology.org/GO.tools.microarray.shtml>. The page describes the **ermineJ** tool and the **FatiGO** tool. The **FatiGO** section includes the following text:

**FatiGO**  
Bioinformatics Department at the Centro de Investigacion Principe Felipe (Spain)  
[PubMed abstract]  
FatiGO assigns representative functional information (under-represented or over-represented Gene Ontology terms) to a given set of genes. Statistical significance is obtained using multiple-testing correction. FatiGO has been designed to be used by biologists with little or no informatics background. A command-line interface is available for users who wish to script the use of ermineJ. Several different methods for scoring gene sets are implemented, with a focus on methods that don't rely on simple "over-representation" measures.

**FuncAssociate**  
Roth Computational Biology Laboratory, Harvard Medical School  
[PubMed abstract]  
FuncAssociate is a web-based tool that accepts as input a list of genes, and returns a list of GO attributes that are over- (or under-) represented among the genes in the input list. Only those over- (or under-) representations that are statistically significant, after correcting for multiple hypotheses testing, are reported. Currently 10 organisms are supported. In addition to the input list of genes, users may specify a) whether this list should be regarded as ordered or unordered; b) the universe of genes to be considered by

*Al-Shahrour et al., 2004 Bioinformatics (3rd most cited paper in computing sciences. Source: ISI Web of knowledge.)*

*Al-Shahrour et al., 2005 Bioinformatics. Al-Shahrour et al., 2005 NAR*

*Al-Shahrour et al., 2006 NAR. Al-Shahrour et al., 2007 BMC Bioinformatics*

*Al-Shahrour et al., 2007 NAR*

# Compilation of tools for functional interpretation of sets of genes

Tool	Statistical model	Correction for multiple experiments	Functional labels	Site (web-based applications)	Reference
Babelomics	Fisher's exact test, t-test, Kolmogorov-Smirnov	FDR, q-value	GO, KEGG, protein domains, swissprot keywords, Transfac motifs, CisRed motifs, chromosomal location, tissues, bioentities (text-mining)	<a href="http://www.babelomics.org">http://www.babelomics.org</a>	(Al-Shahrour et al., 2006; Al-Shahrour et al., 2005)
BayGO	hypergeometric	bayesian	GO		(Vencio et al., 2006)
DAVID / EASEonline	Fisher's exact test	Bonferroni	GO, pathways, diseases, protein domains, interactions	<a href="http://david.abcc.ncifcrf.gov/">http://david.abcc.ncifcrf.gov/</a>	(Dennis et al., 2003; Hosack et al., 2003)
FatiGO+	Fisher's exact test	step-down minP, FDR	GO, KEGG, protein domains, swissprot keywords, Transfac motifs, CisRed motifs, chromosomal location, tissues	<a href="http://www.fatigo.org">http://www.fatigo.org</a>	(Al-Shahrour et al., 2004)
FuncSpec	hypergeometric	Bonferroni	GO, phenotypes, protein interactions, etc. (only for yeast)	<a href="http://funspec.med.utoronto.ca/">http://funspec.med.utoronto.ca/</a>	(Robinson et al., 2002)
GeneMerge	hypergeometric	Bonferroni	GO, KEGG, chromosomal location, other	<a href="http://genemerge.bioteam.net/">http://genemerge.bioteam.net/</a>	(Castillo-Davis & Hartl, 2003)
GO-TermFinder	hypergeometric	Bonferroni	GO		(Boyle et al., 2004)
GoMiner	Fisher's exact test	FDR	GO		(Zeeberg et al., 2003; Zeeberg et al., 2005)
Gostat	X2 Fisher's exact test	FDR, Holm	GO	<a href="http://gostat.wehi.edu.au/">http://gostat.wehi.edu.au/</a>	(Beissbarth & Speed, 2004)
GoSurfer	X2	q-value	GO		(Zhong et al., 2004)
GOToolBox	hypergeometric, binomial, Fisher's exact test	Bonferroni	GO	<a href="http://crfb.univ-mrs.fr/GOToolBox/index.php">http://crfb.univ-mrs.fr/GOToolBox/index.php</a>	(Martin et al., 2004)
Ontology Traverser	hypergeometric	FDR	GO	<a href="http://franklin.imgen.bcm.tmc.edu/rho-old/services/OntologyTraverser/">http://franklin.imgen.bcm.tmc.edu/rho-old/services/OntologyTraverser/</a>	(Young et al., 2005)
Onto-Tools	X2, binomial, hypergeometric Fisher's exact test	Sidak, Holm, Bonferroni, FDR	GO, KEGG	<a href="http://vortex.cs.wayne.edu/projects.htm">http://vortex.cs.wayne.edu/projects.htm</a>	(Draghici et al., 2003; Khatri et al., 2005)
FuncAssociate	Fisher's exact test	--	GO	<a href="http://llama.med.harvard.edu/cgi/func/funcassociate">http://llama.med.harvard.edu/cgi/func/funcassociate</a>	(Berriz et al., 2003)
GOTM	hypergeometric	--	GO	<a href="http://bioinfo.vanderbilt.edu/gotm/">http://bioinfo.vanderbilt.edu/gotm/</a>	(Zhang et al., 2004)
CL ENCH	Hypergeometric, X2, binomial	--	GO (only for <i>A. thaliana</i> )	--	(Shah & Fedoroff, 2004)





BABELOMICS

# Functional terms



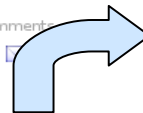
FatiGO+

Help | [References](#) | [Send comments](#)

[search](#)

**[compare](#)**

[genomics](#)



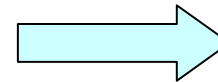
Organism

Organism



List of genes #1

genes list #1

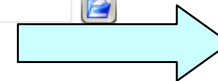


Gene List1

or genes list file #1

List of genes #2

genes list #2



Gene List2

or genes list file #2

Functional annotation

Gene Ontology: cellular component ☒

Gene Ontology: biological process ☒

Gene Ontology: molecular function ☒

InterPro motifs ☒

KEGG pathways ☒

SwissProt keywords ☒

Chemical terms bioalma ☒

Diseases terms bioalma ☒

Gene expression in Tissues ☒

Transcription factors ☒

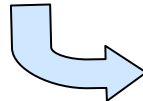
cisRED: cis-regulatory element ☒

E-mail (optional)

Project name (optional)

Submit

Run



Biological process  
Molecular function  
Cellular component  
KEGG pathways  
Interpro motifs  
Swissprot keywords  
Bioentities from literature (Marmite)  
Gene Expression (TMT)  
Transcription Factor binding sites  
Cis-regulatory elements (CisReD)



# GEPAS



# SOTA +

*Sati*

**GEPAS - SOTArray : basic form - Microsoft Internet Explorer**

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección: [http://gepas.bioinfo.ochoa.fib.es/cgi-bin/sotarray?dir=28031\\_vsudqkf&dir=data](http://gepas.bioinfo.ochoa.fib.es/cgi-bin/sotarray?dir=28031_vsudqkf&dir=data)

**GEPAS**  
Gene Expression Pattern Analysis Suite v2.0  
Bioinformatics Department - CIPF

normalization | preprocessing | clustering | supervised classification | differential expressions | functional annotation | cgh arrays | viewers

Tools > Clustering > SOTArray

**SOTArray : basic form**

Help ?  
References ?  
Send comments ?

**Data file** data already in server

**Cluster conditions** - None -

**Distance between genes** Correlation Coeff. (linear)

**End training conditions**

Unrestricted Grow ☐ threshold 0.00

Resource Threshold ☐ abs 0.00

Variability Threshold (abs)

Variability Threshold (%) ☐ % 90

Unconditional training ☐ Unconditional

Cycles before stopping

Switch mode Switch to Advanced Mode

Submit Run

Other servers available at:

**GEPAS - SOTArray : results - Microsoft Internet Explorer**

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección: <http://gepas.bioinfo.ochoa.fib.es/cgi-bin/sotarray>

**GEPAS**  
Gene Expression Pattern Analysis Suite v2.0  
Bioinformatics Department - CIPF

normalization | preprocessing | clustering | supervised classification | differential expressions | functional annotation | cgh arrays | viewers

Tools > Clustering > SOTArray

**SOTArray : results**

Using data already in Server...

Calculating the variability threshold... 0.465  
End of training in 0 min 0.10 sec.  
Total Resource: 22.4112 in 33 nodes  
Extracting profiles of SOTA leaves...  
Extracting clusters...

Input File: data.list  
Output File: data.sot  
Rich Full Newick File: data.nw  
Codebook File: data.cob  
Leaf Profile File: data.clusters.txt -> Send to the Preprocessor  
Cluster Text File: data.clu  
Cluster HTML File: data.clu.html

Other servers available at:

**GEPAS - SotaTree : basic form - Microsoft Internet Explorer**

Archivo Edición Ver Favoritos Herramientas Ayuda

Dirección: [http://gepas.bioinfo.ochoa.fib.es/cgi-bin/sotatree?dir=28031\\_vsudqkf&dir=data#sotatree](http://gepas.bioinfo.ochoa.fib.es/cgi-bin/sotatree?dir=28031_vsudqkf&dir=data#sotatree)

**GEPAS**  
Gene Expression Pattern Analysis Suite v2.0  
Bioinformatics Department - CIPF

normalization | preprocessing | clustering | supervised classification | differential expressions | functional annotation | cgh arrays | viewers

Tools > Viewers > SotaTree

**SotaTree : basic form**

Help ?  
References ?  
Send comments ?

**Data** data already in server

**Profiles** data - 11/06/2005 17:01:28 GMT

**Codebook vectors** codebook vectors already in server

Draw profiles as: Lines ☒ Histogram ☐ None ☐

Adjust profile scale ☒

Set the same scale for all the profiles ☒

Labels

Text Labels ☐

Frequency Labels ☐

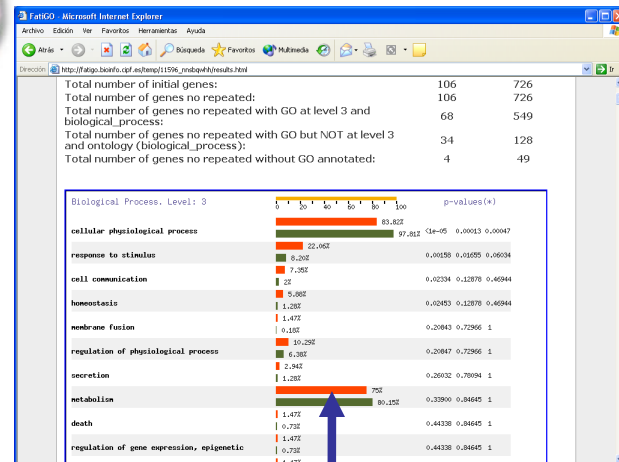
Percentage Labels ☐

Label keys:

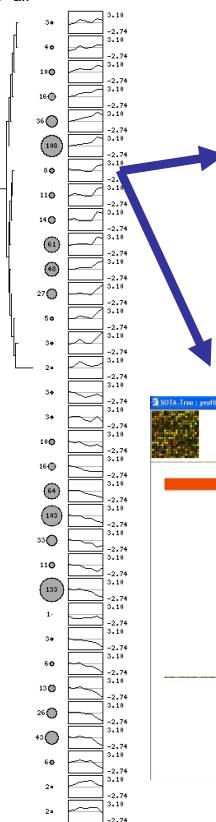
Clear form New

Advanced mode Switch

Submit Run



data - 11/06/2005 17:01:28 GMT





# Biological processes shown by the genes differentially expressed among PTL-LB

Total number of initial genes:  
 Total number of genes no repeated:  
 Total number of Cluster IDs retired - their current Cluster IDs  
 Total number of genes no repeated with current Cluster IDs:  
 Total number of genes no repeated with GO at level 3 and biological\_process:  
 Total number of genes no repeated with GO but NOT at level 3 and ontology:  
 Total number of genes no repeated without GO annotated:

Cluster Query	Cluster Reference
162	4764
129	4731
7 - 23	449 - 1627
145	5909
88	2610

Gene Ontology Term

response to external stimulus



36.36%

response to stress

11.65%

21.59%

6.86%

signal transduction

39.77%

26.05%

cell motility

9.09%

3.79%

resistance to pathogenic bacteria

1.14%

0.04%

viral replication

1.14%

0.15%

cell death

9.09%

5.75%

regulation of gene expression, epigenetic

1.14%

0.10%

Obvious? NO

2) You now know that there are no other co-variables (e.g. age, sex, etc)

3) If you do not have previously a strong biological hypothesis, now you have an explanation

0.1702 0.9912 1 1

0.1806 0.9940 1 1

# Weaknesses of the two-steps, functional enrichment approach

Low sensitivity of conventional gene selection methods

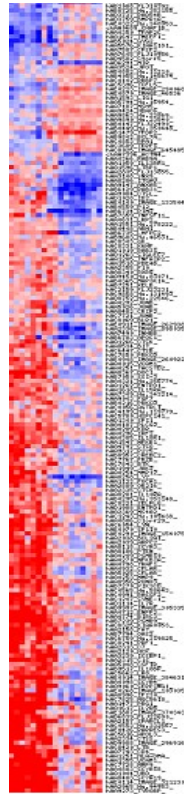
A B

A

8 with impaired tolerance (**IGT**)  
+ 18 with type 2 diabetes mellitus (**DM2**)

B

17 with normal tolerance to glucose (**NTG**)

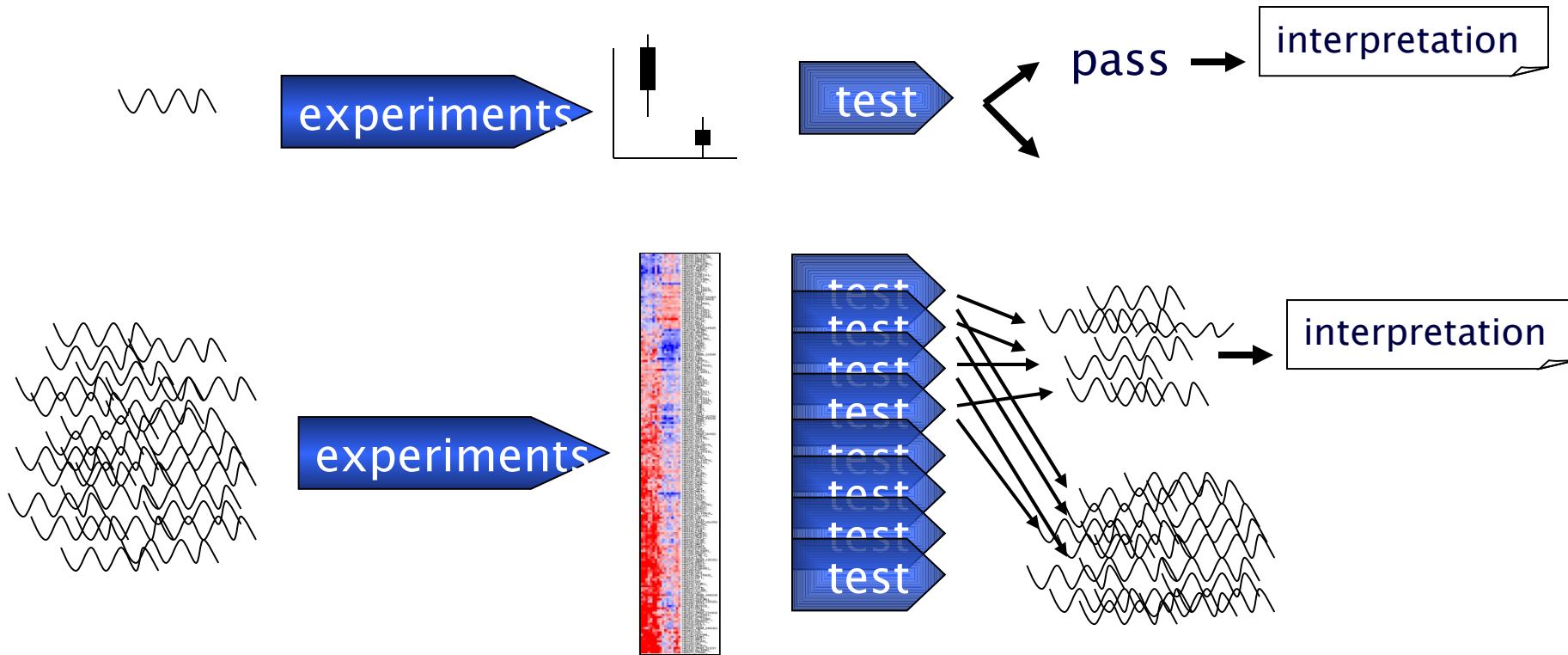


Instability of molecular signatures. Variable selection with microarray data can lead to many solutions that are equally good from the point of view of prediction rates, but that share few common genes (Ein-Dor 2006 PNAS)

Platform comparison. There are still some concerns with the cross-platform coherence of results. Paradoxically, despite the fact that gene-by-gene results are not always the same, the biological themes emerging from the different platforms are increasingly consistent (Bammiller 2005 Nat Methods)

(Mootha et al., 2003)

# Functional enrichment approach reproduces pre-genomics paradigms

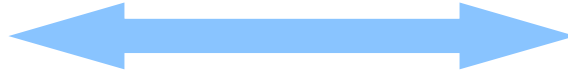
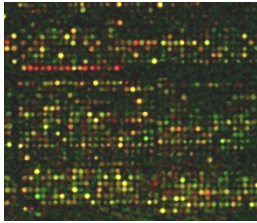


Context and cooperation between genes is ignored

# Functional genomics.

## Historic perspective and future

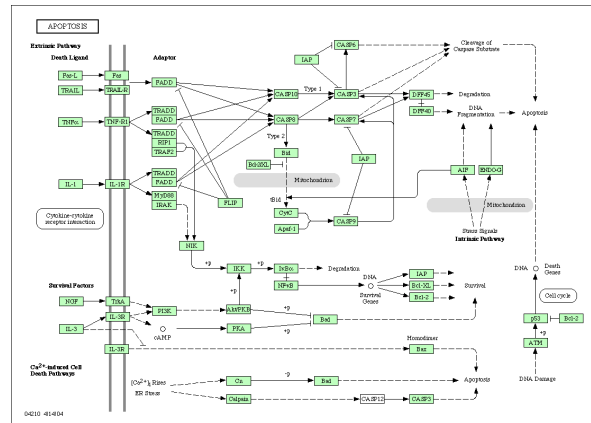
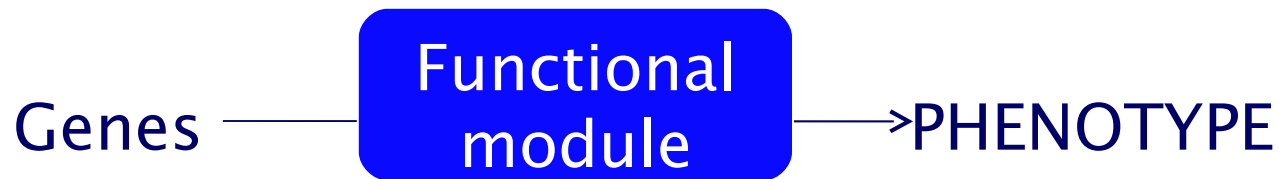
Differences at phenotype level are the visible cause of differences at molecular level which, in many cases, can be detected by measuring the levels of gene expression. The same holds for different experiments, treatments, strains, etc.



- Classification of phenotypes / experiments. **Sensitivity**
- Selection of differentially expressed genes **Specificity**
- Biological roles the genes are carrying out in the cell. **Interpretation**
- **Reformulating the questions.** Are we asking the proper questions? What are the real bricks that account for the cellular behaviour and for the phenotype or the response to environmental stimuli? The genes or other higher level units?

# What is the basic functional component in the cell?

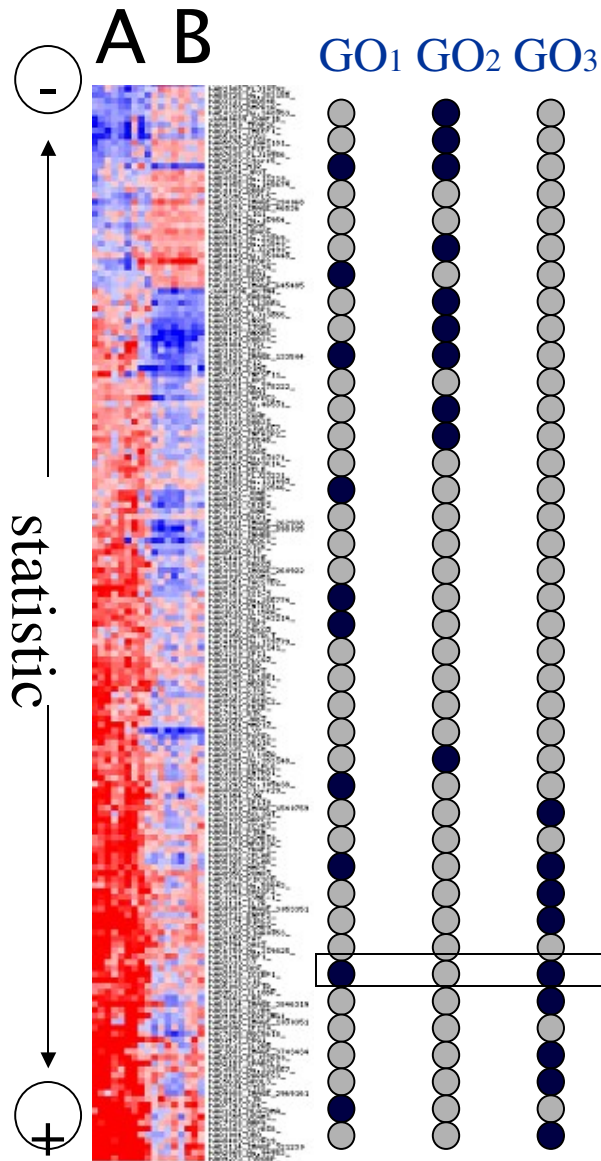
Genes —————> PHENOTYPE



e.g. KEGG  
pathway defines a  
functional module



# Cooperative activity of genes can be detected and related to a macroscopic observation



**Ranking:** A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

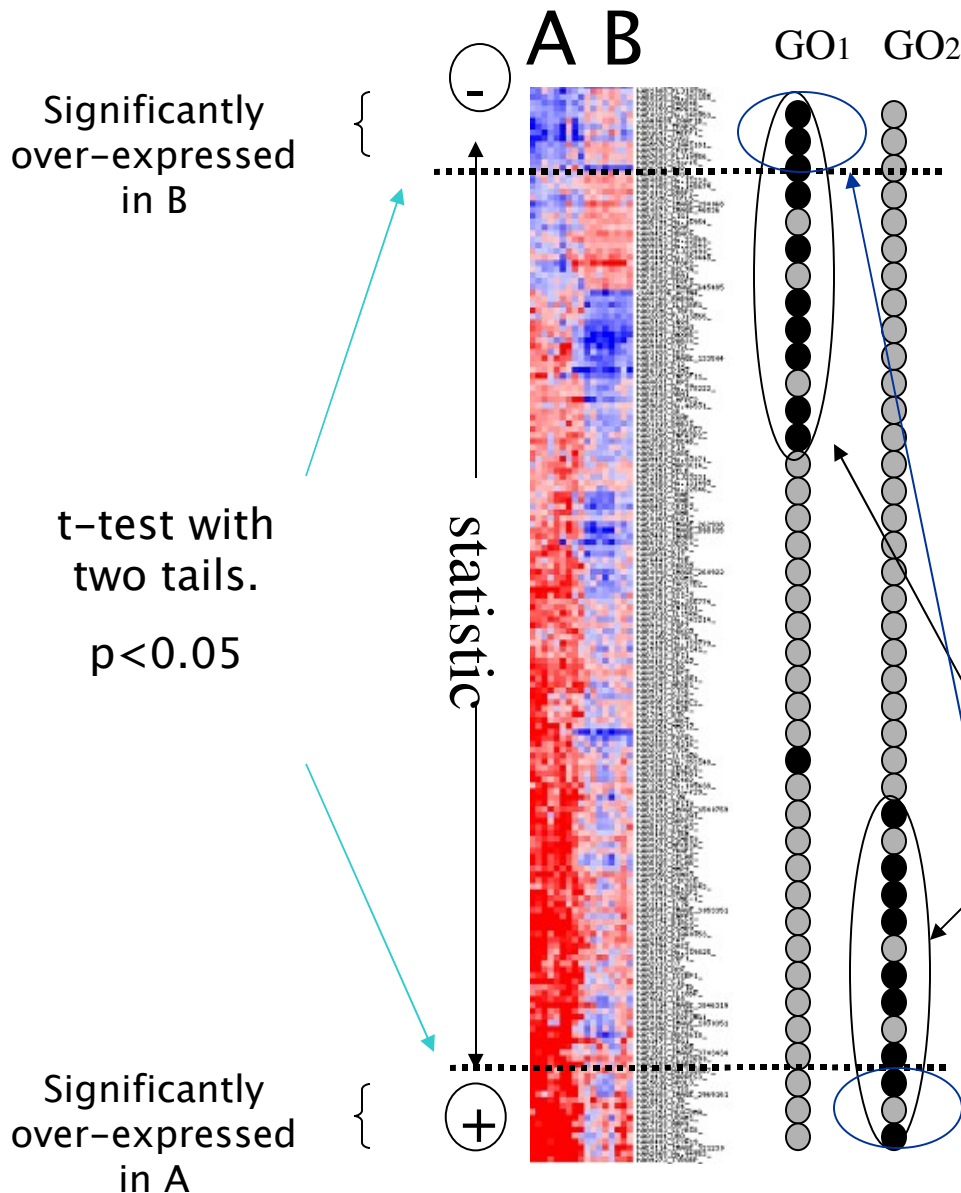
**Distribution of GO:** Rows GO1, GO2 and GO3 represent the position of the genes belonging to three different GO terms across the ranking.

The first GO term is completely uncorrelated with the arrangement, while GOs 2 and 3 are clearly associated to high expression in the experimental conditions **B** and **A**, respectively.

Note that genes can be multi-functional

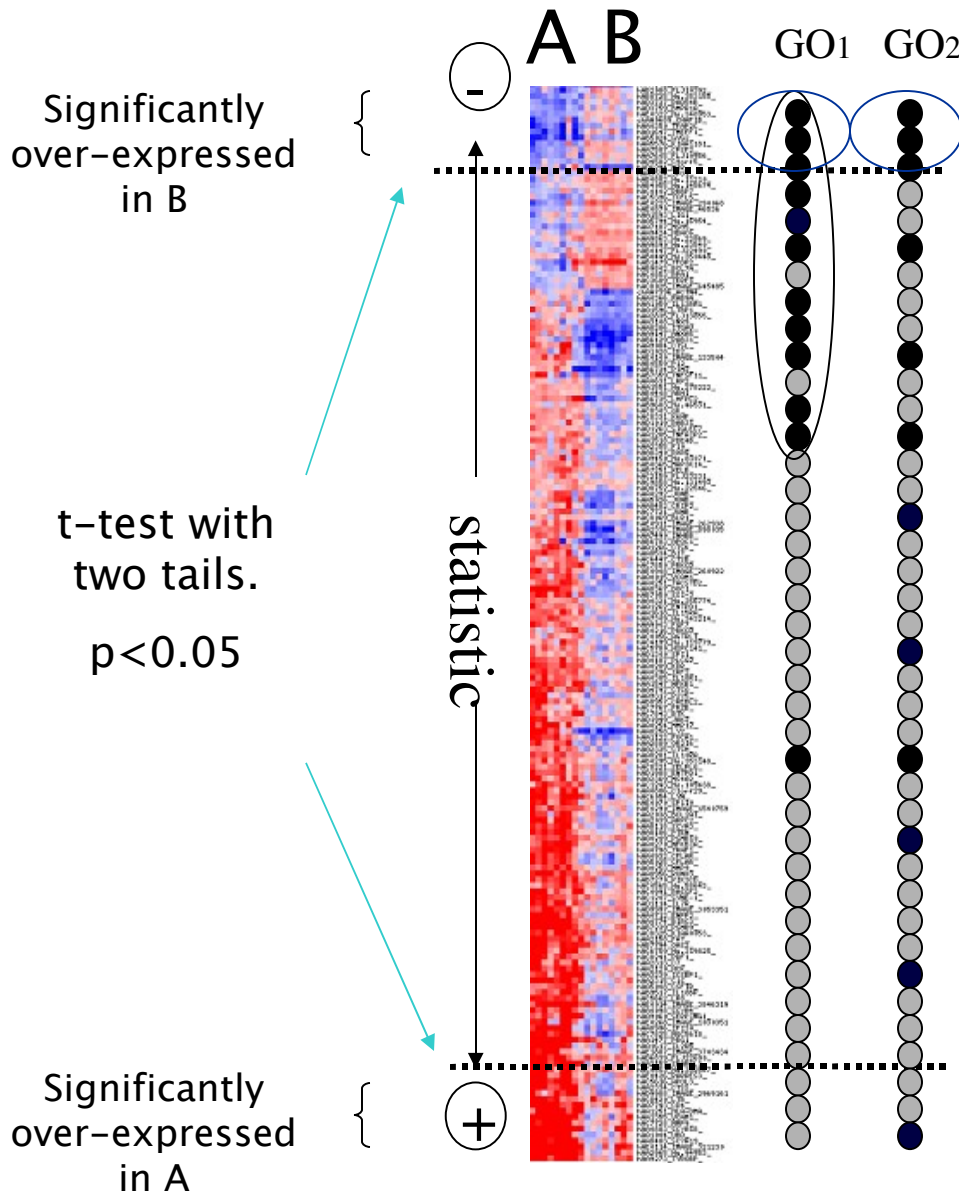


# A previous step of gene selection causes loss of information and makes the test insensitive



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

# A previous step of gene selection causes loss of information and makes the test insensitive



The main problem is that the two-steps approach cannot distinguish between these two different cases.

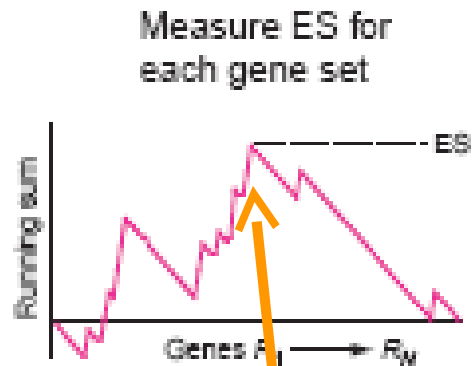
We put both sides of the partition into two bags and destroy the structure of the data.

	up	down
GO	3	9
no GO	0	25

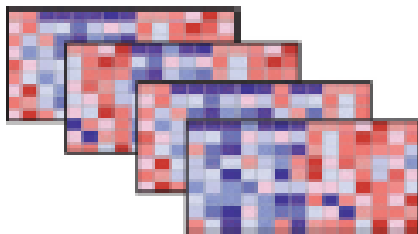
Same contingency table for GO<sub>1</sub> and GO<sub>2</sub> !!

# Gene-set enrichment methods

## GSEA



Permute class labels  
(1,000 times)



A B

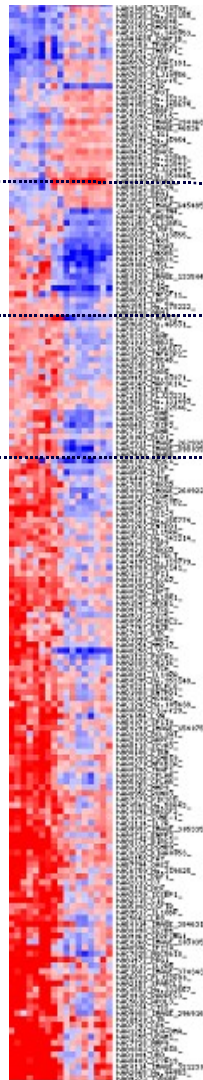
-

↑

statistic

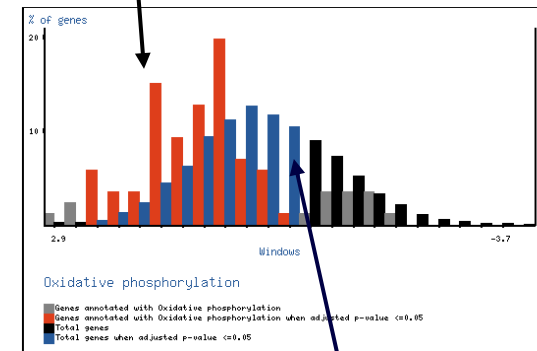
↓

+



## FatiScan

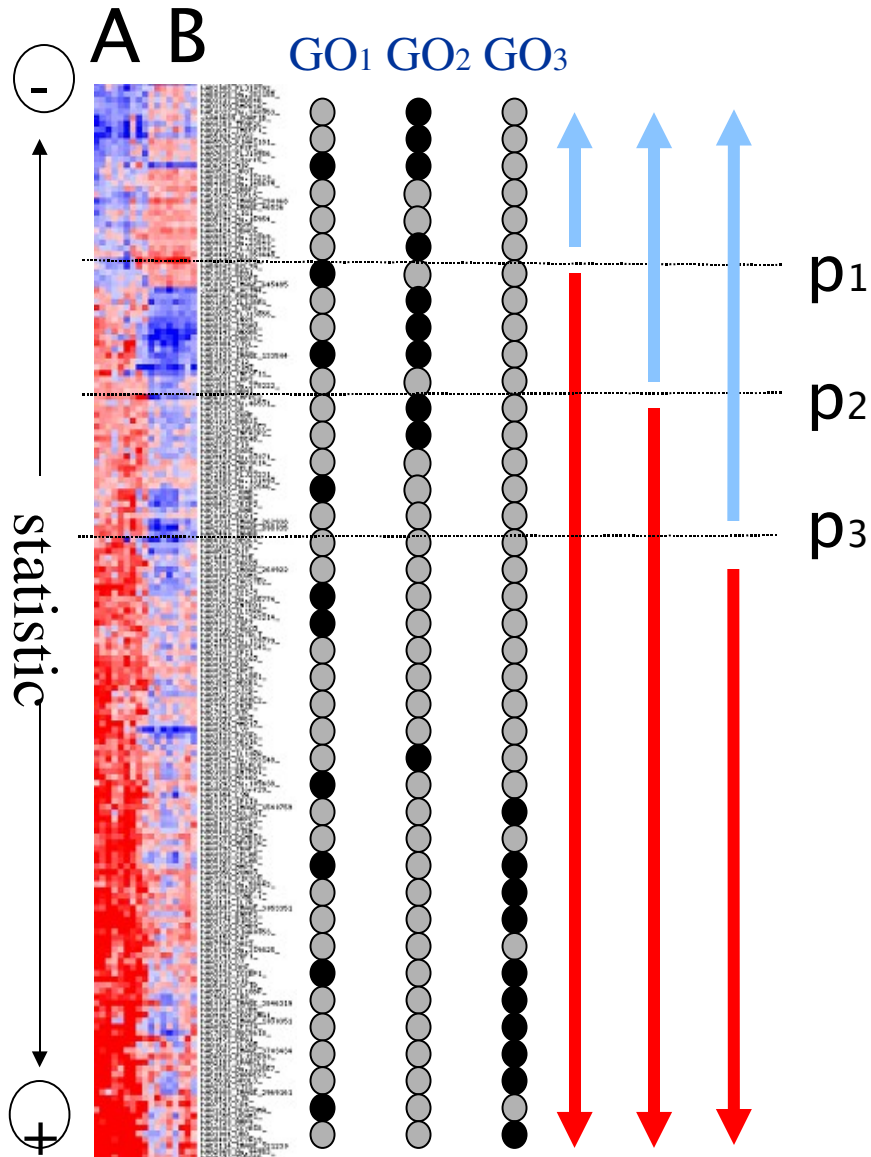
Gene set



background

Independent of the  
experimental  
design

# FatiScan, a segmentation test, provides an easy approach to directly testing functional terms



E.g., term  $GO_2$ ,  
partition  $p_1$

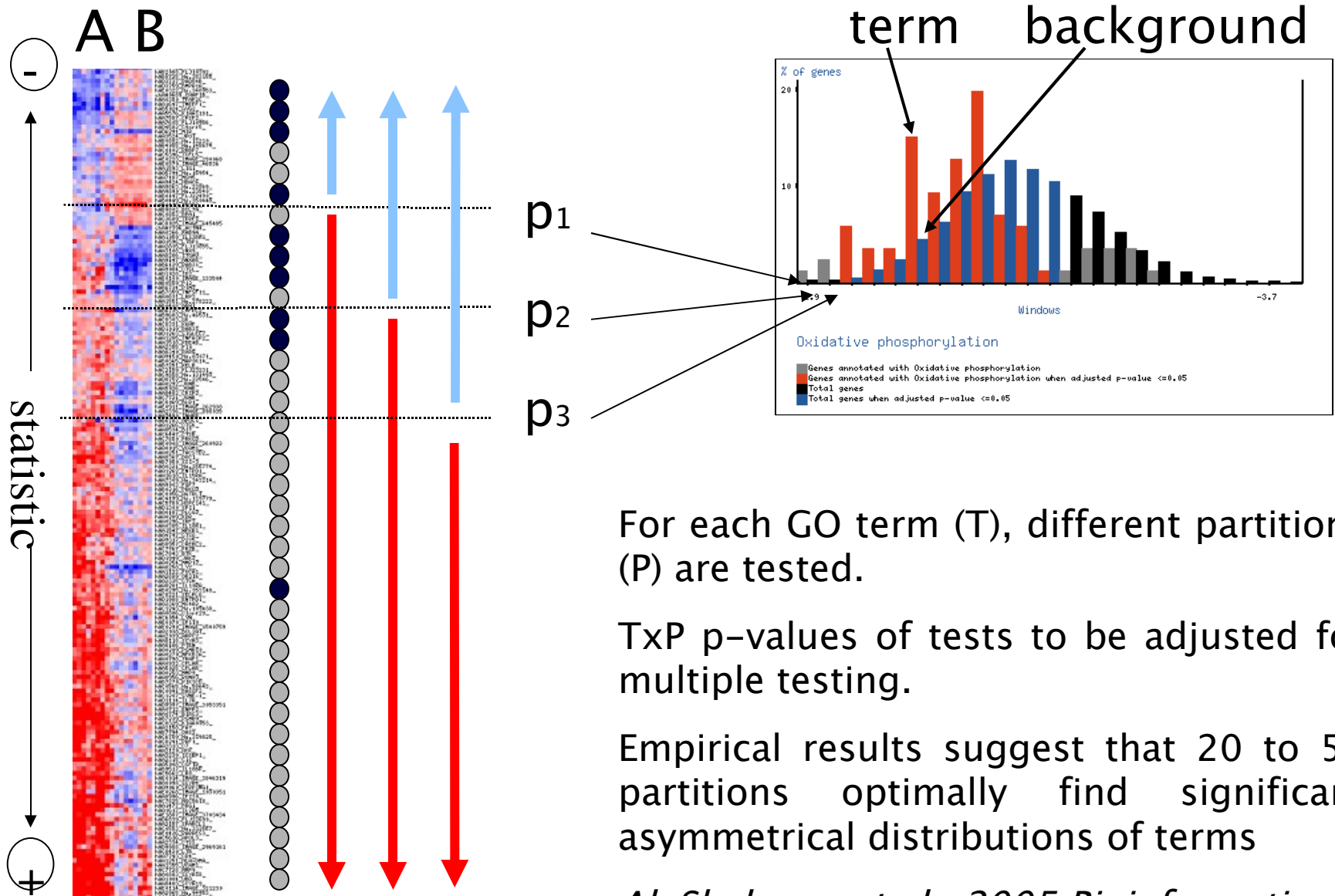
	up	down
GO	4	6
no GO	2	30

GOs can be directly tested by a segmentation test. A series of partitions of the list are performed ( $p_1$ ,  $p_2$ ,  $p_3$ ...) and the GO terms for each functional class in the upper part are compared to the corresponding ones in the lower part by a Fisher test. Asymmetrical distributions of terms towards the extremes of the list will produce significant values of the test.

Finally, p-values are adjusted by FDR

*Al-Shahrour et al., 2005 Bioinformatics*

# Obtaining significant results

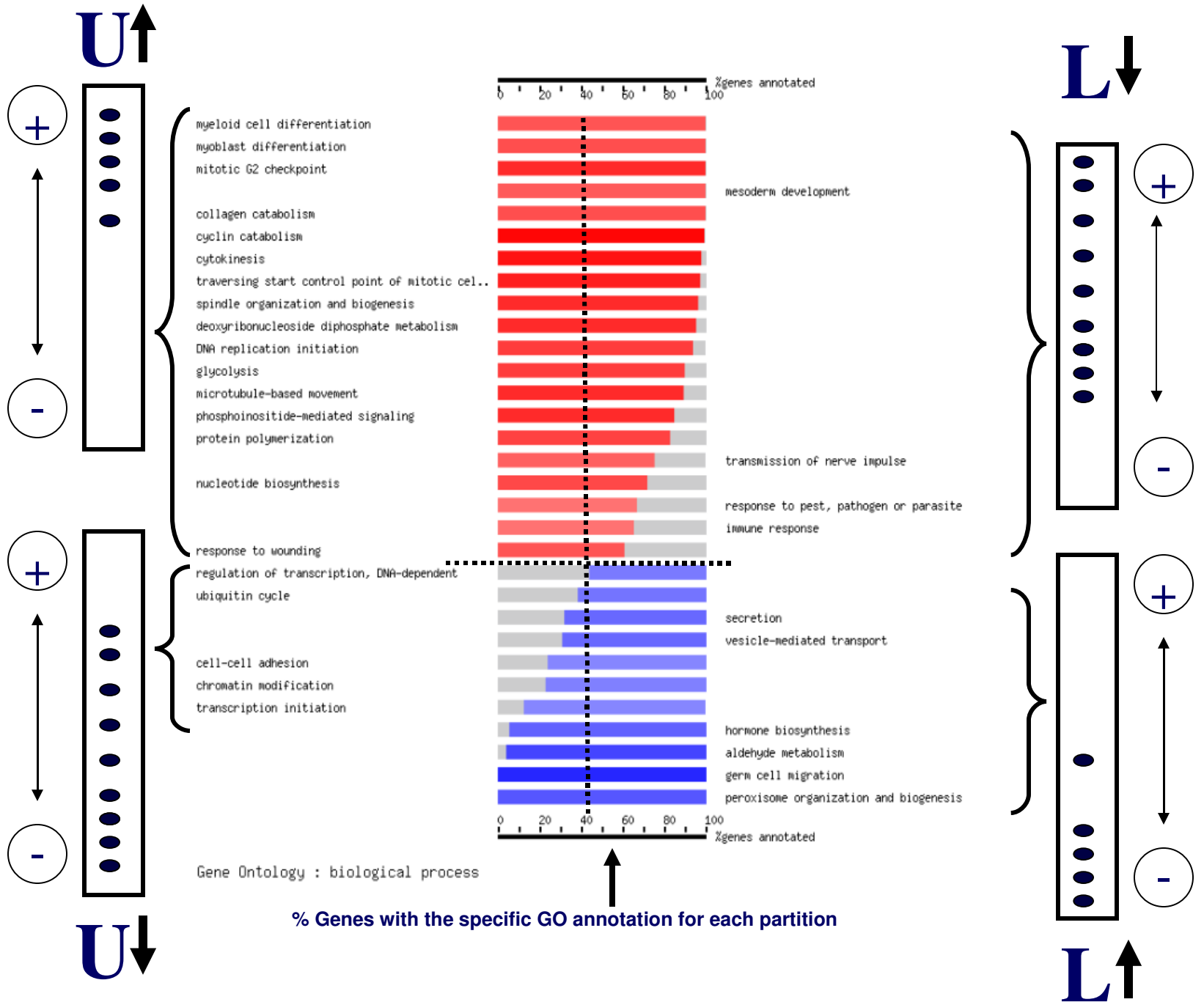


For each GO term (T), different partitions (P) are tested.

TxP p-values of tests to be adjusted for multiple testing.

Empirical results suggest that 20 to 50 partitions optimally find significant asymmetrical distributions of terms

*Al-Shahrour et al., 2005 Bioinformatics*



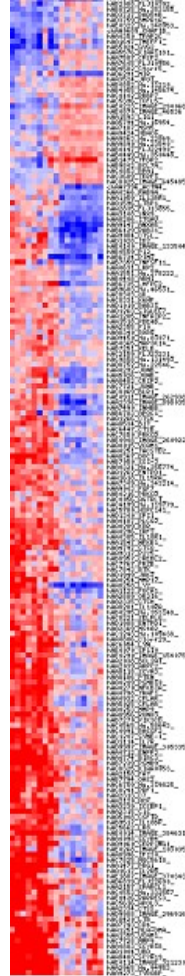


# Case study: functional differences in a class comparison experiment

**A**

8 with impaired tolerance (IGT) + 18 with type 2 diabetes mellitus (DM2)

**A B**



No one single gene shows **significant** differential expression upon the application of a t-test

	Healthy vs diabetic	Functional class	Repository		
			GO	KEGG	Swissprot keyword
Up-regulated		Oxidative phosphorylation	X	X	
		ATP synthesis		X	
		Ribosome		X	
		Ubiquinone			X
		Ribosomal protein			X
		Ribonucleoprotein			X
		Mitochondrion	X		X
		Transit peptide			X
		Nucleotide biosynthesis	X		
		NADH dehydrogenase (ubiquinone) activity	X		
Down-regulated		Nuclease activity	X		
		Insulin signalling pathway		X	

**B**

17 with normal tolerance to glucose (NTG)

Nevertheless, many pathways, and functional blocks are **significantly** activated/deactivated



# Beyond discrete variables: Survival data

Microarrays  
34 samples from  
tumours of  
hypopharyngeal  
cancer (GEO  
GDS1070)



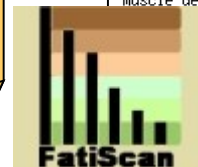
Gene  
selection

**Cox Proportional-Hazards model** to  
study how the  
expression of each  
gene across  
patients is related  
to their survival

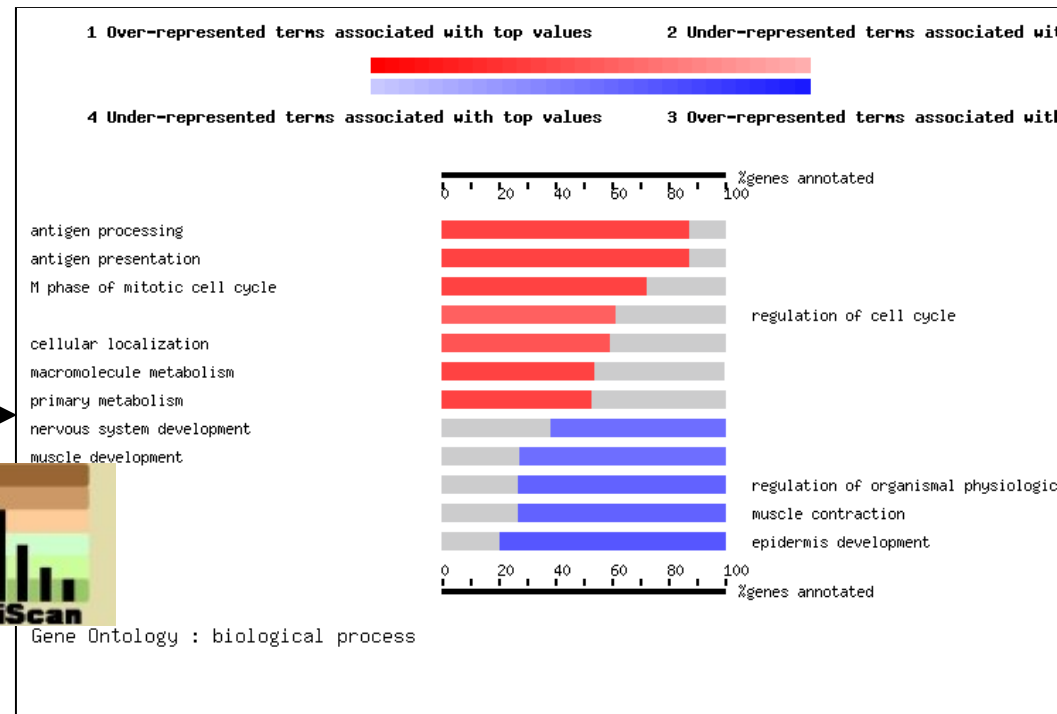
Since FatiScan depends only on a list of  
ordered genes, and not on the original  
experimental values, it can be applied to  
different experimental designs

**- Survival**

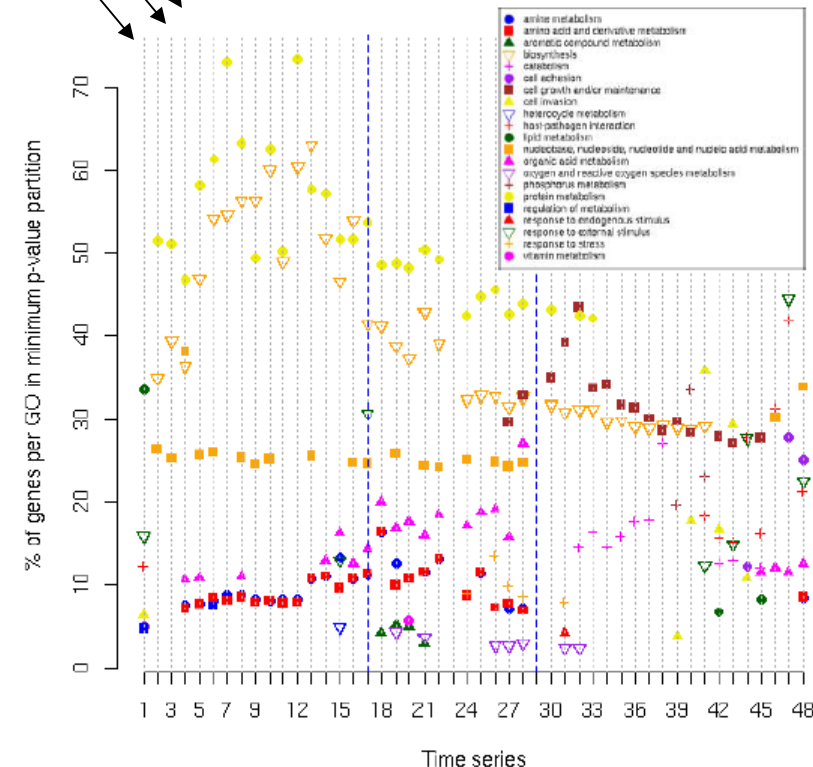
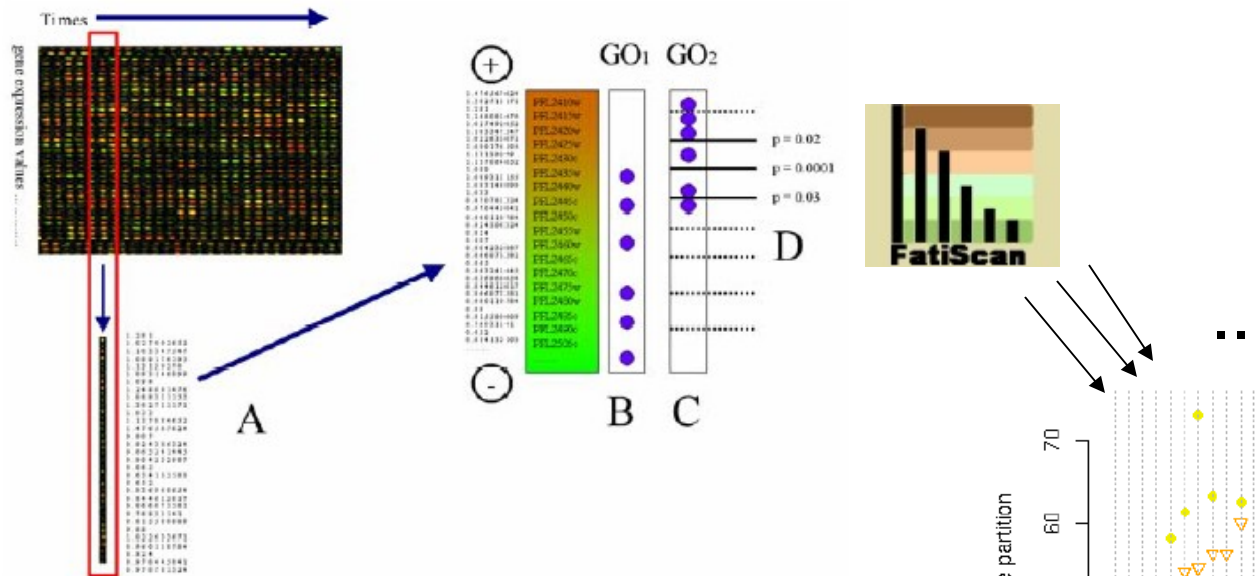
Gen	risk
Gen1	5.8
Gen2	5.6
Gen3	5.4
Gen4	5.2
Gen5	5.2
Gen6	5.0
.....	.....
.....	.....
.....	.....
Gen1000	-6.0
Gen1001	-6.3



**+ Survival**



# Functional analysis of a time series in *P. falciparum*



- Genes at each time point are ranked from highest (red) to lowest (green) relative expression with respect to time 1.

- For each list of ranked genes generated in any time point, the significant over-represented GO terms in the tail corresponding to the highest expression values are recorded.

- The partitions used to decide that a given term is significantly over-represented in the upper tail of the list with respect to the lower part are used for the graphical representation.

# Comparison of gene set methods at a glance

Healthy vs diabetic	Functional class	Repository				Method			
		GO	KEGG	Swissprot keyword	Defined in GSEA	FatiScan	GSEA	PAGE	Tian et al.
Up- regulated	Oxidative phosphorylation	+	+		+	yes	yes	yes	yes
	ATP synthesis		+			yes	-	-	-
	Ribosome		+			yes	-	-	-
	Ubiquinone			+		yes	-	-	-
	Ribosomal protein			+		yes	-	-	-
	Ribonucleoprotein			+		yes	-	-	-
	Mitochondrion	+		+	+	yes	yes	yes	yes
	Transit peptide			+		yes	-	-	-
	Nucleotide biosynthesis	+			+	yes	yes	yes	yes
Dow- regulated	NADH dehydrogenase (ubiquinone) activity	+				yes	-	-	-
	Nuclease activity	+				yes	-	-	-
	Insulin signalling pathway		+			yes	-	-	-

Terms from distinct repositories, reported by different methods in the diabetes dataset (Mootha et al., 2003)

Still one more problem...  
are functional modules defining  
real co-expression classes?

Not a naïve and trivial question.

Functional enrichment methods and gene set analysis methods rely on the assumption that the modules tested do **coexpress**

There are tens of thousands GO terms and hundreds of KEGG pathways



3034

Normalization

	A1	A2	A3	...	A3034
Probeset1	g11	g12	g13	...	
Probeset2	g21	g22	g23	...	
Probeset3	g31	g32	g33	...	
...	...	...	...	...	...
Probeset54675	...	...	...	...	...

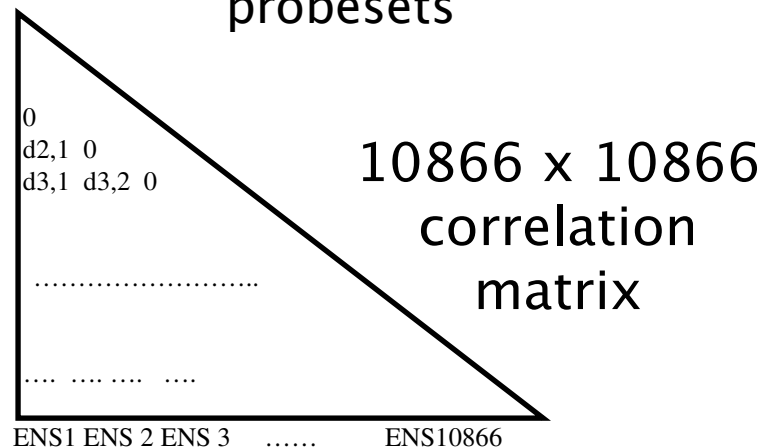
54675  
probesets

	A1	A2	A3	...	A3034
ENS1	g11	g12	g13	...	
ENS2	g21	g22	g23	...	
ENS3	g31	g32	g33	...	
...	...	...	...	...	...
ENS10866	...	...	...	...	...

10866  
transcripts

ENS1  
ENS2  
ENS3

ENS10866



10000  
random  
sampled  
modules

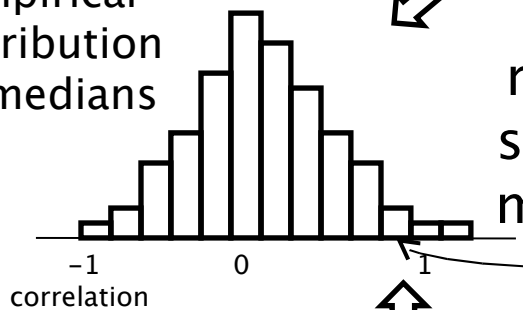
Median A

ENS x  
ENS y  
ENS z

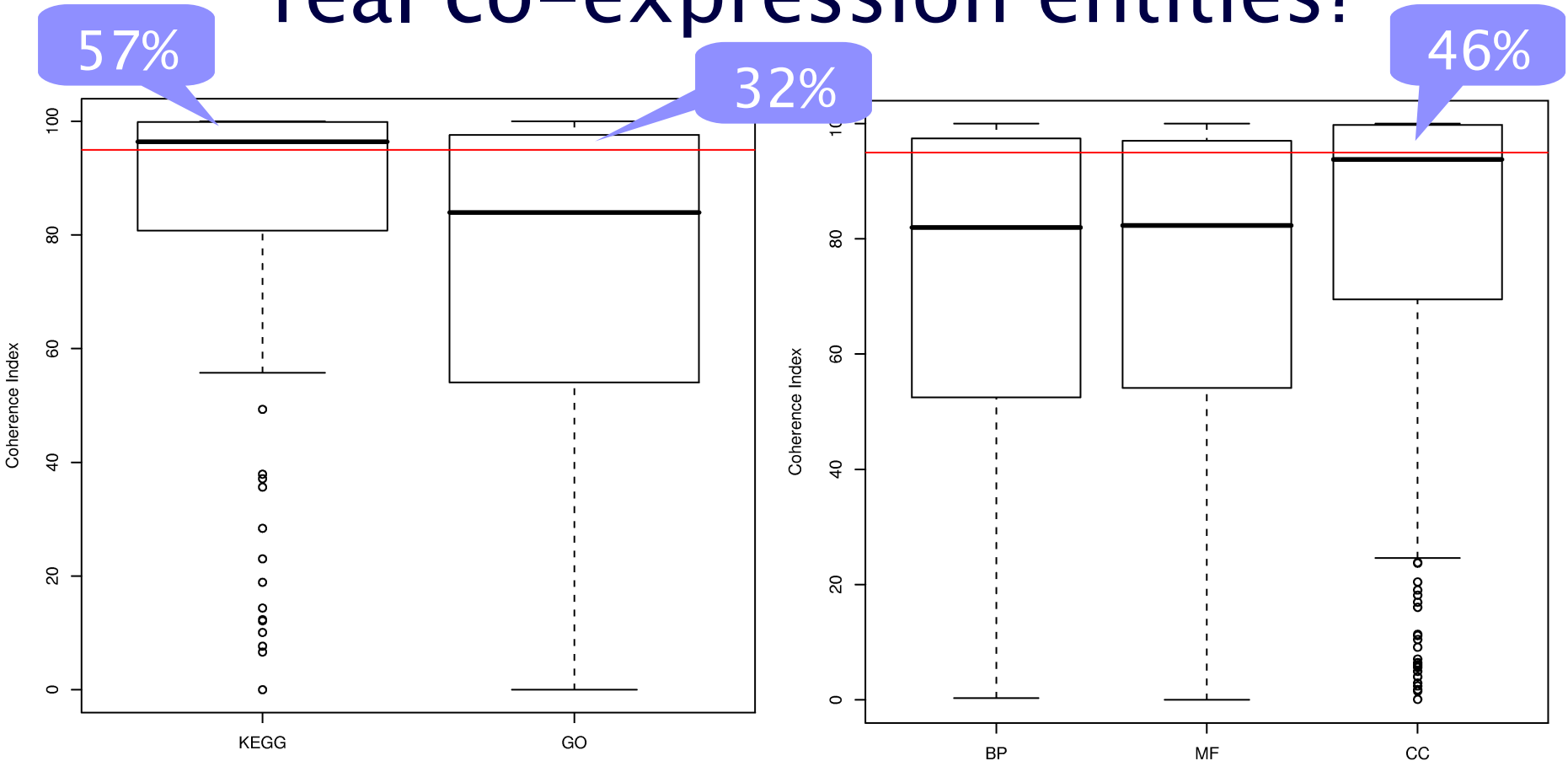
Functional  
module A

p-value

Empirical  
distribution  
of medians



# But are functional modules defining real co-expression entities?



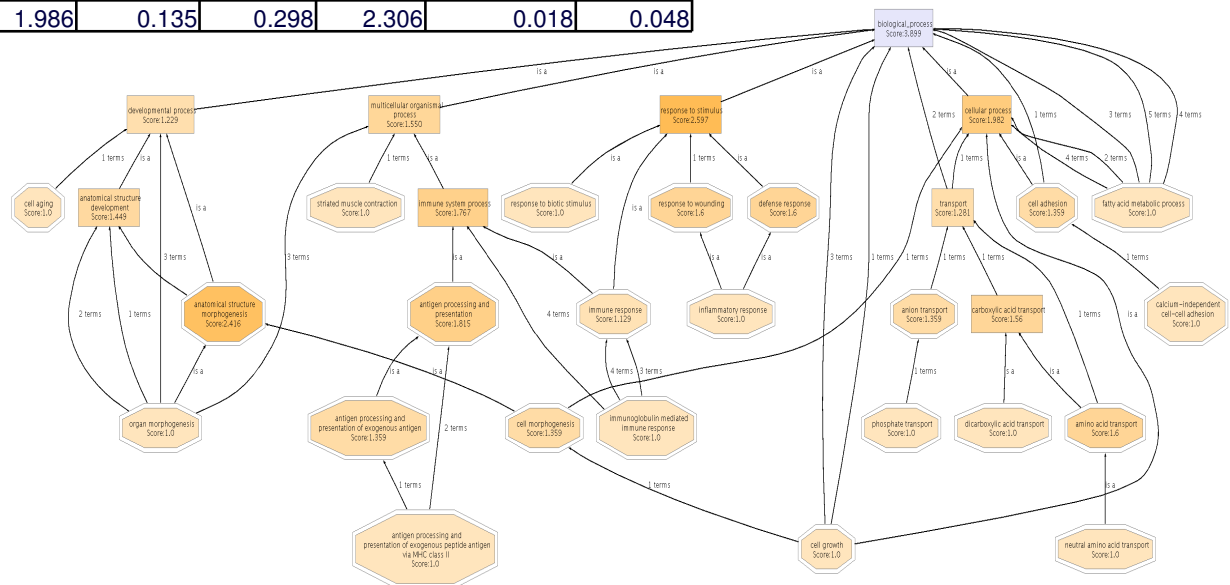
Coherence index:  $(1 - \text{p-value}) * 100$ .

CI > 95% means internal co-expression significantly higher than random co-expression

# Weighting gene module membership by co-expression

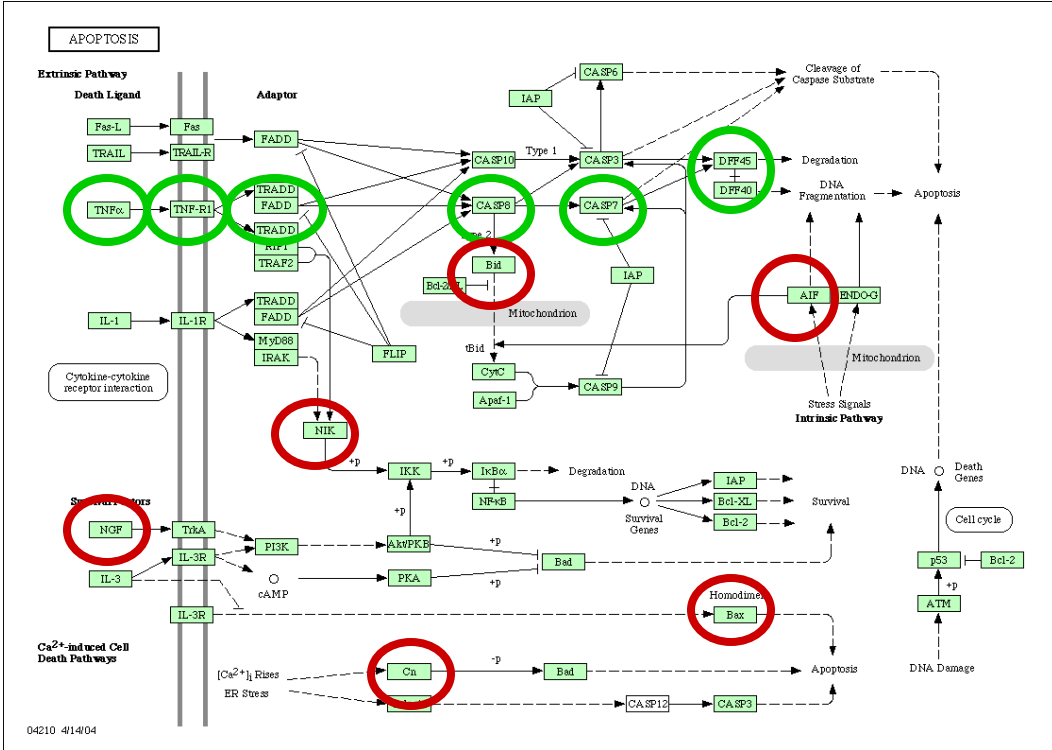
KEGG pathway	Unweighted test			Weighted test		
	statistic	p-value	adjusted p-value	statistic	p-value	Adjusted p-value
Caprolactam degradation	2.741	0.059	0.289	3.124	0.003	0.034
<b>Cell cycle</b>	<b>2.588</b>	<b>0</b>	<b>0</b>	2.711	0	0
Maturity onset diabetes of the young	2.517	0.075	0.289	2.734	0.008	0.034
RNA polymerase	2.497	0.077	0.289	2.657	0.009	0.034
One carbon pool by folate	2.497	0.077	0.289	2.766	0.007	0.034
Urea cycle and metabolism of amino groups	2.497	0.077	0.289	2.674	0.009	0.034
Heparan sulfate biosynthesis	2.478	0.078	0.289	2.818	0.006	0.034
Alanine and aspartate metabolism	2.386	0.087	0.289	2.497	0.012	0.04
Amyotrophic lateral sclerosis (ALS)	2.386	0.087	0.289	2.91	0.005	0.034
beta-Alanine metabolism	2.318	0.094	0.289	2.668	0.009	0.034
Basal transcription factors	2.125	0.116	0.298	2.431	0.014	0.04
Benzoate degradation via CoA ligation	2.072	0.123	0.298	2.468	0.013	0.04
Limonene and pinene degradation	1.986	0.135	0.298	2.306	0.018	0.048

Very simple weight schema:  
 $W=2$  if correlation is positive  
 $W=0.5$  if negative  
 $W=1$  if not in the class





# Future directions



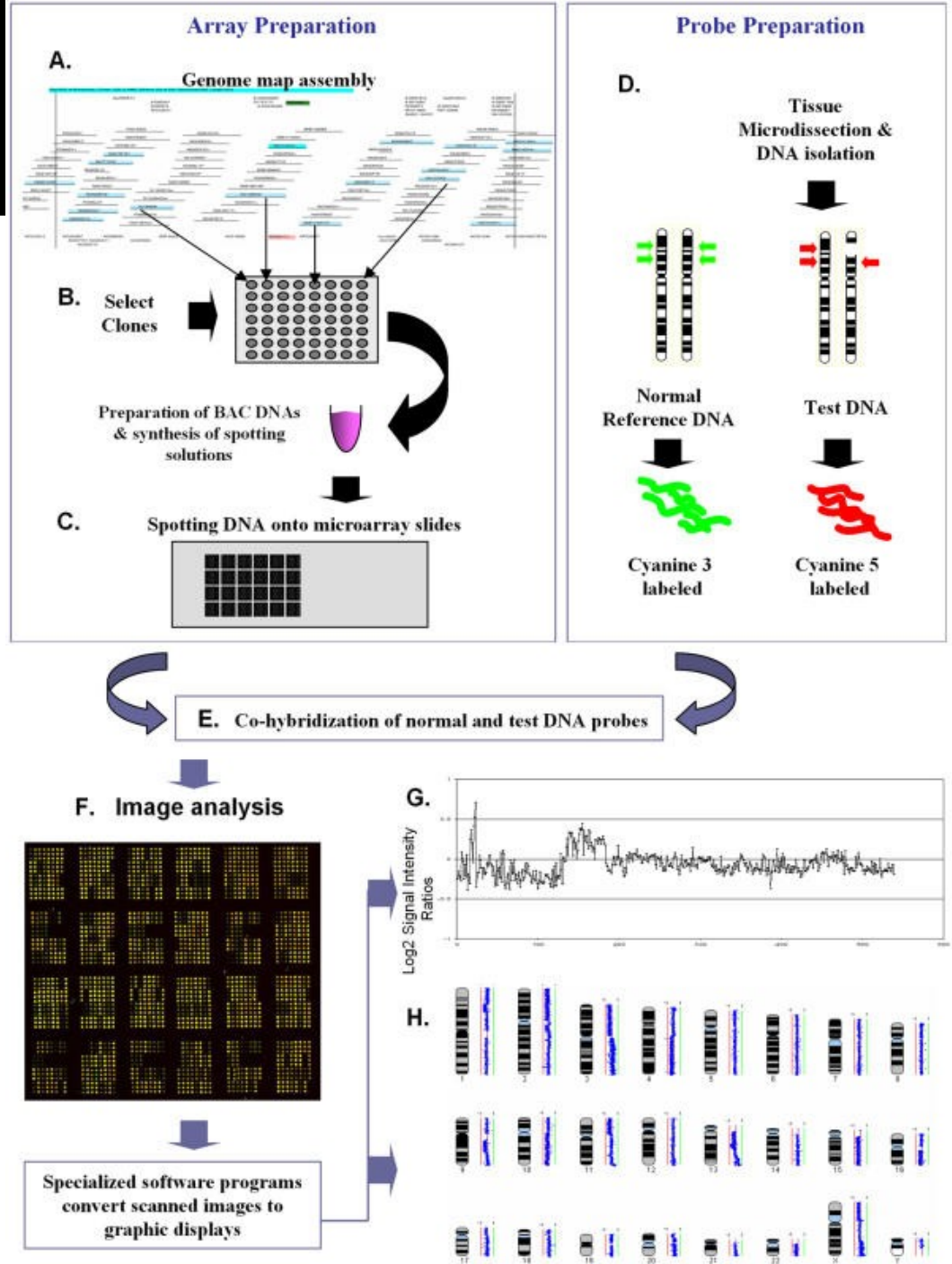
Testing hierarchies is  
better  
Functions and  
pathways are  
correlated.  
Testing models will  
increase our  
sensitivity

# Pathways are not categorical variables

In general (systems) biology is behind. Our questions must be inspired directly by biology

# Array CGH

A new way of  
studying copy  
number  
alterations with  
virtually a few  
bases resolution...



Jump over

# ISA CGH (*In silico* array-CGH)

Estimating copy number, correlation copy number – expression and the minimum common amplified / lost region

InSilicoArray  
CGH

SPECIE

Select the species

Homo Sapiens

GENOMIC DATAFILE

Upload the genomic data

Browse...

Hybrid Example

See Example

Upload the class labels (optional)

Browse...

Upload the gender labels (default: males)

Browse...

Define the copy number method

Smoothing

Define the excluded autosomal and female sex chromosome bands

Autosomal Median

Define the excluded male sex chromosome bands

Half Autosomal Median

EXPRESSSION DATAFILE

Upload your expression data

Browse...

Hybrid Example

See Example

POSITIONS DATAFILE

Upload the data with the positions of your entries (optional)

Browse...

Hybrid Example

See Example

PARAMETERS

Set the scale value... (default: max. value)

Reference lines... (default: none)

DETAILED

One array / All chromosomes

Number of the selected array...

Limit to chromosomes (optional)

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

X

Y

All arrays / One chromosome

Chr.

Choose output type

Text

Run

InSilicoArray

PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

Download & Webmaster

InSilicoArray  
CGH

InSilicoArray  
CGH

InSilicoArray  
CGH

InSilicoArray  
CGH

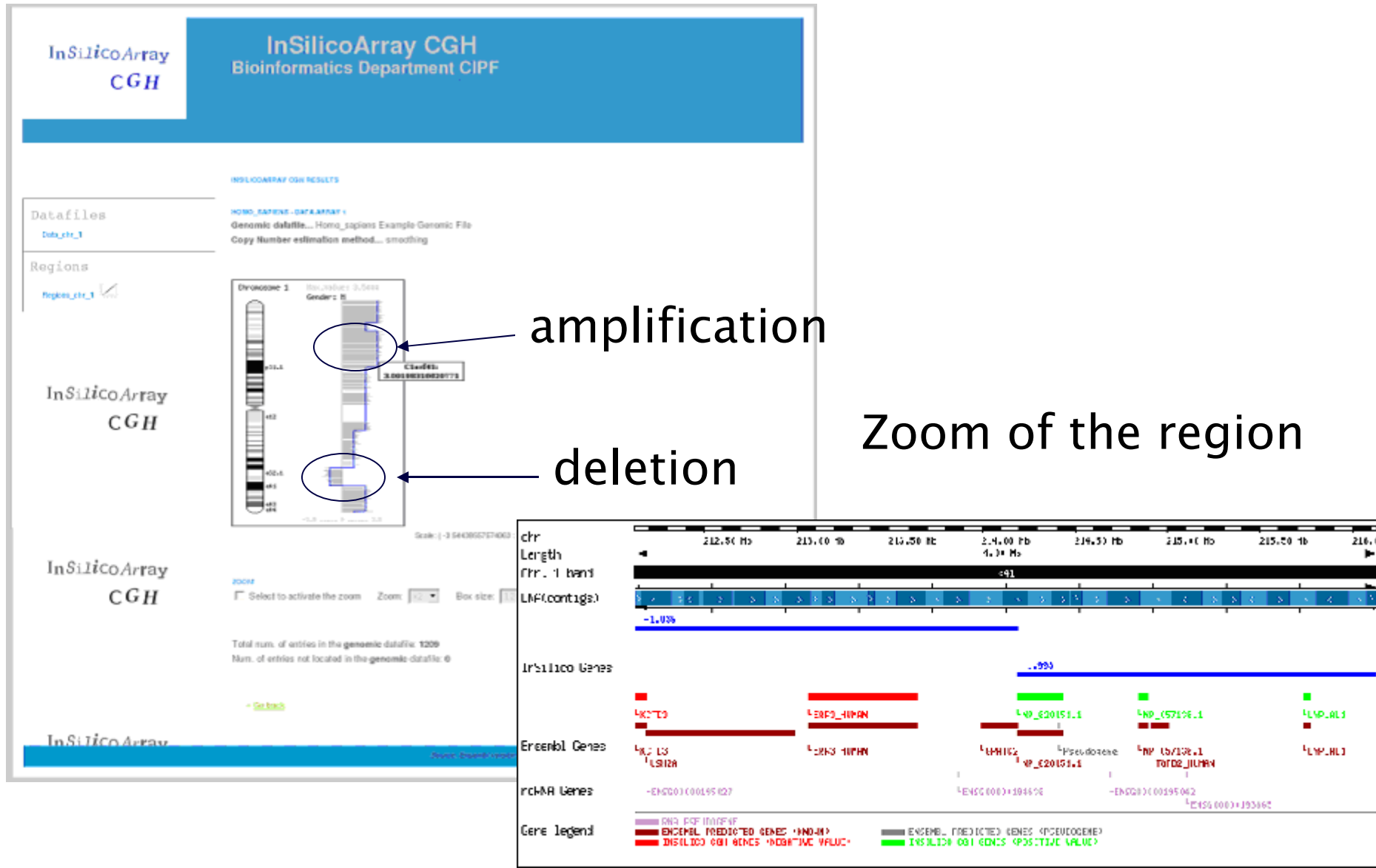
InSilicoArray  
CGH

InSilicoArray

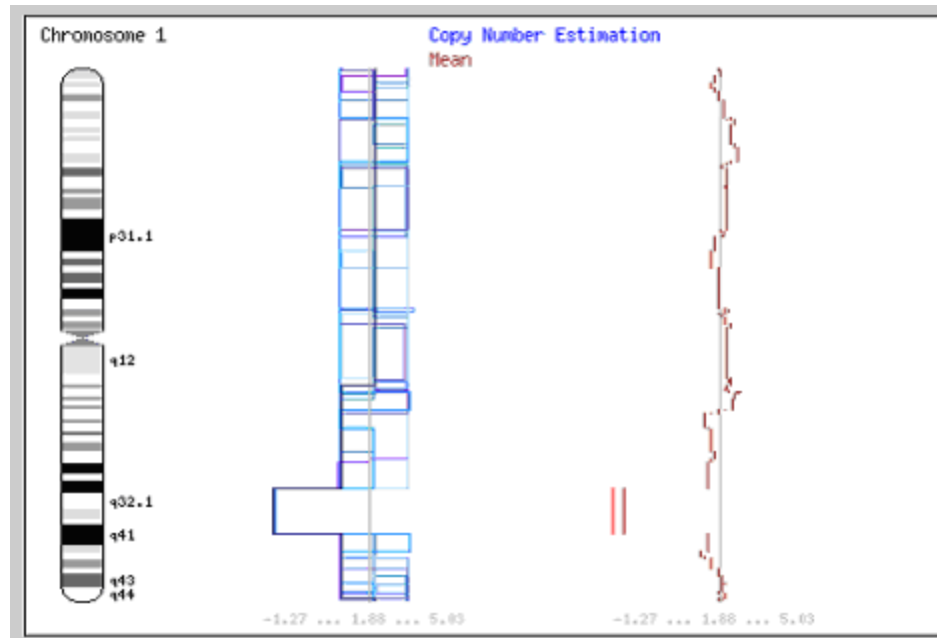
PRINCIPE FELIPE  
CENTRO DE INVESTIGACION

Download & Webmaster

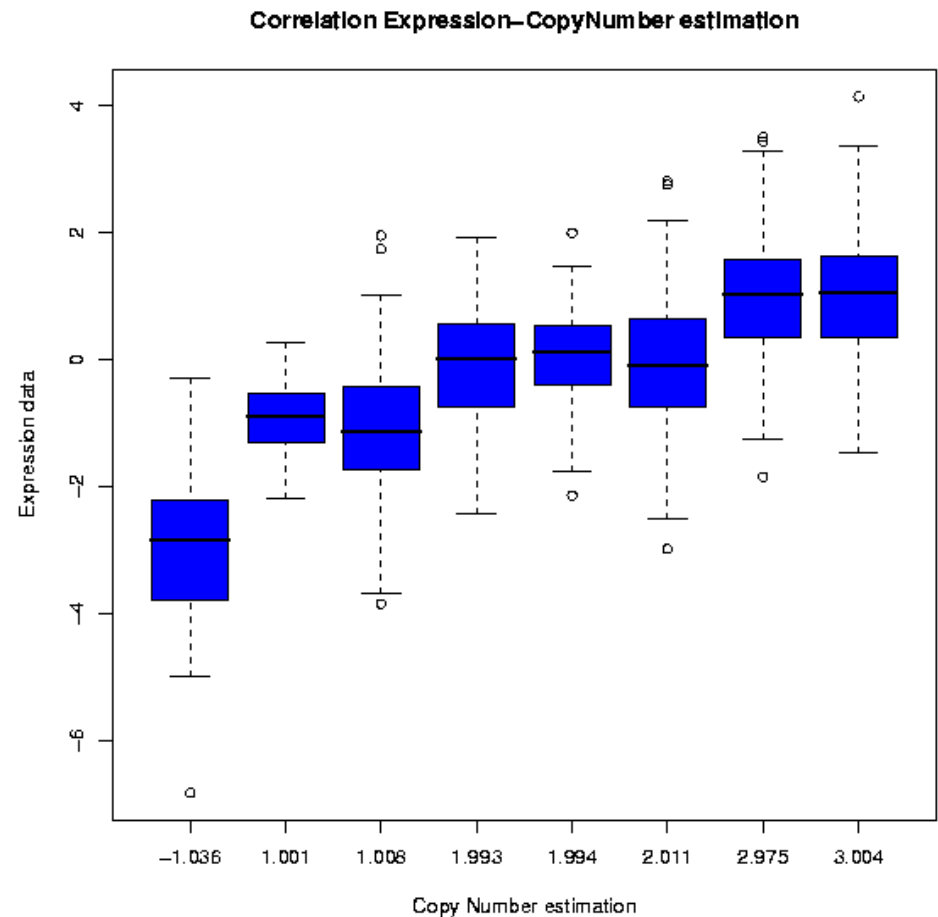
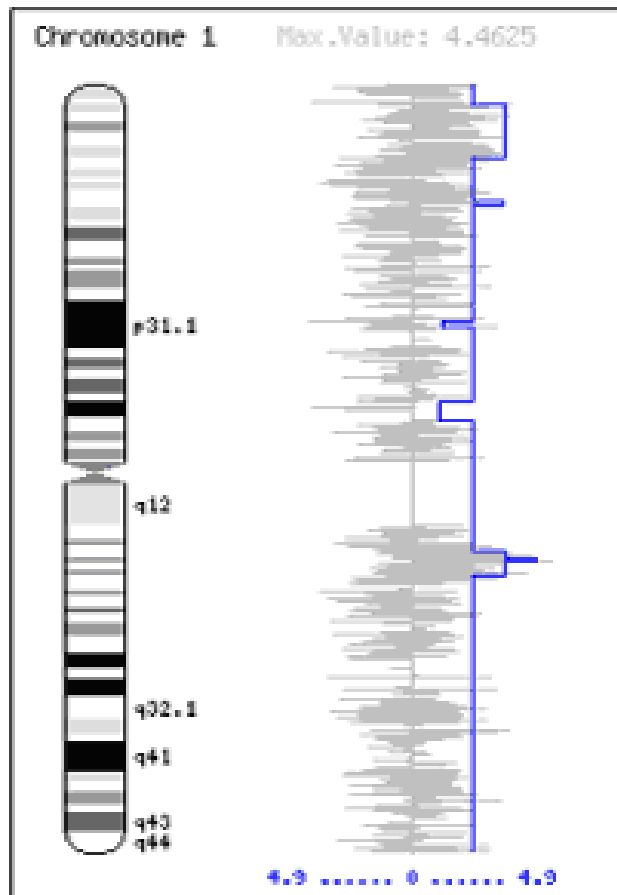
# Estimating copy number



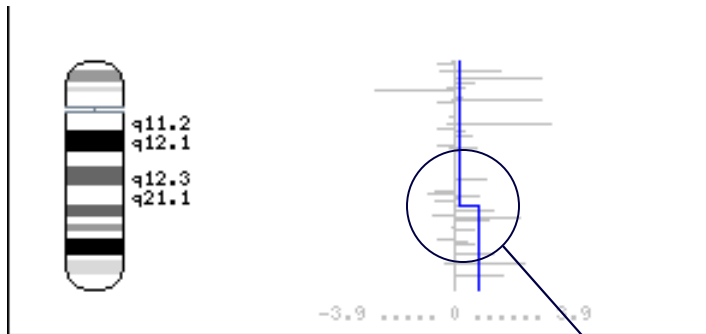
# Minimum region with consistent losses or gains



# Correlation copy number to expression value



# Array-CGH. DAS server

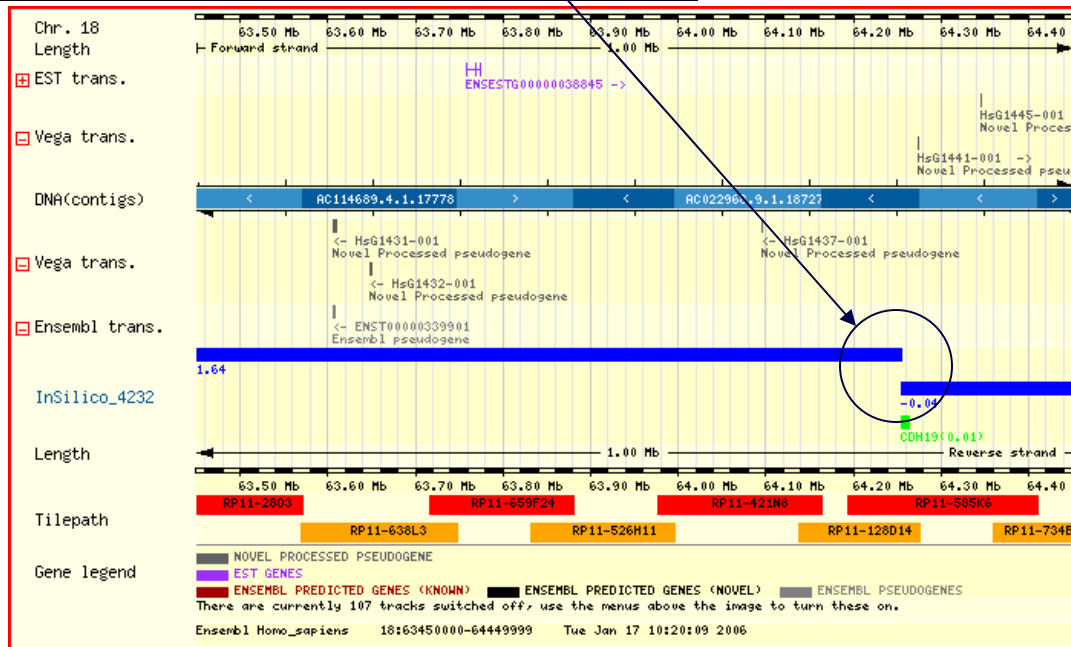


Detection of copy number alterations (two new methods)

Relationship expression / copy number alteration

Functional annotation of altered regions

DAS server





# What to use

## Web tools gain popularity

- Interactive
- Heavy calculations at server side
- Large databases at server side
- Always the last version

Microarray data analysis webtools with at least 10 citations<sup>1</sup>.

Web tool	URL	Citations <sup>1</sup>
GEPAS	<a href="http://www.gepas.org">http://www.gepas.org</a>	252
ExpressionProfiler	<a href="http://www.ebi.ac.uk/expressionprofiler">http://www.ebi.ac.uk/expressionprofiler</a>	46
caGEDA	<a href="http://bioinformatics.upmc.edu/GEDA.html">http://bioinformatics.upmc.edu/GEDA.html</a>	30
GenePublisher	<a href="http://www.cbs.dtu.dk/services/GenePublisher">http://www.cbs.dtu.dk/services/GenePublisher</a>	25
ExpressYourself	<a href="http://bioinfo.mbb.yale.edu/expressyourself">http://bioinfo.mbb.yale.edu/expressyourself</a>	24
RACE	<a href="http://race.unil.ch/">http://race.unil.ch/</a>	22
ArrayPipe	<a href="http://www.pathogenomics.ca/arraypipe">http://www.pathogenomics.ca/arraypipe</a>	19
VAMPIRE	<a href="http://genome.ucsd.edu/microarray/">http://genome.ucsd.edu/microarray/</a>	17
MIDAW	<a href="http://muscle.cribi.unipd.it/midaw/">http://muscle.cribi.unipd.it/midaw/</a>	13
t-profiler	<a href="http://www.t-profiler.org">http://www.t-profiler.org</a>	12
CARMAweb	<a href="https://carmaweb.genome.tugraz.at">https://carmaweb.genome.tugraz.at</a>	10

1) Scholar Google citations over all the references of the tool.

# GEPAS

www.gepas.org - Mozilla Firefox

Archivo Editar Ver Historial Marcadores Herramientas Ayuda

http://www.gepas.org/

Comenzar a usar Fire... Últimas noticias Google Académico

www.gepas.org

**GEPAS**  
GENE EXPRESSION PATTERN ANALYSIS SUITE

You are using the **new GEPAS 4.0**, previous [version v3.1 here](#)  
For GEPAS v4.0, your browser has to accept cookies

log in | register  
You are an anonymous user

Bioinformatics Department - CIPF

Tools Documentation Datasets News About us

**Lastest News**

**October 2007**  
New release GEPAS v4.0, including new user-interface, session, project and jobs management, new normalization, preprocessing, clustering, differential expression, predictors, viewers, InSilicoArray CGH, Babelomics Suite

**July 2007**  
A new ID convertor added to the GEPAS package

**June 2007**  
New beta version: GEPAS v4.0, including new user-interface, session, project and jobs management,...

**September 2006**  
A new tool was added to the GEPAS package: Prophet, a tool for building a class predictor

**March 2006**  
Computing journal award to the best R&D project

**February 2006**  
New releases: GEPAS v3.0


**New Release v4.0**

Gene Expression Pattern Analysis Suite (GEPAS) is one of the most complete integrated packages of tools for microarray data analysis available over the web. GEPAS maintained effort to offer a platform for gene expression data analysis to the scientific community, which has uninterruptedly been running since 2001. During its evolution to keep pace with the new interests and trends in the ever changing world of microarray data analysis.

GEPAS is designed to provide an intuitive although powerful web-based interface that offers diverse analysis options from the early step of preprocessing (Affymetrix and two-colour microarray experiments and other preprocessing options), to the final step of the functional profiling of the experiment (using Gene Ontology, PubMed abstracts etc.), which include different possibilities for clustering, gene selection, class prediction and array-comparative genomic hybridization

GEPAS is extensively used by researchers of many countries. See [usage map](#).

Did you know...?

 GEPAS is also the name of a herb: *Sarcandra glabra*.

*Sarcandra glabra* is a warm herb having strong detoxifying properties, helping to clear heat and toxic material. It can especially enhance cellular energy production. It also has nonspecific anti-inflammatory activity, and has been used to promote circulation.

GEPAS, 2002-2007, Bioinformatics Department, CIPF, Avda. Autopista del Saler 16, 46013 Valencia, Spain, + 34 96 328 96 80

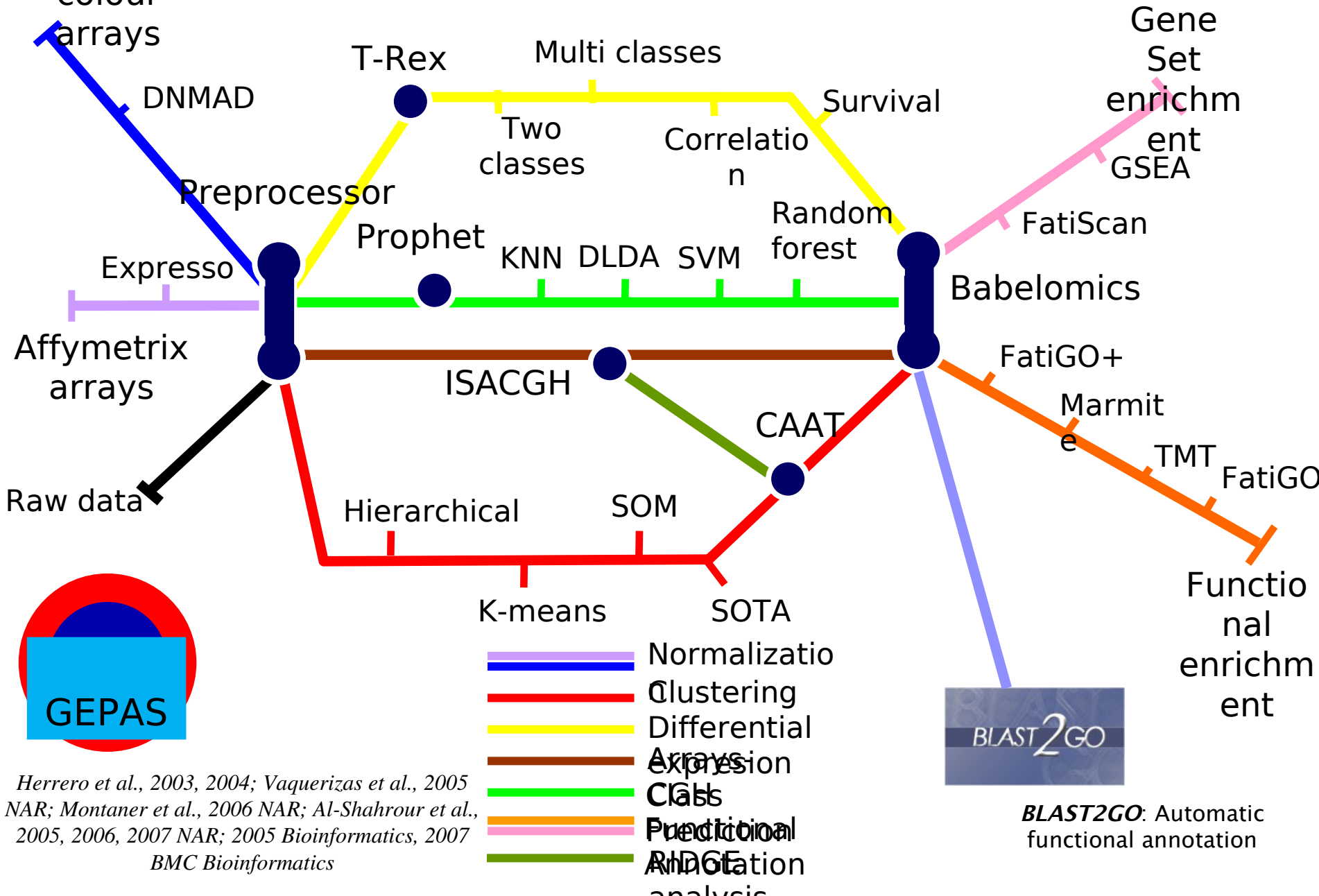
Encontrar:  Siguiente Anterior Resaltar todo

javascript:makeGetRequest('/cgi-bin/tools.cgi',null,'center');

Inicio Explorado... Microsoft O... EndNote 9 - G... 19 Microsoft ... Bioinformatics ... www.gepas.or... Microsoft Powe... ES 23:15

Since october 2007, GEPAS 4.0

# GEPAS



Herrero et al., 2003, 2004; Vaquerizas et al., 2005  
 NAR; Montaner et al., 2006 NAR; Al-Shahrour et al.,  
 2005, 2006, 2007 NAR; 2005 Bioinformatics, 2007  
 BMC Bioinformatics

# Some numbers

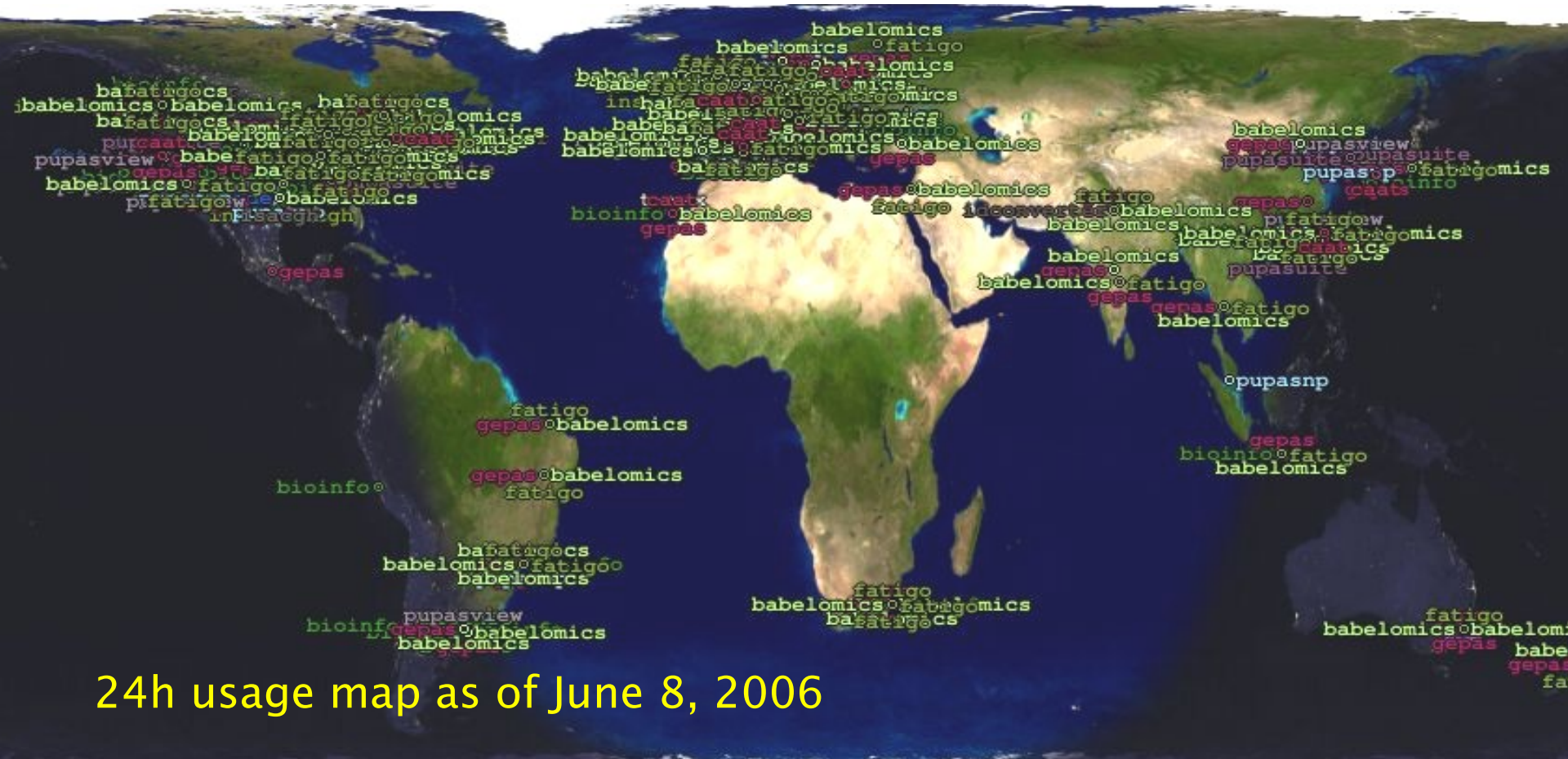
297 papers cite GEPAS during the last three years

260 papers cite Babelomics/FatiGO

*(source ISI Web of Knowledge,  
December2007)*

More than 150,000 experiments  
analysed during the last year.

More than 500 experiments per day.



## 24h usage map as of June 8, 2006

# Web tools for functional profiling

Web tools with 10 or more Scholar Google citations

Tool	URL	Analysis type	References	Citations
GSEA	<a href="http://www.broad.mit.edu/gsea/">http://www.broad.mit.edu/gsea/</a>	GSA	(3,33)	1013
DAVID	<a href="http://www.DAVID.niaid.nih.gov">http://www.DAVID.niaid.nih.gov</a>	FE	(34)	504
GOMiner	<a href="http://discover.nci.nih.gov/gominer/">http://discover.nci.nih.gov/gominer/</a>	FE	(35,36)	408
<b><i>Babelomics</i></b>	<b><i><a href="http://www.babelomics.org">http://www.babelomics.org</a></i></b>	FE, GSA	<b><i>(11-13,29)</i></b>	<b><i>402</i></b>
MAPPFinder	<a href="http://www.GenMAPP.org">http://www.GenMAPP.org</a>	FE	(37)	379
GOSats	<a href="http://gostat.wehi.edu.au/">http://gostat.wehi.edu.au/</a>	FE	(27)	249
Ontotools	<a href="http://vortex.cs.wayne.edu/ontoexpress/">http://vortex.cs.wayne.edu/ontoexpress/</a>	FE	(38,40-43)	223
GOTM	<a href="http://genereg.ornl.gov/gotm/">http://genereg.ornl.gov/gotm/</a>	FE	(44)	164
FunSpec	<a href="http://funspec.med.utoronto.ca">http://funspec.med.utoronto.ca</a> webcite	FE	(45)	100
GeneMerge	<a href="http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html">http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html</a>	FE	(46)	96
FuncAssociate	<a href="http://llama.med.harvard.edu/Software.html">http://llama.med.harvard.edu/Software.html</a>	FE, GSA	(39)	91
GOToolBox	<a href="http://gin.univ-mrs.fr/GOToolBox">http://gin.univ-mrs.fr/GOToolBox</a>	FE	(28)	74
GFINDER	<a href="http://www.medinfopoli.polimi.it/GFINDER/">http://www.medinfopoli.polimi.it/GFINDER/</a>	FE	(47,48)	49
WebGestalt	<a href="http://bioinfo.vanderbilt.edu/webgestalt/">http://bioinfo.vanderbilt.edu/webgestalt/</a>	FE	(49)	46
GOAL	<a href="http://microarrays.unife.it">http://microarrays.unife.it</a>	GSA	(50)	25
Pathway Explorer	<a href="https://pathwayexplorer.genome.tugraz.at/">https://pathwayexplorer.genome.tugraz.at/</a>	FE	(51)	25
PLAGE	<a href="http://dulci.biostat.duke.edu/pathways/">http://dulci.biostat.duke.edu/pathways/</a>	GSA	(52)	18
t-profiler	<a href="http://www.t-profiler.org/">http://www.t-profiler.org/</a>	GSA	(53)	12
WebBayGO	<a href="http://blasto.iq.usp.br/~tkoide/BayGO/">http://blasto.iq.usp.br/~tkoide/BayGO/</a>	FE	(54)	10

# Other tools (non-commercial)

To cover more specific analysis requirements

Bioconductor: <http://www.bioconductor.org>

BRB tools: <http://linus.nci.nih.gov/BRB-ArrayTools.html>

TM4 (MeV): <http://www.tm4.org/mev.html>



# The bioinformatics department at the Centro de Investigación Príncipe Felipe (Valencia, Spain)...

Joaquín Dopazo  
Eva Alloza  
Leonardo Arbiza  
Fátima Al-Shahrour  
Jordi Burguet  
Emidio Capriotti  
Lucía Conde  
Ana Conesa  
Hernán Dopazo  
Toni Gabaldon  
Francisco García  
Stefan Goetz  
Jaime Huerta  
Marc Martí  
Ignacio Medina  
Pablo Minguez  
David Montaner  
Joaquín Tárraga  
Peio Ziarsolo



INSTITUTO NACIONAL  
DE BIOINFORMÁTICA



...the INB, National Institute of Bioinformatics  
(Functional Genomics Node) and the CIBER-ER  
Network of Centers for Rare Diseases

