MDA VII

Joaquín Dopazo

Department of Bioinformatics and Genomics,
Centro de Investigación Príncipe Felipe (CIPF),
Functional Genomics Node, (INB), and
Bioinformatics Group (CIBERER)
Valencia, Spain.

http://www.babelomics.org
http://bioinfo.cipf.es

Valencia, 21 March 2011

MINISTERIO DE CIENCIA E INNOVACIÓN

F.E.D.E.R.

INB

ciberer

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Who we are

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

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

ciberer

MINISTERIO DE CIENCIA E INNOVACIÓN

F.E.D.E.R.

ProyectoDeméter

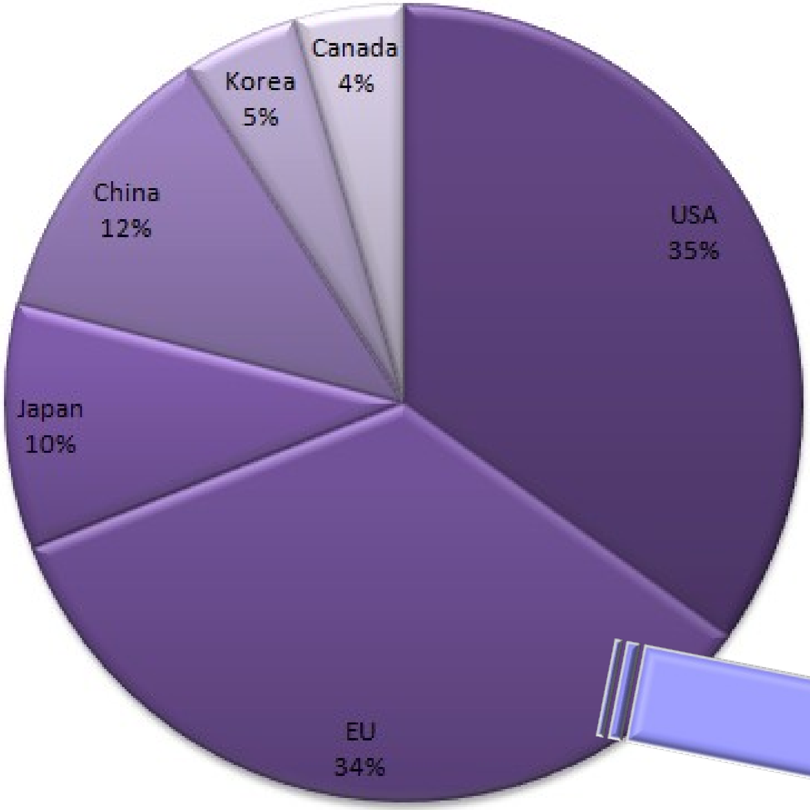
Rticc

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
AFFILIATED REGIONAL INSTITUT

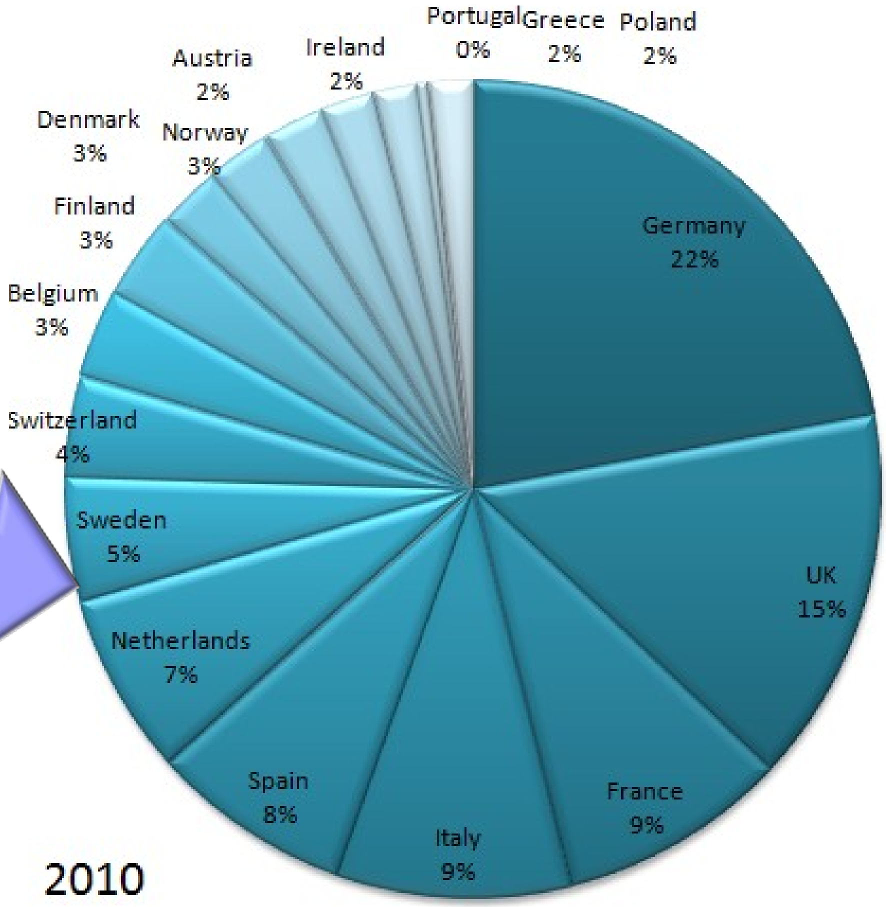
Some bibliographic data Microarray publications



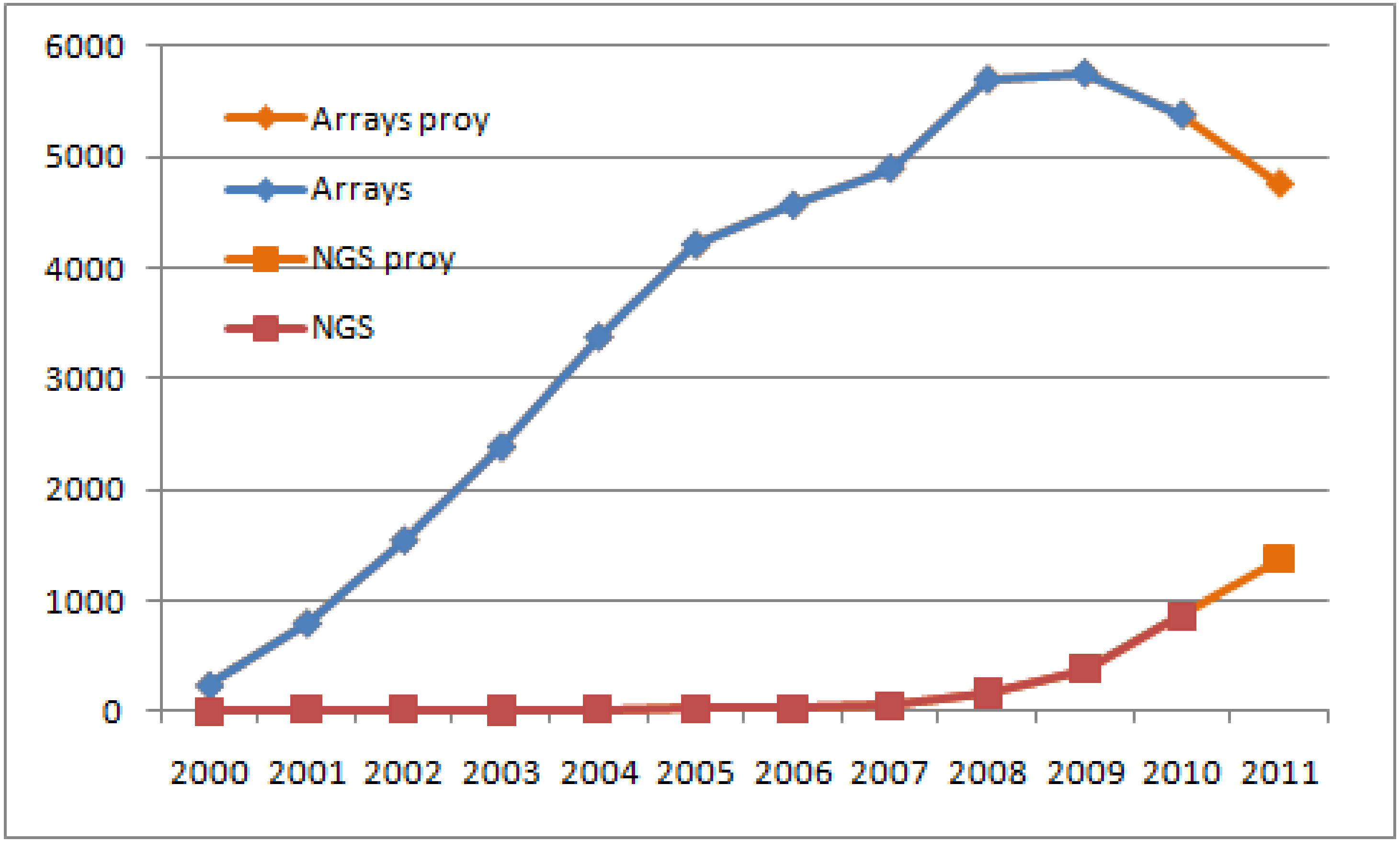
2010 Worlwide

Source Pubmed. Query: 2009[Entrez Date] AND country[Affiliation]AND microarray[Title/Abstract]

2010 Europe



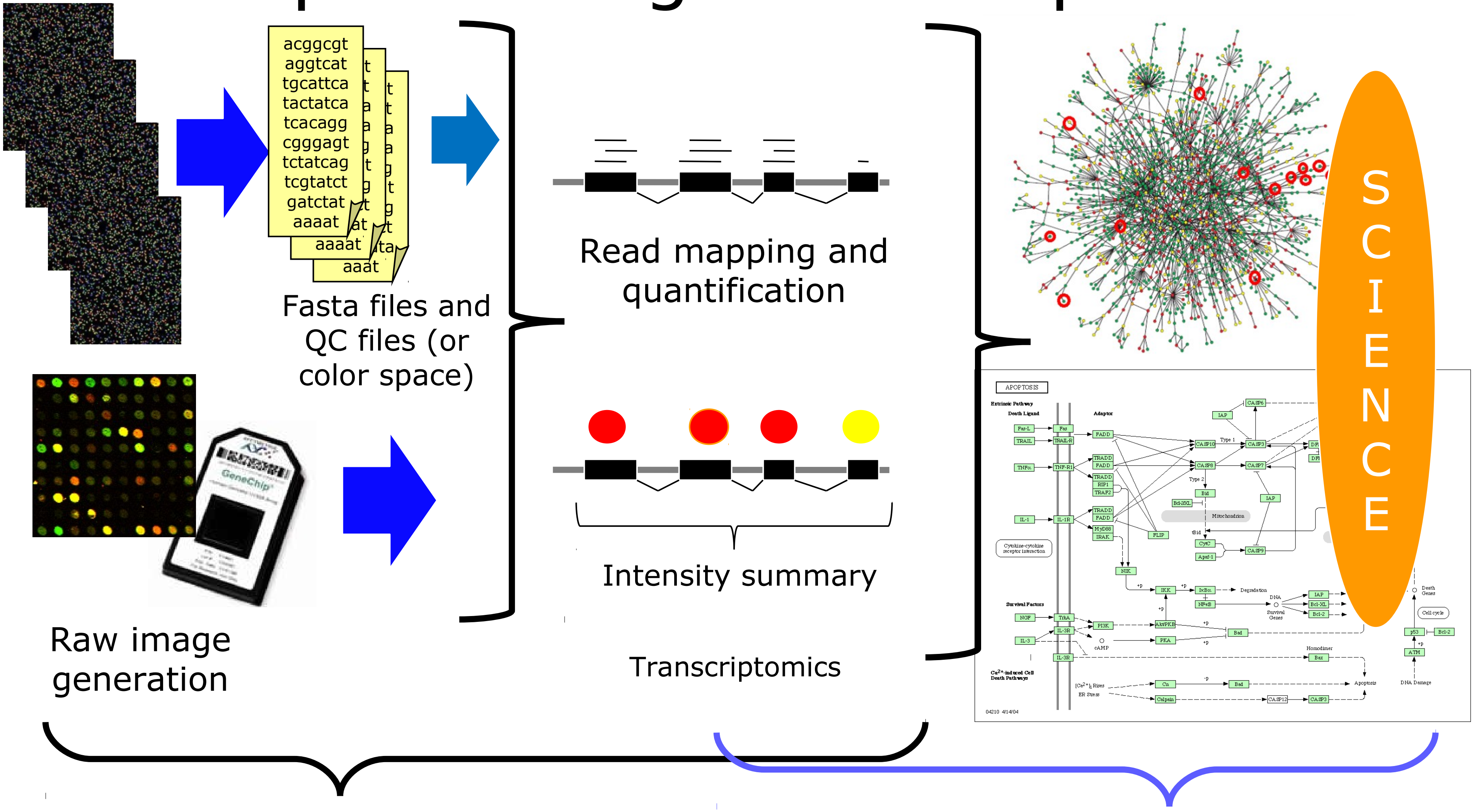
Evolution of the papers published in microarray and next gen technologies



Source Pubmed. Query: "high-throughput sequencing"[Title/Abstract] OR "next generation sequencing"[Title/Abstract] OR "rna seq"[Title/Abstract]) AND year[Publication Date]

Projections 2011 based on January and February

Genomic data, the double challenge: Data processing and interpretation



Raw image generation

Fasta files and QC files (or color space)

Read mapping and quantification

Intensity summary

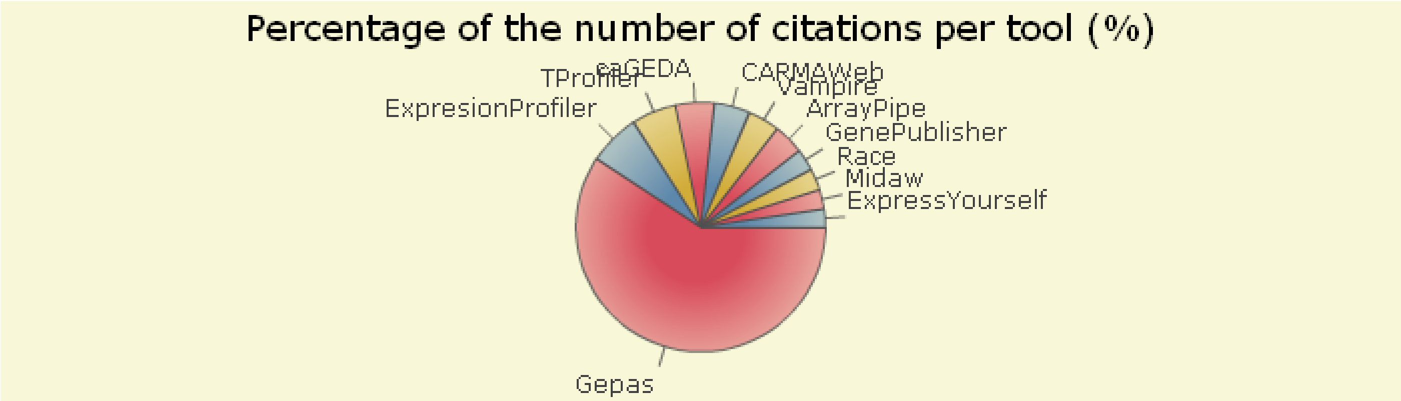
Transcriptomics

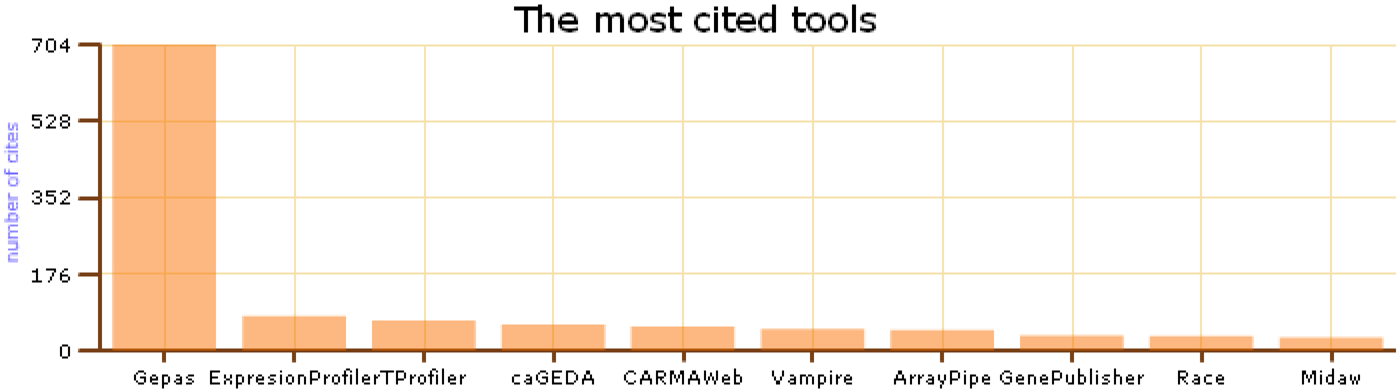
SCIENCE

Technology driven

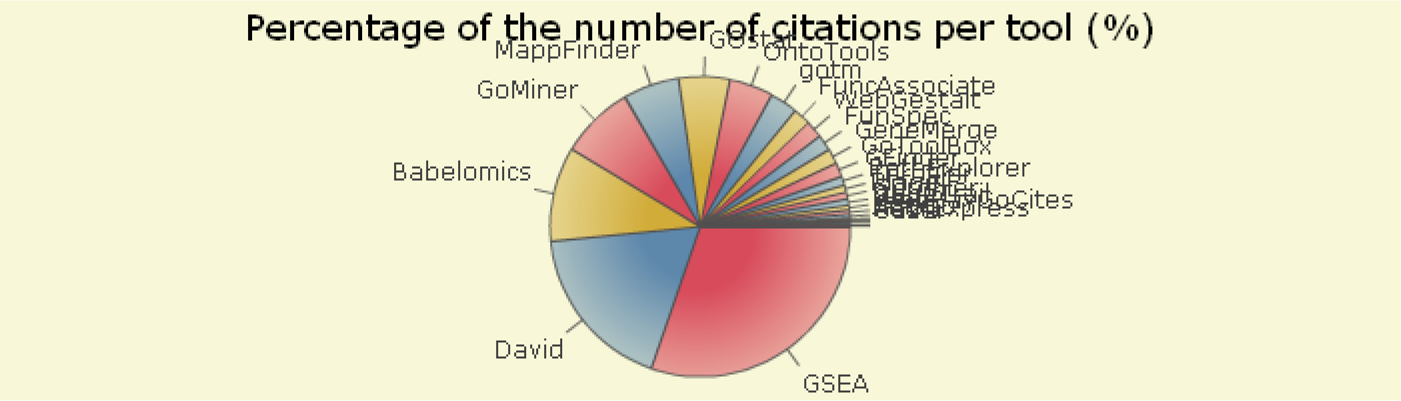
Hypothesis driven

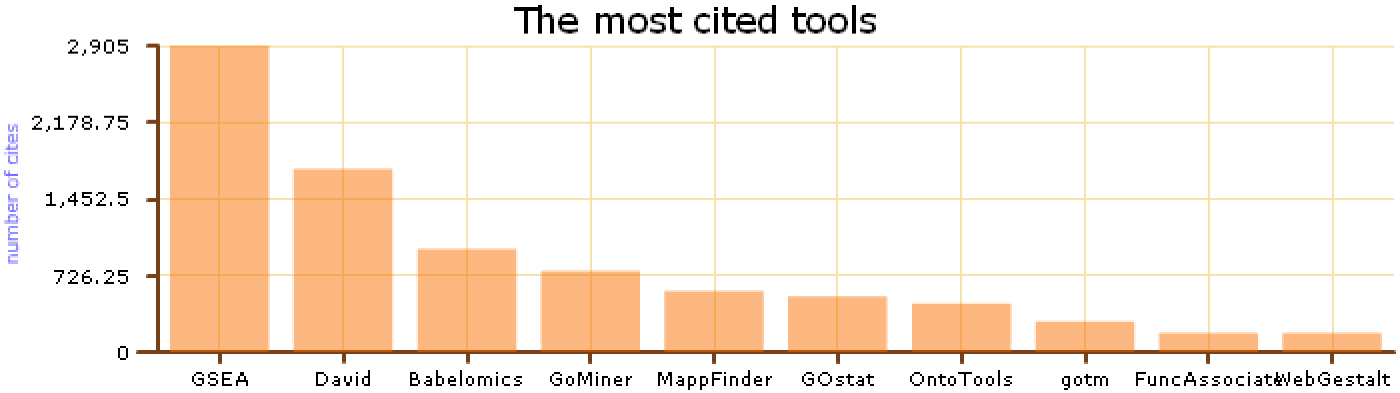
Tools for gene expression analysis





Tools for functional profiling





Some numbers

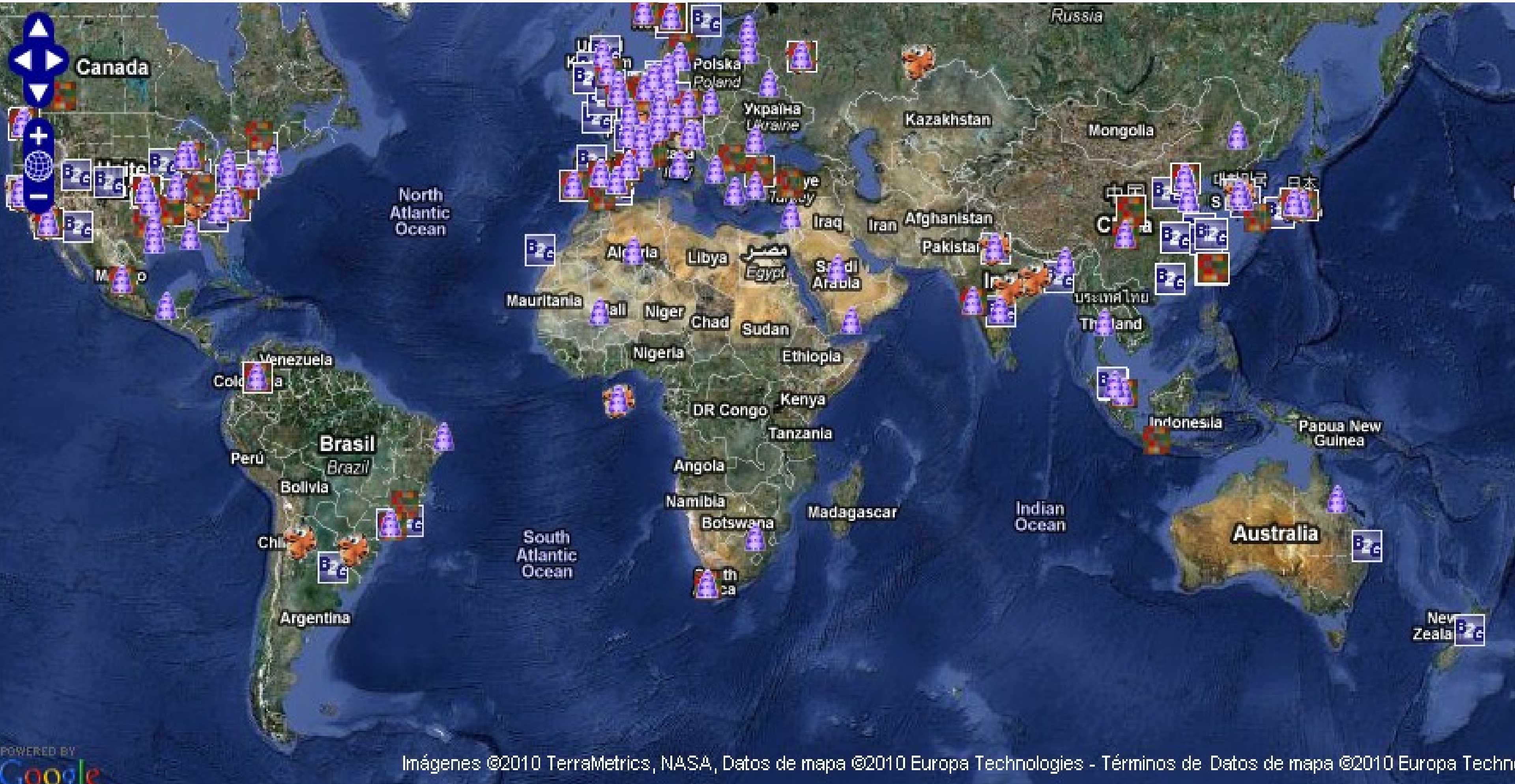
451 papers cite GEPAS (215 are SOTA cites)

632 papers cite Babelomics (442 are FatiGO cites)

(source ISI Web of Knowledge, May 2010)

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

More than 1000 experiments per day.

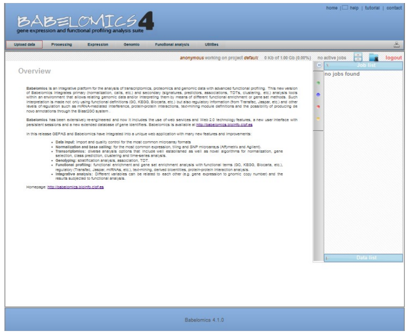


Structure of the course

Theoretical

Monday	Tuesday	Wednesday	Thursday	Friday
Course Reception Course Overview	Transcript Assembly	Statistical Reminder	Differential expression for microarrays	Gene-Set Methodologies
Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
Introduction to Linux Introduction to NGS Technologies	Extracting RNAseq Counts in NGS Studies	Microarray Data Normalization	Predictors	Biological Networks
LUNCH	LUNCH	LUNCH	LUNCH	LUNCH
NGS Data Preprocessing	Differential Expression in RNAseq Studies	Genomic SNP data analysis	Clustering Methods	Closing
Coffee break	Coffee break	Coffee break	Coffee break	
NGS Read Mapping	Functional annotation	Genome Wide Association Studies	Biological Databases	

Theoretical and Hands-on:



Overview

Babelomics is an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. This new version of Babelomics integrates primary (normalization, calls, etc.) and secondary (signatures, predictors, associations, TDTs, clustering, etc.) analysis tools within an environment that allows relating genomic data and/or interpreting them by means of different functional enrichment or gene set methods. Such interpretation is made not only using functional definitions (GO, KEGG, Biocarta, etc.) but also regulatory information (from Transfac, Jaspar, etc.) and other levels of regulation such as miRNA-mediated interference, protein-protein interactions, text-mining module definitions and the possibility of producing de novo annotations through the Blast2GO system .

Babelomics has been extensively re-engineered and now it includes the use of web services and Web 2.0 technology features, a new user interface with persistent sessions and a new extended database of gene identifiers. Babelomics is available at http://babelomics.bioinfo.cipf.es

In this release GEPAS and Babelomics have integrated into a unique web application with many new features and improvements:

- **Data input:** import and quality control for the most common microarray formats
- **Normalization and base calling:** for the most common expression, tiling and SNP microarrays (Affymetrix and Agilent).
- **Transcriptomics:** diverse analysis options that include well established as well as novel algorithms for normalization, gene selection, class prediction, clustering and time-series analysis.
- **Genotyping:** stratification analysis, association, TDT.
- **Functional profiling:** functional enrichment and gene set enrichment analysis with functional terms (GO, KEGG, Biocarta, etc.), regulatory (Transfac, Jaspar, miRNAs, etc.), text-mining, derived bioentities, protein-protein interaction analysis.
- **Integrative analysis:** Different variables can be related to each other (e.g. gene expression to gnomic copy number) and the results subjected to functional analysis.

Homepage: http://babelomics.bioinfo.cipf.es

no jobs found

Babelomics 4.1.0

Background

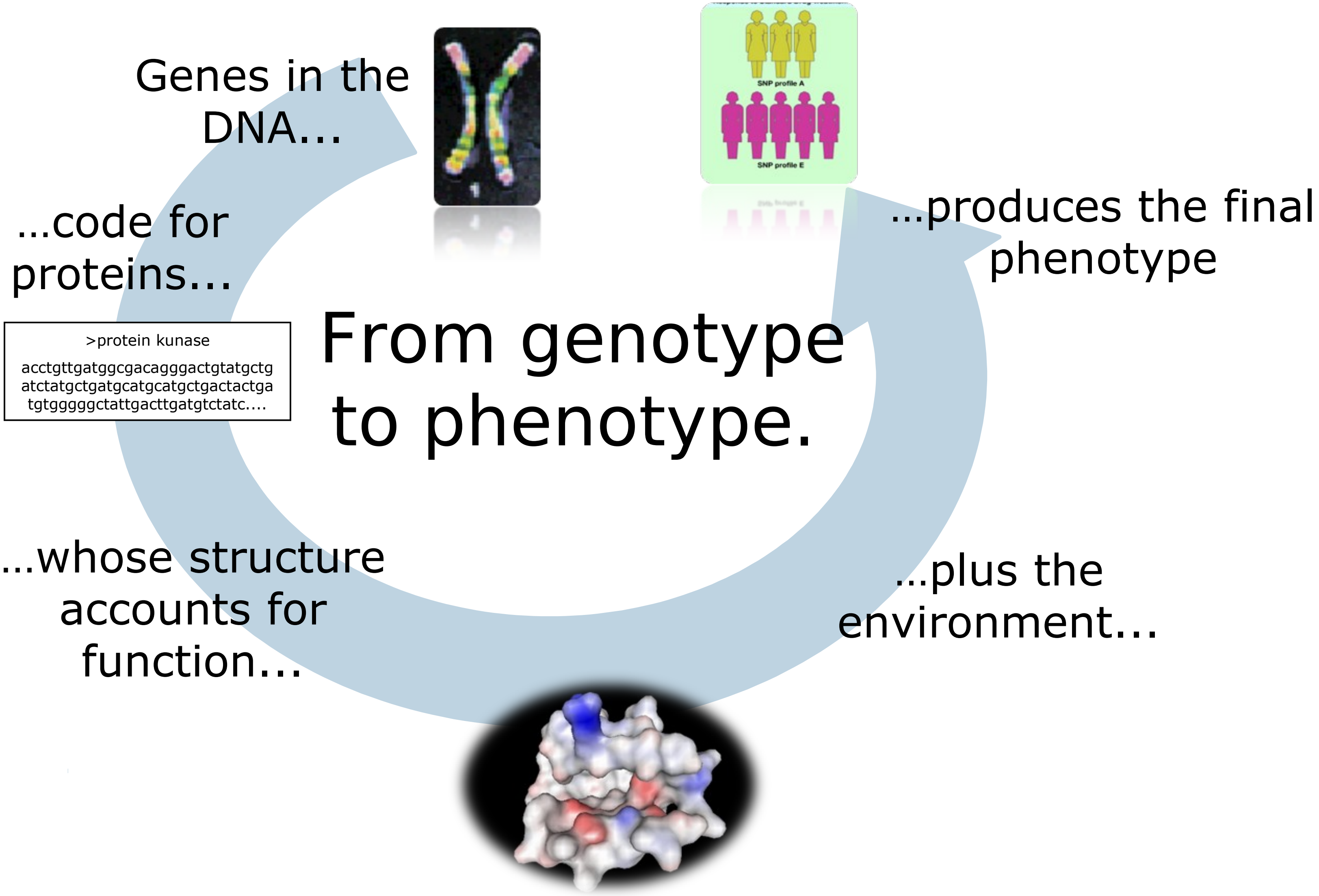
The road of excess leads to the palace of wisdom

(*William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker*)

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…

Where do we come from?
The pre-genomics paradigm



Genes in the DNA...

...code for proteins...

>protein kunase

acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgcatgctgactactga
tgtgggggctattgacttgatgtctatc....

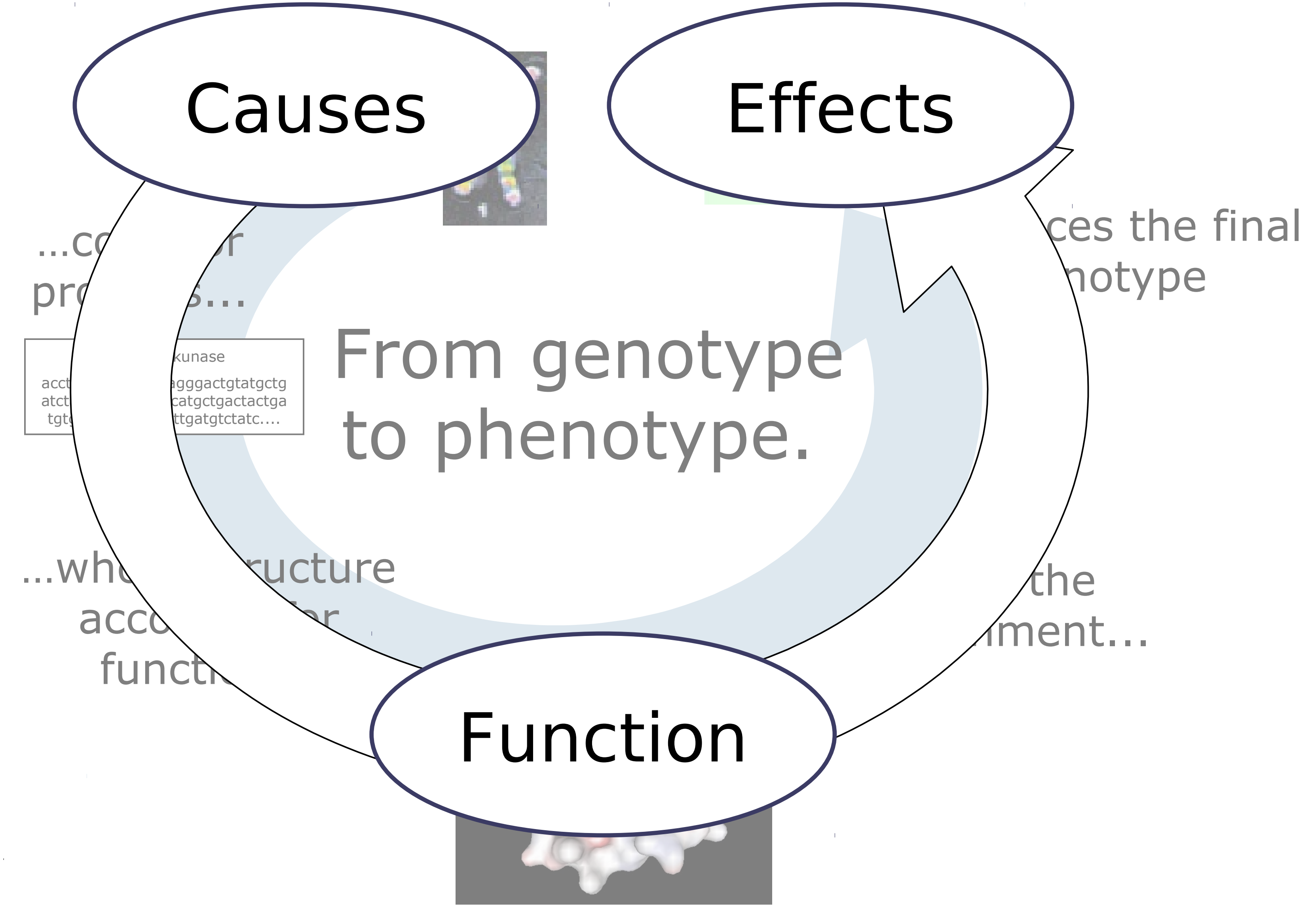
...whose structure accounts for function...

...plus the environment...

...produces the final phenotype

From genotype to phenotype.

Reduccionistic approach to link causes (genome) to effects (phenotype) through actions (function)



From genotype to phenotype.

(in the post-genomics scenario)

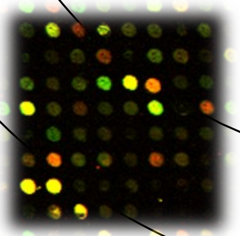
Genes in the DNA...

Next Generation Sequencing
10^9bp per round

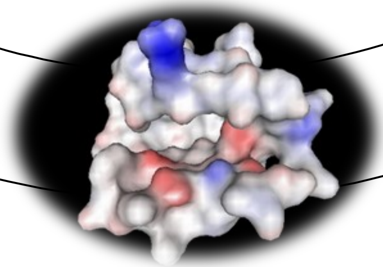
...with its complex variability...

12 million SNPs in exonic regions

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

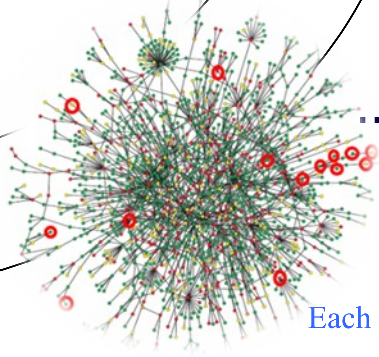


...code for proteins...



That undergo post-translational modifications, somatic recombination...
100K-500K proteins

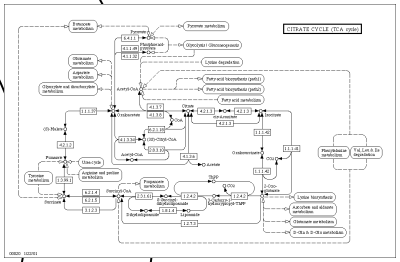
...that account for function if...



...in cooperation with other proteins...

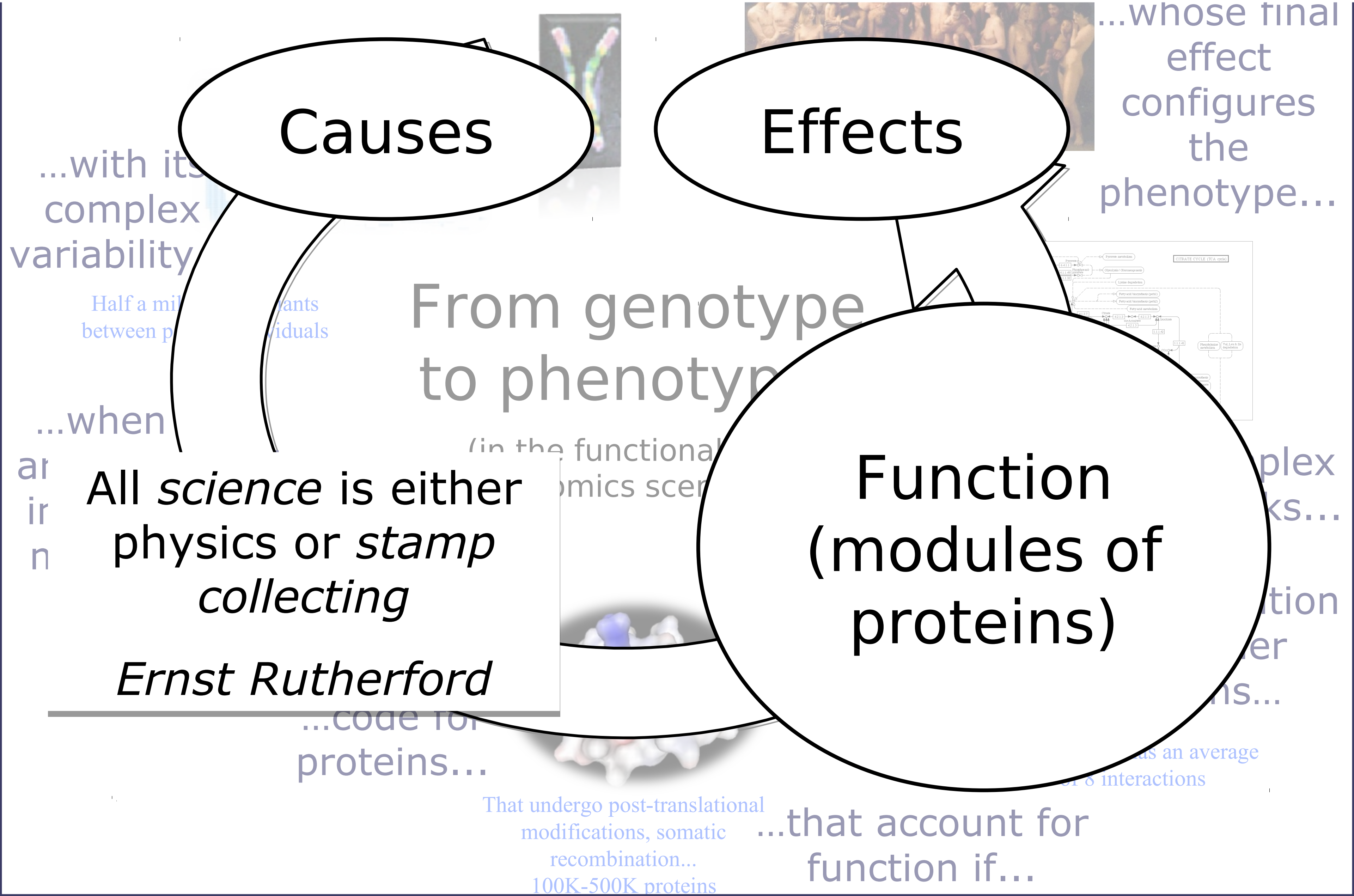
Each protein has an average of 8 interactions

...conforming complex interaction networks...



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

Holistic approach. Causes and effects remain essentially the same. The concept of function has changed



Causes

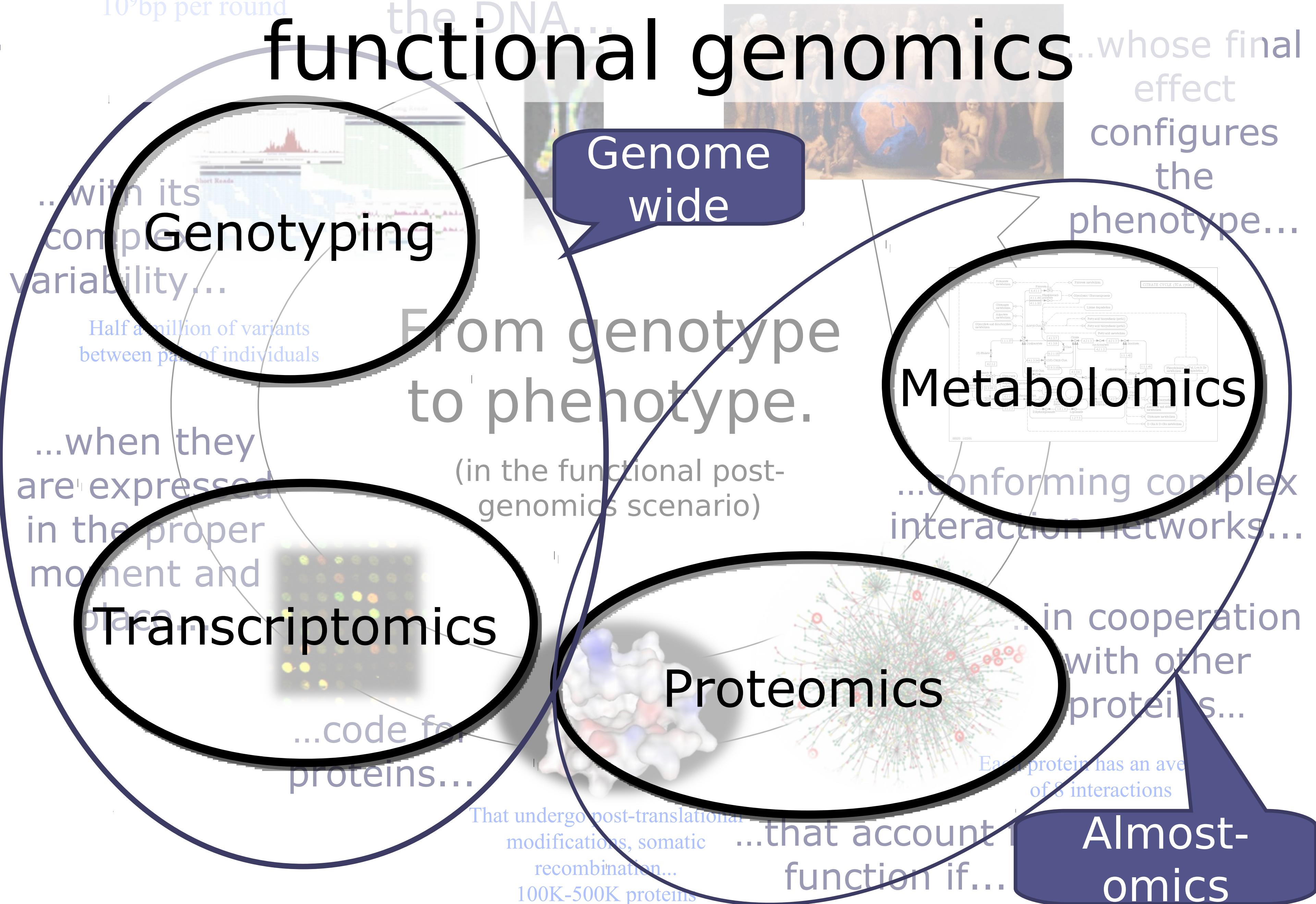
Effects

Function (modules of proteins)

> All *science* is either physics or *stamp collecting*
>
> *Ernst Rutherford*

...whose final effect configures the phenotype...
...with its complex variability
Half a mi ants
between p viduals
From genotype to phenotyp
(in the functional
omics scen
...when
ar
ir
n
plex
ks...
tion
er
ns...
as an average
of 8 interactions
...code for proteins...
That undergo post-translational modifications, somatic recombination...
100K-500K proteins
...that account for function if...

High-throughput data for functional genomics



Next Generation Sequencing

10^9bp per round

Genes in the DNA...

...with its complex variability...

Half a million of variants between pairs of individuals

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

...code for proteins...

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

100K-500K proteins

...that account f

function if...

...in cooperation with other proteins...

Each protein has an ave of 8 interactions

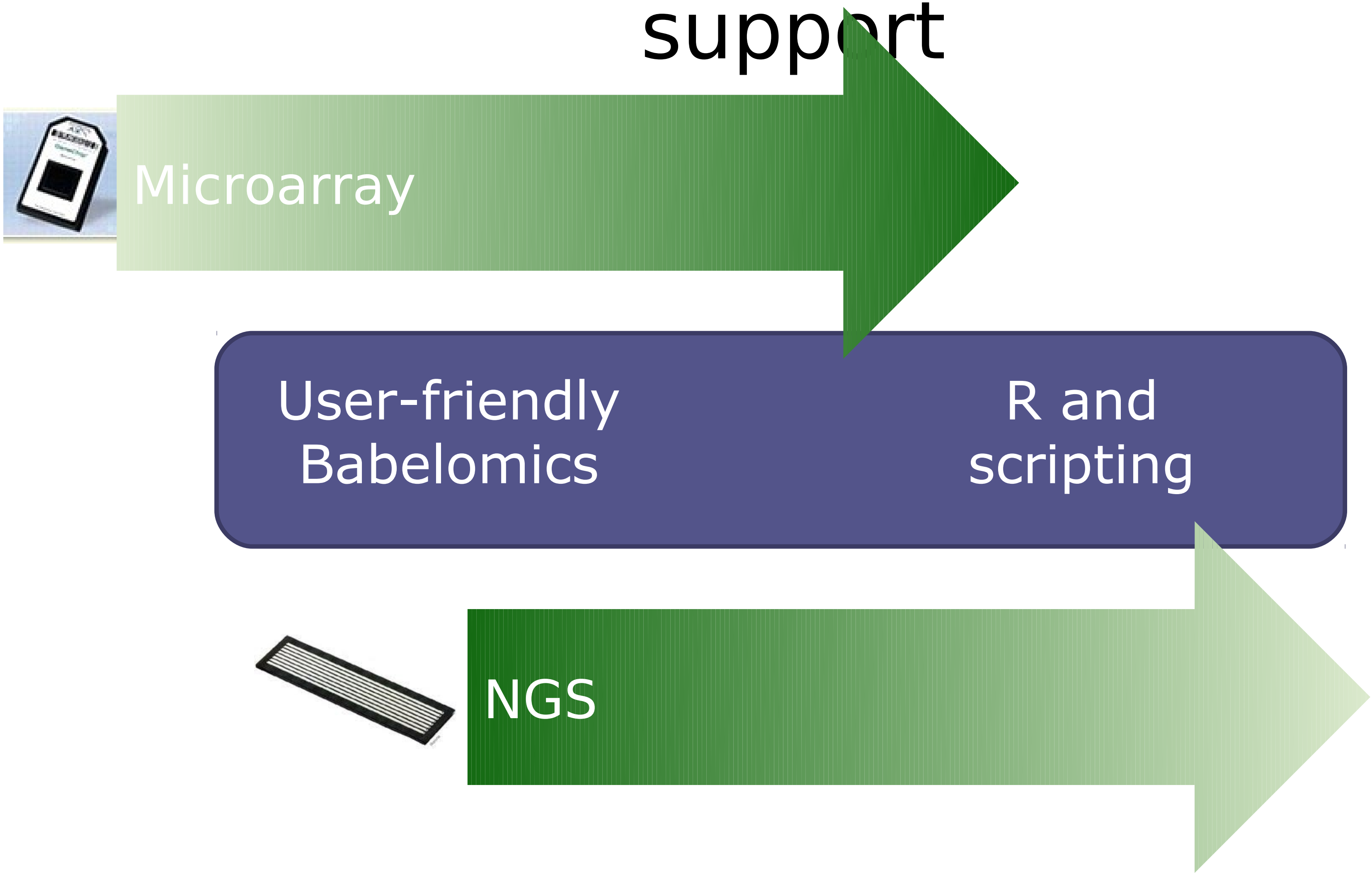
...conforming complex interaction networks...

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

From genotype to phenotype.

(in the functional post-genomics scenario)

Technologies for transcriptomics and genotyping and the corresponding bioinformatics support



Microarray

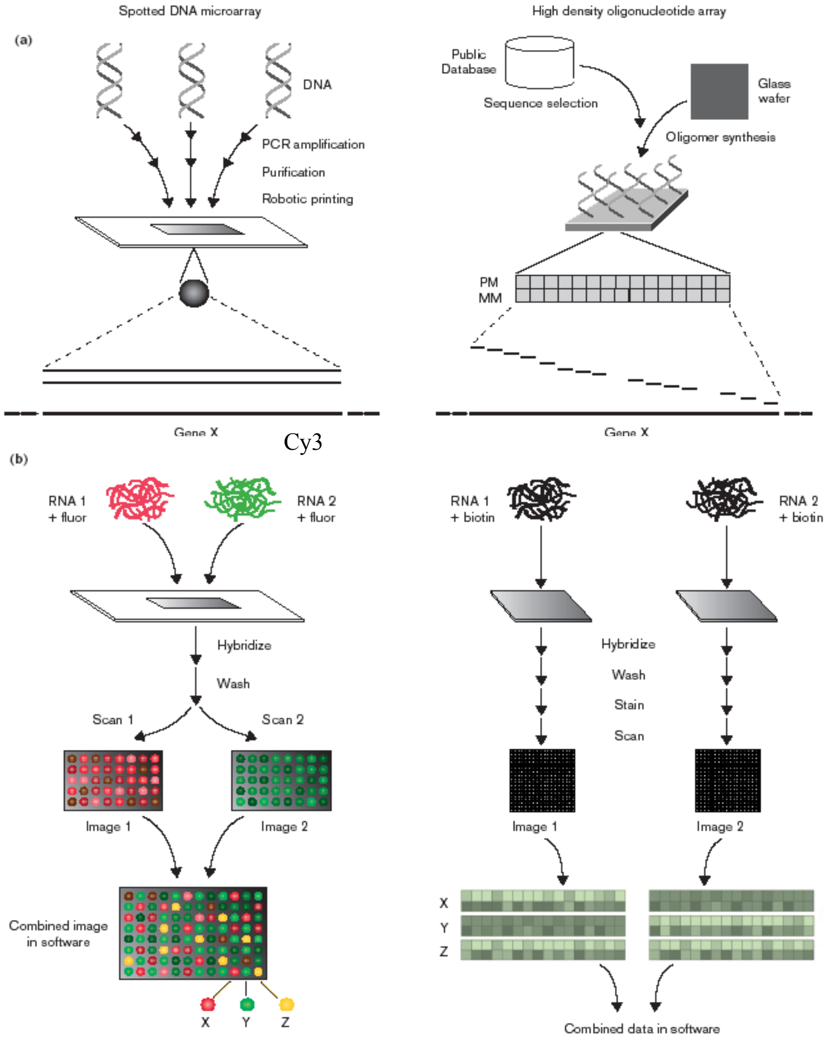
User-friendly Babelomics

R and scripting

NGS

DNA expression microarrays. Strategies of hybridization

Cy5

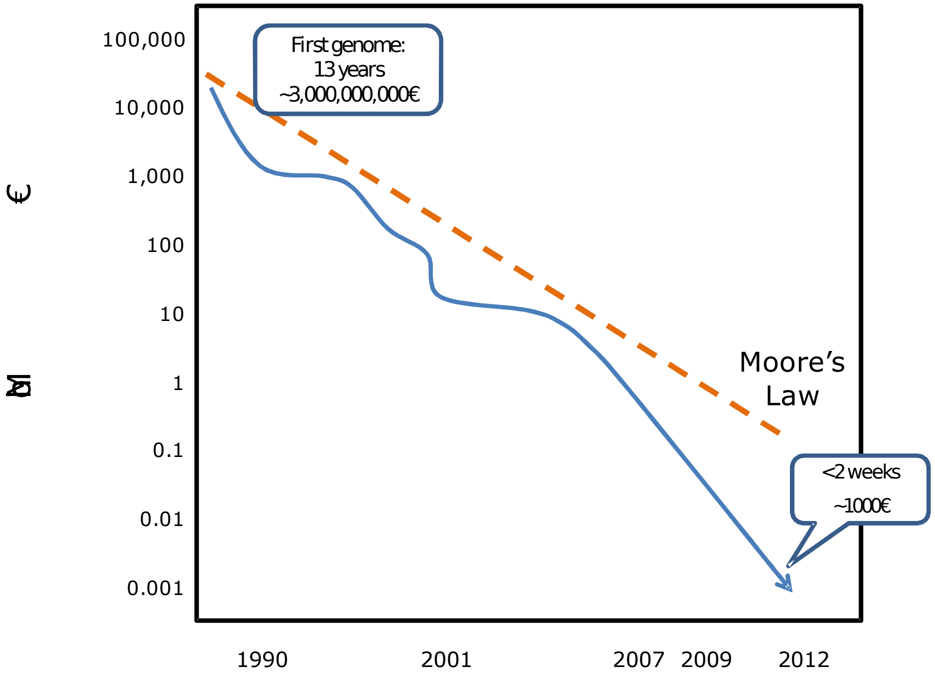


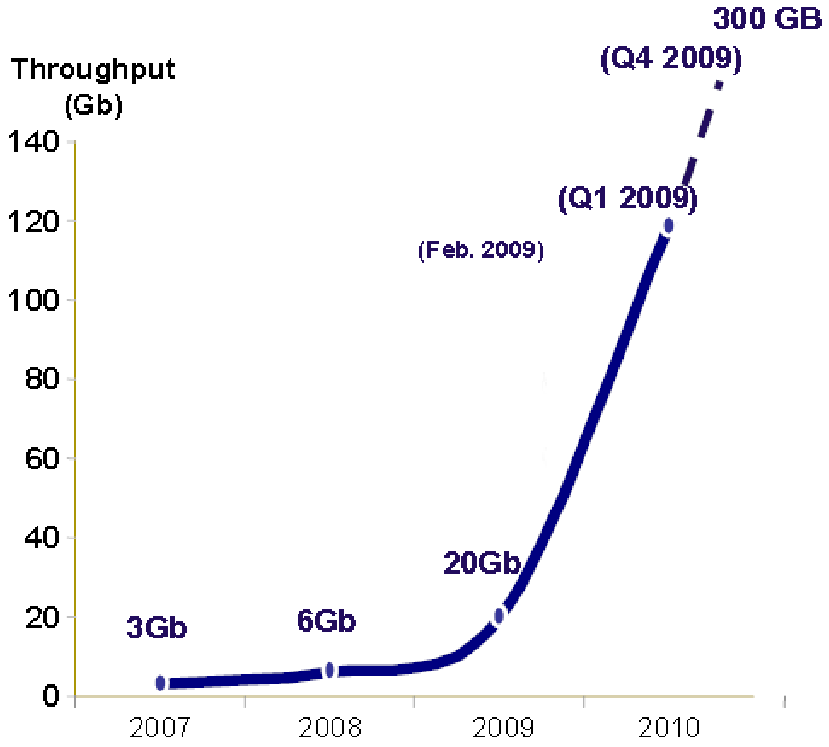
Cy3

Competitive hybridization (two colors)

One color

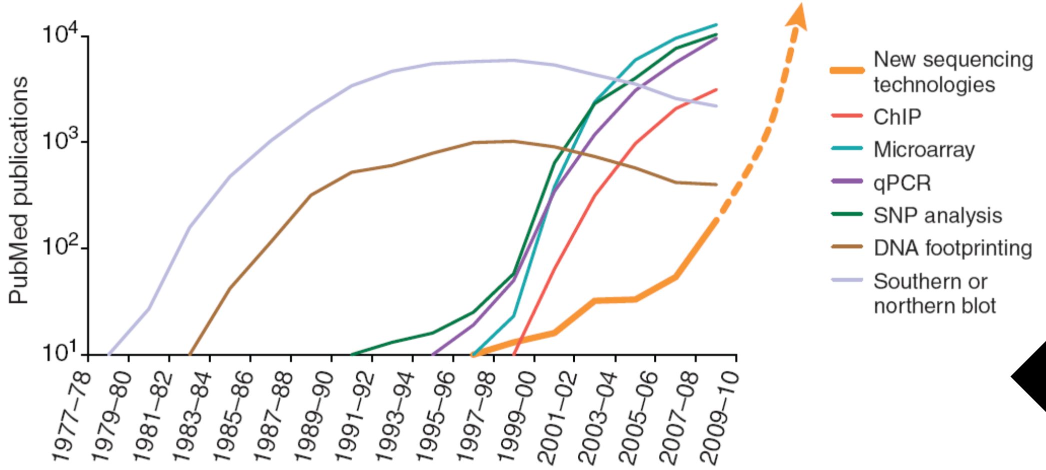
Next generation sequencing technologies are here





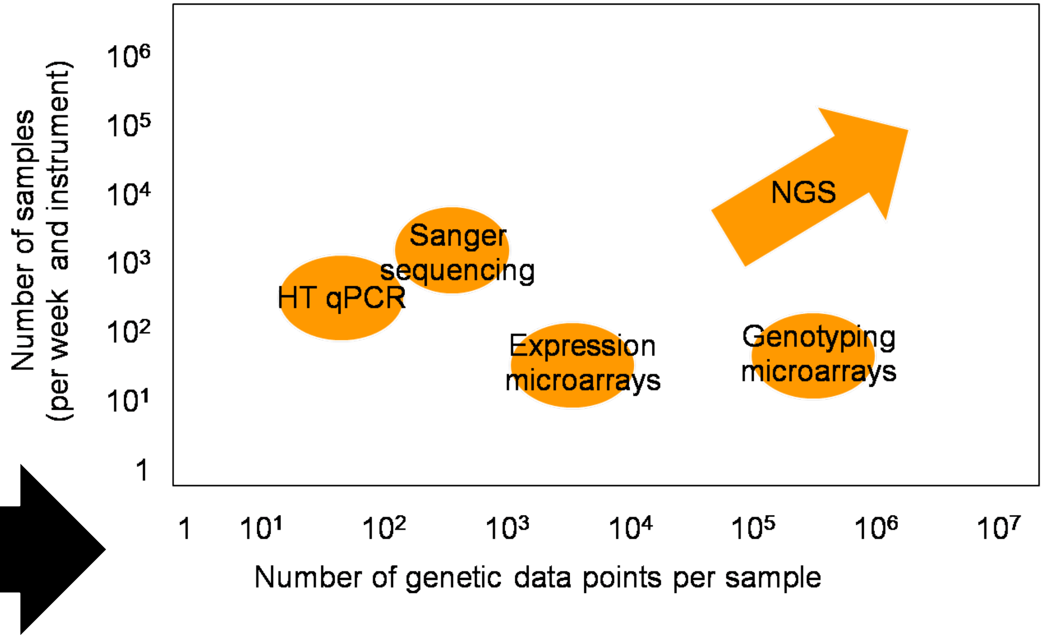
The cost goes down, while the amount of data to manage and its complexity raise exponentially.

Next generation sequencing technologies are here

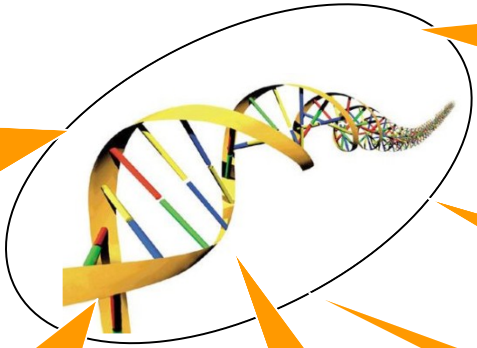


Observed and expected trend of publications in which NGS is being used.

Relative throughput of the different technologies. NGS emerges with a potential of data production that will, eventually wipe out conventional HT technologies in the years coming



Some of the most common applications of NGS



RNA-seq
Transcriptomics:
Quantitative
Descriptive
(alternative
splicing)
miRNA

Resequencing:
Mutation calling
Profiling

De novo
sequencing

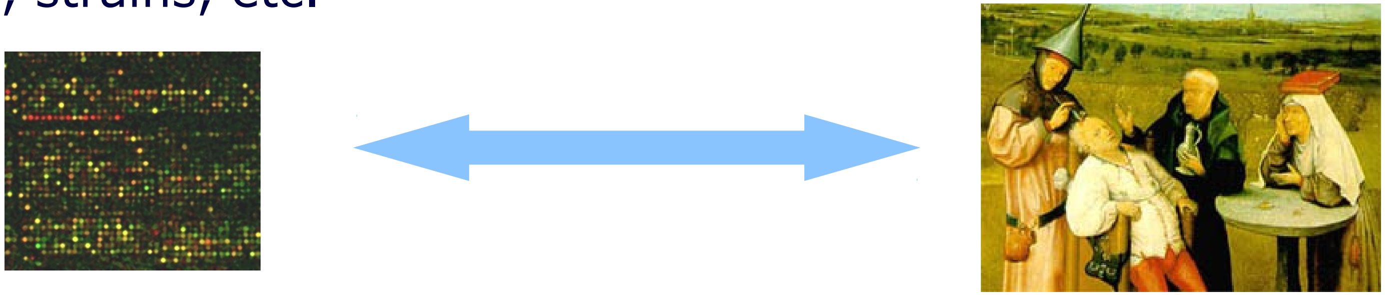
Copy number
variation

Chip-seq
Protein-DNA interactions
Active transcription
factor binding sites

Metagenomics
Metatranscriptomics

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 we 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 we 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)

Primary analysis

- Transform images corresponding to hybridization intensities (microarrays) or to read counts (NGS) into numbers

- Convert all the measurements to a common scale that makes them comparable across experiments.

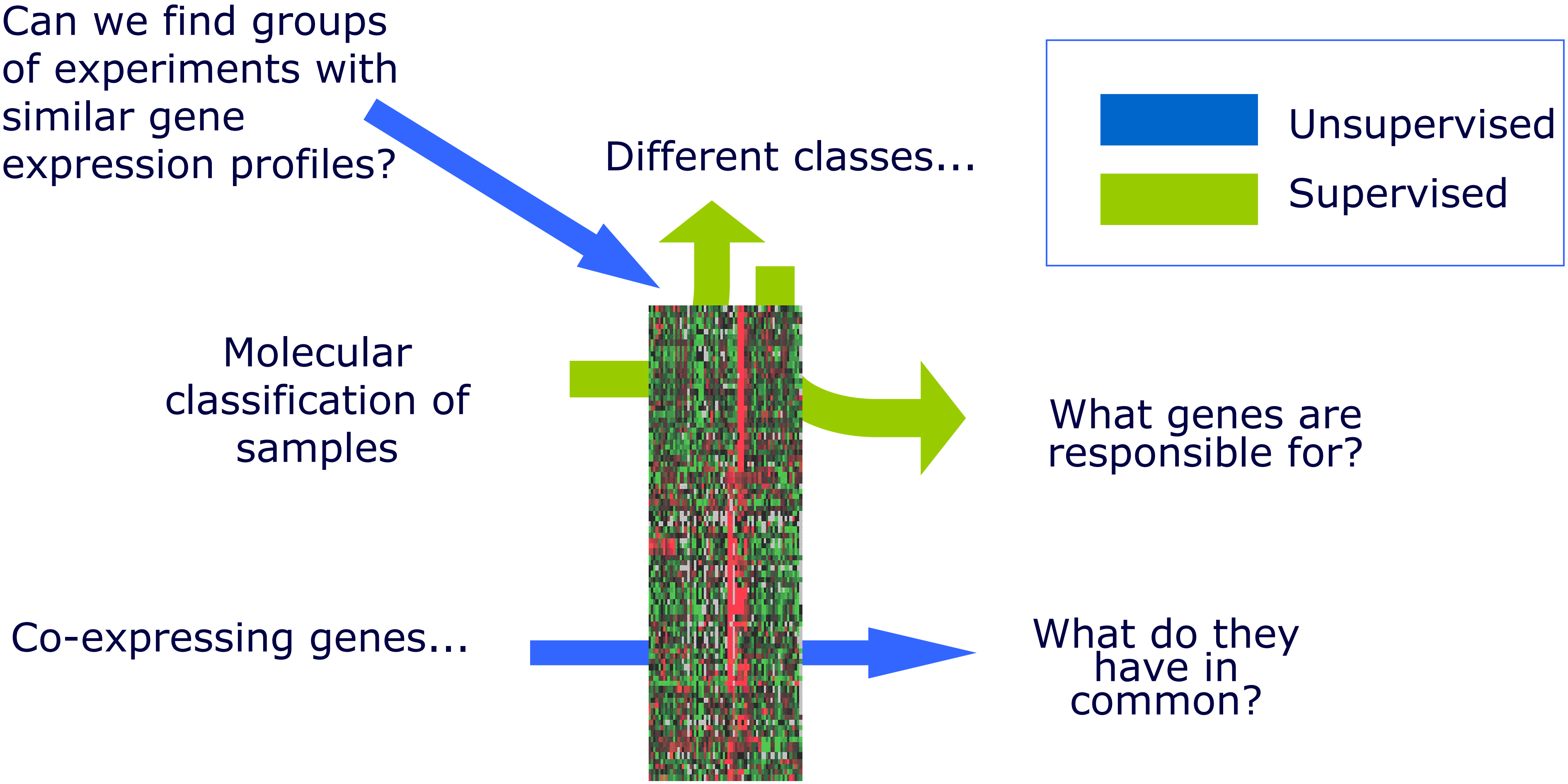
Secondary analysis

Once the measurements are in a common, comparable scale the results can be studied.

Different studies can be made that include class discovery, classification, gene selection, variant calling, etc.

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...

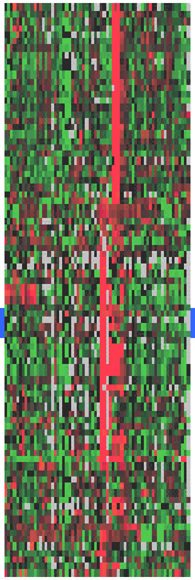
What do they have in common?

Unsupervised problem: class discovery

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

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

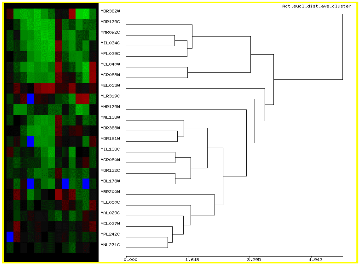
Co-expressing genes...

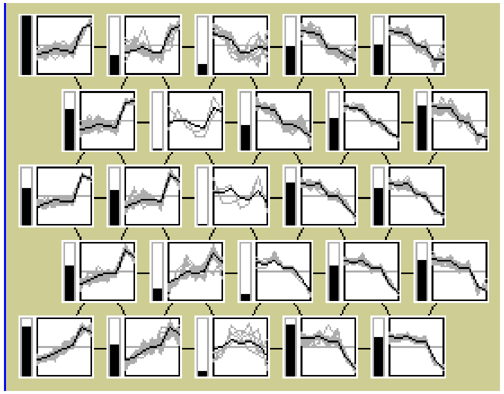


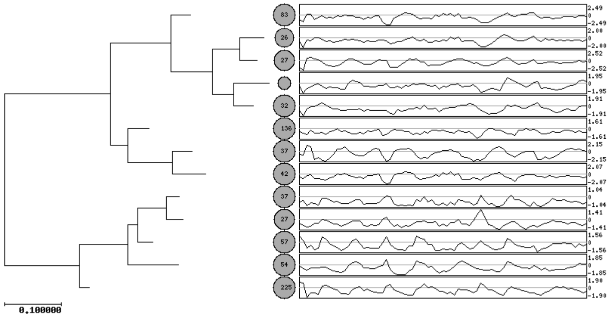
- What genes co-express?
- How many different expression patterns do we have?
- What do they have in common?
- Etc.

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

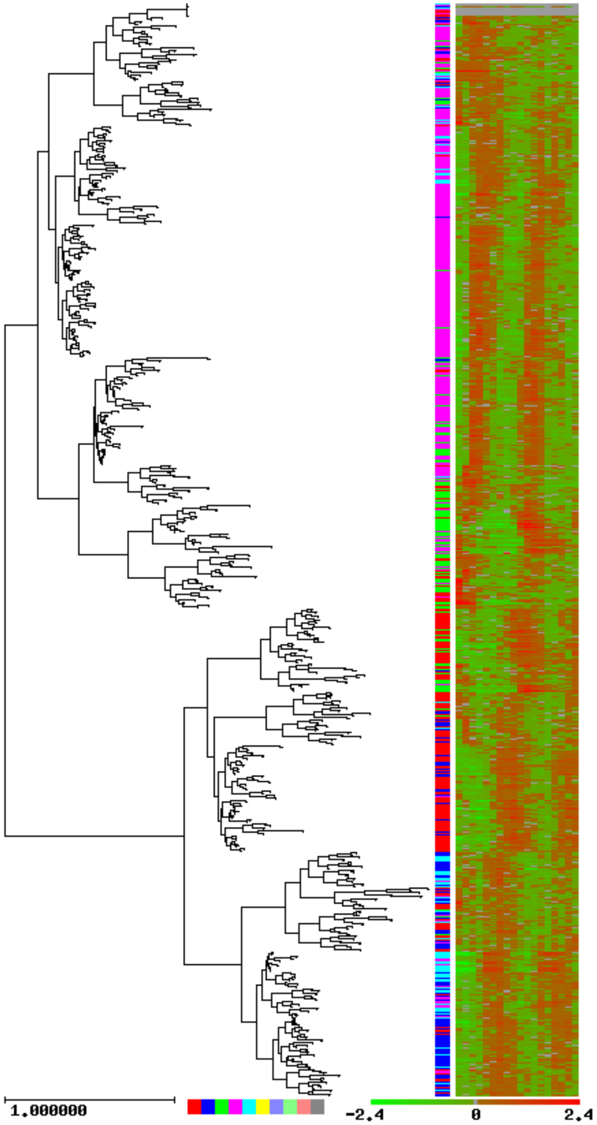
Non hierarchical	hierarchical
K-means, PCA	UPGMA
SOM	SOTA
	Different levels of information







An unsupervised problem: clustering of genes.



- Gene clusters are previously unknown

- Distance function

- Cluster gene expression patterns based uniquely on their similarities.

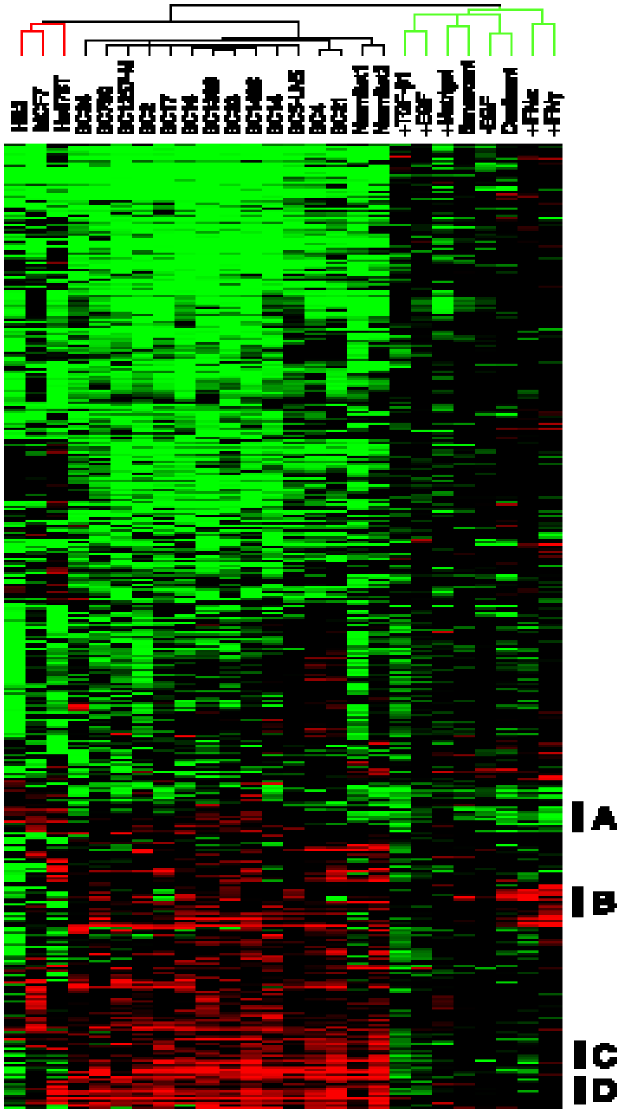
- Results are subjected to further interpretation (if possible)

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)

Clustering of experiments: The problems

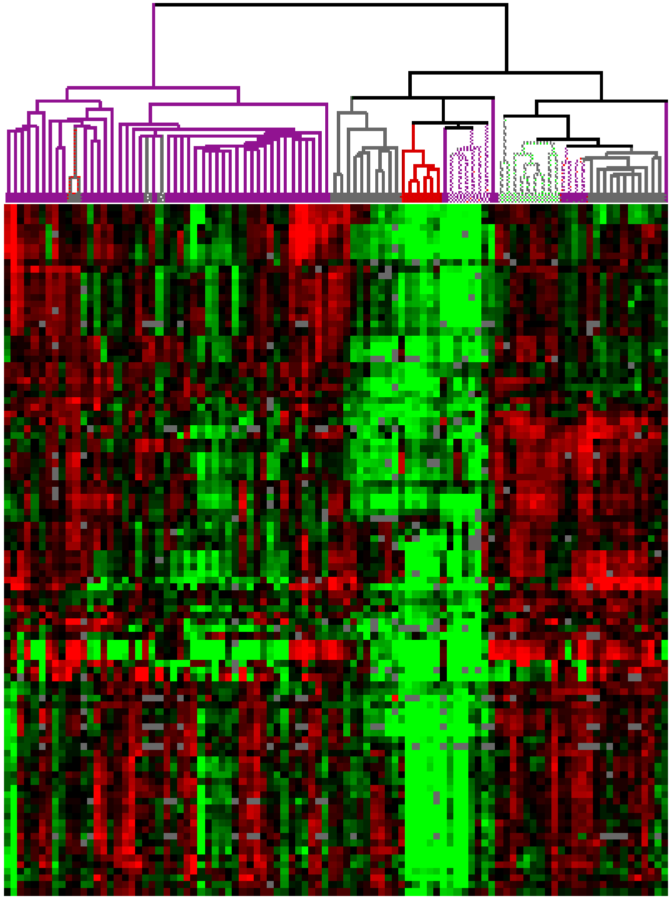
Any gene (regardless its relevance for the classification) has the same weight in the comparison.

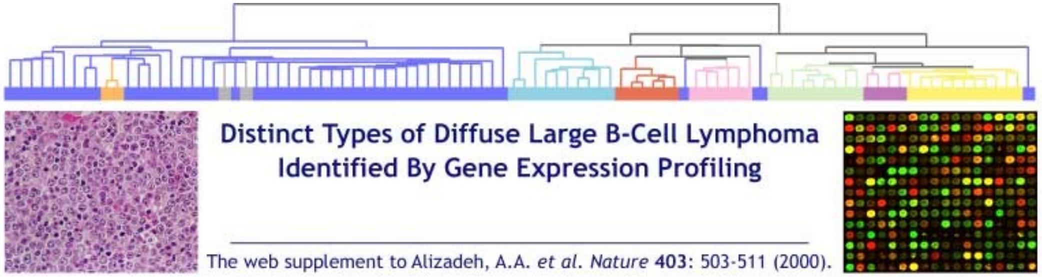
If relevant genes are not in overwhelming majority we will find:

Noise

and/or

irrelevant trends



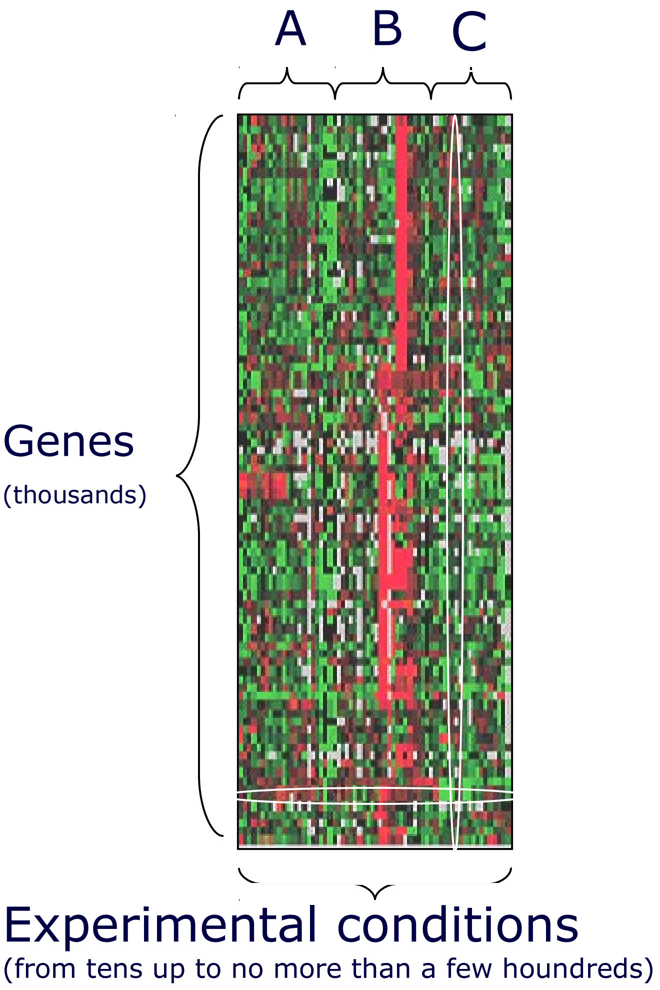


Distinct Types of Diffuse Large B-Cell Lymphoma Identified By Gene Expression Profiling

The web supplement to Alizadeh, A.A. *et al.* *Nature* **403**: 503-511 (2000).

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?

Class prediction

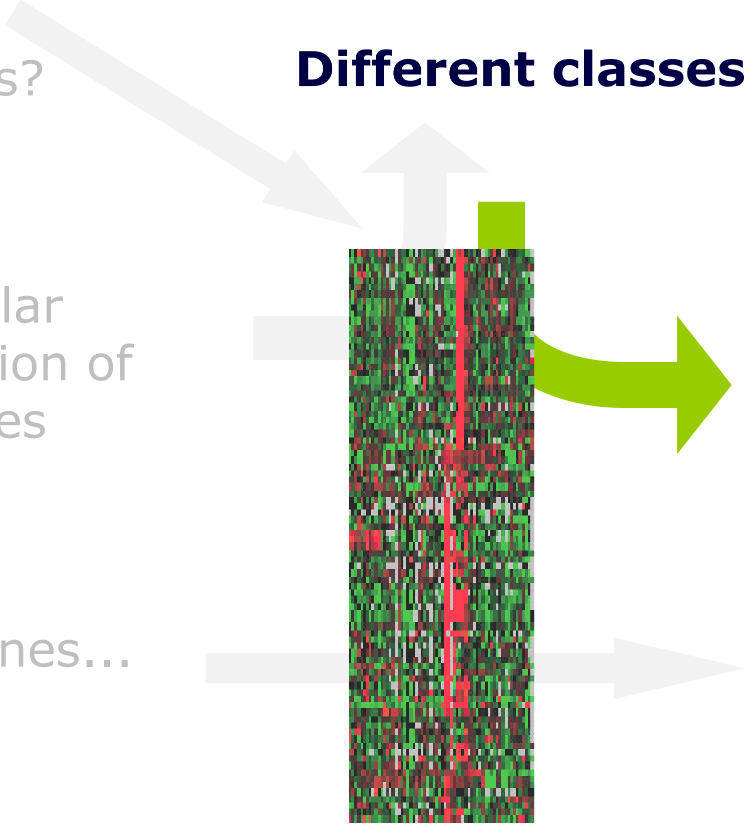
And

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

Gene selection

Studies must be hypothesis driven.

gene selection



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

Different classes...

Molecular classification of samples

What genes are responsible for?

Co-expressing genes...

What do they have in common?

Gene selection.

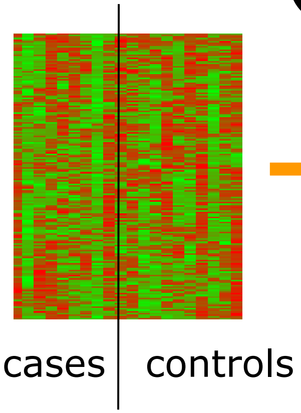
The simplest way: univariant gene-by-gene.
Other multivariant approaches can be used

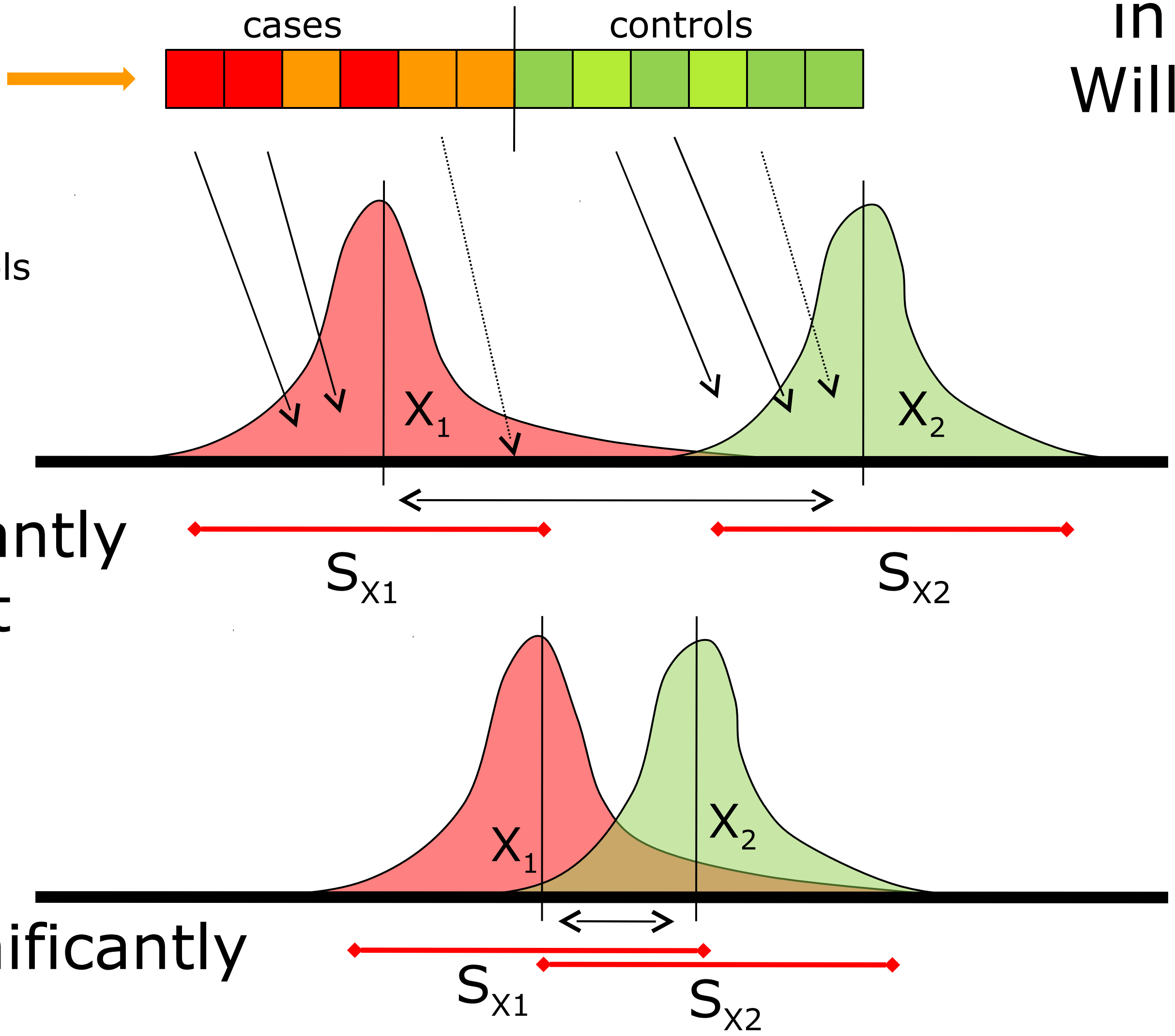
- **One class**
  - Limma
- **Two classes**
  - T-test
  - Limma
  - Fold-change
- **Multiclass**
  - Anova
  - Limma

- **Continuous variable (e.g. level of a metabolite)**
  - Pearson
  - Spearmam
  - Regression
- **Survival**
  - Cox model
- **Time Course**

Gene selection

The t-statistic was introduced in 1908 by William Sealy Gosset



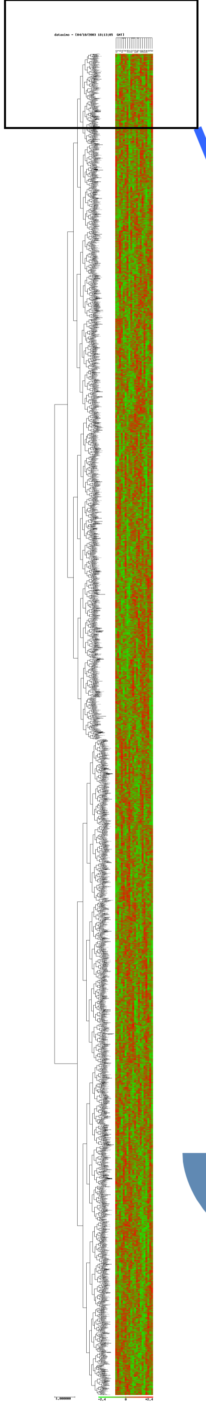


Significantly different

Non significantly different

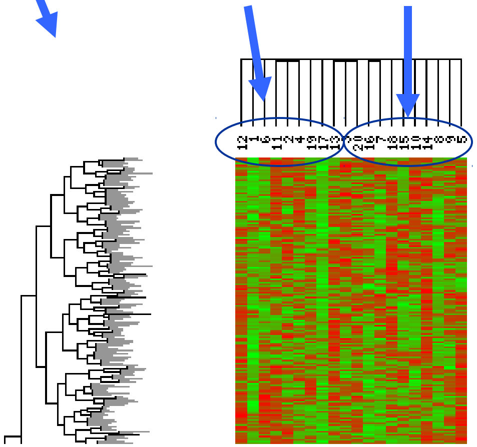
$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$ being $$S_{X_1X_2} = \sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{2}}.$$

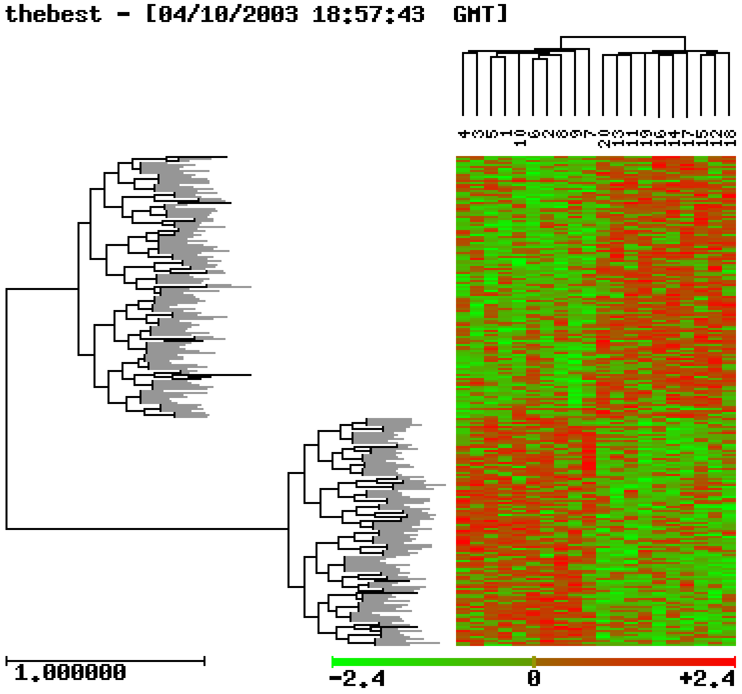
A simple problem: gene selection for class discrimination



~15,000 genes

Case(10)/control(10)

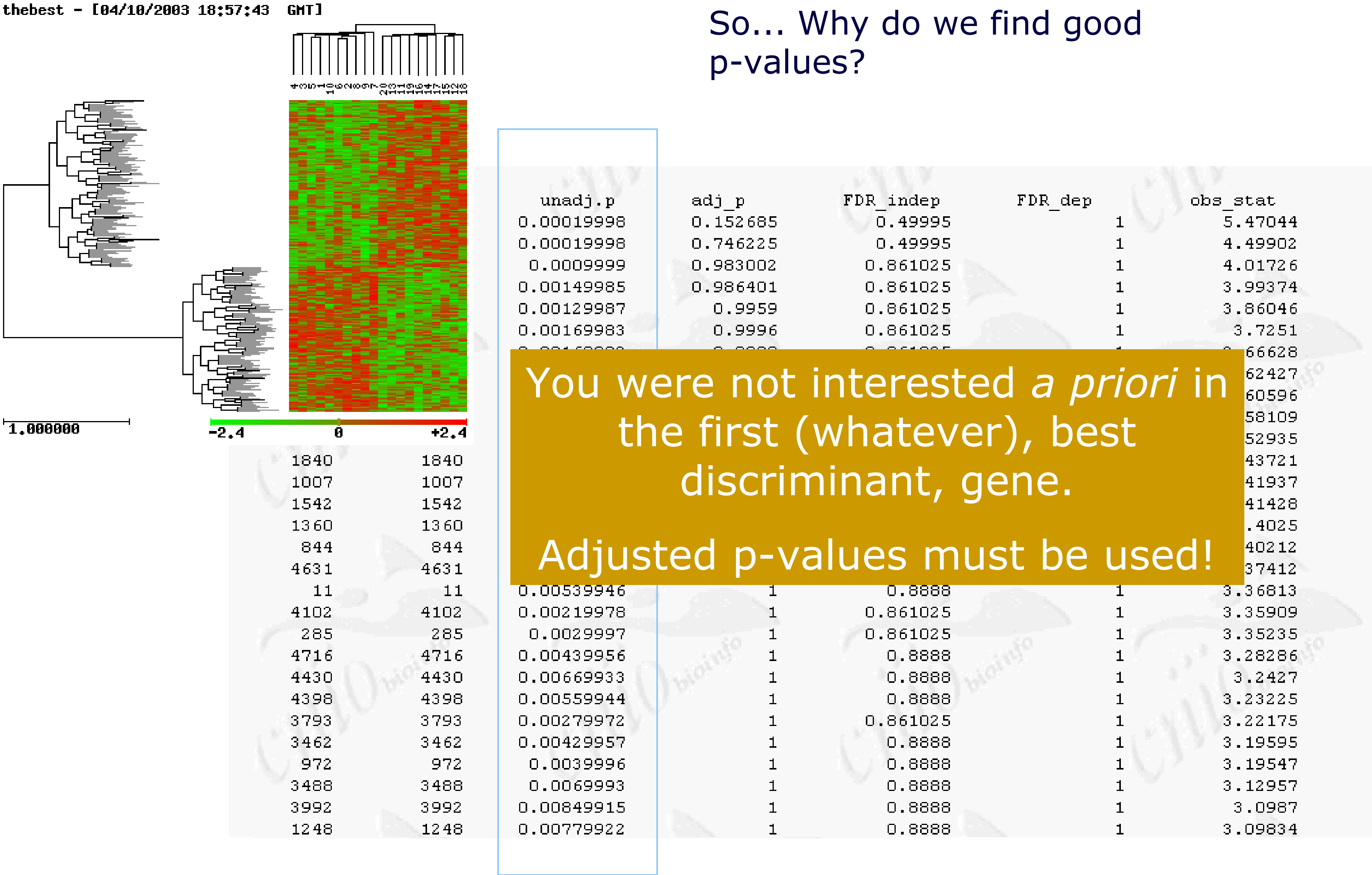




Genes differentially expressed among classes (t-test), with p-value < 0.05

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

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

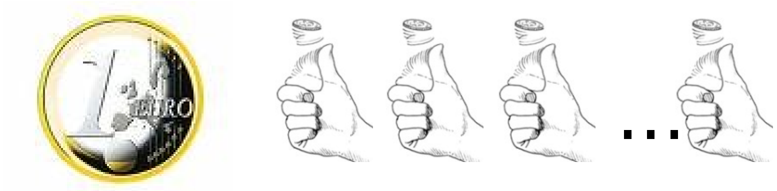


		unadj.p	adj_p	FDR_indep	FDR_dep	obs_stat
		0.00019998	0.152685	0.49995	1	5.47044
		0.00019998	0.746225	0.49995	1	4.49902
		0.0009999	0.983002	0.861025	1	4.01726
		0.00149985	0.986401	0.861025	1	3.99374
		0.00129987	0.9959	0.861025	1	3.86046
		0.00169983	0.9996	0.861025	1	3.7251
					1	66628
						62427
						60596
						58109
						52935
1840	1840					43721
1007	1007					41937
1542	1542					41428
1360	1360					.4025
844	844					40212
4631	4631					37412
11	11	0.00539946	1	0.8888	1	3.36813
4102	4102	0.00219978	1	0.861025	1	3.35909
285	285	0.0029997	1	0.861025	1	3.35235
4716	4716	0.00439956	1	0.8888	1	3.28286
4430	4430	0.00669933	1	0.8888	1	3.2427
4398	4398	0.00559944	1	0.8888	1	3.23225
3793	3793	0.00279972	1	0.861025	1	3.22175
3462	3462	0.00429957	1	0.8888	1	3.19595
972	972	0.0039996	1	0.8888	1	3.19547
3488	3488	0.0069993	1	0.8888	1	3.12957
3992	3992	0.00849915	1	0.8888	1	3.0987
1248	1248	0.00779922	1	0.8888	1	3.09834

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

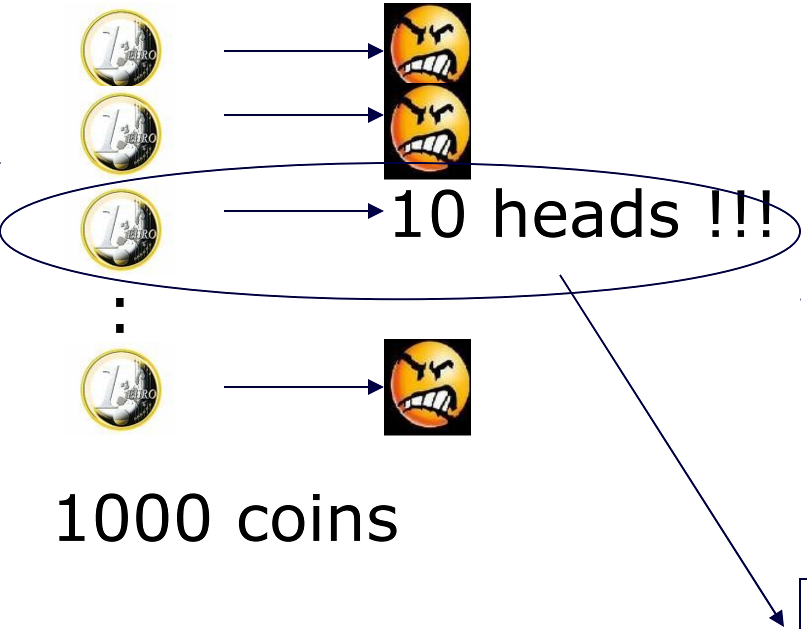
Adjusted p-values must be used!

On the problem of multiple testing



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

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



10 heads !!!

1000 coins

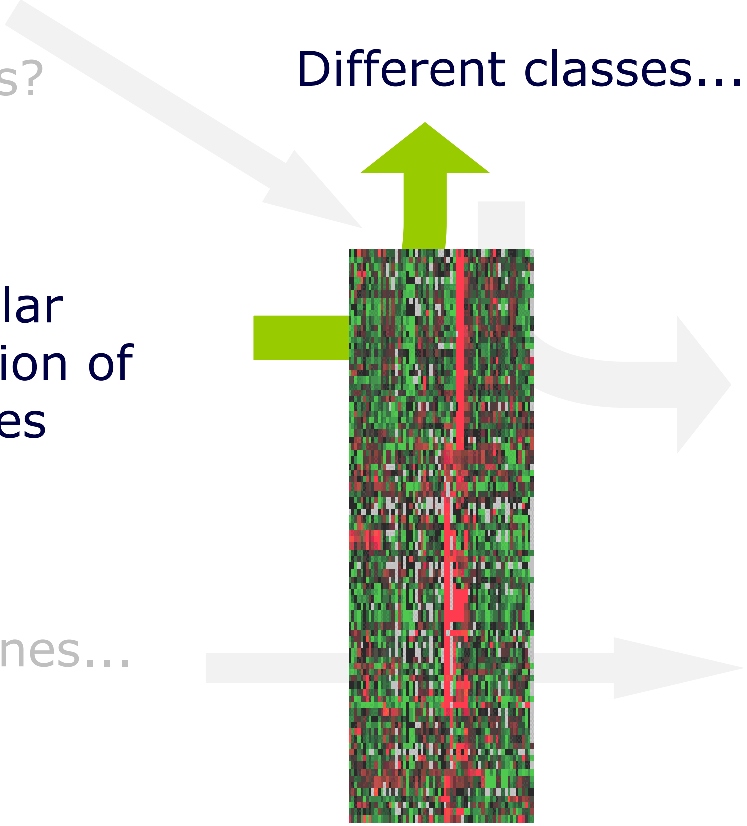
$$P= 1-(1-0.5^{10})^{1000}=0.62$$

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

Will you still use this coin for betting?

Studies must be hypothesis driven.

sample classification



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

Different classes...

Molecular classification of samples

What genes are responsible for?

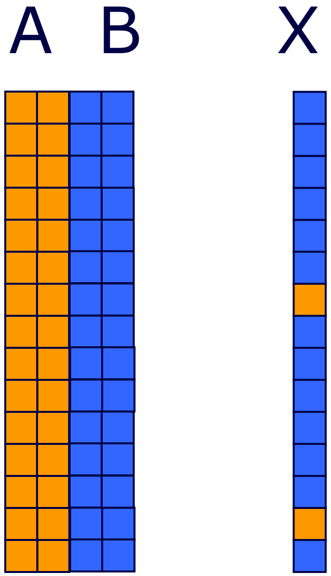
Co-expressing genes...

What do they have in common?

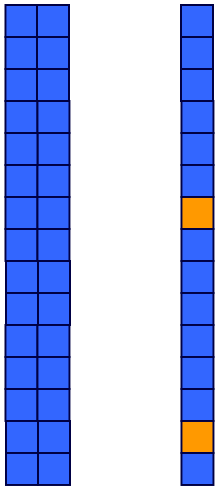
Of predictors and molecular signatures

What is a predictor?

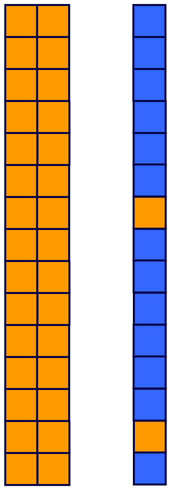
Intuitive notion:



Is X, A or B?



Diff (B, X) = 2



Diff (A, X) = 13

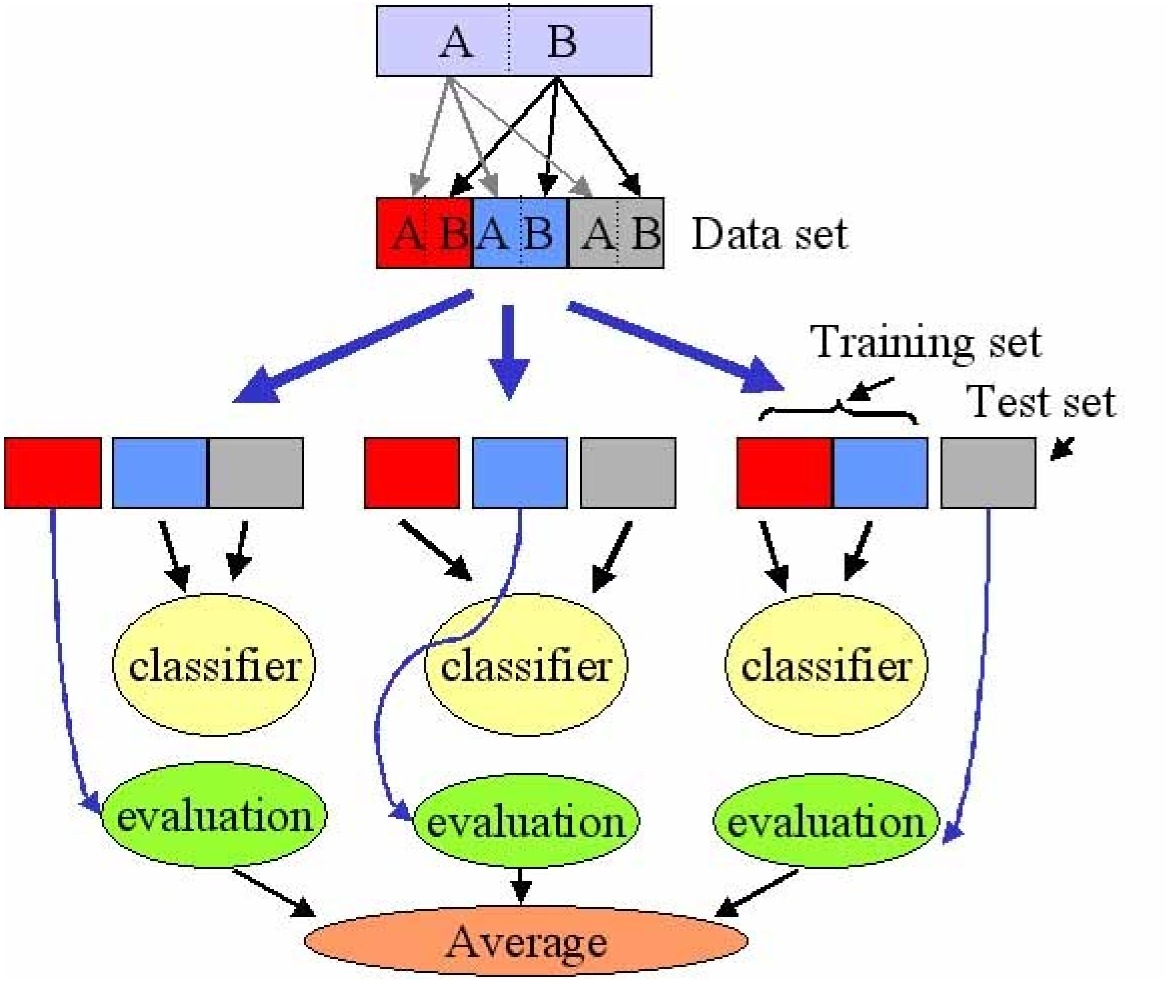
Most probably X belongs to class B

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

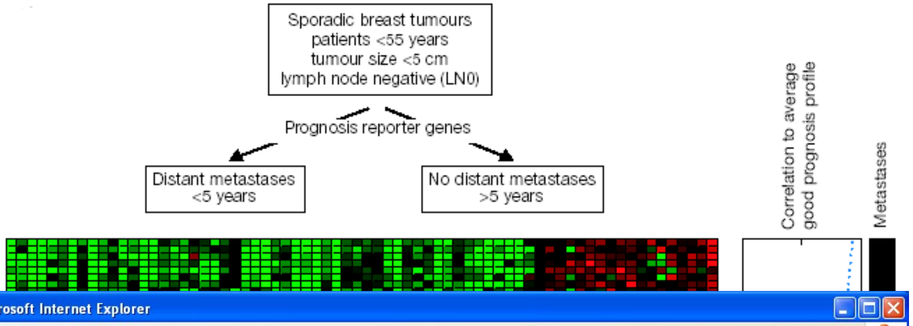
Cross-validation

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

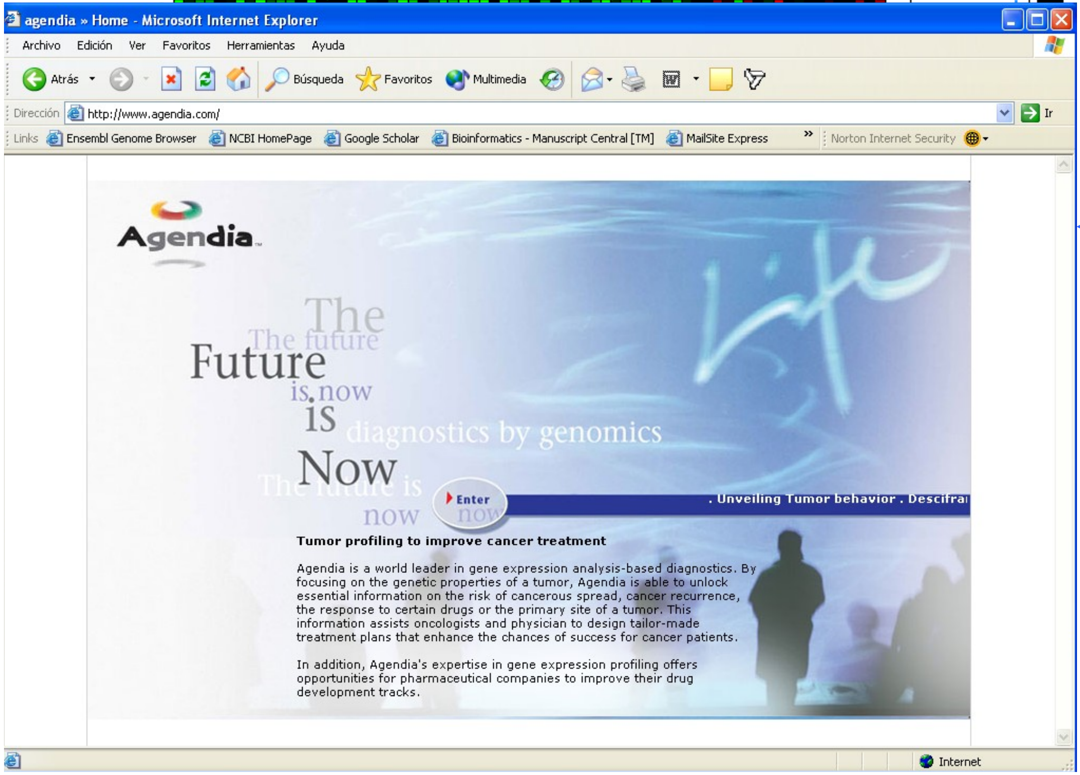
Typical are three-fold, ten-fold and leave-one-out (LOO), in case of few samples for the training



Predictor of clinical outcome in breast cancer


Sporadic breast tumours
patients <55 years
tumour size <5 cm
lymph node negative (LN0)
Prognosis reporter genes
Distant metastases <5 years
No distant metastases >5 years
Correlation to average good prognosis profile
Metastases

Genes are arranged to their correlation eith the pronostic groups



Tumor profiling to improve cancer treatment

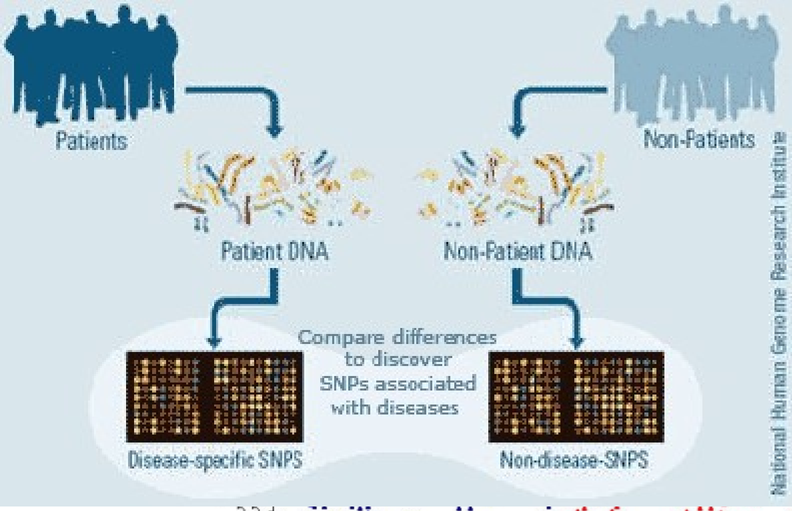
Agendia is a world leader in gene expression analysis-based diagnostics. By focusing on the genetic properties of a tumor, Agendia is able to unlock essential information on the risk of cancerous spread, cancer recurrence, the response to certain drugs or the primary site of a tumor. This information assists oncologists and physician to design tailor-made treatment plans that enhance the chances of success for cancer patients.

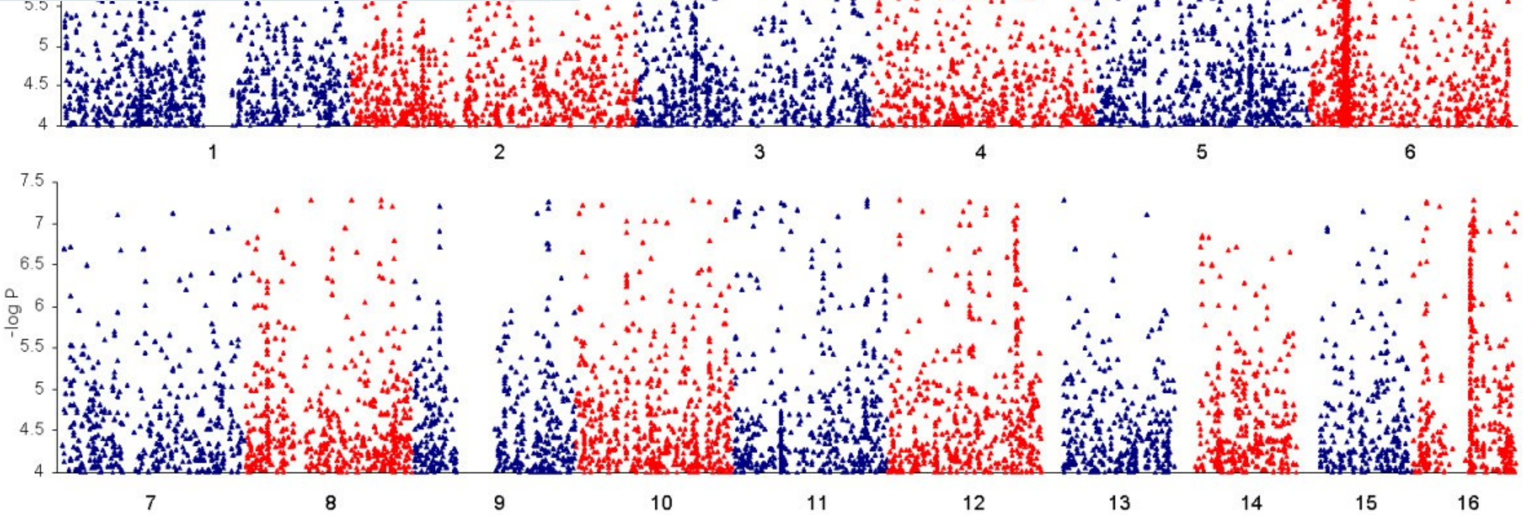
In addition, Agendia's expertise in gene expression profiling offers opportunities for pharmaceutical companies to improve their drug development tracks.

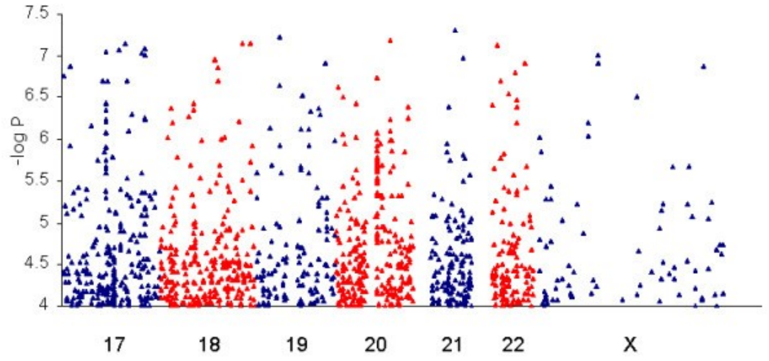
Pronostic classifier with optimal accuracy

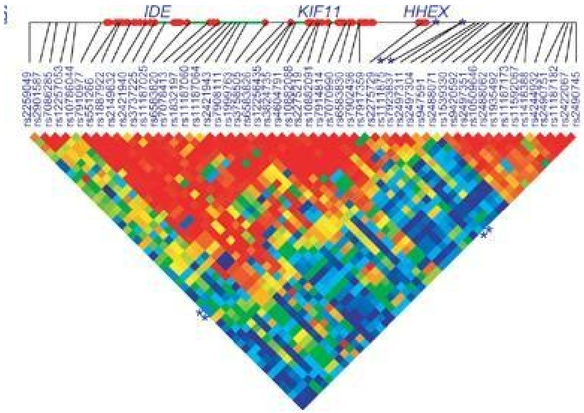
van't Veer et al., Nature, 2002

Genotyping


Patients
Non-Patients
Patient DNA
Non-Patient DNA
Compare differences to discover SNPs associated with diseases
Disease-specific SNPS
Non-disease-SNPS
National Human Genome Research Institute

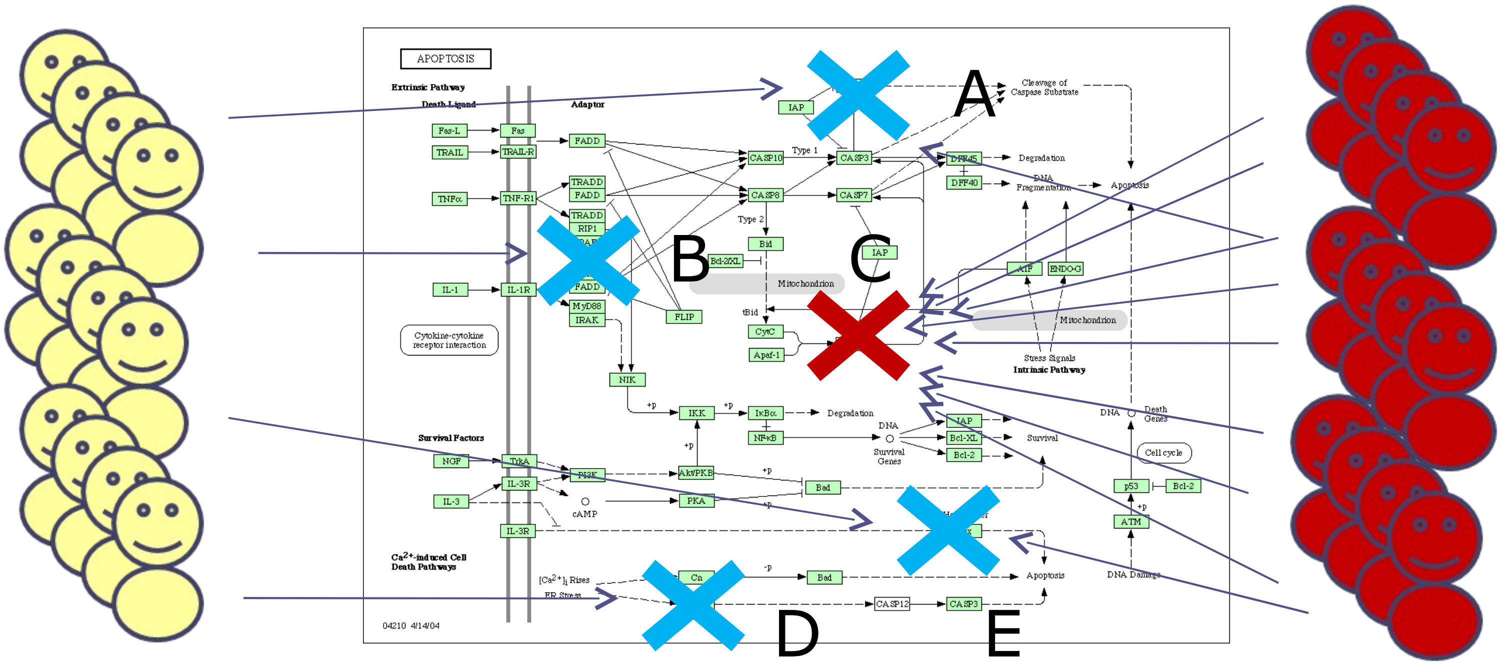

5.5
5
4.5
4
1
2
3
4
5
6
7.5
7
6.5
6
5.5
5
4.5
4
-log P
7
8
9
10
11
12
13
14
15
16


7.5
7
6.5
6
5.5
5
4.5
4
-log P
17
18
19
20
21
22
X


IDE
KIF11
HHEX

Genotyping to find mutations associated to diseases

The simplest case: monogenic disease



Controls

Cases

Gene A	1 0 0 0 0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0 0 0 0 0
Gene B	0 0 0 1 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0
Gene C	0 0 0 0 0 0 0 0 0 0 0 0	1 1 1 1 1 1 1 1 1 1 1 1
Gene D	0 0 0 0 0 0 0 0 0 0 1 0	0 0 0 0 0 0 0 0 0 0 0 0
Gene E	0 0 0 0 0 1 0 0 0 0 0 0	0 0 0 0 0 0 0 1 0 0 0 0

The real life in GWAS

Our analysis of Hirschsprung´s disease

54 trios of short-segment Hirschsprung´s disease Affy 6.0 (1million SNPs)

Conventional TDT test reports only 4 significant SNPs mapping only on one gene: RET, already knowk to be associated to the disease

This is not a matter of sample size: an example of GWAS in Breast Cancer.

The CGEMS initiative. (Hunter et al. Nat Genet 2007)

1145 cases 1142 controls. Affy 500K

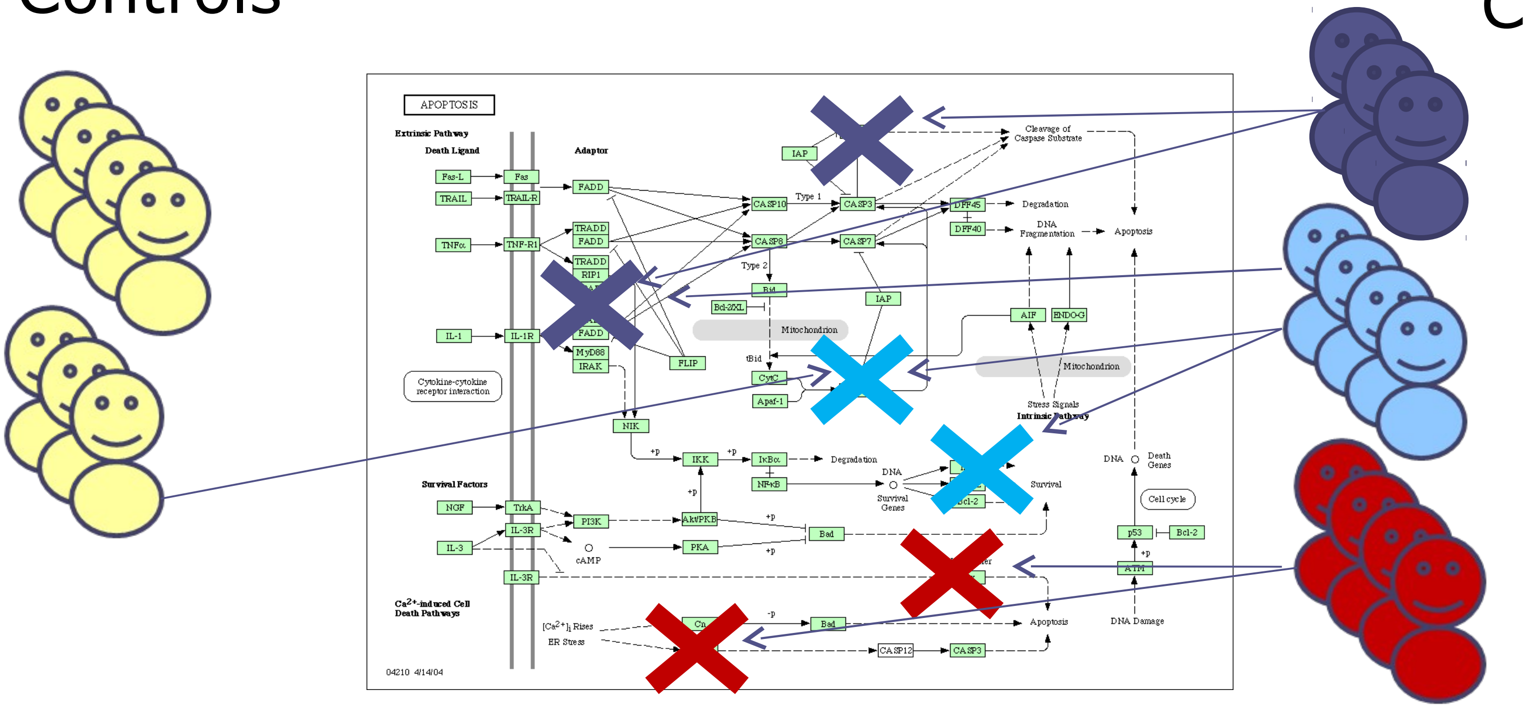
Conventional association test reports only significant 4 SNPs mapping only on one gene: FGFR2

Conclusions: conventional tests are not providing much resolution. What is the problem with them? Are there solutions?

Clear individual gene associations are difficult to find in multifactorial diseases

Controls

Cases



The cases of the multifactorial disease will have different mutations (or combinations). Many cases have to be used to obtain significant associations to many markers. The only common element is the pathway (yet unknow) affected.

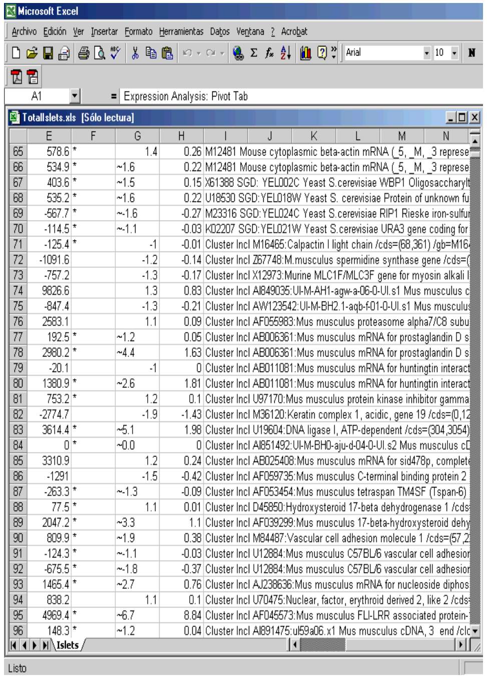
Functional profiling of genome-scale experiments in the post-genomic era

My data...

How are structured?

What are these groups?

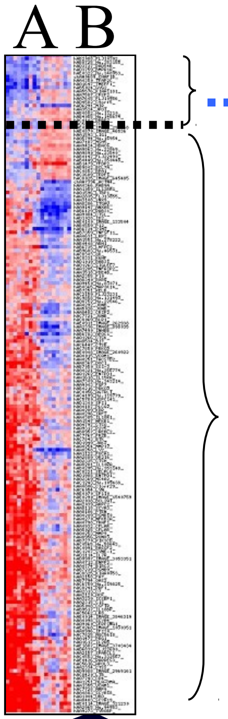
What is this gen?


Microsoft Excel
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ? Acrobat
Arial
10
N
Islets
Listo

A1 = Expression Analysis: Pivot Tab

Totallslets.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: YEL002C Yeast S.cerevisiae WBP1 Oligosaccharylt					
68	535.2	*	~1.6	0.22	U18530 SGD:YEL018W Yeast S. cerevisiae Protein of unknown fu					
69	-567.7	*	~-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 M16465:Calpactin I light chain /cds=(68,361) /gb=M16					
72	-1091.6		-1.2	-0.14	Cluster Incl Z67748:M.musculus spermidine synthase gene /cds=(					
73	-757.2		-1.3	-0.17	Cluster Incl X12973:Murine MLC1F/MLC3F gene for myosin alkali l					
74	9826.6		1.3	0.83	Cluster Incl AI849035:UI-M-AH1-agw-a-06-0-UI.s1 Mus musculus c					
75	-847.4		-1.3	-0.21	Cluster Incl AW123542:UI-M-BH2.1-aqb-f-01-0-UI.s1 Mus musculus					
76	2583.1		1.1	0.09	Cluster Incl AF055983:Mus musculus proteasome alpha7/C8 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 AI851492:UI-M-BH0-aju-d-04-0-UI.s2 Mus musculus cD					
85	3310.9		1.2	0.24	Cluster Incl AB025408:Mus musculus mRNA for sid478p, complete					
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 AF053454:Mus musculus tetraspan TM4SF (Tspan-6)					
88	77.5	*	1.1	0.01	Cluster Incl D45850:Hydroxysteroid 17-beta dehydrogenase 1 /cds					
89	2047.2	*	~3.3	1.1	Cluster Incl AF039299:Mus musculus 17-beta-hydroxysteroid dehy					
90	809.9	*	~1.9	0.38	Cluster Incl M84487:Vascular cell adhesion molecule 1 /cds=(57,2					
91	-124.3	*	~-1.1	-0.03	Cluster Incl U12884:Mus musculus C57BL/6 vascular cell adhesior					
92	-675.5	*	~-1.8	-0.37	Cluster Incl U12884:Mus musculus C57BL/6 vascular cell adhesior					
93	1465.4	*	~2.7	0.76	Cluster Incl AJ238636:Mus musculus mRNA for nucleoside diphos					
94	838.2		1.1	0.1	Cluster Incl U70475:Nuclear, factor, erythroid derived 2, like 2 /cds					
95	4969.4	*	~6.7	8.84	Cluster Incl AF045573:Mus musculus FLI-LRR associated protein-					
96	148.3	*	~1.2	0.04	Cluster Incl AI891475:ul59a06.x1 Mus musculus cDNA, 3 end /clo					

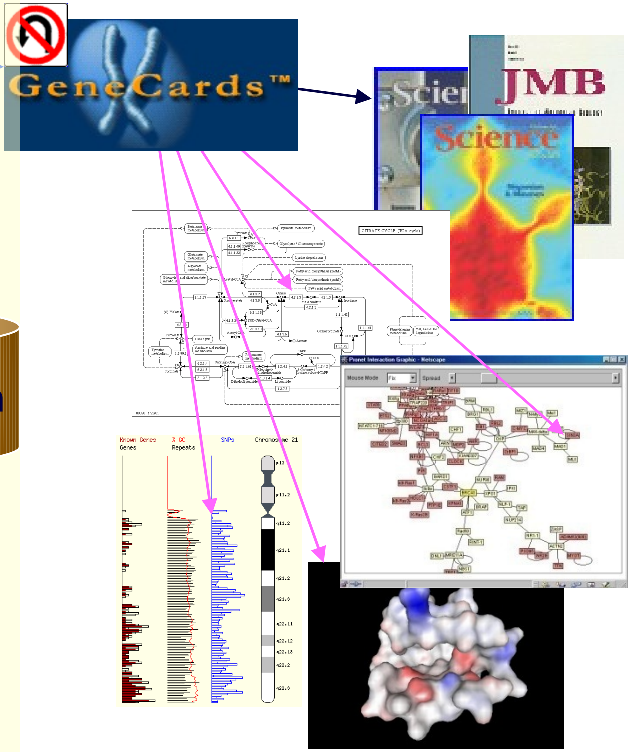


?

Cell cycle...

DBs Information

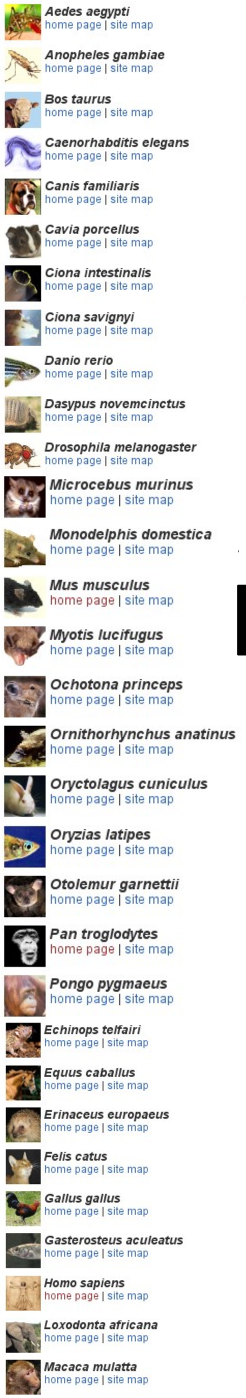
I M19380:Calmodulin 3 /cds=(109,558) /gb=M19380 /gi=469419
I AI842328:UI-M-AM1-afz-b-11-0-UI.s1 Mus musculus cDNA, 3
I AJ242663:Mus musculus mRNA for cathepsin Z precursor (cts
I U12620:Dipeptidylpeptidase 4 /cds=(117,2399) /gb=U12620 /g
I M13444:Mouse alpha-tubulin isotype M-alpha-4 mRNA, compl
I U11027:Mus musculus C57BL/6J Sec61 protein complex gam
I J03928:Phosphofructokinase, liver, B-type /cds=(42,2384) /gb=
I Z67745:M.musculus mRNA for phosphatase 2A catalytic subu
I U80932:Serine/threonine kinase 6 /cds=(48,1235) /gb=U80932
I U47024:Maternal embryonic message 3 /cds=(137,2401) /gb=
I AF075136:Mus musculus Sin3-associated protein (sap30) mR
I M25944:Mouse carbonic anhydrase II (CAII) mRNA, 3 end /cd
I X74671:Neurofibromatosis 2 /cds=(576,2366) /gb=X74671 /gi=
I M12848:Mouse myb proto-oncogene mRNA encoding 71 kd m
I AW125458:UI-M-BH2.2-aqm-a-07-0-UI.s1 Mus musculus cDN
I U84903:Ribosomal protein L23 /cds=(61,501) /gb=U84903 /gi=
I U35141:Mus musculus retinoblastoma-binding protein (mRbAp
I U19521:Mus musculus vesicle transport protein (munc-18c) m
I M15268:Aminolevulinic acid synthase 2, erythroid /cds=(0,176
I M25149:Tissue specific transplantation antigen P91A /cds=(0,
I X66449:Calcyclin /cds=(159,428) /gb=X66449 /gi=50271 /ug=M



Analysis

Functional profiling

Links



Aedes aegypti
home page | site map

Anopheles gambiae
home page | site map

Bos taurus
home page | site map

Caenorhabditis elegans
home page | site map

Canis familiaris
home page | site map

Cavia porcellus
home page | site map

Ciona intestinalis
home page | site map

Ciona savignyi
home page | site map

Danio rerio
home page | site map

Dasypus novemcinctus
home page | site map

Drosophila melanogaster
home page | site map

Microcebus murinus
home page | site map

Monodelphis domestica
home page | site map

Mus musculus
home page | site map

Myotis lucifugus
home page | site map

Ochotona princeps
home page | site map

Ornithorhynchus anatinus
home page | site map

Oryctolagus cuniculus
home page | site map

Oryzias latipes
home page | site map

Otolemur garnettii
home page | site map

Pan troglodytes
home page | site map

Pongo pygmaeus
home page | site map

Echinops telfairi
home page | site map

Equus caballus
home page | site map

Erinaceus europaeus
home page | site map

Felis catus
home page | site map

Gallus gallus
home page | site map

Gasterosteus aculeatus
home page | site map

Homo sapiens
home page | site map

Loxodonta africana
home page | site map

Macaca mulatta
home page | site map

Genome Annotation

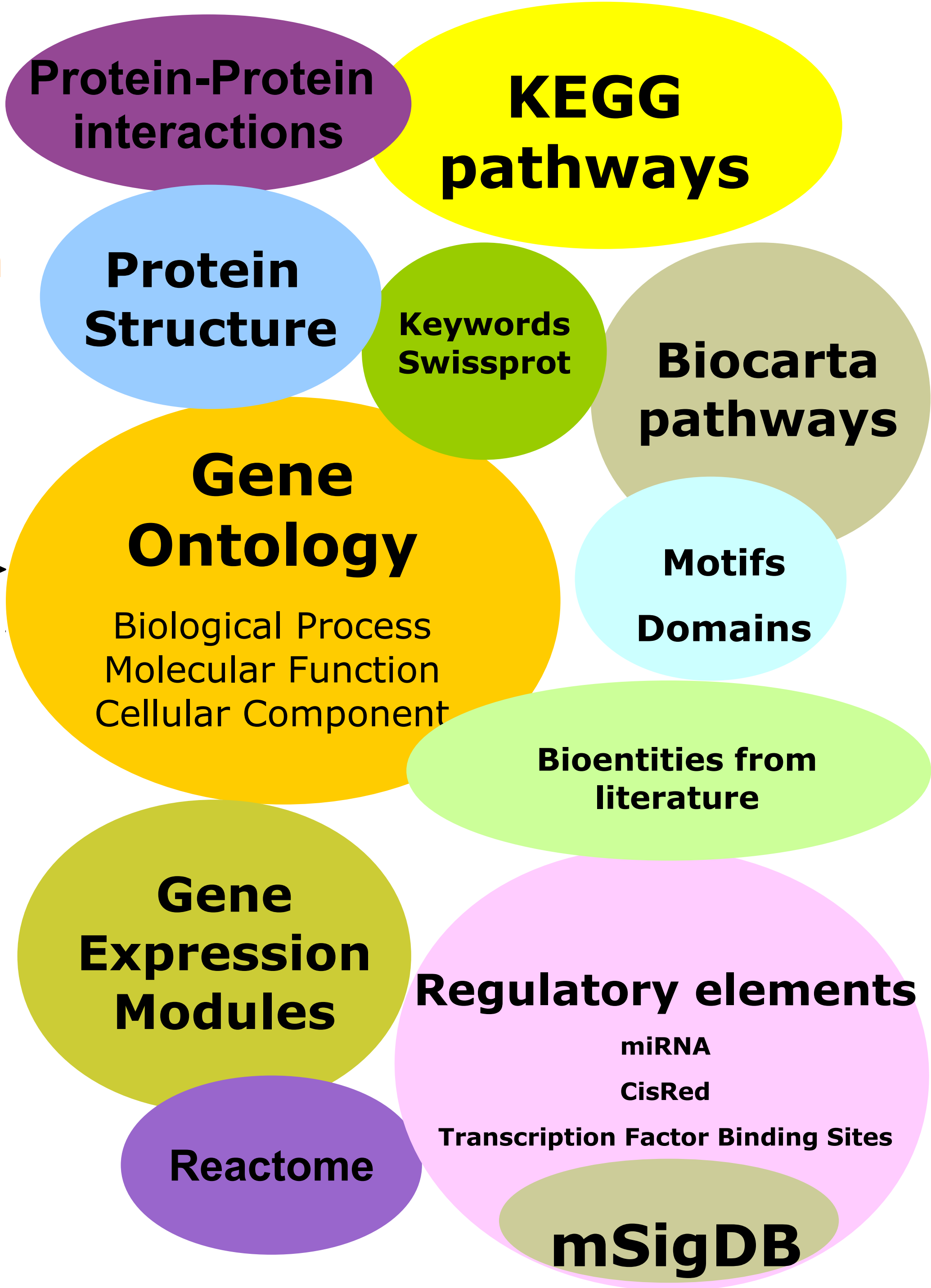
Structural Annotation

Functional Annotation

Biological Databases

Gene Annotation

Gene Set Annotation



Protein-Protein interactions

KEGG pathways

Protein Structure

Keywords Swissprot

Biocarta pathways

Gene Ontology

Biological Process
Molecular Function
Cellular Component

Motifs

Domains

Bioentities from literature

Gene Expression Modules

Regulatory elements

miRNA

CisRed

Transcription Factor Binding Sites

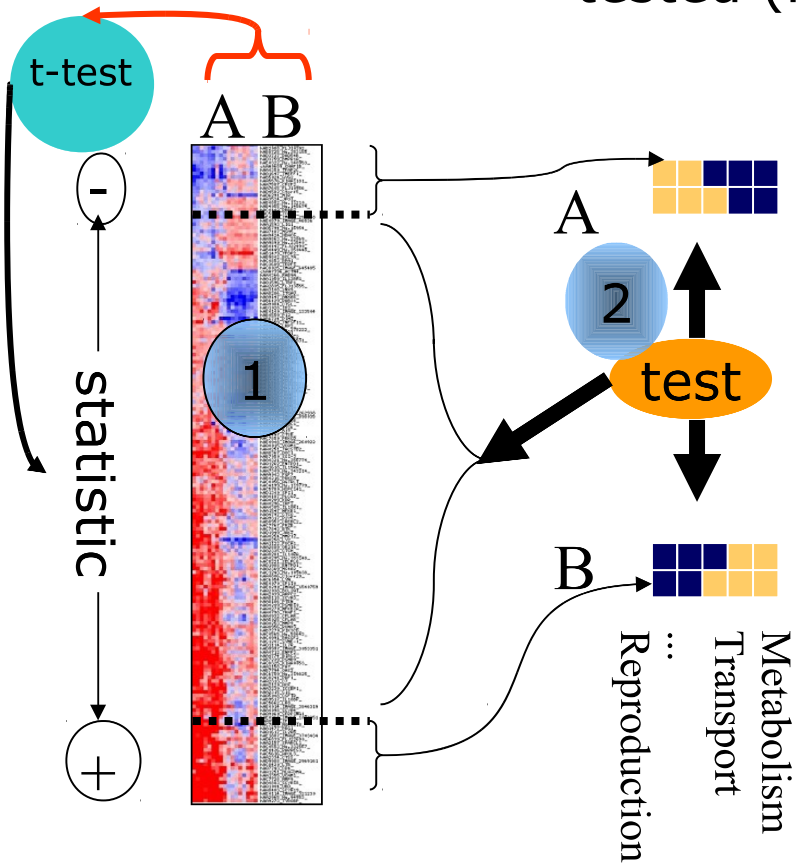
Reactome

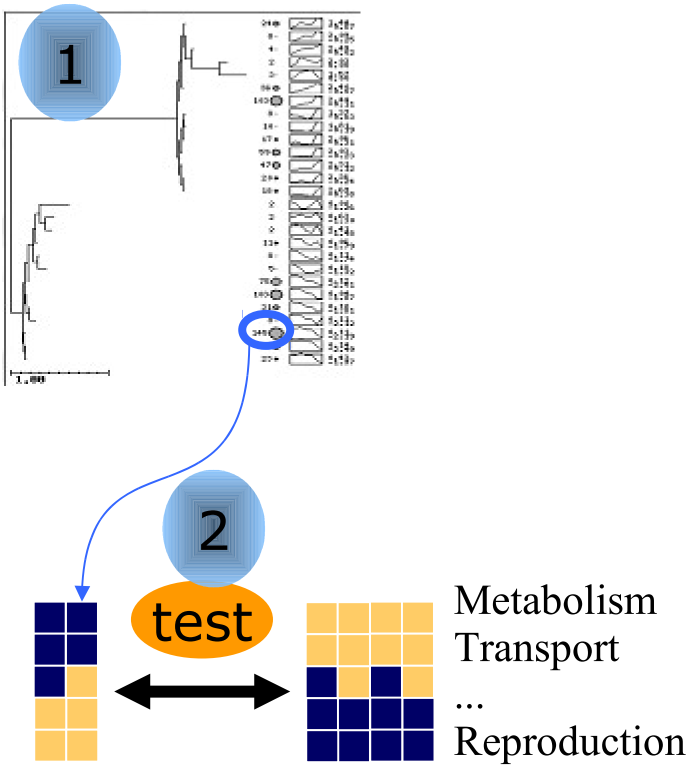
mSigDB

Two-steps functional interpretation

1 Genes are selected based on their experimental values and...

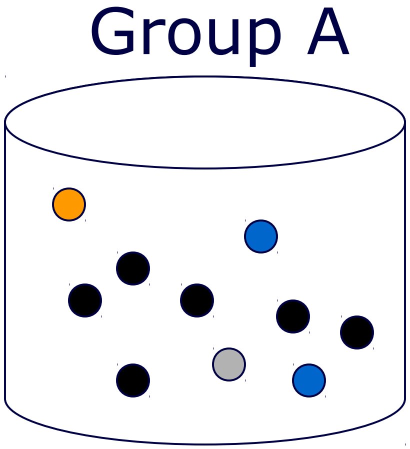
2 Enrichment in functional terms is tested (FatiGO, GoMiner, etc.)



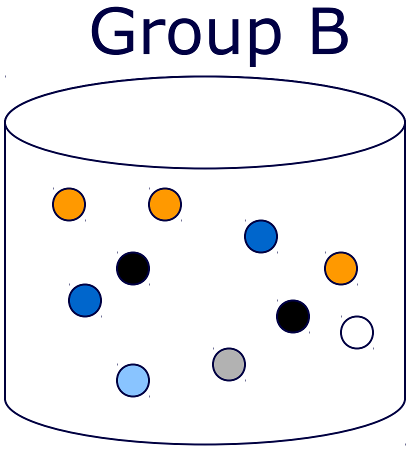


Testing two GO terms

(remember, we have to test thousands)



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

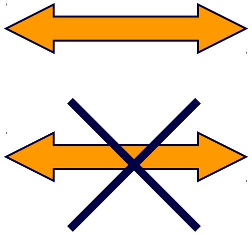


	Biosynthesis	Other
A	6	4
B	2	8

The popular Fisher's test

Cell cycle 60% ●

Apoptosis 20% ●



Cell cycle 20% ●

Apoptosis 20% ●

Genes in group A have significantly to do with cell cycle, but not with apoptosis.

GO terms found in sets of 50 genes

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

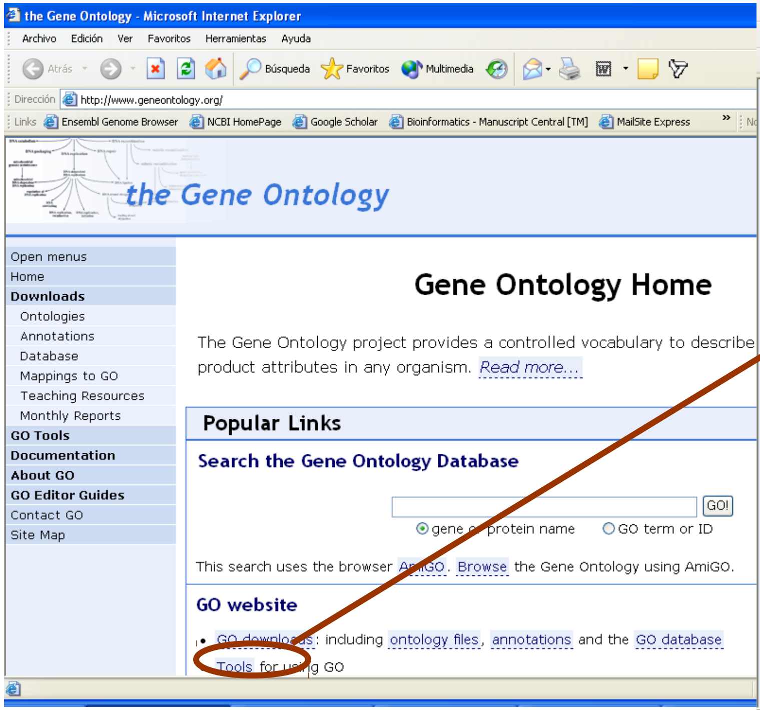
Each row corresponds to a random selection of 50 genes from the *E. coli* genome, compared with respect to the rest of the genome.

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

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


the Gene Ontology - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Multimedia
Dirección http://www.geneontology.org/
Links Ensembl Genome Browser NCBI HomePage Google Scholar Bioinformatics - Manuscript Central [TM] MailSite Express
the Gene Ontology
Open menus
Home
Downloads
Ontologies
Annotations
Database
Mappings to GO
Teaching Resources
Monthly Reports
GO Tools
Documentation
About GO
GO Editor Guides
Contact GO
Site Map
GO!
○ gene or protein name ○ GO term or ID

Gene Ontology Home

The Gene Ontology project provides a controlled vocabulary to describe product attributes in any organism. Read more...

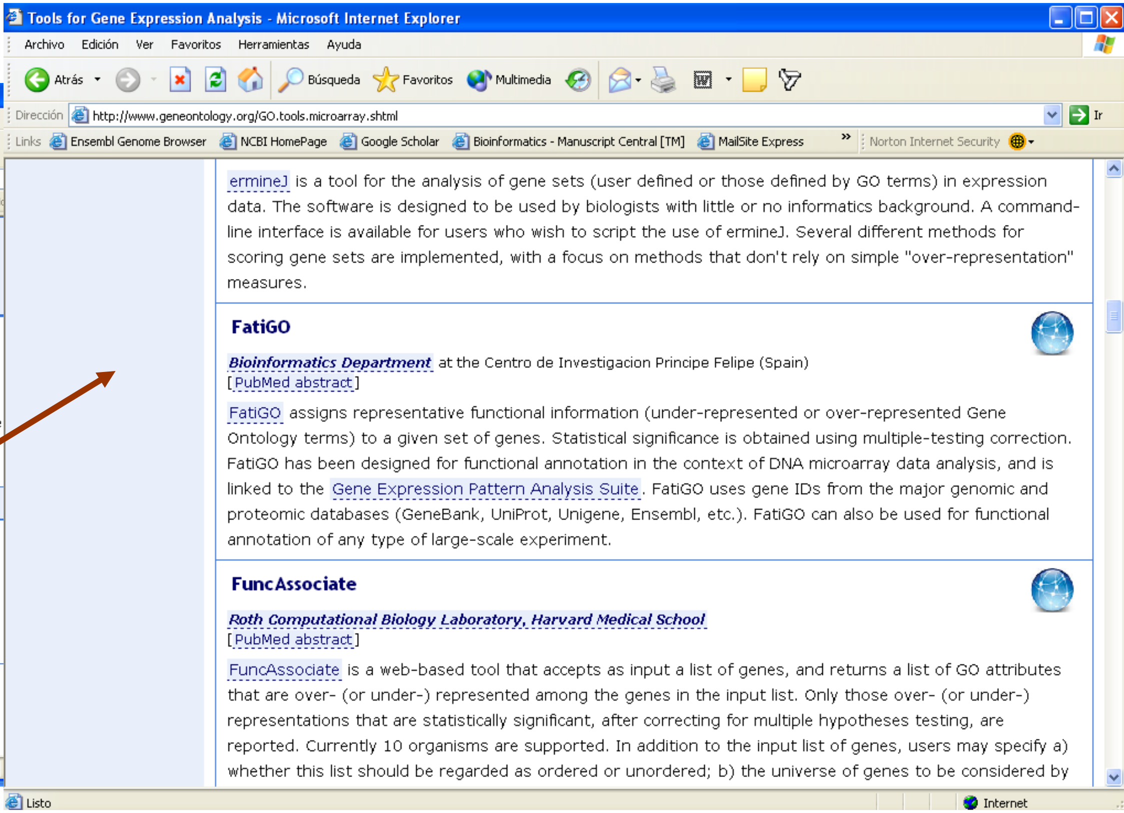
Popular Links

Search the Gene Ontology Database

This search uses the browser AmiGO. Browse the Gene Ontology using AmiGO.

GO website

- GO downloads: including ontology files, annotations and the GO database
- Tools for using GO


Tools for Gene Expression Analysis - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Multimedia
Dirección http://www.geneontology.org/GO.tools.microarray.shtml
Ir
Links Ensembl Genome Browser NCBI HomePage Google Scholar Bioinformatics - Manuscript Central [TM] MailSite Express
Norton Internet Security
Listo
Internet

ermineJ is a tool for the analysis of gene sets (user defined or those defined by GO terms) in expression data. The software is designed to be used by biologists with little or no informatics background. A command-line interface is available for users who wish to script the use of ermineJ. Several different methods for scoring gene sets are implemented, with a focus on methods that don't rely on simple "over-representation" measures.

FatiGO

Bioinformatics Department at the Centro de Investigacion Principe Felipe (Spain)
[PubMed abstract]

FatiGO assigns representative functional information (under-represented or over-represented Gene Ontology terms) to a given set of genes. Statistical significance is obtained using multiple-testing correction. FatiGO has been designed for functional annotation in the context of DNA microarray data analysis, and is linked to the Gene Expression Pattern Analysis Suite. FatiGO uses gene IDs from the major genomic and proteomic databases (GeneBank, UniProt, Unigene, Ensembl, etc.). FatiGO can also be used for functional annotation of any type of large-scale experiment.

FuncAssociate

Roth Computational Biology Laboratory, Harvard Medical School
[PubMed abstract]

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

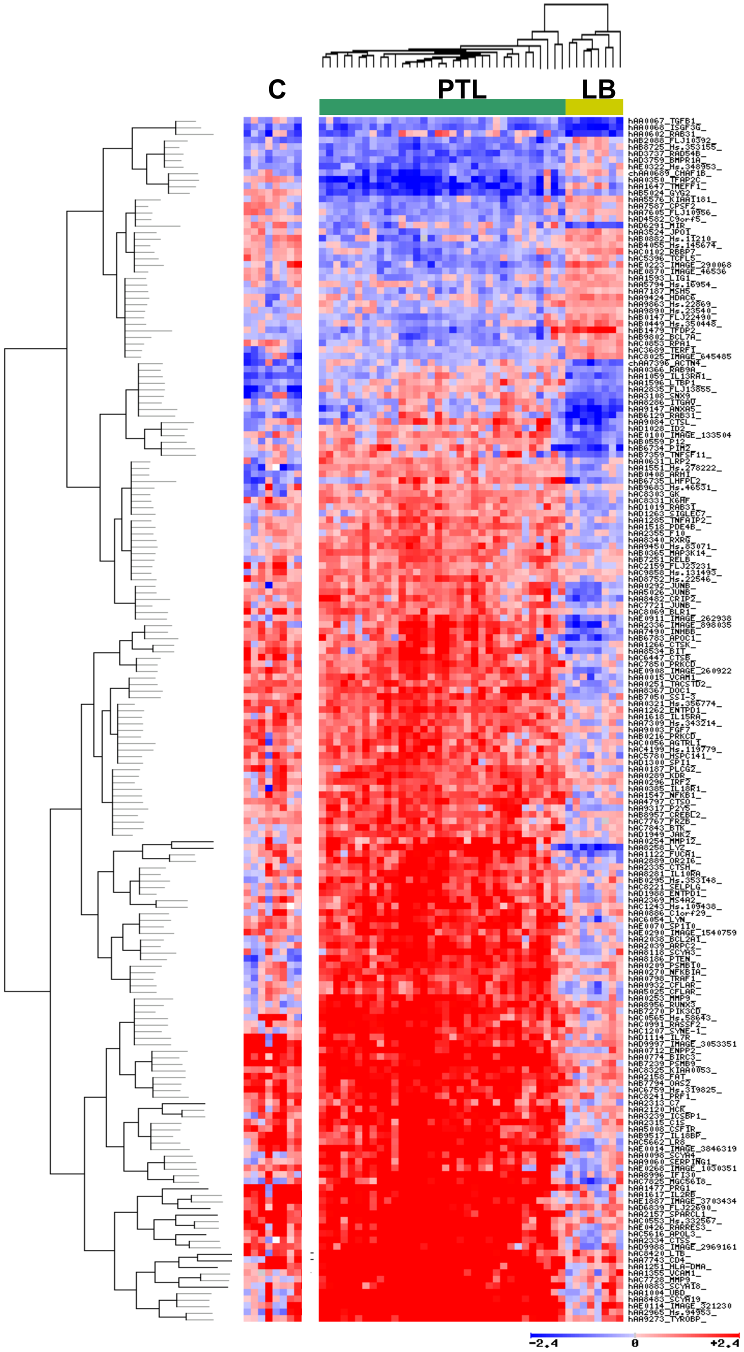
Al-Shahrour et al., 2004 Bioinformatics ***(3rd most cited paper in computing sciences. Source: ISI Web of knowledge.)***

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

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

Al-Shahrour et al., 2007 NAR

Understanding why genes differ in their expression between two different conditions



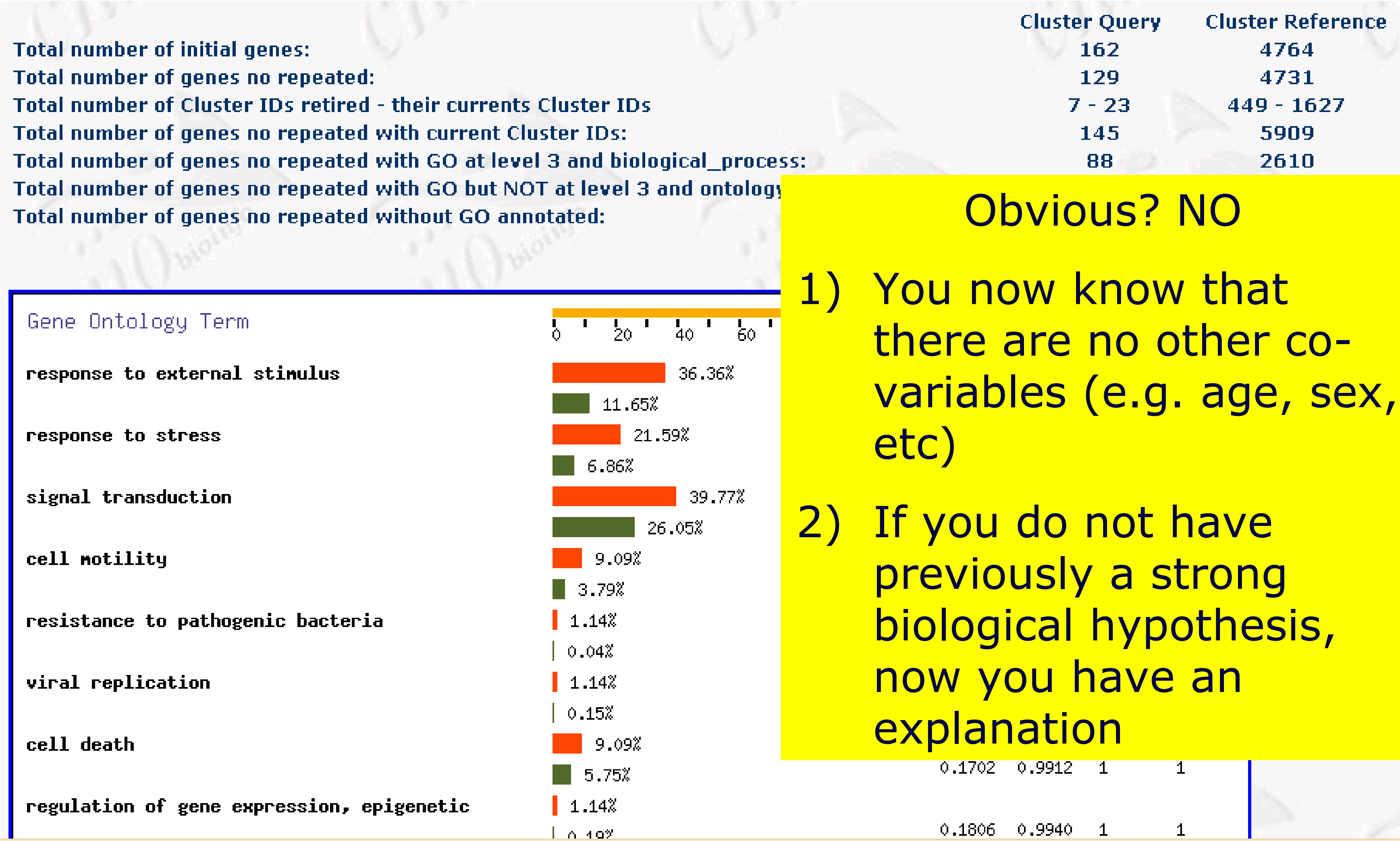
Limphomas from mature lymphocytes (LB) and precursor T-lymphocyte (PTL).

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



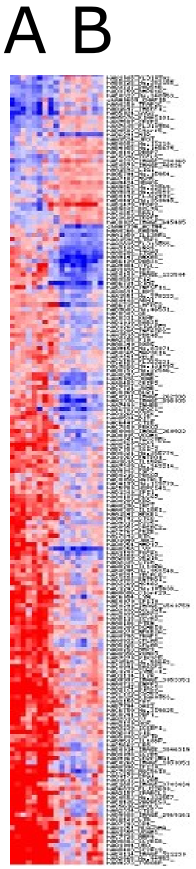
	Cluster Query	Cluster Reference
Total number of initial genes:	162	4764
Total number of genes no repeated:	129	4731
Total number of Cluster IDs retired - their currents Cluster IDs	7 - 23	449 - 1627
Total number of genes no repeated with current Cluster IDs:	145	5909
Total number of genes no repeated 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

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

Weaknesses of the two-steps, functional enrichment approach

Low sensitivity of conventional gene selection methods



A

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

B

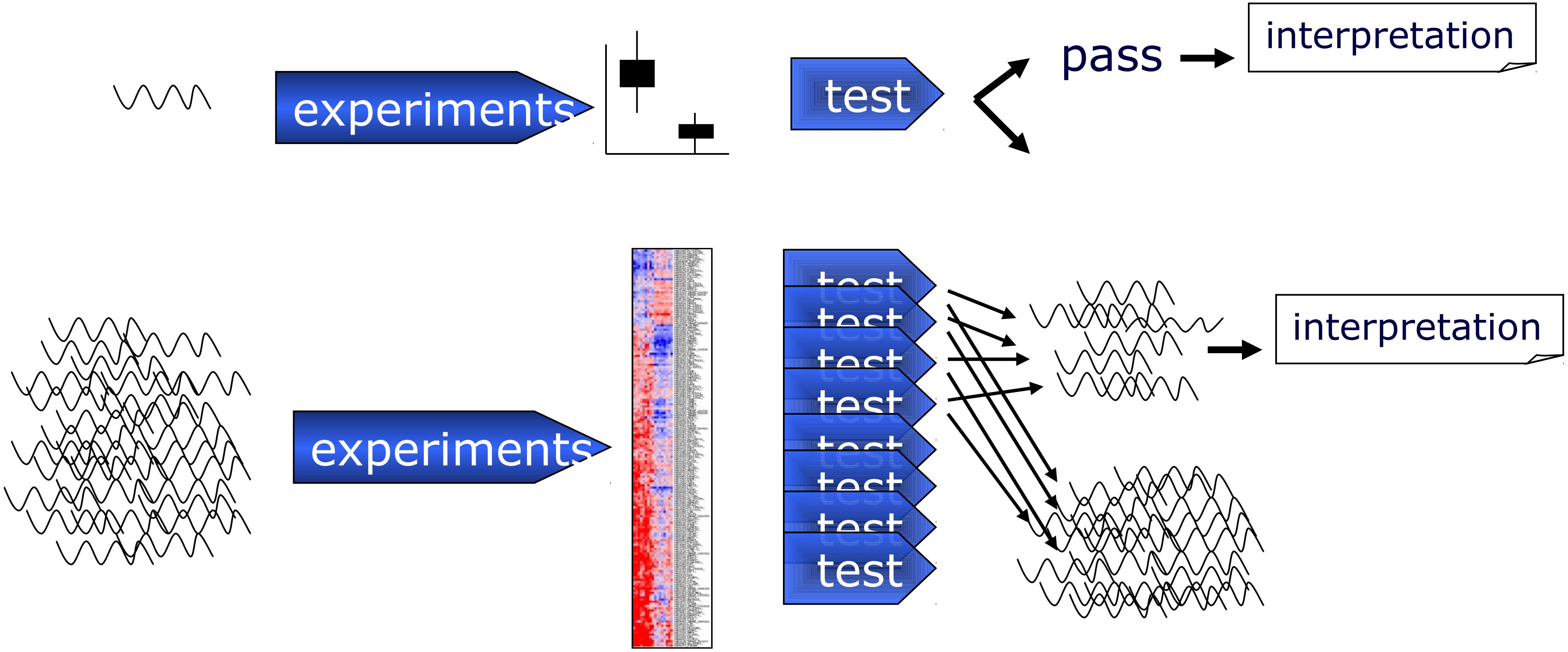
17 with normal tolerance to glucose (**NTG**)

(Mootha et al., 2003)

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

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

Functional enrichment approach reproduces pre-genomics paradigms



experiments

test

pass

interpretation

experiments

test

test

test

test

test

test

test

test

interpretation

Context and cooperation between genes is ignored

So, what is wrong with what we are doing?

We seek for the functions activated/deactivated in our experiment

To find them we firstly seek for genes activated/deactivated one at a time (independently)

Then we look among them for enrichment in functions (cooperative activities) using a second test that consider functions independent.

Therefore… is all wrong with this. The test we conduct is implicitly answering a question different to the one we want to ask.

So, what is wrong with what we are doing? (II)

This testing strategy is very strict in controlling:

Type I error (α): reject the null hypothesis when the null hypothesis is true, (false positive)

Type II error (β): fail to reject the null hypothesis when the null hypothesis is false (false negative)

But, we forget about

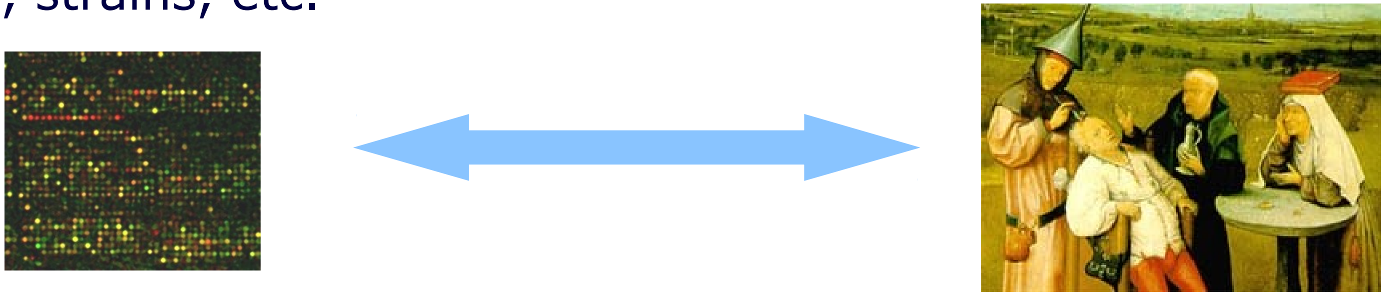
Type III error : get the right answer having asked the wrong question!

The testing strategy we are conducting is implicitly answering a question different to the one we want to ask.

Functional genomics.

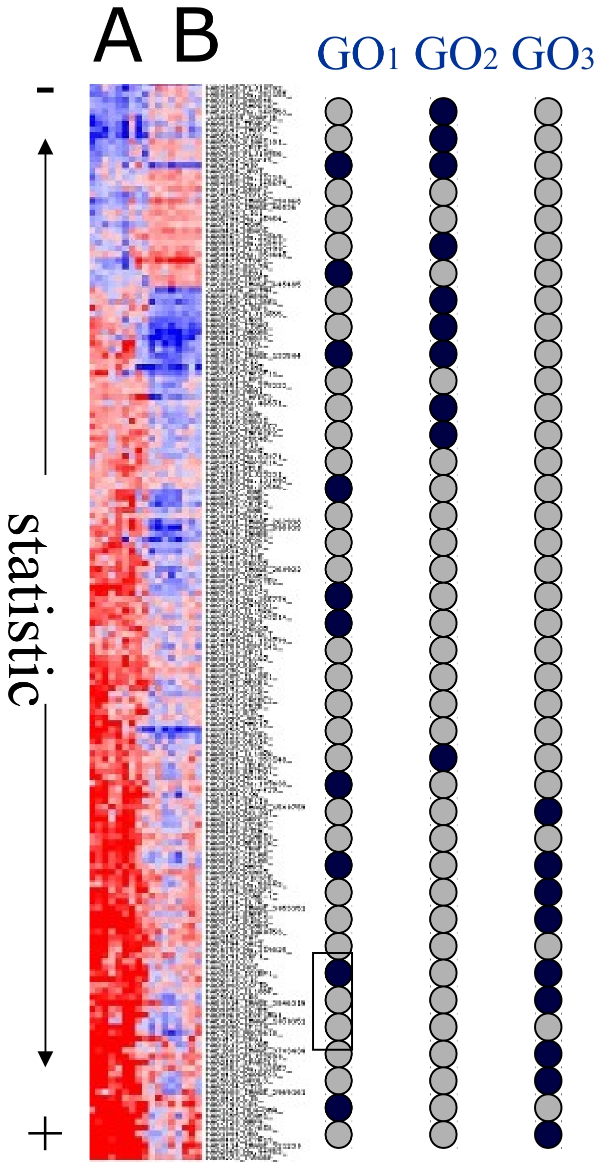
Historic perspective and future

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



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

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

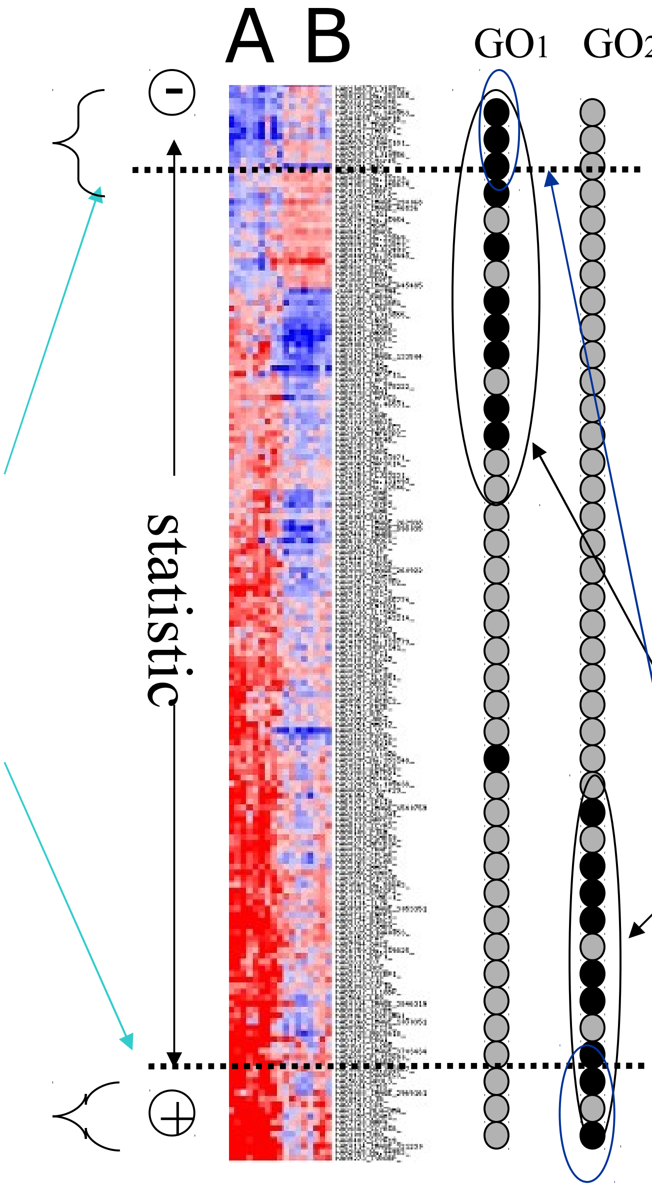
A previous step of gene selection causes loss of information and makes the test insensitive

Significantly over-expressed in B

t-test with two tails.

p<0.05

Significantly over-expressed in A



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

GSA case study: functional differences in a class comparison experiment

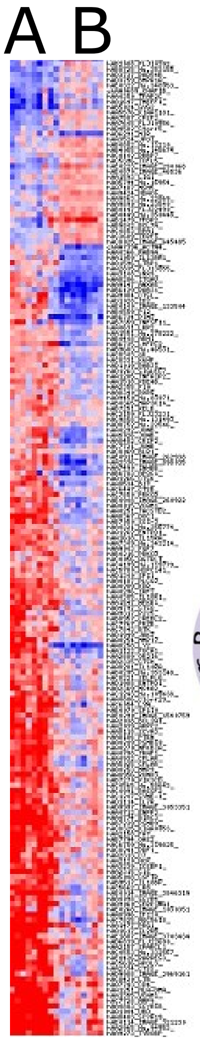
A

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

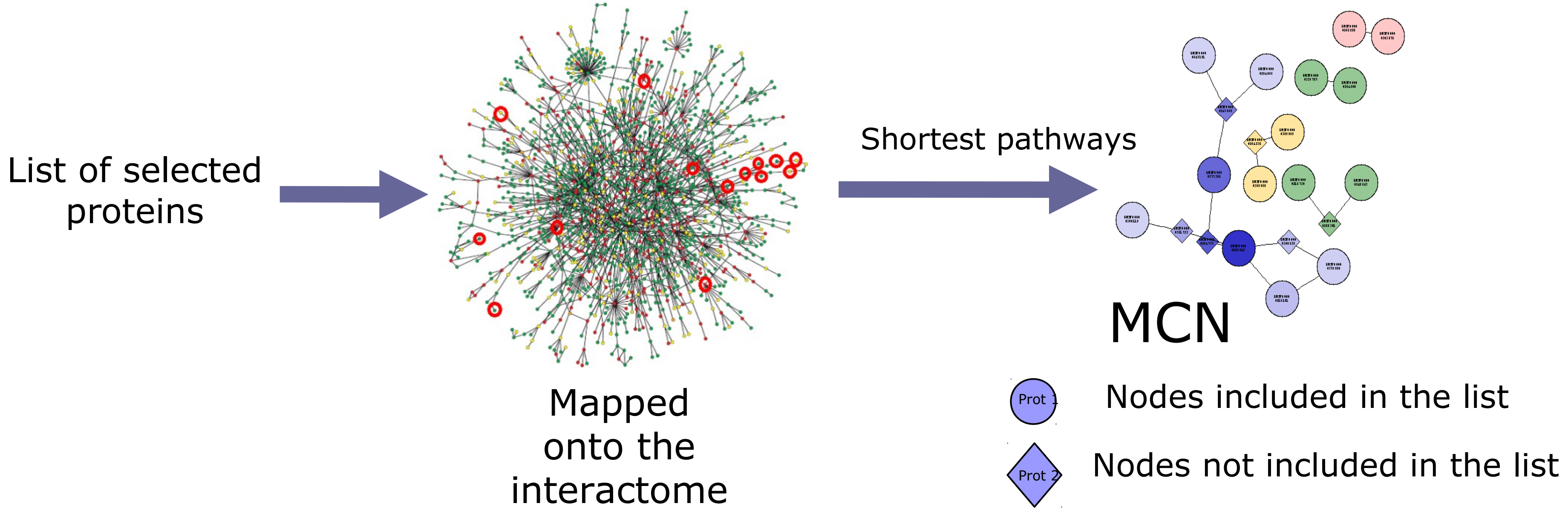
Healthy vs diabetic	Functional class	GO	KEGG
Up-regulated	Oxidative phosphorylation	X	X
	ATP synthesis		X
	Ribosome		X
	Mitochondrion	X	
	Nucleotide biosynthesis	X	
	NADH dehidrogenase (ubiquinone activity)	X	
	Nuclease activity	X	
Down-regulated	Insulin signalling pathway		X

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

Protein-protein interaction networks

Evaluation of the cooperative behaviour of a list of genes

Shortest pathways between all pairs of nodes in the list.
The minimum connection network (MCN)



List of selected proteins

Shortest pathways

Mapped onto the interactome

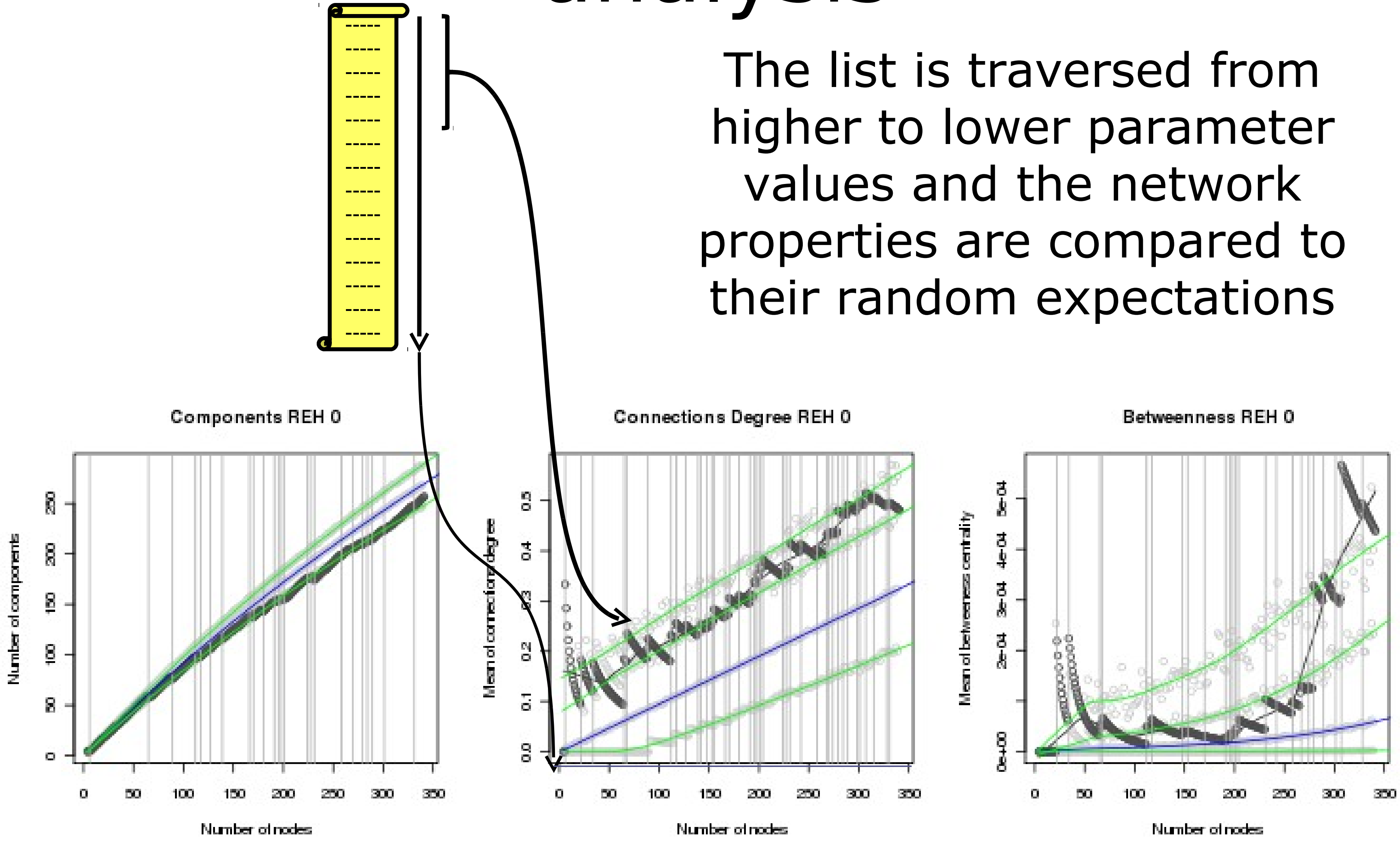
MCN

Nodes included in the list

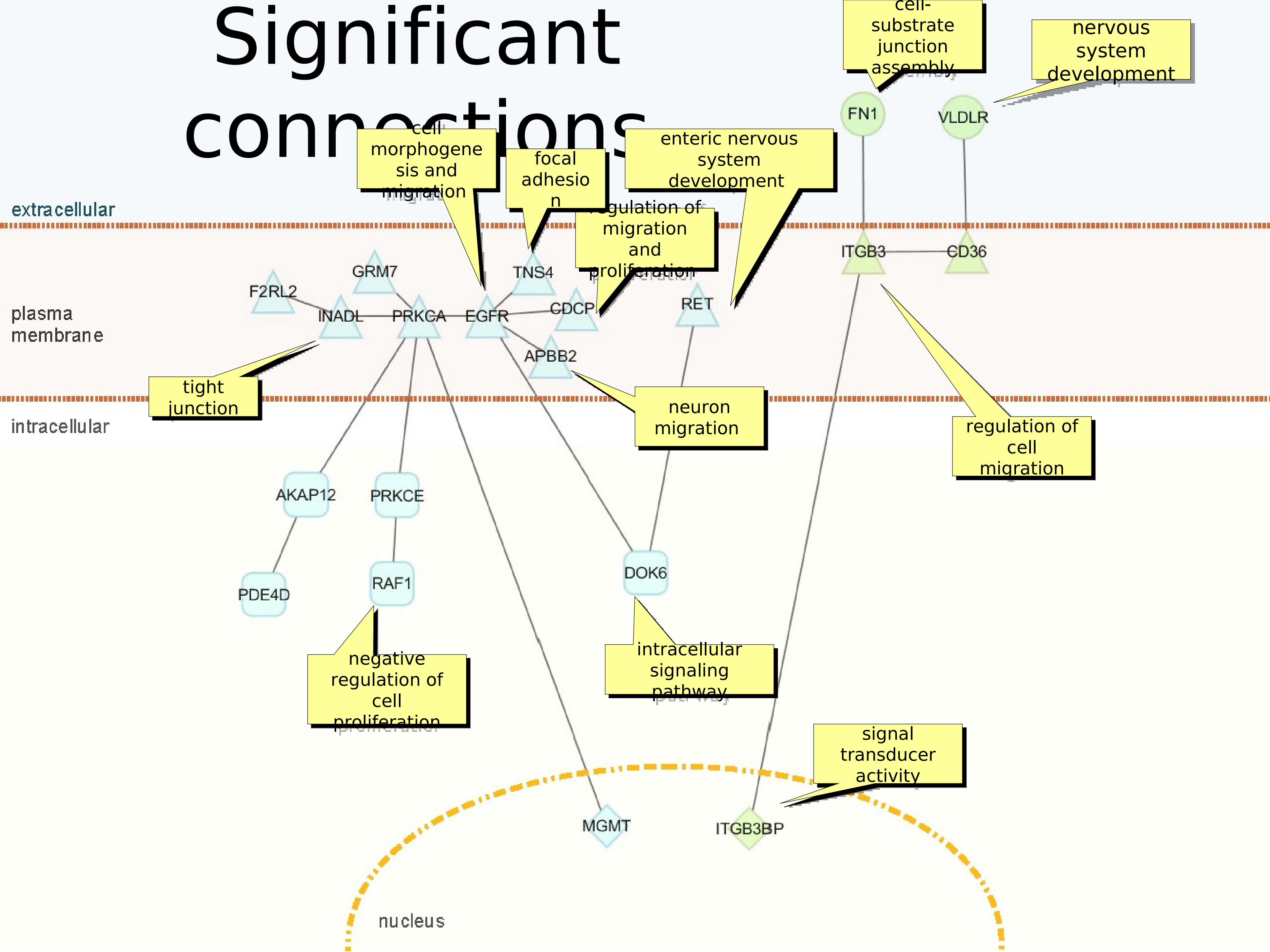
Nodes not included in the list

Gene-set-like network analysis

The list is traversed from higher to lower parameter values and the network properties are compared to their random expectations



Significant connections



cell-substrate junction assembly

nervous system development

cell morphogenesis and migration

focal adhesion

enteric nervous system development

regulation of migration and proliferation

tight junction

neuron migration

regulation of cell migration

negative regulation of cell proliferation

intracellular signaling pathway

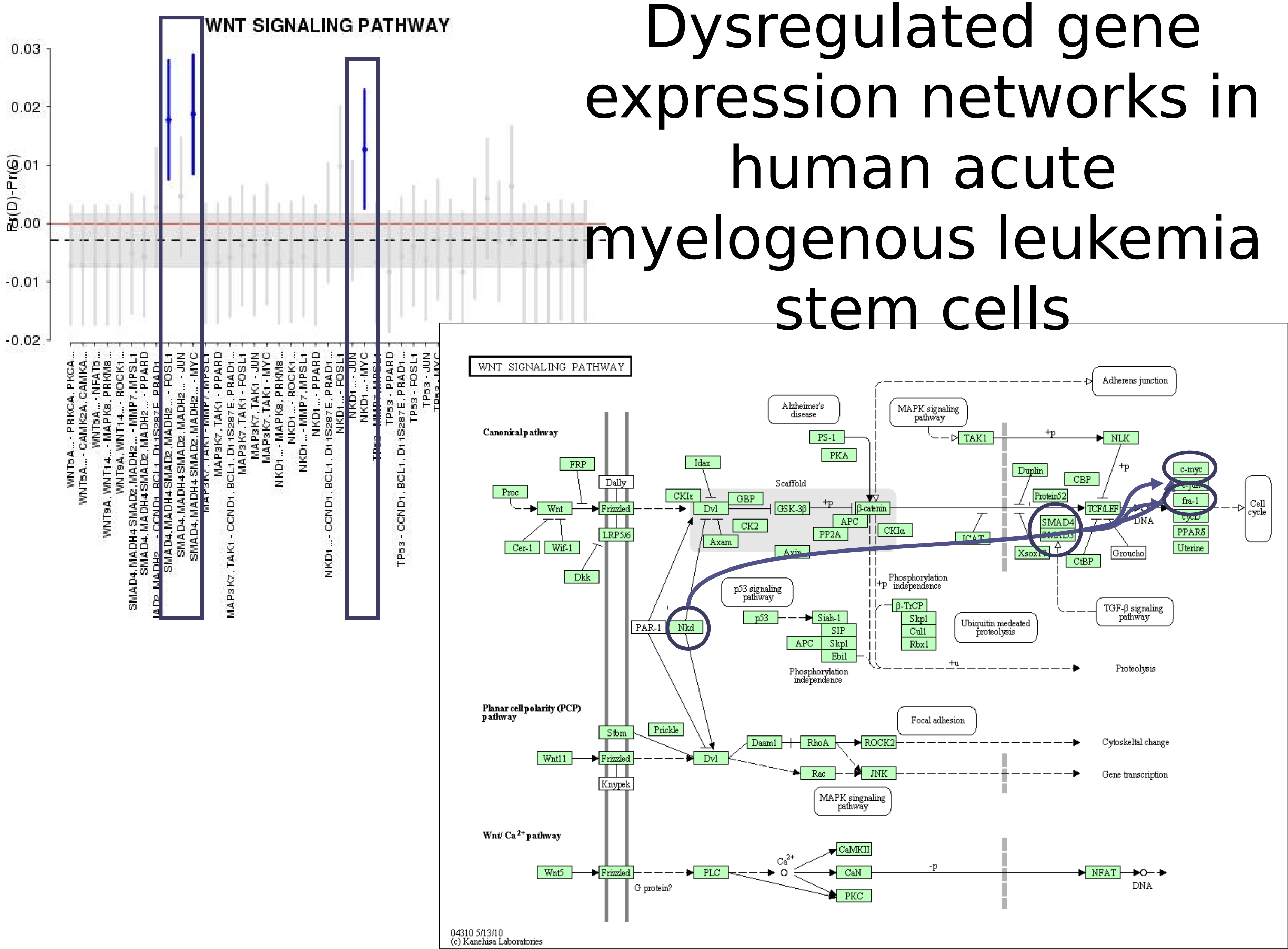
signal transducer activity

What is next?

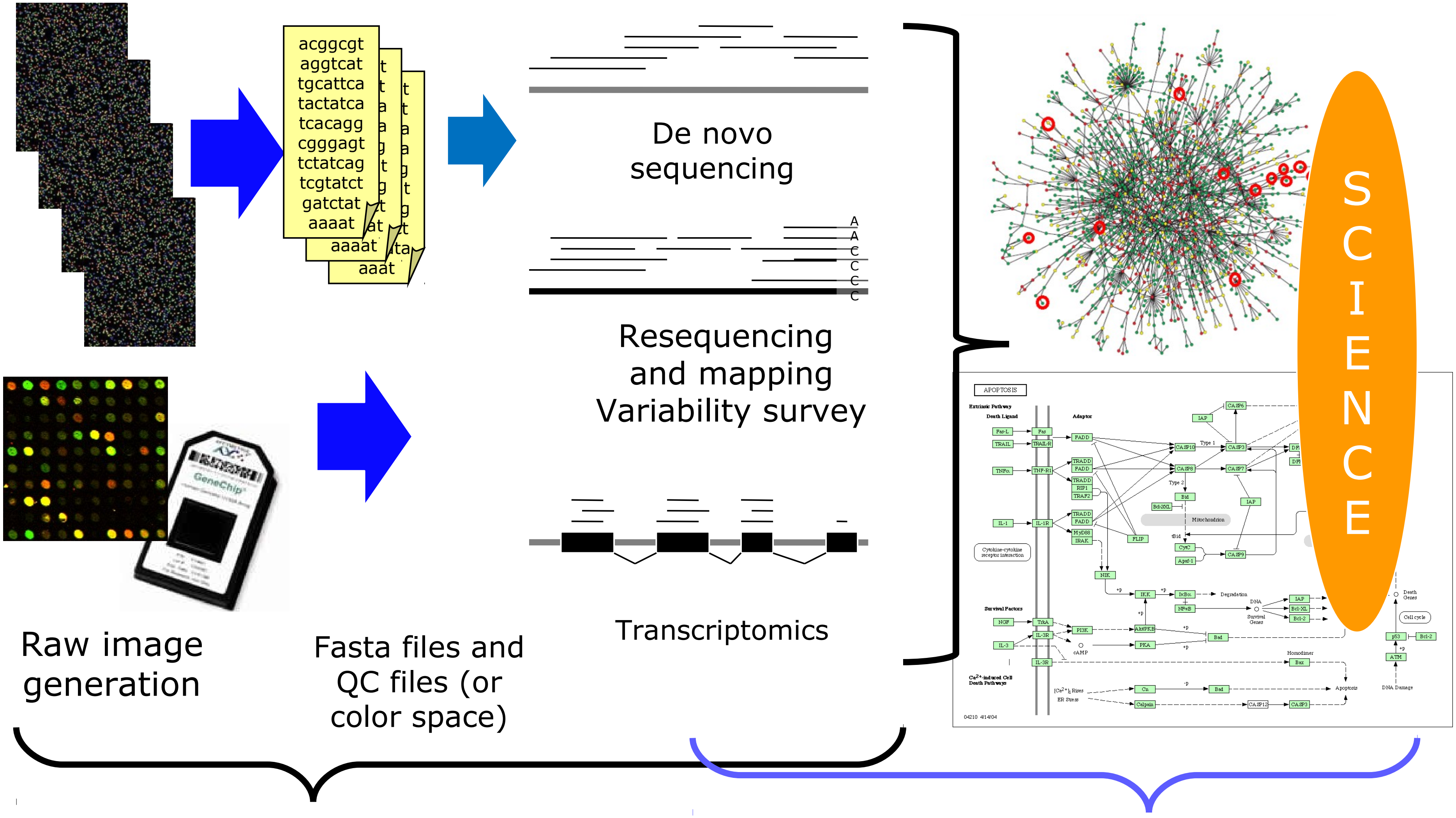
Functional classes have internal structure. Exploiting function and internal structure by modeling pathways

Method	Gene-based selection	Function-based selection	Function	Relationships among components
Functional enrichment	X		X	
Gene-set analysis		X	X	
Network enrichment	X			X
Network enrichment analysis		X		X
Pathway modeling		X	X	X

Example: Dysregulated gene expression networks in human acute myelogenous leukemia stem cells



Pipeline general of analysis



acggcgt
aggtcat
tgcattca
tactatca
tcacagg
cgggagt
tctatcag
tcgtatct
gatctat
aaaat

De novo sequencing

Resequencing and mapping Variability survey

Transcriptomics

Raw image generation

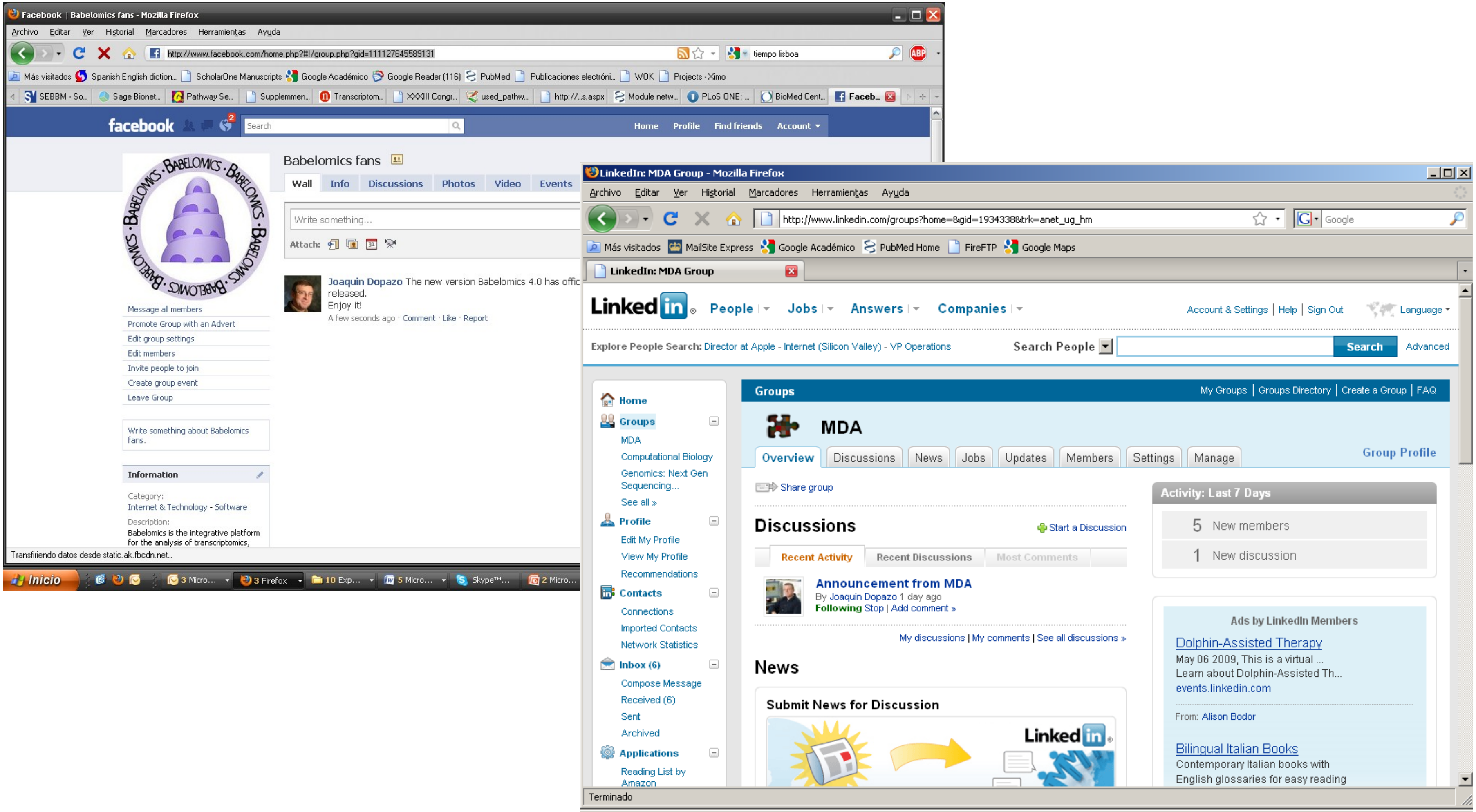
Fasta files and QC files (or color space)

Technology driven

Hypothesis driven

SOCIAL:

MDA group in Linked-in
Babelomics group in Facebook


Facebook | Babelomics fans - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://www.facebook.com/home.php?#!/group.php?gid=111127645589131
tiempo lisboa
Más visitados Spanish English diction... ScholarOne Manuscripts Google Académico Google Reader (116) PubMed Publicaciones electróni... WOK Projects - Ximo
SEBBM - So... Sage Bionet... Pathway Se... Supplemmen... Transcriptom... XXIII Congr... used_pathw... http://...s.aspx Module netw... PLoS ONE: ... BioMed Cent... Faceb...
facebook
Search
Home Profile Find friends Account
Wall Info Discussions Photos Video Events
Write something...
Attach:
Message all members
Promote Group with an Advert
Edit group settings
Edit members
Invite people to join
Create group event
Leave Group
Write something about Babelomics fans.
Information
Transfiriendo datos desde static.ak.fbcdn.net...
Inicio 3 Micro... 3 Firefox 10 Exp... 5 Micro... Skype™... 2 Micro...
LinkedIn: MDA Group - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://www.linkedin.com/groups?home=&gid=1934338&trk=anet_ug_hm
Google
Más visitados MailSite Express Google Académico PubMed Home FireFTP Google Maps
LinkedIn: MDA Group
Linked in
People Jobs Answers Companies
Account & Settings | Help | Sign Out Language
Search People
Search
Advanced
Home
Groups
MDA
Computational Biology
Genomics: Next Gen Sequencing...
See all »
Profile
Edit My Profile
View My Profile
Recommendations
Contacts
Connections
Imported Contacts
Network Statistics
Inbox (6)
Compose Message
Received (6)
Sent
Archived
Applications
Reading List by Amazon
Groups
My Groups | Groups Directory | Create a Group | FAQ
Overview Discussions News Jobs Updates Members Settings Manage
Group Profile
Share group
Start a Discussion
Recent Activity Recent Discussions Most Comments
My discussions | My comments | See all discussions »
Terminado

Babelomics fans

Joaquin Dopazo The new version Babelomics 4.0 has offic
released.
Enjoy it!
A few seconds ago · Comment · Like · Report

Category:
Internet & Technology - Software

Description:
Babelomics is the integrative platform for the analysis of transcriptomics,

Explore People Search: Director at Apple - Internet (Silicon Valley) - VP Operations

MDA

Discussions

Announcement from MDA
By Joaquin Dopazo 1 day ago
Following Stop | Add comment »

News

Submit News for Discussion

Activity: Last 7 Days

5 New members

1 New discussion

Ads by LinkedIn Members

Dolphin-Assisted Therapy
May 06 2009, This is a virtual ...
Learn about Dolphin-Assisted Th...
events.linkedin.com

From: Alison Bodor

Bilingual Italian Books
Contemporary Italian books with
English glossaries for easy reading

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

INB

...the Medical Genome Project (Sevilla)

ciberer

MINISTERIO
DE CIENCIA
E INNOVACIÓN

F.E.D.E.R.

ProyectoDeméter

Rticc

JUNTA DE ANDALUCIA
CONSEJERÍA DE SALUD

MGP Medical Genome Project

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY
AFFILIATED REGIONAL INSTITUT