

Microarray data analysis

Concluding remarks

Department of Bioinformatics and Genomics, (BIG)
Centro de Investigación Príncipe Felipe (CIPF), and
Functional genomics node, (INB),
Valencia, Spain.

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

<http://www.babelomics.org>

<http://bioinfo.cipf.es>



INB INSTITUTO NACIONAL



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Structure of the course

Theoretical

Hands-on **GEPAS**

Introduction

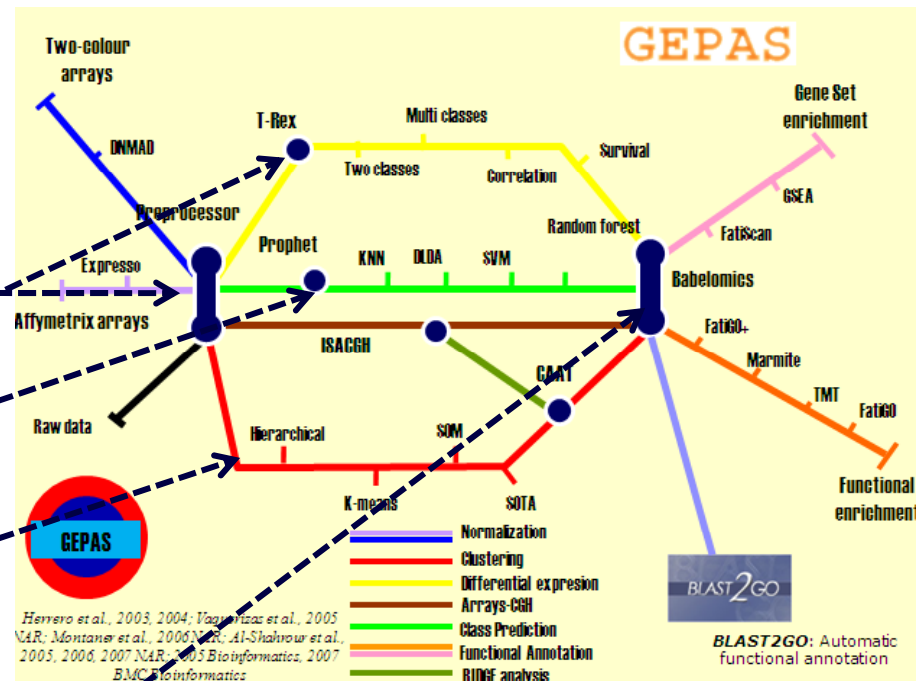
Normalization

Gene selection

Predictors

Clustering

Functional interpretation



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

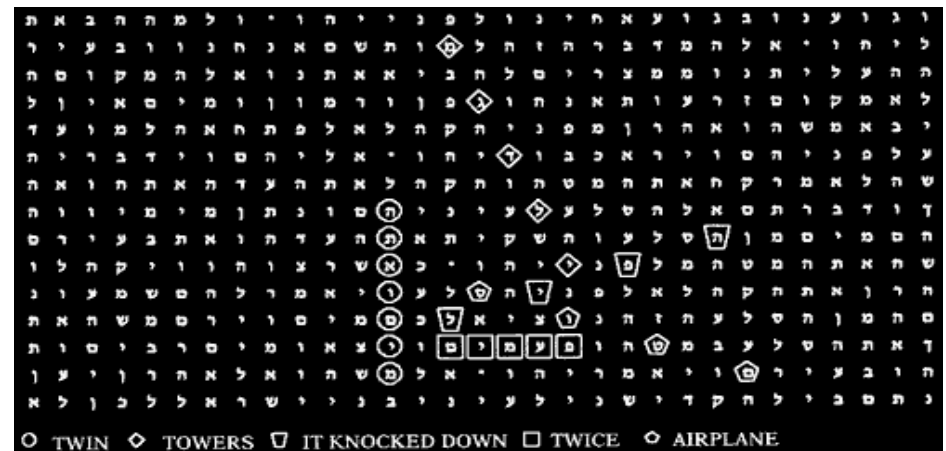
Is gene A involved in process B?

Experiment \longrightarrow (sometimes) test \longrightarrow Hypothesis

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

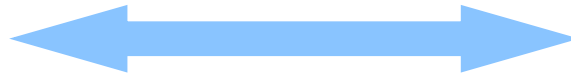
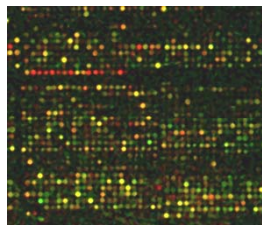
Sure, but... Is it real? (many hypotheses are rejected while this one is accepted *a posteriori*: numerology)

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



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**)

Studies must be hypothesis driven.

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

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

Different classes...

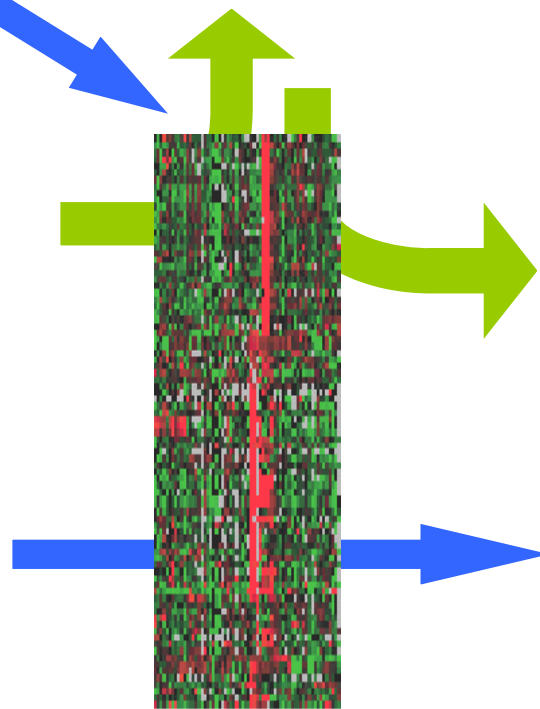
Unsupervised
Supervised

Molecular classification of samples

What genes are responsible for?

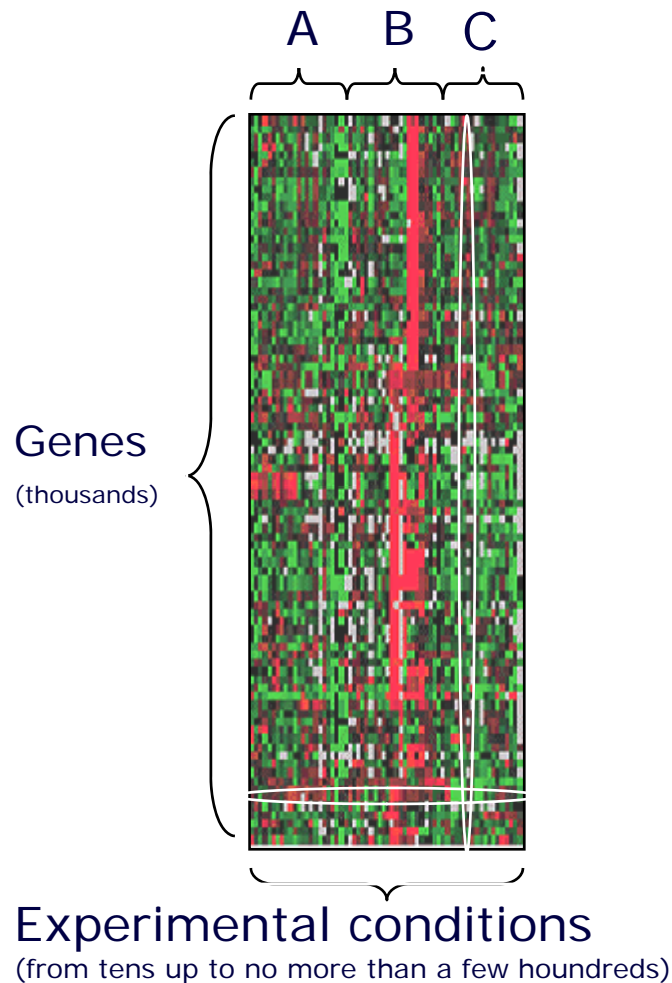
Co-expressing genes...

What do they have in common?



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?

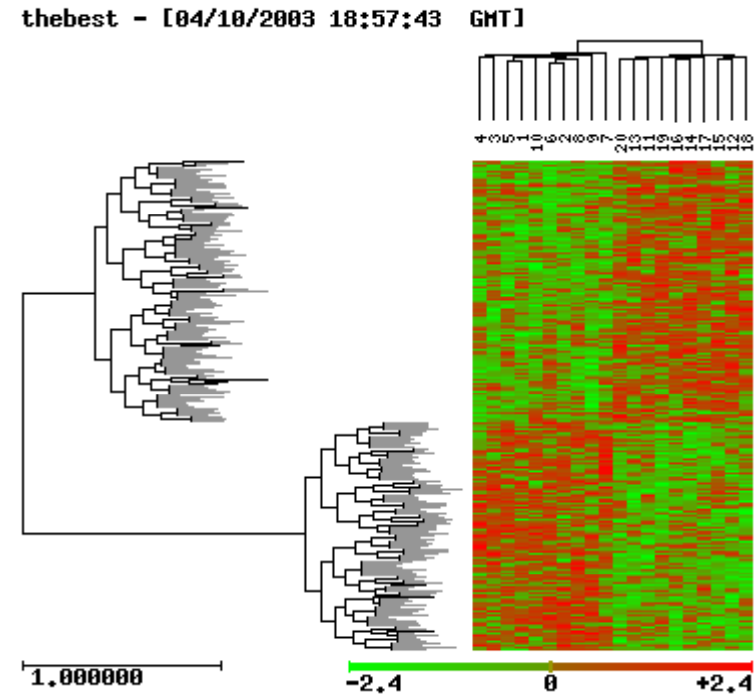
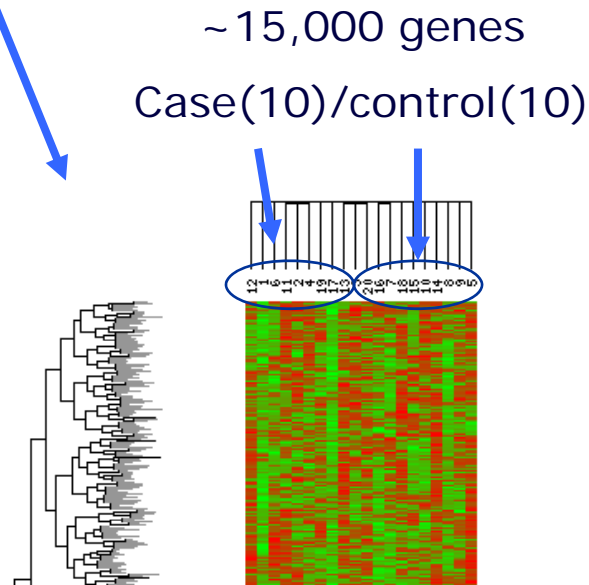
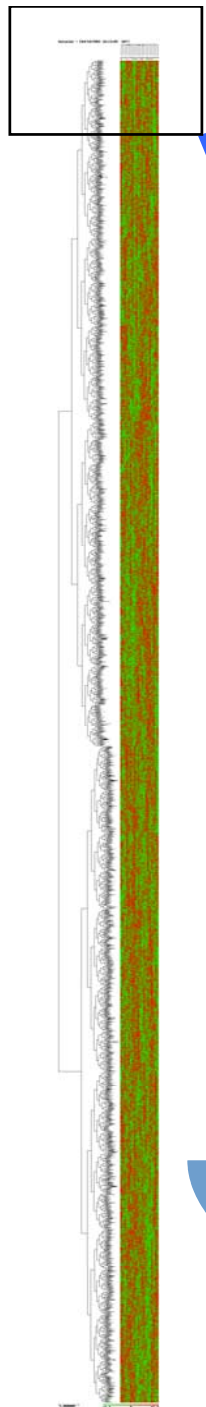
And

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

Class prediction

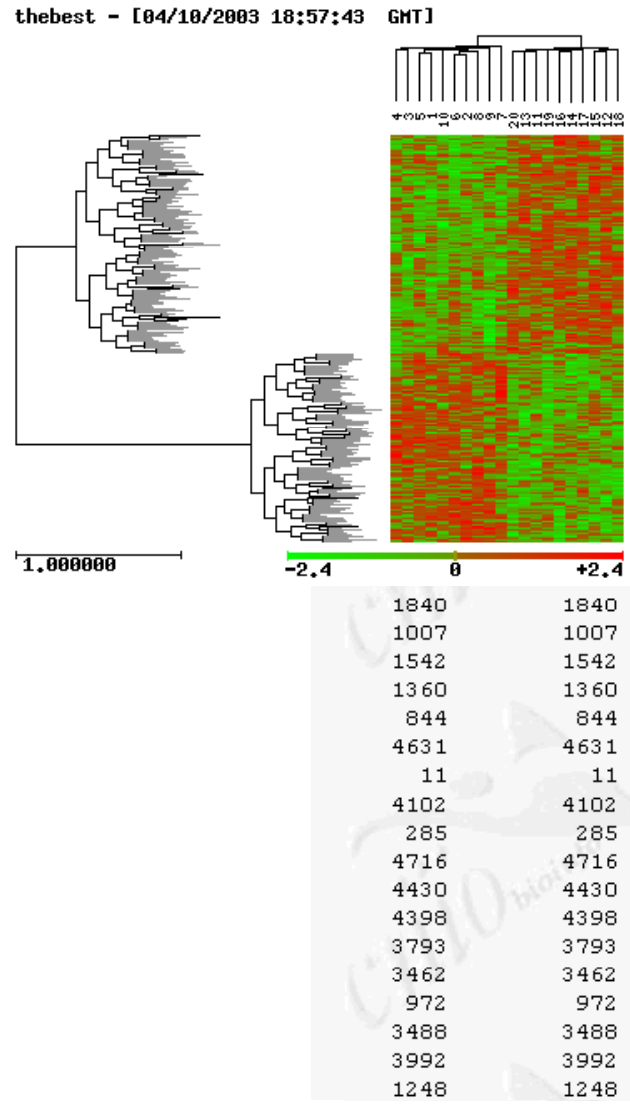
Gene selection

A simple problem: gene selection for class discrimination



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

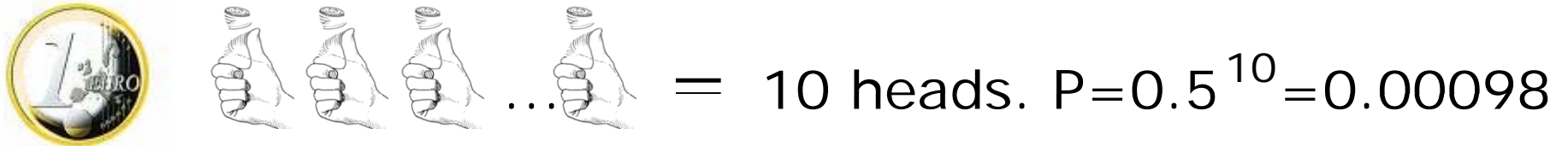


So... Why do we find good p-values?

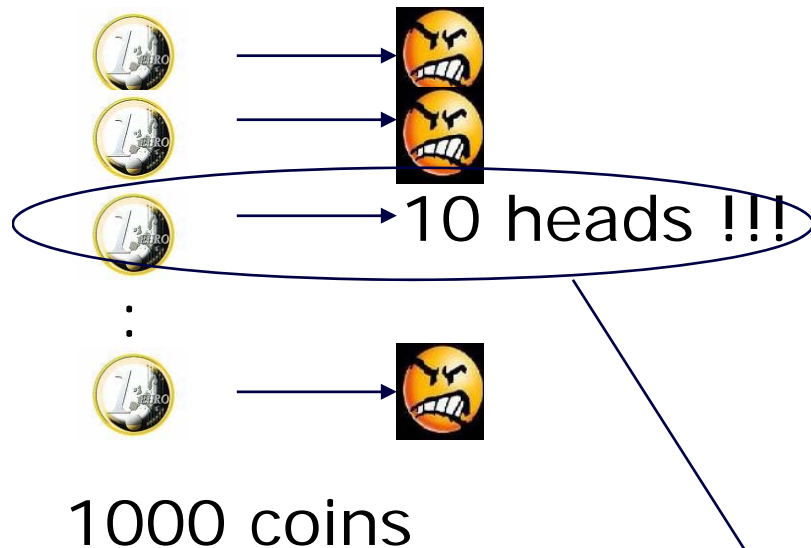
unadj_p	adj_p	FDR_indep	FDR_dep	obs_stat
0.00019998	0.152685	0.49995	1	5.47044
0.00019998	0.746225	0.49995	1	4.49902
0.0009999	0.983002	0.861025	1	4.01726
0.00149985	0.986401	0.861025	1	3.99374
0.00129987	0.9959	0.861025	1	3.86046
0.00169983	0.9996	0.861025	1	3.7251
0.00169983	0.9996	0.861025	1	3.66628
0.00169983	0.9996	0.861025	1	3.62427
0.00169983	0.9996	0.861025	1	3.60596
0.00169983	0.9996	0.861025	1	3.58109
0.00169983	0.9996	0.861025	1	3.52935
0.00169983	0.9996	0.861025	1	3.43721
0.00169983	0.9996	0.861025	1	3.41937
0.00169983	0.9996	0.861025	1	3.41428
0.00169983	0.9996	0.861025	1	3.4025
0.00169983	0.9996	0.861025	1	3.40212
0.00169983	0.9996	0.861025	1	3.37412
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



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



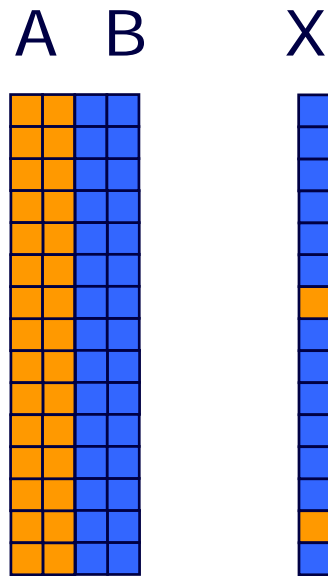
$$P = 1 - (1 - 0.5^{10})^{1000} = 0.62$$

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

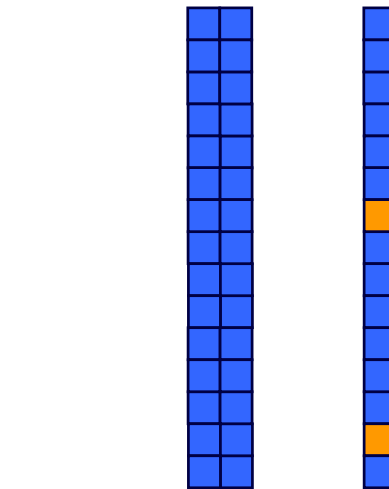
Will you still use this coin for betting?

Of predictors and molecular signatures

What is a predictor?

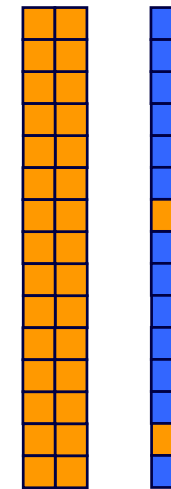


Is X, A
or B?



Diff (B, X) = 2

Intuitive notion:



Diff (A, X) = 13

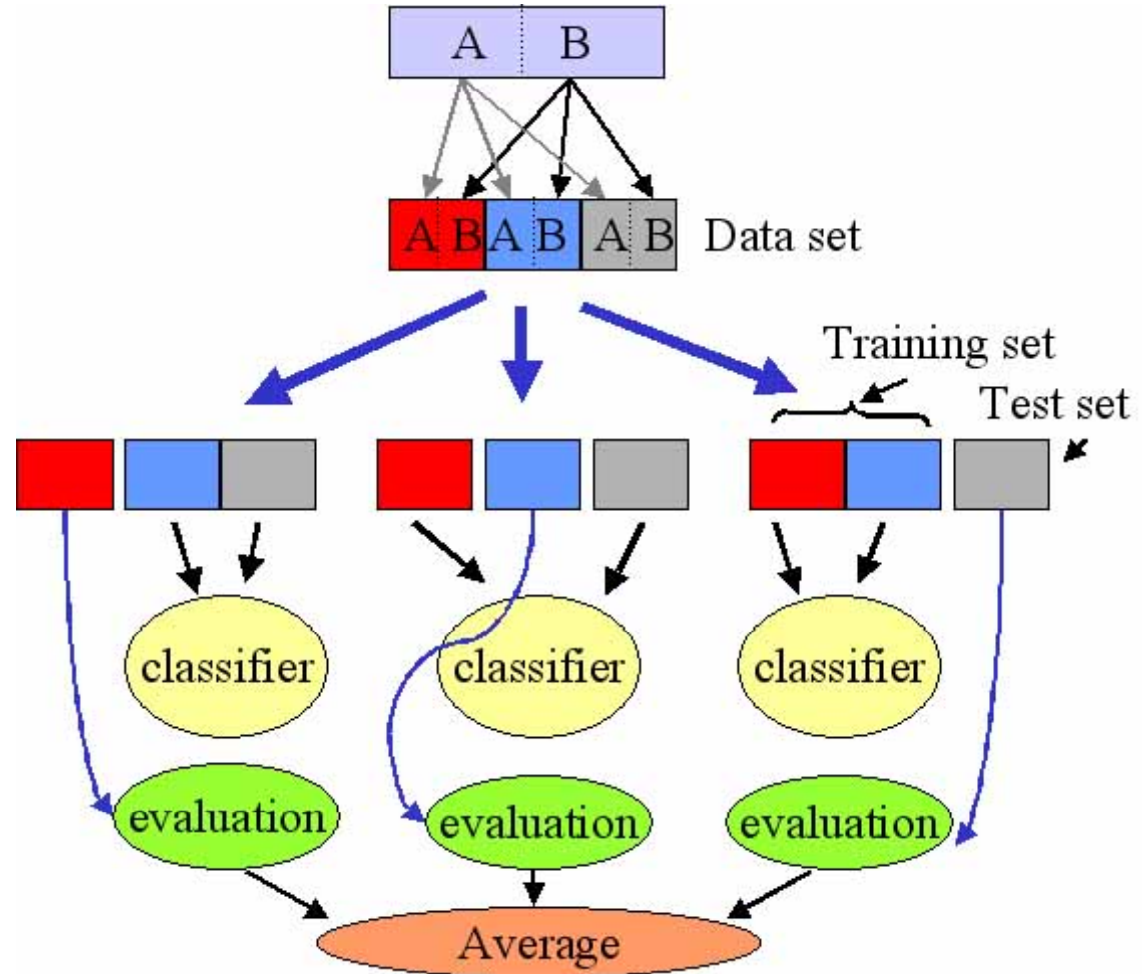
Most probably X belongs to class B

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



Studies must be hypothesis driven.

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

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

Different classes...

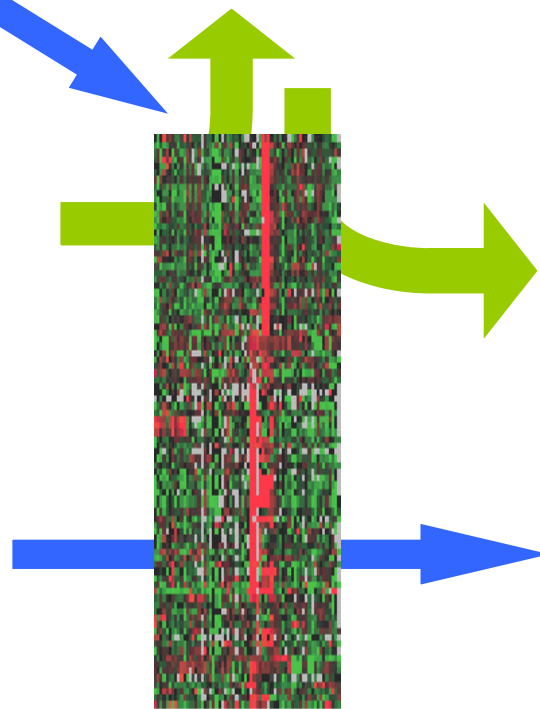
Unsupervised
Supervised

Molecular classification of samples

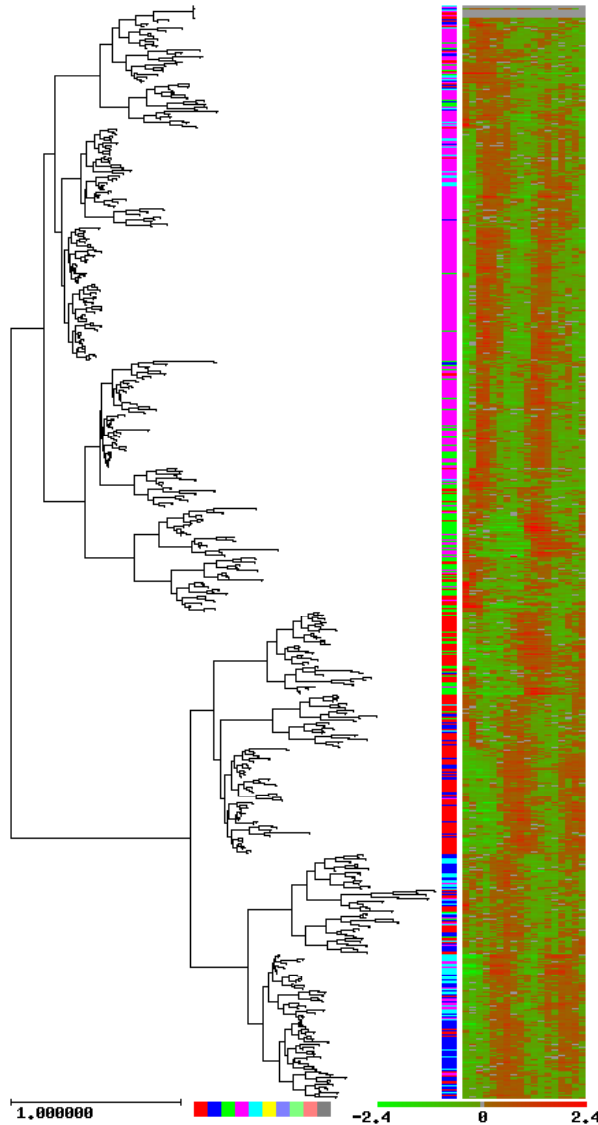
What genes are responsible for?

Co-expressing genes...

What do they have in common?



An unsupervised problem: clustering of genes.



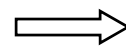
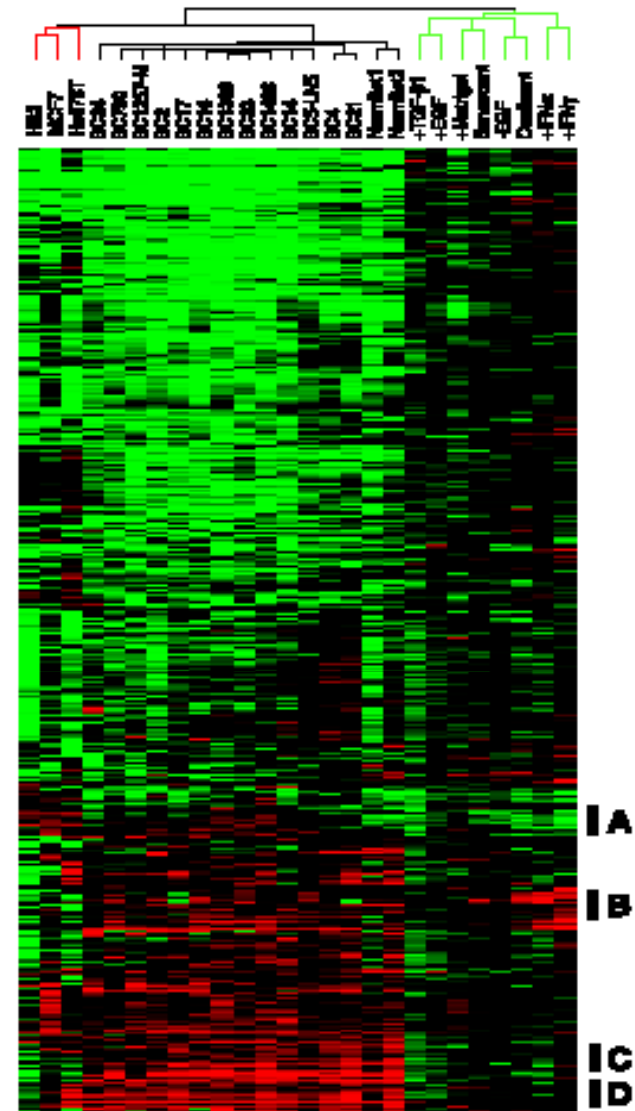
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

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)

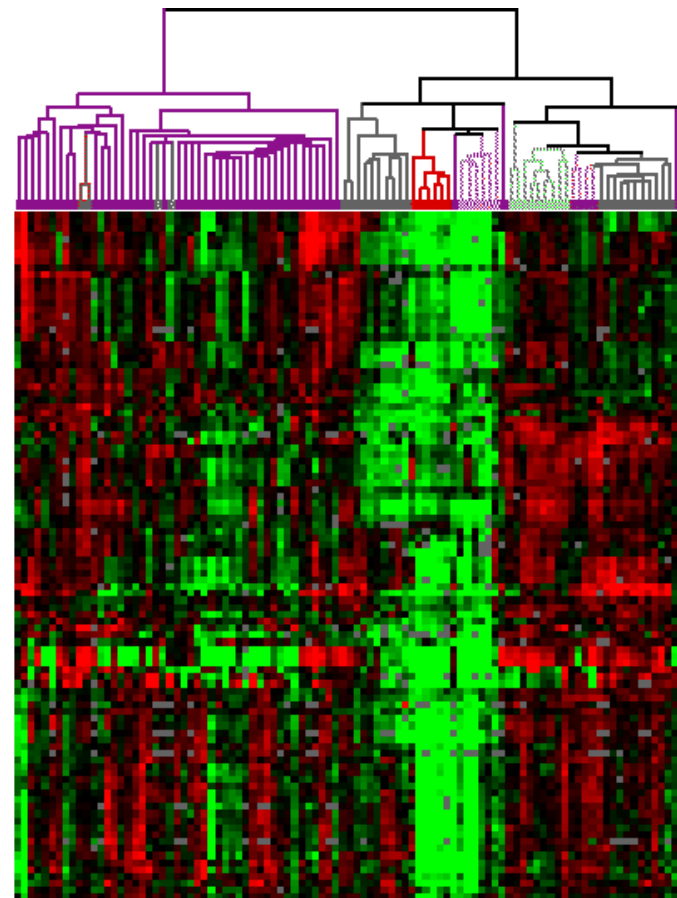
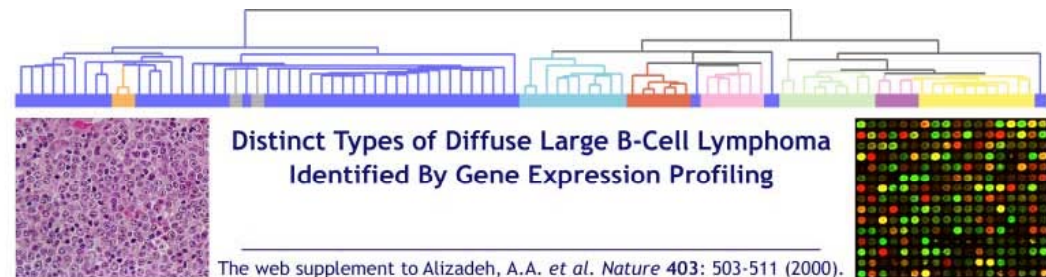
Clustering of experiments: The problems

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



Functional interpretation of genome-scale experiments in the post genomics era

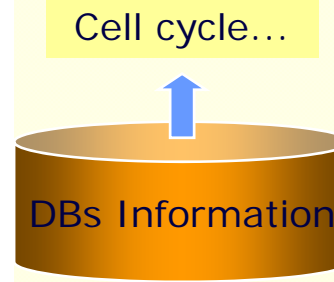
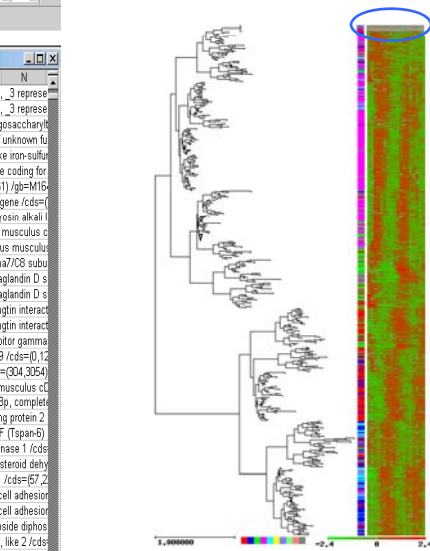
My data...

How are structured?

What are these groups?

What is this gen?

	E	F	G	H	I	J	K	L	M	N
65	578.6*		1.4	0.26	M12461	Mouse cytoplasmic beta-actin mRNA (L5_M_3 repress				
66	534.9*		-1.6	0.22	M12461	Mouse cytoplasmic beta-actin mRNA (L5_M_3 repress				
67	403.6*		-1.5	0.15	X61388	SGD: YEL002C: Yeast S. cerevisiae WBP1 Oligosaccharyl				
68	535.2*		-1.6	0.22	U10530	SGD: YEL016W: Yeast S. cerevisiae Protein of unknown fu				
69	-567.7*		-1.6	-0.27	M23316	SGD: YEL024C: Yeast S. cerevisiae RPI1 Rieske iron-sulfur				
70	-114.5*		-1.1	-0.03	K02207	SGD: YEL021W: Yeast S. cerevisiae URA3 gene coding for				
71	-125.4*		-1.1	-0.01	Cluster Incl M16465	Calpactin I light chain (cds=69,361) (gib=M16				
72	-1091.6		-1.2	-0.14	Cluster Incl Z87746	M. musculus spermidine synthase gene (cds=1				
73	-727.2		-1.3	-0.17	Cluster Incl X12573	M. musculus MLC17/MLC17P gene for myosin alkali				
74	9036.6		1.3	0.83	Cluster Incl A84935	U1-M-AH1-agw-a-06-B-U1 s1 Mus musculus c				
75	-847.4		-1.3	-0.21	Cluster Incl AW129542	U1-M-BH2-1-aph-f01-Q-U1 s1 Mus musculus				
76	2983.1		1.1	0.09	Cluster Incl AF058993	Mus musculus proteasome alpha7/O3 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 U97170	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 I, ATP-dependent (cds=304,3054)				
84	0*		-0.0	0	Cluster Incl AB51492	U1-M-BH0-ajj-d-04-Q-U1 s2 Mus musculus c				
85	3310.9		1.2	0.24	Cluster Incl AB025408	Mus musculus mRNA for sid47bp, complet				
86	-1291		-1.5	-0.42	Cluster Incl AF059735	Mus musculus C-terminal binding protein 2				
87	-263.3*		-1.3	-0.09	Cluster Incl AF059735	Mus musculus tetraspan TM4SF (Tspan-6)				
88	77.5*		1.1	0.01	Cluster Incl D45950	Hydroxysteroid 17-beta-dehydrogenase 1 (cds				
89	2047.2*		-3.3	1.1	Cluster Incl AF038299	Mus musculus 17-beta-hydroxysteroid dehy				
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 adhesio				
92	-675.5*		-1.8	-0.37	Cluster Incl U12884	Mus musculus C57BL/6 vascular cell adhesio				
93	1465.4*		-2.7	0.76	Cluster Incl A238636	Mus musculus mRNA for nucleoside diphos				
94	836.2		1.1	0.1	Cluster Incl U70475	Nuclear, factor, erythroid derived 2, like 2 (cds				
95	4969.4*		-6.7	8.84	Cluster Incl AF045673	Mus musculus FLALRR associated protein-				
96	148.3*		-1.2	0.04	Cluster Incl AB81475	u59a06.x1 Mus musculus cDNA, 3' end (cds=				



- I M19380 Calmodulin 3 (cds=109,558) (gib=M19380) (gib=469419)
- I AB42328 U1-M-AH1-afz-b-11-Q-U1 s1 Mus musculus cDNA, 3'
- I A242693 Mus musculus mRNA for cathepsin Z precursor (cds
- I U12620 Dipeptidylpeptidase 4 (cds=117,2999) (gib=U12620) (g
- I M13444 Mouse alpha-tubulin isotype M-alpha-4 mRNA, comp
- I U11027 Mus musculus C57BL/6J Sec63 protein complex gam
- I X33529 Phosphothio kinase, liver, D type (cds=162,2304) (gib=
- I 257745 M. musculus mRNA for phosphothio kinase 2A catalytic subu
- I U80932 Serine/threonine kinase 6 (cds=48,1235) (gib=U80932)
- I U47024 Maternal embryonic message 3 (cds=137,2401) (gib=
- I AF075136 Mus musculus Sin3-associated protein (vas30) mR
- I M25944 Mouse carbonic anhydrase II (CAII) mRNA, 3' end (cd
- I X74871 Neurofibromatosis 2 (cds=576,2366) (gib=X74871) (gib=
- I M12048 Mouse myb proto-oncogene mRNA encoding 71 kd m
- I A41125498 U1-M-BH2-3-agg-a07-Q-U1 s1 Mus musculus cDNA
- I U84903 Ribosomal protein L23 (cds=61,501) (gib=U84903) (gib=
- I U35141 Mus musculus retinoblastoma-binding protein (mRb) p
- I U19621 Mus musculus vesicle transport protein (musv-18c) m
- I M15265 Adenosine nucleic acid synthase 2, erythroid (cds=0,178)
- I M25149 Tissue specific transplantation antigen P91A (cds=0,
- I X68449 Calcyclin (cds=159,428) (gib=X68449) (gib=50271) (gib=



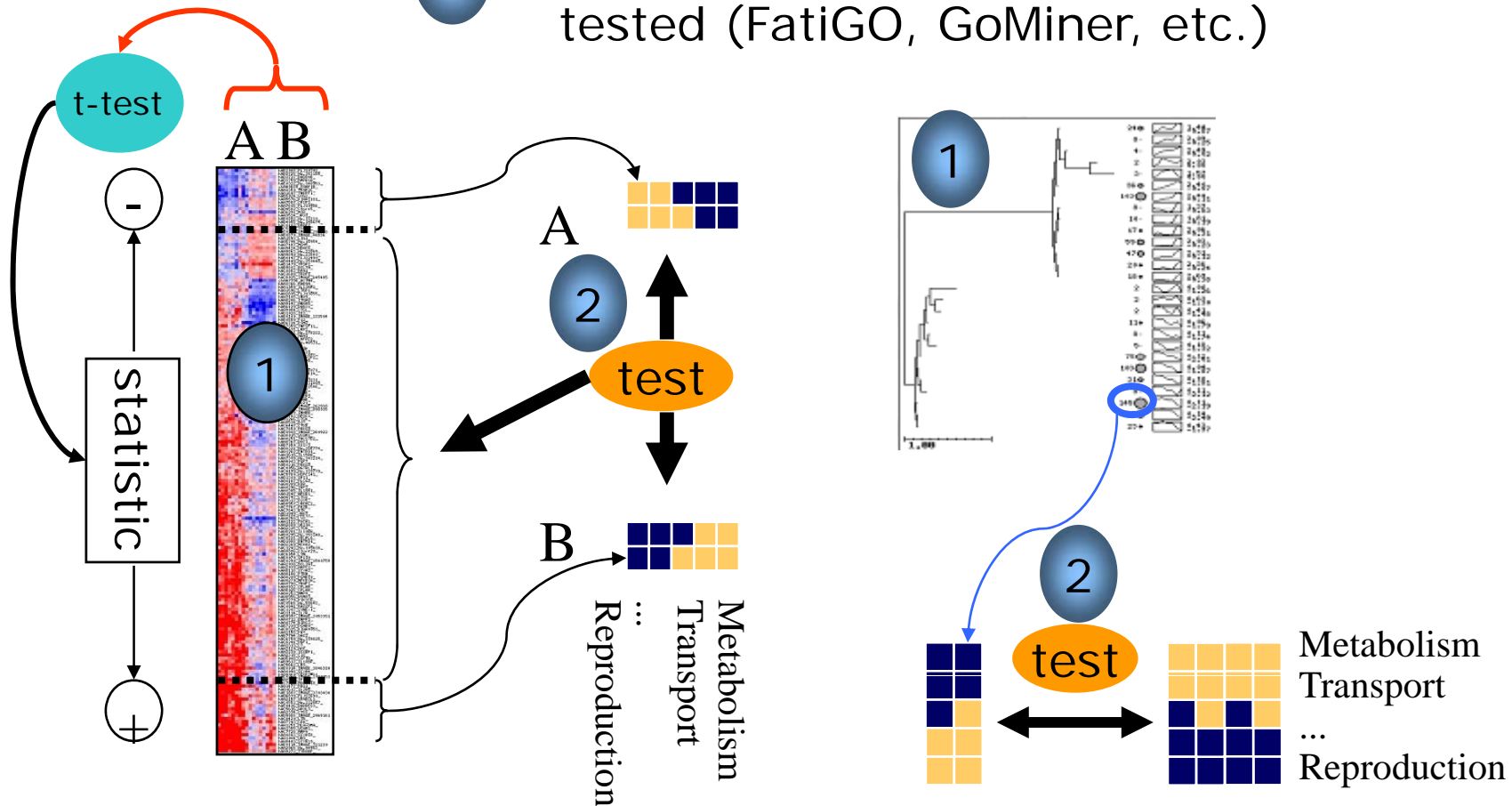
Analysis

Information mining

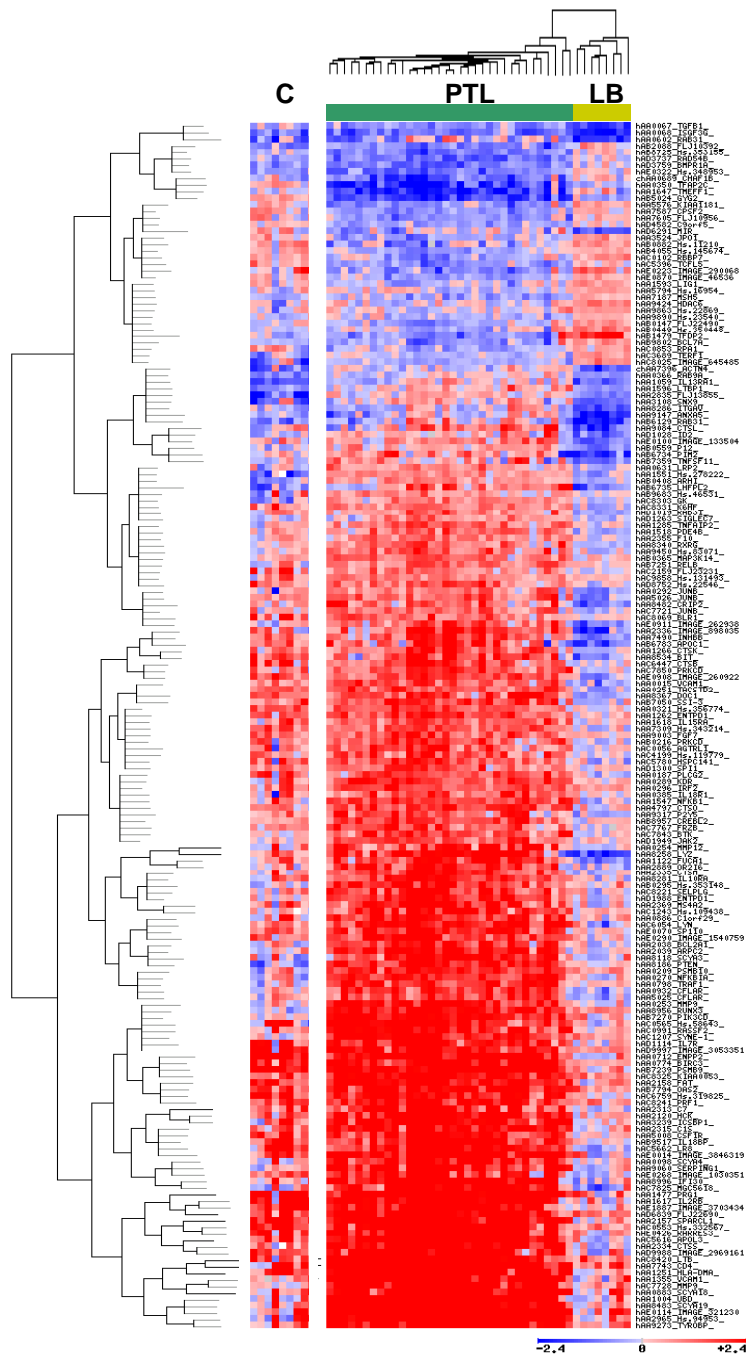
Links

Two-steps functional interpretation

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



Understanding why genes differ in their expression between two different conditions



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

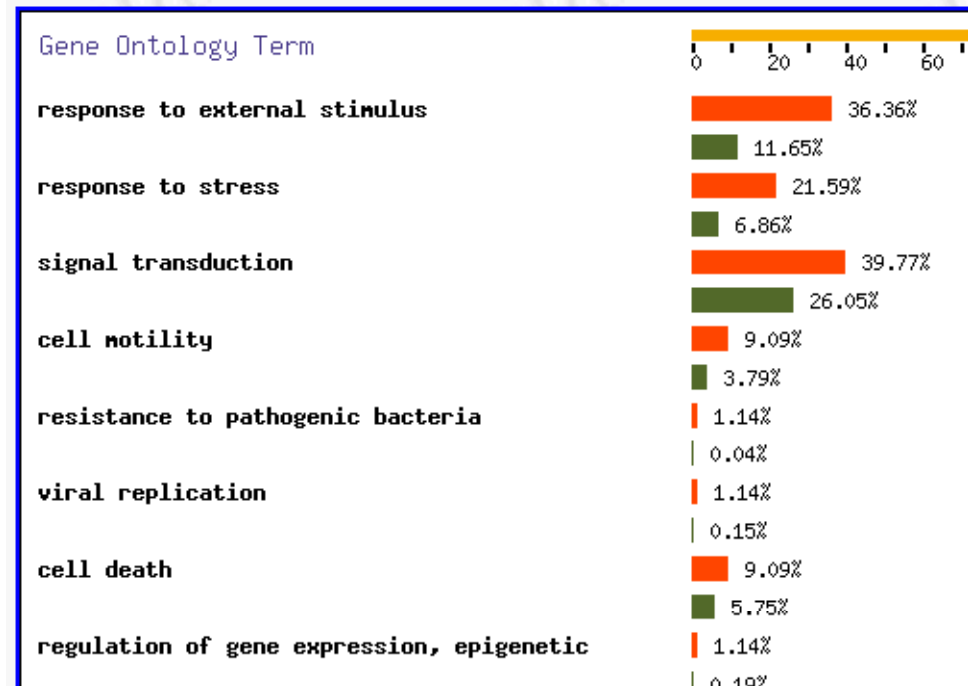
Genes differentially expressed, selected among the ~7000 genes in the CNIO oncochip

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

	Cluster Query	Cluster Reference
Total number of initial genes:	162	4764
Total number of genes no repeated:	129	4731
Total number of Cluster IDs retired - their currents Cluster IDs	7 - 23	449 - 1627
Total number of genes no repeated with current Cluster IDs:	145	5909
Total number of genes no repeated with GO at level 3 and biological_process:	88	2610
Total number of genes no repeated with GO but NOT at level 3 and ontology:		
Total number of genes no repeated without GO annotated:		

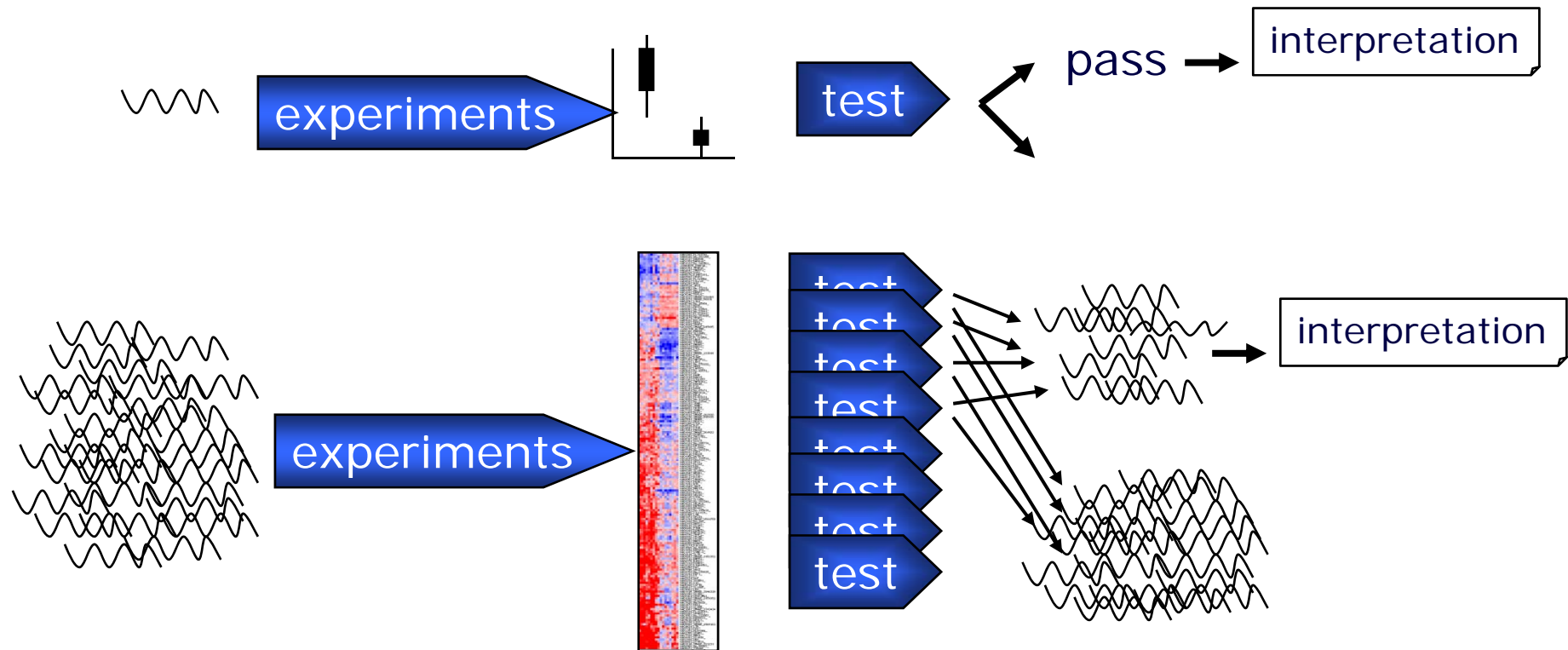


Obvious? NO

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

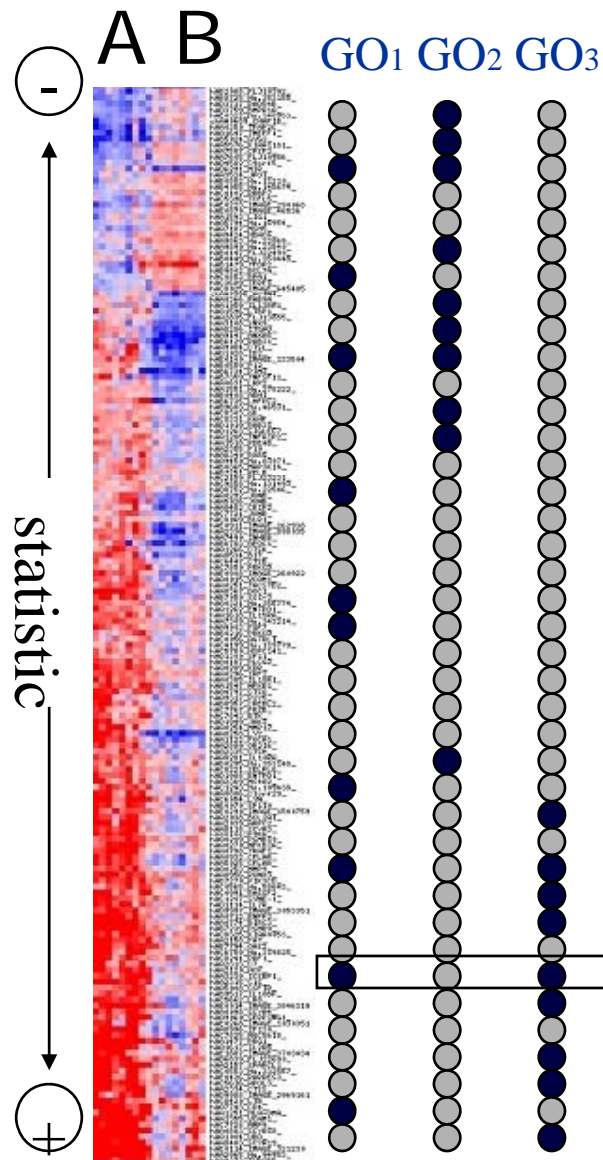
0.1702	0.9912	1	1
0.1806	0.9940	1	1

Two-steps approach reproduces pre-genomics paradigms



Context and cooperation between genes is ignored

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



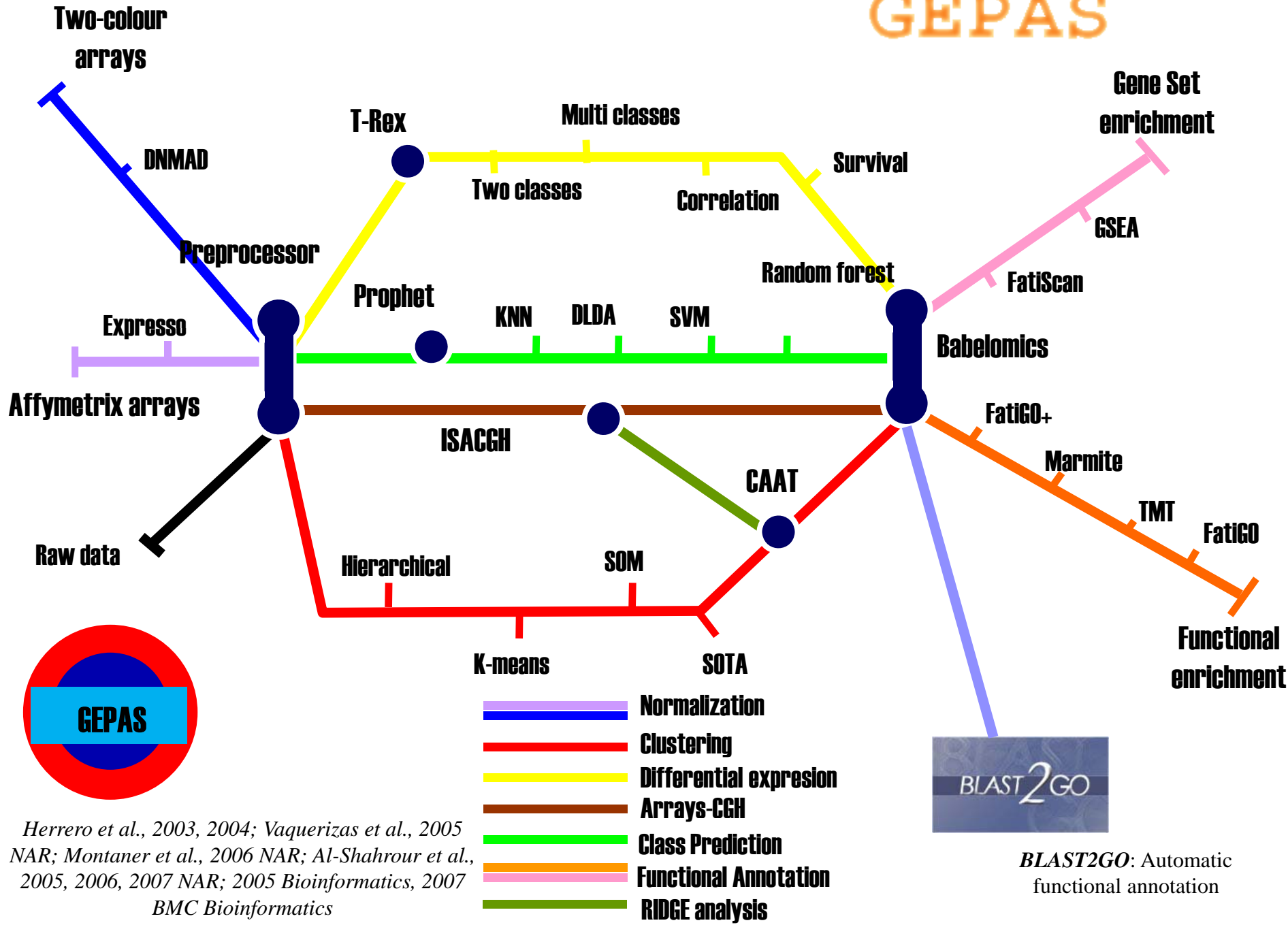
Ranking: A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

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

GEPAS



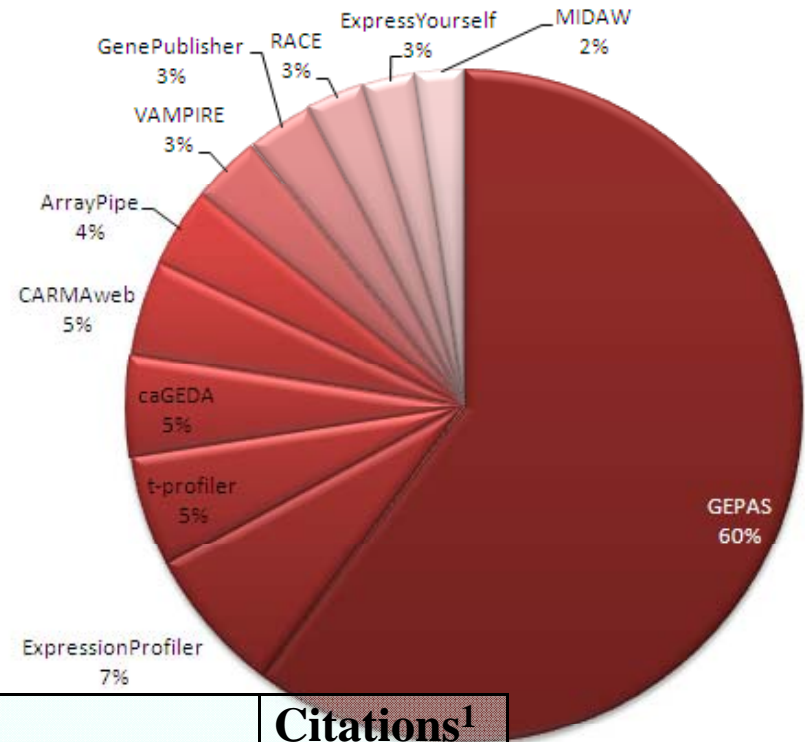
Herrero et al., 2003, 2004; Vaquerizas et al., 2005 NAR; Montaner et al., 2006 NAR; Al-Shahrour et al., 2005, 2006, 2007 NAR; 2005 Bioinformatics, 2007 BMC Bioinformatics

Successful reception by the scientific community

GEPAS: currently is the most cited web-based platform for transcriptomic analysis (646 scholar google citations)

Babelomics. Third most cited platform (903 scholar google citations; FatiGO is amongst the 50 most cited papers in Bioinformatics)

Microarray data analysis webtools with at least 10 citations¹.



Web tool	URL	Citations ¹
GEPAS	http://www.gepas.org	646
ExpressionProfiler	http://www.ebi.ac.uk/expressionprofiler	76
t-profiler	http://www.t-profiler.org	58
caGEDA	http://bioinformatics.upmc.edu/GEDA.html	52
CARMAweb	https://carmaweb.genome.tugraz.at	49
ArrayPipe	http://www.pathogenomics.ca/arraypipe	41
VAMPIRE	http://genome.ucsd.edu/microarray/	36
GenePublisher	http://www.cbs.dtu.dk/services/GenePublisher	34
RACE	http://race.unil.ch/	30
Express Yourself	http://bioinfo.mbb.yale.edu/expressyourself	27
MIDAW	http://muscle.cribi.unipd.it/midaw/	25

Approximately
1000 users per day
1500 registered users (6 months)

Publications
2008 – 6
2007 – 6
2006 – 5
2005 – 5

1) Scholar Google citations over all the references of the tool (September 2009). See <http://bioinfo.cipf.es/docus/tools-citations/microarrays/>

Web tools for functional profiling

Web tools with 10 or more Scholar Google citations

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

Data analysis at the department of bioinformatics, CIPF

- Outsourcing and consulting for conventional or routine analysis tasks
- Collaborations on mutual interest basis
- Soon GEPAS/Balelomics mirror available
- Future: data storage (with backup) and VIP access to the tools

The Bioinformatics and Genomics Department at the Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain, and...

Joaquín Dopazo
Eva Alloza
Leonardo Arbiza
Fátima Al-Shahrour
Davide Bau
Emidio Capriotti
Jose Carbonell
Ana Conesa
Adriana Cucchi
Hernán Dopazo
Pablo Escobar
Francisco García
Stefan Goetz
Martina Marbà
Marc Martí
Ignacio Medina
Pablo Minguez
David Montaner
Marina Naval
Luis Pulido
Javier Santoyo
Patricia Sebastian
François Serra
Sonia Tarazona
Joaquín Tárraga



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

