

Microarray data analysis using GEPAS and Babelomics

**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>



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...

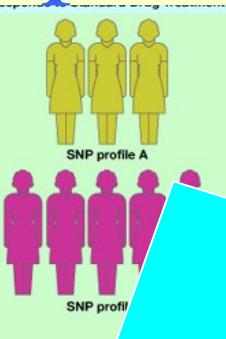
The pre-genomics paradigm

Genes in the DNA...



...code for proteins...

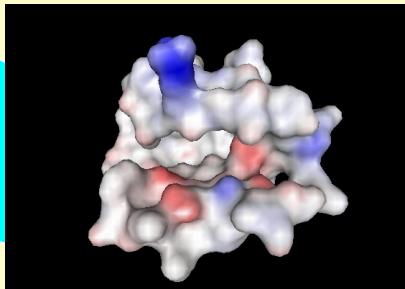
>protein kinase
acctgttgatggcacaggactgtatgtatc
tatgtctgtatcatgtatgtactgtatgtggg
ggctattgtacttgatgtatcatac....



...produces the final phenotype

From genotype to 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

Genes in the DNA...

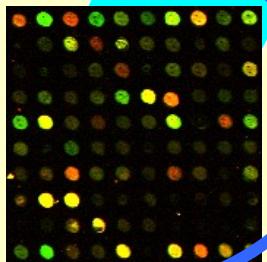
>protein kinase

```
acctgttgtatggccgacaggactgtatgtatgc  
tatgttgtatgtatgtatgtactgtatgtatgg  
ggctttatgtactgtatgtatcata...
```



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

A typical tissue is expressing among 5000 and 10000 genes



...code for proteins...

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

100K-500K proteins

...whose structures account for function...

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

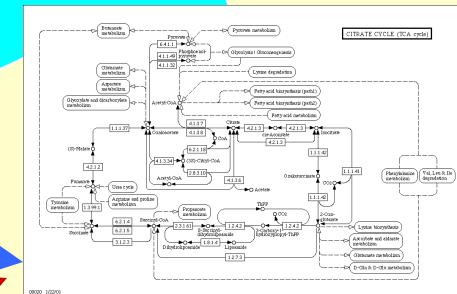
10 million SNPs

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

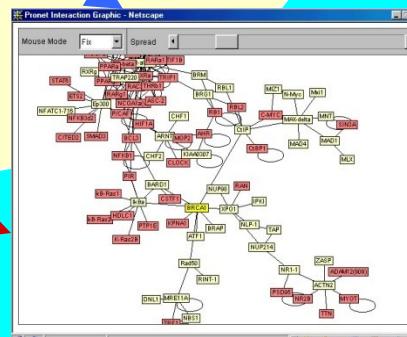
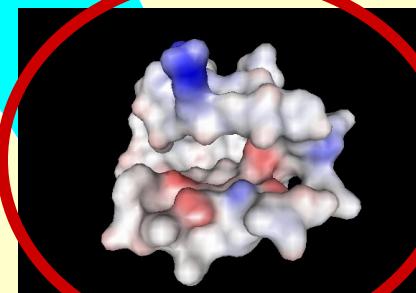


From genotype to phenotype

(in the functional post-genomics scenario)



...conforming complex interaction networks...

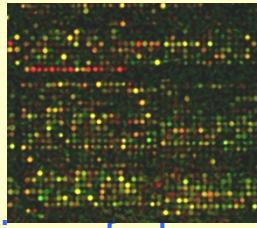


...in cooperation with other proteins...

Each protein has an average of 8 interactions

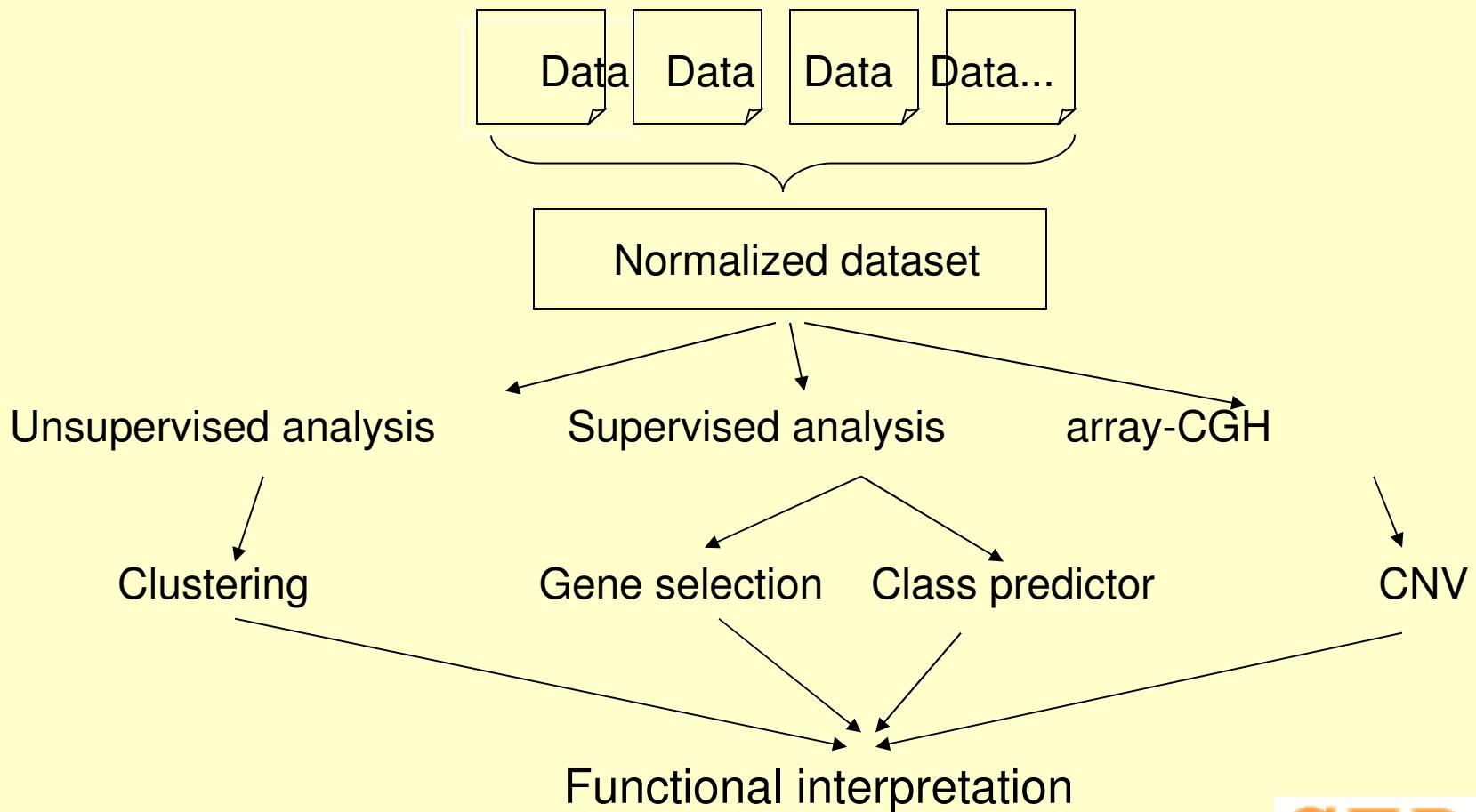
Gene expression profiling. Historic perspective

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**)

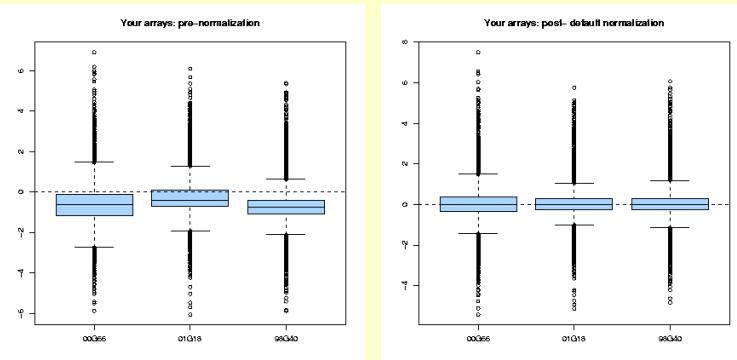
General pipeline for the (most common) analyses



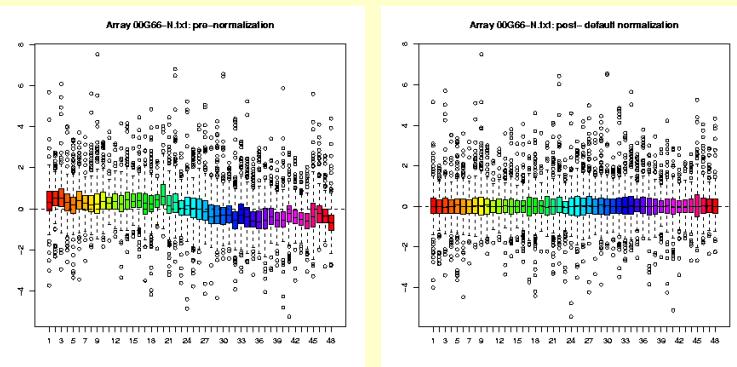
Of course, there are more possible analyses (e.g. reverse engineering of networks, ChIP-on-Chip, etc.)

Normalisation

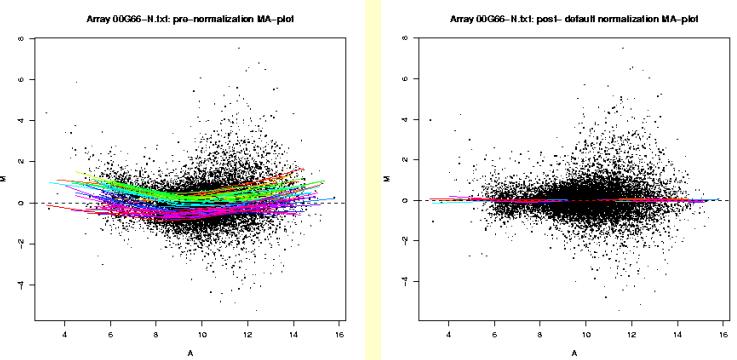
A



B



C

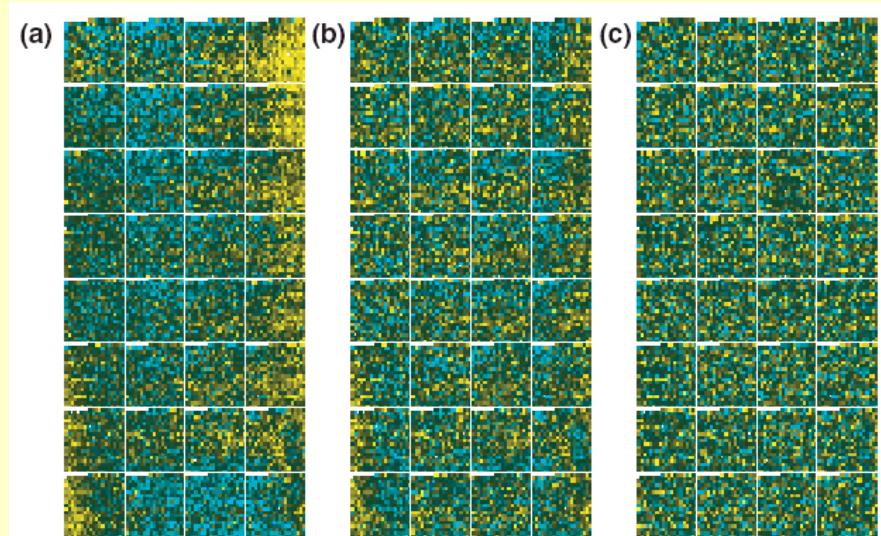


↑ Before (left) and after (right) normalisation. A) BoxPlots, B)
BoxPlots of subarrays and C) MA plots (ratio versus intensity)

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

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



On data, probes and hybridisations

Gene expression

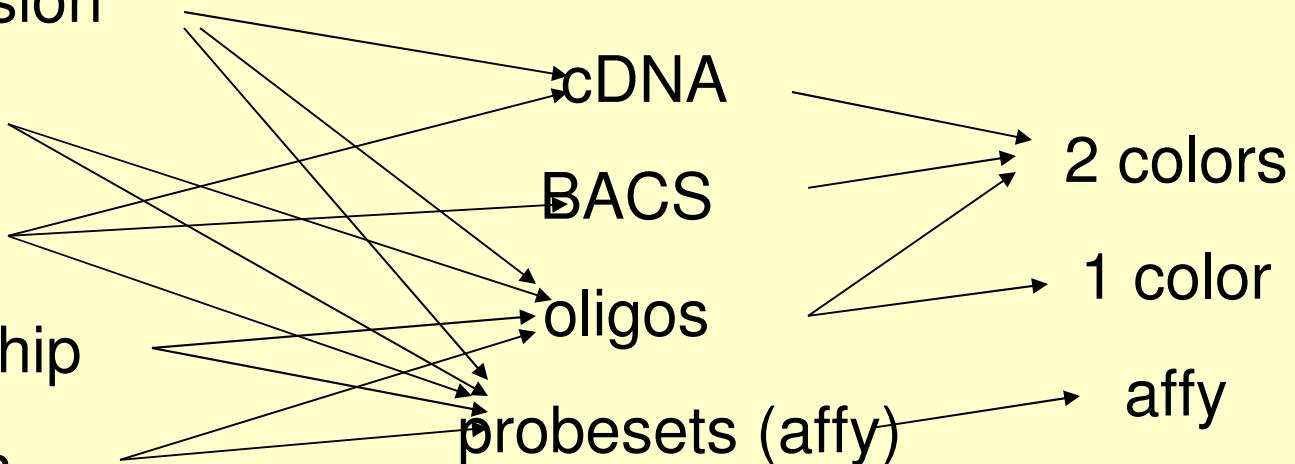
exon

CNV

ChIP-on-Chip

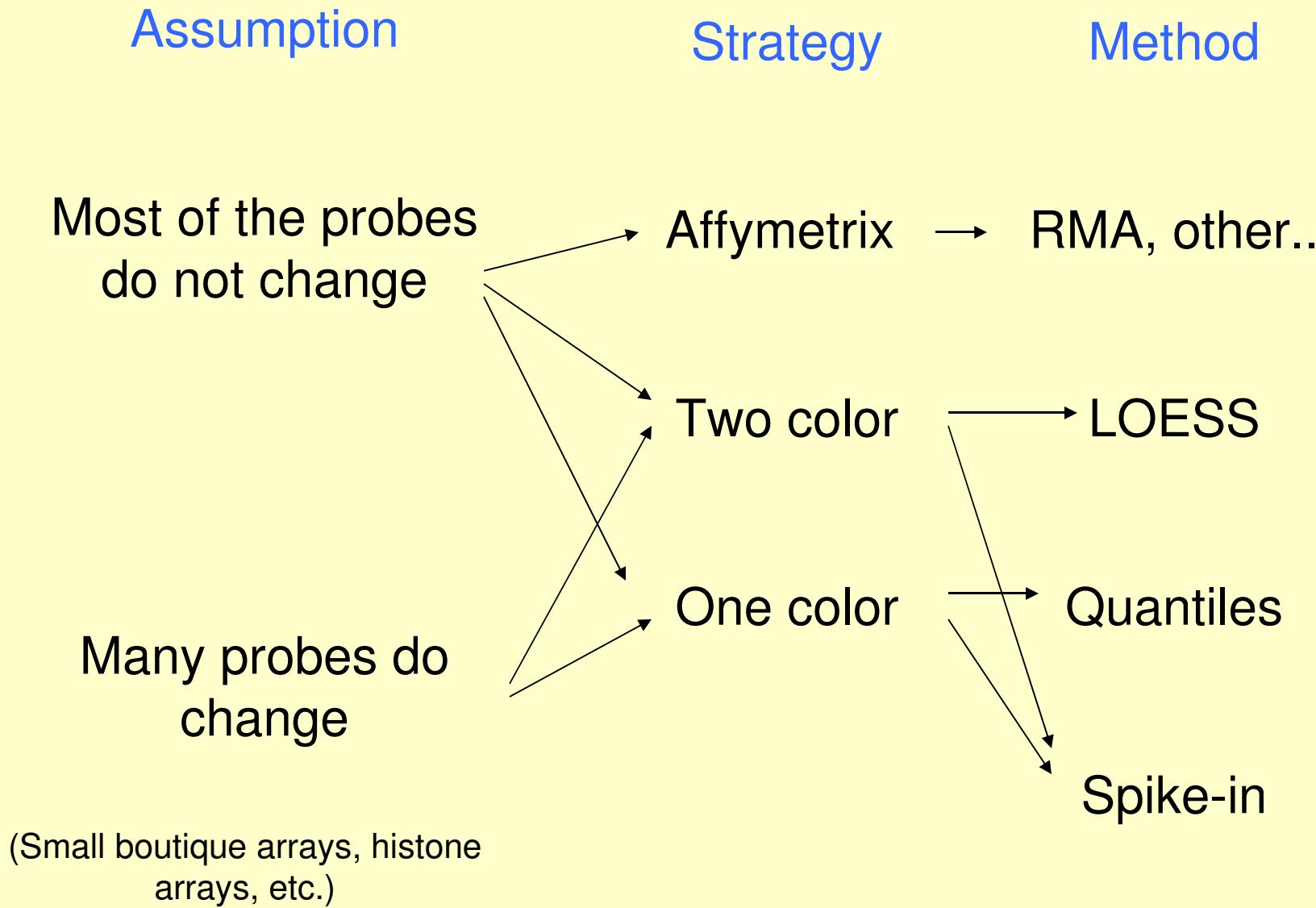
Histones

Etc...



Only a simplified picture. There are more players such as Illumina, small custom-made arrays, etc.

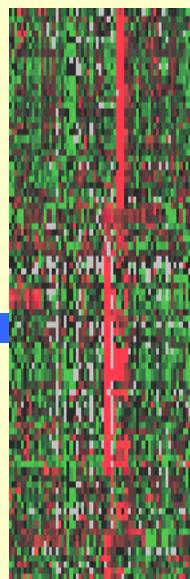
Common normalisation strategies



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

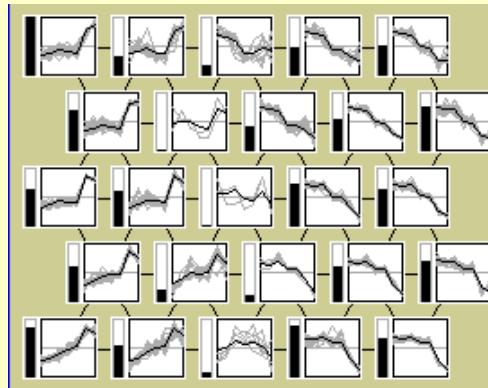
Non hierarchical

hierarchical

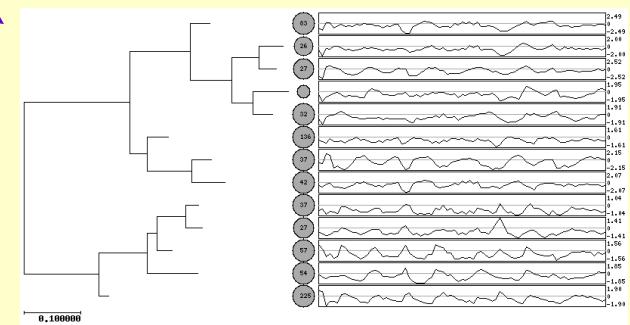
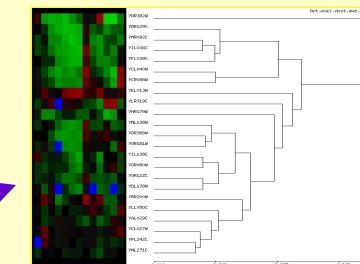
K-means

UPGMA

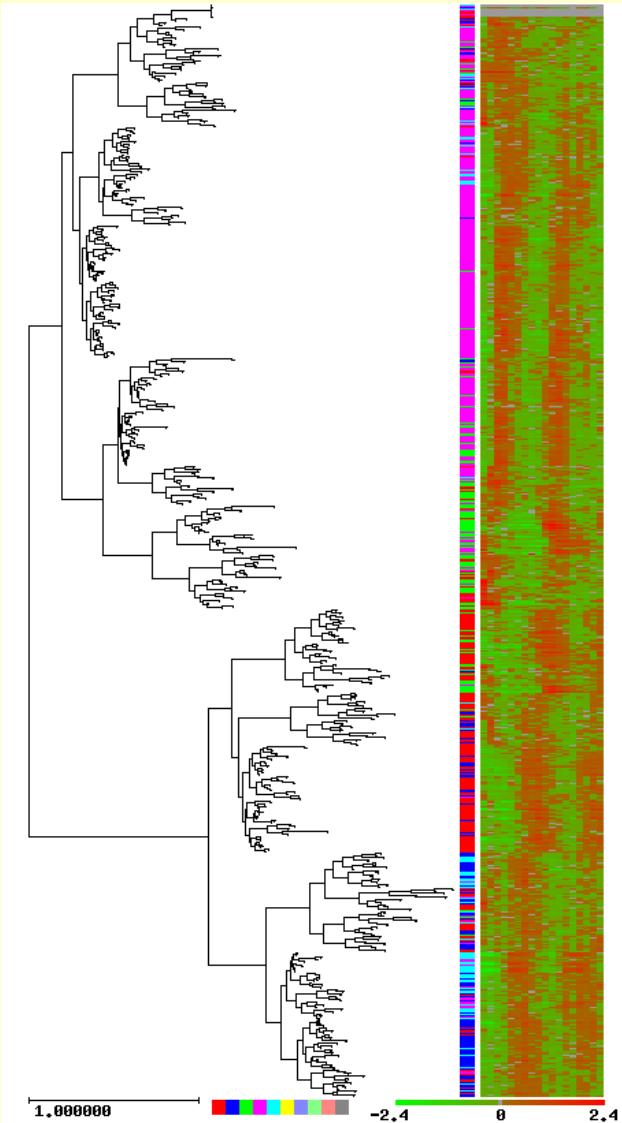
SOM



Different levels of information



An unsupervised problem: clustering of genes.



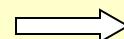
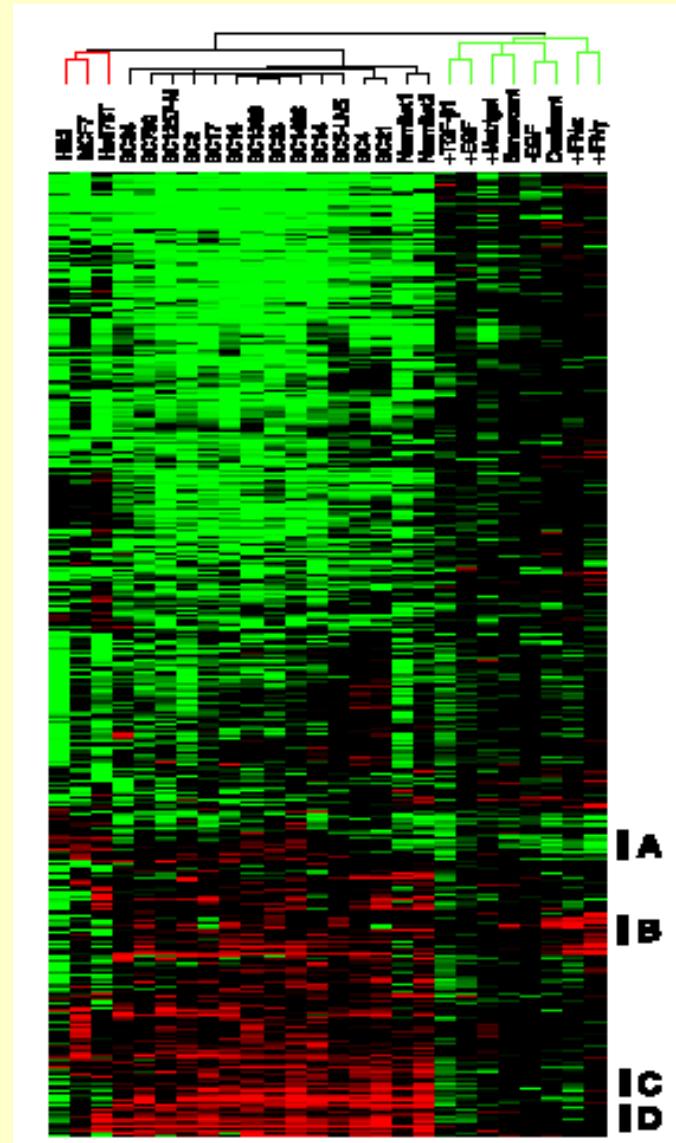
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

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)

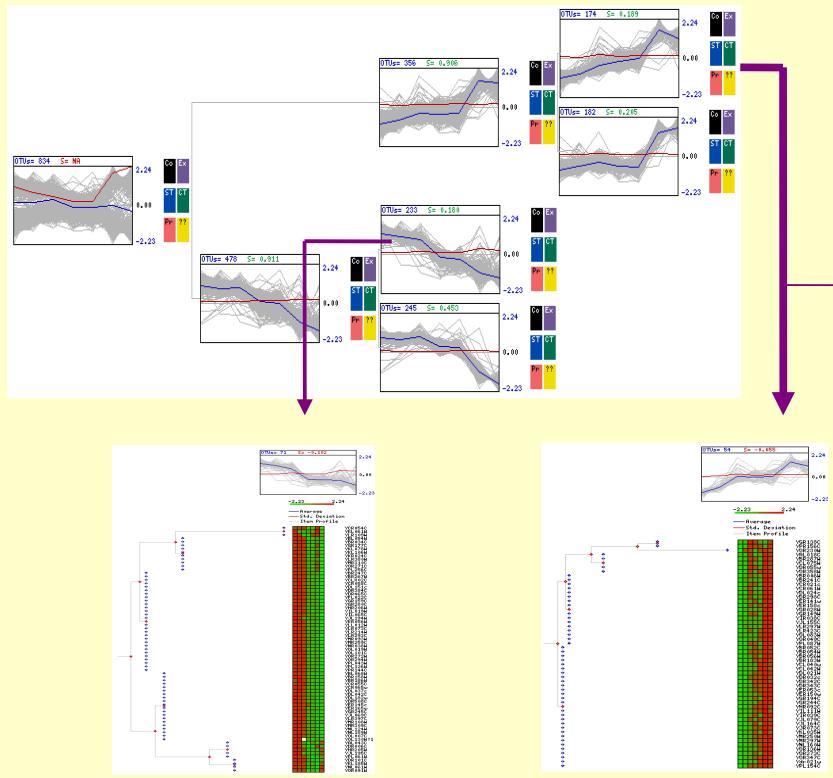
CAAT: Interactive tree browsing

Summary Trees

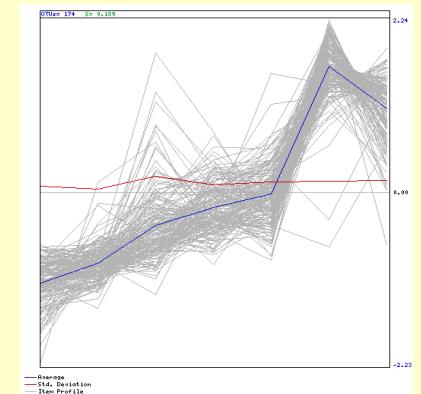
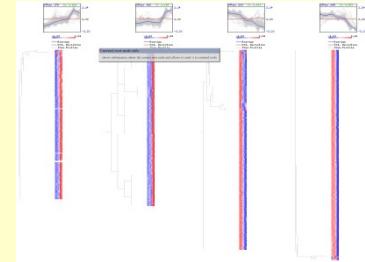
Complete Trees

Cluster validation & analysis

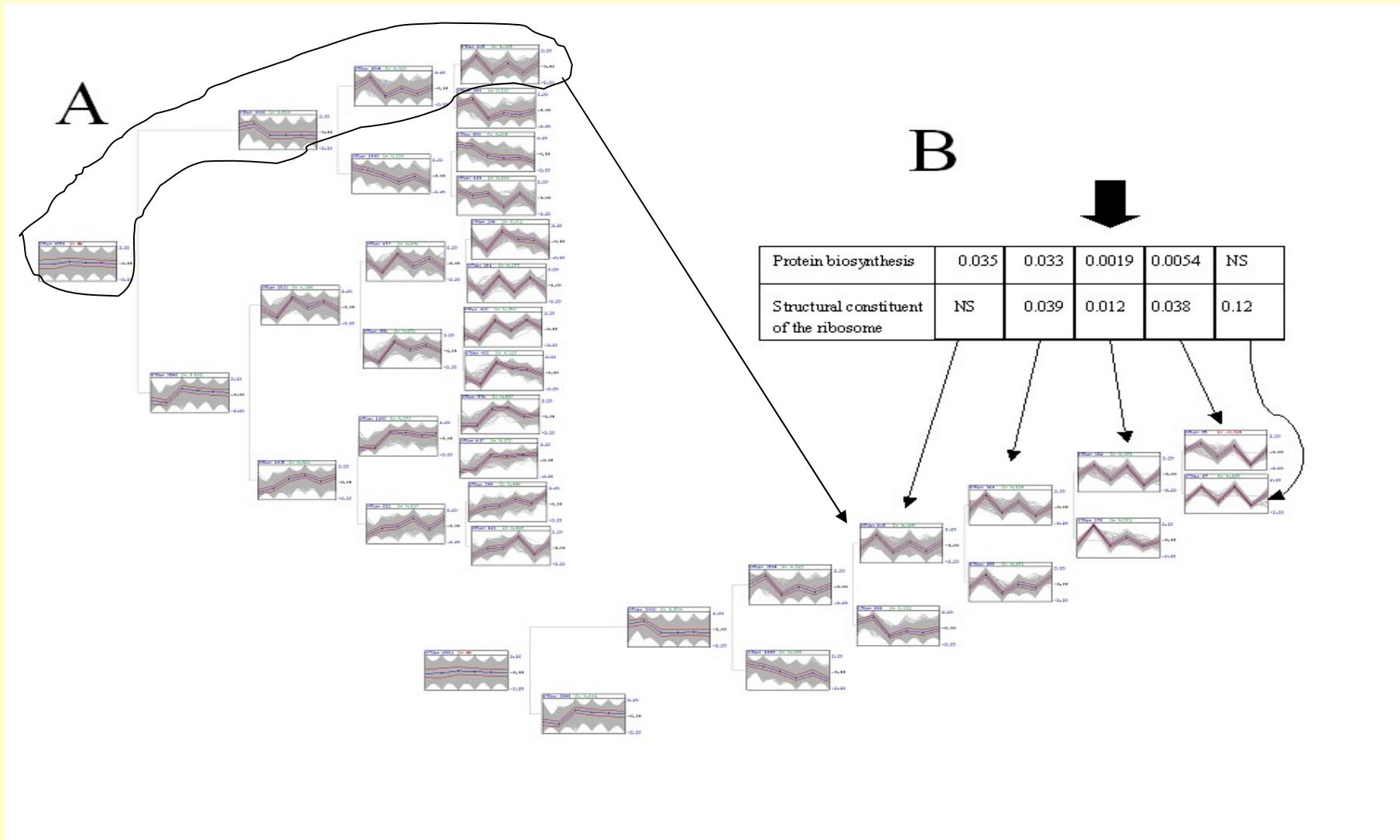
Cluster extraction



Internal_Node_4	
Jmol	Send To FatIGO
Pupa Suite	Send To PinesSNP
InSilicoArray	
CGH	
Send To inSilicoCGH	
ID	4
Distance to Parent	0.022530
Silhouette Index	0.056803
IntraCluster Distance	0.173553
InterCluster Distance	0.154305
Cluster Variance	0.165871
Tree Level	4
Number of items	25
Raw info file	NodeInfo.txt
Partial tree starting at this node (newick)	Extract
Cluster item names	Cluster.txt
Partial array for this cluster	PartialArry.txt
Show Contrary Item List?	Do it!
This Node contains [25] items:	
ID	Name(click to show)
8	YGR138C
9	YPR156C
10	YOR230W
12	YAL018C
13	YBR287W
14	YCL075W
15	YDR055w
16	YOR358W
18	YBR006W
19	YBR241C
20	YCR021c

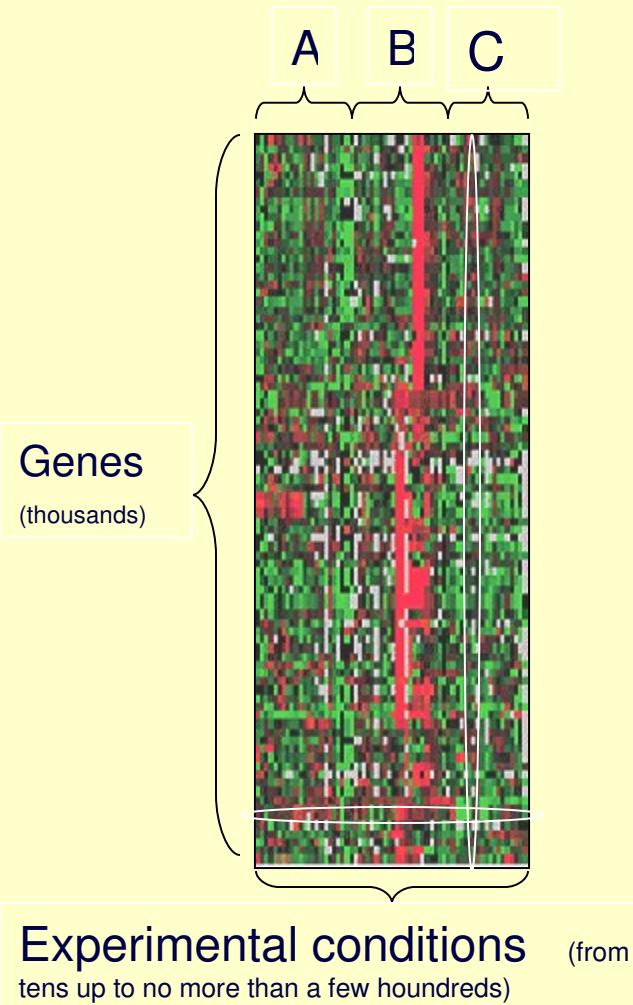


Cluster quality external measures: Functional interpretation



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.



Problems:

How can classes A, B, C... be distinguished based on the corresponding profiles of gene expression?

How a continuous phenotypic trait (resistance to drugs, survival, etc.) can be predicted?

And

Which genes among the thousands analysed are relevant for the classification?

Class prediction

Gene selection

Gene selection.

The simplest way: univariate gene-by-gene.
Other multivariate approaches can be used

- **Two classes**

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

- **Multiclass**

- Anova
- Clear

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

- Pearson
- Spearmam
- Regression

- **Survival**

- Cox model

The screenshot shows a Microsoft Internet Explorer window titled "GEPAS - T-Rex : form - Microsoft Internet Explorer". The address bar shows the URL "http://t-rex.bioinfo.cipf.es/cgi-bin/t-rex.cgi". The top menu includes Archivo, Edición, Ver, Favoritos, Herramientas, Ayuda. The toolbar includes Back, Forward, Stop, Home, Search, Favorites, Multimedia, Mail, Print, Copy, Paste, Cut, Find, and Help. The page header "GEPAS" is displayed, along with "Gene Expression Pattern Analysis Suite v3.0 Bioinformatics Department - CIPF". A navigation menu at the top right includes Tools, Documentation, dataSets, Publications, and About. Below the menu, links for normalization, preprocessing, clustering, supervised classification, differential expressions, functional annotation, cgh arrays, and viewers are listed. A "Tools > Differential expression > T-Rex" link is highlighted. The main content area is titled "T-Rex : form" and features a "two classes" tab selected. It contains fields for "Expression data" (with "Examinar..." button), "Class labels" (with "Examinar..." button), "Test" (radio buttons for t-test, Bayes, Data adaptive, CLEAR test, and CLEAR test), "Significance level" (set to 0.05), "Image appearance" (radio buttons for Standardize, yes/no, Rows 100, Scale -3/+3), "Project name (optional)", "E-mail (optional)", and buttons for "Reset" and "Submit (Run)". Below the form, a "References:" section lists several scientific publications:

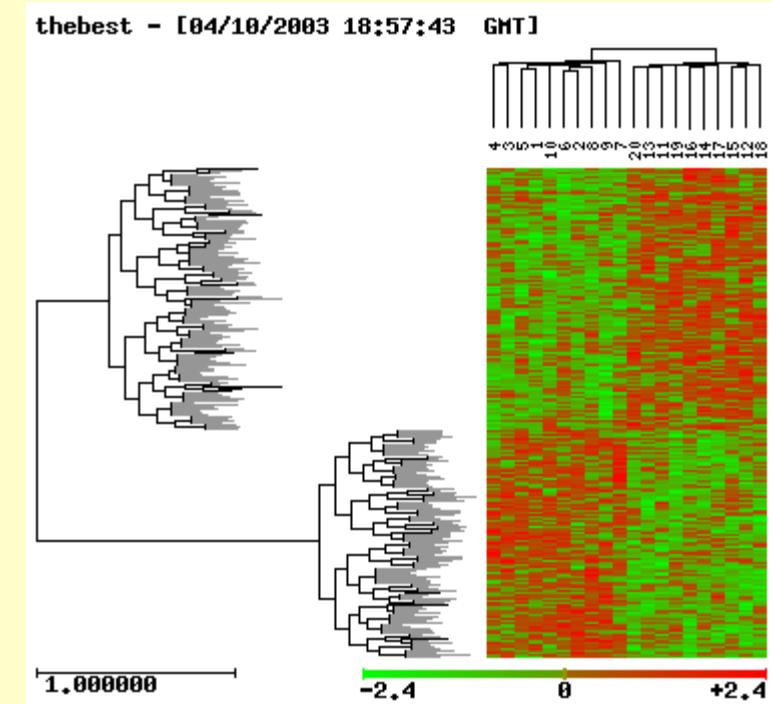
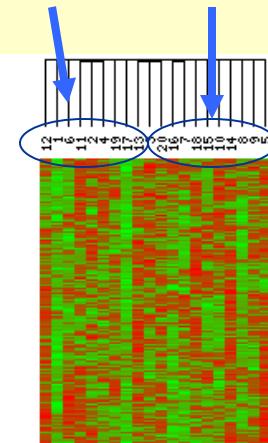
- Vaquerizas J.M., Conde L., Yankilevich P., Cabezon A., Minguez P., Diaz-Uriarte R., Al-Shahrour F., Herrero J. & Dopazo J. (2005). [Gepas: an experiment-oriented pipeline for the analysis of microarray gene expression data](#). Nucleic Acids Research 33 (Web Server issue): W616-W620.
- Herrero J.,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

At the bottom, there is a footer for the Centro de Investigación Príncipe Felipe, CIPF - Avda. Autopista del Saler, 16 - 46013 Valencia - Spain - +34 96 328 96 80, and an Internet link.

The T-rex tool

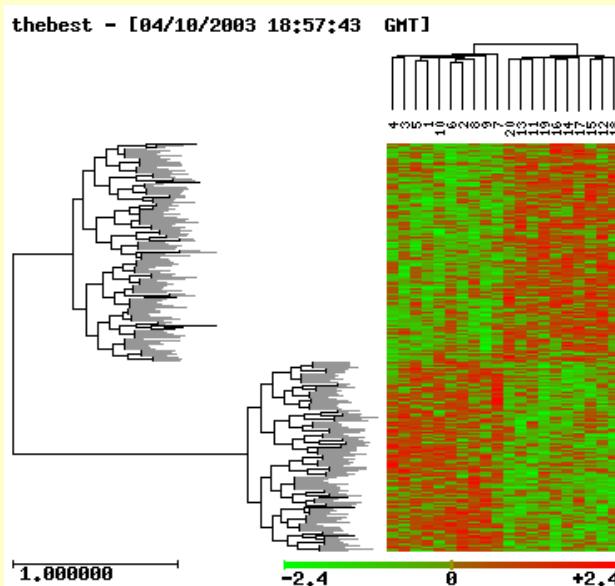
A simple problem: gene selection for class discrimination

~15,000 genes
Case(10)/control(10)



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



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	0.9999	0.861025	1	3.66628
				62427
				60596
				58109
				52935
				43721
				41937
				41428
				.4025
				40212
				37412
				3.36813
				3.35909
				3.35235
				3.28286
				3.2427
				3.23225
				3.22175
				3.19595
				3.19547
				3.12957
				3.0987
				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



...



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

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



10 heads !!!

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

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

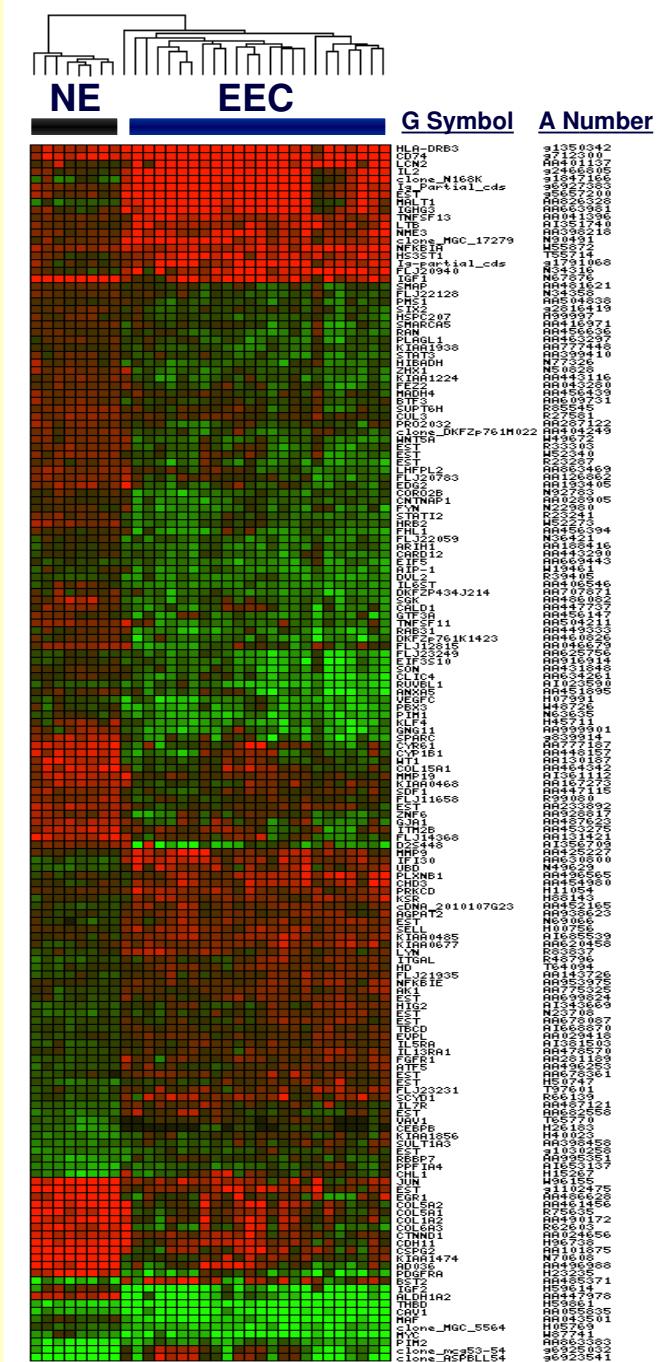
1000 coins

Will you still use this coin
for betting?

Genes differentially expressed between normal endometrium (ne) and endometrioid endometrial carcinomas (eec)

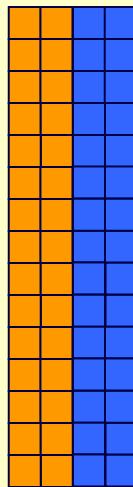
Hierarchical Clustering of 86 genes with different expression patterns between Normal Endometrium and Endometrioid Endometrial Carcinoma (FDR adjusted $p < 0.05$) selected among the ~7000 genes in the CNIO oncochip

Moreno et al., 2003 Cancer Research 63, 5697-5702



Of predictors and molecular signatures

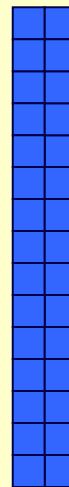
A B X



Is X, A or
B?



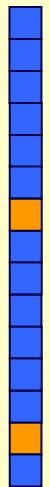
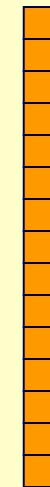
Diff (B, X) = 2



What is a predictor?
Intuitive notion:



Diff (A, X) = 13



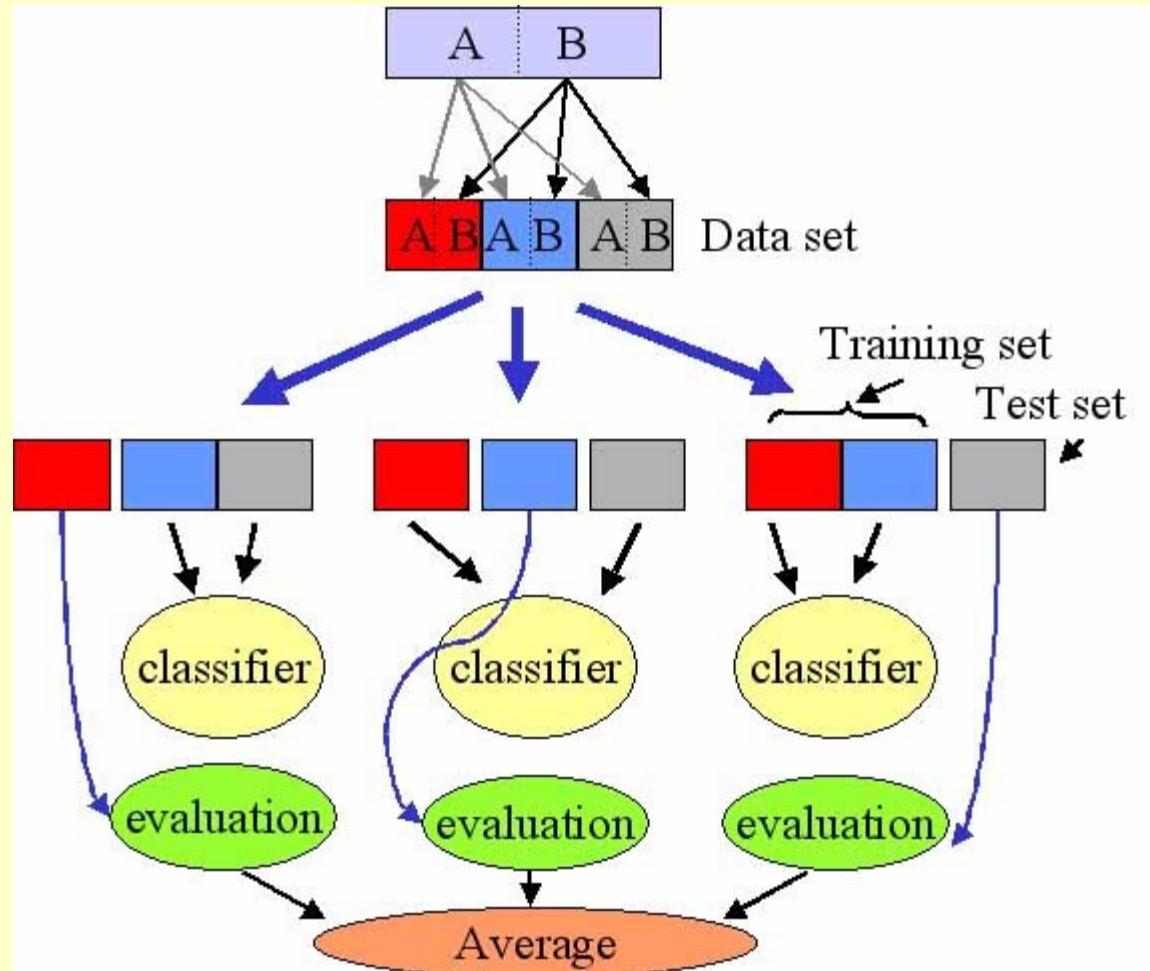
Most probably X belongs to class B

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



Predictors

- Gene selection

- F-ratio

- Wilcoxon test

- Predictors

- SVM

- KNN

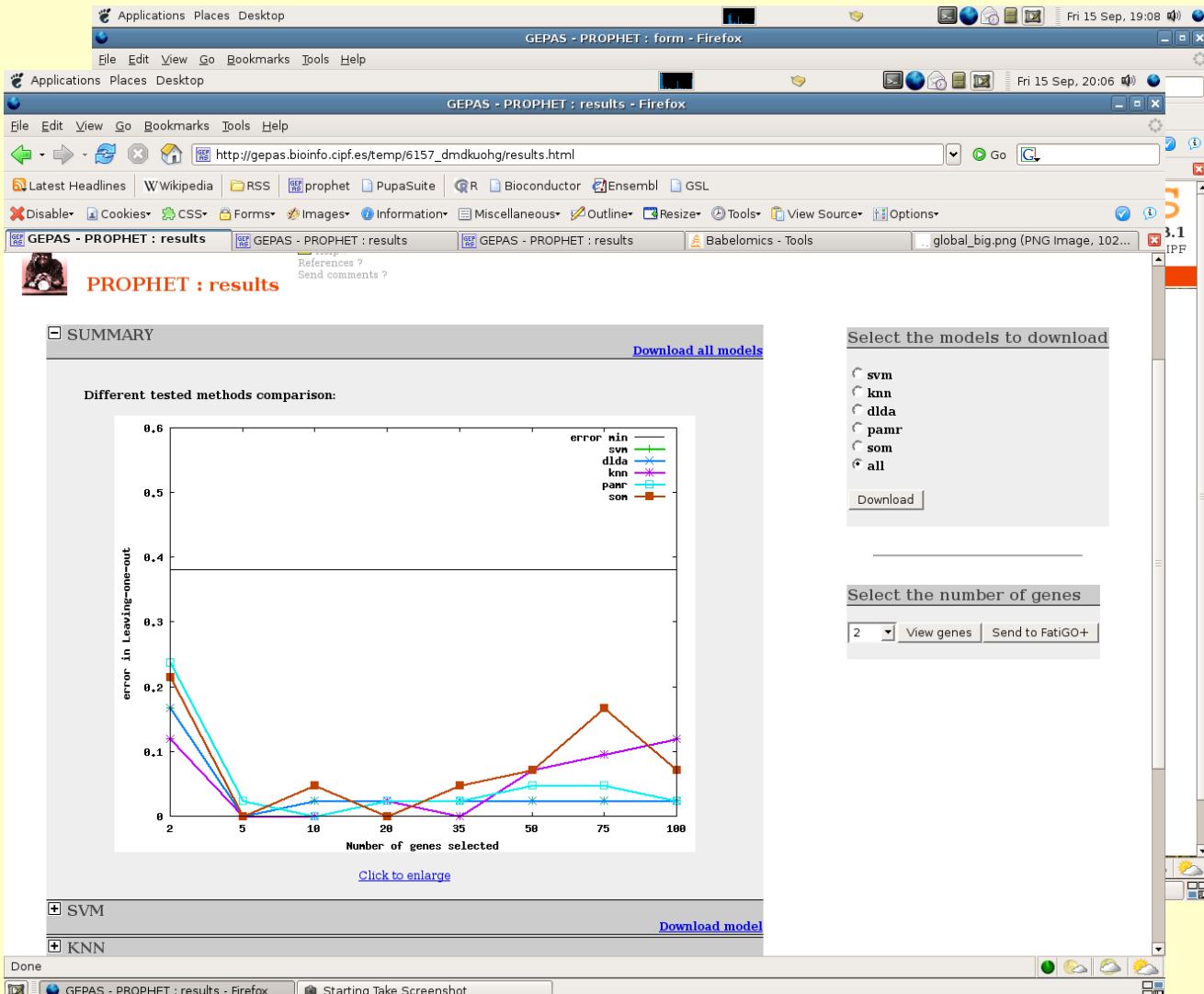
- DLDA

- PAM

- SVM

- Unbiased CV error

- Confusion matrices



The prophet tool: the only class predictor on the web based on genes

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

My data...

How are
structured?

What are
these
groups?

What is this
gen?

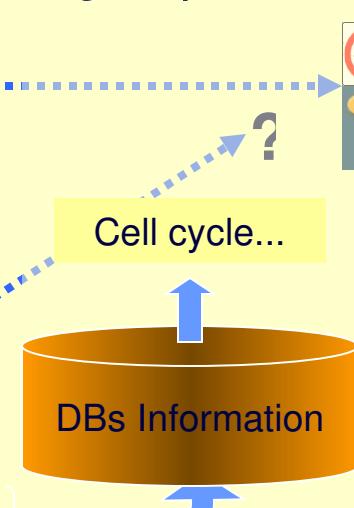
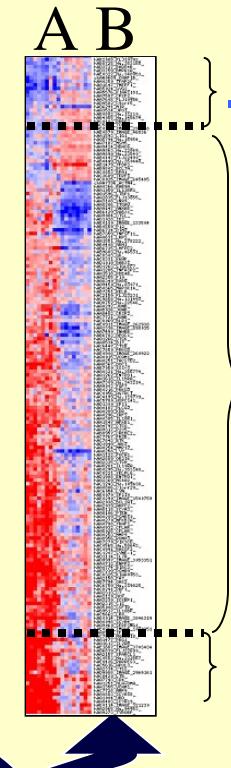
Microsoft Excel

Archivo Edición Ver Insertar Formato Herramientas Datos Verágina Ayuda Archivo

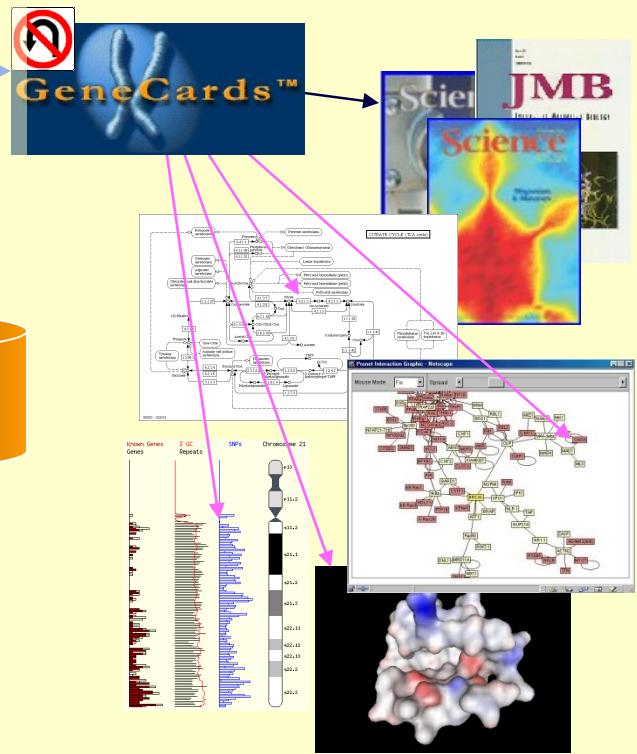
Expression Analysis: Pivot Tab

Total.xls [Sólo lectura]

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	X61388	SGD: YEL020C Yeast S.cerevisiae WBP1 Oligosaccharyl				
68	532.2 *	~1.6	0.22	X1630	SGD: YEL021W Yeast S.cerevisiae Protein of unknown fu				
69	587.7 *	~1.6	0.27	M23316	SGD: YEL024C Yeast S.cerevisiae RIP1 Rieske iron-sulfur				
70	~114.5 *	~1.1	0.03	K02207	SGD: YEL021W Yeast S.cerevisiae URA3 gene coding for				
71	~125.4 *	~1	-0.01	Cluster Incl	M16495 Calpastatin I light chain /cds=(68,36) /gi=M16				
72	~1091.6	~1.2	-0.14	Cluster Incl	Z57749 M.musculus spermidine synthase gene /cds=(
73	~757.2	~1.3	-0.17	Cluster Incl	Z12973 Murine MLC1F/MLC3F gene for myosin alkali				
74	9626.6	1.3	0.63	Cluster Incl	A649035 U-M-AH1-agc-a-06-U1 s1 Mus musculus c				
75	~847.4	~1.3	-0.21	Cluster Incl	AW12542 U-M-BH2-1-qb-F01-0-U1 s1 Mus musculus c				
76	~2693.1	1.1	0.09	Cluster Incl	AF059893 Mus musculus proteasome alpha/70 subu				
77	192.5 *	~1.2	0.05	Cluster Incl	AB006361 Mus musculus mRNA for prostaglandin D s				
78	2801.2	~4.4	1.63	Cluster Incl	AB006361 Mus musculus mRNA for prostaglandin D s				
79	~201	-1	0.02	Cluster Incl	AB011081 Mus musculus mRNA for huntingtin interact				
80	1389.9 *	~2.6	1.81	Cluster Incl	AB011081 Mus musculus mRNA for huntingtin interact				
81	753.2 *	1.2	0.1	Cluster Incl	U97170 Mus musculus protein kinase inhibitor gamma				
82	~2774.7	~1.9	-1.43	Cluster Incl	M51220 keratin complex 1, acidic, gene 19 /cds=(0,12)				
83	3614.4 *	~5.1	1.98	Cluster Incl	U19604 DNA ligase I, ATP-dependent /cds=(0,3054)				
84	0 *	~0.0	0.02	Cluster Incl	AB5142 U-M-BH2-1-qb-D-U1 s2 Mus musculus c				
85	3310.9	1.2	0.24	Cluster Incl	AB029408 Mus musculus mRNA for slab7b, complete				
86	~1291	~1.5	-0.42	Cluster Incl	AF059735 Mus musculus C-terminal binding protein 2				
87	~263.3 *	~1.3	-0.02	Cluster Incl	AF034545 Mus musculus tetraspan TM4SF (Tspan-6)				
88	77.5 *	1.1	0.01	Cluster Incl	D46892 Hydroxysteroid 17-beta dehydrogenase 1 /cds=				
89	2047.2 *	~3.3	1.1	Cluster Incl	AF039299 Mus musculus 17-beta-hydroxysteroid dehy				
90	808.9 *	~1.9	0.38	Cluster Incl	M84407 Vascular cell adhesion molecule 1 /cds=(57,2				
91	~124.3 *	~1.1	-0.03	Cluster Incl	U12841 Mus musculus C57BL/6 vascular cell adhesio				
92	~675.5 *	~1.8	-0.37	Cluster Incl	U12841 Mus musculus C57BL/6 vascular cell adhesio				
93	1465.4 *	~2.7	0.76	Cluster Incl	A23863 Mus musculus mRNA for nucleoside diphos				
94	838.2	1.1	0.1	Cluster Incl	U70476 Nuclear factor, erythroid derived 2, like 2 /cds=				
95	4969.4 *	~6.7	8.84	Cluster Incl	AF045673 Mus musculus FLJ11RR associated protein				
96	148.3 *	~1.2	0.04	Cluster Incl	AB91475 Y59d06.x1 Mus musculus cDNA, 3 end /cds=				



Functional profiling

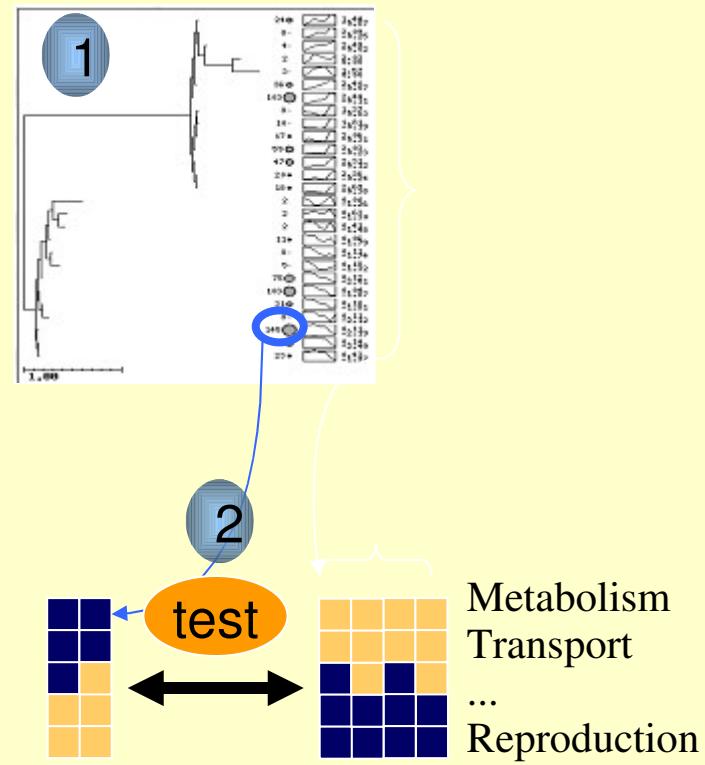
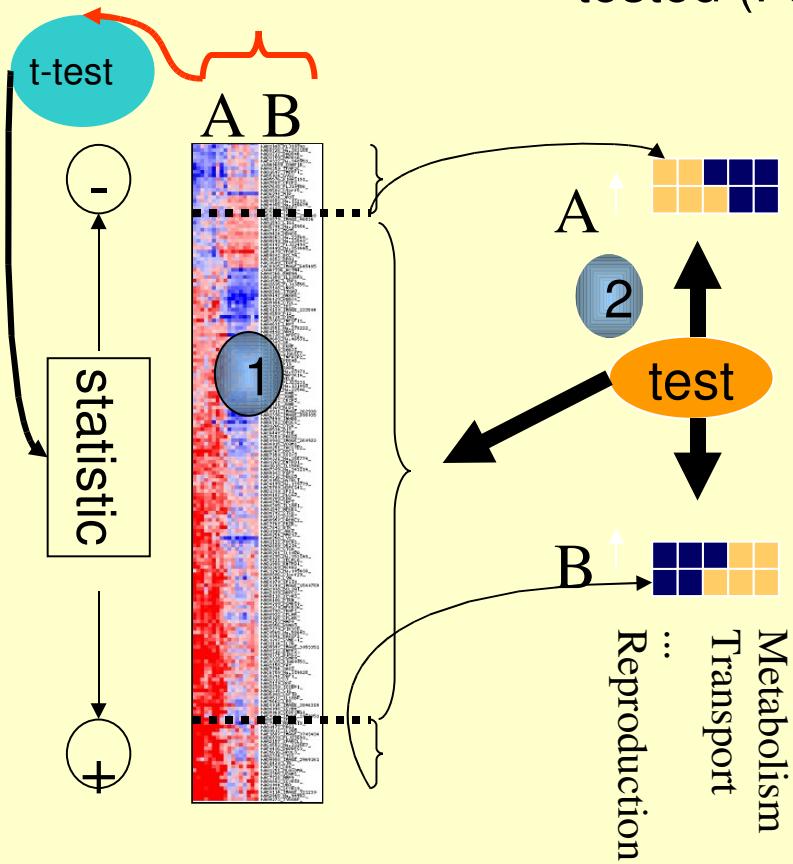


Analysis

Links

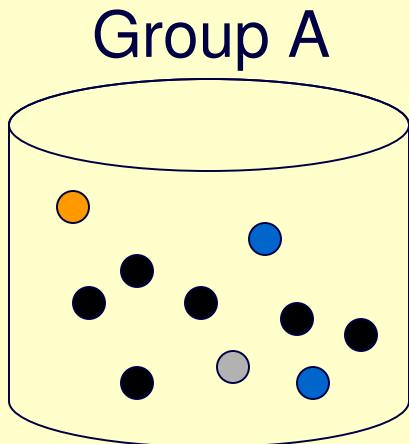
Functional enrichment (two steps)

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

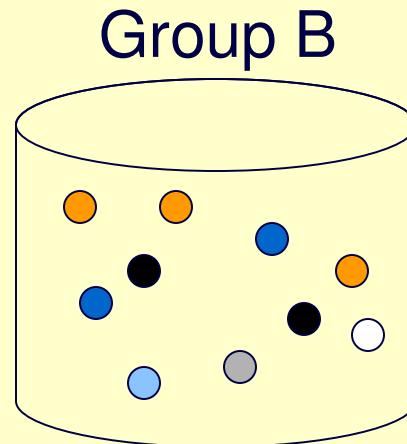


Testing two GO terms

(remember, we have to test thousands)



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



Biosynthesis Other

A	6	4
B	2	8

A
B

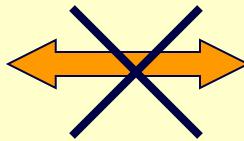
The popular Fisher's test

Biosynthesis 60% ●



Biosynthesis 20% ●

Sporulation 20% ●



Sporulation 20% ●

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

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 screenshot shows two side-by-side web pages. On the left is the 'Gene Ontology Home' page, featuring a sidebar with links like 'Downloads', 'Annotations', and 'Mappings to GO'. Below it is a search bar for 'GOI' (Gene or protein name) and a link to the 'GO website'. A red arrow points from the 'GO website' link on the left to the 'FatiGO' section on the right. On the right is the 'Tools for Gene Expression Analysis - Microsoft Internet Explorer' page, specifically the 'GO-tools.microarray.shtml' page. This page contains sections for 'ermineJ', 'FatiGO', and 'FuncAssociate', each with a brief description and a link to a 'PubMed abstract'.

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

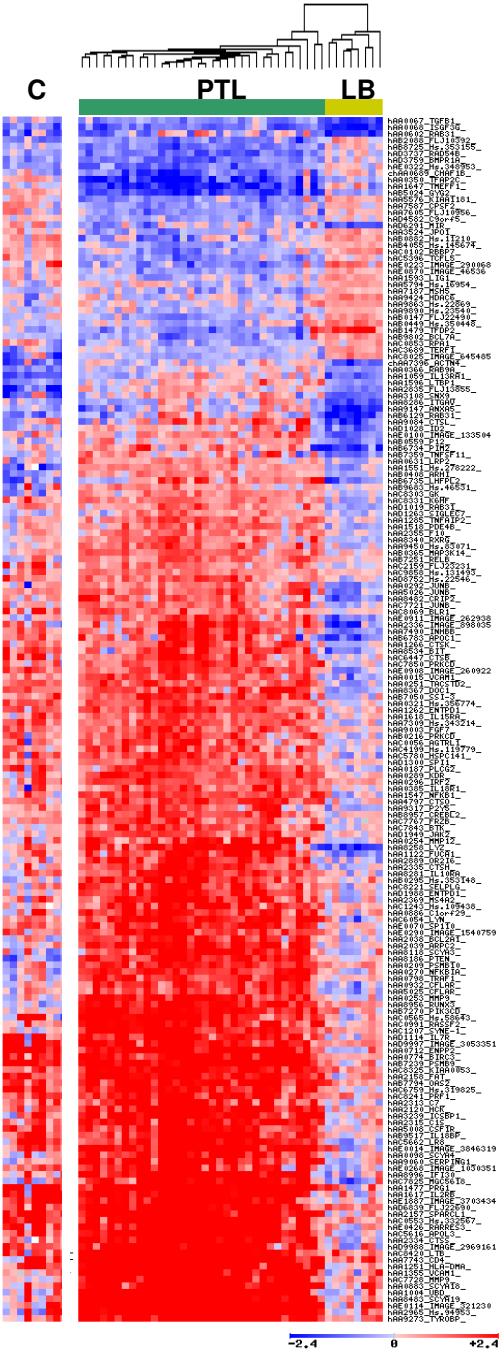
Al-Shahrour et al., 2006 NAR

Al-Shahrour et al., 2007 NAR; Al-Shahrour et al., 2007, BMC Bioinformatics

Functional terms

The screenshot shows the FatiGO+ web interface. At the top, there are logos for Bioinformatics, PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN, and BABELOMICS. Below the header, there are tabs for Home, Tools, Tutorials, Papers, and About. The main area has a logo for FatiGO plus and navigation links for search, compare, and genomics. A large blue arrow points from the 'Organism' input field to a box labeled 'Organism'. Another blue arrow points from the 'List of genes #1' or 'genes list file #1' input fields to a box labeled 'Gene List1'. A third blue arrow points from the 'List of genes #2' or 'genes list file #2' input fields to a box labeled 'Gene List2'. On the left, under 'Functional annotation', there is a list of categories with checkboxes: 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, and cisRED: cis-regulatory element. At the bottom, there are fields for 'E-mail (optional)', 'Project name (optional)', and a 'Submit' button, with a 'Run' button below it.

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



Understanding why genes differ in their expression between two different conditions

Lymphomas from mature lymphocytes (LB) and precursor T-lymphocyte (PTL).

Genes differentially expressed, selected among the ~7000 genes in the CNIO oncochip

Genes differentially expressed among both groups were mainly related to immune response (activated in mature lymphocytes)

Martinez et al., Clinical Cancer Research.
10: 4971-4982.

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 currents Cluster IDs

Total number of genes no repeated with current Cluster IDs:

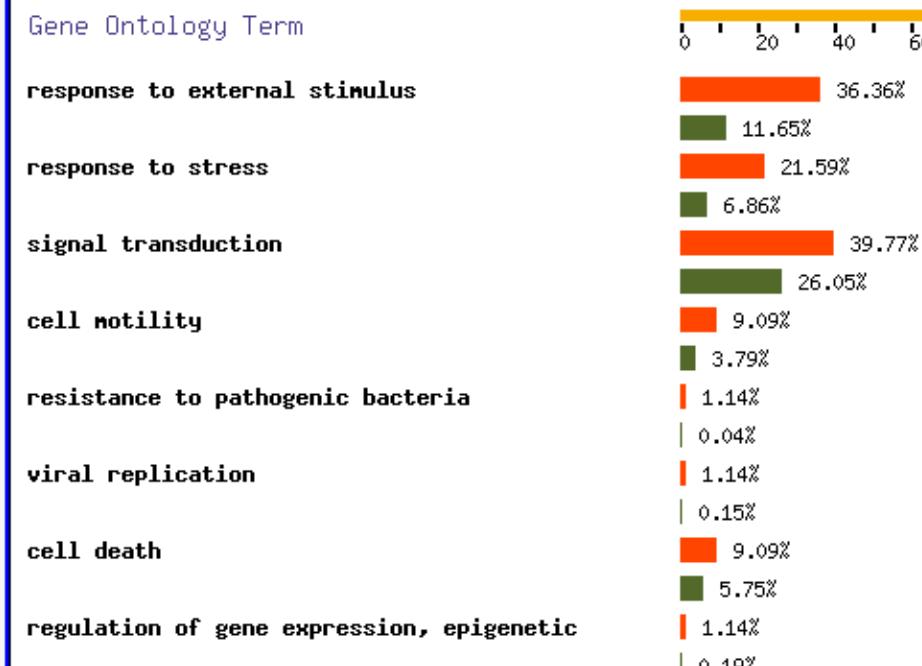
Total number of genes no repeated with GO at level 3 and biological_process:

Total number of genes no repeated with GO but NOT at level 3 and ontology:

Total number of genes no repeated without GO annotated:

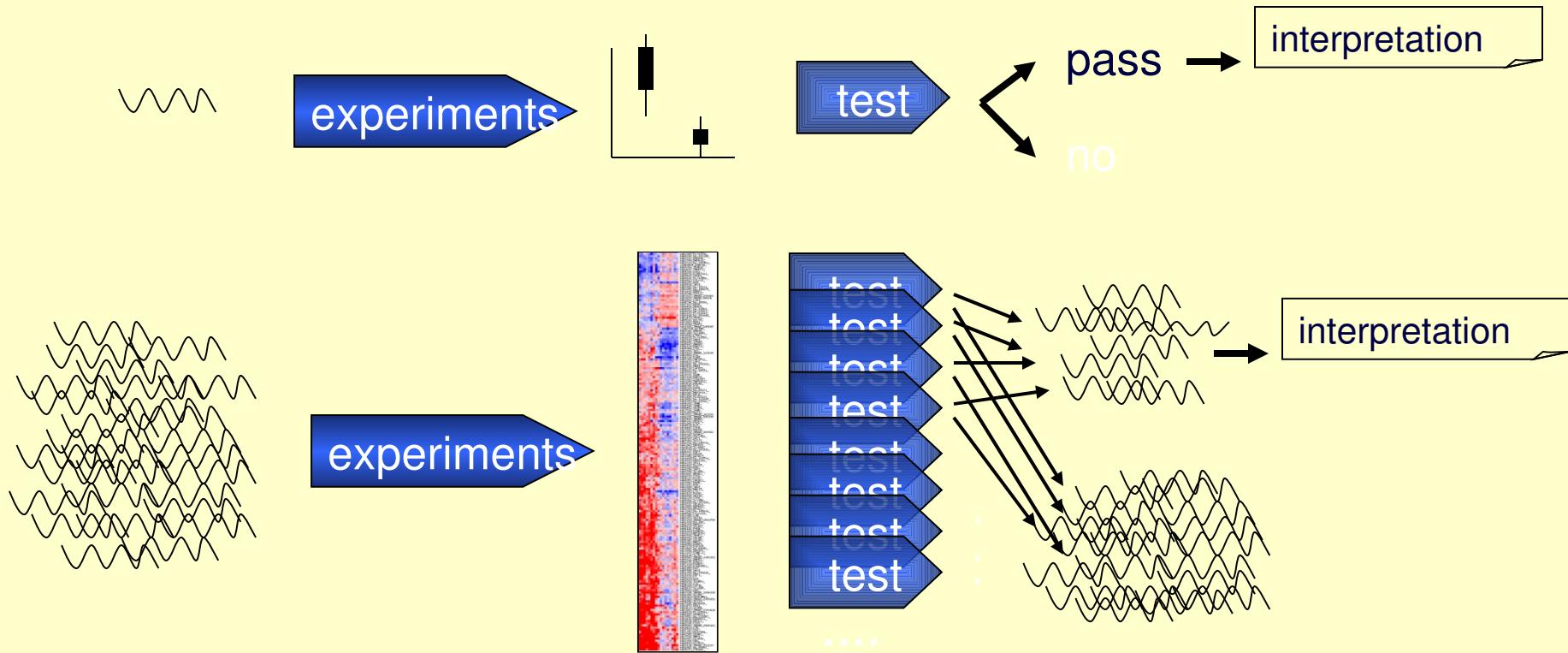
Cluster Query	Cluster Reference
162	4764
129	4731
7 - 23	449 - 1627
145	5909
88	2610

Obvious? NO



- 2) If you do not have previously a strong biological hypothesis, now you have an explanation
- 3) You now know that there are no other co-variables (e.g. age, sex, etc)

Conventional gene selection reproduces pre-genomics paradigms



Genes do not operate alone.

Context and cooperation between genes is ignored

Now: 22240 (NCBI build 35 12/04)

50-70% display alternative splicing

25%-60% unknown

Transfrags

Genes in the DNA...

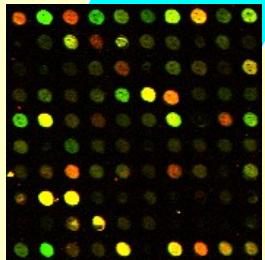
>protein kinase

```
acctgtgtatggcgacaggactgtatgtatc  
tatgtgtatgtatcatgtgtactgtatgtatgg  
ggcttattgtactgtatgtatcata....
```



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

A typical tissue is expressing among 5000 and 10000 genes

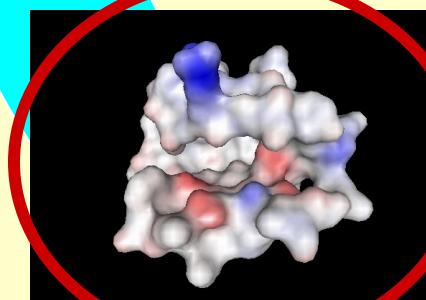


...code for proteins...

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

100K-500K proteins

...whose structures account for function...



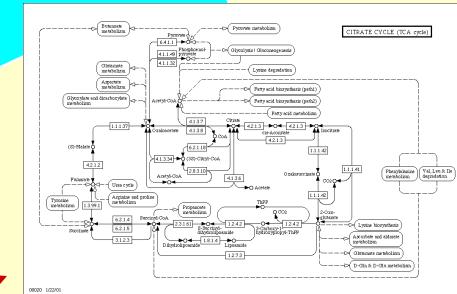
...which can be different because of the variability.



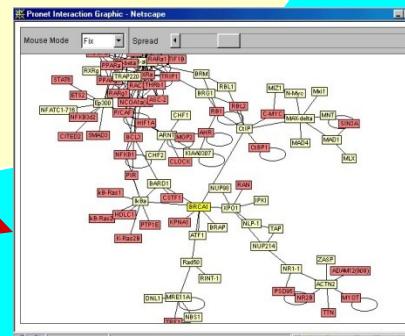
10 million SNPs

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

Function is carried out by gene sets



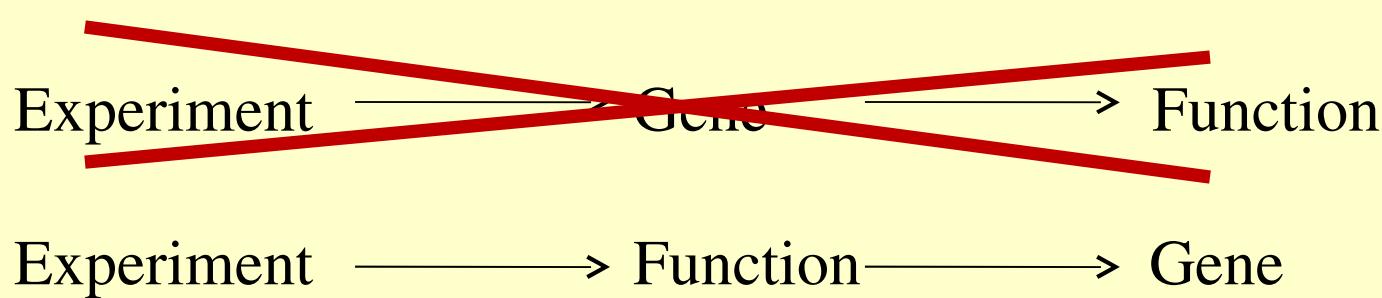
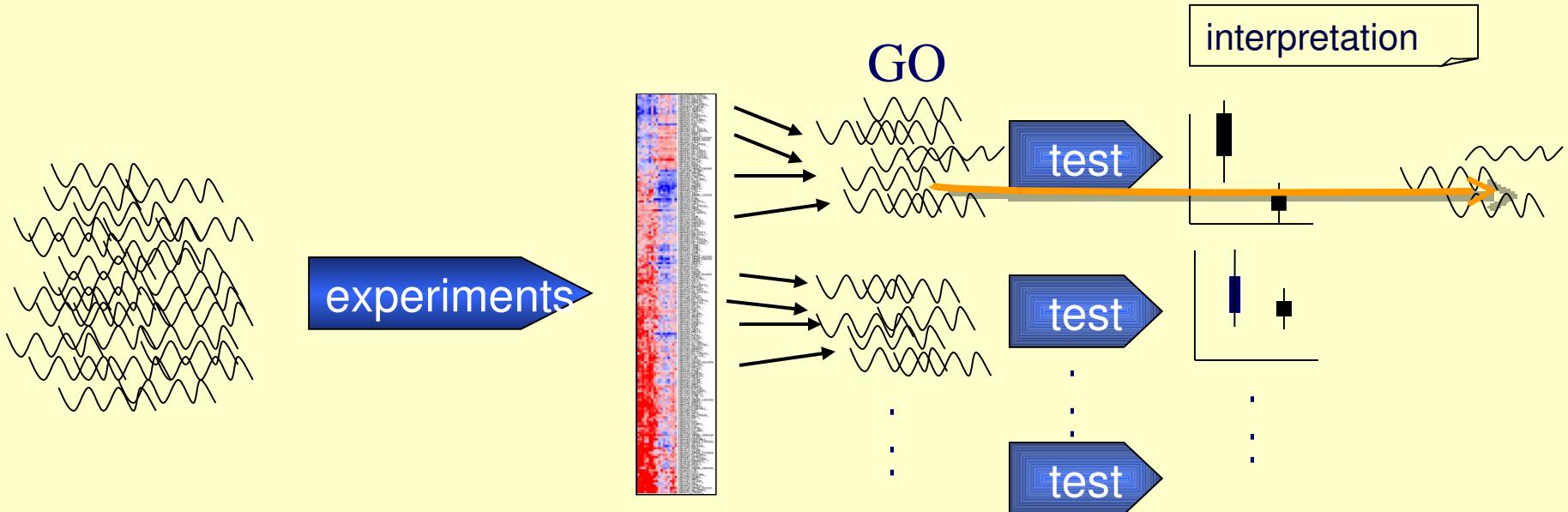
...conforming complex interaction networks...



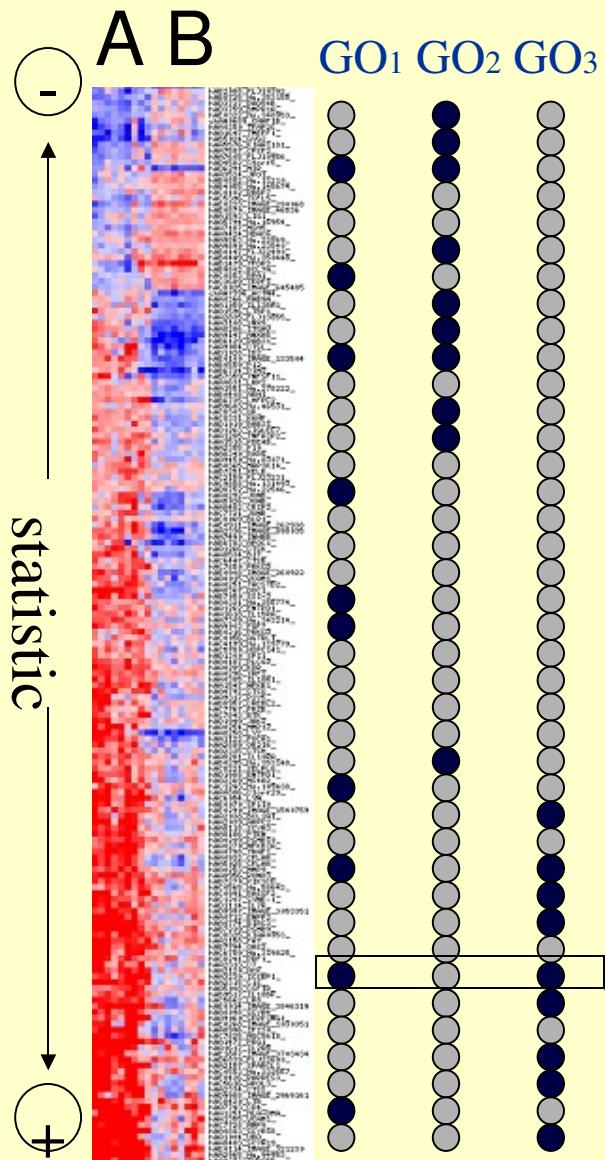
...in cooperation with other proteins...

Each protein has an average of 8 interactions

Hypothesis inspired in biology (functionality carries out by gene-sets) must be tested



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



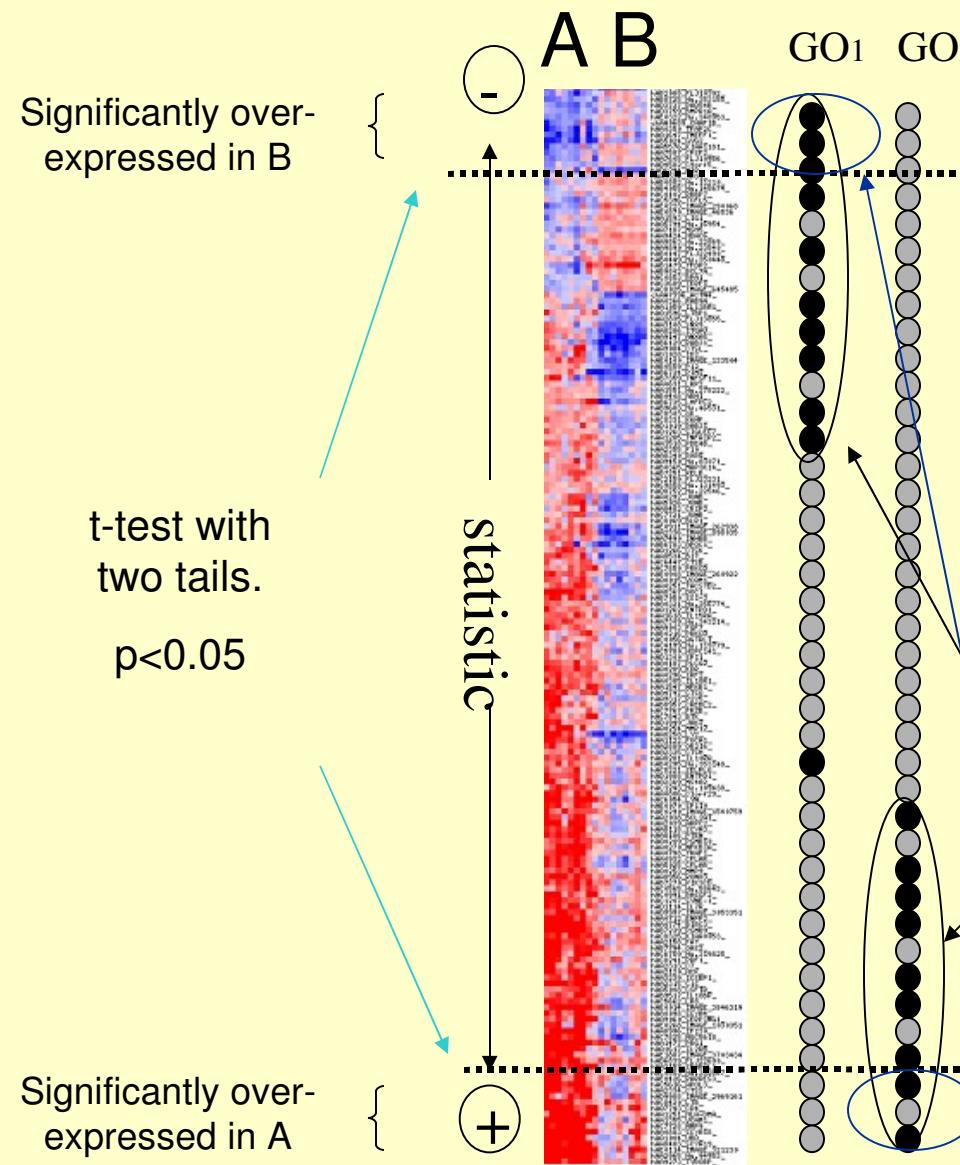
Ranking: A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

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



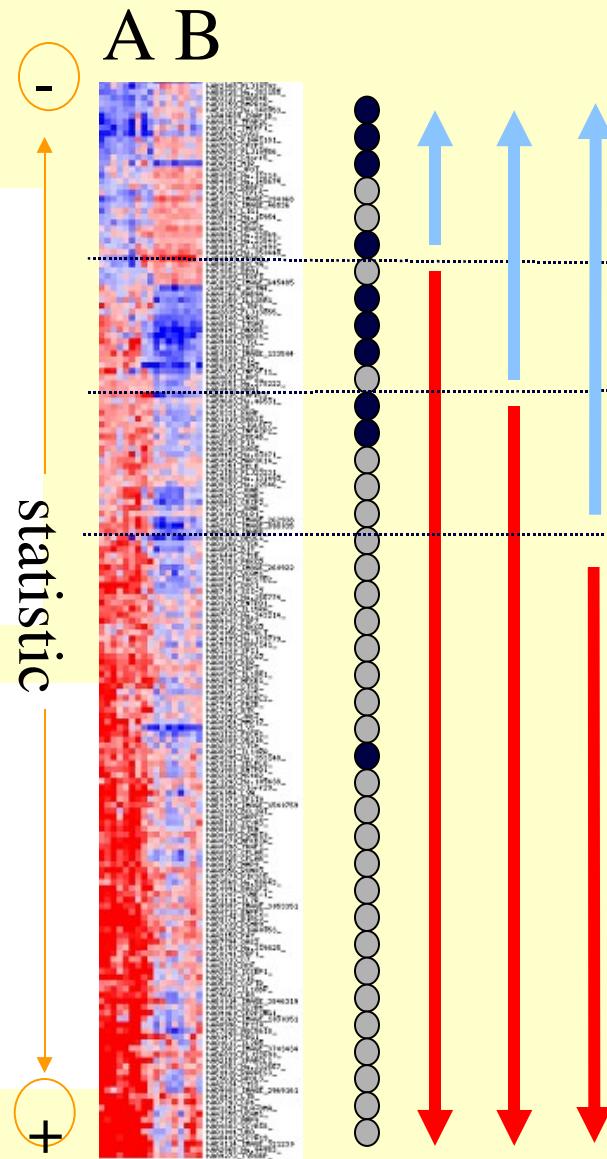
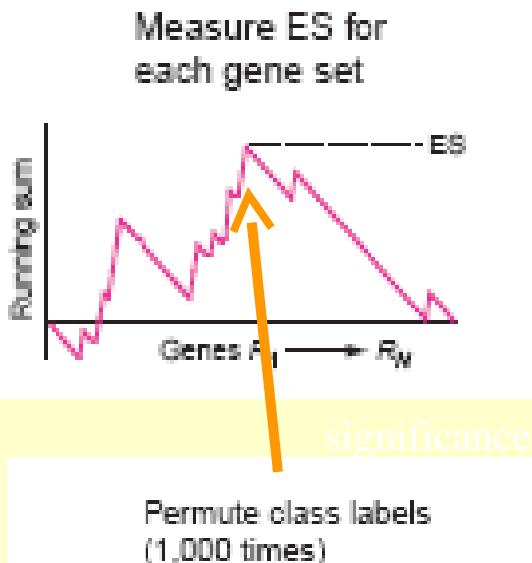
If a threshold based on the experimental values is applied, and the resulting selection of genes compared for overabundance of a functional term, this might not be found.

Classes expressed as blocks in A and B

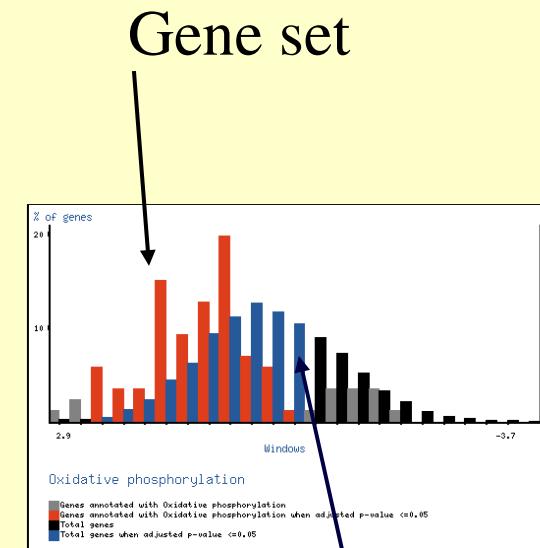
Very few genes selected to arrive to a significant conclusion on GOs 1 and 2

Gene-set enrichment methods

GSEA



FatiScan

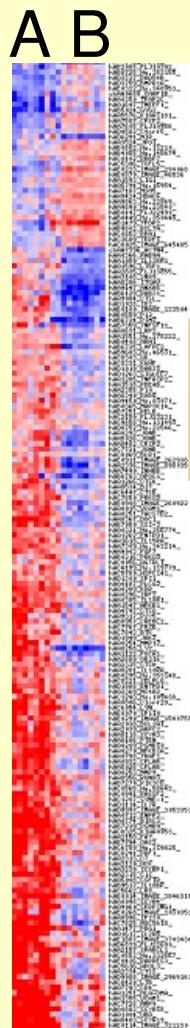


Independent of the experimental design

Case study: functional differences in a class (case/control) comparison experiment

A

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



B

17 with normal tolerance to glucose (NTG)

(Mootha et al., 2003)

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

	Healthy vs diabetic	Functional class	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		
Dow-regulated		Nuclease activity	X		
		Insulin signalling pathway		X	

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)



GEPAS
t-rex tool

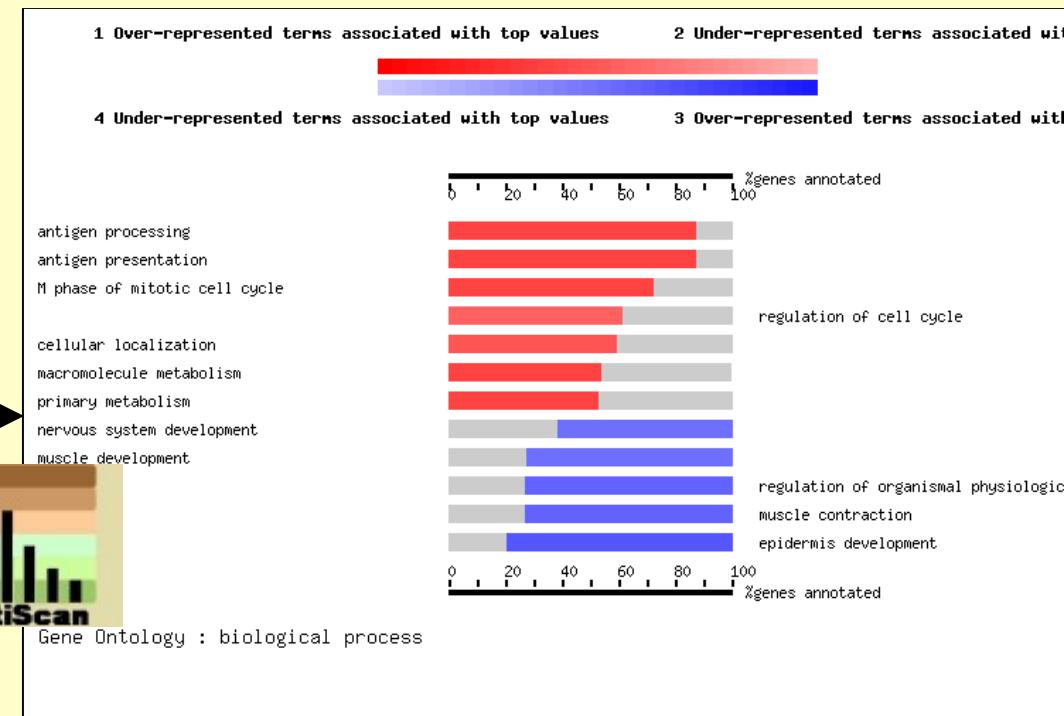
Cox Proportional-Hazards model to study how the expression of each gene across patients is related to their survival

- 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

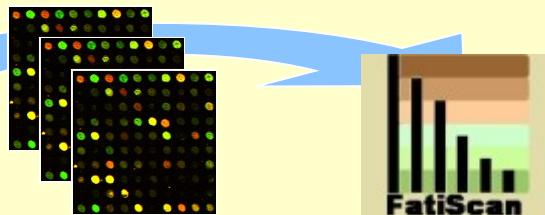
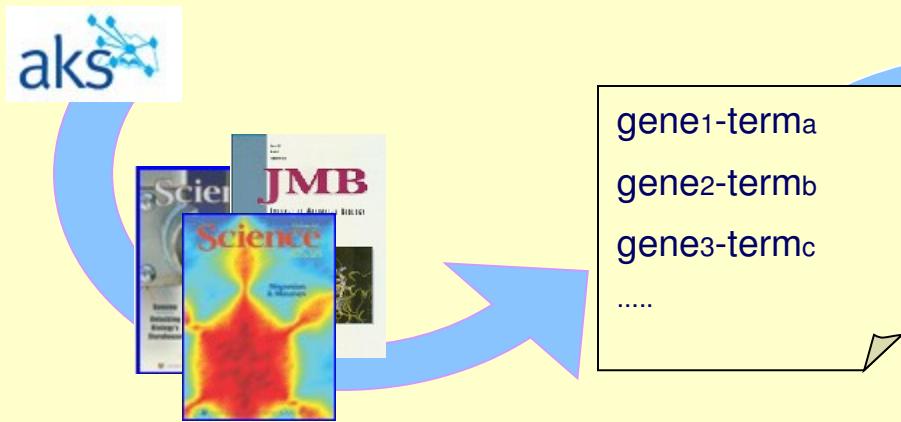


+ Survival



Beyond the classical repositories: Using text-mining-derived functional terms

Text-mining methods allows extracting informative terms (**bioentities**) with different functional, chemical, clinical, etc. meanings, that can be associated to genes.



Compound	bioentity(chemical product)	adjusted p-value (FDR)
1 a-methyl-L-p-tyrosine	ACETYLCHOLINE	0.0237
	CAMP	0.0176
	N-ETHYLMALEIMIDE	0.0324
caffeine	PHOSPHATIDYLINOSITOL	0.0191
fluorouracil	PURINE	0.0327
fluorouridine	HYDROGEN PEROXIDE	0.0355
	GLUTATHIONE	0.0237
	PHOSPHATIDYLINOSITOL	0.0176
	NITROUS OXIDE	0.0227
	GALNAC	0.0191
	DOPAMINE	0.0237
pergolide methanesulfonate	PHOSPHATIDYLINOSITOL	0.0237
sulmazole	CHOLINE	0.0237

List of **bioentities** found to be significantly over-represented in the treatment of AML cells with an specific compound (Stegmaier, K., et al. 2004, Nat Genet, 36, 257-63)

Babelomics suite for functional interpretation

<http://www.babelomics.org>



FatiGO: Fast transference of Information using **Gene Ontology**.



FatiGOplus: an extension of FatiGO for **InterPro motifs**, **KEGG pathways** and **SwissProt keywords** , **transcription factors (TF)**, **gene expression in tissues**, **bioentities from scientific literature**, **cir-regulatory elements CisRed**.



Tissues Mining Tool: compares reference values of **gene expression in tissues** to your results.



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



FatiScan: Detects blocks of functionally related genes (**GO terms**, **InterPro motifs**, **KEGG pathways** and **SwissProt keywords** , **transcription factors (TF)**, **cir-regulatory elements CisRed**, etc.) with significant coordinate (although modest) over- or under-expression using a segmentation test.



GSEA: Detects blocks of functionally related genes (GO, KEGG, etc.) with significant coordinate (although modest) over- or under-expression using a modified Kolmogorov-Smirnov test.

Comparison of threshold-free 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
Down-regulated	NADH dehydrogenase (ubiquinone) activity	+				yes	-	-	-
	Nuclease activity	+				yes	-	-	-
Insulin signalling pathway			+			yes	-	-	-

GSEA 2003
FatiScan 2005
PAGE 2005
Tian 2005

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

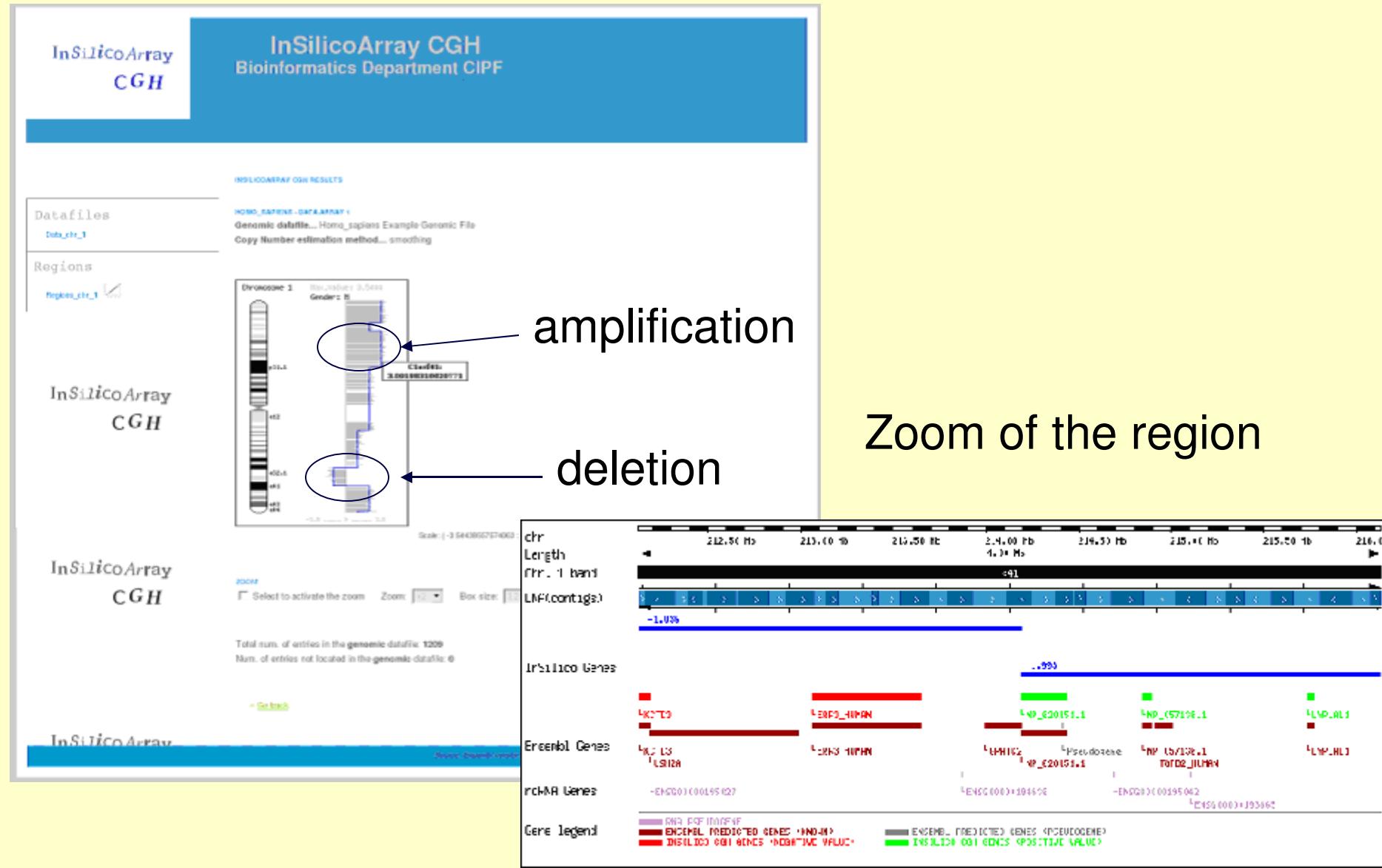
ISA CGH (*In silico* Array-CGH)

- Estimating copy number variation
- Correlation copy number – expression
- Minimum common amplified / lost region

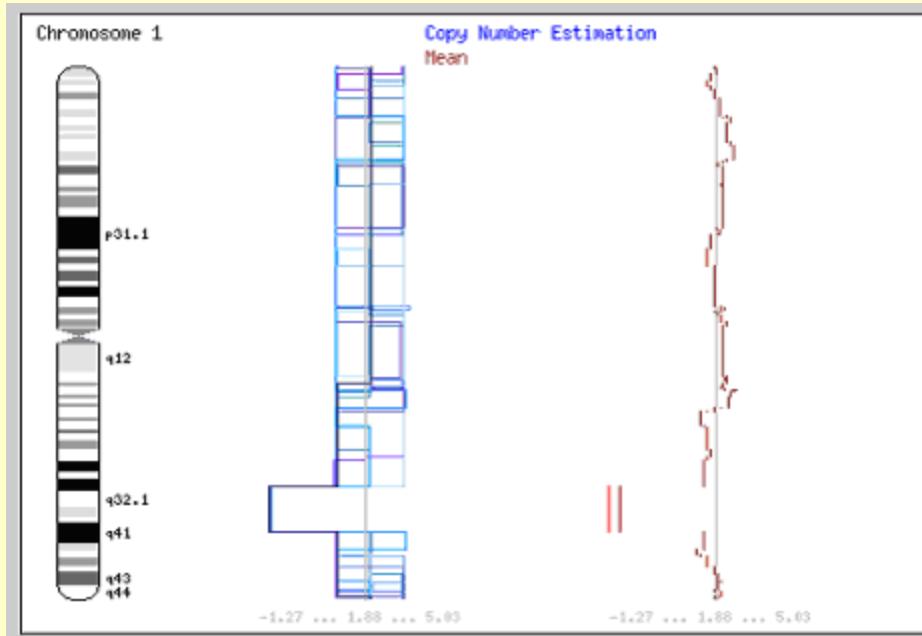
The screenshot shows the InSilicoArray CGH web application interface. The top navigation bar includes the logo "InSilicoArray CGH" and the text "Bioinformatics Department CIPF". The main interface is divided into several sections:

- SPECIE:** Select the species (Home Sapiens).
- GENOMIC DATAFILE:** Upload the genomic data (Browse...). Buttons for "Upload Example" and "See Example" are available.
- Define the copy number method:** Smoothing.
- Define the sexed autosomal and female sex chromosome baseline:** Autosomal Median.
- Define the sexed male sex chromosome baseline:** Half Autosomal Median.
- EXPRESSION DATAFILE:** Upload your expression data (Browse...). Buttons for "Upload Example" and "See Example" are available.
- POSITIONS DATAFILE:** Upload the data with the positions of your entries (optional) (Browse...). Buttons for "Upload Example" and "See Example" are available.
- PARAMETERS:** Set the scale value... (default: max, value) and Reference lines... (default: none).
- DRAWING:**
 - 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 X Y
 - All arrays / One chromosome:
Chr:
- Choose output type:** Text.
- Run** button.
- PRINCIPE FELIPE CENTRO INVESTIGACION** logo.

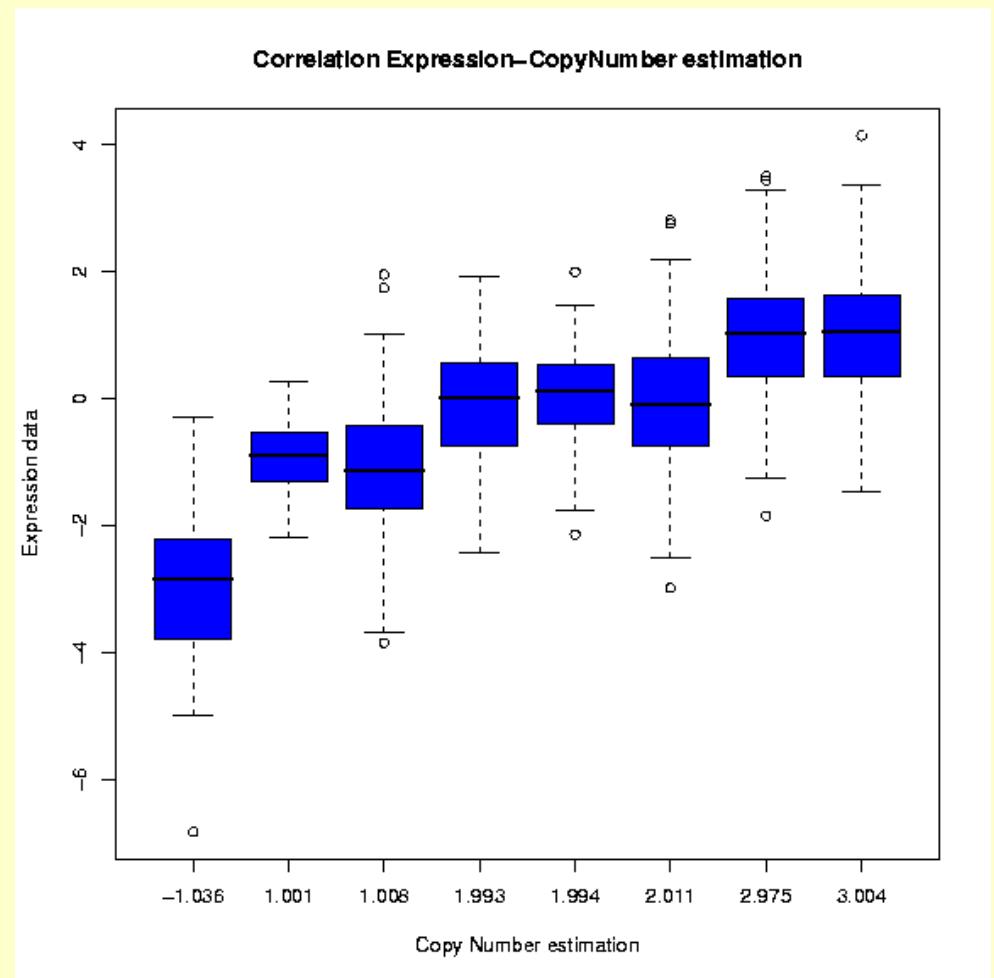
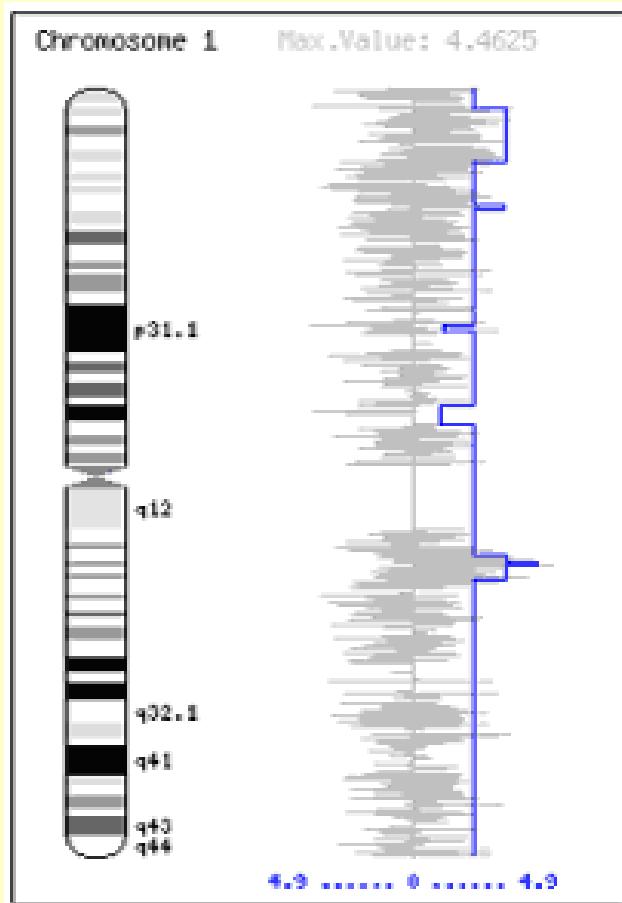
Estimating copy number



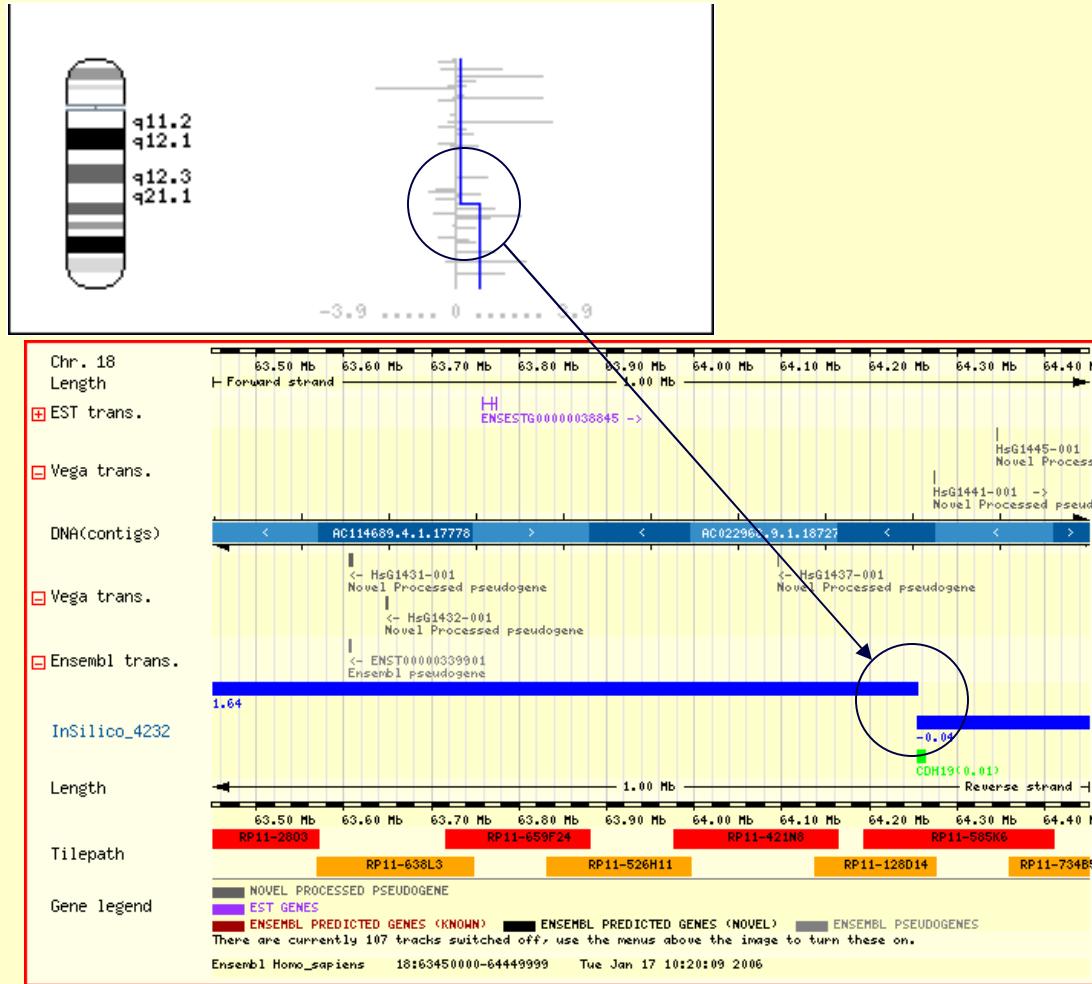
Minimum region with consistent losses or gains



Correlation copy number to expression value



Array-CGH. DAS server



Detection of copy number alterations (several methods)

Relationship expression / copy number alteration

Functional annotation of altered regions

DAS server

GEPAS & Babelomics

The screenshot shows the GEPAS web interface. At the top, there's a navigation bar with links to 'Archivo', 'Edición', 'Ver', 'Favoritos', 'Herramientas', and 'Ayuda'. Below the bar, there's a toolbar with icons for back, forward, search, and file operations. The main content area features a large heatmap with green and yellow dots. To the right of the heatmap, there's a 'Bioinformatics' logo and a 'GEPAS' logo. Below these are sections for 'GEPAS. New Release v3.0', 'What's new ?', and 'about GEPAS'. A footer at the bottom includes logos for BBVA, RIEX, and the Príncipe Felipe Centro de Investigación.

Al-Shahrour et al. Bioinformatics (2004)

Al-Shahrour et al. NAR (2005)

Al-Shahrour et al. Bioinformatics (2005)

Al-Shahrour et al. NAR (2006)

Dopazo OMICS (2006)

Al-Shahrour et al. NAR (2007)

Al-Shahrour et al. BMC Bioinformatics(2007)

Herrero et al. Bioinformatics (2001)

Herrero et al. NAR (2003)

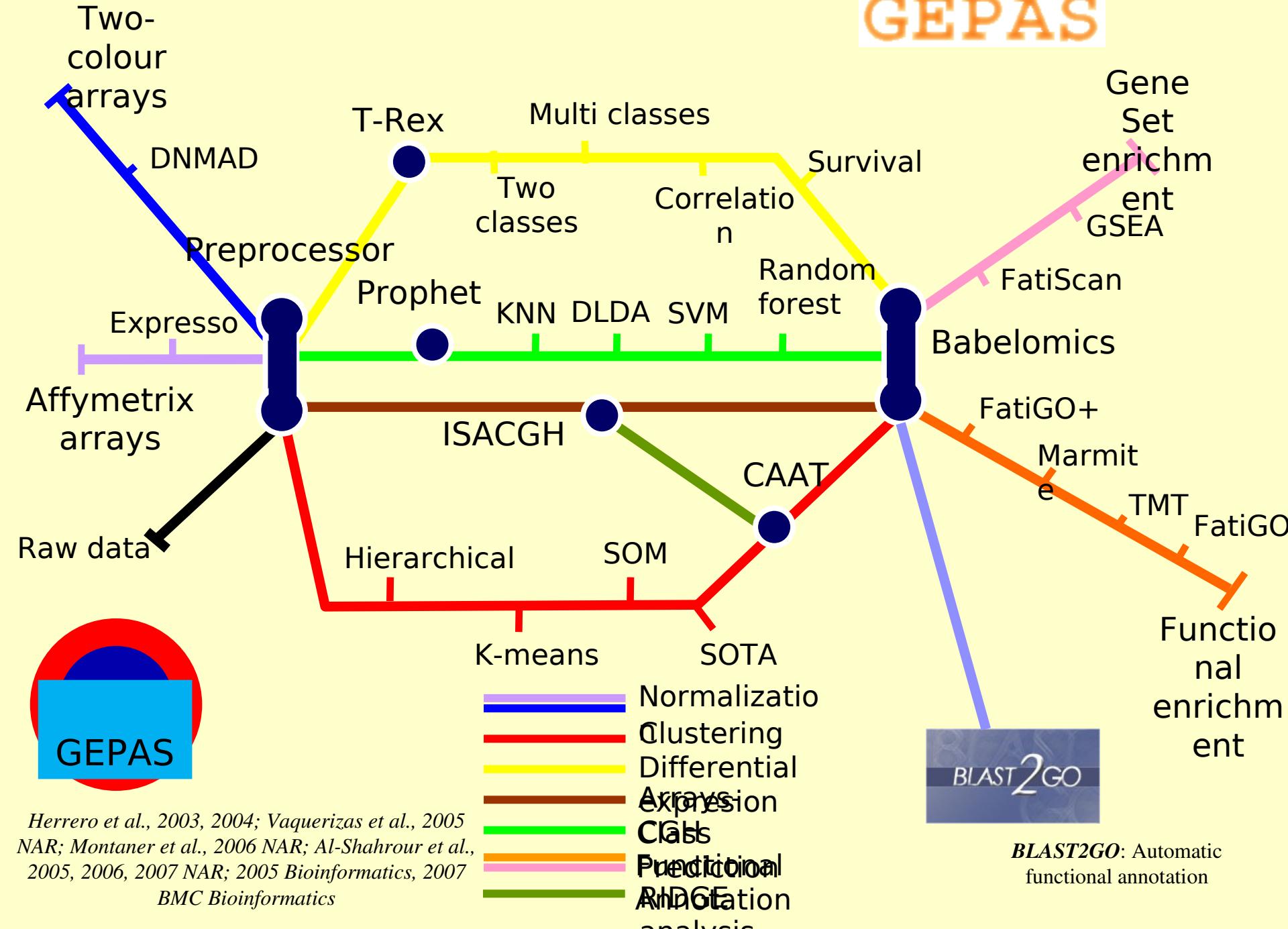
Herrero et al. NAR (2004)

Vaquerizas et al. NAR (2005)

Montaner et al. NAR (2006)

The screenshot shows the Babelomics web interface. At the top, there's a navigation bar with links to 'Archivo', 'Edición', 'Ver', 'Favoritos', 'Herramientas', and 'Ayuda'. Below the bar, there's a toolbar with icons for back, forward, search, and file operations. The main content area features a 'Bioinformatics' logo and a 'BABELOMICS' logo. Below these are sections for 'Babelomics, the systems biology way to functional annotation of genome-scale experiments' and 'Tools', 'Tutorials', 'Papers', and 'About'. A detailed description of BABELOMICS follows, mentioning its name after Borges' 'The Library of Babel', its purpose for functional annotation, and its tools like Fatigo+, Tissue Mining Tool, and MARMITE. A note at the bottom states 'Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments.'

GEPAS



Some numbers

More than 100,000 experiments analysed during 2005.

More than 500 experiments per day.



24h usage map as of June 8, 2006

New features for the next major release (February 2008)

New data types (tiling arrays, ChIP on Chip)

New normalization methods

Gene selection with many variables (classes, gender, age, etc.)

Gene selection in time series

New cluster visualization facilities

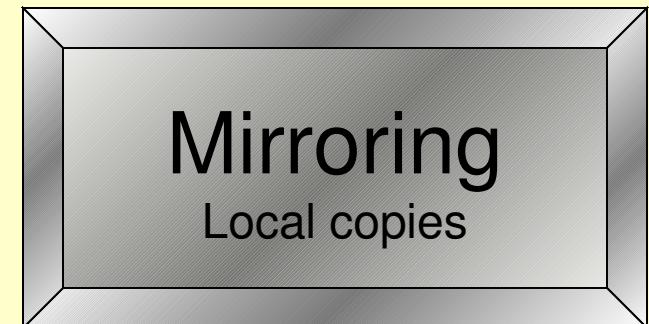
New tests for functional profiling

Integration with interactome data

Pathway visualization and analysis

New environment for visual programmation

Based on webservices



Courses

<http://bioinfo.cipf.es/docus/courses/courses.html>

Bioinformatics @ CIPF - Windows Internet Explorer

http://bioinfo.cipf.es/docus/courses/courses.html

Archivo Edición Ver Favoritos Herramientas Ayuda

VIII Jornadas de Bioinformática dotProject Bioinformatics @ CIPF Página Herramientas

PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN Bioinformatics PRINCIPE FELIPE CENTRO DE INVESTIGACIÓN

Courses

10th-14th March 2008 Course on Microarray Data Analysis Centro de Investigación Príncipe Felipe - Valencia

8th-13th October 2007 Perspectivas Bioinformáticas de la Genómica Comparativa, Funcional y Estructural Facultad de Ciencias Exactas y Naturales (FCEyN)Universidad de Buenos Aires (UBA), Argentina

1st-3rd October 2007 Course on Microarray Data Analysis Graduate School of Biological, Medical and Veterinary Sciences. University of Cambridge. UK

24th-26th September 2007 Course on Microarray Data Analysis NBN, Cape Town, South Africa

25th-27th June 2007 Practical Microarray Data Analysis Instituto Gulbenkian de Ciência (IGC), Lisbon, Portugal

4th-8th June 2007 Curso de Doctorado en Bioinformática Universidad de Alicante. Alicante

21st - 25th May 2007 Second Course on Molecular Evolution, Phylogenetics and Phylogenomics Centro de Investigación Príncipe Felipe - Valencia

26th April - 2nd May 2007 VI - BioSapiens European School in Bioinformatics Centro de Investigación Príncipe Felipe - Valencia

**CIPF, Valencia
(the week of Fallas)**

Possibilities of collaboration

Courses

- University of Cambridge (Feb. & Oct.), Valencia (March), Cape Town, Lisbon, etc.
- On demand (at your place)

Data analysis

- Consulting
- Outsourcing

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
Josete Carbonell
Ana Conesa
Hernán Dopazo
Toni Gabaldon
Francisco García
Stefan Goetz
Jaime Huerta
Marina Marcet
Marc Martí
Ignacio Medina
Pablo Minguez
David Montaner
François Serra
Joaquín Tárraga
Peio Ziarolo



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

