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

INB
INSTITUTO NACIONAL DE BIOINFORMÁTICA

National Institute of Bioinformatics,
Functional Genomics node

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Background

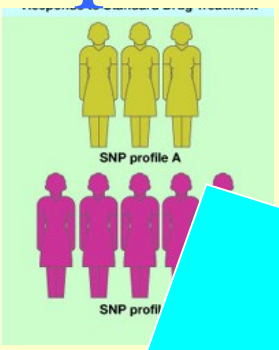
Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.
Sydney Brenner, 1980

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...

…produces the final phenotype

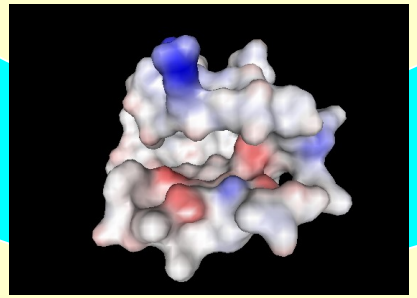
>protein kunase

acctgttgatggcgacagggactgtatgctgatc
tatgctgatgcatgcatgctgactactgatgtggg
ggctattgacttgatgtctatc....

From genotype to phenotype.

…whose structure accounts for function...

…plus the environment...



From genotype to phenotype

(in the functional post-genomics scenario)

Now: 22240 (NCBI build 35 12/04)
50-70% display alternative splicing
25%-60% unknown
Transfrags

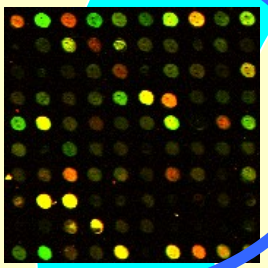
Genes in the DNA...

>protein kunase

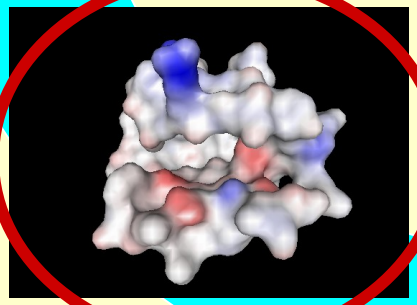
acctgttgatggcgacagggactgtatgctgatc
tatgctgatgcatgcatgctgactactgatgtggg
ggctattgacttgatgtctatc....

…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...

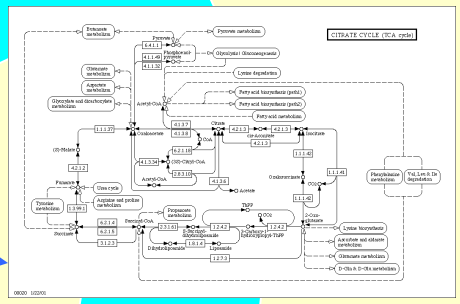
100K-500K proteins

…whose structures account for function...

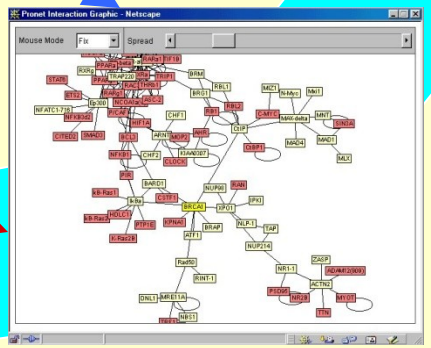
…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype...



…conforming complex interaction networks...

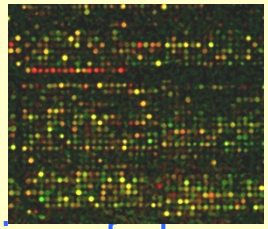


…in cooperation with other proteins…

Each protein has an average of 8 interactions

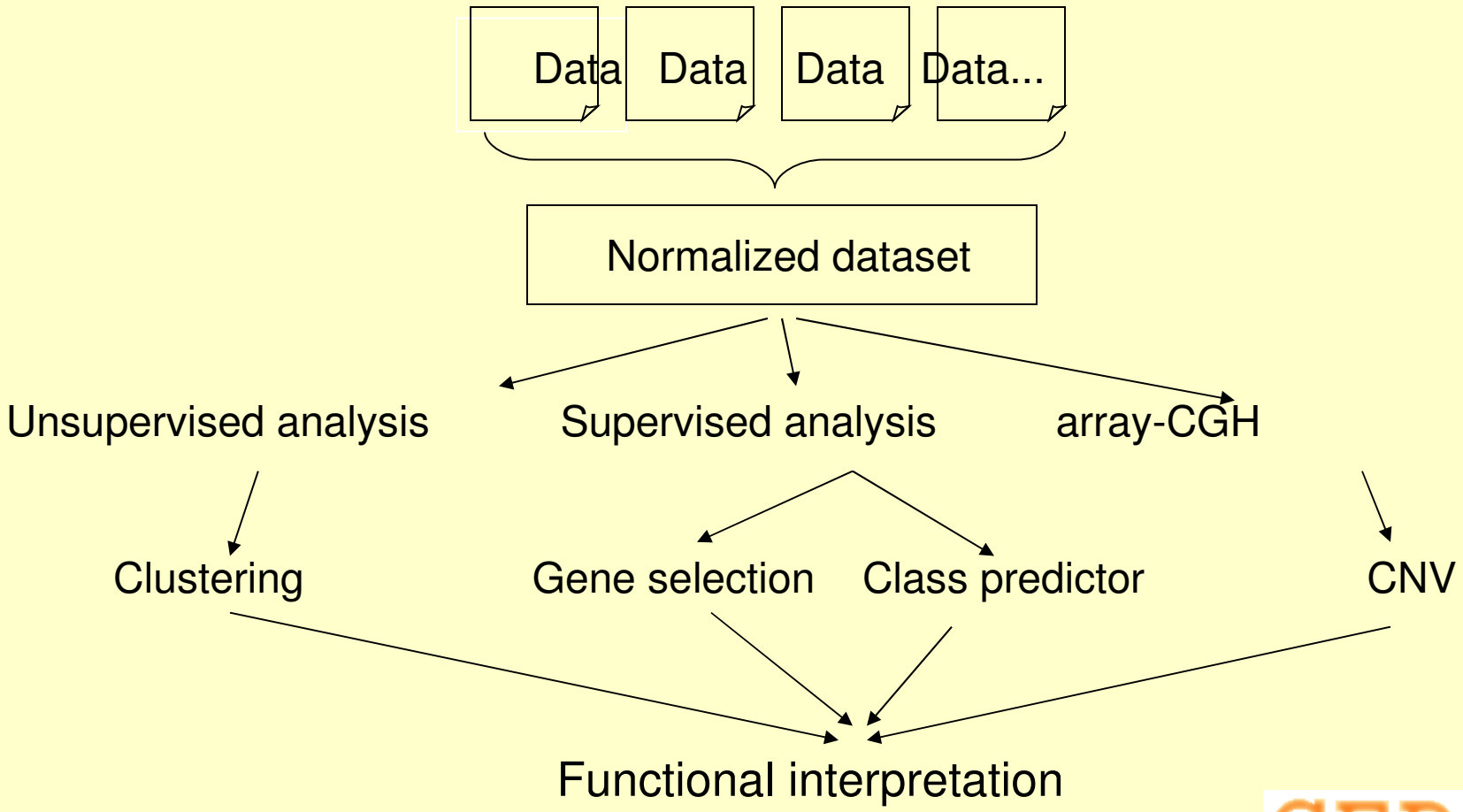
Gene expression profiling. Historic perspective

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.)

GEPAS

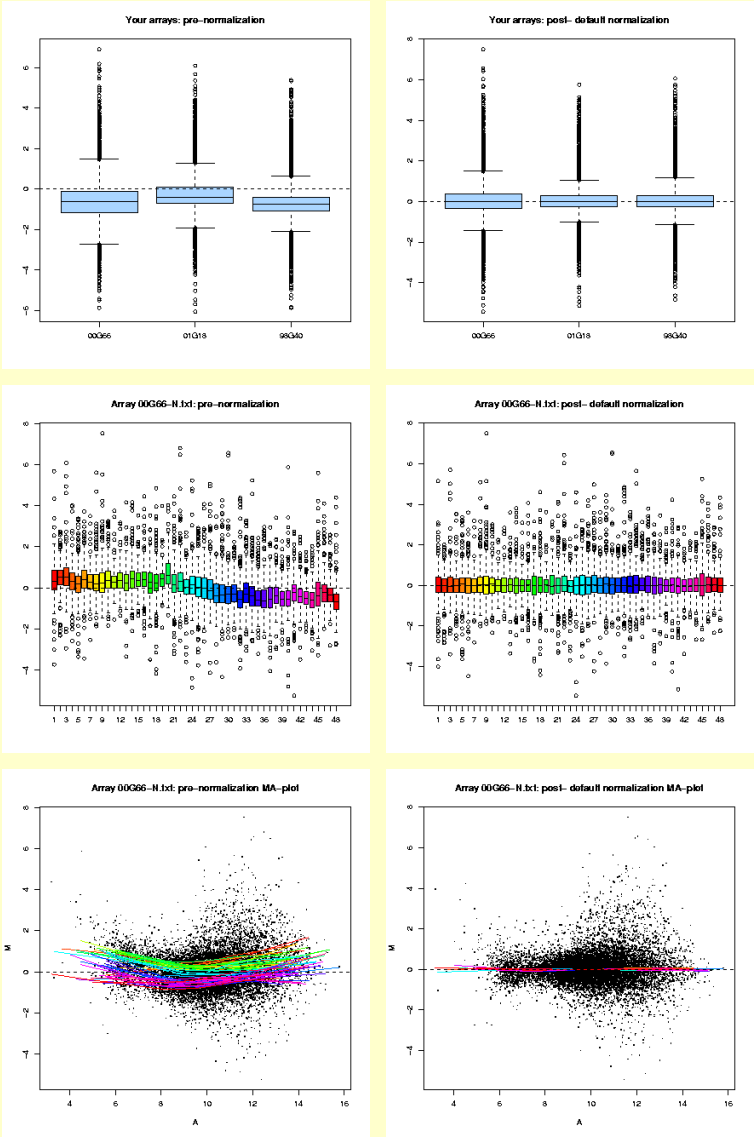
BABELOMICS

Normalisation

A

B

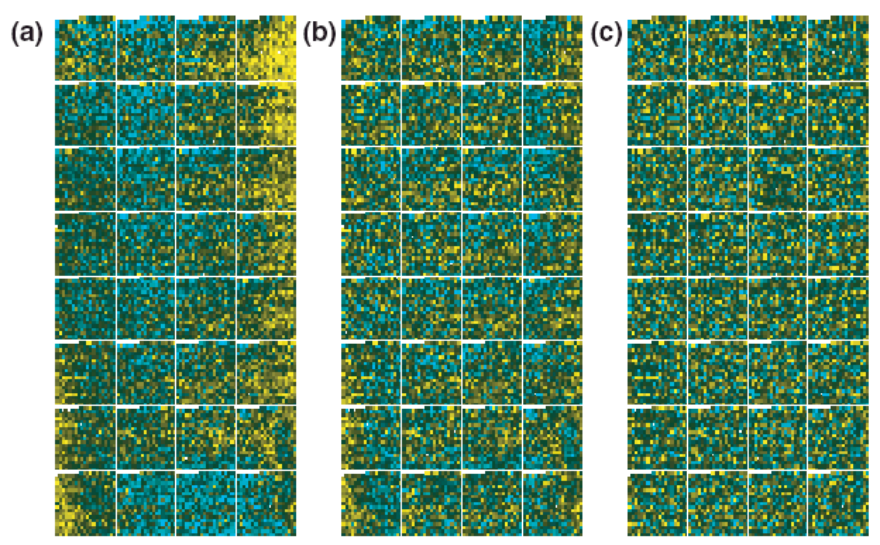
C



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

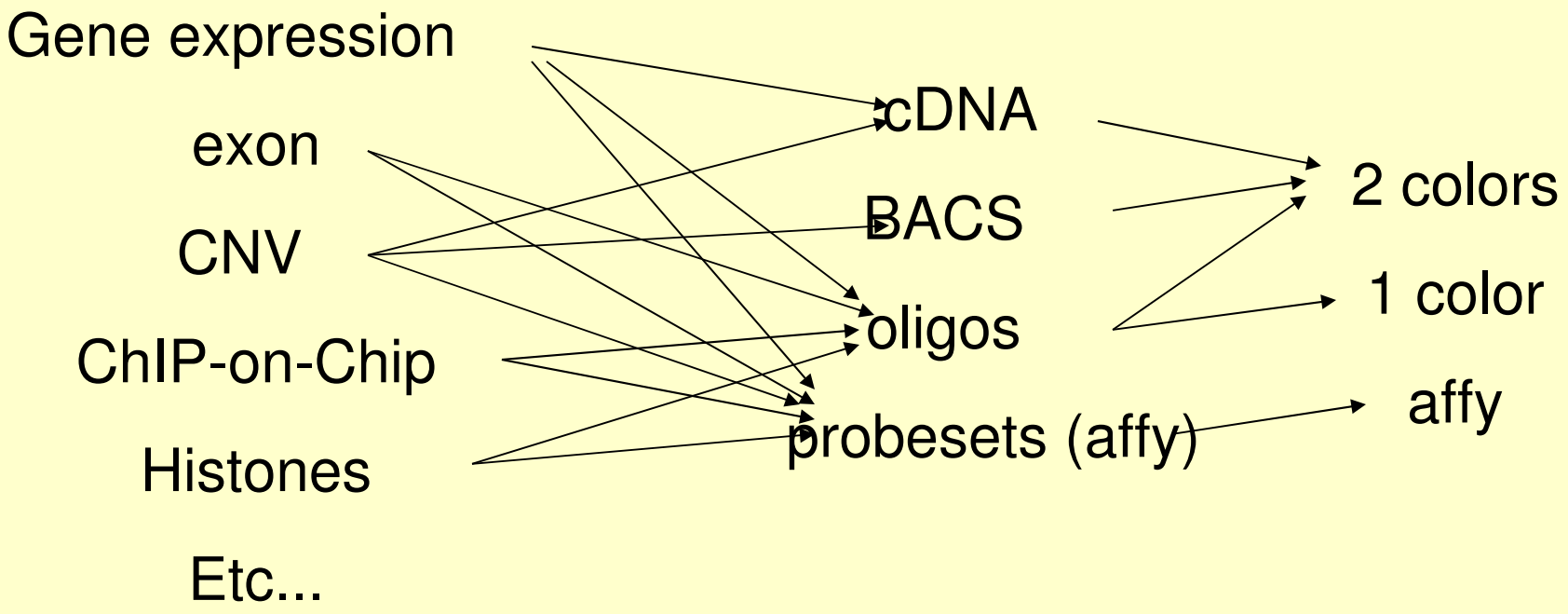
There are many sources of error that can affect and seriously bias the interpretation of the results. Differences in the efficiency of labelling, the hybridisation, local effects, etc.

Normalisation is a necessary step before proceeding with the analysis



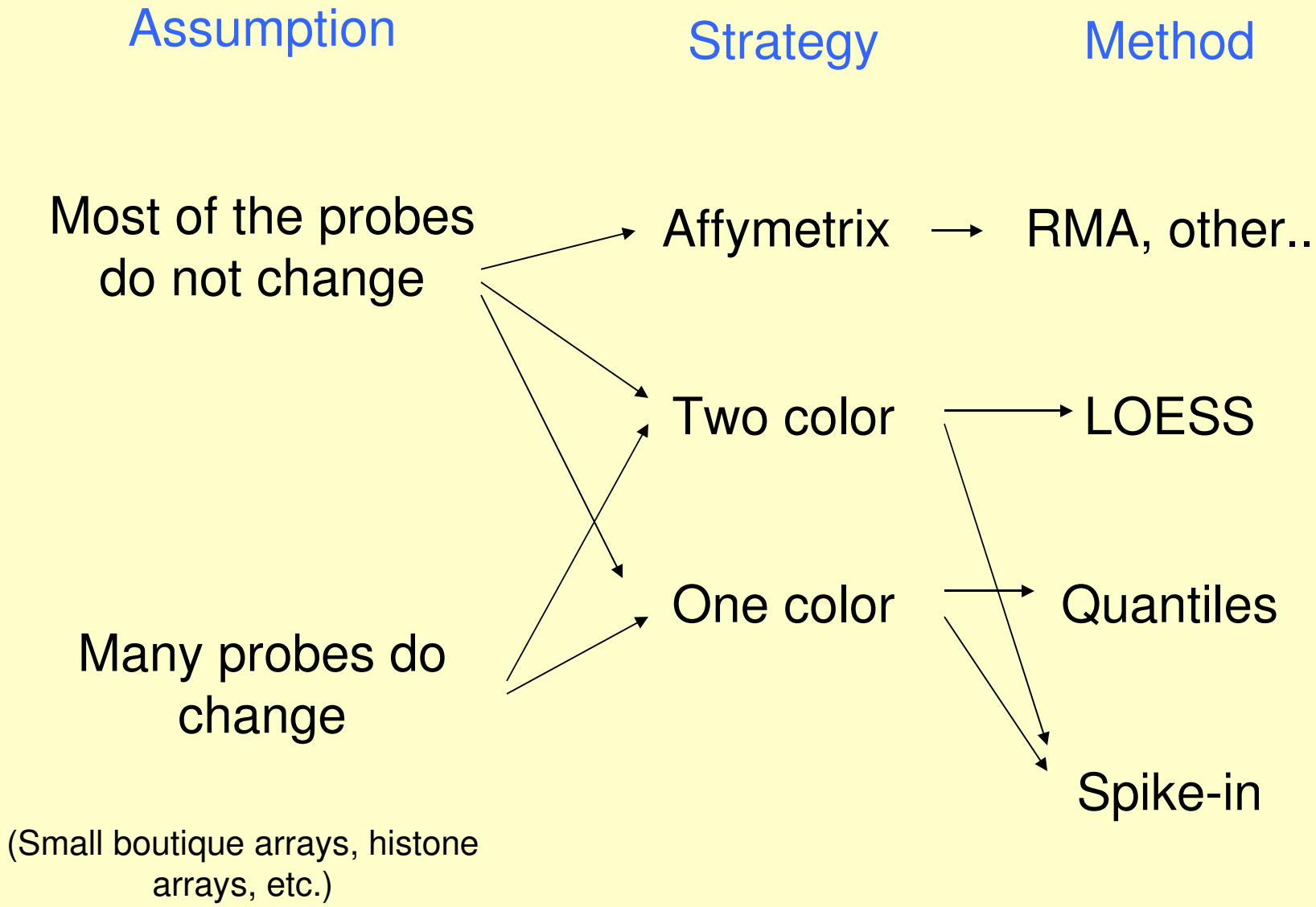
(a) After normalization by average (b) after print-tip lowess normalization (c) after normalisation taking into account spatial effects

On data, probes and hybridisations



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

Common normalisation strategies

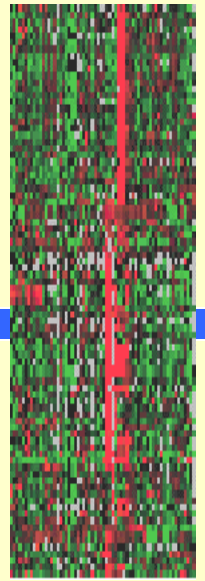


Unsupervised problem: class discovery

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?

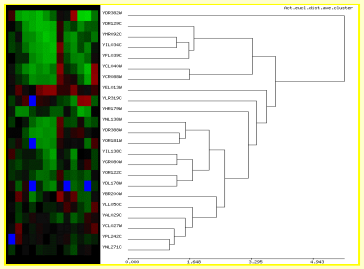
Co-expressing genes...

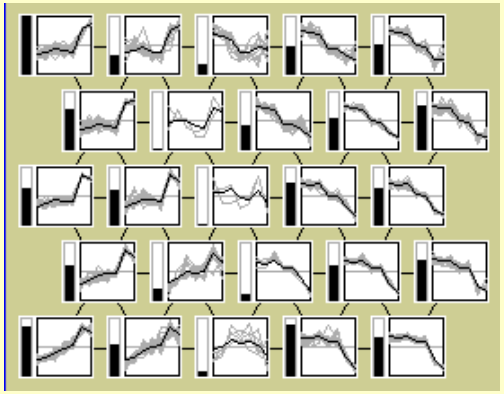


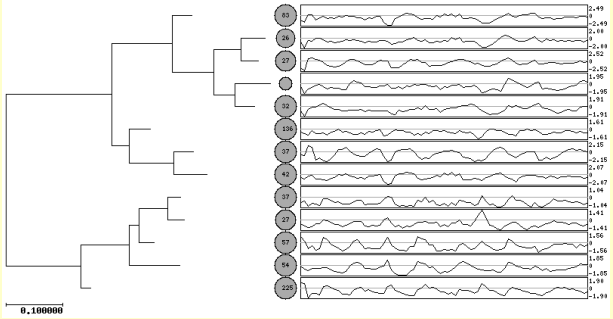
- What genes co-express?
- How many different expression patterns do we have?
- What do they have in common?
- Etc.

Unsupervised clustering methods: Method + distance: produce groups of items based on its global similarity

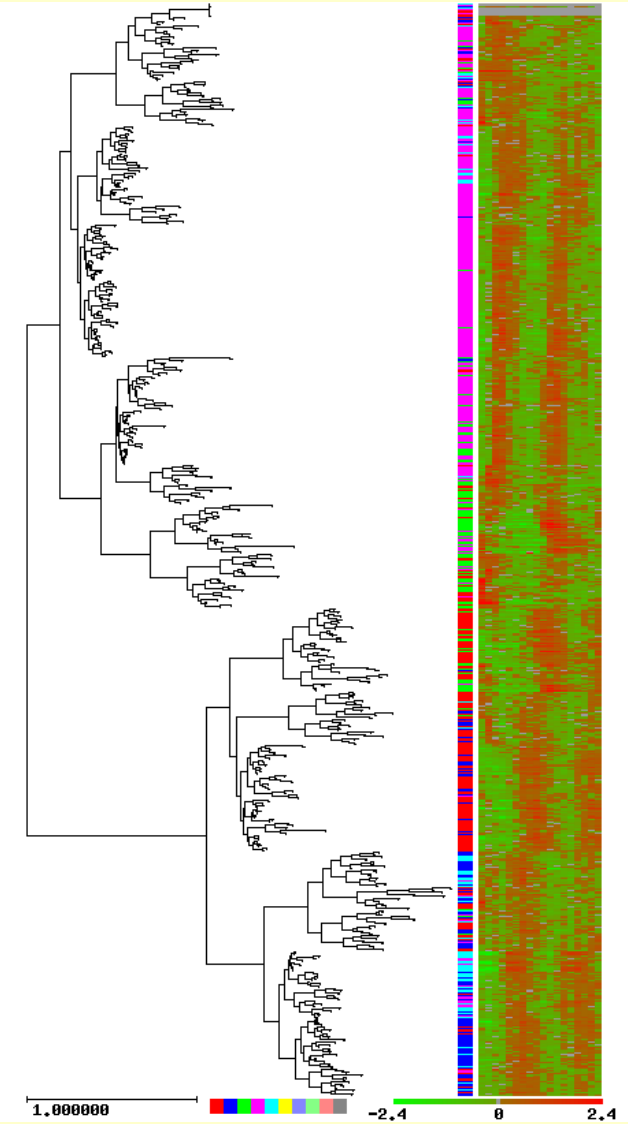
Non hierarchical	**hierarchical**
K-means	UPGMA
SOM	SOTA
	Different levels of information







An unsupervised problem: clustering of genes.



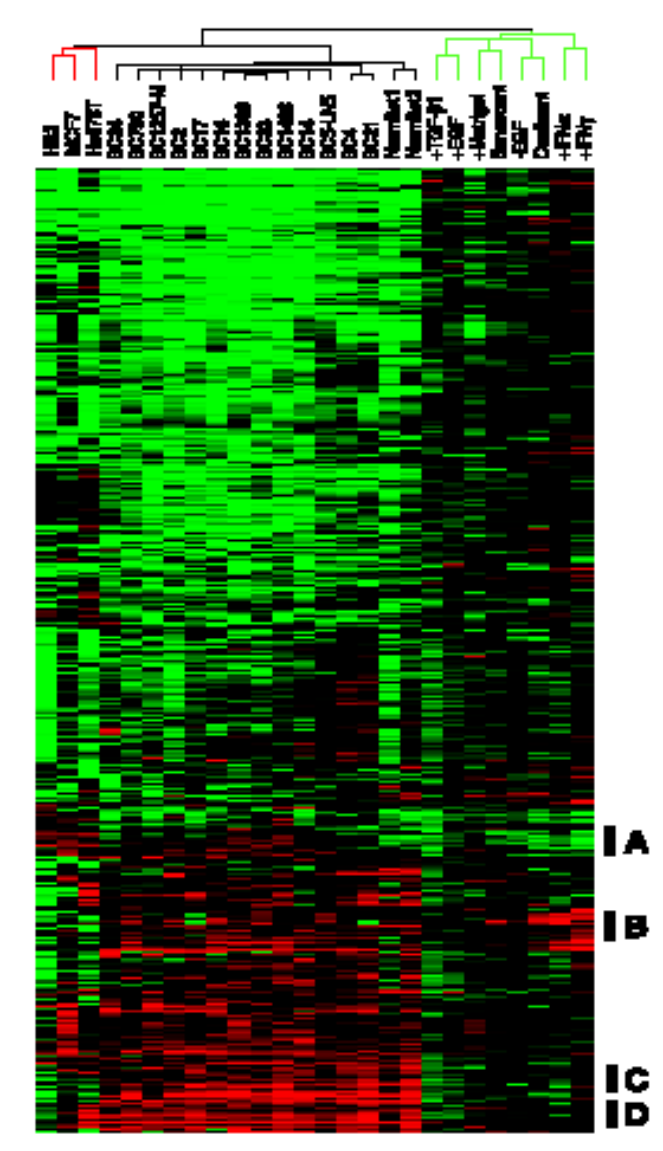
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

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

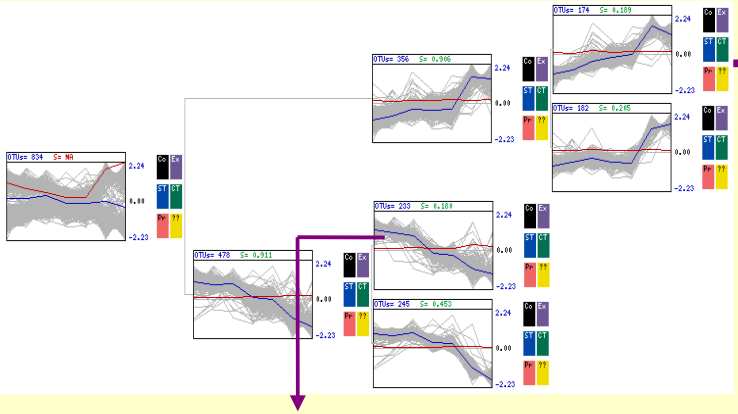
Overview of the combined *in vitro* and breast tissue specimen cluster diagram. A scaled-down representation of the 1,247-gene cluster diagram The black bars show the positions of the clusters discussed in the text: (*A*) proliferation-associated, (*B*) IFNregulated, (*C*) B lymphocytes, and (*D*) stromal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

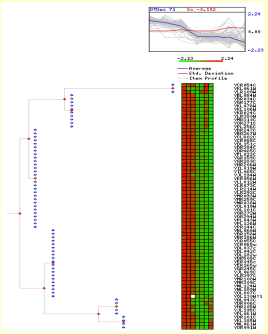


Perou et al., PNAS 96 (1999)

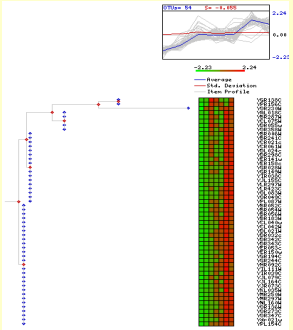
CAAT: Interactive tree browsing

Summary Trees

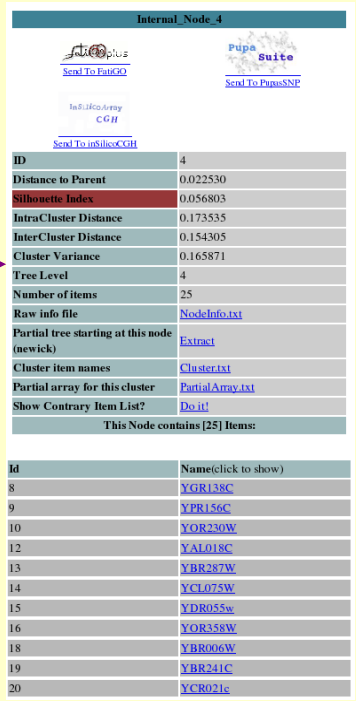




Complete Trees



Cluster validation & analysis



Internal_Node_4

Send To FatiGO

Send To PupasSNP

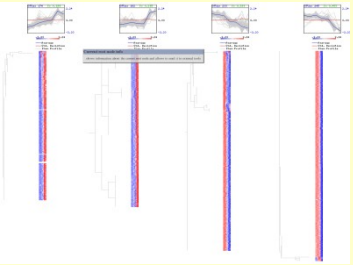
Send To inSilicoCGH

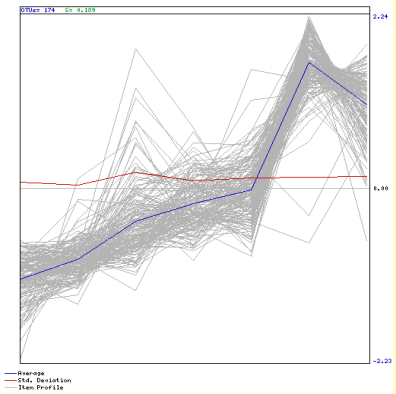
ID	4
Distance to Parent	0.022530
Silhouette Index	0.056803
IntraCluster Distance	0.173535
InterCluster Distance	0.154305
Cluster Variance	0.165871
Tree Level	4
Number of items	25
Raw info file	NodeInfo.txt
Partial tree starting at this node (newick)	Extract
Cluster item names	Cluster.txt
Partial array for this cluster	PartialArray.txt
Show Contrary Item List?	Do it!

This Node contains [25] Items:

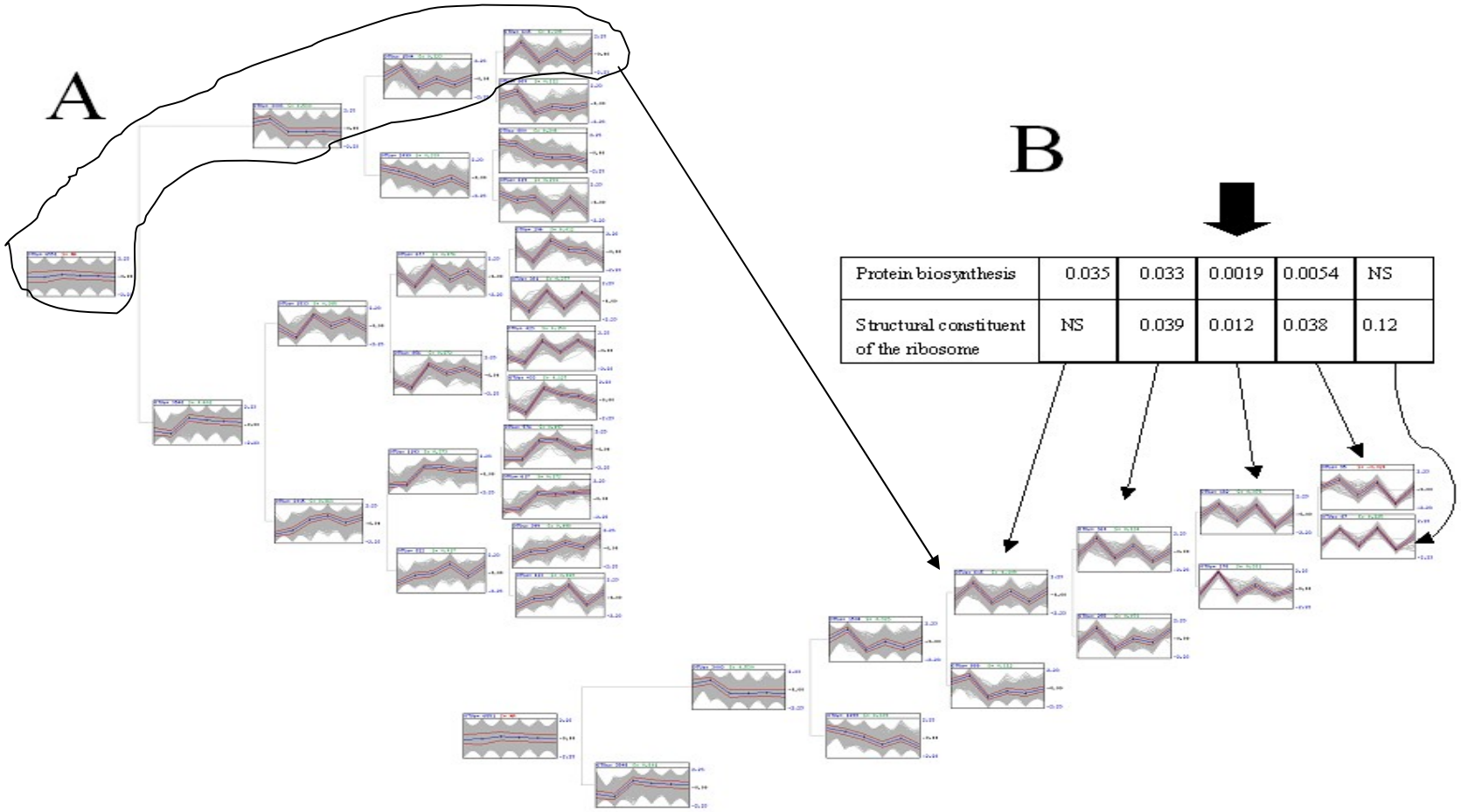
Id	Name(click to show)
8	YGR138C
9	YPR156C
10	YOR230W
12	YAL018C
13	YBR287W
14	YCL075W
15	YDR055w
16	YOR358W
18	YBR006W
19	YBR241C
20	YCR021c

Cluster extraction





Cluster quality external measures: Functional interpretation



Protein biosynthesis	0.035	0.033	0.0019	0.0054	NS
Structural constituent of the ribosome	NS	0.039	0.012	0.038	0.12

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

Information on classes (<u>defined on criteria external to the gene expression measurements</u>) 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?

Class prediction

And

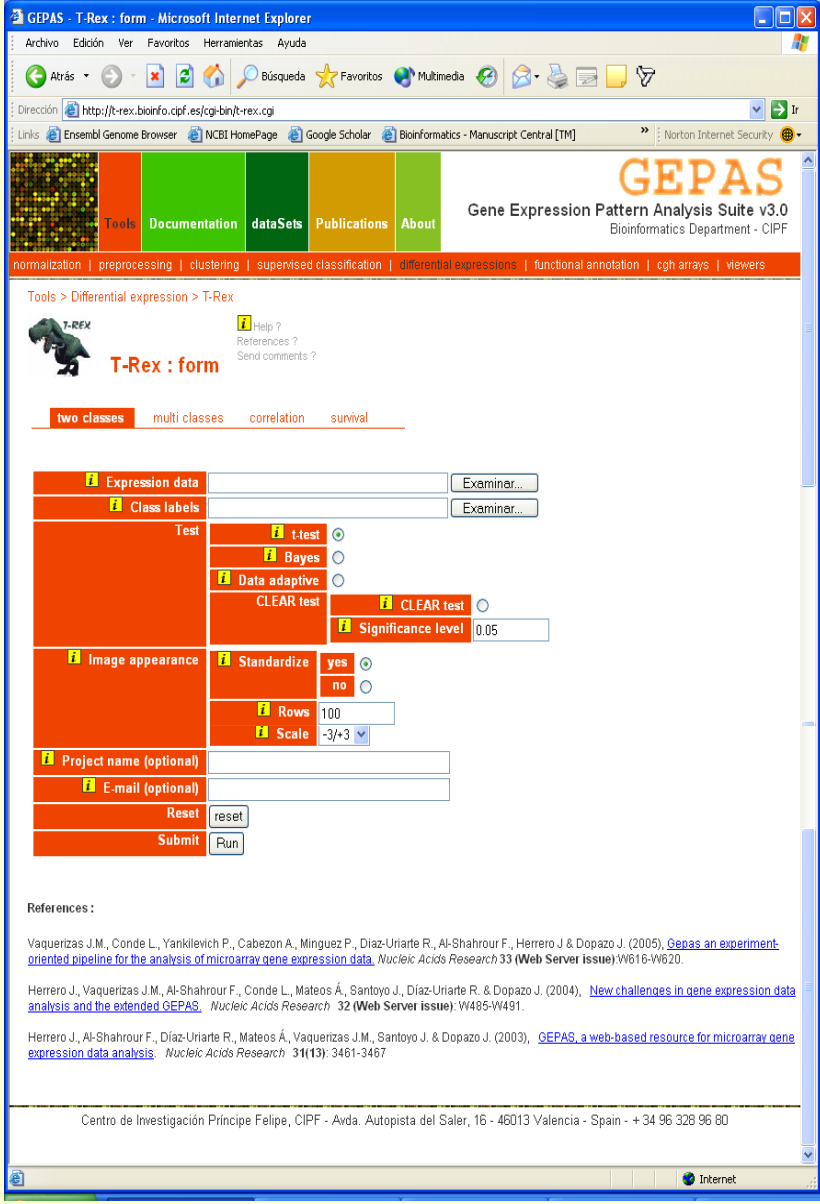
Which genes among the thousands analysed are relevant for the classification?

Gene selection

Gene selection.

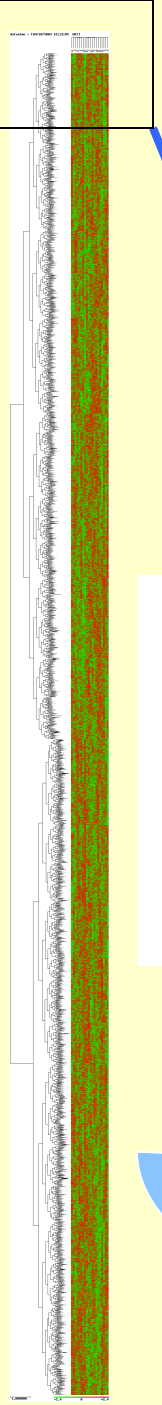
The simplest way: univariant gene-by-gene.
Other multivariant approaches can be used

- **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



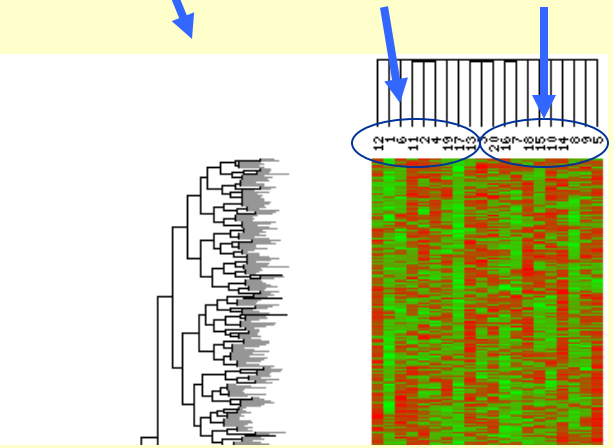
The T-rex tool

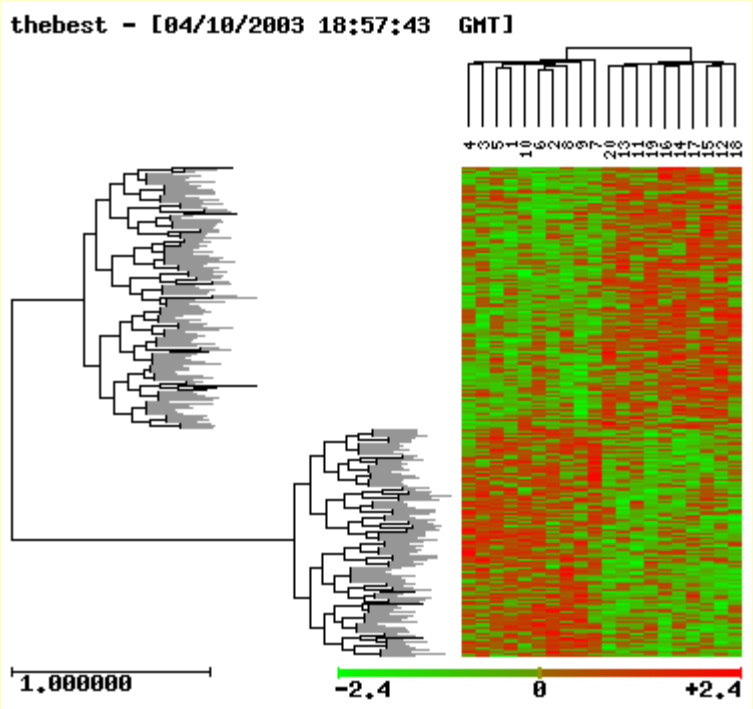
A simple problem: gene selection for class discrimination



~15,000 genes

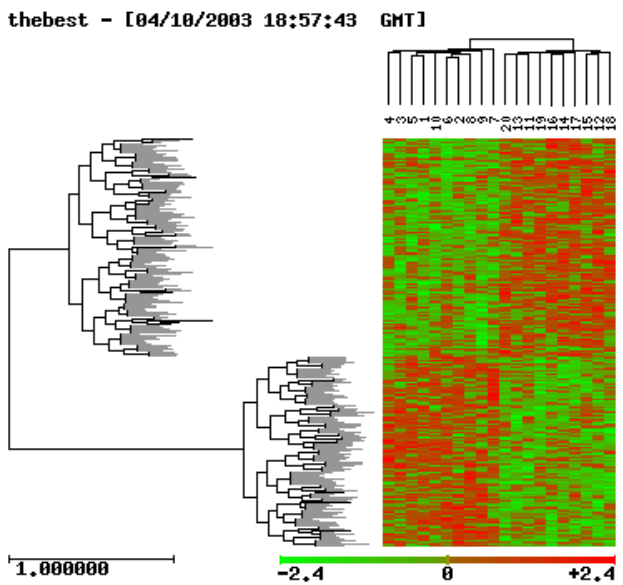
Case(10)/control(10)





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?

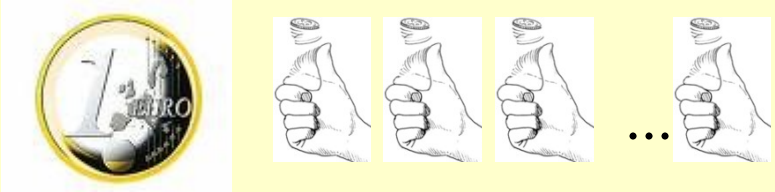
unadj.p	adj_p	FDR_indep	FDR_dep	obs_stat
0.00019998	0.152685	0.49995	1	5.47044
0.00019998	0.746225	0.49995	1	4.49902
0.0009999	0.983002	0.861025	1	4.01726
0.00149985	0.986401	0.861025	1	3.99374
0.00129987	0.9959	0.861025	1	3.86046
0.00169983	0.9996	0.861025	1	3.7251
				66628
				62427
				60596
				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

1840	1840
1007	1007
1542	1542
1360	1360
844	844
4631	4631
11	11
4102	4102
285	285
4716	4716
4430	4430
4398	4398
3793	3793
3462	3462
972	972
3488	3488
3992	3992
1248	1248

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

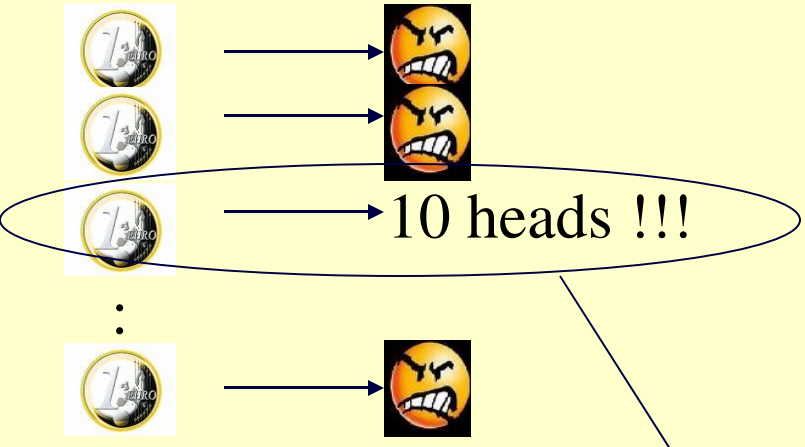
Adjusted p-values must be used!

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



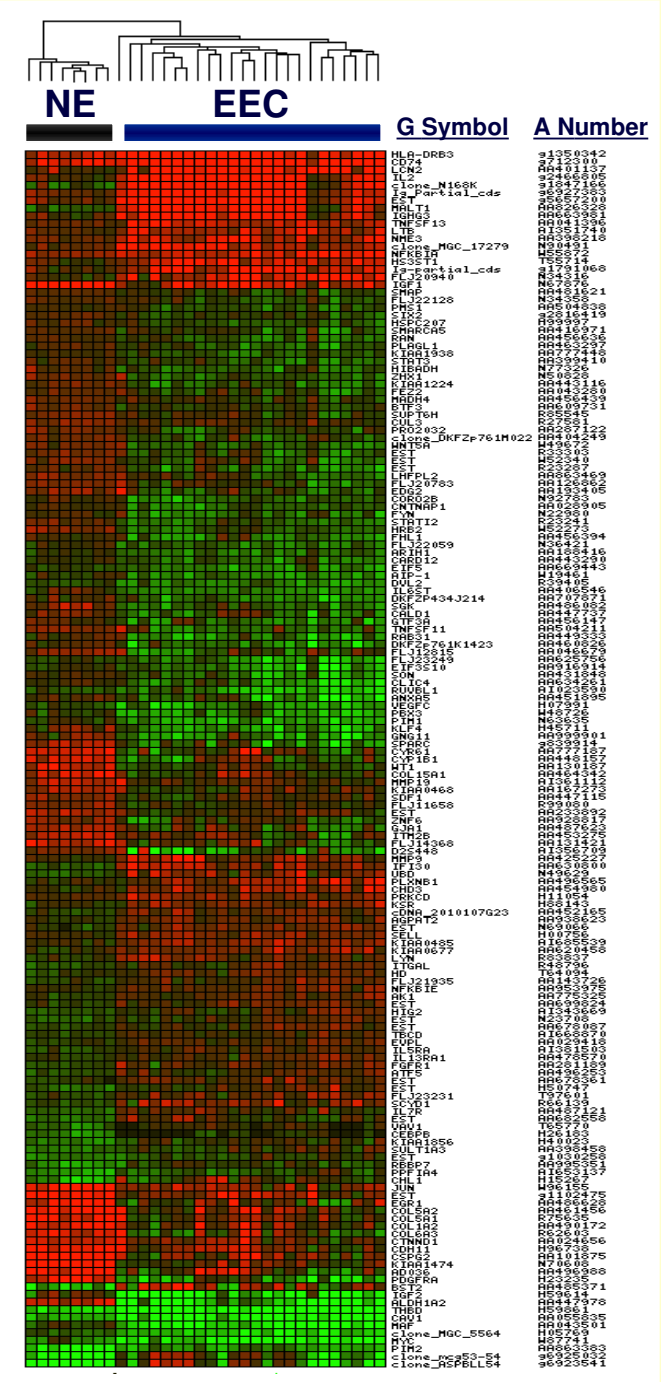
1000 coins

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

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?

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



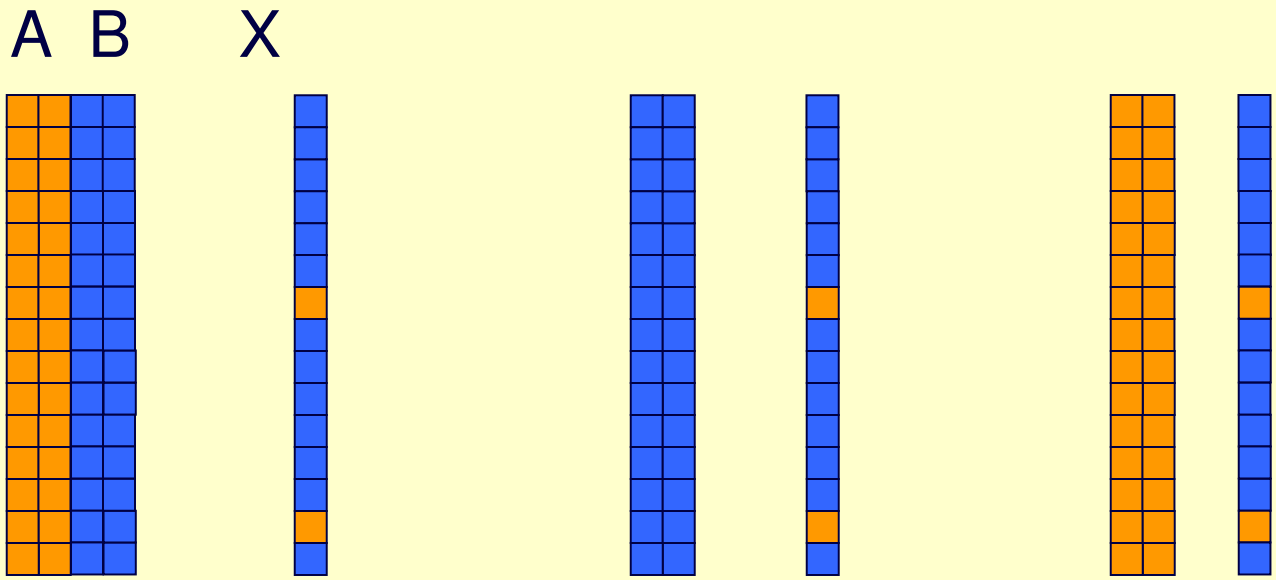
Hierarchical Clustering of 86 genes with different expression patterns between Normal Endometrium and Endometrioid Endometrial Carcinoma (FDR adjusted p<0.05) selected among the ~7000 genes in the CNIO oncochip

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

Of predictors and molecular signatures

What is a predictor?

Intuitive notion:



Is X, A or B?

Diff (B, X) = 2

Diff (A, X) = 13

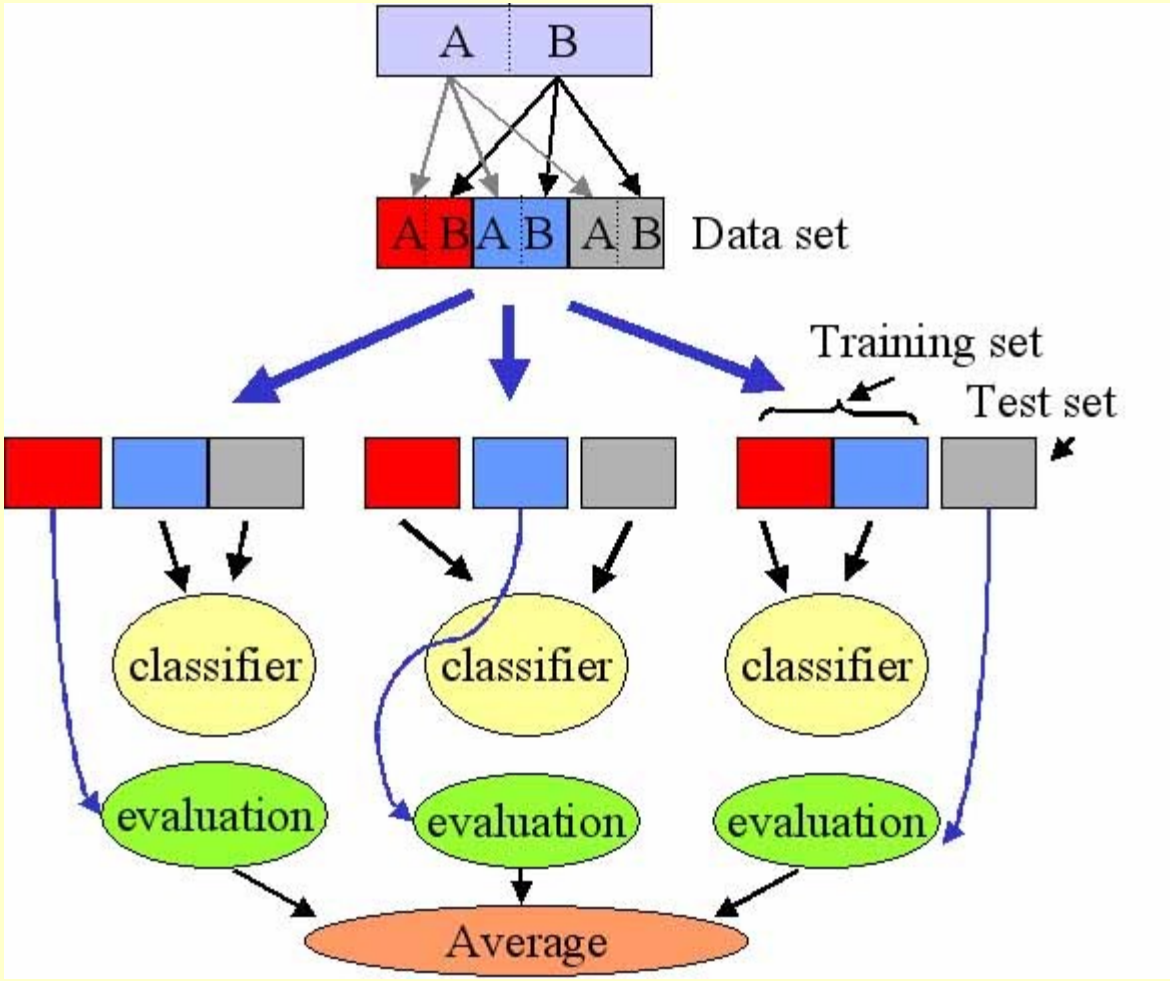
Most probably X belongs to class B

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

Cross-validation

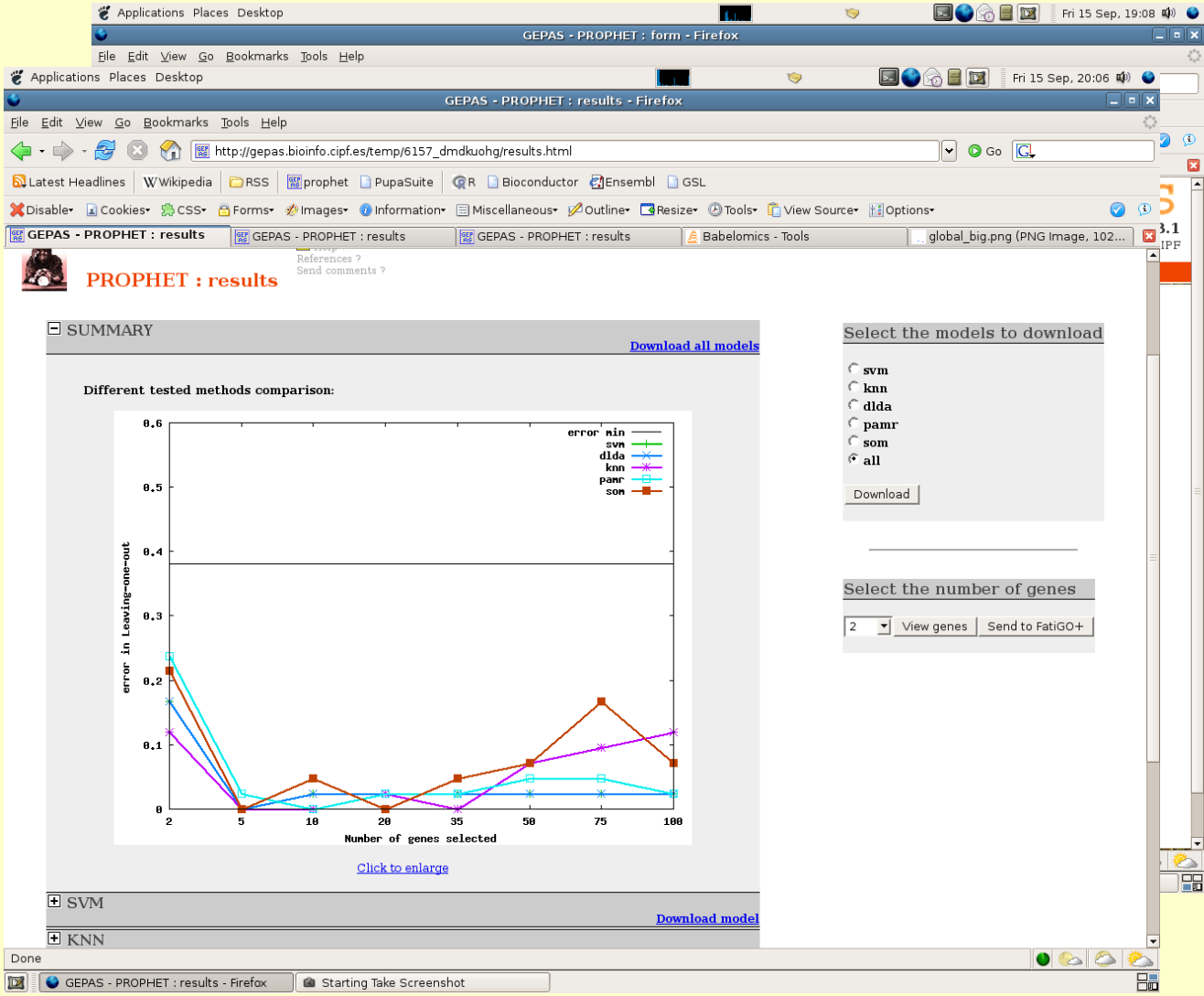
The efficiency of a classifier can be estimated through a process of cross-validation.

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

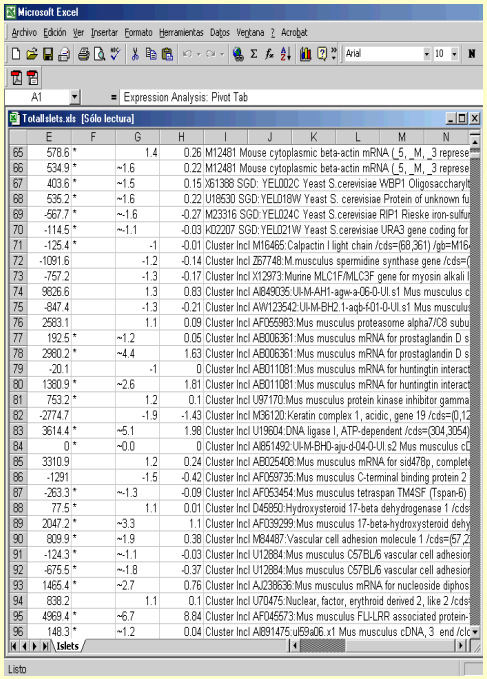
Functional profiling of genome-scale experiments in the post-genomic era

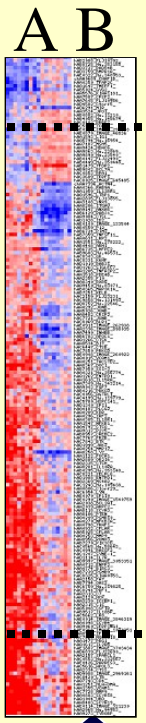
My data...

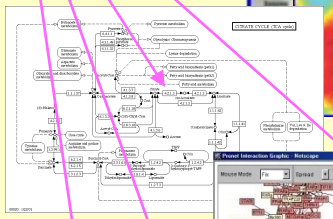
How are structured?

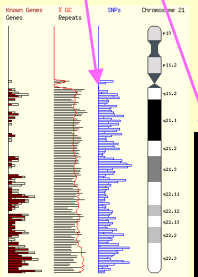
What are these groups?

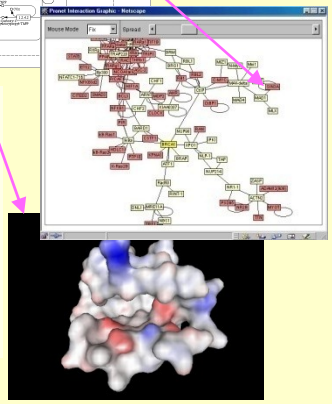
What is this gen?





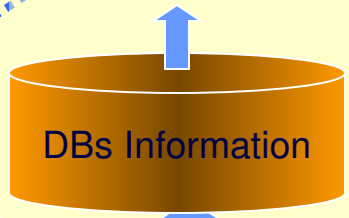


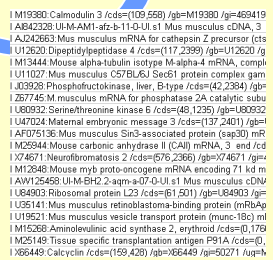




?

Cell cycle...

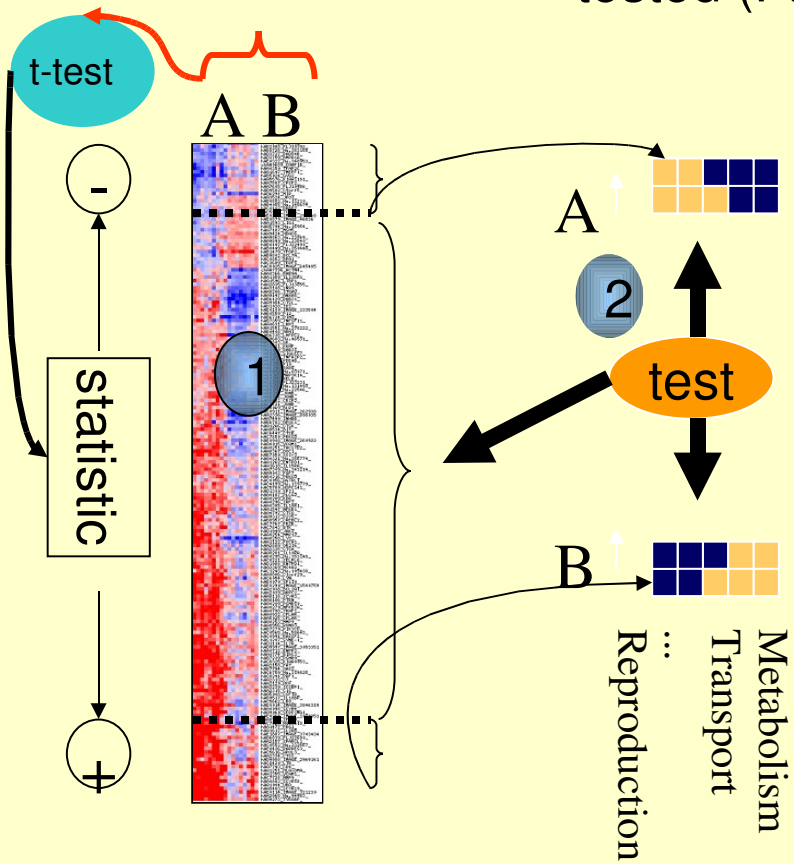


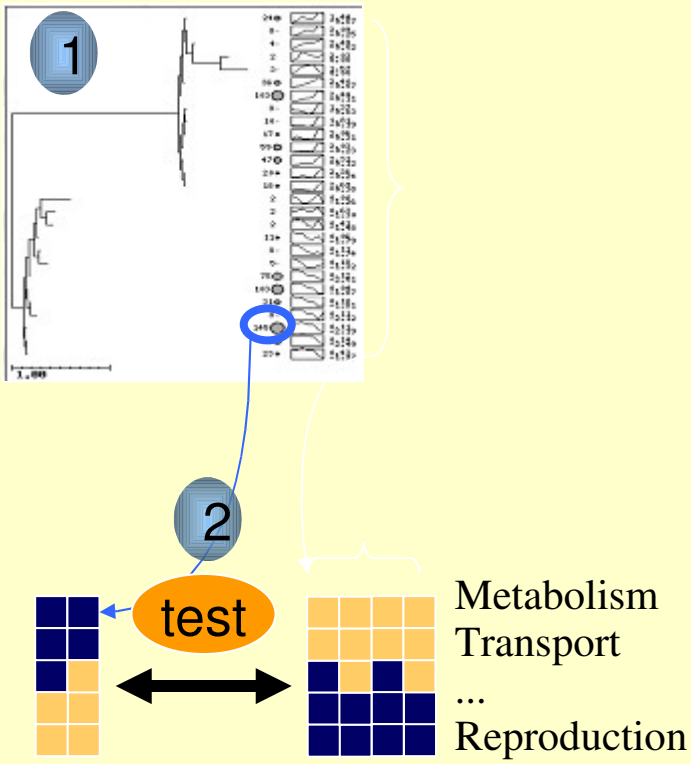


Analysis *Functional profiling* *Links*

Functional enrichment (two steps)

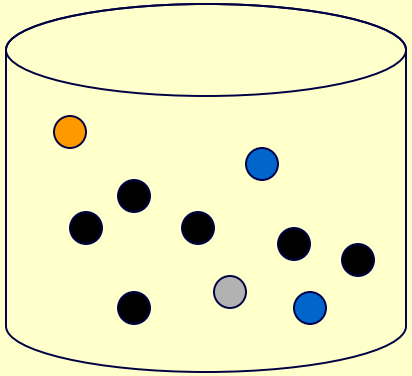
1. Genes are selected based on their experimental values and...
2. Enrichment in functional terms is tested (FatiGO, GoMiner, etc.)





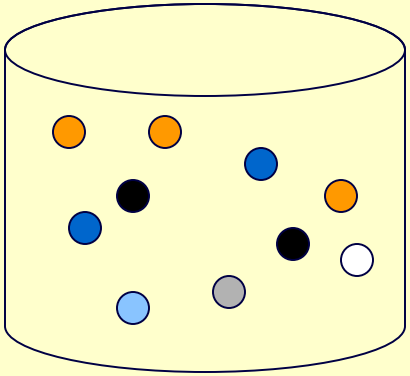
Testing two GO terms
(remember, we have to test thousands)

Group A



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

Group B



	Biosynthesis	Other
A	6	4
B	2	8

The popular Fisher's test

Biosynthesis 60% ● ⟷ Biosynthesis 20% ●

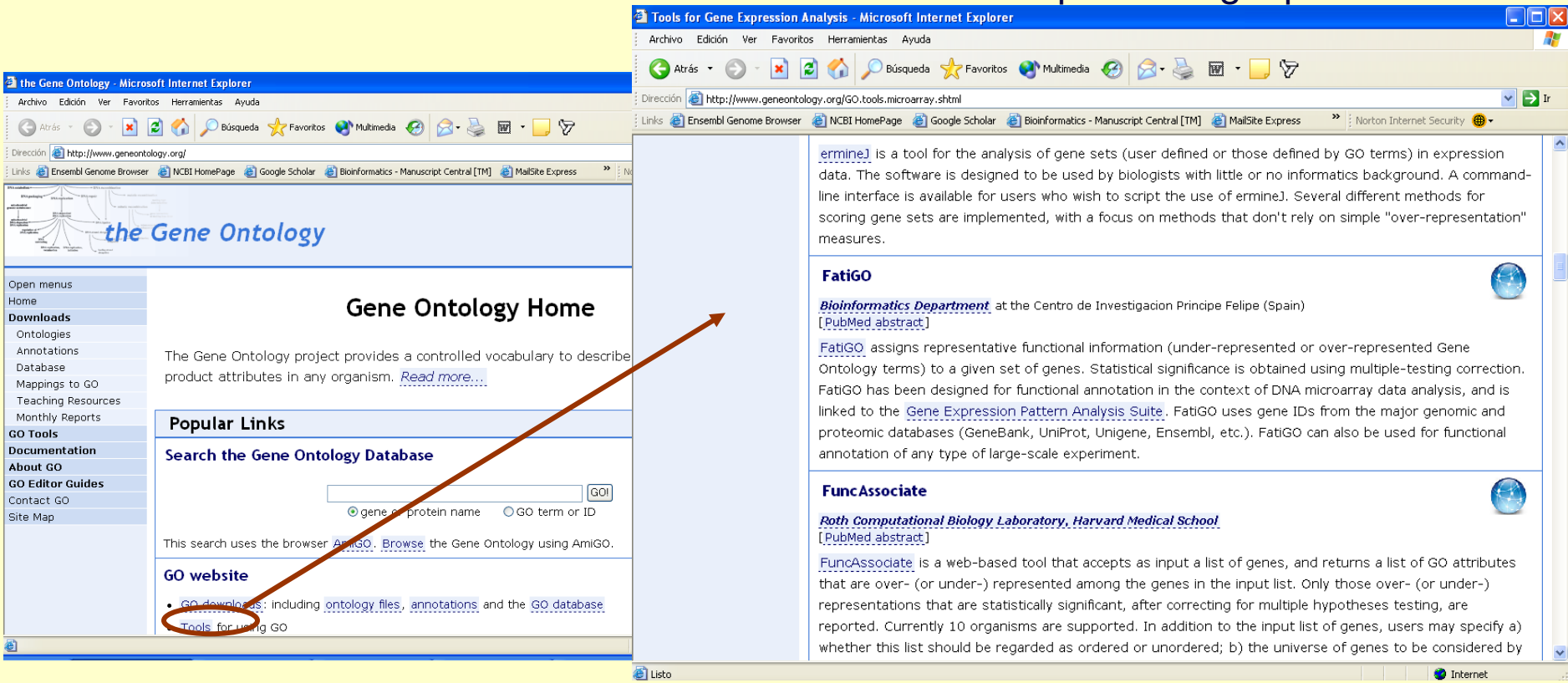
Sporulation 20% ● ~~⟷~~ Sporulation 20% ●

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

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

FatiGO: GO-driven data analysis

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



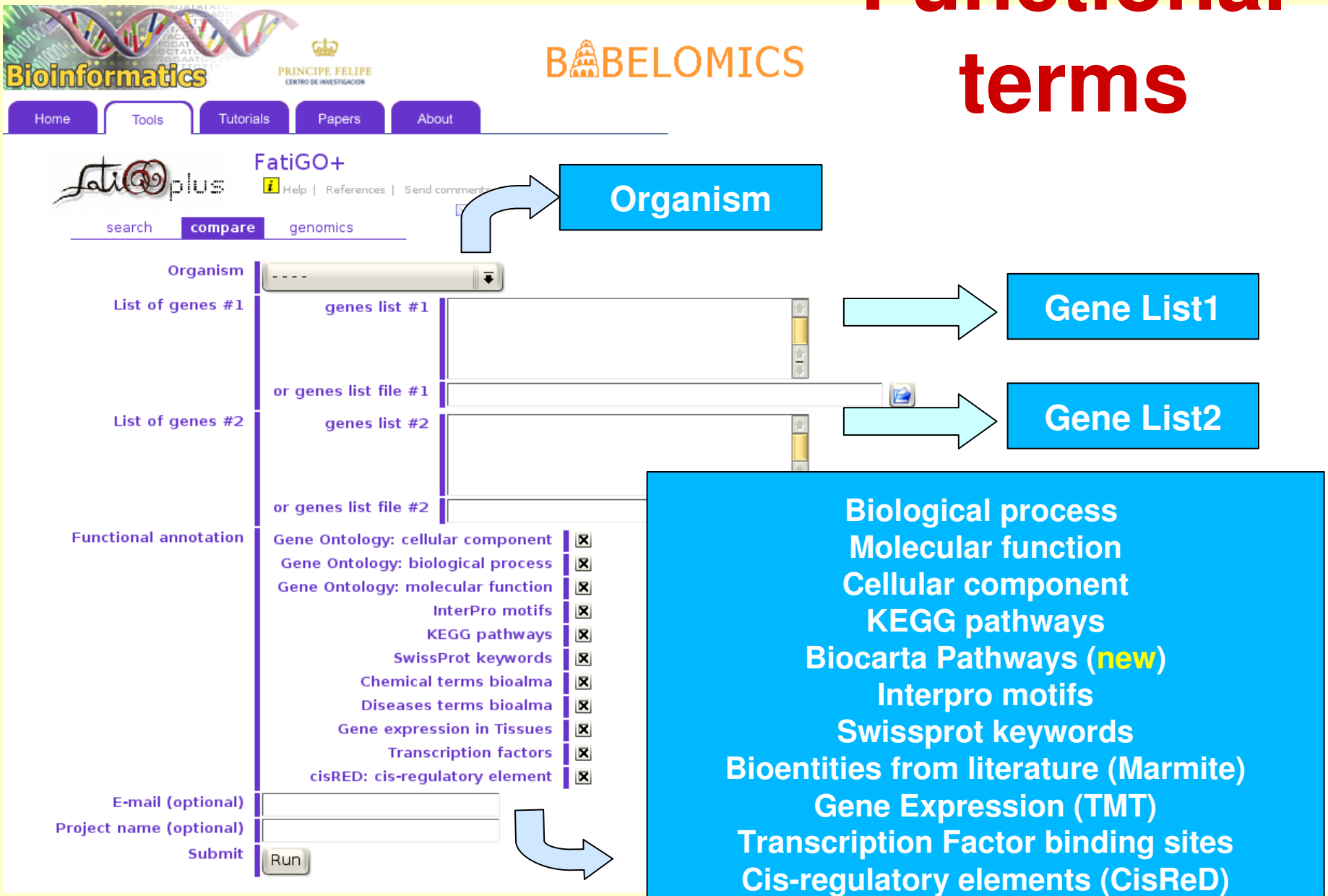
Al-Shahrour et al., 2004 Bioinformatics ***(3rd most cited paper in computing sciences. Source: ISI Web of knowledge.)***

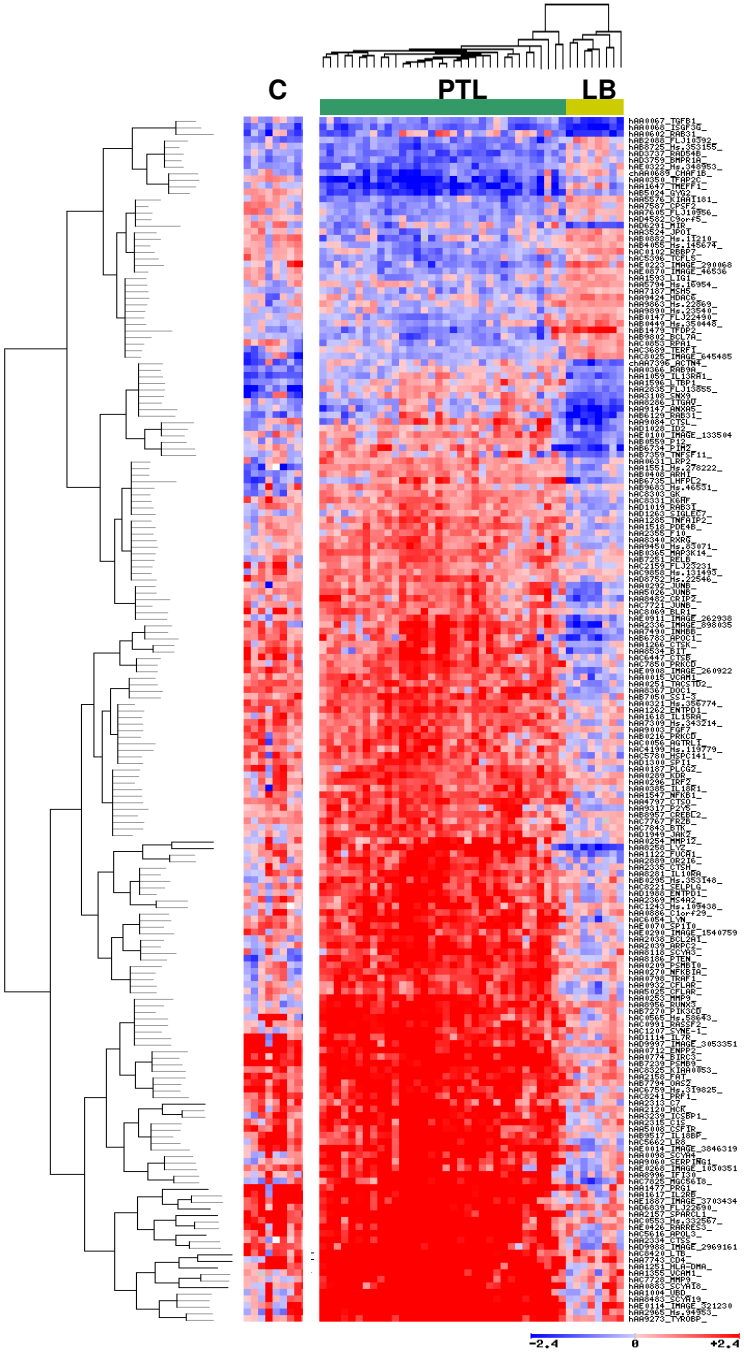
Al-Shahrour et al., 2005 Bioinformatics; Al-Shahrour et al., 2005 NAR

Al-Shahrour et al., 2006 NAR

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

Functional terms





Understanding why genes differ in their expression between two different conditions

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

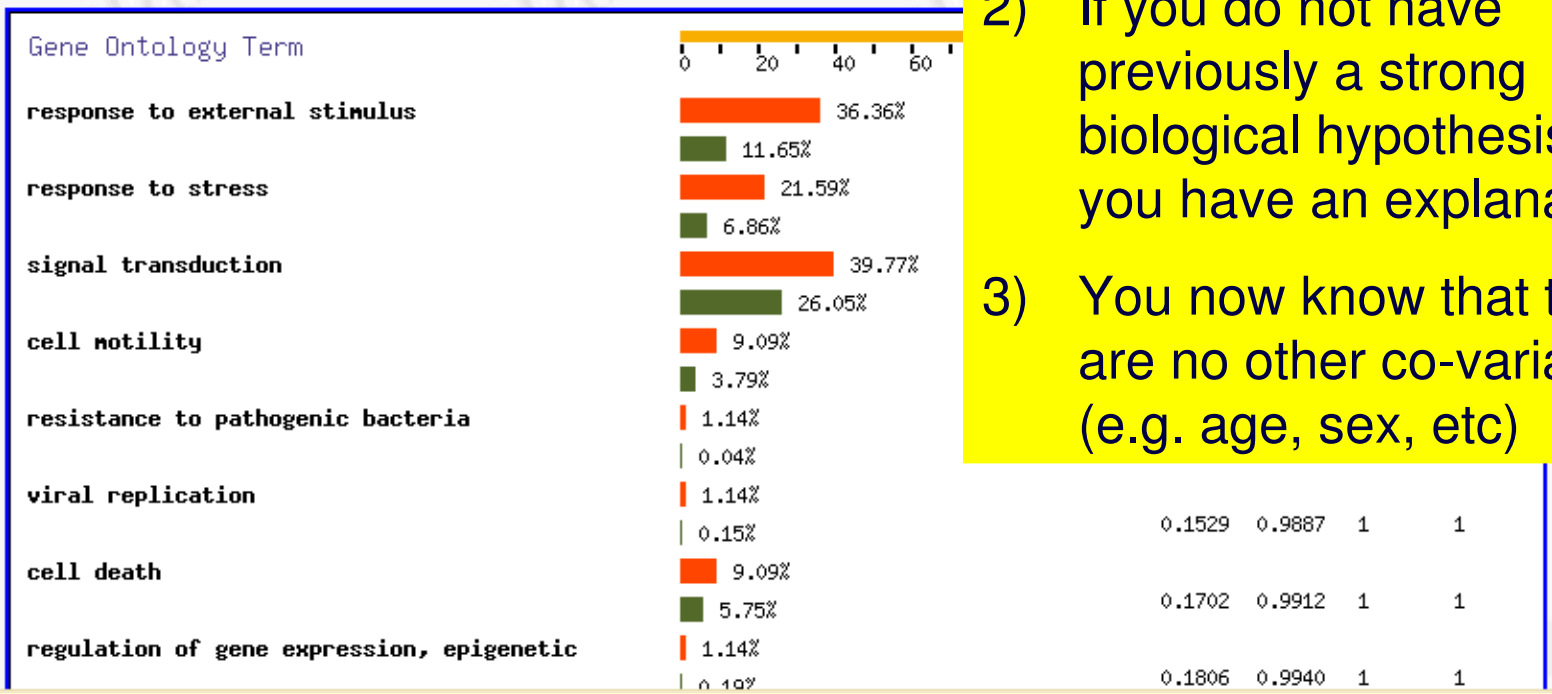
Genes differentially expressed, selected among the ~7000 genes in the CNIO oncochip

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:		

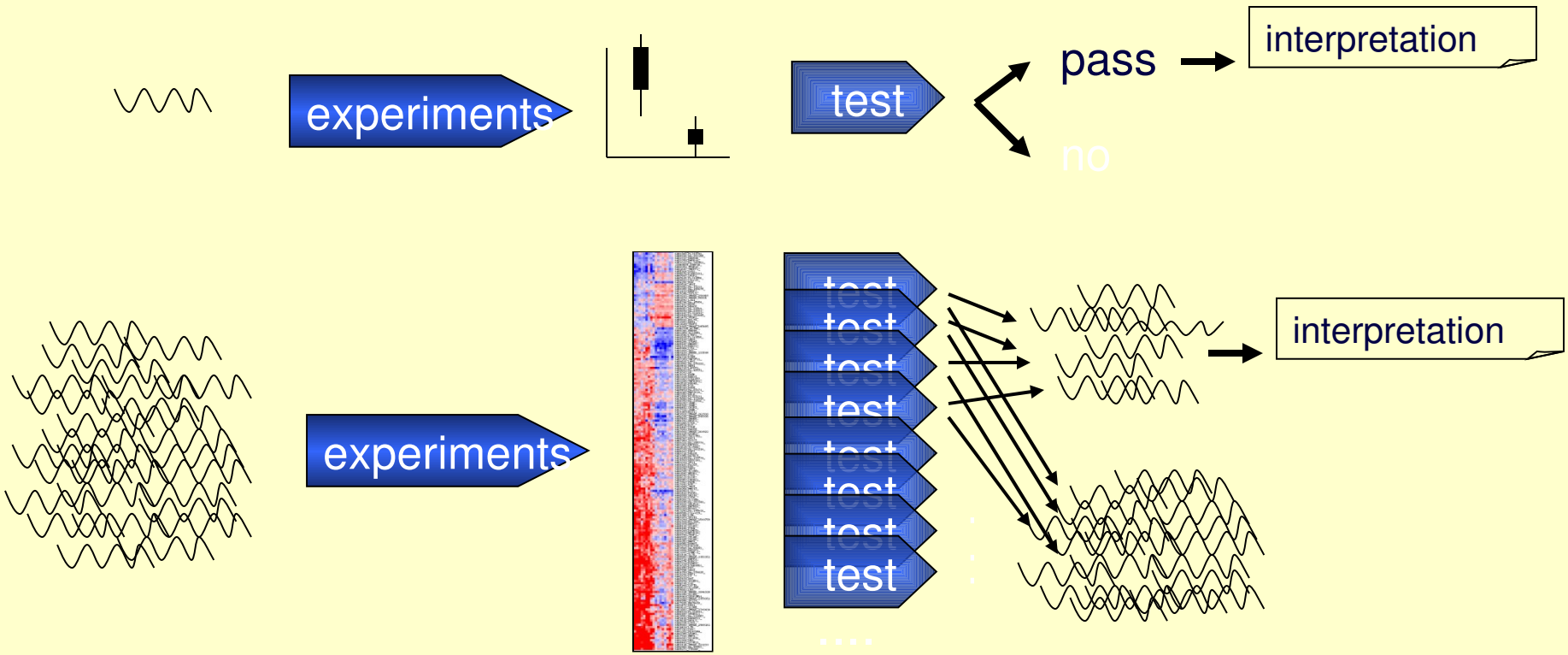


Obvious? NO

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

Function is carried out by gene sets

Now: 22240 (NCBI build 35 12/04)
50-70% display alternative splicing
25%-60% unknown
Transfrags

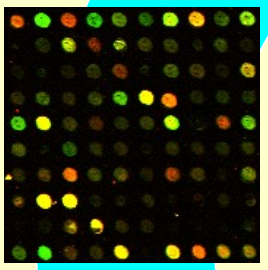
Genes in the DNA...

>protein kunase

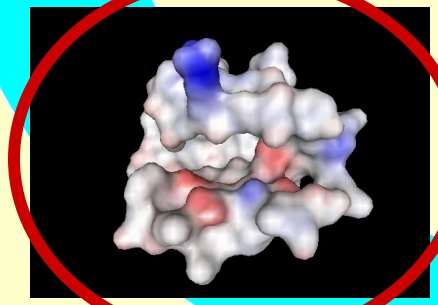
acctgttgatggcgacagggactgtatgctgatc
tatgctgatgcatgcatgctgactactgatgtggg
ggctattgacttgatgtctatc....

…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...

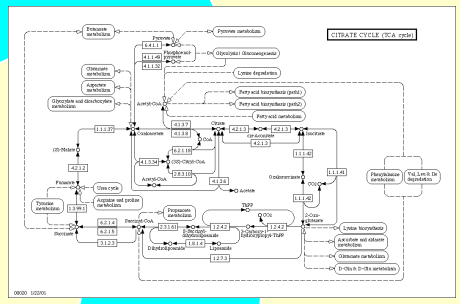
100K-500K proteins

…whose structures account for function...

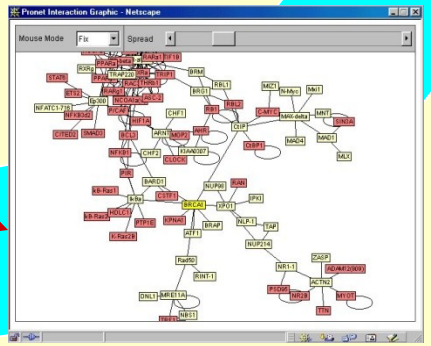
…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype...



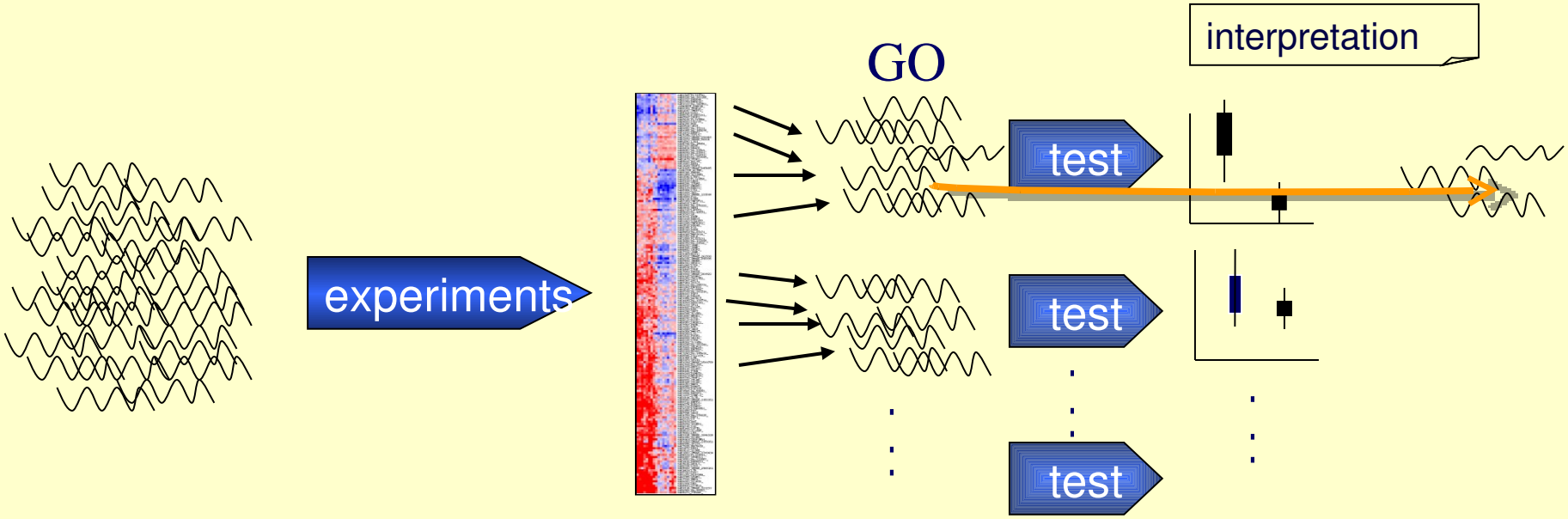
…conforming complex interaction networks...



…in cooperation with other proteins…

Each protein has an average of 8 interactions

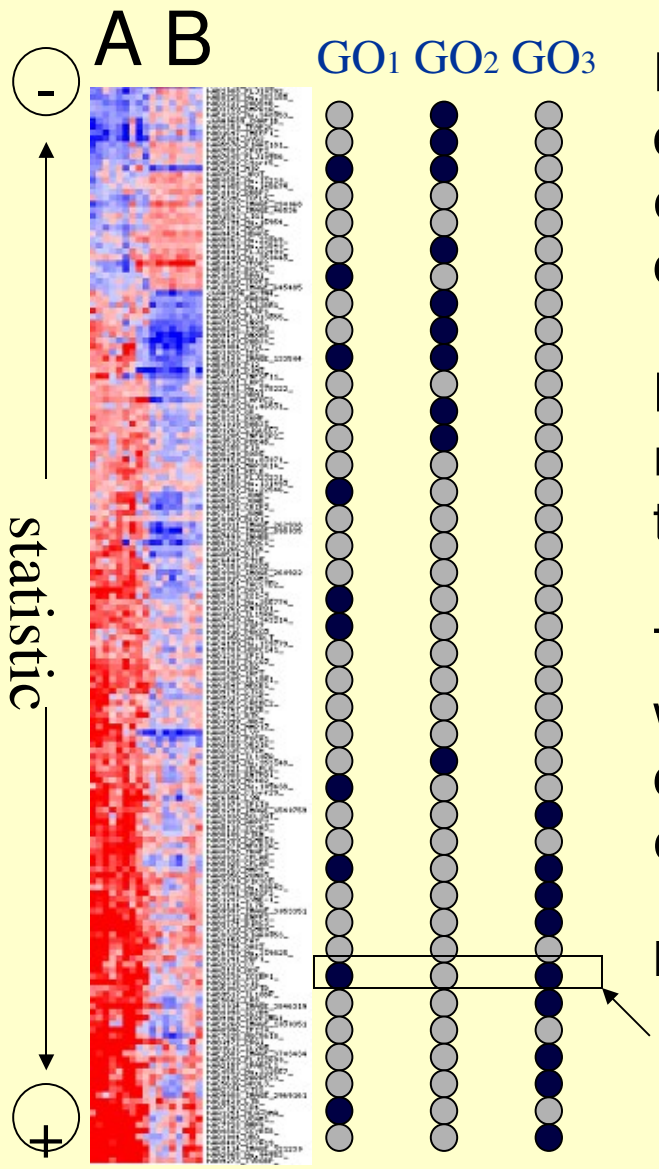
Hypothesis inspired in biology (functionality carries out by gene-sets) must be tested



Experiment → ~~Gene~~ → Function

Experiment → Function → Gene

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



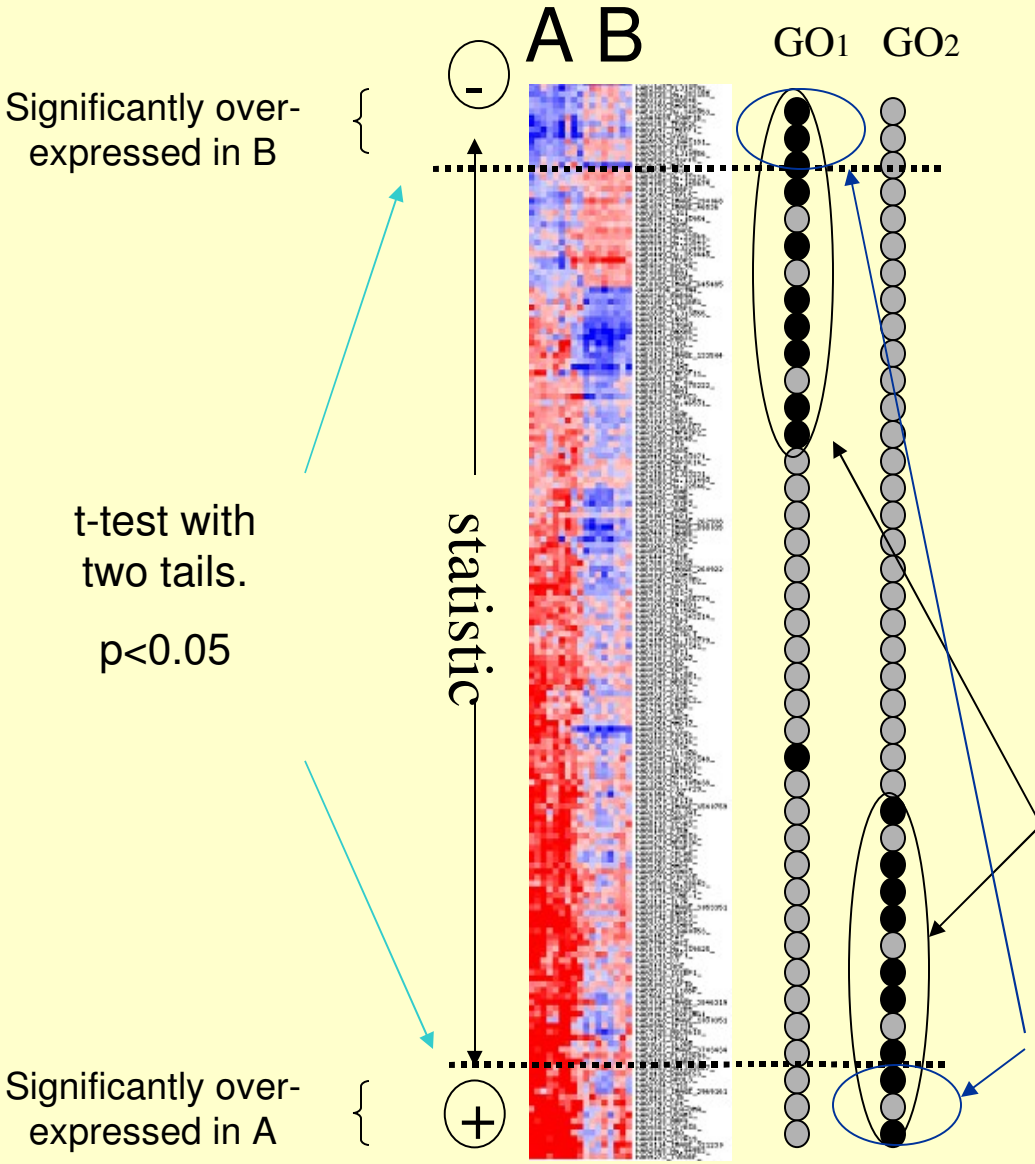
Ranking: A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

Distribution of GO: Rows GO**1**, GO**2** and GO**3** 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



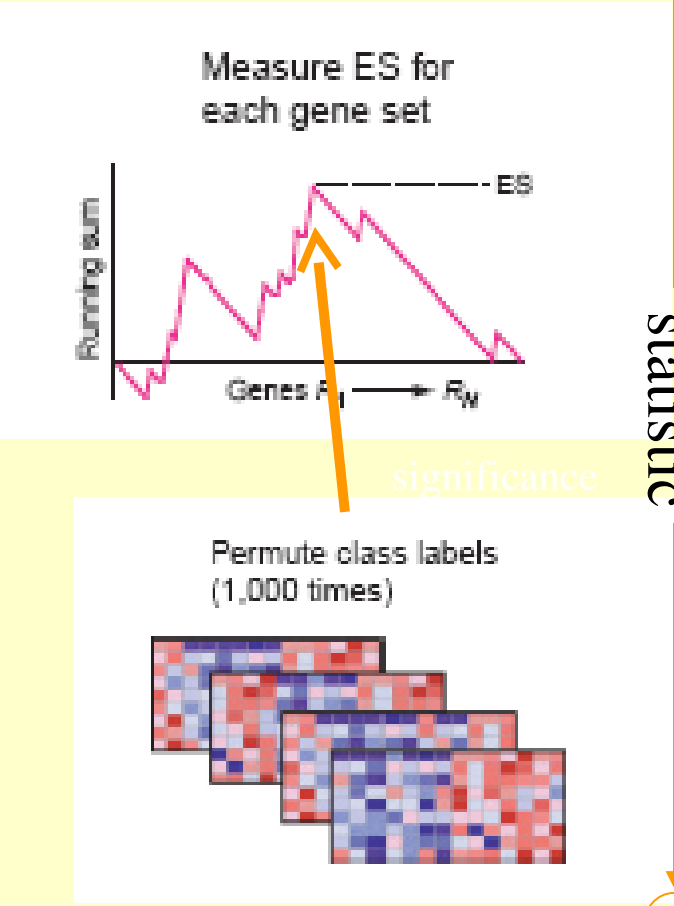
If a threshold based on the experimental values is applied, and the resulting selection of genes compared for over-abundance of a functional term, this migh not be found.

Classes expressed as blocks in A and B

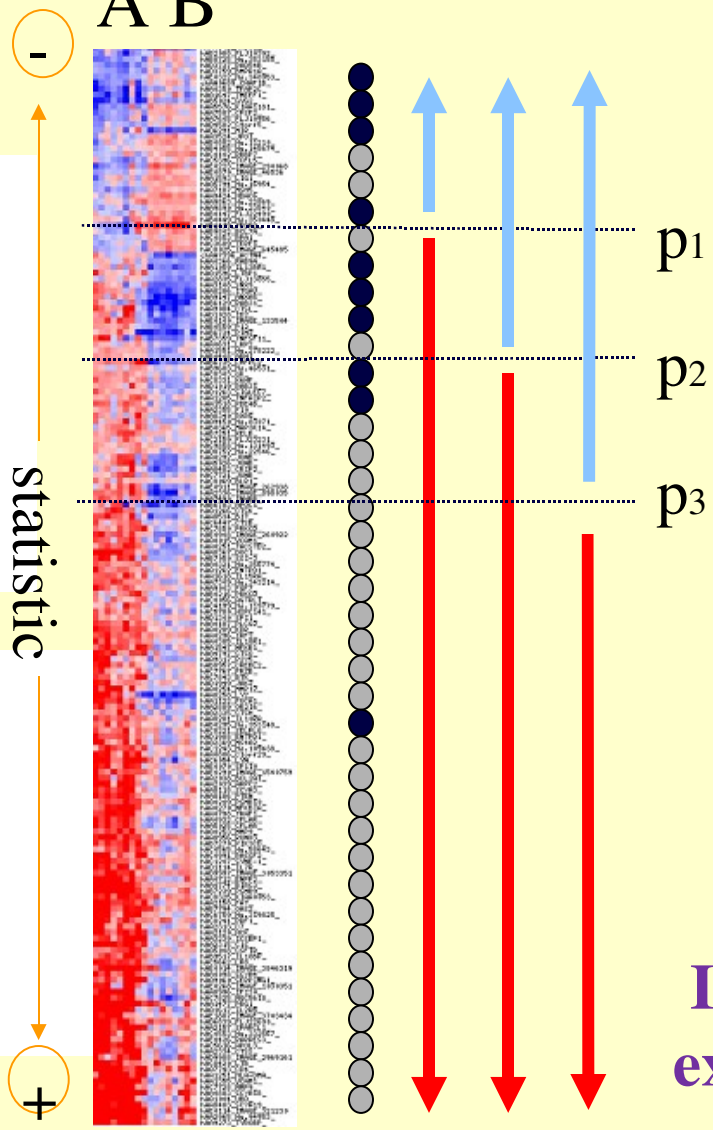
Very few genes selected to arrive to a significant conclussion on GOs 1 and 2

Gene-set enrichment methods

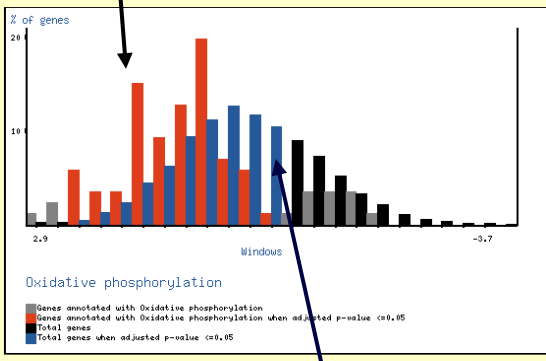
GSEA



A B



FatiScan



Gene set

background

Independent of the experimental design

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

A

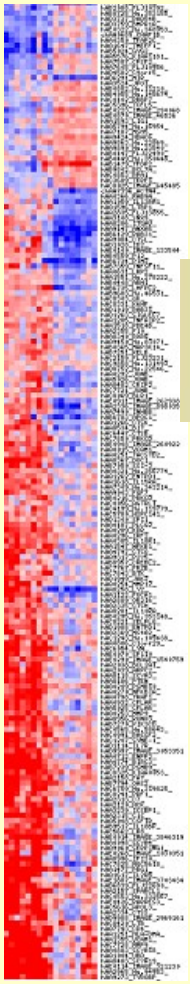
8 with impaired tolerance (**IGT**) + 18 with type 2 diabetes mellitus (**DM2**)

B

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

(Mootha et al., 2003)

A B



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

		Repository		
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 dehidrogenase (ubiquinone) activity	X		
	Nuclease activity	X		
Dow-regulated	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)

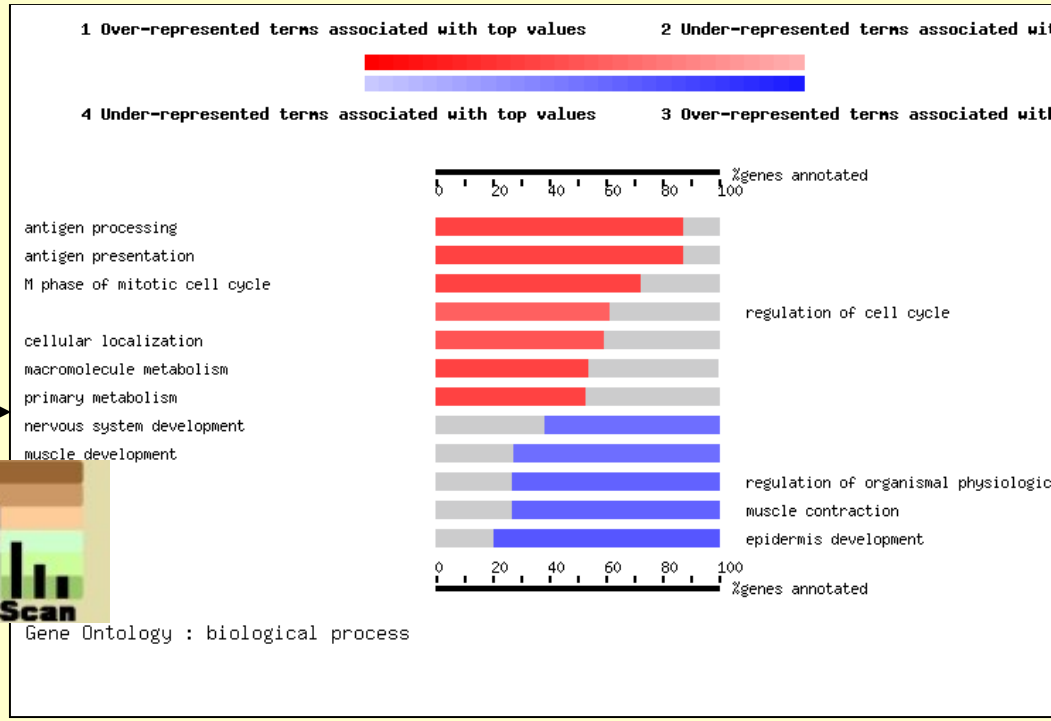
GEPAS
t-rex tool

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

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

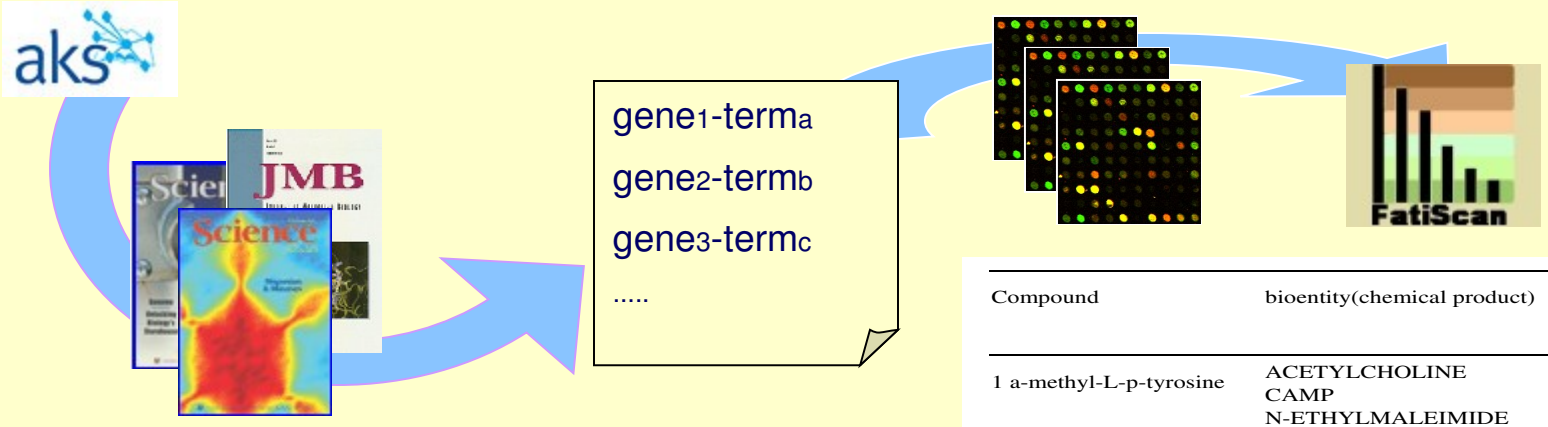
Gen	risk
Gen1	5.8
Gen2	5'6
Gen3	5.4
Gen4	5.2
Gen5	5.2
Gen6	5.0
……	….
……	….
……	….
Gen1000	-6.0
Gen1001	-6.3



+ Survival

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

		Repository				Method			
Healthy vs diabetic	Functional class	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
	NADH dehidrogenase (ubiquinone) activity	+				yes	-	-	-
	Nuclease activity	+				yes	-	-	-
Dow-regulated	Insulin signalling pathway		+			yes	-	-	-

GSEA 2003

FatiScan 2005

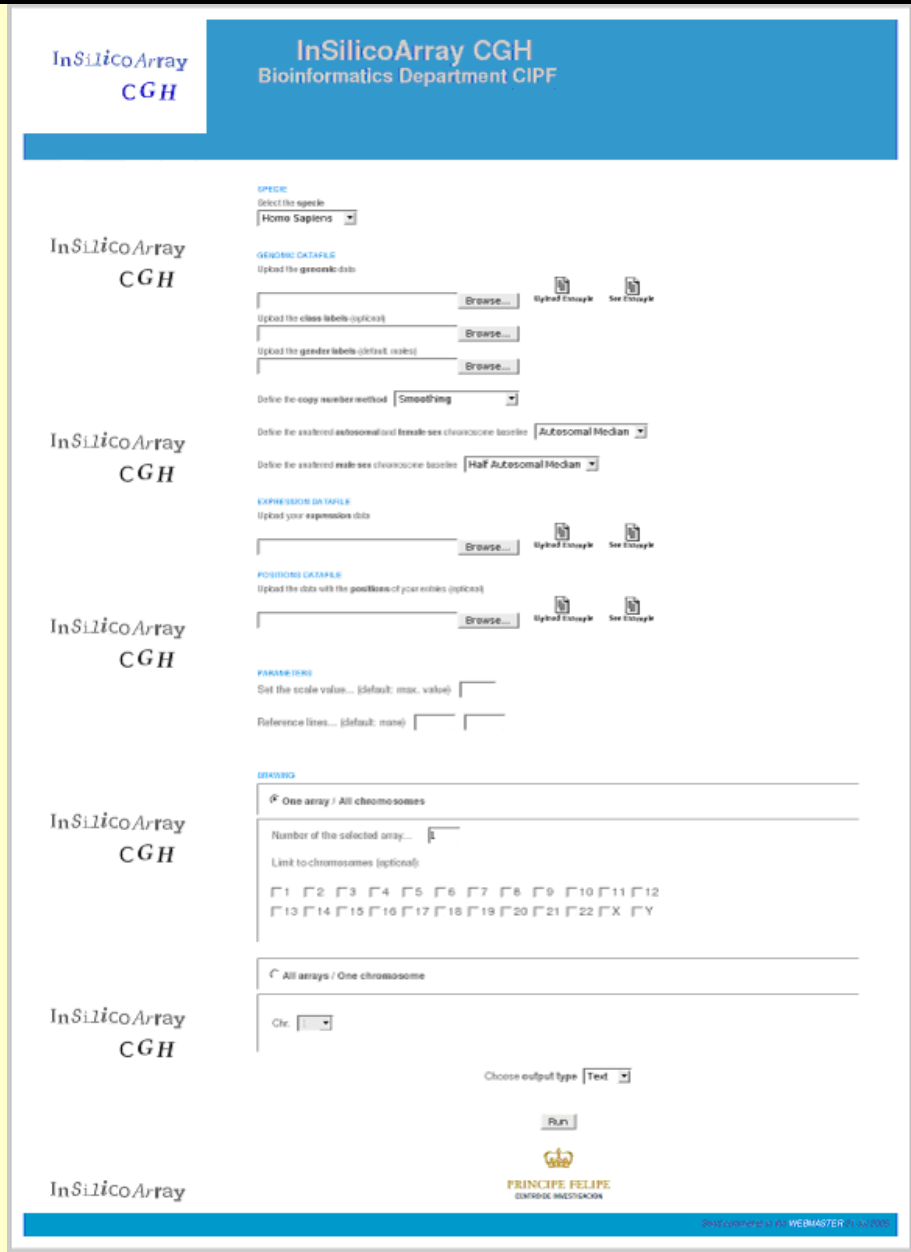
PAGE 2005

Tian 2005

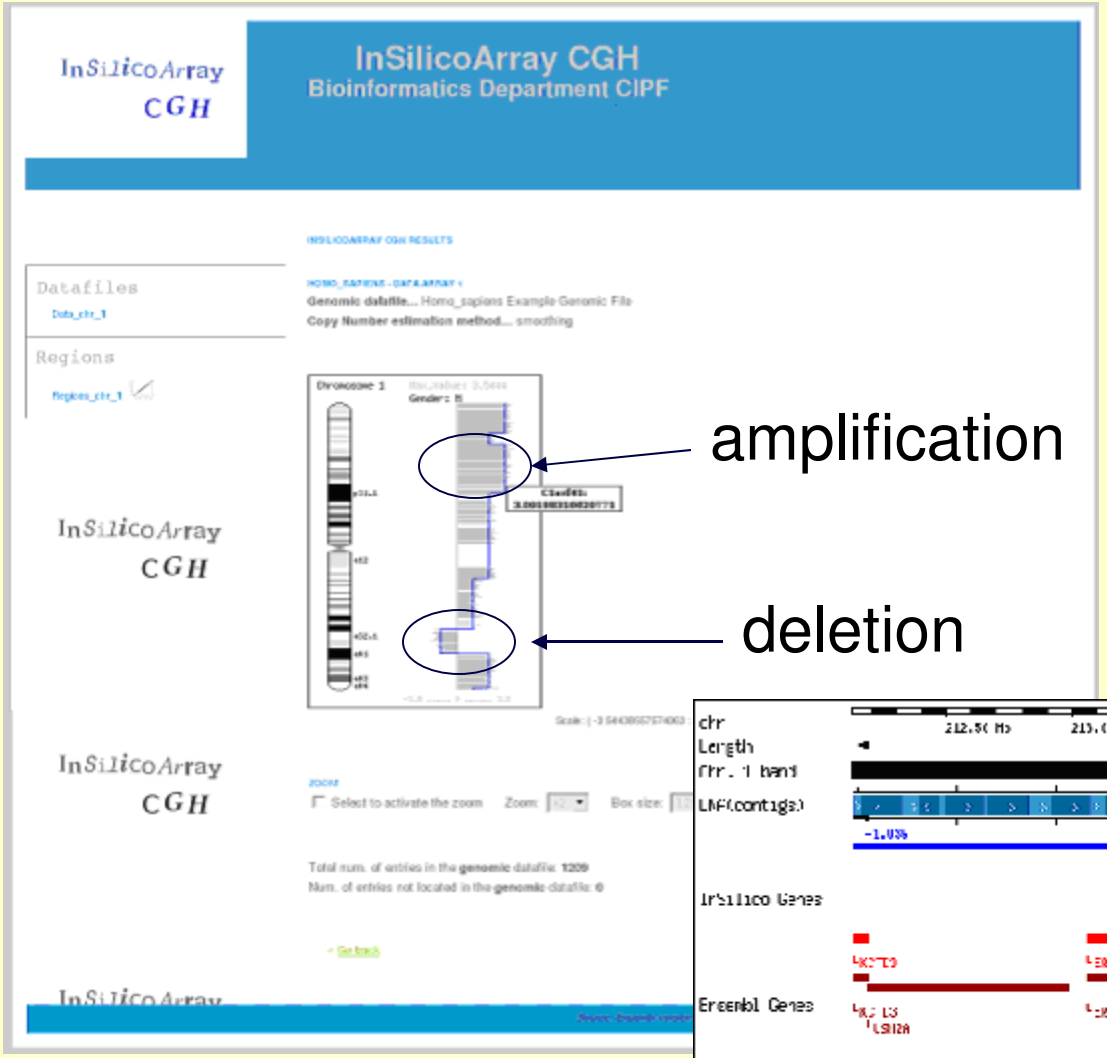
Terms from distinct repositories, reported by different methods in the diabetes dataset (Mootha et al., 2003)

ISA CGH (*In silico* Array-CGH)

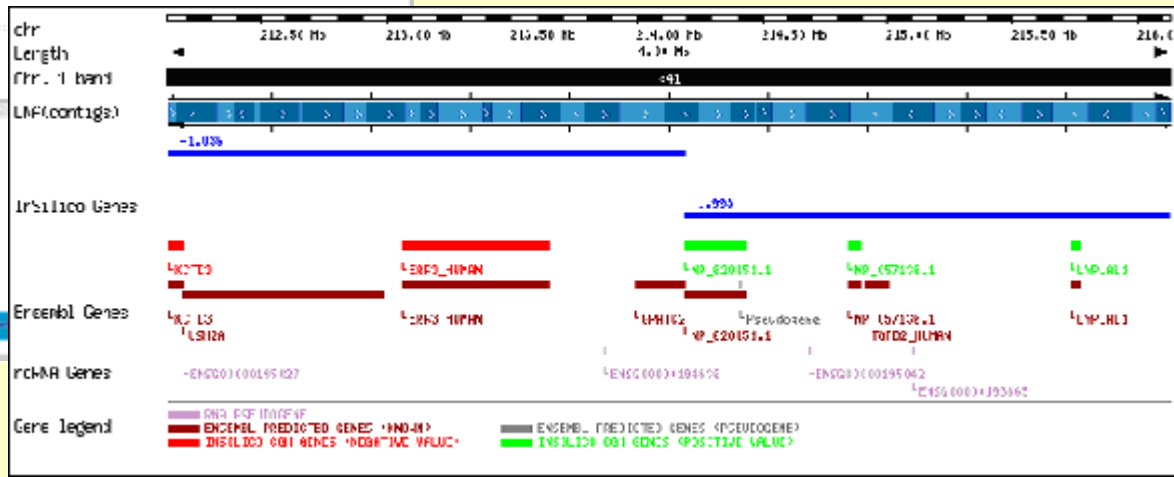
- Estimating copy number variation
- Correlation copy number – expression
- Minimum common amplified / lost region



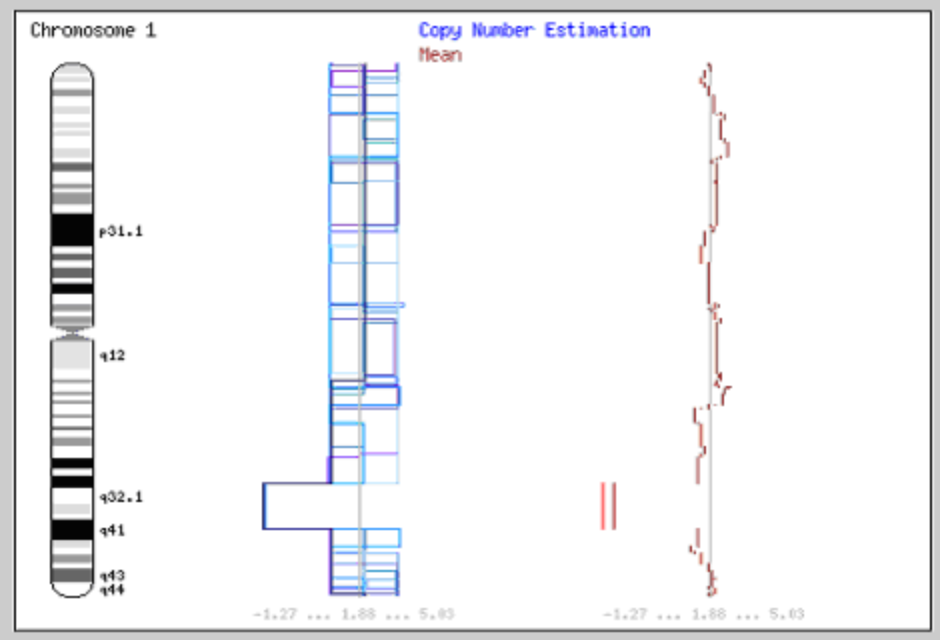
Estimating copy number



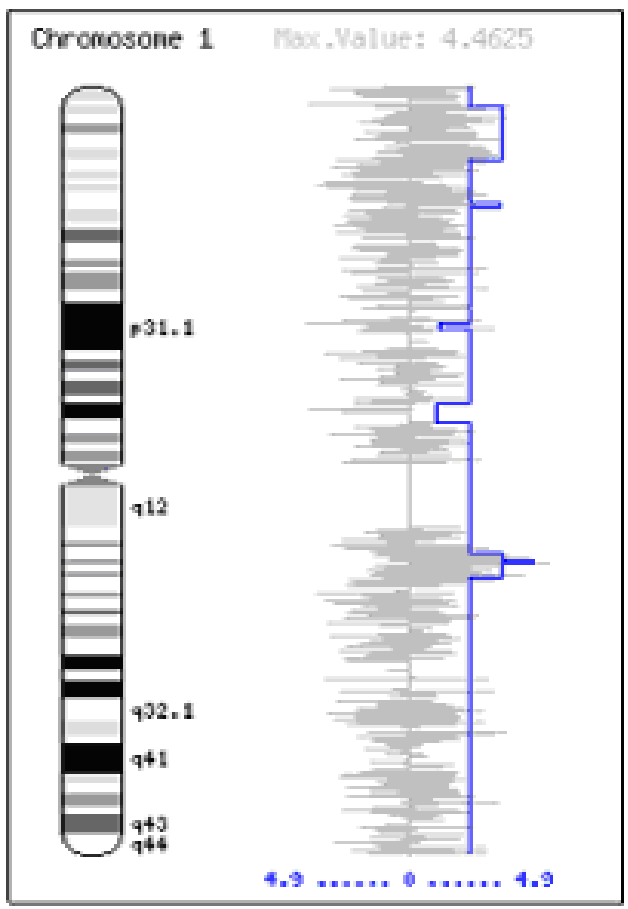
Zoom of the region

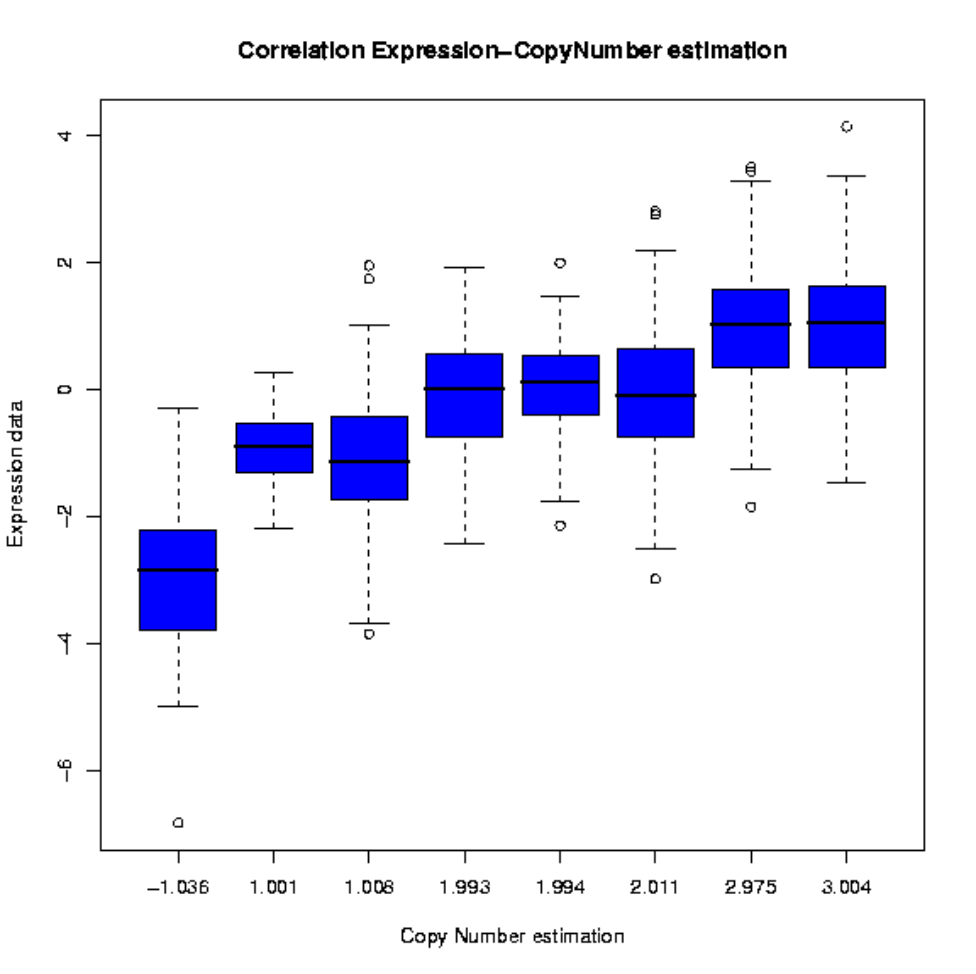


Minimum region with consistent losses or gains

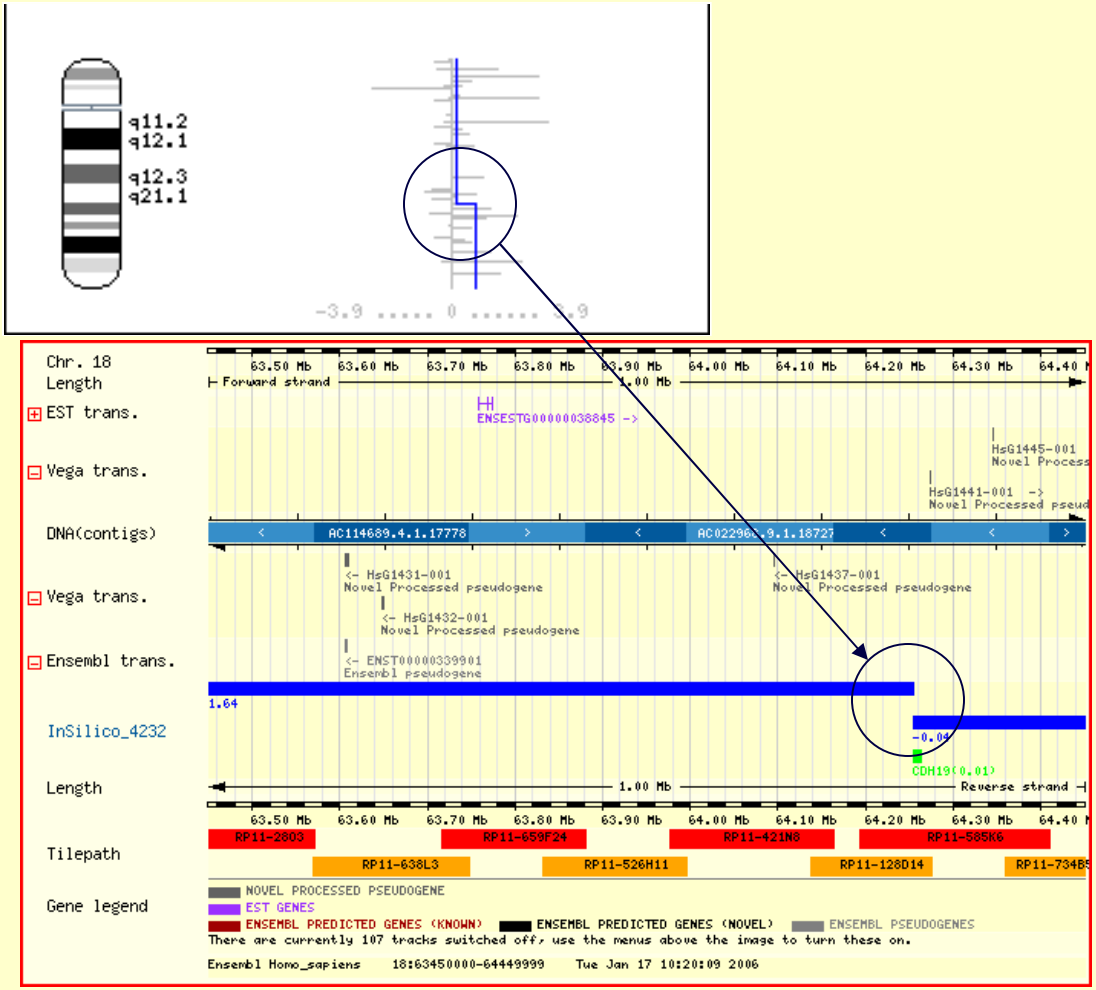


Correlation copy number to expression value





Array-CGH. DAS server



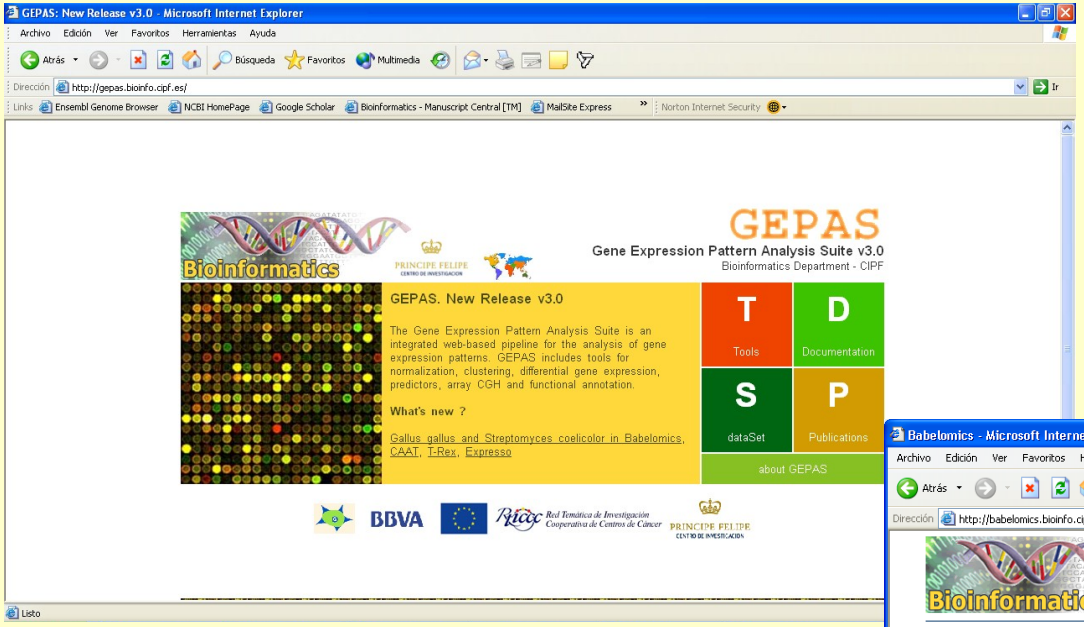
Detection of copy number alterations (several methods)

Relationship expression / copy number alteration

Functional annotation of altered regions

DAS server

GEPAS & Babelomics



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)

Al-Shahrour et al. Bioinformatics (2004)

Al-Shahrour et al. NAR (2005)

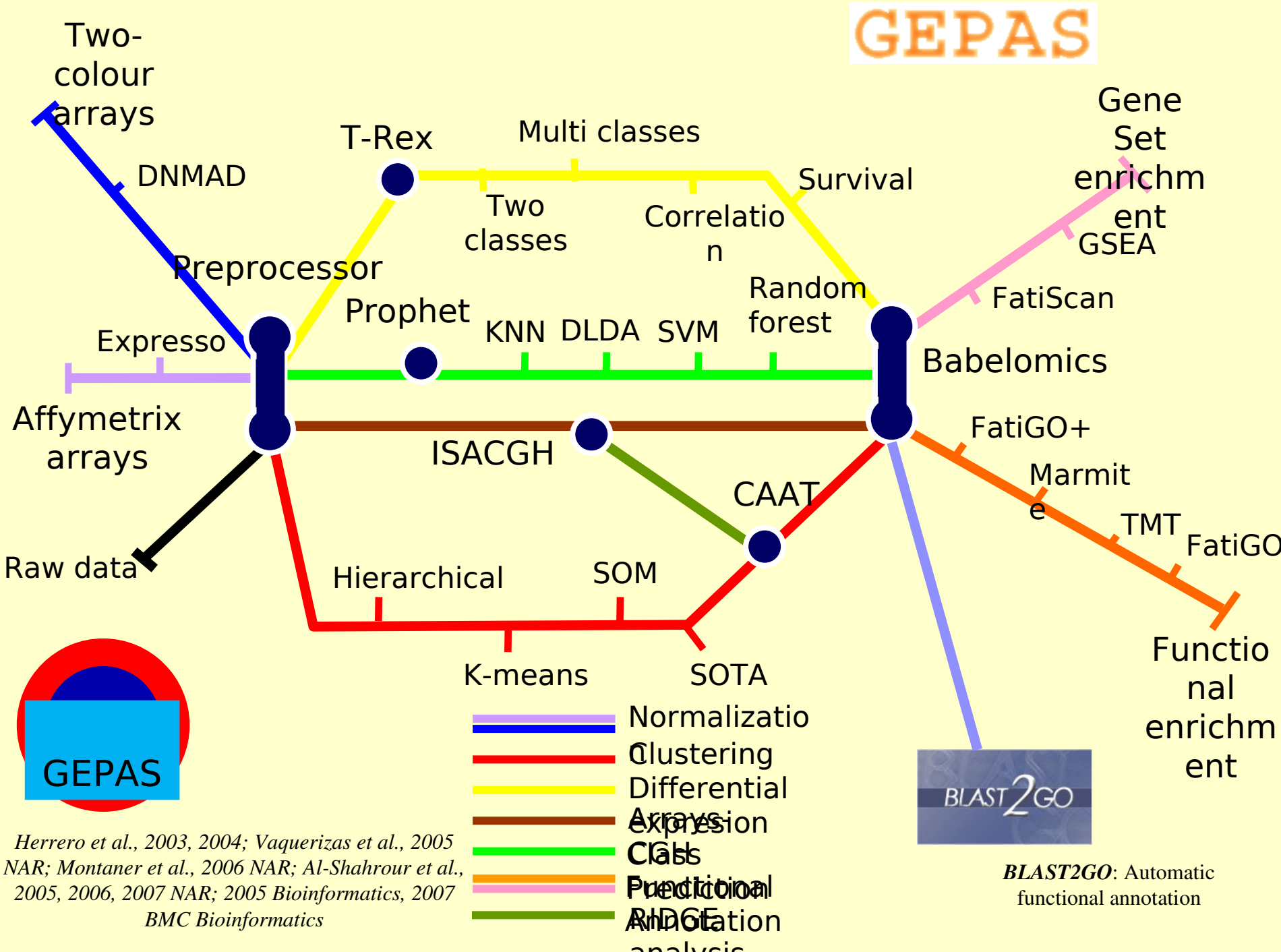
Al-Shahrour et al. Bioinformatics (2005)

Al-Shahrour et al. NAR (2006)

Dopazo OMICS (2006)

Al-Shahrour et al. NAR (2007)

Al-Shahrour et al. BMC Bioinformatics(2007)


GEPAS
Two-colour arrays
DNMAD
T-Rex
Multi classes
Two classes
Correlation
Survival
Gene Set enrichment
GSEA
Preprocessor
Prophet
KNN
DLDA
SVM
Random forest
FatiScan
Expresso
Babelomics
Affymetrix arrays
ISACGH
FatiGO+
Marmite
TMT
FatiGO
CAAT
Raw data
Hierarchical
SOM
K-means
SOTA
Functional enrichment
GEPAS
Normalization
Clustering
Differential Expression
Arrays CGH
Class Prediction
Functional Annotation
RIDGE analysis
BLAST2GO
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
BLAST2GO: Automatic functional annotation

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 Bioinformáti... | dotProject | Bioinformatics @ CIPF

Página Herramientas

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Bioinformatics

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Courses

Google Calendar

CIPF, Valencia

(the week of Fallas)

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

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 Ziarsolo

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