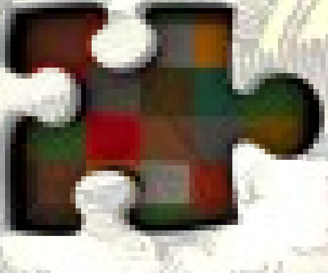
Massive Data Analysis



Introduction

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

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

INB
INSTITUTO NACIONAL

MINISTERIO DE EDUCACIÓN Y CIENCIA

RTICC

ciberer

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

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

INB

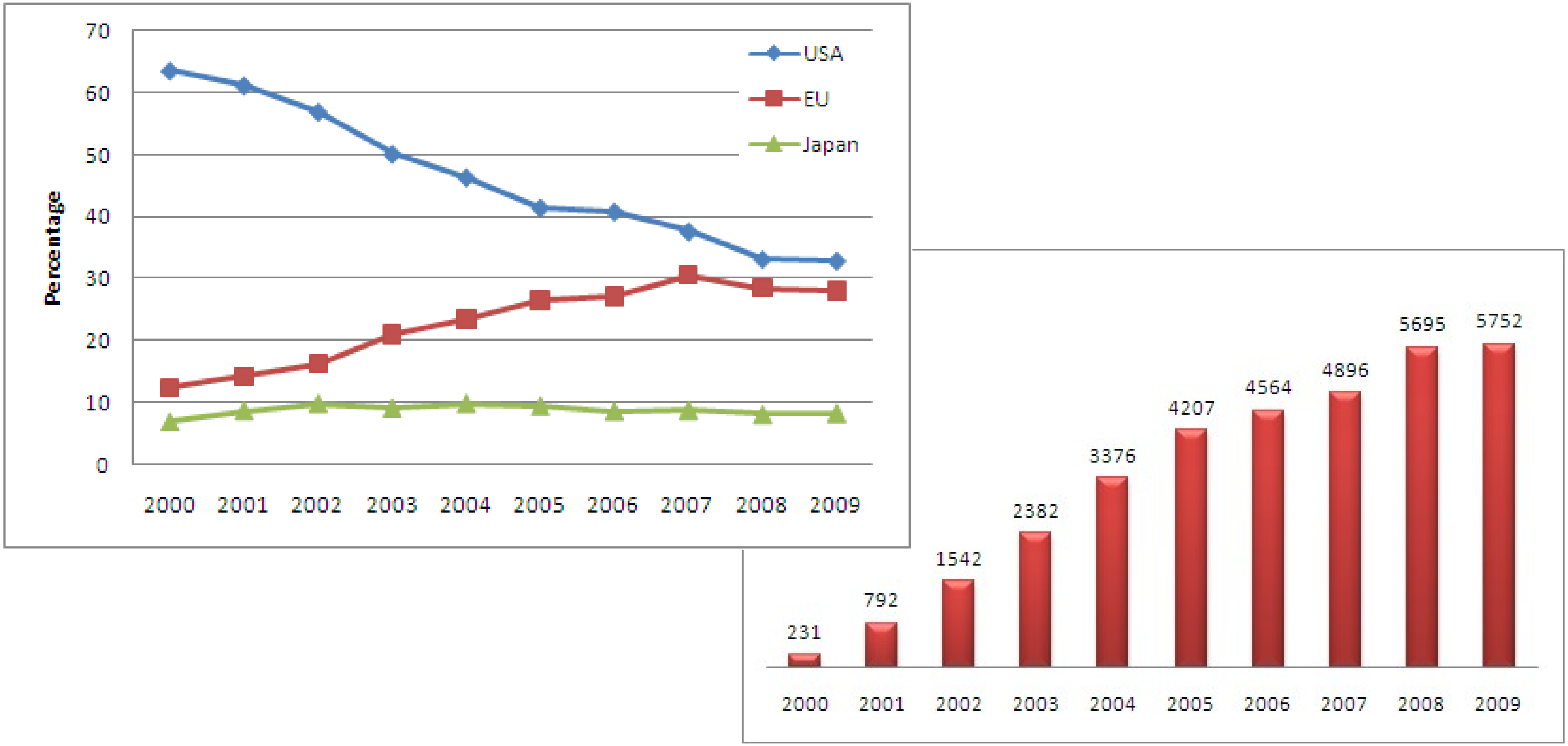
Rticc

ciberer

MGP Medical Genome Project

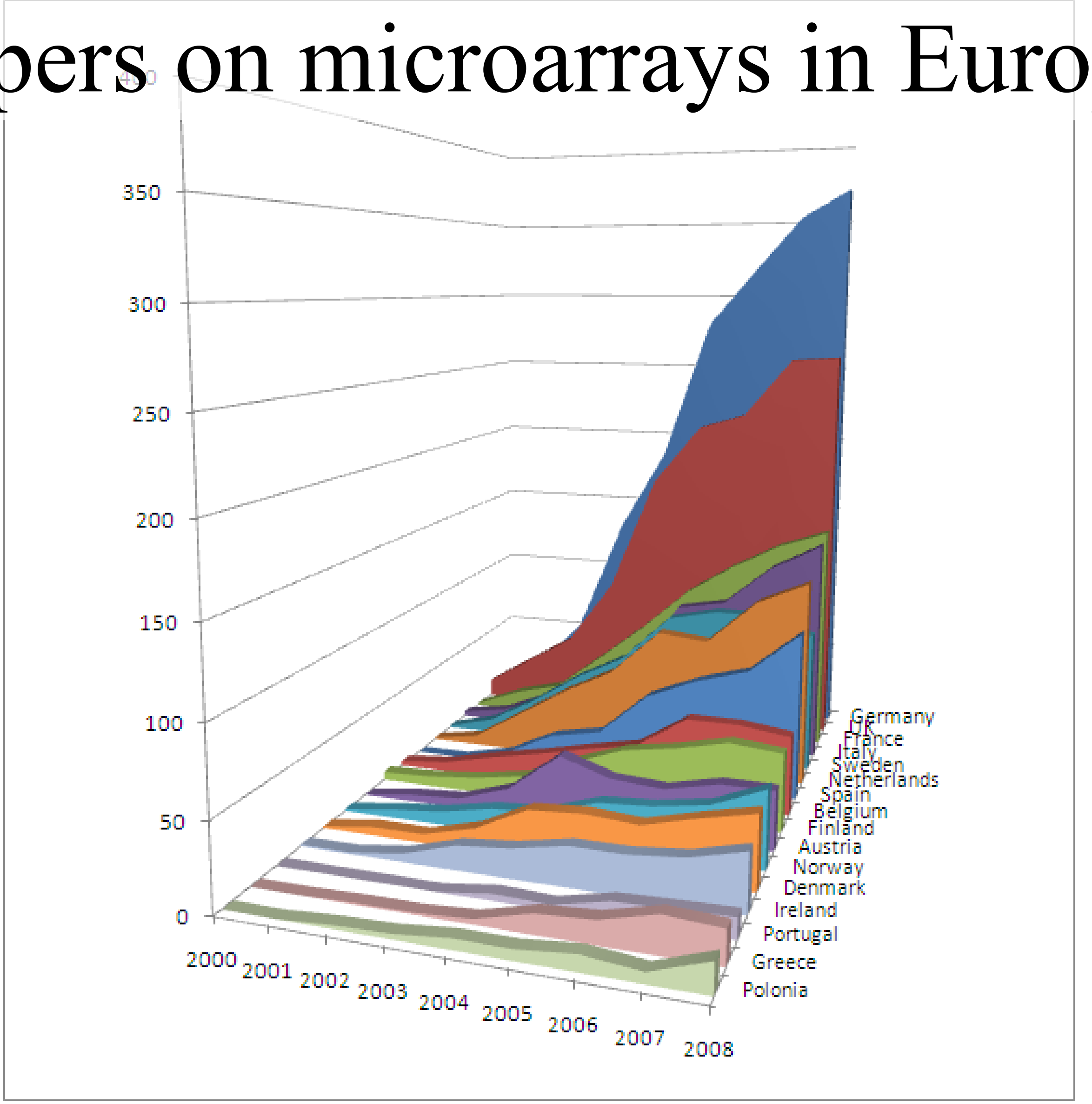
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Evolution of the percentages of published papers on microarrays



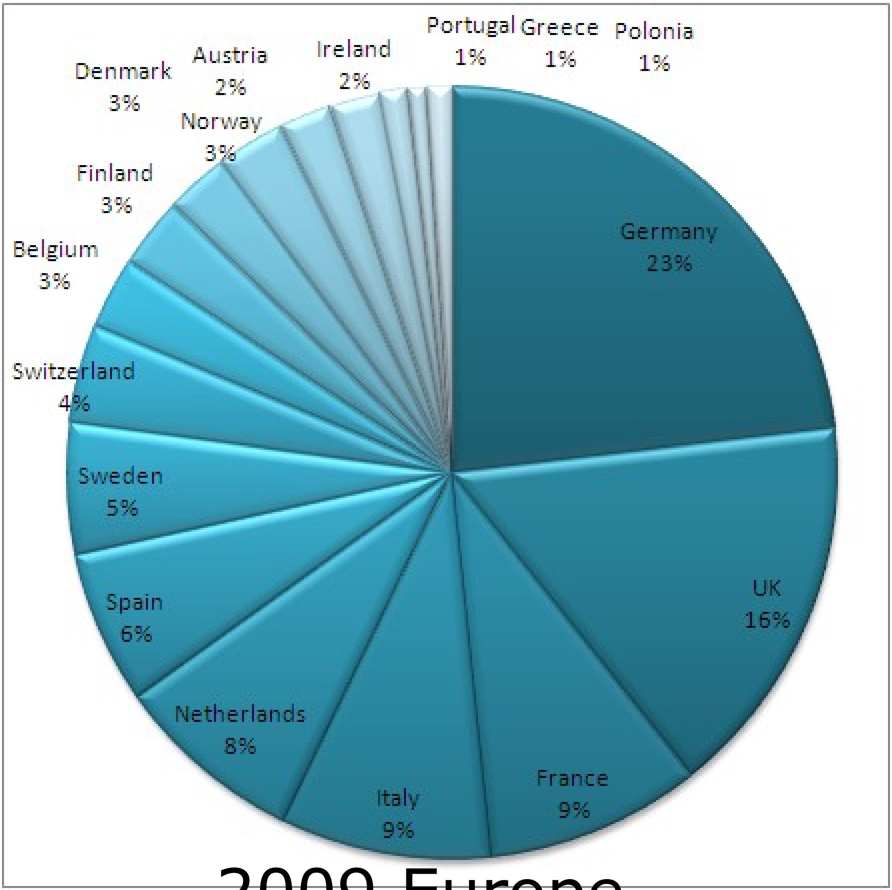
Source Pubmed. Query:
date[Entrez Date] AND country[Affiliation]AND microarray[Title/Abstract]

Evolution of the percentages of published papers on microarrays in Europe



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

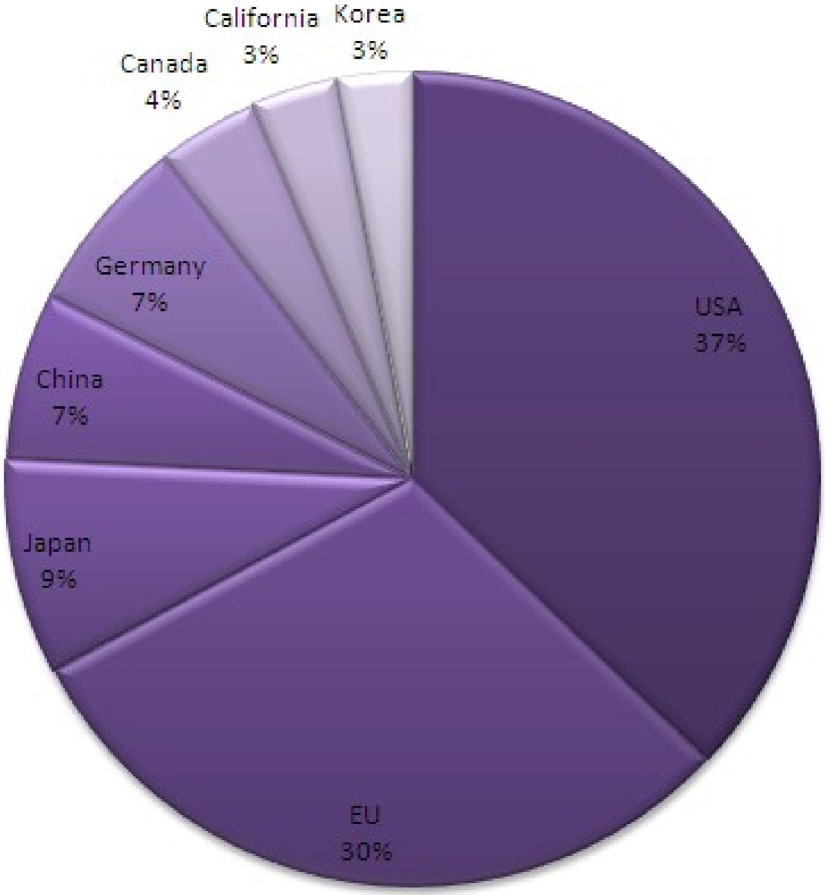
Microarray publications



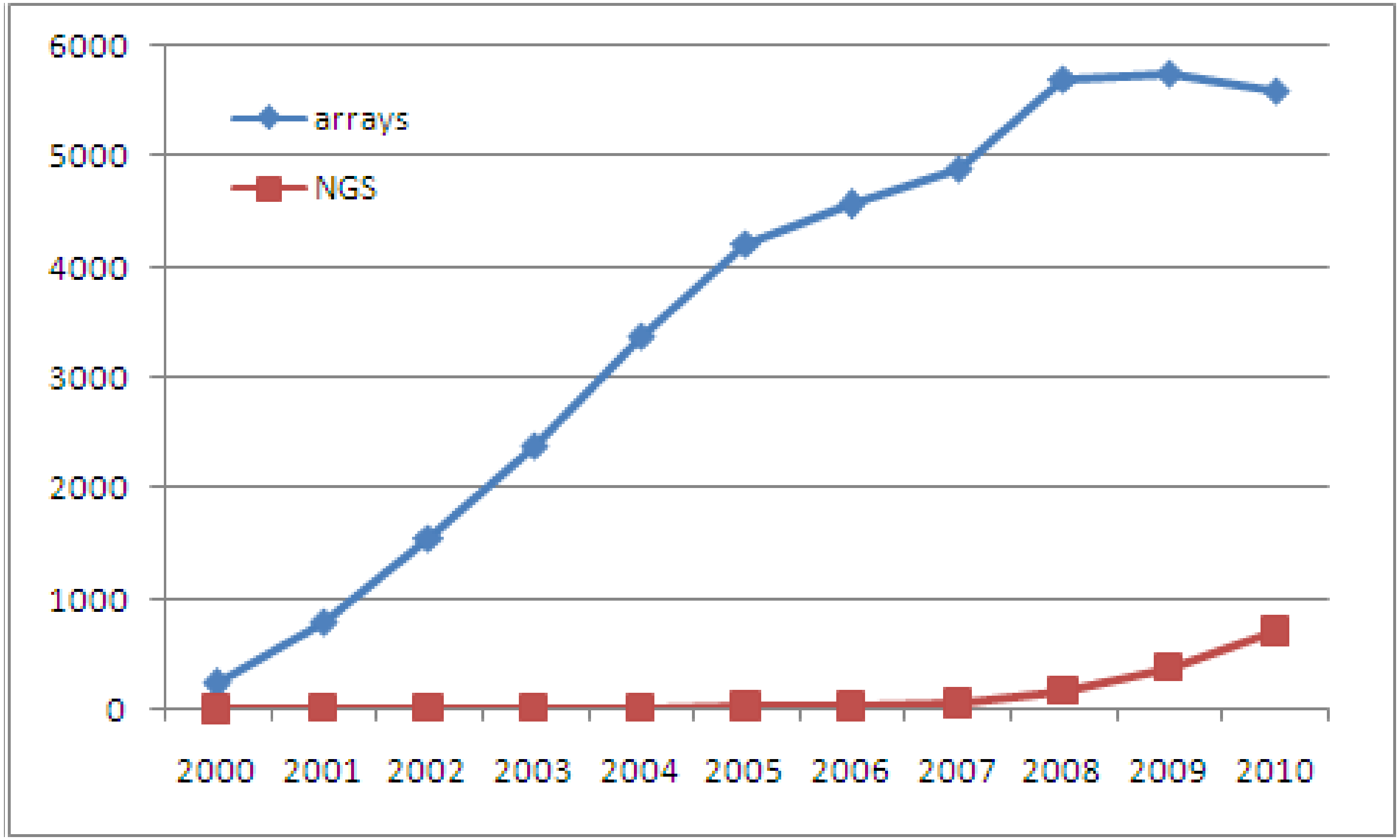
2009 Europe

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

2009 Worldwide



Trends in publications



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

Some numbers

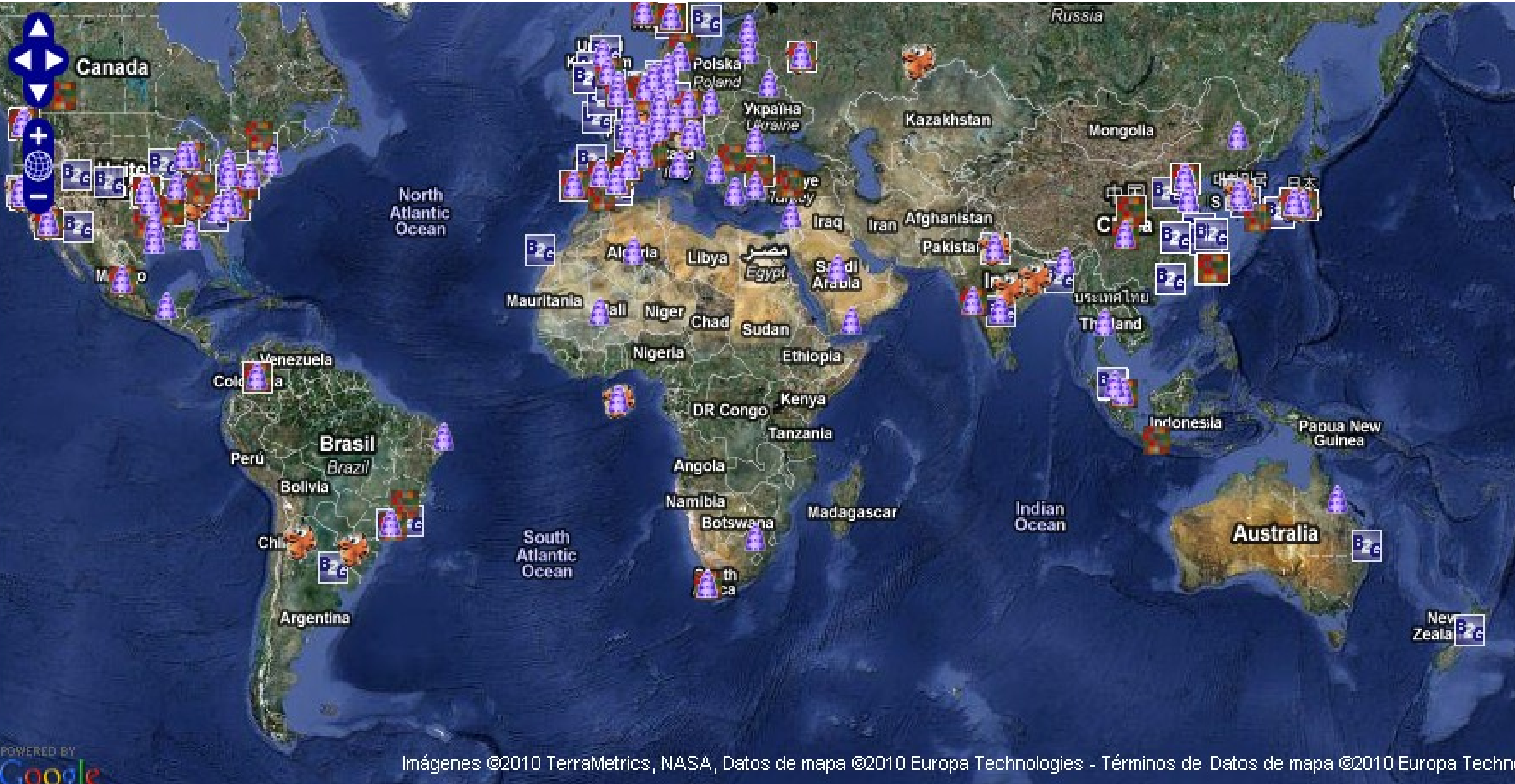
451 papers cite GEPAS (215 are SOTA cites)

632 papers cite Babelomics (442 are FatiGO cites)

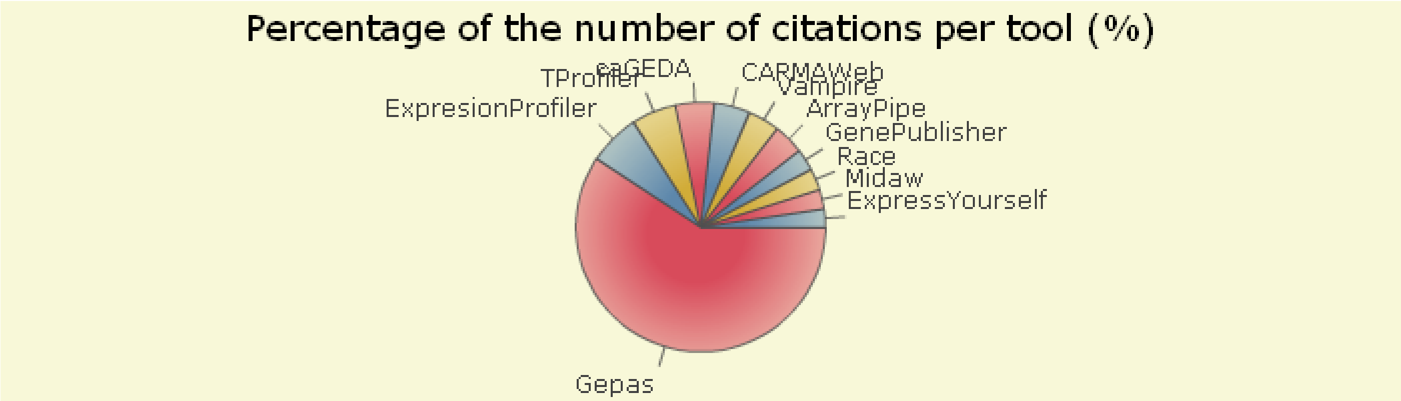
(source ISI Web of Knowledge, May 2010)

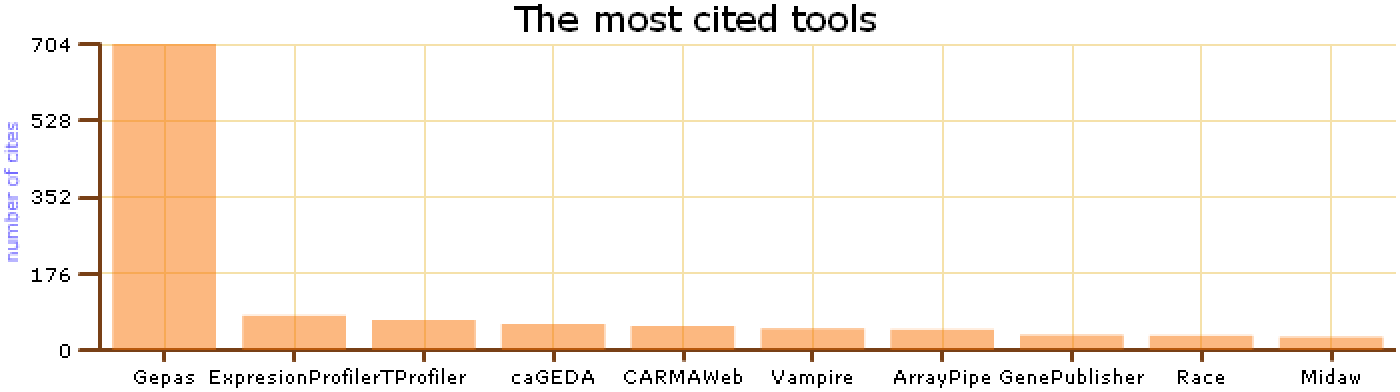
More than 150,000 experiments analysed during the last year.

More than 1000 experiments per day.

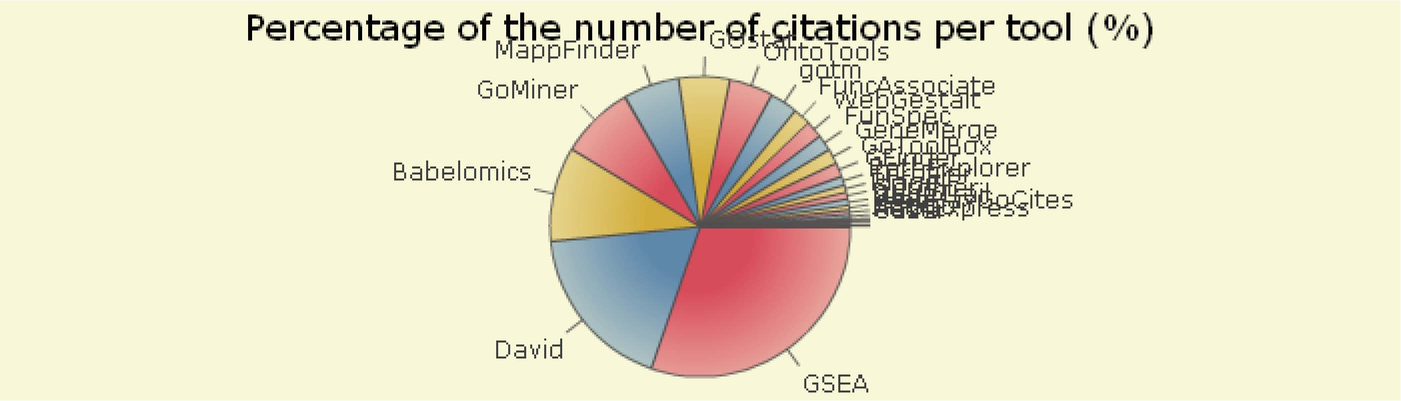


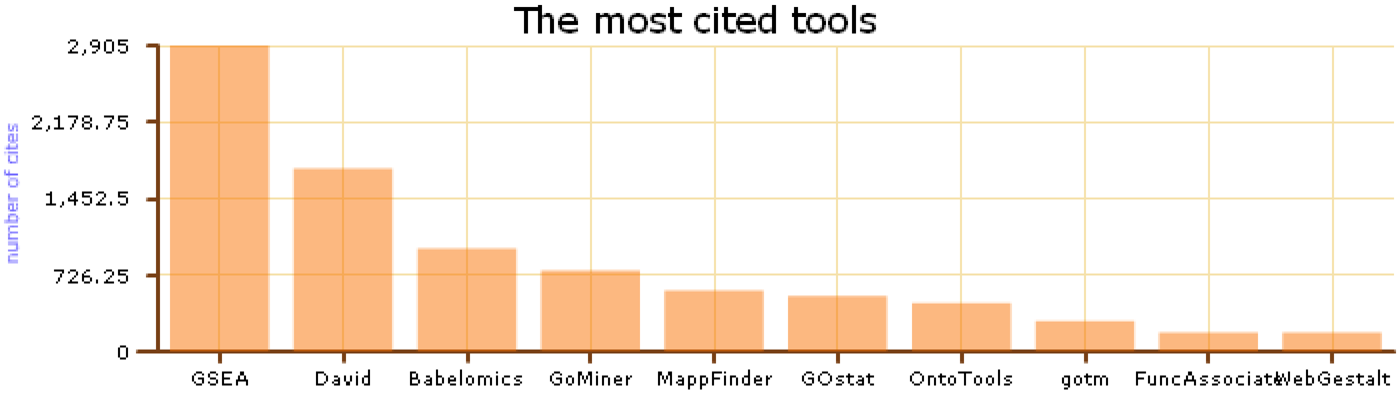
Tools for gene expression analysis





Tools for functional profiling


Percentage of the number of citations per tool (%)


The most cited tools

Structure of the course

Theoretical

Hands-on **GEPAS**

Introduction

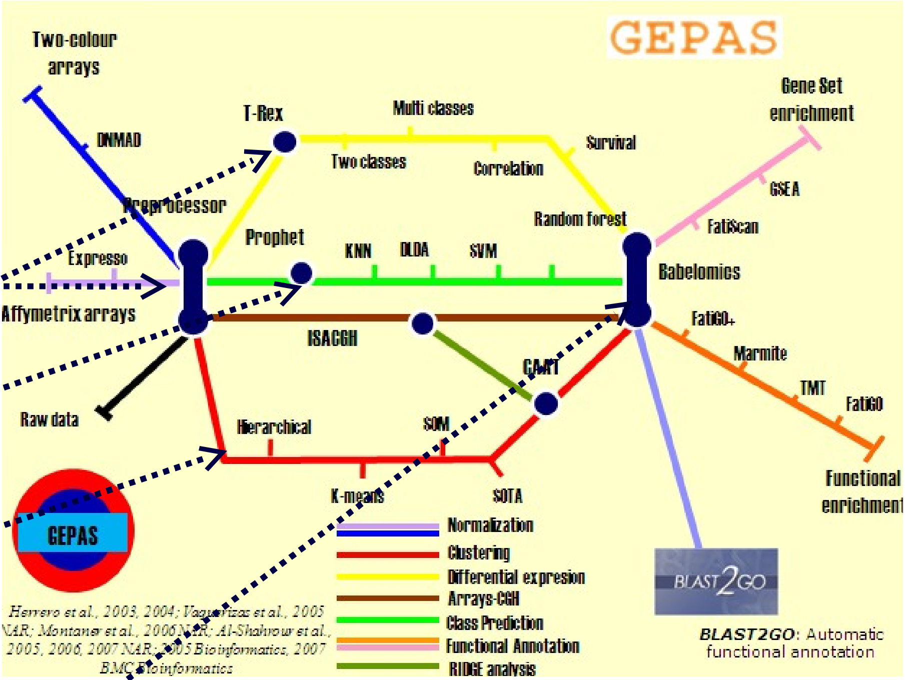
Normalization

Gene selection

Predictors

Clustering

Functional interpretation



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

Background

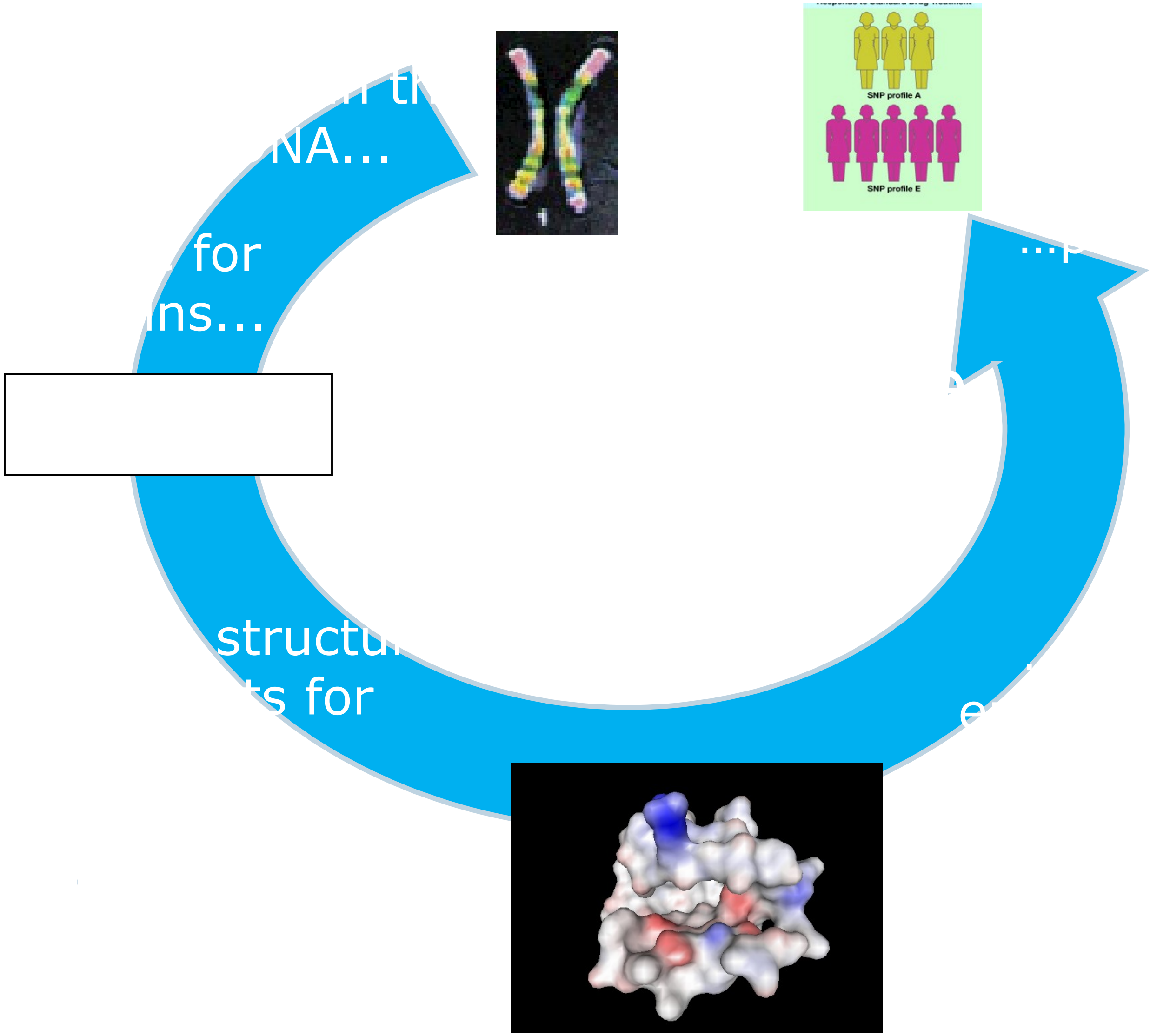
The road of excess leads to the palace of wisdom

(*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



n th
NA...

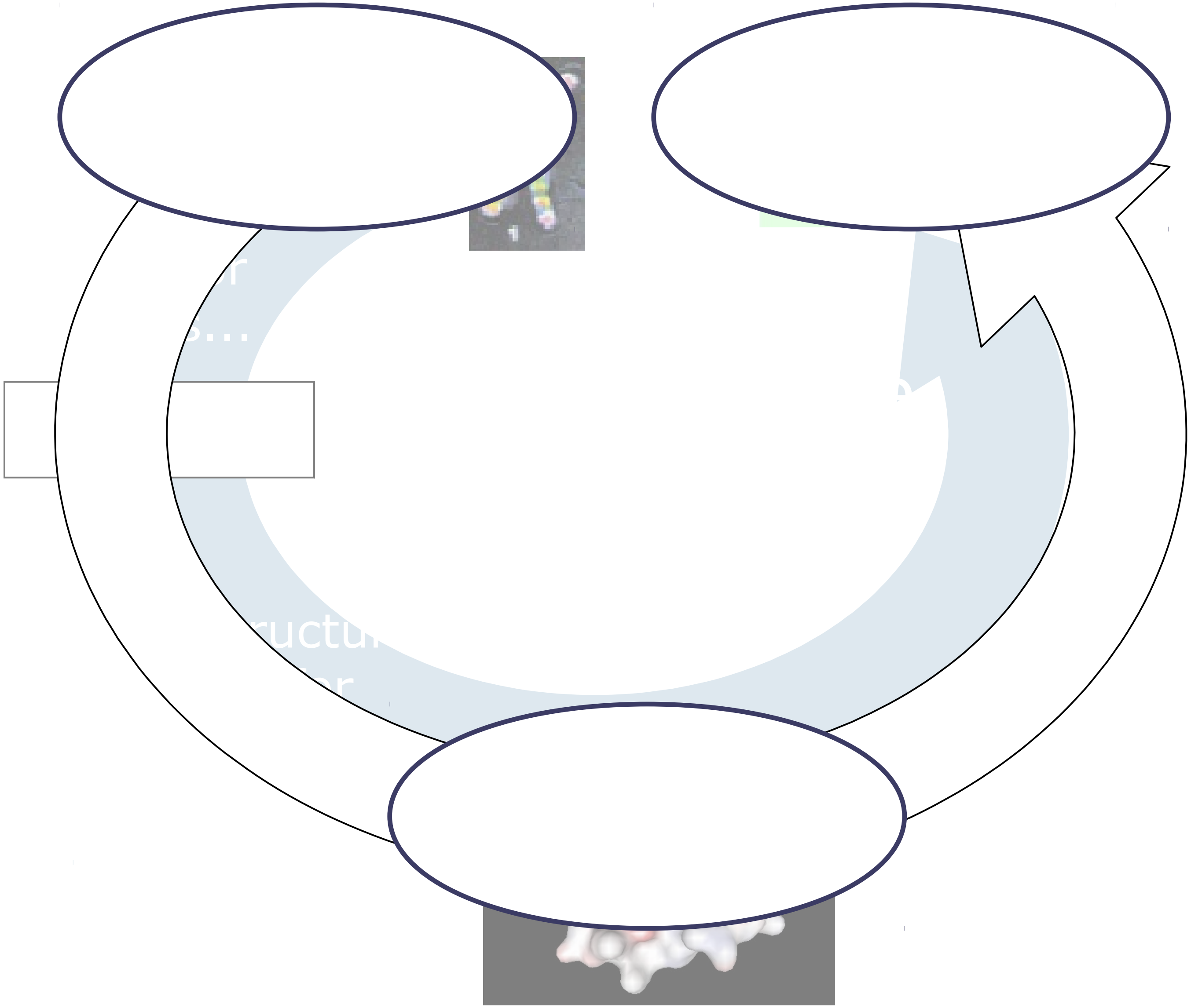
for
ns...

structu
ts for

...p

e

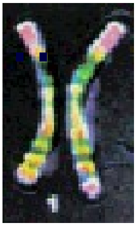
Reduccionistic approach to link causes (genome) to effects (phenotype) through actions (function)



From genotype to phenotype.

(in the functional post-genomics scenario)

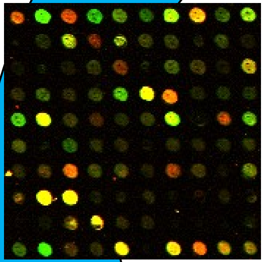
Genes in the DNA...



>protein kunase
acctgttgatggcgacagggactgtatgctgatct
atgctgatgcatgcatgctgactactgatgtgggg
gctattgacttgatgtctatc....

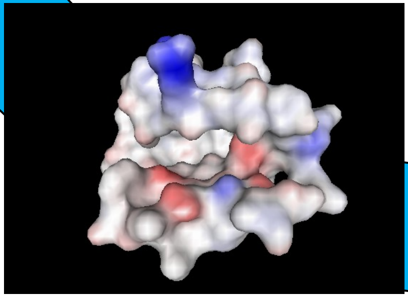
Next Generation Sequencing
10^9bp per round
(10^{10} expected soon)

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



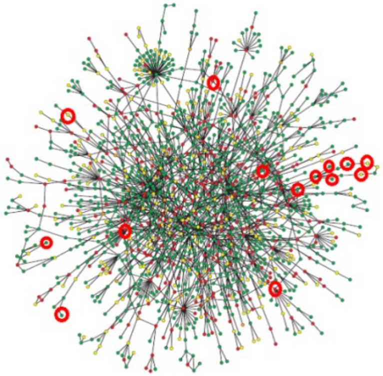
A typical tissue is expressing among 5000 and 10000 genes

...code for proteins...



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

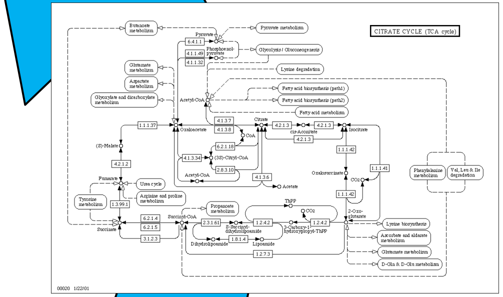
...that account for function if...



...in cooperation with other proteins...

Each protein has an average of 8 interactions

...conforming complex interaction networks...

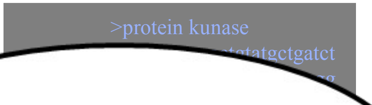


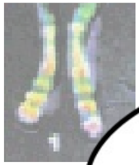
...whose final effect configures the phenotype...

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

15 million SNPs

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


>protein kunase
tgtatgctgatct



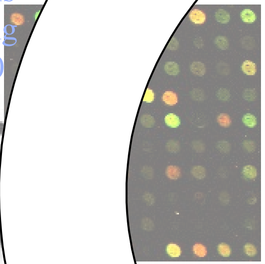
...whose final effect configures the phenotype...

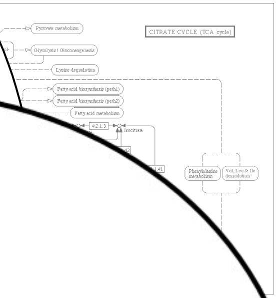
Causes

Effects

...when ex proper n place

A typical tissue is expressing among 5000 and 10000 genes





From genotype to phenotyp

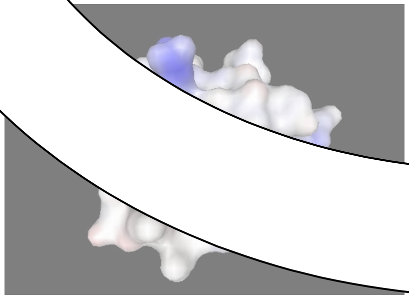
(in the functional genomics scen

Function (modules of proteins)

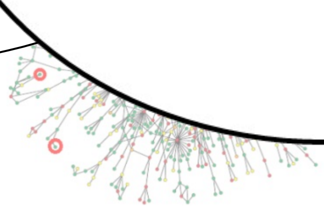
...code proteins.

plex ks...

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



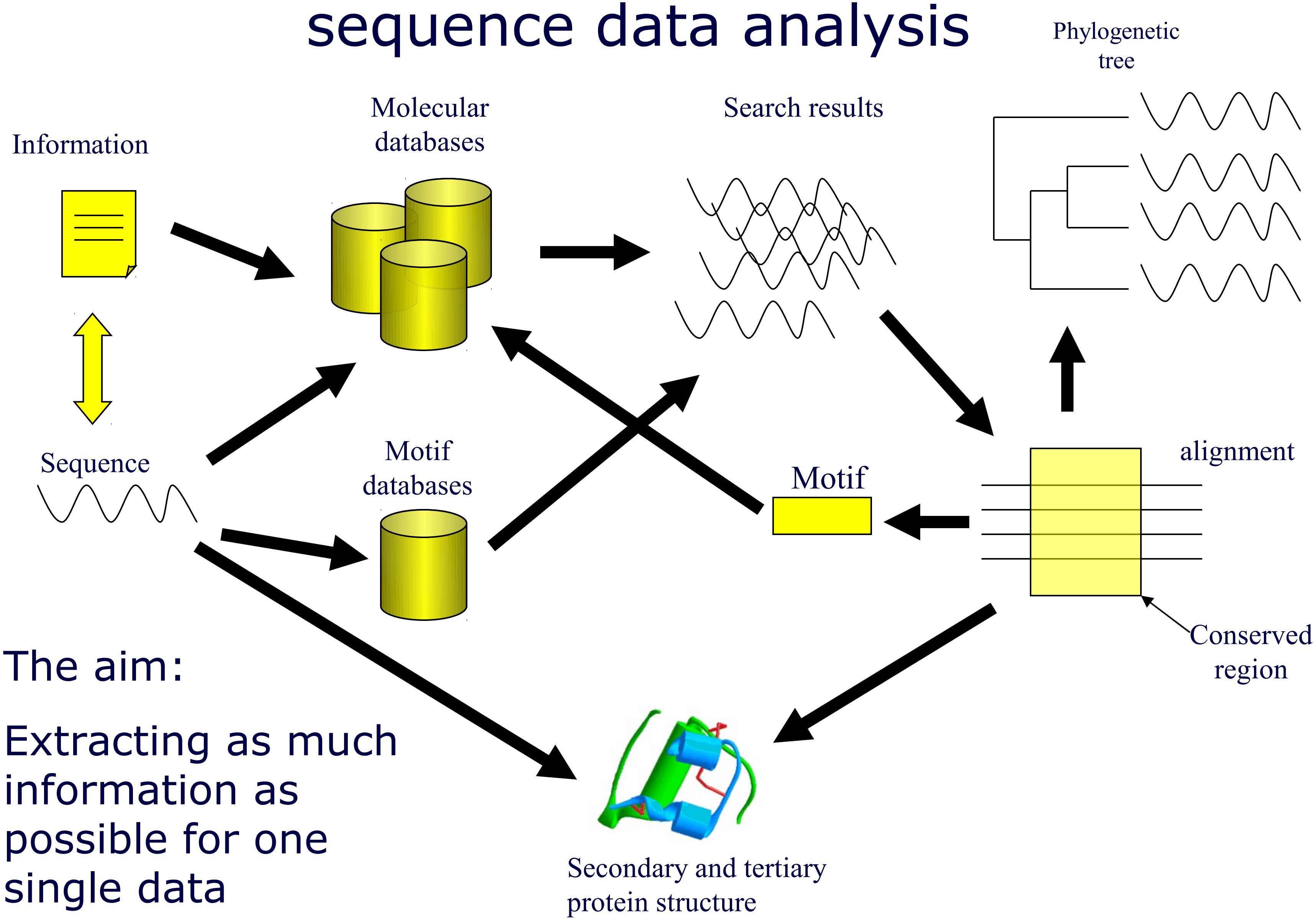
ration other proteins...



...whose structures account for function...

Each protein has an average of 8 interactions

Bioinformatics tools for pre-genomic sequence data analysis



Phylogenetic tree

Information

Molecular databases

Search results

Sequence

Motif databases

Motif

alignment

Conserved region

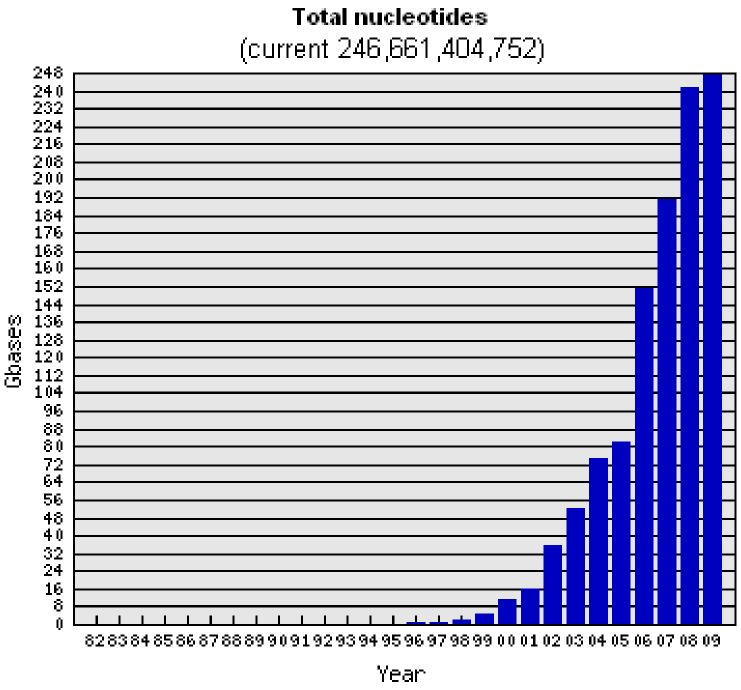
Secondary and tertiary protein structure

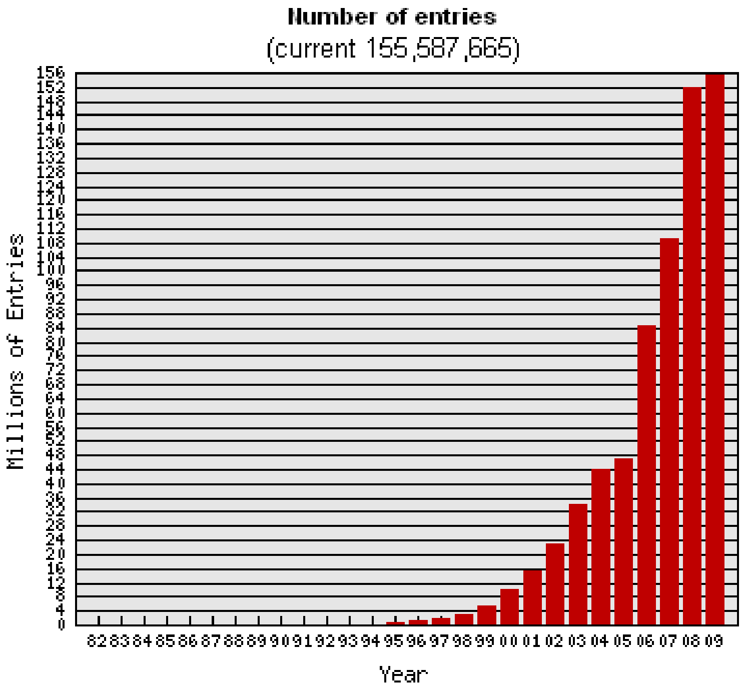
The aim:

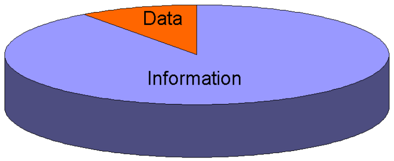
Extracting as much information as possible for one single data

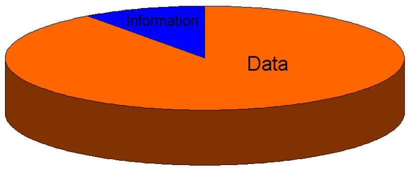
Post-genomic vision

EMBL database growth (March 2009)









Genome scale data and a note of caution on associations, correlations or patterns discovered:

Genome-wide technologies allows us to produce vast amounts of data.
But... dealing with many data (omic data) increase the occurrence of spurious associations due to chance

Hypothesis → Experiment → test

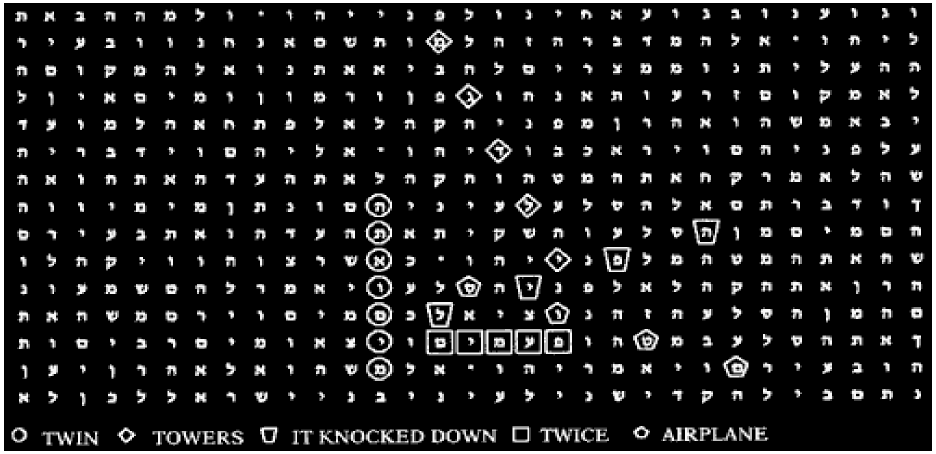
Is gene A involved in process B?

Experiment → (sometimes) test → Hypothesis

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

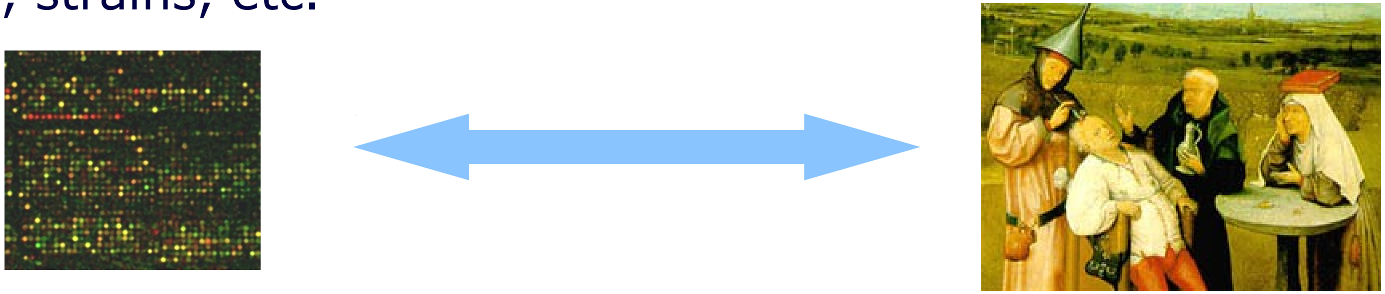
Sure, but... Is it real? (many hypotheses are rejected while this one is accepted *a posteriori*: numerology)

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



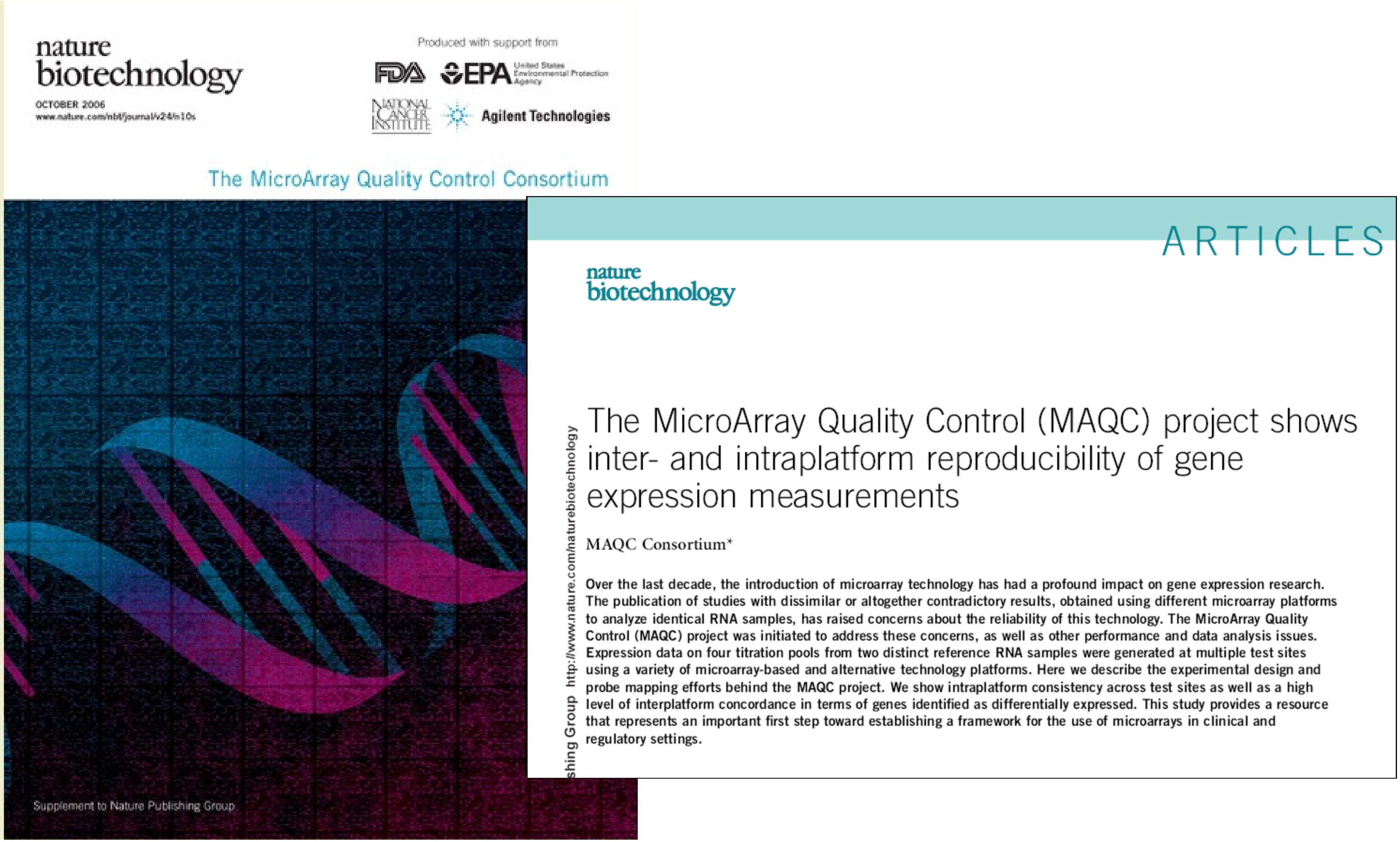
Gene expression profiling. Historic perspective

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



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

Microarrays arrive to an acceptable level of reproducibility



nature biotechnology

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

Produced with support from

FDA

EPA United States Environmental Protection Agency

NATIONAL CANCER INSTITUTE

Agilent Technologies

The MicroArray Quality Control Consortium

Supplement to Nature Publishing Group

ARTICLES

nature biotechnology

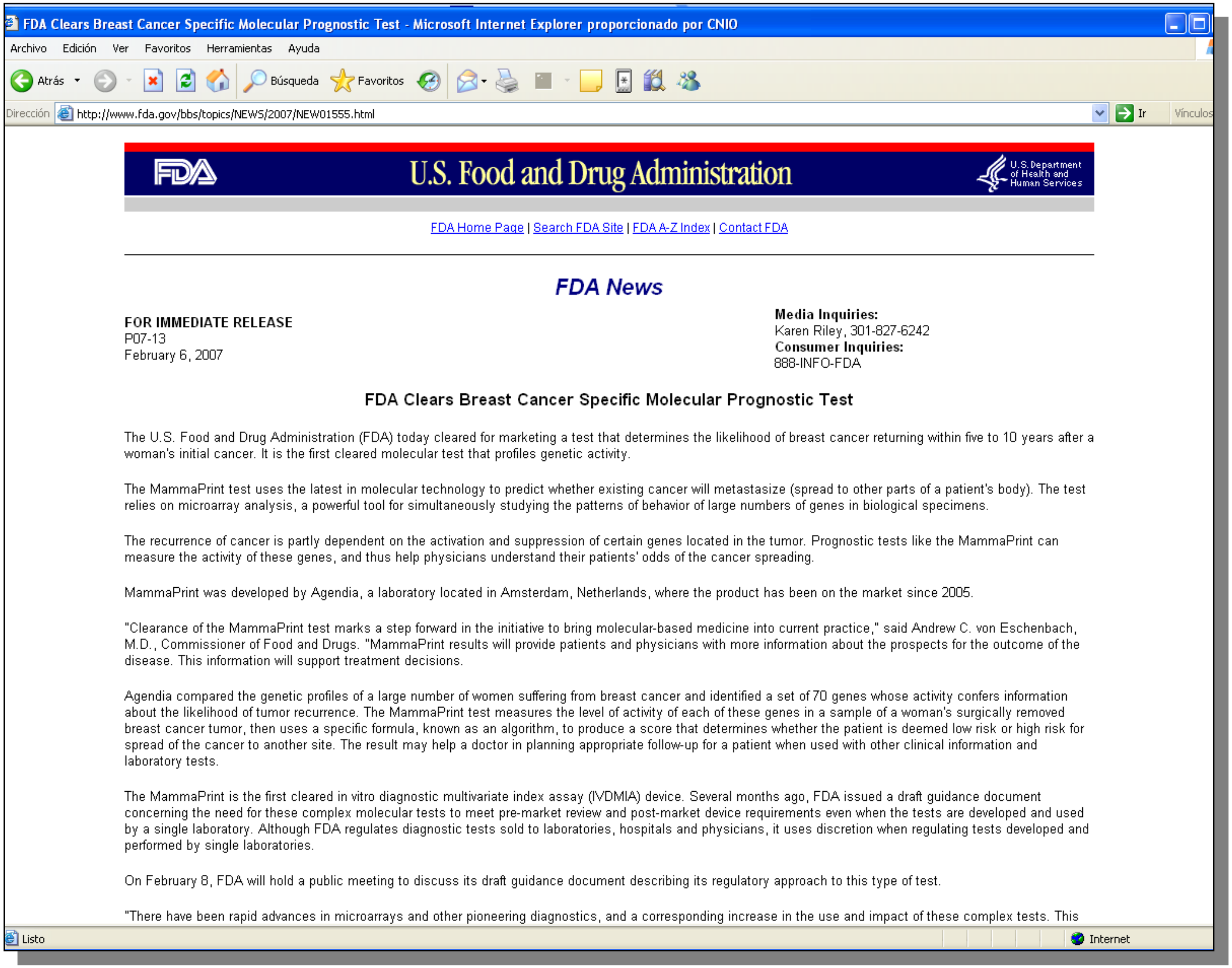
The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements

MAQC Consortium*

Over the last decade, the introduction of microarray technology has had a profound impact on gene expression research. The publication of studies with dissimilar or altogether contradictory results, obtained using different microarray platforms to analyze identical RNA samples, has raised concerns about the reliability of this technology. The MicroArray Quality Control (MAQC) project was initiated to address these concerns, as well as other performance and data analysis issues. Expression data on four titration pools from two distinct reference RNA samples were generated at multiple test sites using a variety of microarray-based and alternative technology platforms. Here we describe the experimental design and probe mapping efforts behind the MAQC project. We show intraplatform consistency across test sites as well as a high level of interplatform concordance in terms of genes identified as differentially expressed. This study provides a resource that represents an important first step toward establishing a framework for the use of microarrays in clinical and regulatory settings.

shing Group http://www.nature.com/naturebiotechnology

FDA approves the first predictor based on microarrays


FDA Clears Breast Cancer Specific Molecular Prognostic Test - Microsoft Internet Explorer proporcionado por CNIO
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás
Búsqueda
Favoritos
Dirección
http://www.fda.gov/bbs/topics/NEWS/2007/NEW01555.html
Ir
Vínculos
FDA
U.S. Food and Drug Administration
U.S. Department of Health and Human Services
FDA Home Page | Search FDA Site | FDA A-Z Index | Contact FDA
Listo
Internet

FDA News

FOR IMMEDIATE RELEASE
P07-13
February 6, 2007

Media Inquiries:
Karen Riley, 301-827-6242
Consumer Inquiries:
888-INFO-FDA

FDA Clears Breast Cancer Specific Molecular Prognostic Test

The U.S. Food and Drug Administration (FDA) today cleared for marketing a test that determines the likelihood of breast cancer returning within five to 10 years after a woman's initial cancer. It is the first cleared molecular test that profiles genetic activity.

The MammaPrint test uses the latest in molecular technology to predict whether existing cancer will metastasize (spread to other parts of a patient's body). The test relies on microarray analysis, a powerful tool for simultaneously studying the patterns of behavior of large numbers of genes in biological specimens.

The recurrence of cancer is partly dependent on the activation and suppression of certain genes located in the tumor. Prognostic tests like the MammaPrint can measure the activity of these genes, and thus help physicians understand their patients' odds of the cancer spreading.

MammaPrint was developed by Agendia, a laboratory located in Amsterdam, Netherlands, where the product has been on the market since 2005.

"Clearance of the MammaPrint test marks a step forward in the initiative to bring molecular-based medicine into current practice," said Andrew C. von Eschenbach, M.D., Commissioner of Food and Drugs. "MammaPrint results will provide patients and physicians with more information about the prospects for the outcome of the disease. This information will support treatment decisions.

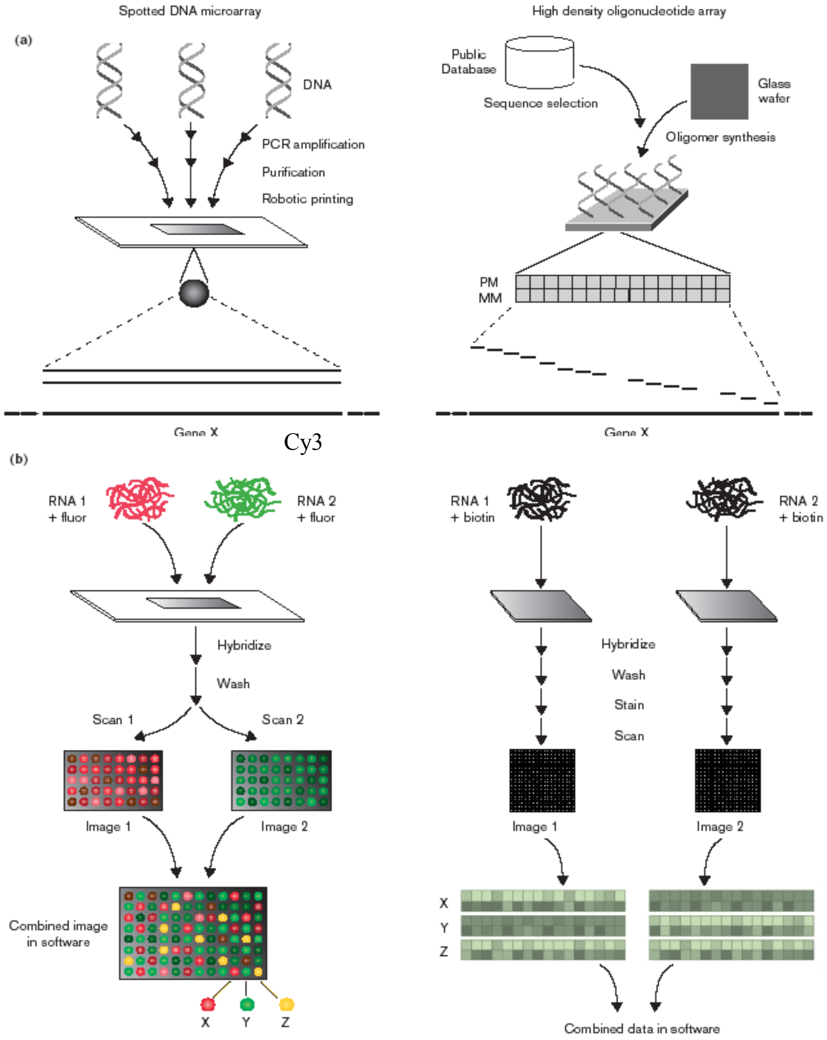
Agendia compared the genetic profiles of a large number of women suffering from breast cancer and identified a set of 70 genes whose activity confers information about the likelihood of tumor recurrence. The MammaPrint test measures the level of activity of each of these genes in a sample of a woman's surgically removed breast cancer tumor, then uses a specific formula, known as an algorithm, to produce a score that determines whether the patient is deemed low risk or high risk for spread of the cancer to another site. The result may help a doctor in planning appropriate follow-up for a patient when used with other clinical information and laboratory tests.

The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device. Several months ago, FDA issued a draft guidance document concerning the need for these complex molecular tests to meet pre-market review and post-market device requirements even when the tests are developed and used by a single laboratory. Although FDA regulates diagnostic tests sold to laboratories, hospitals and physicians, it uses discretion when regulating tests developed and performed by single laboratories.

On February 8, FDA will hold a public meeting to discuss its draft guidance document describing its regulatory approach to this type of test.

"There have been rapid advances in microarrays and other pioneering diagnostics, and a corresponding increase in the use and impact of these complex tests. This

DNA microarrays: the paradigm of a post-genomic technique



Competitive hybridization (two colors)

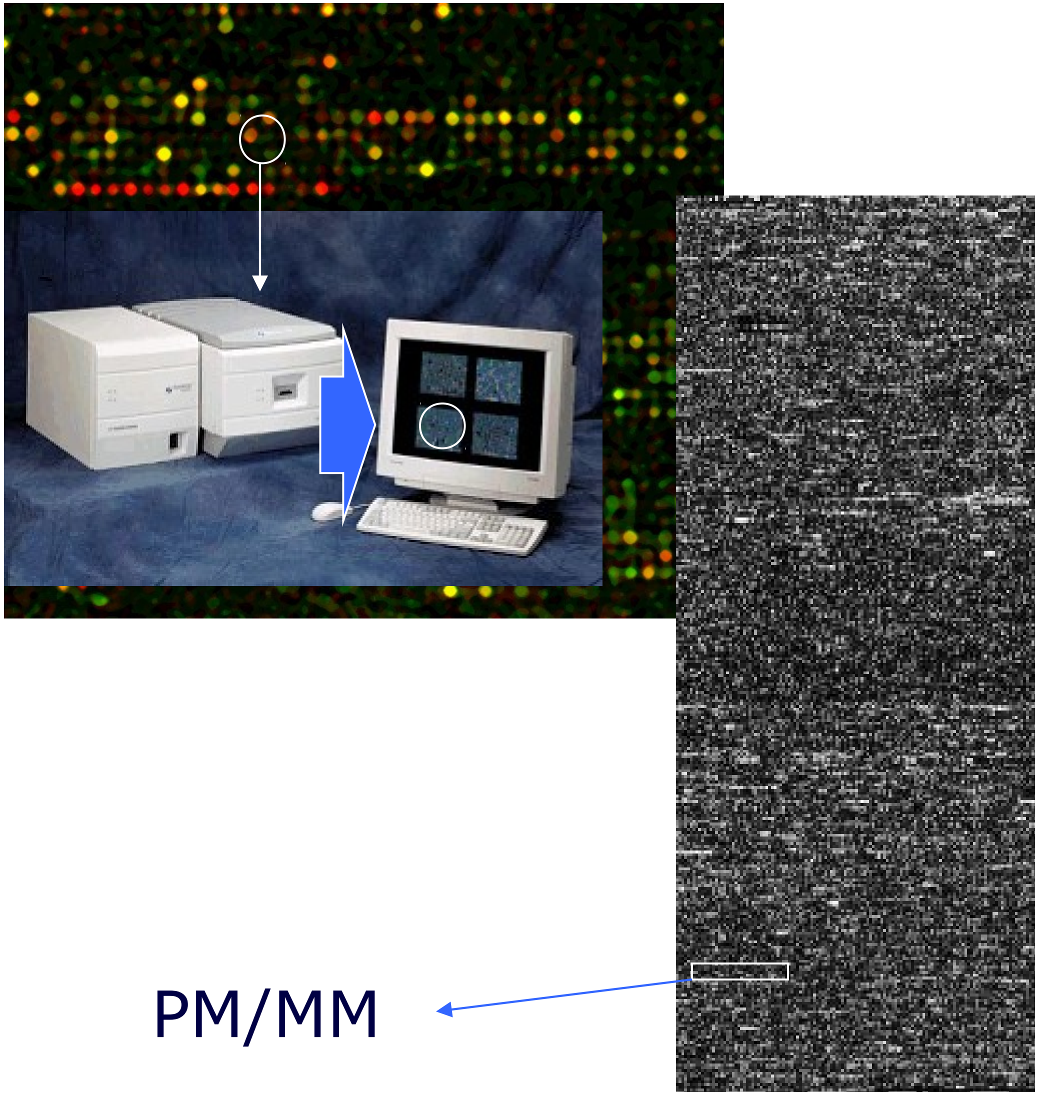
One color

Primary analysis

- Transform images corresponding to hybridization intensities into numbers

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

Transforming images into numbers



PM/MM

Two-color
Test sample labeled red (**Cy5**)
Reference sample labeled green (**Cy3**)
Red : gene overexpressed in test sample
Green : gene underexpressed in test sample
Yellow - equally expressed
red/green - ratio of expression

One color
Intensity of a gene using the probes

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

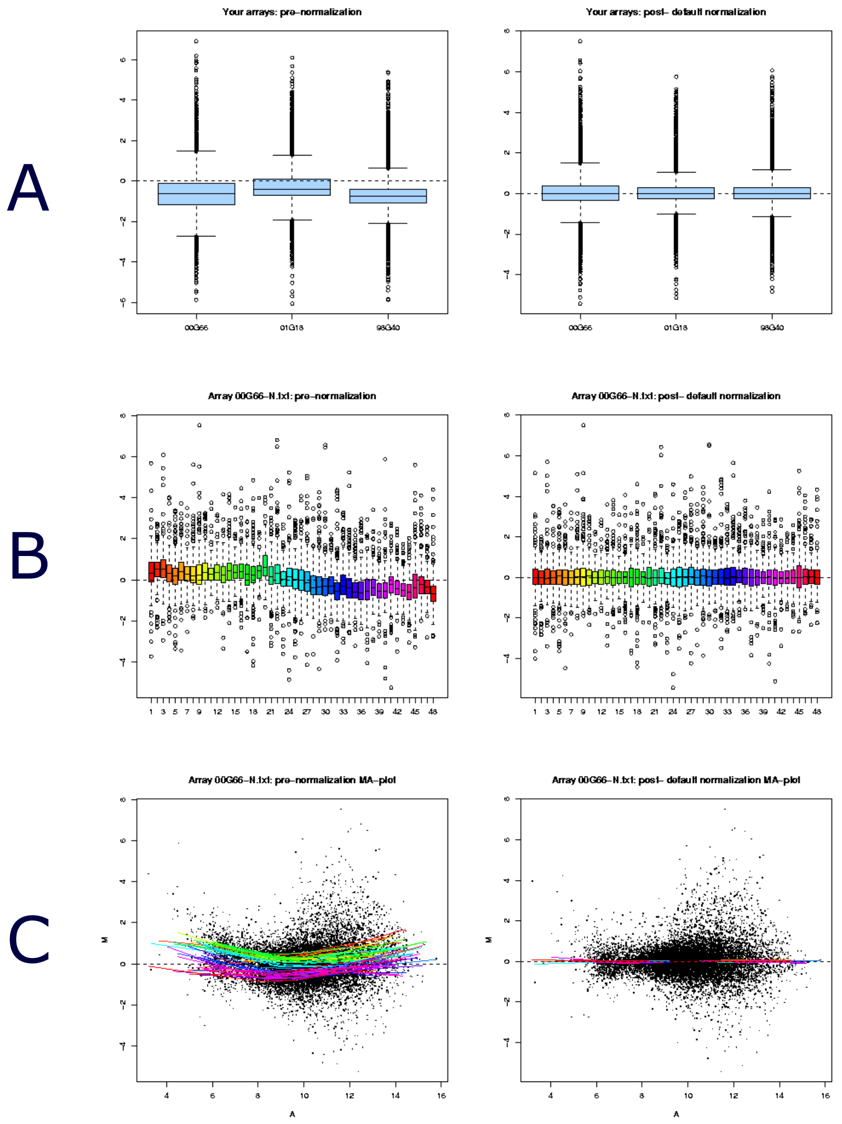
Scanners generate a graphic file.

Software analyzes the file: GenePix Pro (by Axon Instruments, Inc.) or Imagene (By Biodiscovery, Inc.)
There are free systems too: TIGR Spotfinder, ScanAlyze, etc

Normalisation

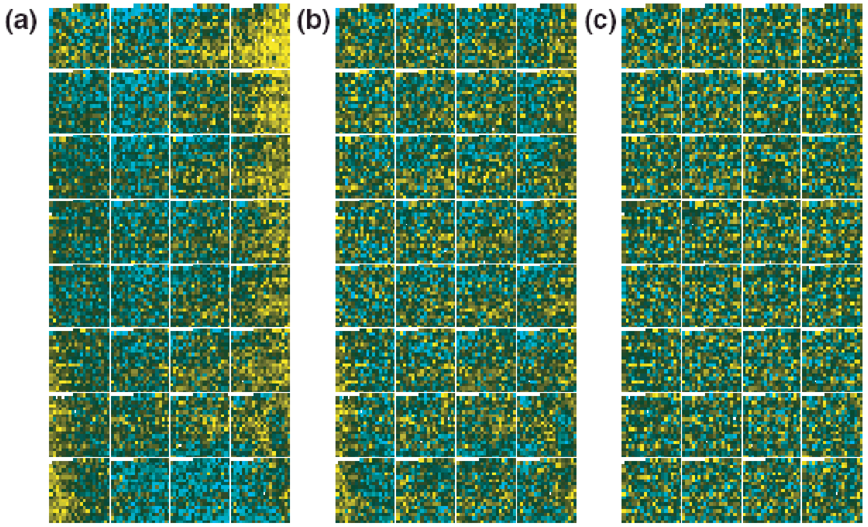
There are many sources of error that can affect and seriously bias the interpretation of the results. Differences in the efficiency of labelling, the hybridisation, local effects, etc.

Normalisation is a necessary step before proceeding with the analysis



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

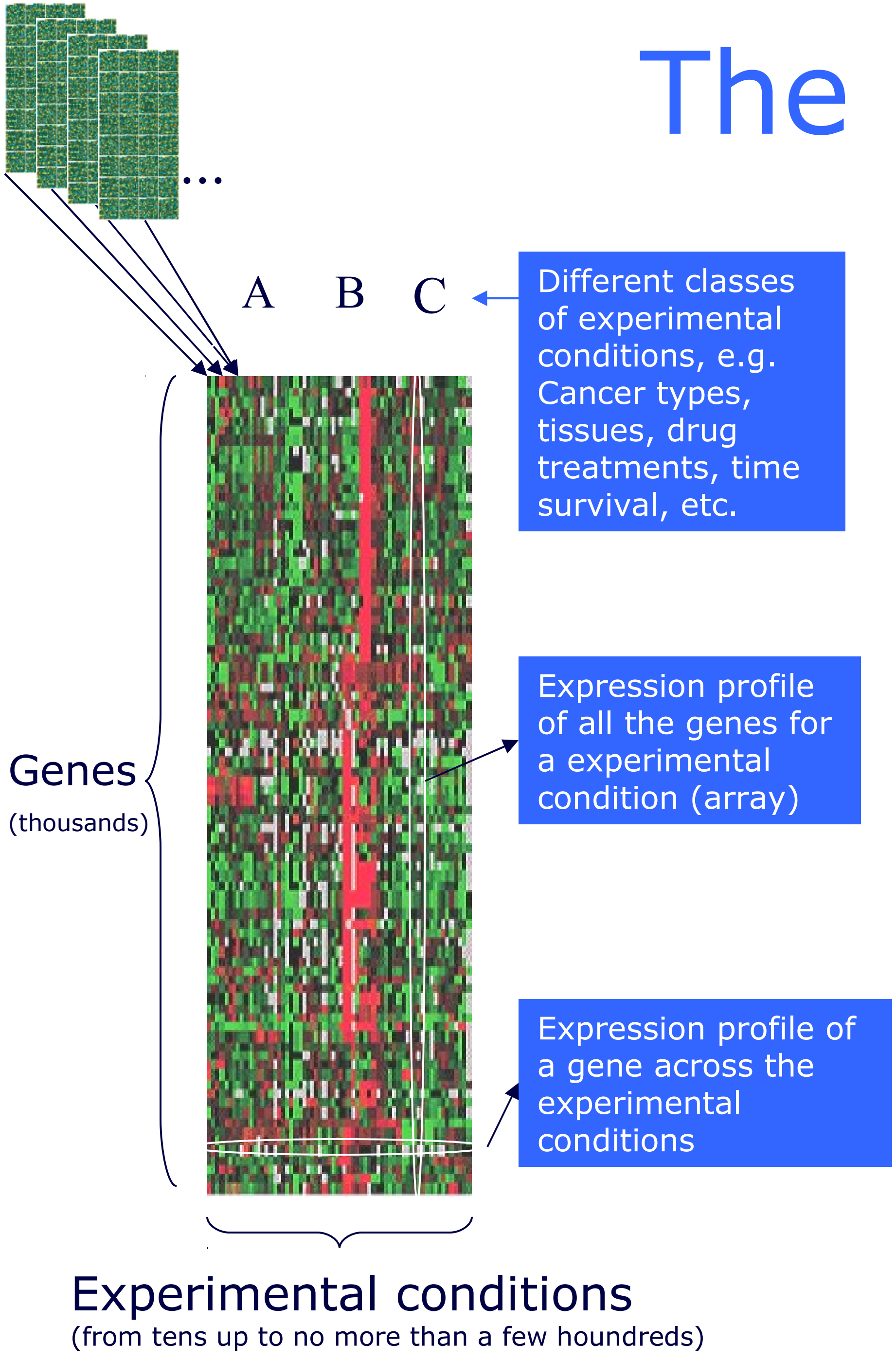
(a) After normalization by average (b) after print-tip lowess normalization (c) after normalisation taking into account spatial effects



Secondary analysis

Once the measurements are in a common, comparable scale the results can be studied.
Diferent studies can be made that include class discovery, classification, gene selection, etc.

The data



Different classes of experimental conditions, e.g. Cancer types, tissues, drug treatments, time survival, etc.

Expression profile of all the genes for a experimental condition (array)

Expression profile of a gene across the experimental conditions

Genes (thousands)

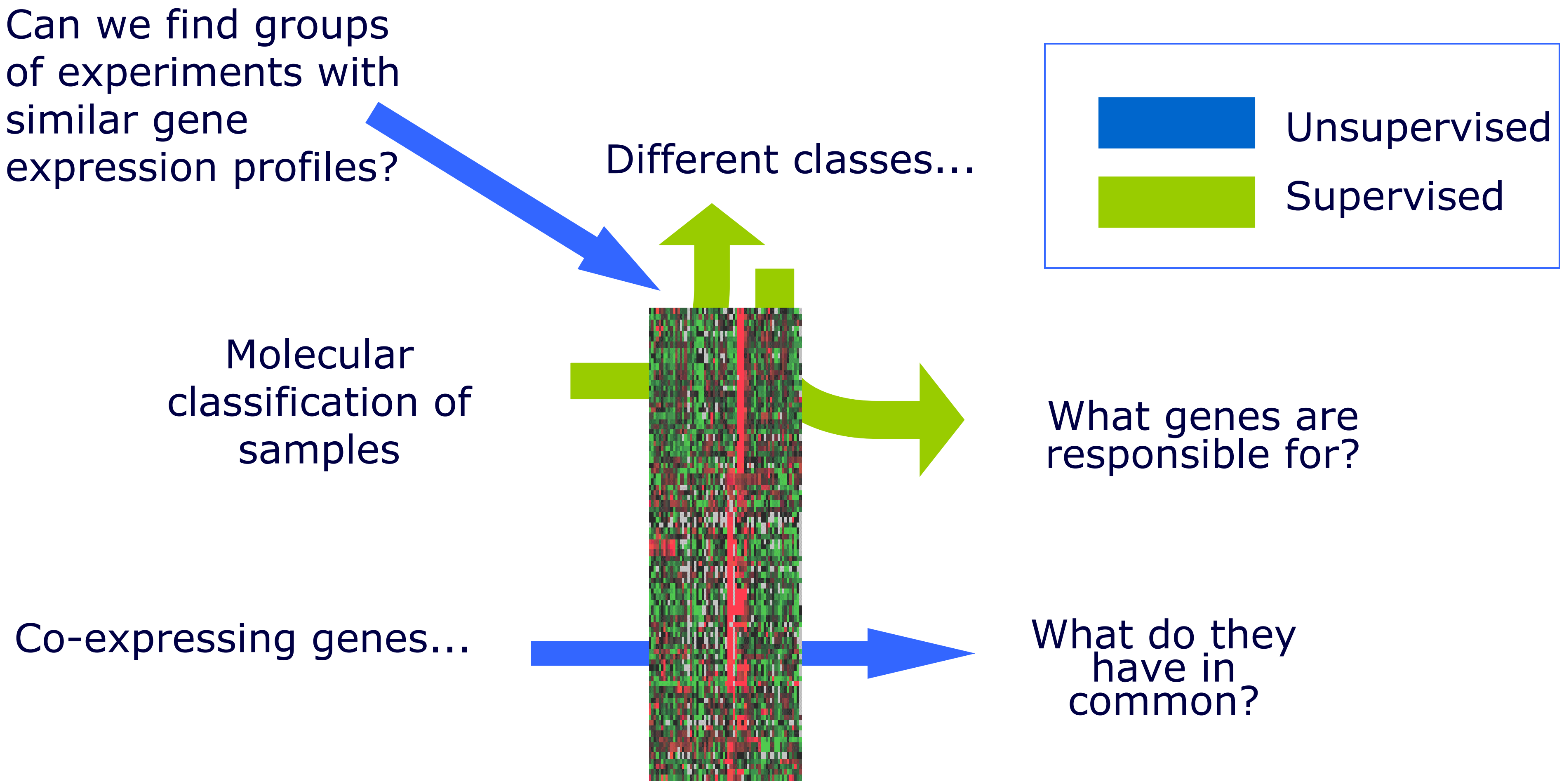
Experimental conditions (from tens up to no more than a few houndreds)

Characteristics of the data:

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

Studies must be hypothesis driven.

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



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

Different classes...

🟦 Unsupervised

🟩 Supervised

Molecular classification of samples

What genes are responsible for?

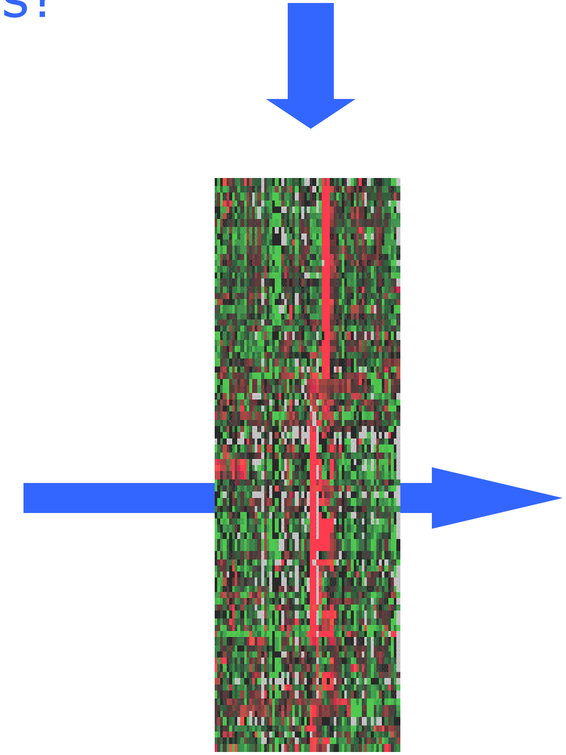
Co-expressing genes...

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?

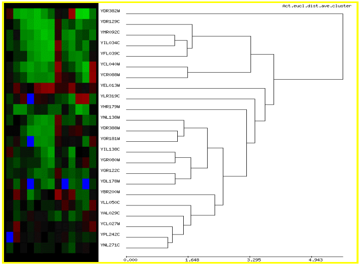


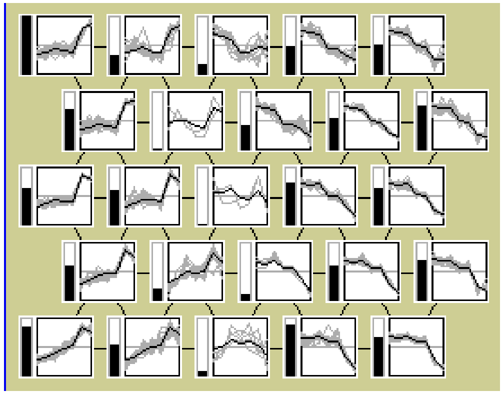
Co-expressing genes...

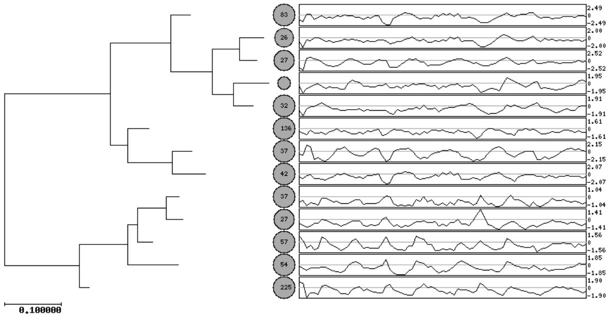
- What genes co-express?
- How many different expression patterns do we have?
- What do they have in common?
- Etc.

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

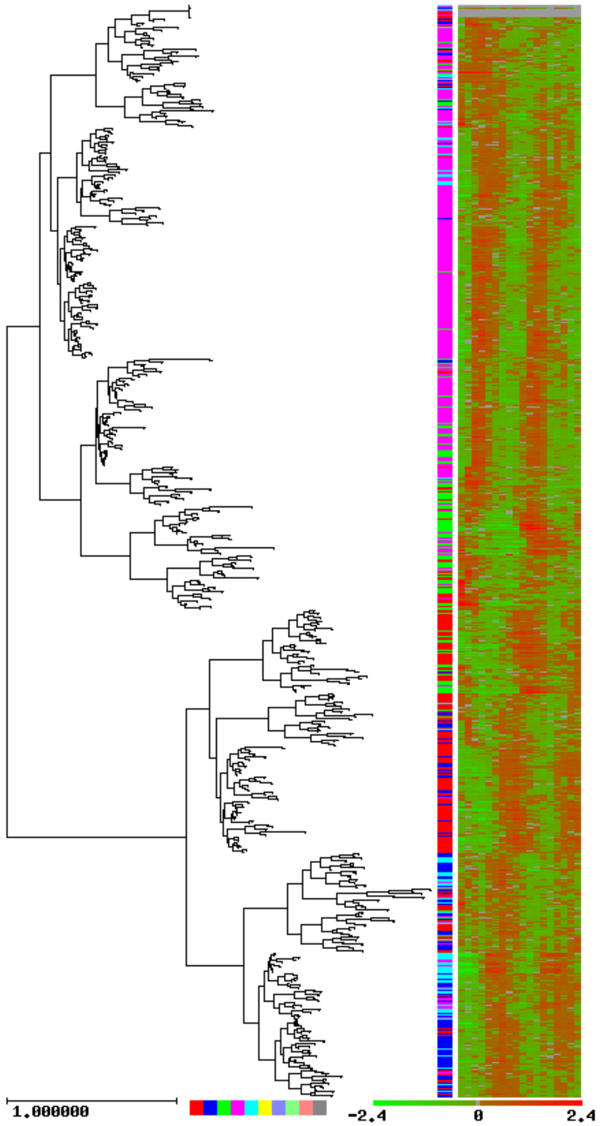
Non hierarchical	hierarchical
K-means, PCA	UPGMA
SOM	SOTA
	Different levels of information







An unsupervised problem: clustering of genes.



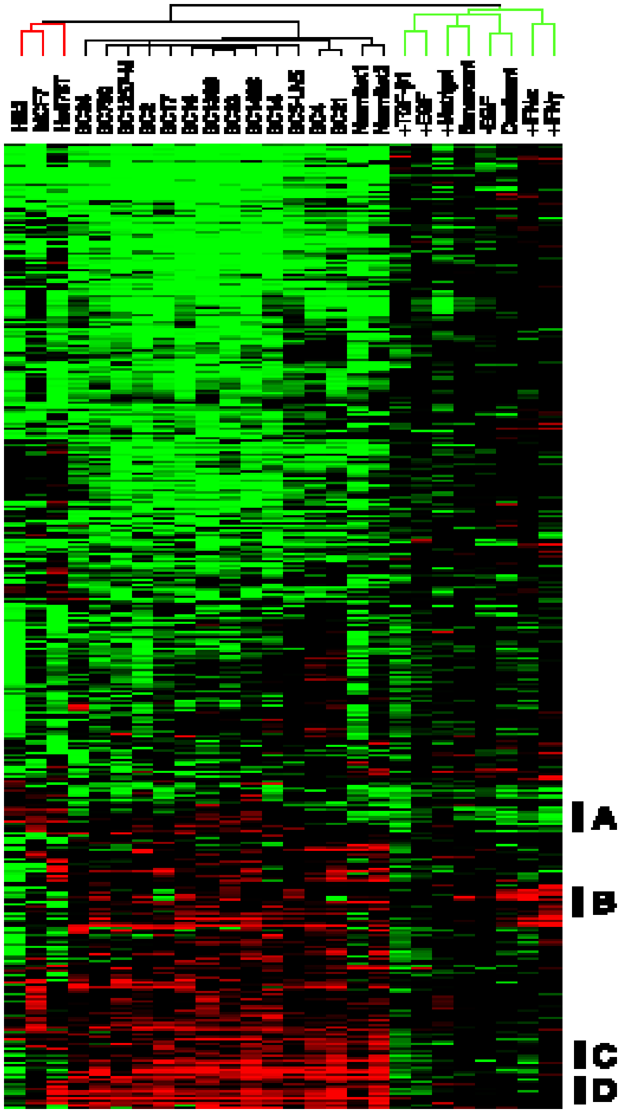
- Gene clusters are previously unknown
- Distance function
- Cluster gene expression patterns based uniquely on their similarities.
- Results are subjected to further interpretation (if possible)

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

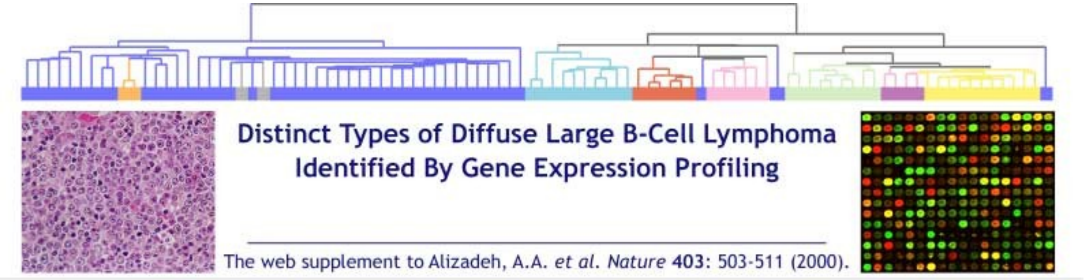
Any gene (regardless its relevance for the classification) has the same weight in the comparison.

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

Noise

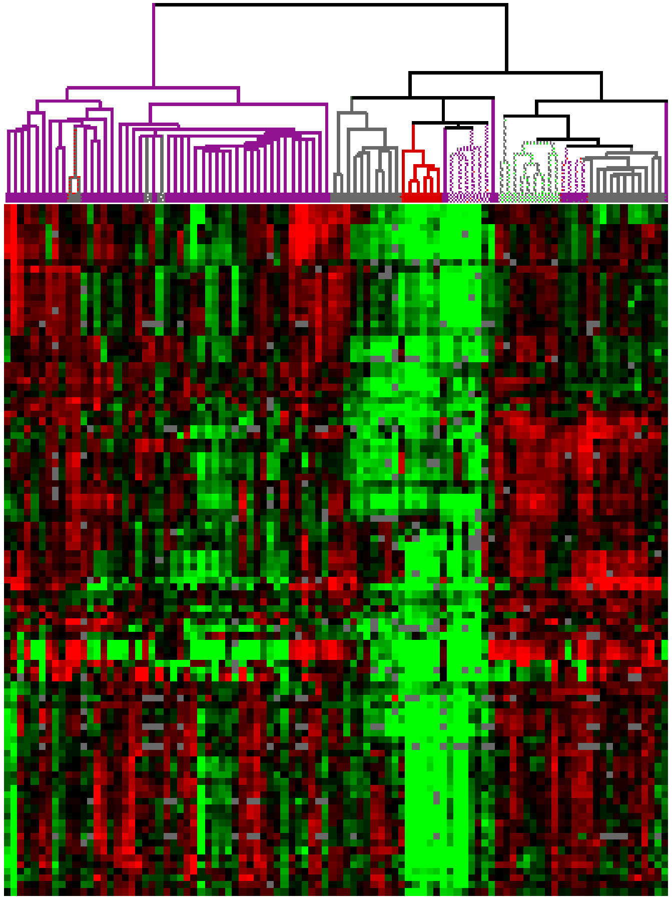
and/or

irrelevant trends



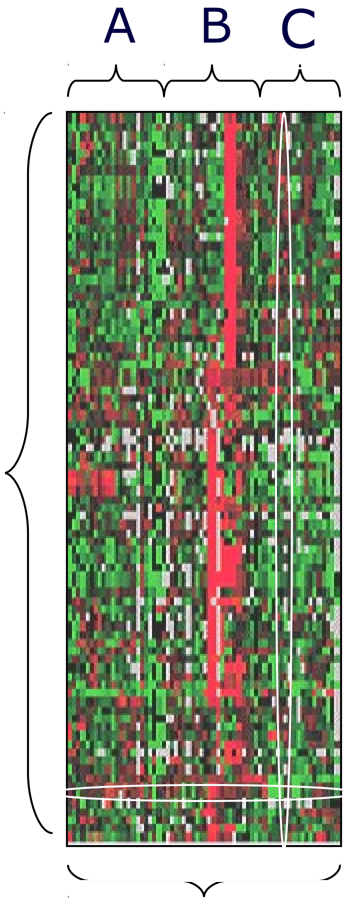
Distinct Types of Diffuse Large B-Cell Lymphoma
Identified By Gene Expression Profiling

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.



A B C

Genes (thousands)

Experimental conditions

(from tens up to no more than a few houndreds)

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

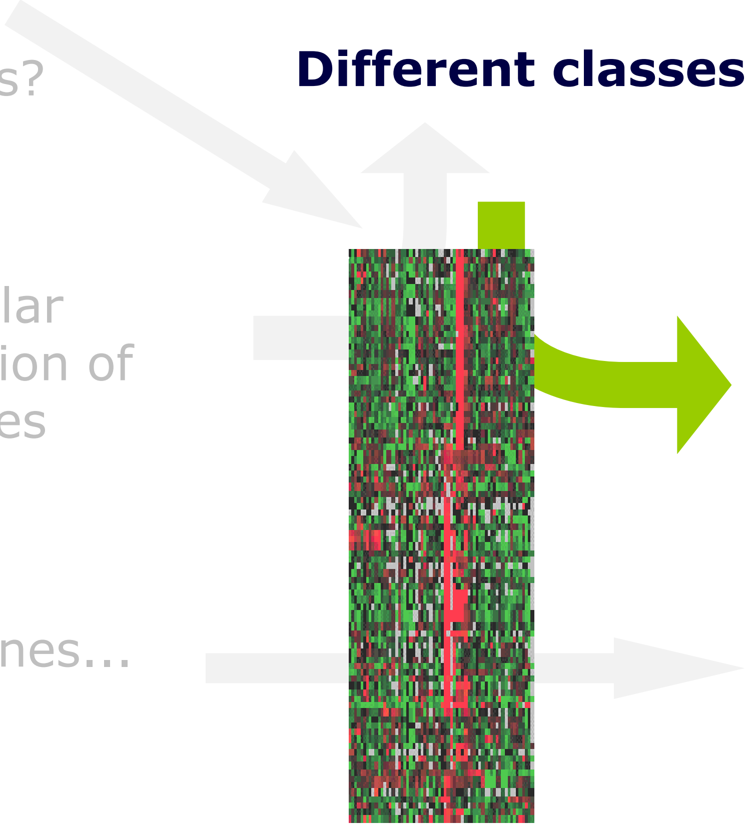
And

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.

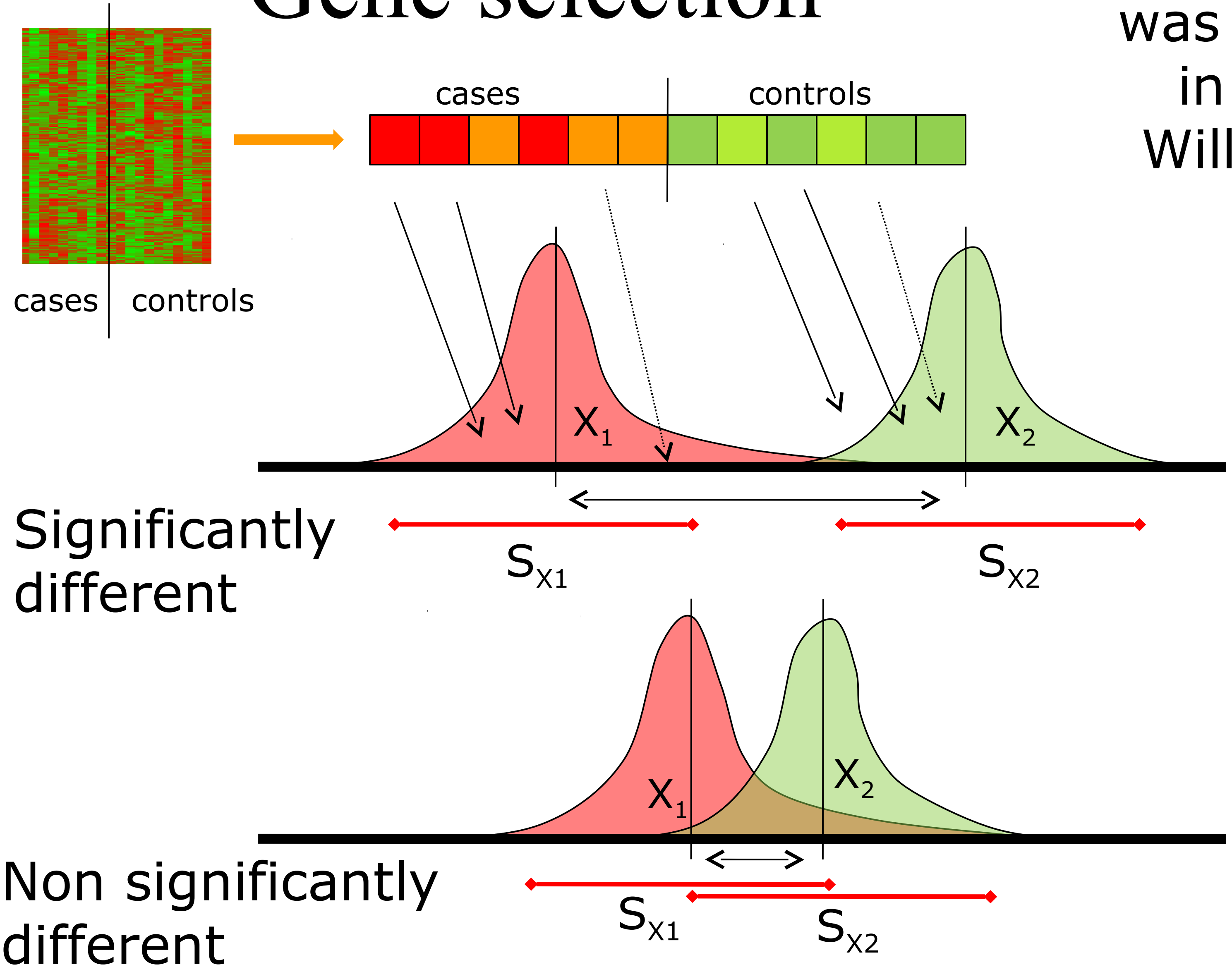
The simplest way: univariant gene-by-gene.
Other multivariant approaches can be used

- **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

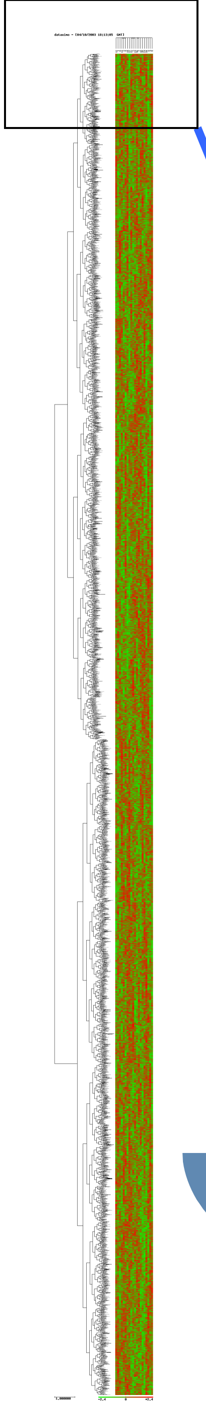
Non significantly different

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}}$$

being

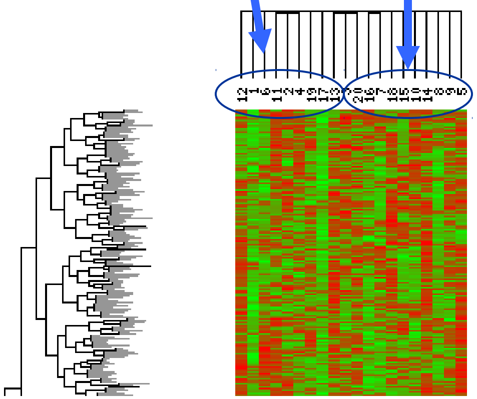
$$S_{X_1X_2} = \sqrt{\frac{S_{X_1}^2 + S_{X_2}^2}{2}}.$$

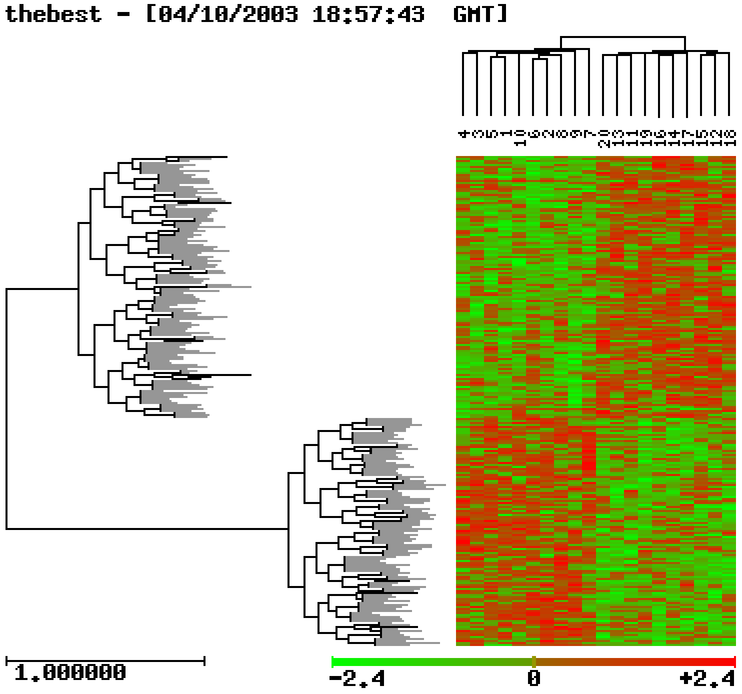
A simple problem: gene selection for class discrimination



~15,000 genes

Case(10)/control(10)

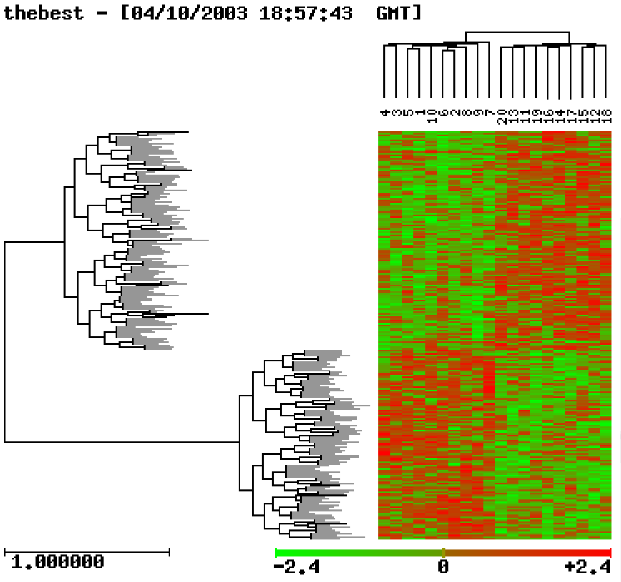




Genes differentially expressed among classes (t-test), with p-value < 0.05

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

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



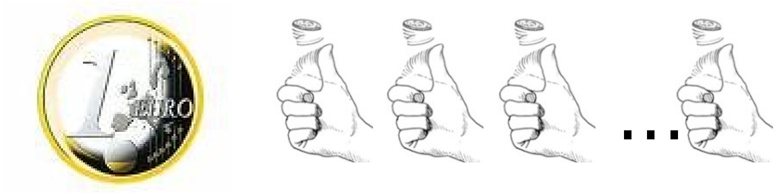
So... Why do we find good p-values?

		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
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.

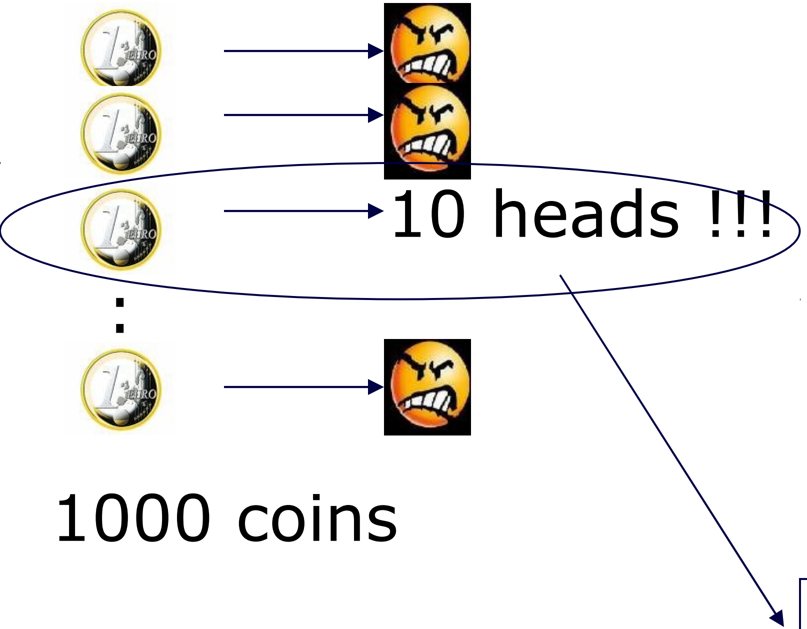
Adjusted p-values must be used!

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

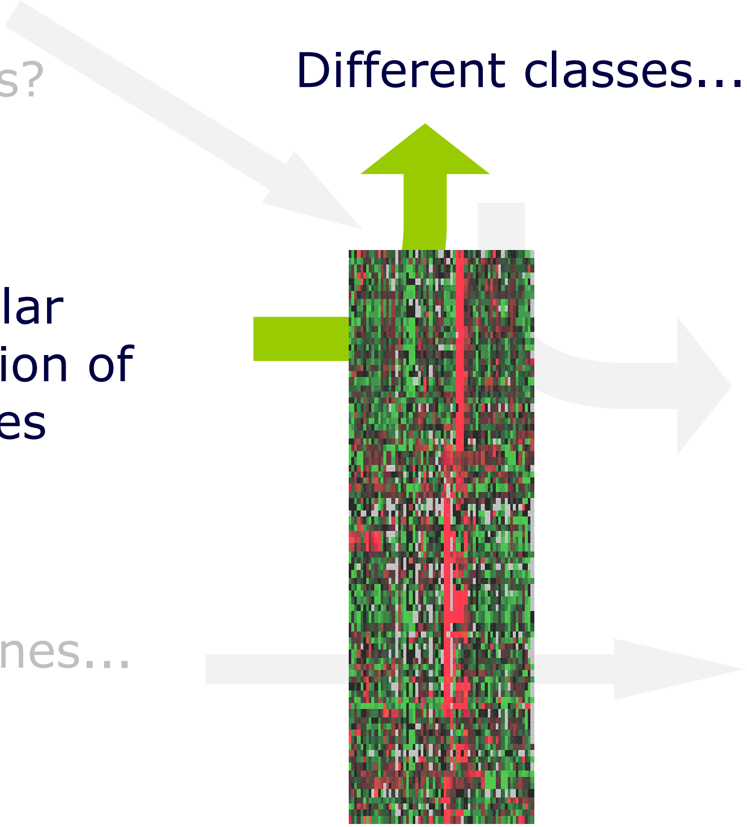
Will you still use this coin for betting?

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?

Context: personalized medicine and what is the future

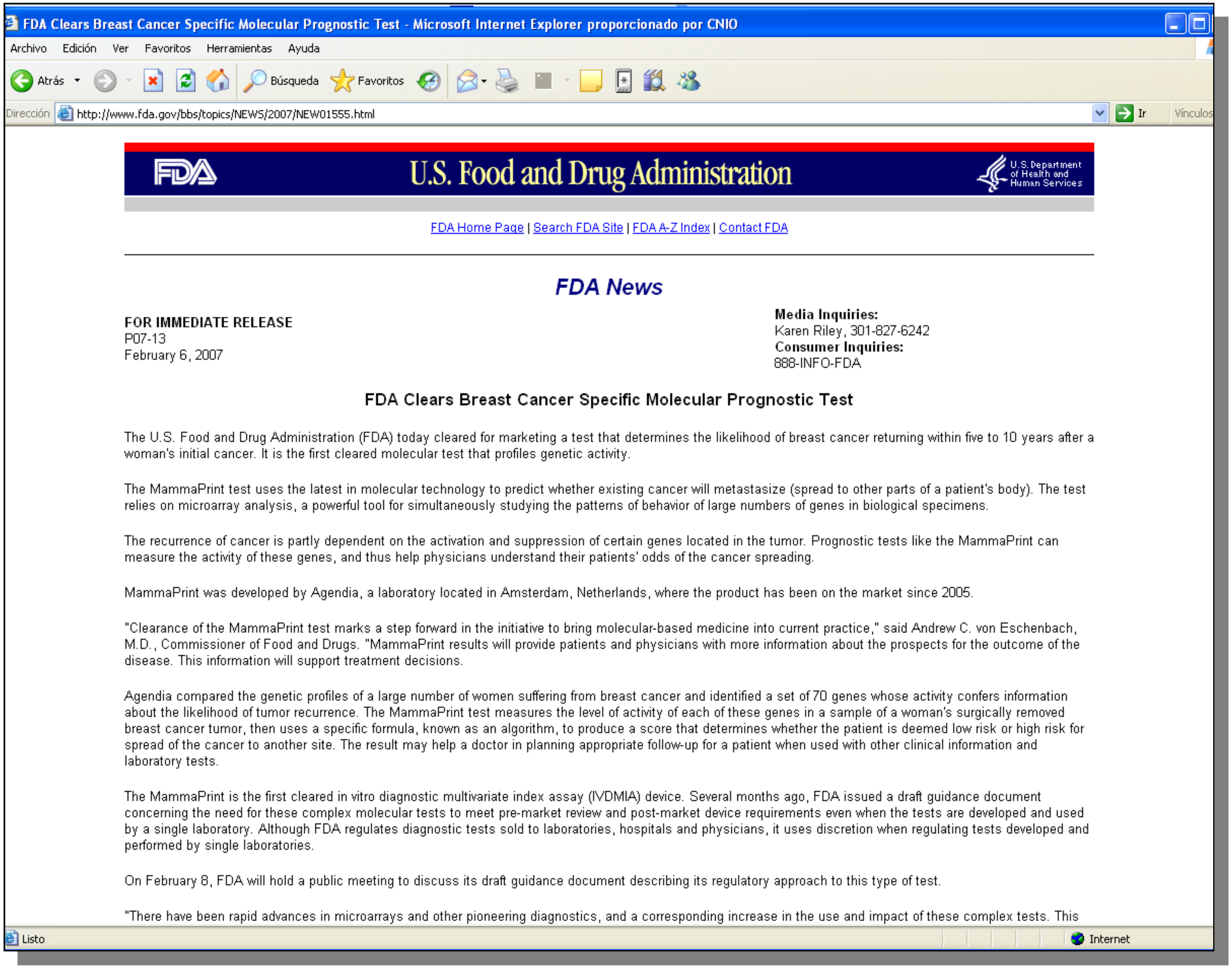
Big challenge for the pharma industry in the 21st century

Driven by academy and regulatory authorities

Relies or pharmacogenomic tests that properly stratifies patients

In the years coming, new tests based on different “omics” methodologies will open new avenues for new personalized drugs and treatments

FDA approves the first predictor based on microarrays


FDA Clears Breast Cancer Specific Molecular Prognostic Test - Microsoft Internet Explorer proporcionado por CNIO
Archivo
Edición
Ver
Favoritos
Herramientas
Ayuda
Atrás
Búsqueda
Favoritos
Dirección
http://www.fda.gov/bbs/topics/NEWS/2007/NEW01555.html
Ir
Vínculos
FDA
U.S. Food and Drug Administration
U.S. Department of Health and Human Services
FDA Home Page | Search FDA Site | FDA A-Z Index | Contact FDA
Listo
Internet

FDA News

FOR IMMEDIATE RELEASE
P07-13
February 6, 2007

Media Inquiries:
Karen Riley, 301-827-6242
Consumer Inquiries:
888-INFO-FDA

FDA Clears Breast Cancer Specific Molecular Prognostic Test

The U.S. Food and Drug Administration (FDA) today cleared for marketing a test that determines the likelihood of breast cancer returning within five to 10 years after a woman's initial cancer. It is the first cleared molecular test that profiles genetic activity.

The MammaPrint test uses the latest in molecular technology to predict whether existing cancer will metastasize (spread to other parts of a patient's body). The test relies on microarray analysis, a powerful tool for simultaneously studying the patterns of behavior of large numbers of genes in biological specimens.

The recurrence of cancer is partly dependent on the activation and suppression of certain genes located in the tumor. Prognostic tests like the MammaPrint can measure the activity of these genes, and thus help physicians understand their patients' odds of the cancer spreading.

MammaPrint was developed by Agendia, a laboratory located in Amsterdam, Netherlands, where the product has been on the market since 2005.

"Clearance of the MammaPrint test marks a step forward in the initiative to bring molecular-based medicine into current practice," said Andrew C. von Eschenbach, M.D., Commissioner of Food and Drugs. "MammaPrint results will provide patients and physicians with more information about the prospects for the outcome of the disease. This information will support treatment decisions.

Agendia compared the genetic profiles of a large number of women suffering from breast cancer and identified a set of 70 genes whose activity confers information about the likelihood of tumor recurrence. The MammaPrint test measures the level of activity of each of these genes in a sample of a woman's surgically removed breast cancer tumor, then uses a specific formula, known as an algorithm, to produce a score that determines whether the patient is deemed low risk or high risk for spread of the cancer to another site. The result may help a doctor in planning appropriate follow-up for a patient when used with other clinical information and laboratory tests.

The MammaPrint is the first cleared in vitro diagnostic multivariate index assay (IVDMIA) device. Several months ago, FDA issued a draft guidance document concerning the need for these complex molecular tests to meet pre-market review and post-market device requirements even when the tests are developed and used by a single laboratory. Although FDA regulates diagnostic tests sold to laboratories, hospitals and physicians, it uses discretion when regulating tests developed and performed by single laboratories.

On February 8, FDA will hold a public meeting to discuss its draft guidance document describing its regulatory approach to this type of test.

"There have been rapid advances in microarrays and other pioneering diagnostics, and a corresponding increase in the use and impact of these complex tests. This

The MicroArray Quality Control (MAQC) Project: An FDA-Led Effort Toward Personalized Medicine

MAQC Website: http://edkb.fda.gov/MAQC/

MAQC-II Objective:

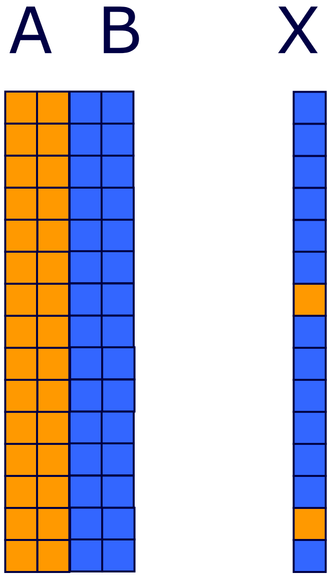
Reaching consensus on the "best practices" (Data Analysis Protocol, DAP) in developing and validating microarray-based predictive models (classifiers) for clinical and preclinical applications.

A international consortium of 36 data analysis teams submitted prediction results from 18,202 models for 6 datasets to the MAQC-II

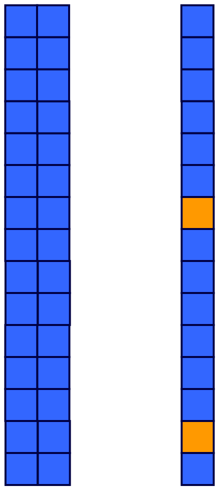
Of predictors and molecular signatures

What is a predictor?

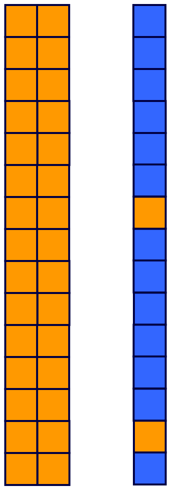
Intuitive notion:



Is X, A or B?



Diff (B, X) = 2



Diff (A, X) = 13

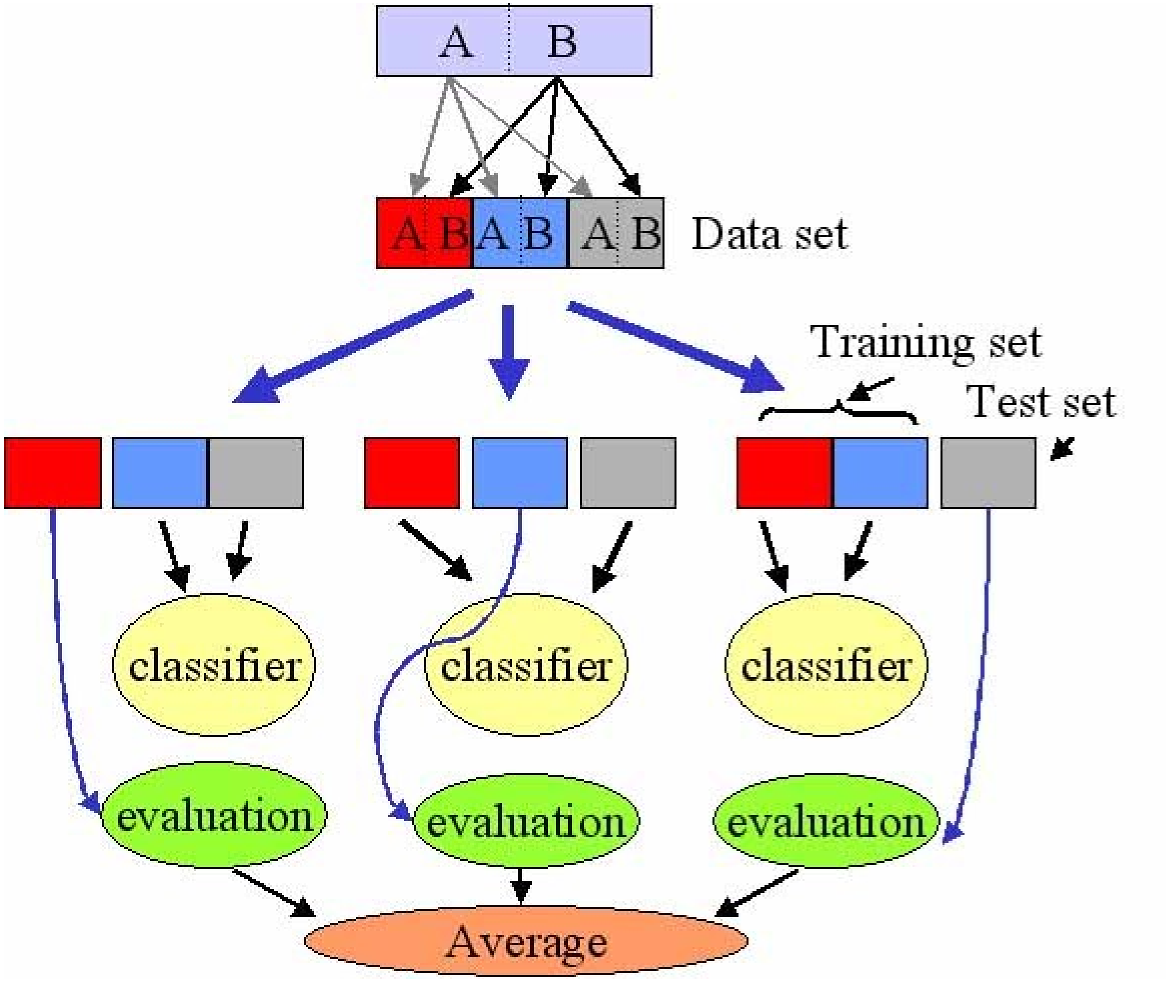
Most probably X belongs to class B

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

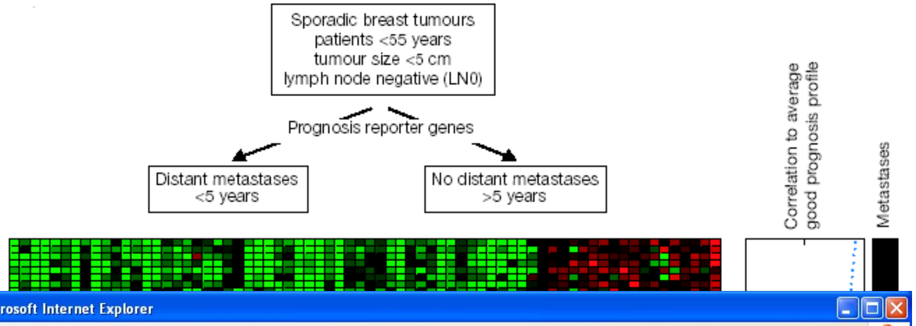
Cross-validation

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

agendia » Home - Microsoft Internet Explorer
Archivo Edición Ver Favoritos Herramientas Ayuda
Atrás Búsqueda Favoritos Multimedia
Dirección http://www.agendia.com/ Ir
Links Ensembl Genome Browser NCBI HomePage Google Scholar Bioinformatics - Manuscript Central [TM] MailSite Express Norton Internet Security
Agendia
The Future is Now
The future is now
diagnostics by genomics
The future is now
Enter
. Unveiling Tumor behavior . Descifrar
Internet

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

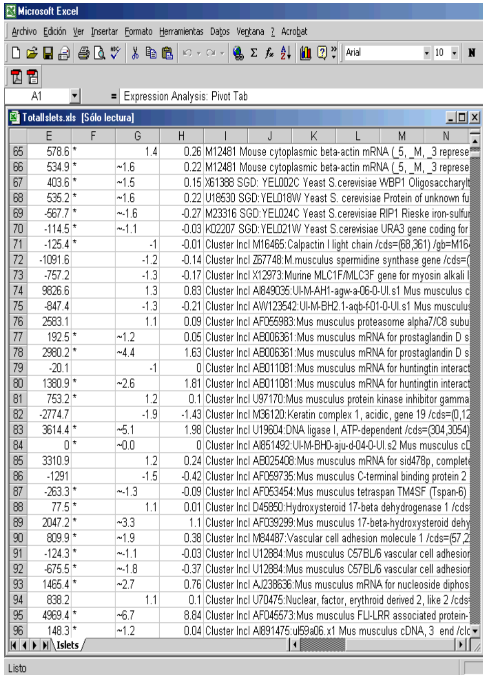
Functional profiling of genome-scale experiments in the post-genomic era

My data...

How are structured?

What are these groups?

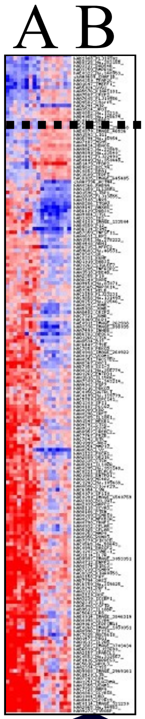
What is this gen?

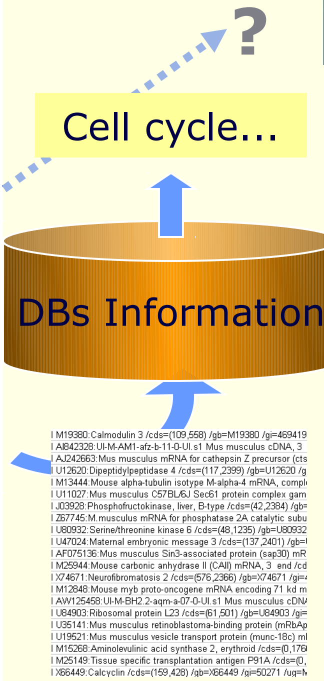


Microsoft Excel
Archivo Edición Ver Insertar Formato Herramientas Datos Ventana ? Acrobat
Arial
10
N
A1
Totallslets.xls [Sólo lectura]
Islets
Listo

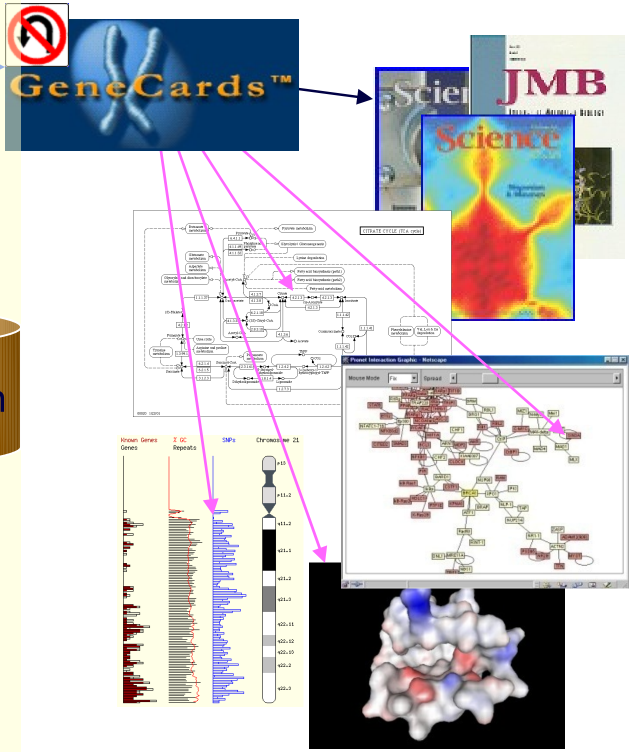
= Expression Analysis: Pivot Tab

	E	F	G	H	I	J	K	L	M	N
65	578.6	*	1.4	0.26	M12481 Mouse cytoplasmic beta-actin mRNA (_5, _M, _3 represe					
66	534.9	*	~1.6	0.22	M12481 Mouse cytoplasmic beta-actin mRNA (_5, _M, _3 represe					
67	403.6	*	~1.5	0.15	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 adhesion					
92	-675.5	*	~-1.8	-0.37	Cluster Incl U12884:Mus musculus C57BL/6 vascular cell adhesion					
93	1465.4	*	~2.7	0.76	Cluster Incl 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 /clc					



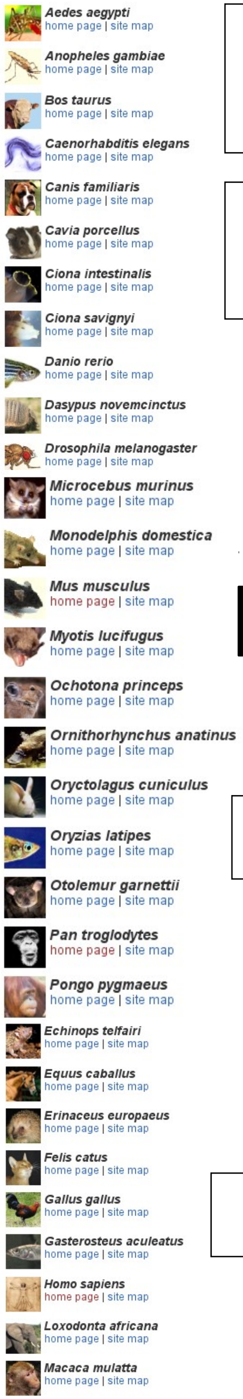


I M19380:Calmodulin 3 /cds=(109,558) /gb=M19380 /gi=469419
I AI842328:UI-M-AM1-afz-b-11-0-UI.s1 Mus musculus cDNA, 3
I AJ242663:Mus musculus mRNA for cathepsin Z precursor (cts
I U12620:Dipeptidylpeptidase 4 /cds=(117,2399) /gb=U12620 /g
I M13444:Mouse alpha-tubulin isotype M-alpha-4 mRNA, compl
I U11027:Mus 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



GeneCards™
Scien
JMB
Science
CITRATE CYCLE (TCA cycle)
Known Genes
SNPs
Chromosome 21
Pronet Interaction Graphic - Netscape

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

Gene Ontology CONSORTIUM

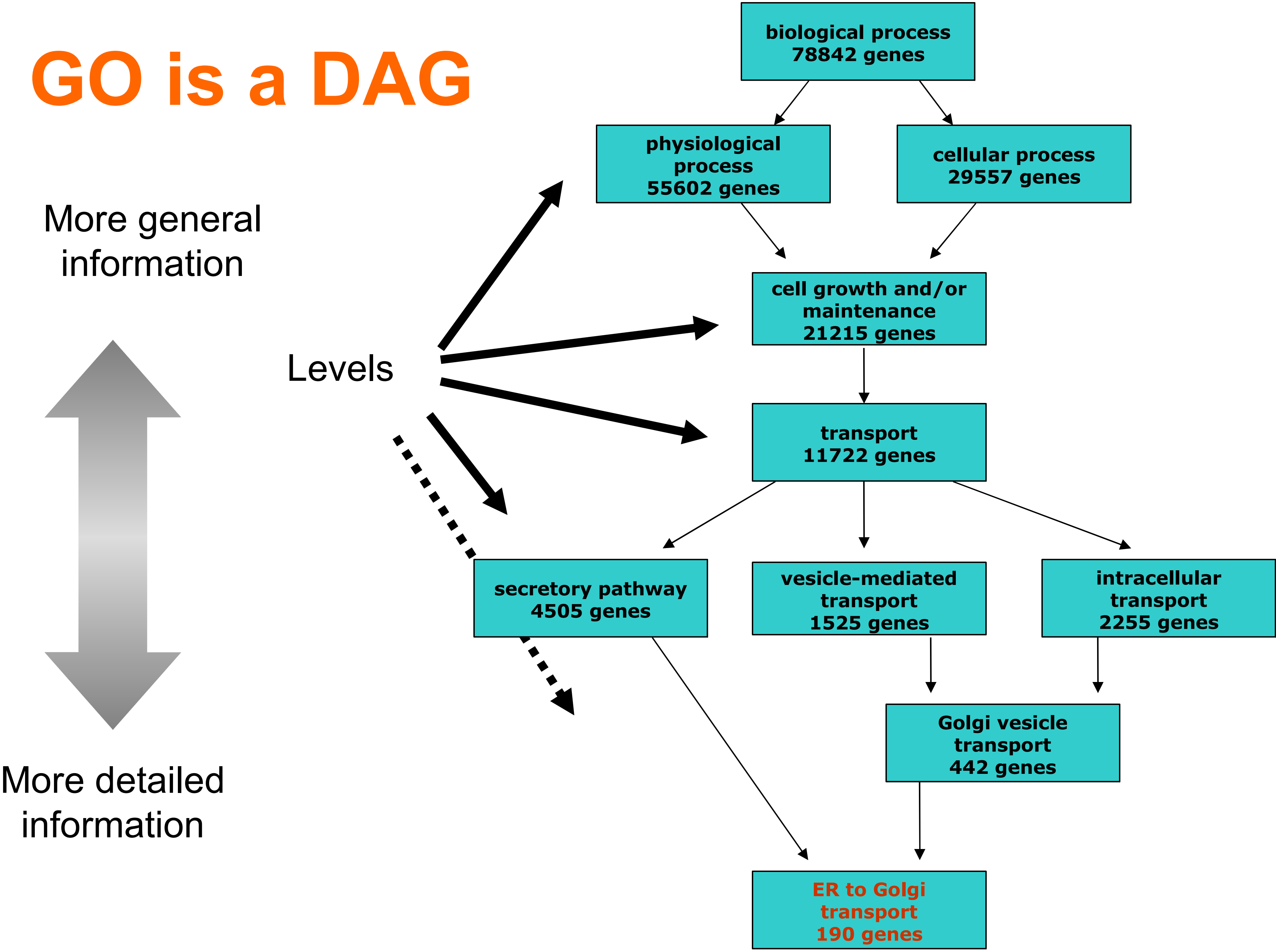
http://www.geneontology.org

- The objective of GO is to provide controlled vocabularies for the description of the molecular function, biological process and cellular component of gene products.

- These terms are to be used as attributes of gene products by collaborating databases, facilitating uniform queries across them.

- The controlled vocabularies of terms are structured

GO is a DAG



biological process
78842 genes

physiological process
55602 genes

cellular process
29557 genes

More general information

cell growth and/or maintenance
21215 genes

Levels

transport
11722 genes

secretory pathway
4505 genes

vesicle-mediated transport
1525 genes

intracellular transport
2255 genes

Golgi vesicle transport
442 genes

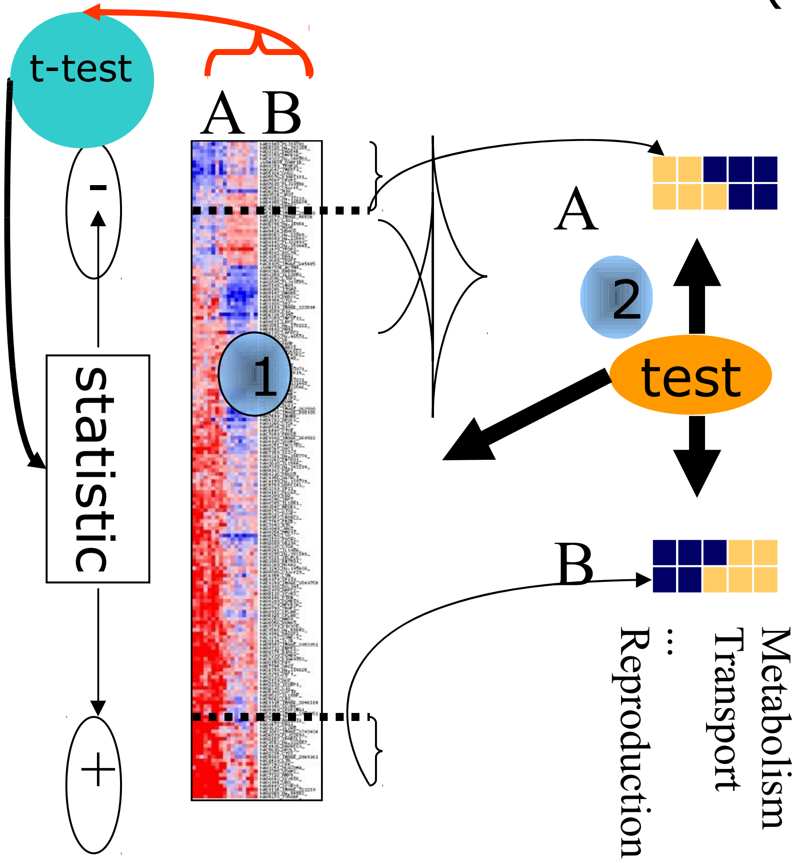
More detailed information

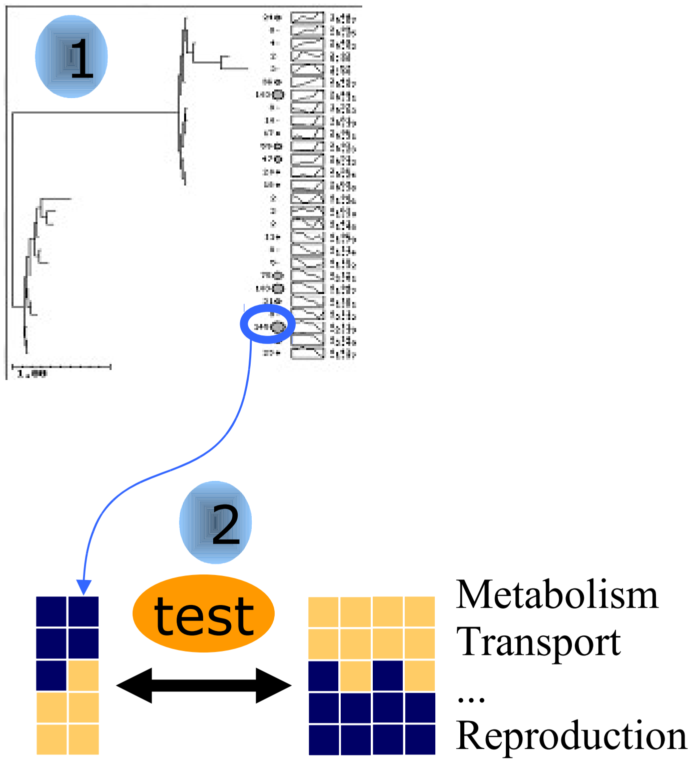
ER to Golgi transport
190 genes

Two-steps functional interpretation

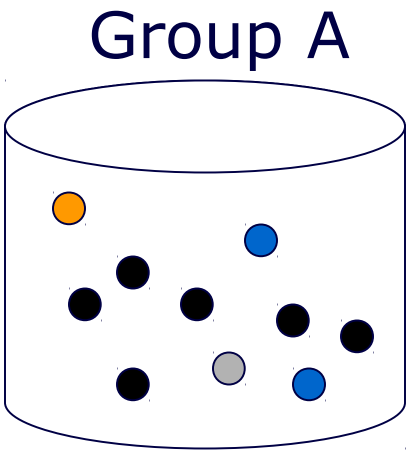
1 Genes are selected based on their experimental values and...

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

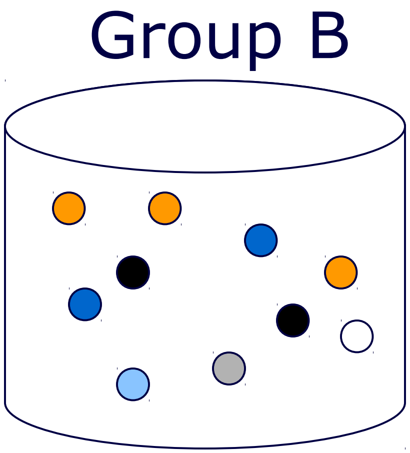




Testing two GO terms
(remember, we have to test thousands)



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



Biosynthesis	Other	
6	4	A
2	8	B

The popular Fisher's test

Biosynthesis 60% ● ⟷ Biosynthesis 20% ●

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

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

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

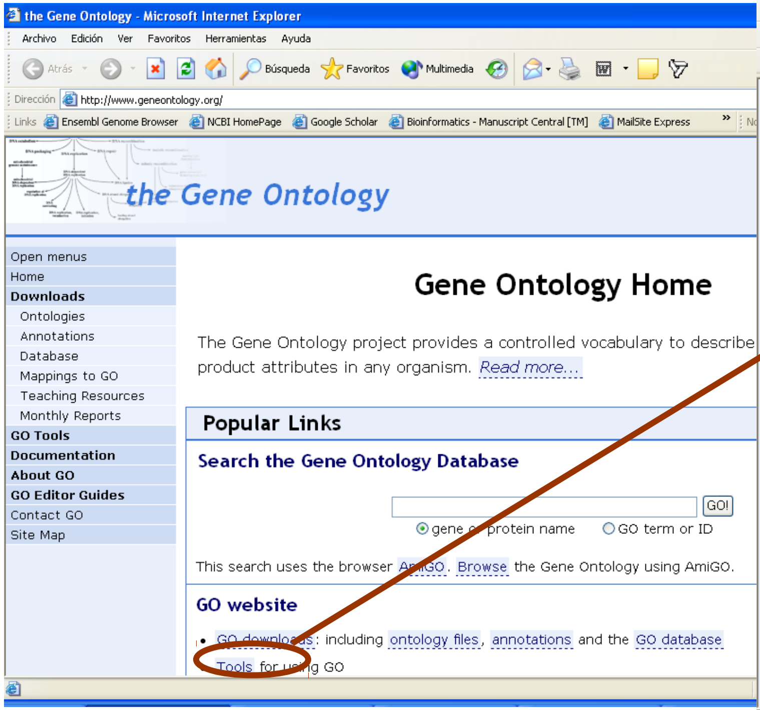
Each row corresponds to a random selection of 50 genes from the *E. coli* genome, compared with respect to the rest of the genome.

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...

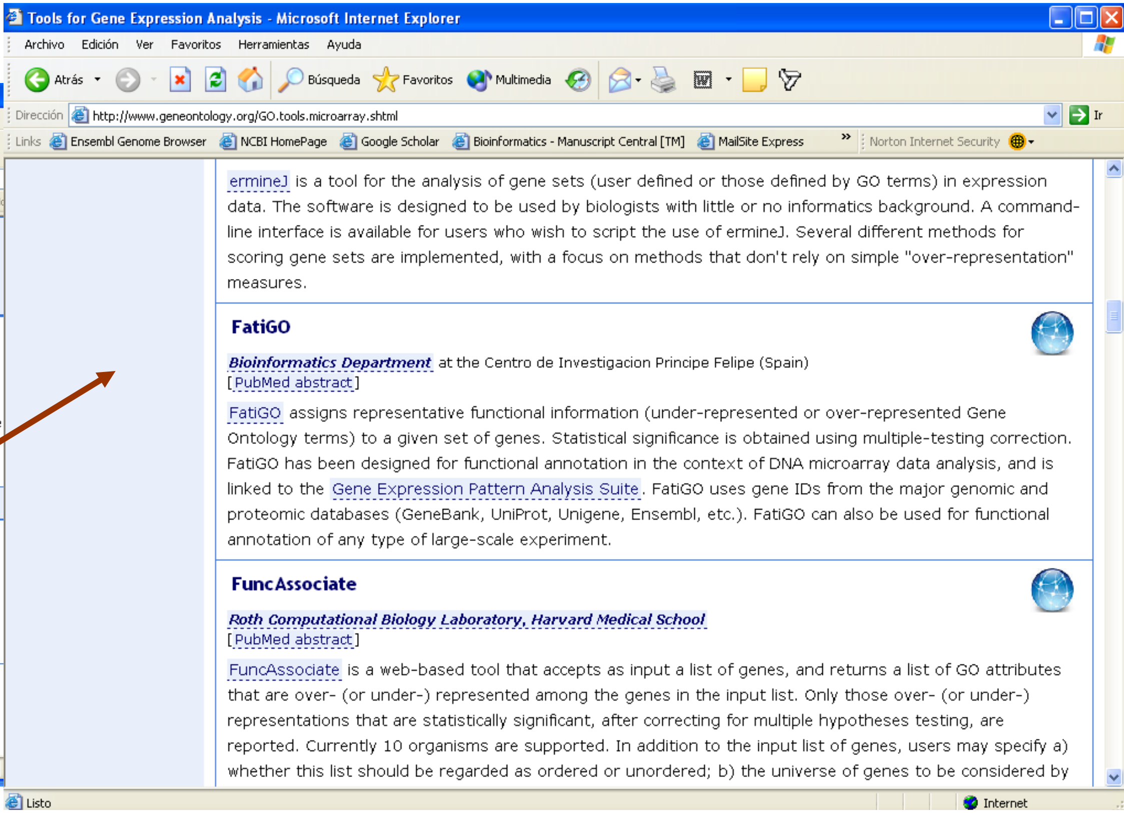
Popular Links

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.)***

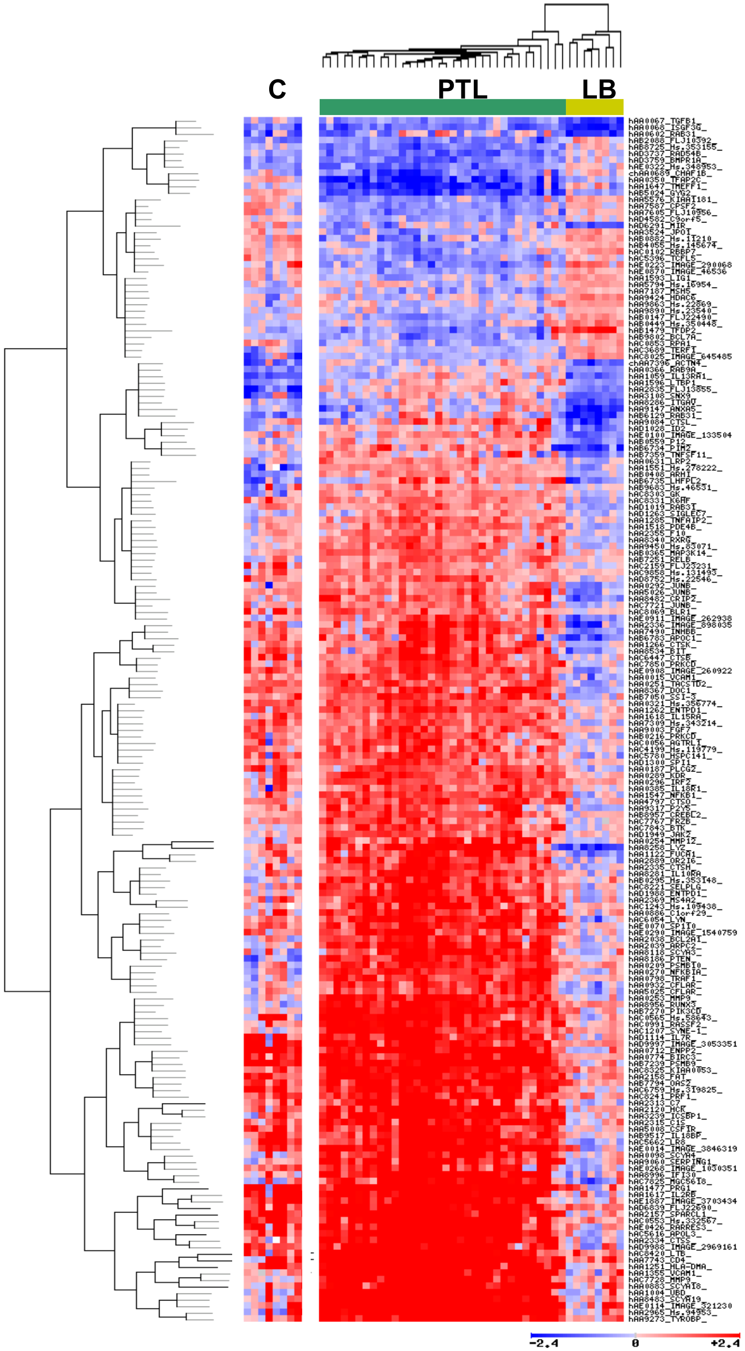
Al-Shahrour et al., 2005 Bioinformatics. Al-Shahrour et al., 2005 NAR

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

Al-Shahrour et al., 2007 NAR

Compilation of tools for functional interpretation of sets of genes

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



Understanding why genes differ in their expression between two different conditions

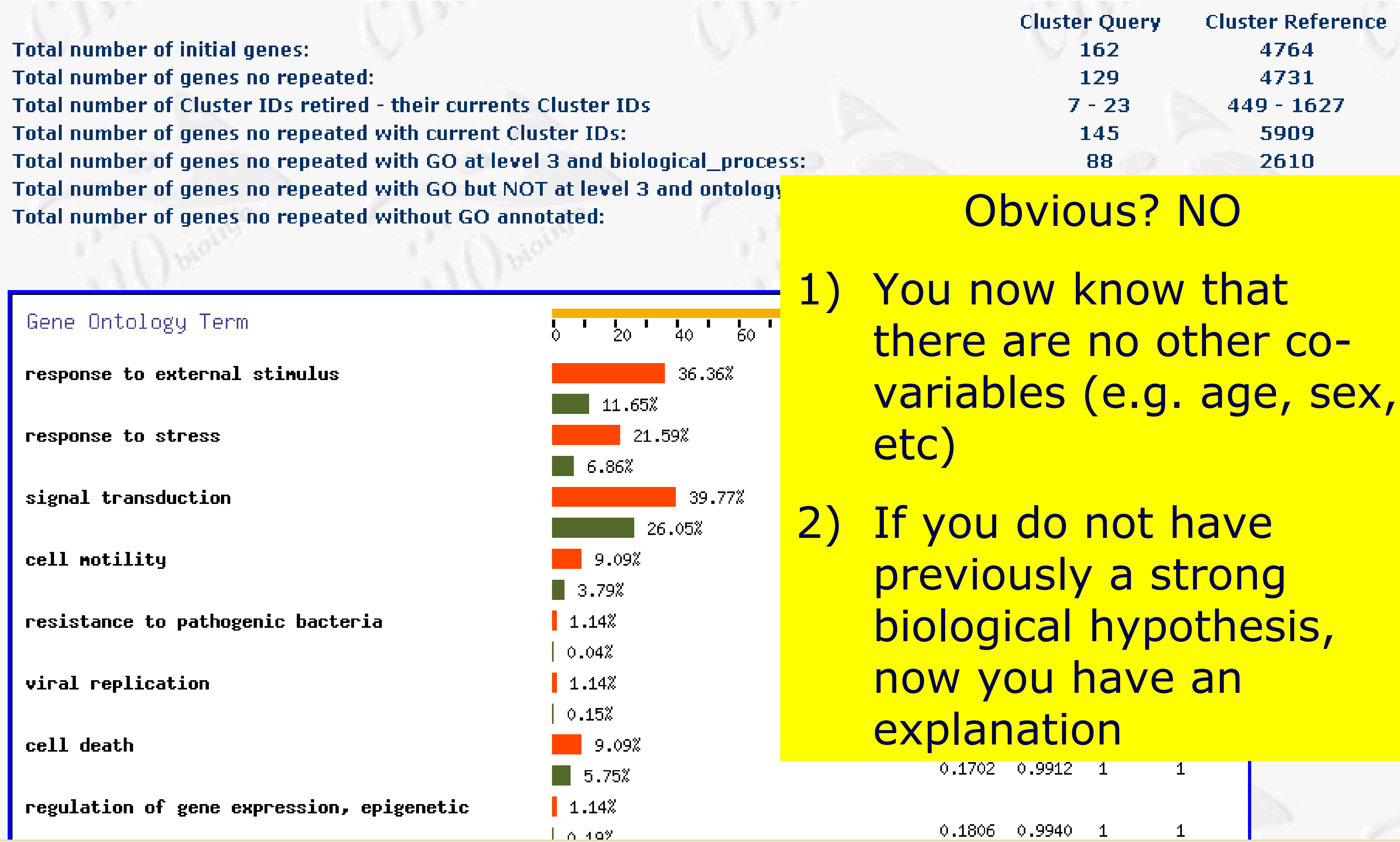
Limphomas from mature lymphocytes (LB) and precursor T-lymphocyte (PTL).

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:		

Gene Ontology Term					
response to external stimulus	36.36% 11.65%				
response to stress	21.59% 6.86%				
signal transduction	39.77% 26.05%				
cell motility	9.09% 3.79%				
resistance to pathogenic bacteria	1.14% 0.04%				
viral replication	1.14% 0.15%				
cell death	9.09% 5.75%	0.1702	0.9912	1	1
regulation of gene expression, epigenetic	1.14% 0.19%	0.1806	0.9940	1	1

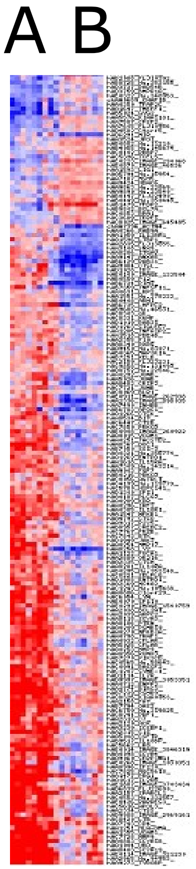
Obvious? NO

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

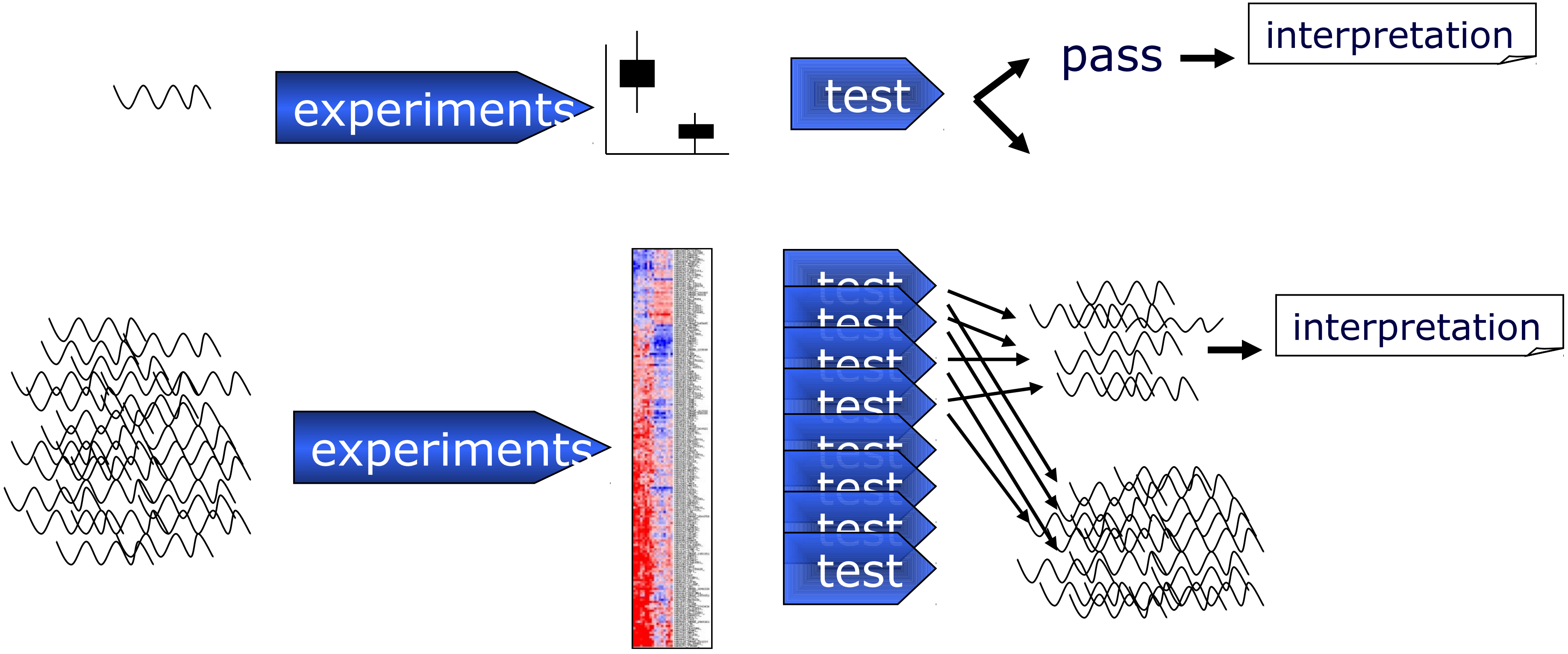
17 with normal tolerance to glucose (**NTG**)

(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)

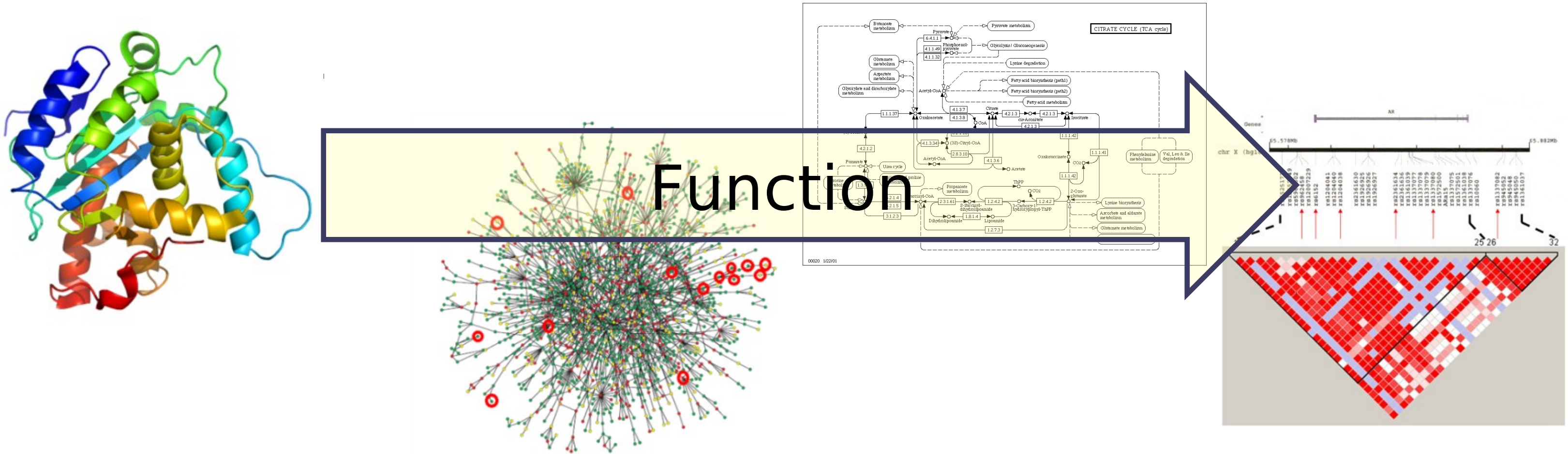
The testing strategy we are conducting is implicitly answering a question different to the one we want to ask.

The true proxies of function

Are we asking the proper questions?

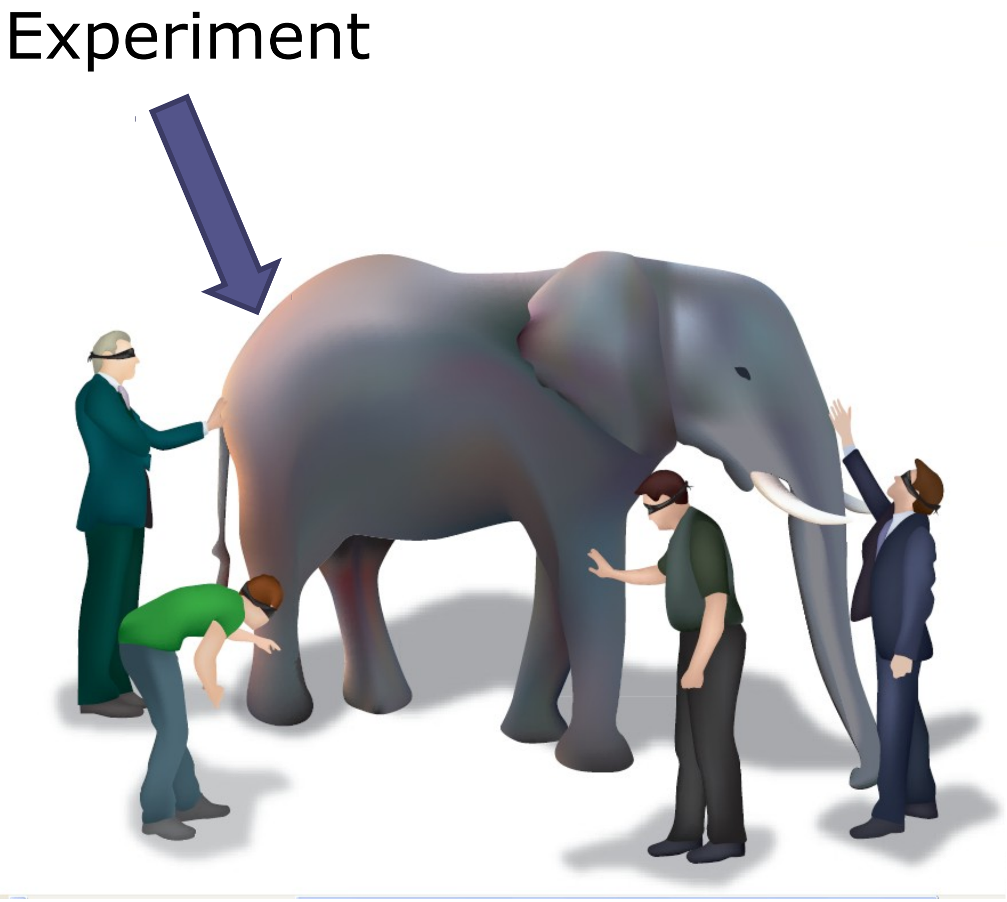
Why do we think in terms of genes?

What are the real bricks that account for the cellular behaviour and for the phenotype or the response to stimulus represented in our experiment? The genes or other higher level units?



Function

What is the entity that accounts for functionality at the cell level?



The wise but blindfolded men could not agree on a description of the elephant’s phenotype

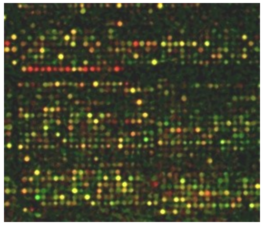
Blindfolded men (**dots in the array**) are the reporters of the individual parts (**genes**), but the reaction (**function altered**) is carried out by the elephant (**functional module, e.g. pathway**)

Therefore, why not to observe the elephant?

Functional genomics.

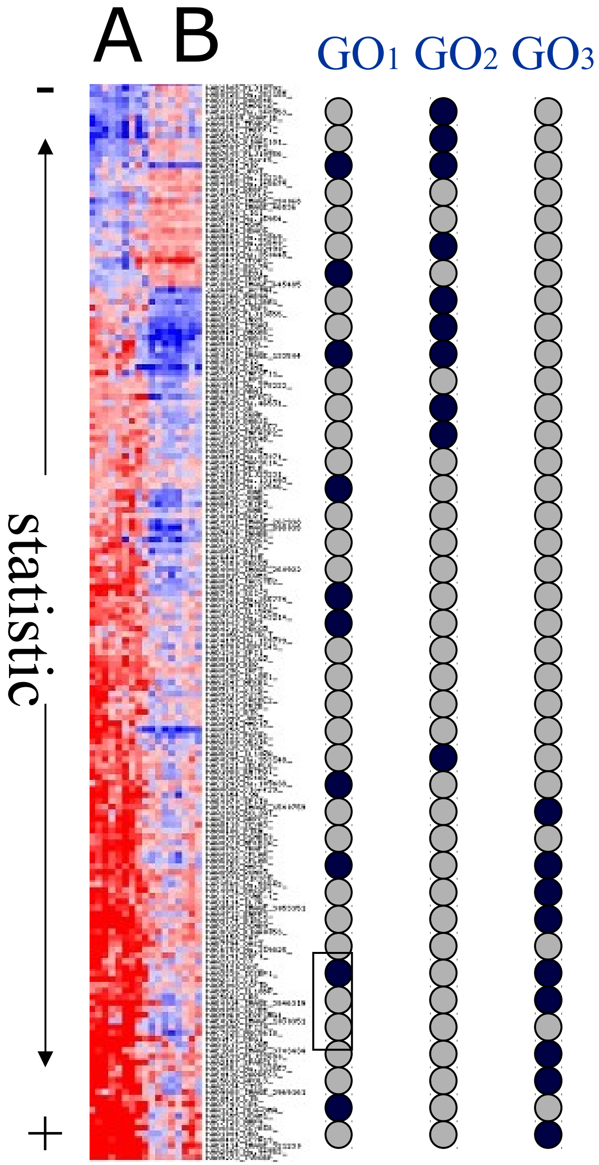
Historic perspective and future

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



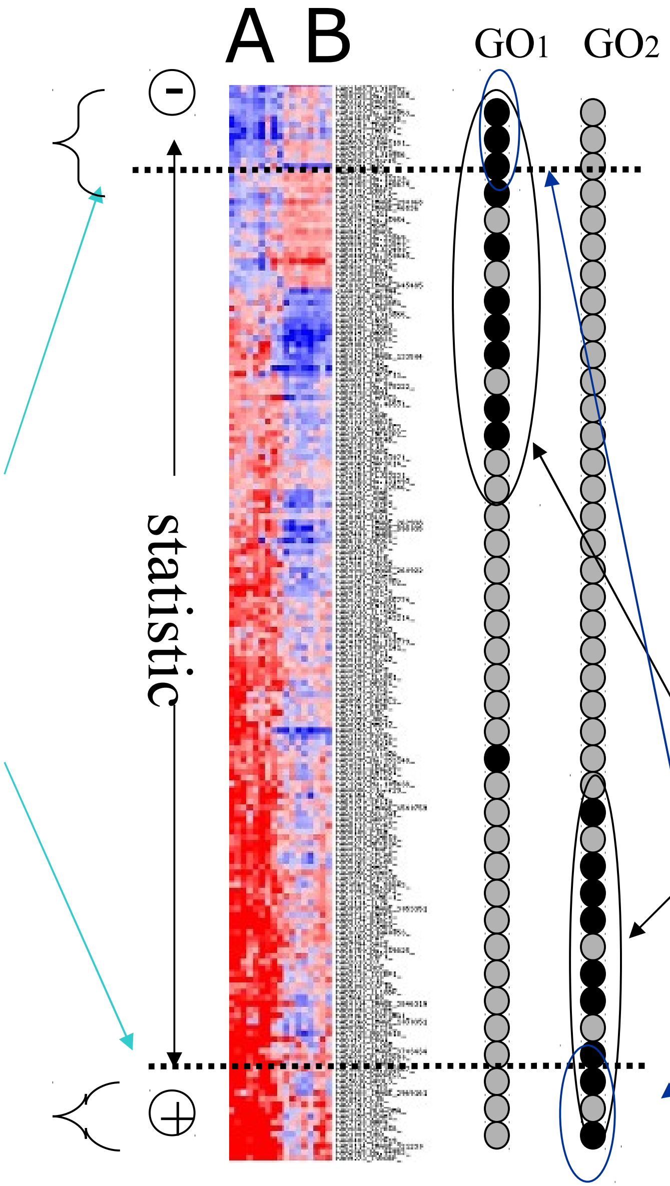
Ranking: A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

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

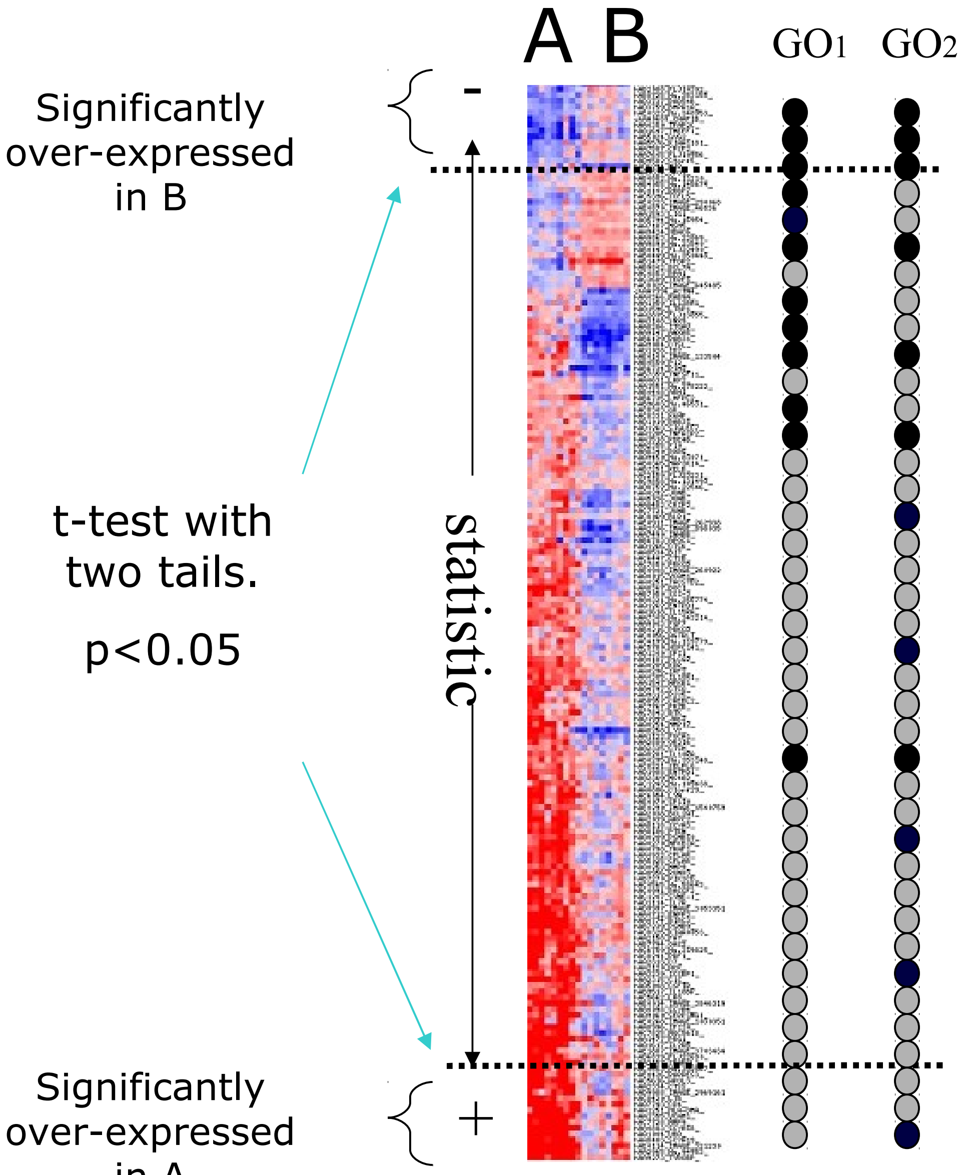
Significantly over-expressed in A

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

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

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

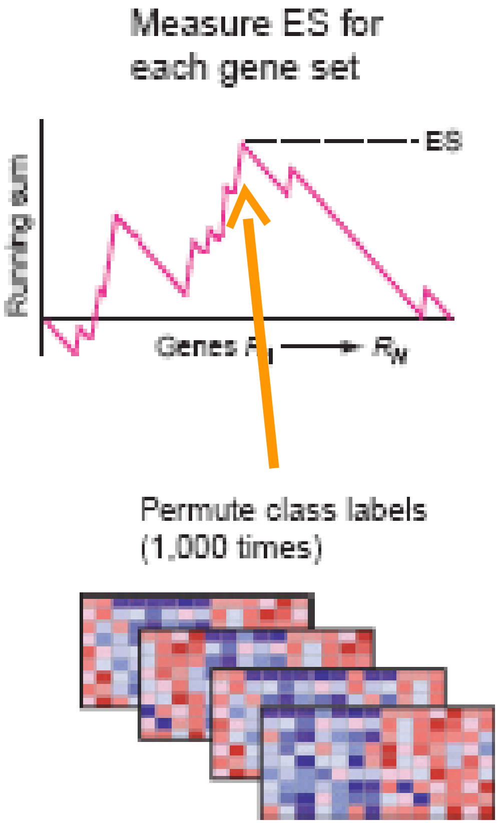
We put both sides of the partition into two bags and destroy the structure of the data.

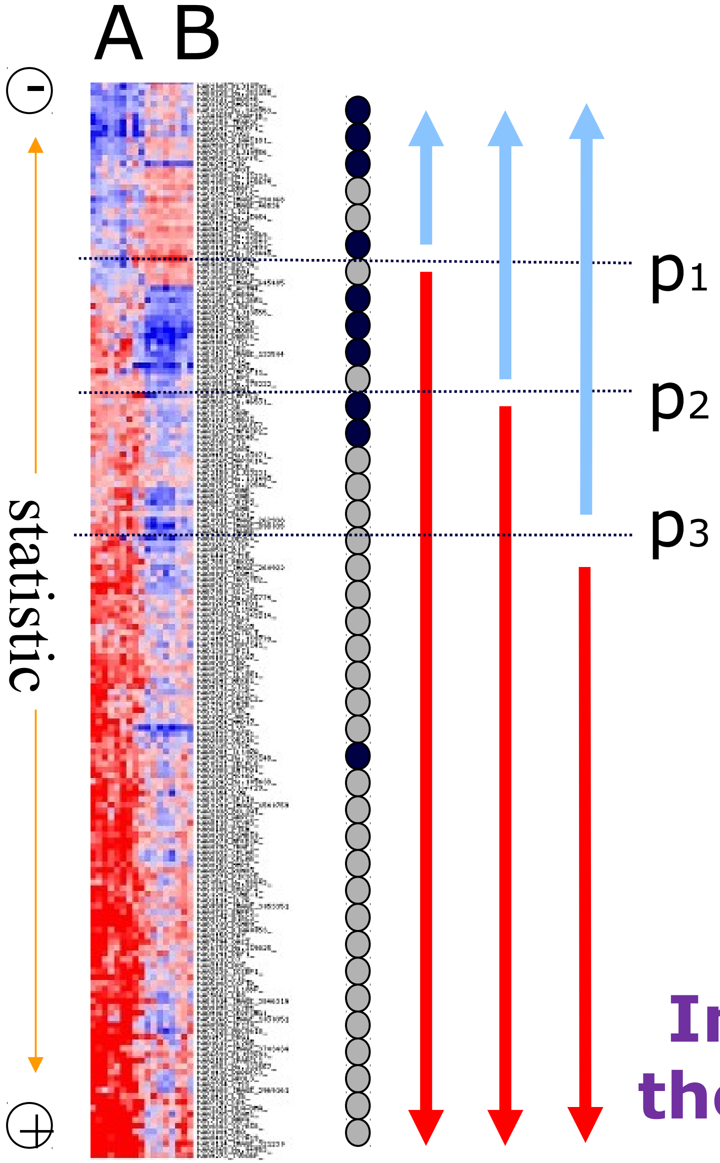
	up	down
GO	3	9
no GO	0	25

Same contingency table for GO_1 and GO_2 !!

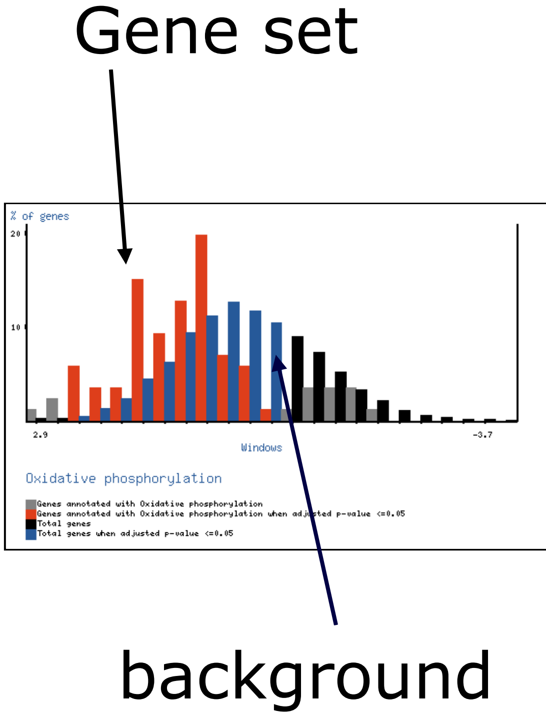
Gene-set enrichment methods

GSEA


Measure ES for each gene set
ES
Running sum
Genes R_1 → R_N
Permute class labels (1,000 times)



FatiScan

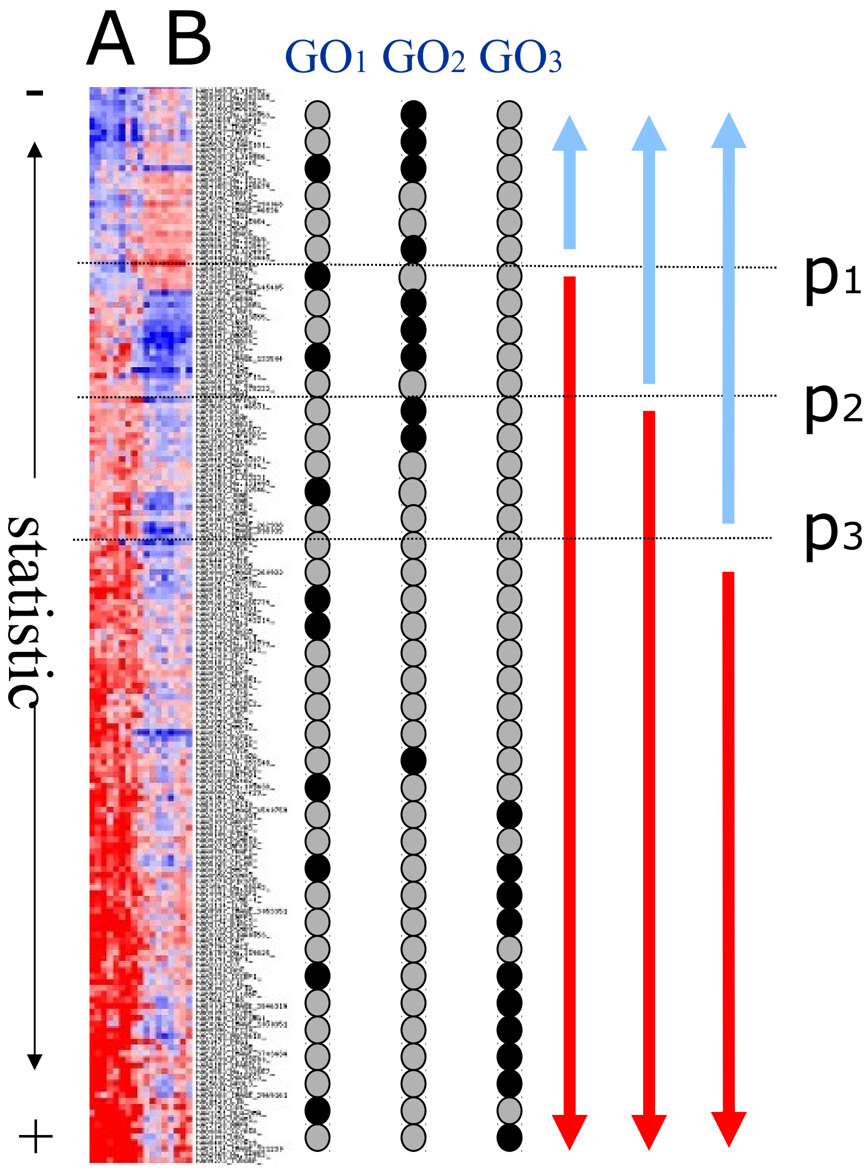


Gene set

background

Independent of the experimental design

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



E.g., term GO$_2$, partition p_1

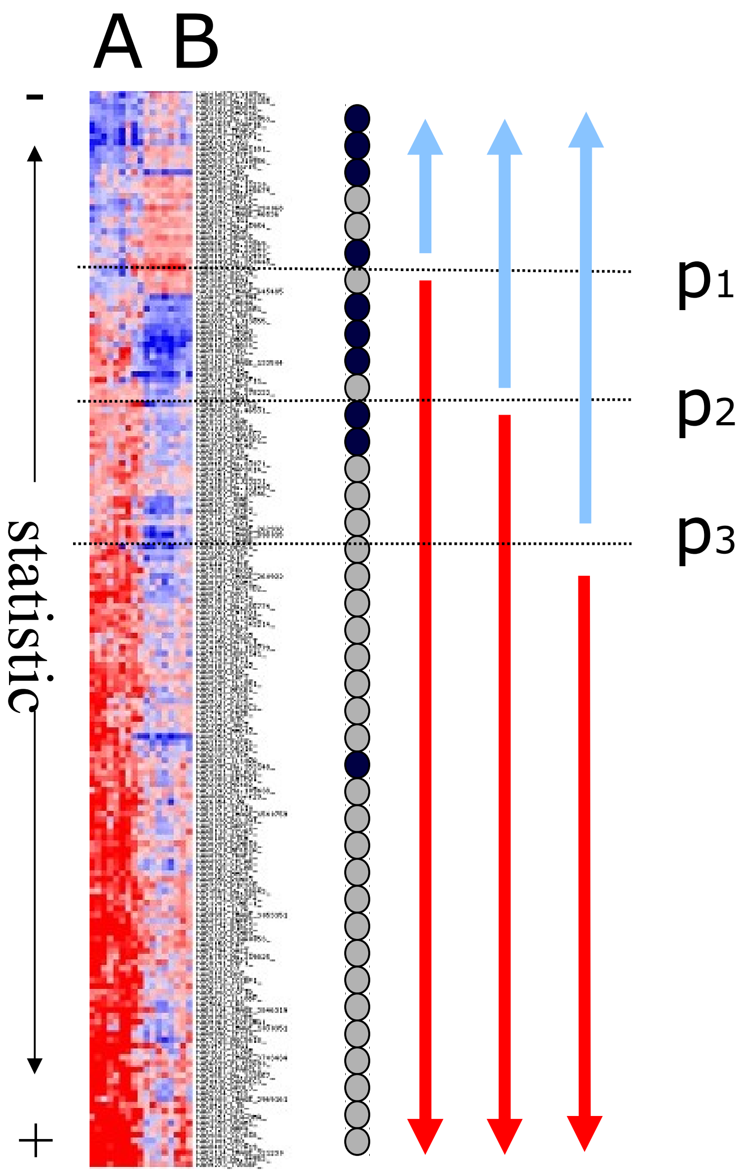
	up	down
GO	4	6
no GO	2	30

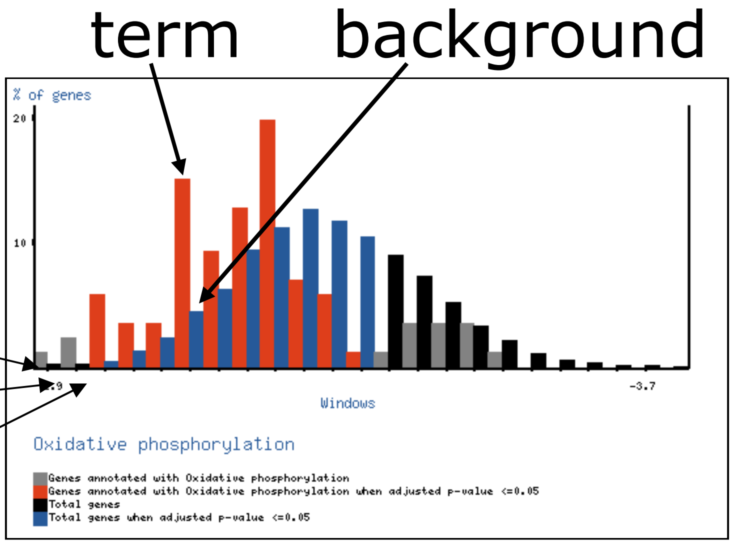
GOs can be directly tested by a segmentation test. A series of partitions of the list are performed (**p_1, p_2, p_3…**) and the GO terms for each functional class in the upper part are compared to the corresponding ones in the lower part by a Fisher test. Asymmetrical distributions of terms towards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

Finally, p-values are adjusted by FDR

Al-Shahrour et al., 2005 Bioinformatics

Obtaining significant results





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

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

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

Al-Shahrour et al., 2005 Bioinformatics

Case study: functional differences in a class comparison experiment

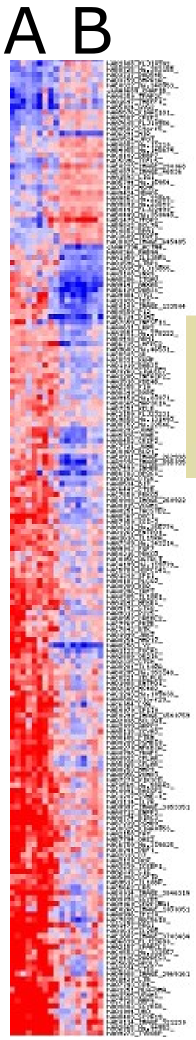
A

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

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

		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)

Gene selection

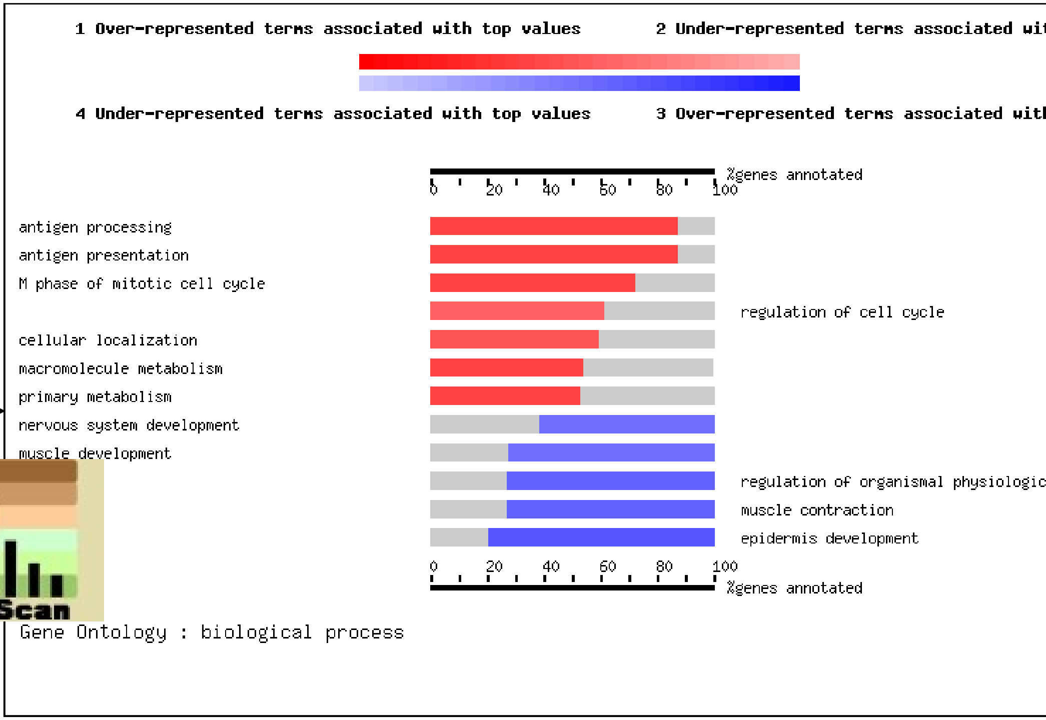
Cox Proportional-Hazards model to study how the expression of each gene across patients is related to their survival

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

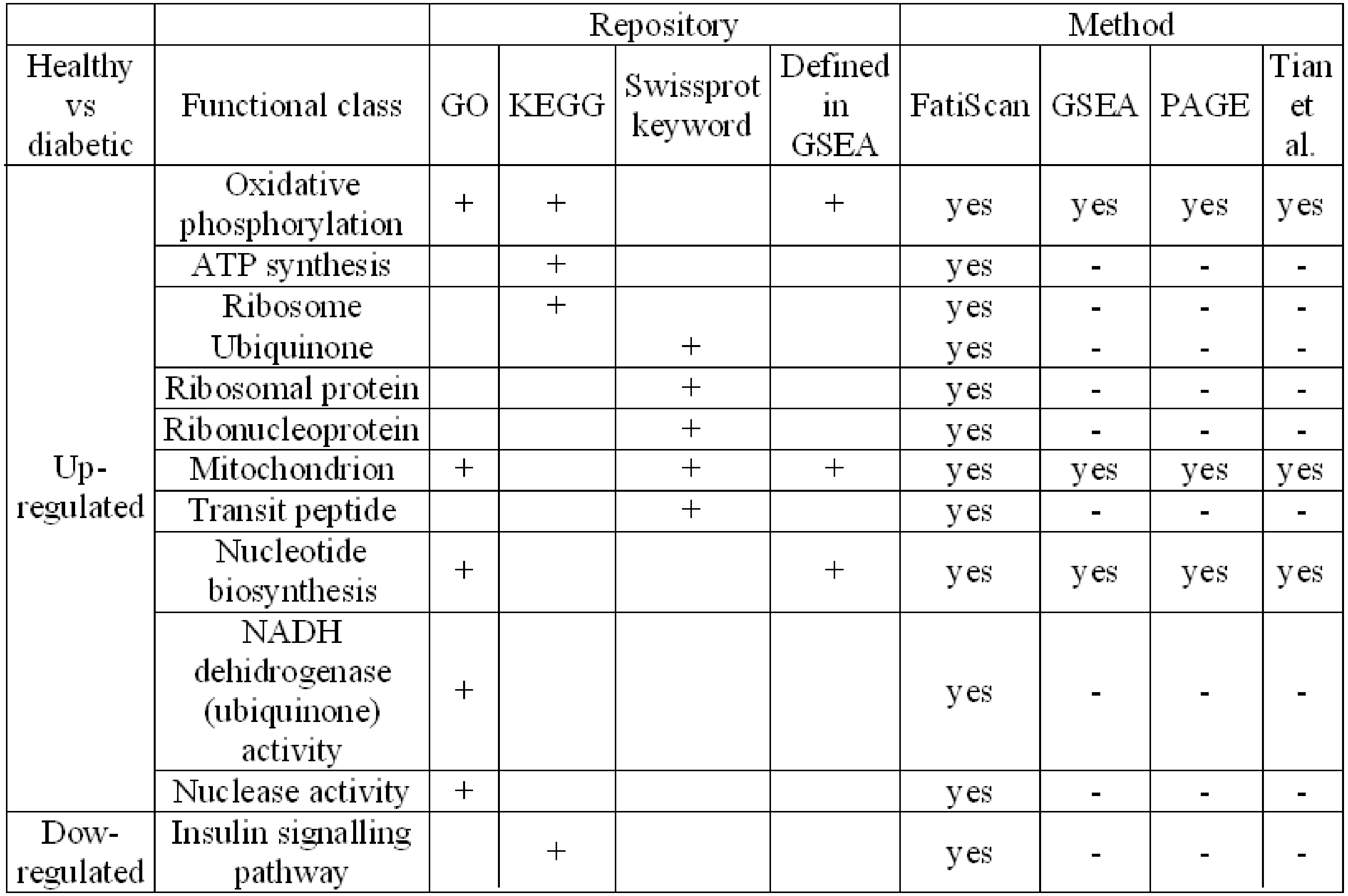
- Survival

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

+ Survival



Comparison of gene set 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	-	-	-

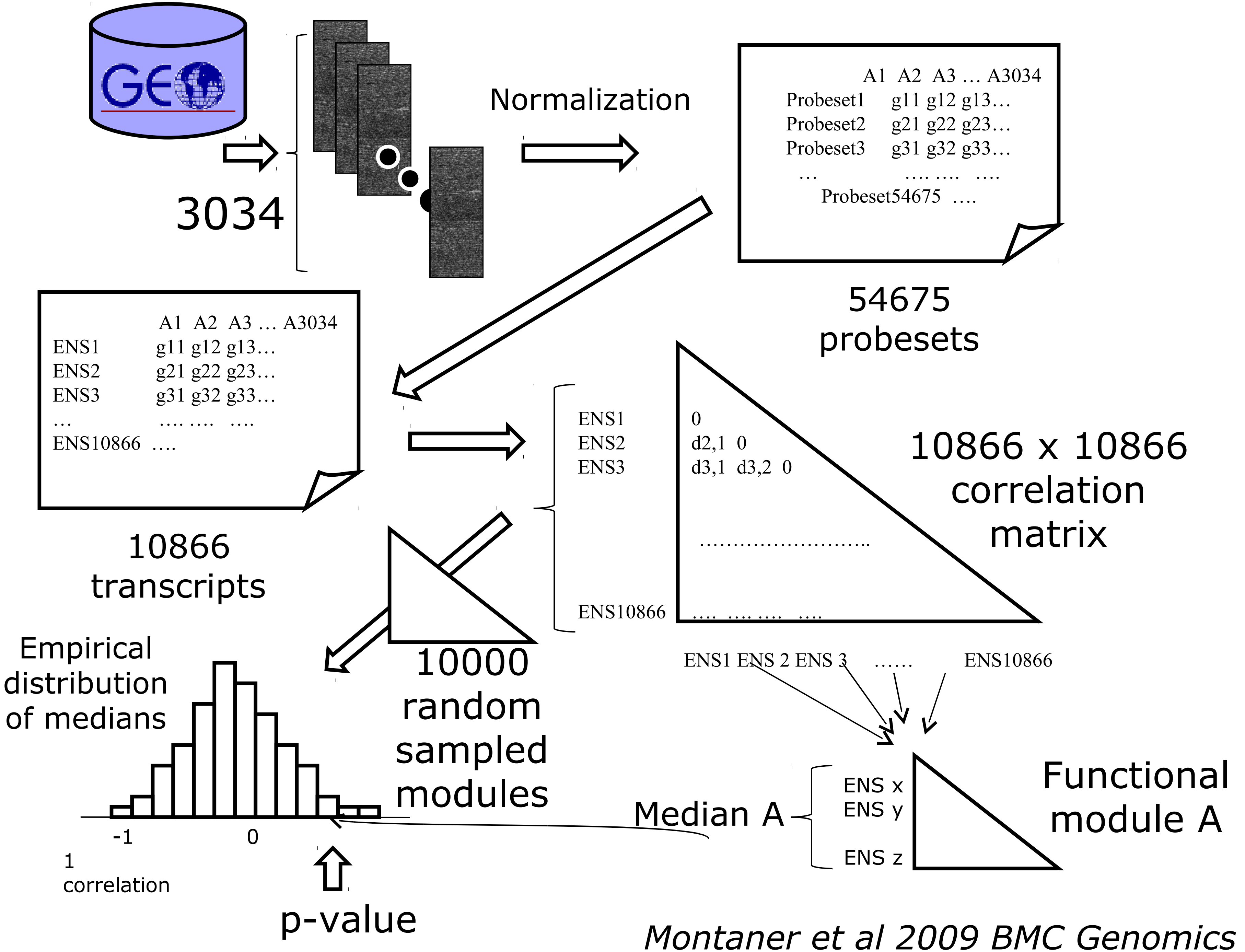
Terms from distinc repositories, reported by different methods in the diabetes dataset (Mootha et al., 2003)

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

Not a naïve and trivial question.

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

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



3034
Normalization
54675 probesets
10866 transcripts
10866 x 10866 correlation matrix
10000 random sampled modules
Empirical distribution of medians
-1
0
1
correlation
p-value
Median A
ENS x
ENS y
ENS z
Functional module A
ENS1 ENS 2 ENS 3 …… ENS10866

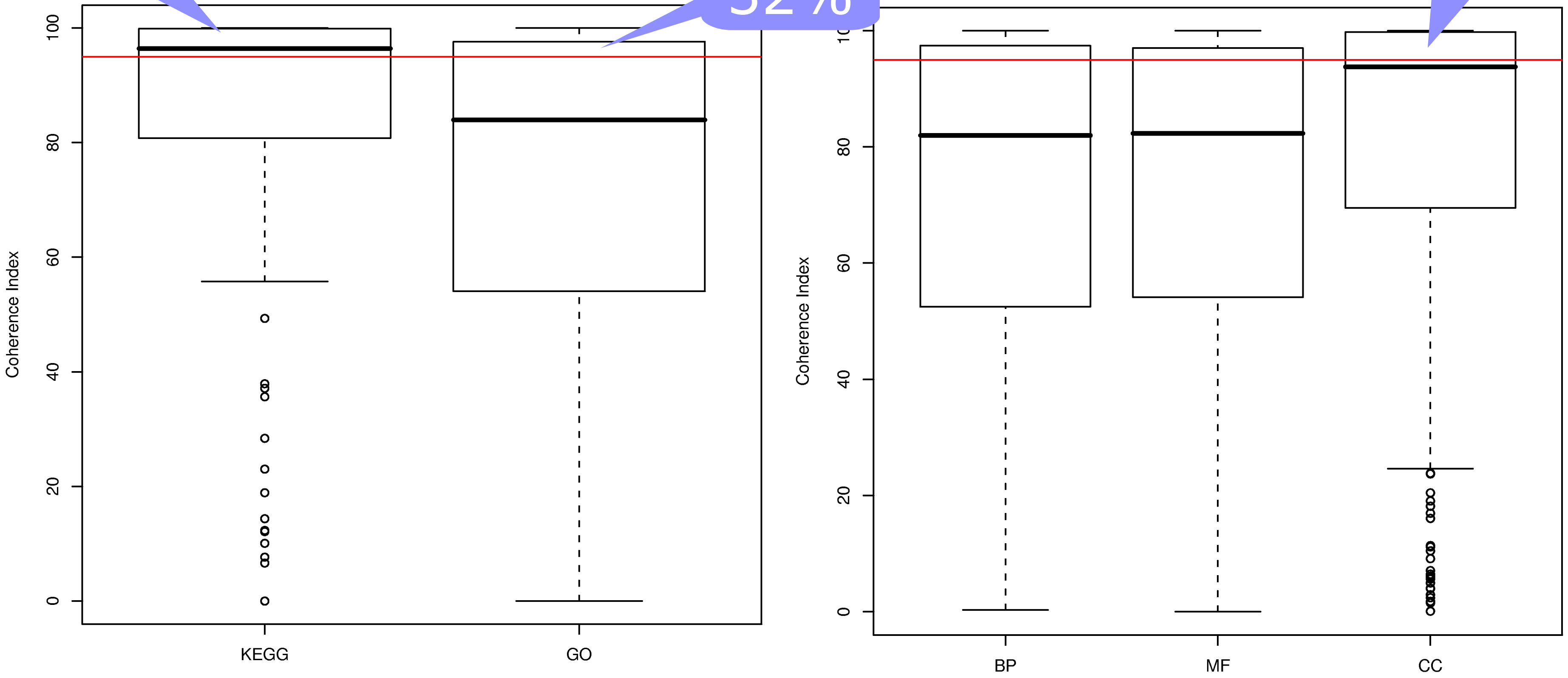
	A1	A2	A3	…	A3034
Probeset1	g11	g12	g13…		
Probeset2	g21	g22	g23…		
Probeset3	g31	g32	g33…		
…	….	….	….		
Probeset54675	….				

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

ENS1	0			
ENS2	d2,1	0		
ENS3	d3,1	d3,2	0	
	……………………..			
ENS10866	….	….	….	….

Montaner et al 2009 BMC Genomics

But are functional modules defining real co-expression entities?



57%

32%

46%

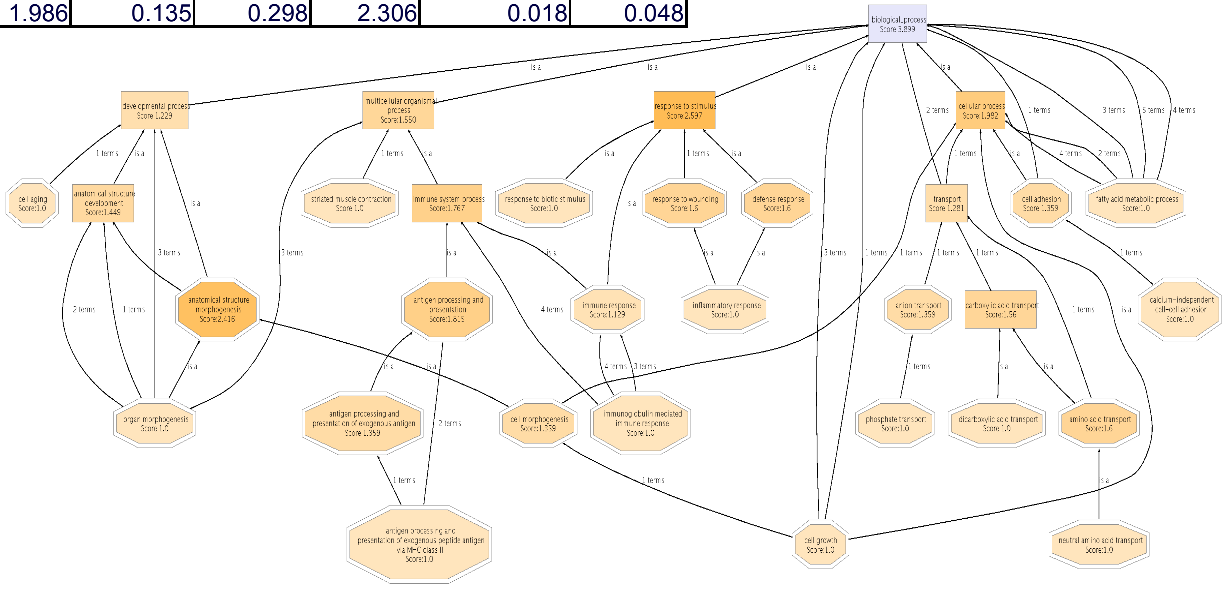
Coherence index: (1-p-value)*100.
CI > 95% means internal co-expression significantly higher than random co-expression

Montaner et al 2009 BMC Genomics

Weighting gene module membership by co-expression

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

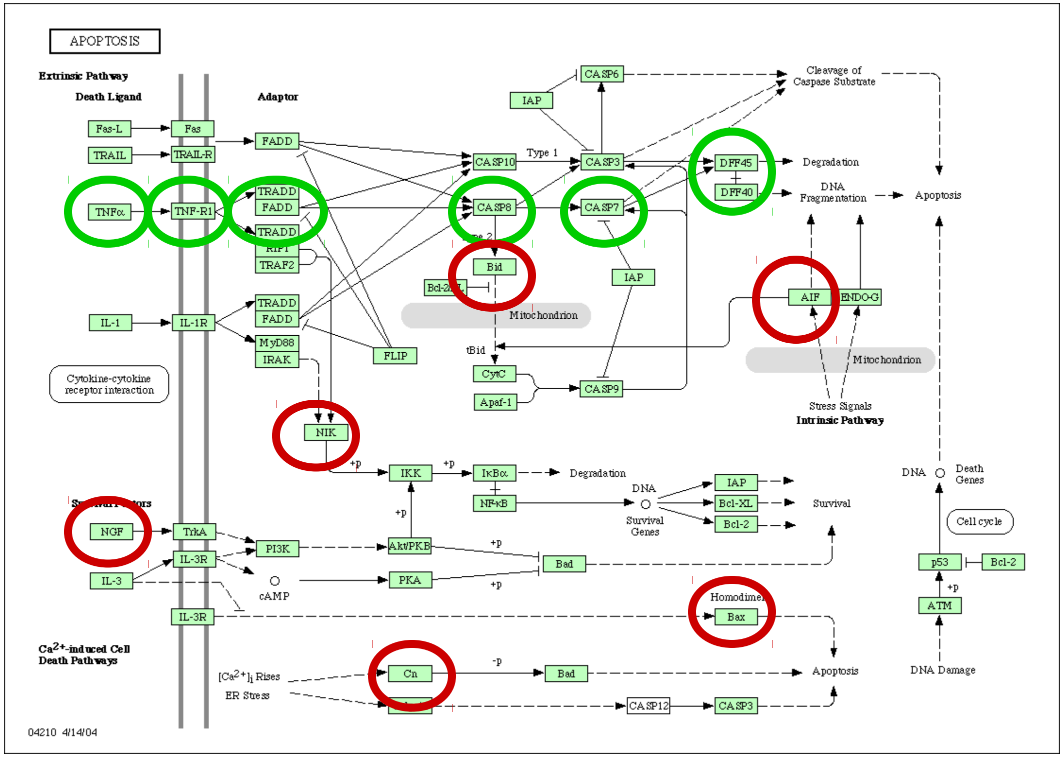
Very simple weight schema:
W=2 if correlation is positive
W=0.5 if negative
W=1 if not in the class


biological_process Score:3.899
developmental process Score:1.229
multicellular organismal process Score:1.550
response to stimulus Score:2.597
cellular process Score:1.982
cell aging Score:1.0
anatomical structure development Score:1.449
striated muscle contraction Score:1.0
immune system process Score:1.767
response to biotic stimulus Score:1.0
response to wounding Score:1.6
defense response Score:1.6
transport Score:1.281
cell adhesion Score:1.359
fatty acid metabolic process Score:1.0
anatomical structure morphogenesis Score:2.416
antigen processing and presentation Score:1.815
immune response Score:1.129
inflammatory response Score:1.0
anion transport Score:1.359
carboxylic acid transport Score:1.56
calcium-independent cell-cell adhesion Score:1.0
organ morphogenesis Score:1.0
antigen processing and presentation of exogenous antigen Score:1.359
cell morphogenesis Score:1.359
immunoglobulin mediated immune response Score:1.0
phosphate transport Score:1.0
dicarboxylic acid transport Score:1.0
amino acid transport Score:1.6
antigen processing and presentation of exogenous peptide antigen via MHC class II Score:1.0
cell growth Score:1.0
neutral amino acid transport Score:1.0

biological_process – combined Graph

Montaner et al 2009 BMC Genomics

Future directions



Testing hierarchies is better

Functions and pathways are correlated.

Testing models will increase our sensitivity

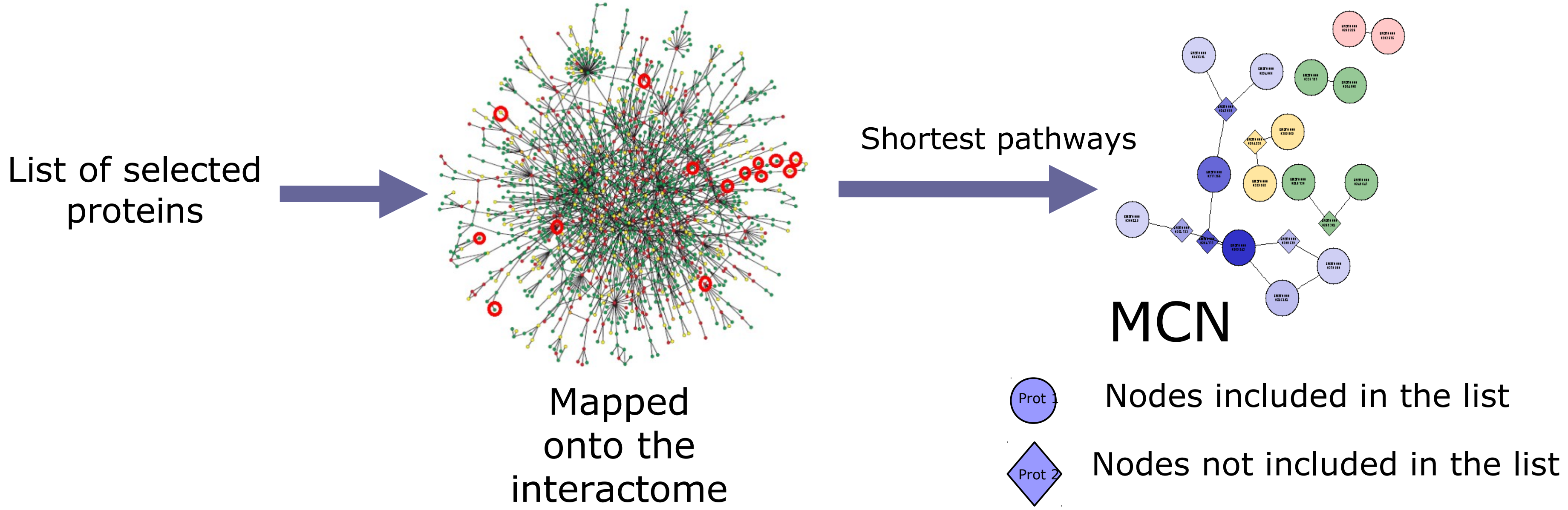
Pathways are not categorical variables

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

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

Mapped onto the interactome

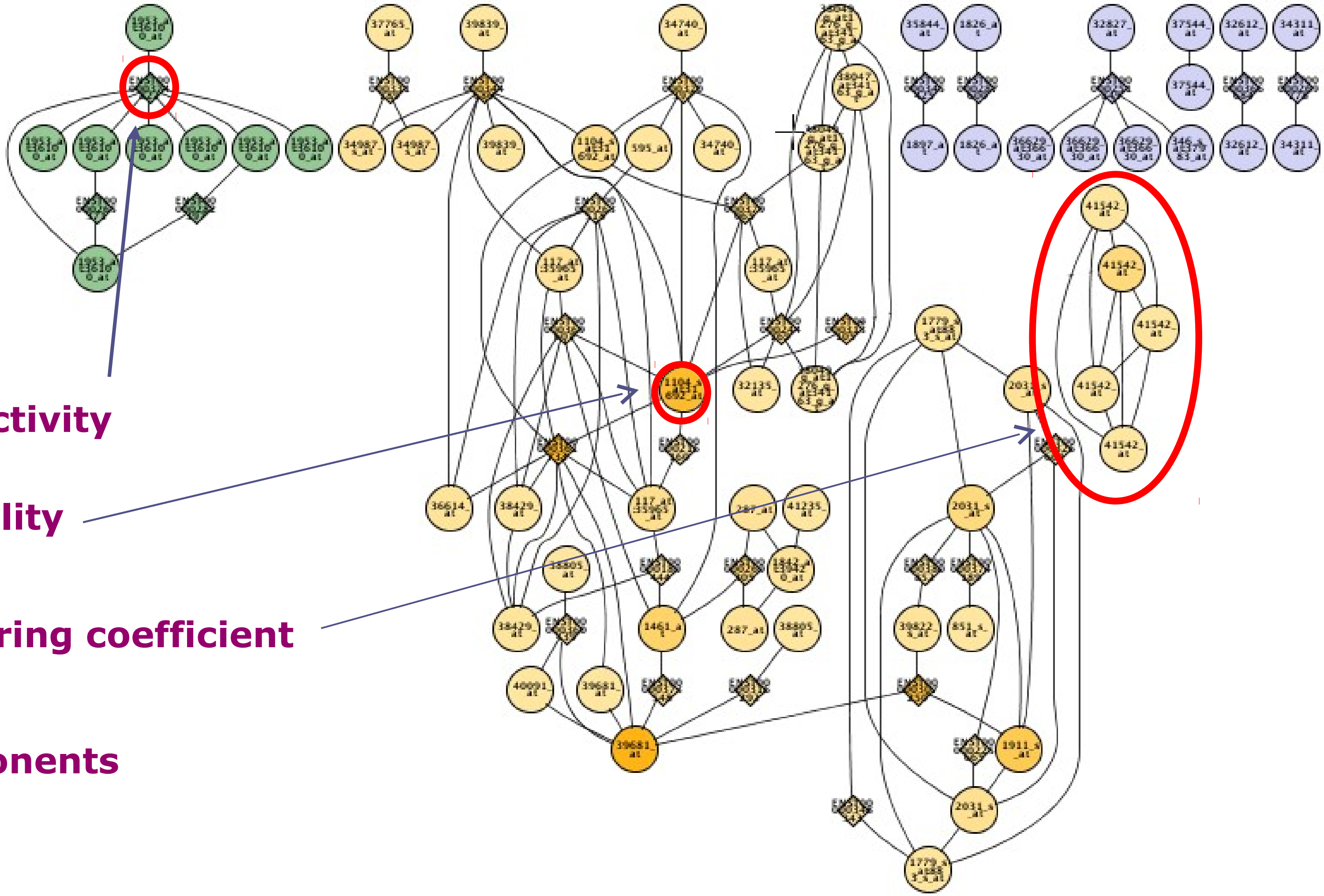
Shortest pathways

MCN

Nodes included in the list

Nodes not included in the list

Network parameters

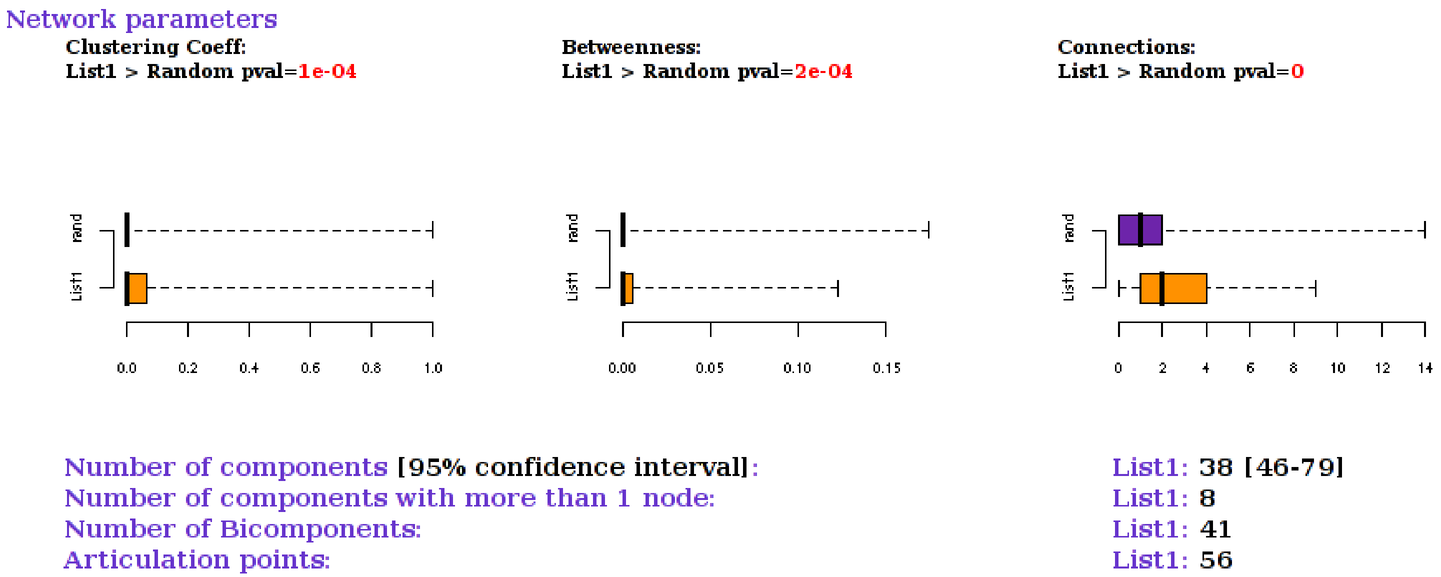


1. **Connectivity**
2. **Centrality**
3. **Clustering coefficient**
4. **Components**

Evaluation of the Minimum Connection Network (MCN)

Parameters to evaluate: connectivity, centrality , clustering coeficient, components

Distribution of the parameterrs' values versus distribution in random MCNs (compared through Kolmogorov-Smirnov tests)



Network parameters

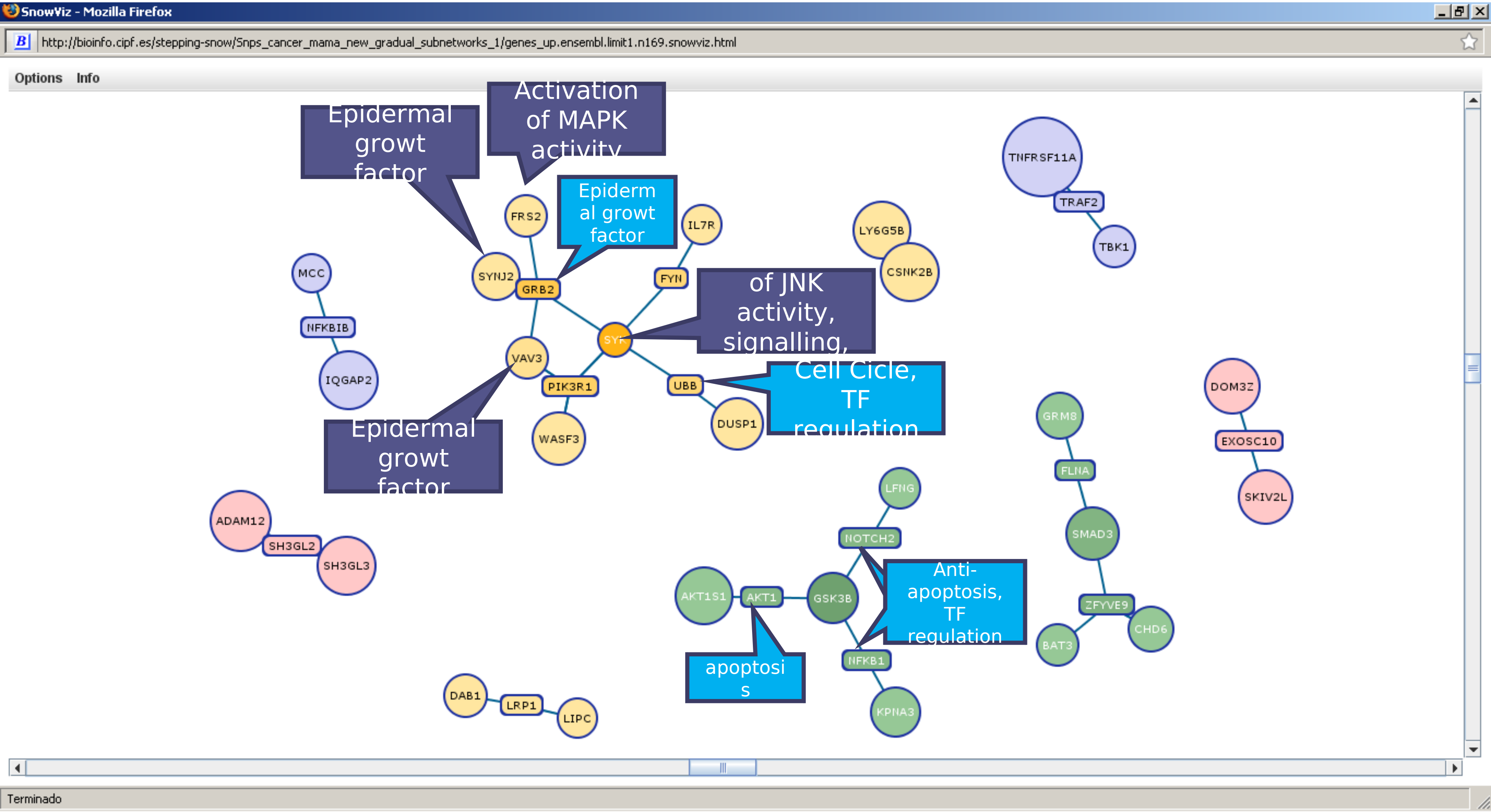
Clustering Coeff:
List1 > Random pval=1e-04

Betweenness:
List1 > Random pval=2e-04

Connections:
List1 > Random pval=0

Number of components [95% confidence interval]:	List1: 38 [46-79]
Number of components with more than 1 node:	List1: 8
Number of Bicomponents:	List1: 41
Articulation points:	List1: 56

Significant connections



SnowViz - Mozilla Firefox

http://bioinfo.cipf.es/stepping-snow/Snps_cancer_mama_new_gradual_subnetworks_1/genes_up.ensembl.limit1.n169.snowviz.html

Epidermal growt factor

Activation of MAPK activity

Epidermal growt factor

of JNK activity, signalling,

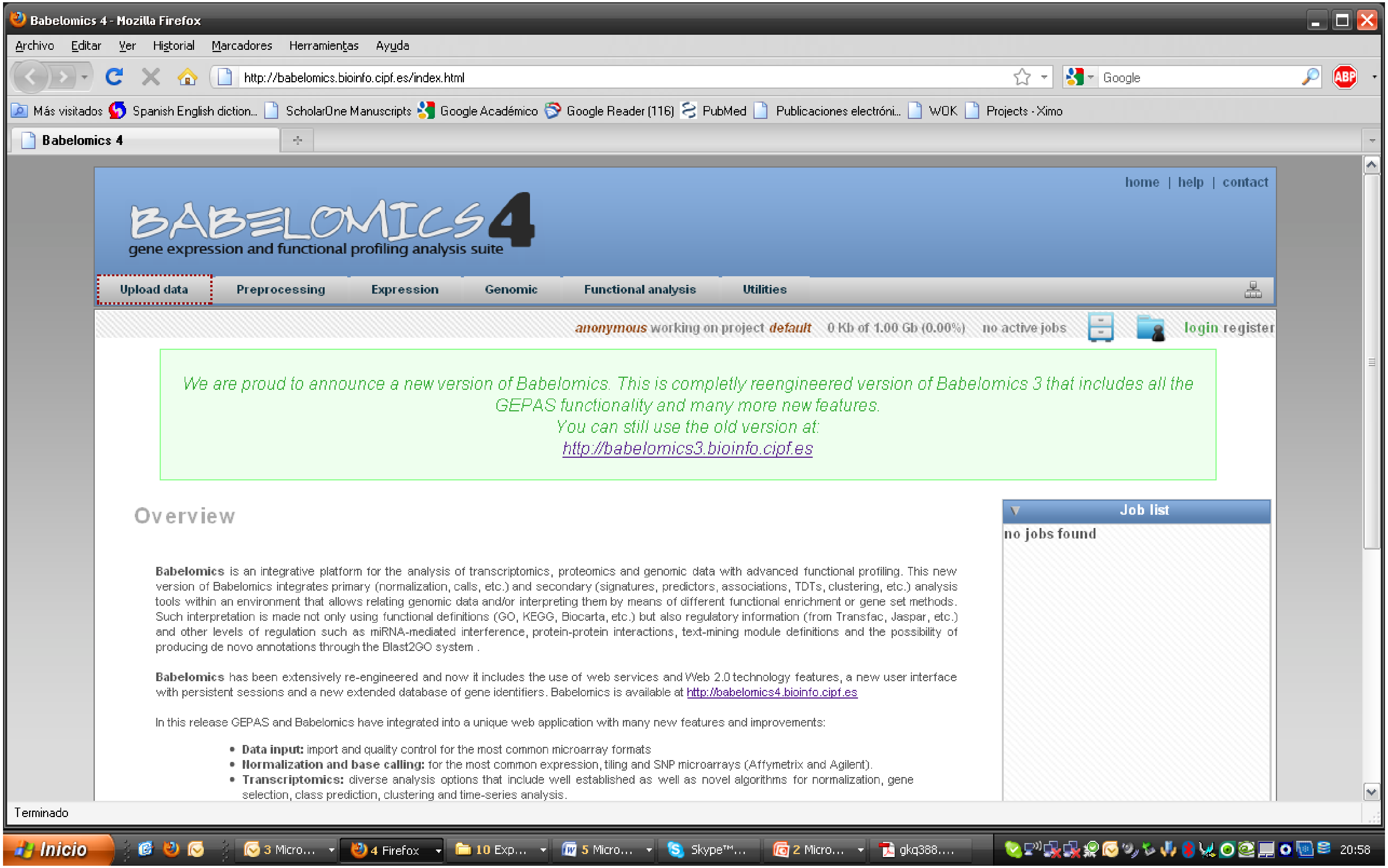
Cell Cicle, TF regulation

Epidermal growt factor

Anti-apoptosis, TF regulation

apoptosis

Babelomics



We are proud to announce a new version of Babelomics. This is completly reengineered version of Babelomics 3 that includes all the GEPAS functionality and many more new features.
You can still use the old version at:
http://babelomics3.bioinfo.cipf.es

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://babelomics4.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.

Since May 1st, Babelomics 4.0

Some numbers

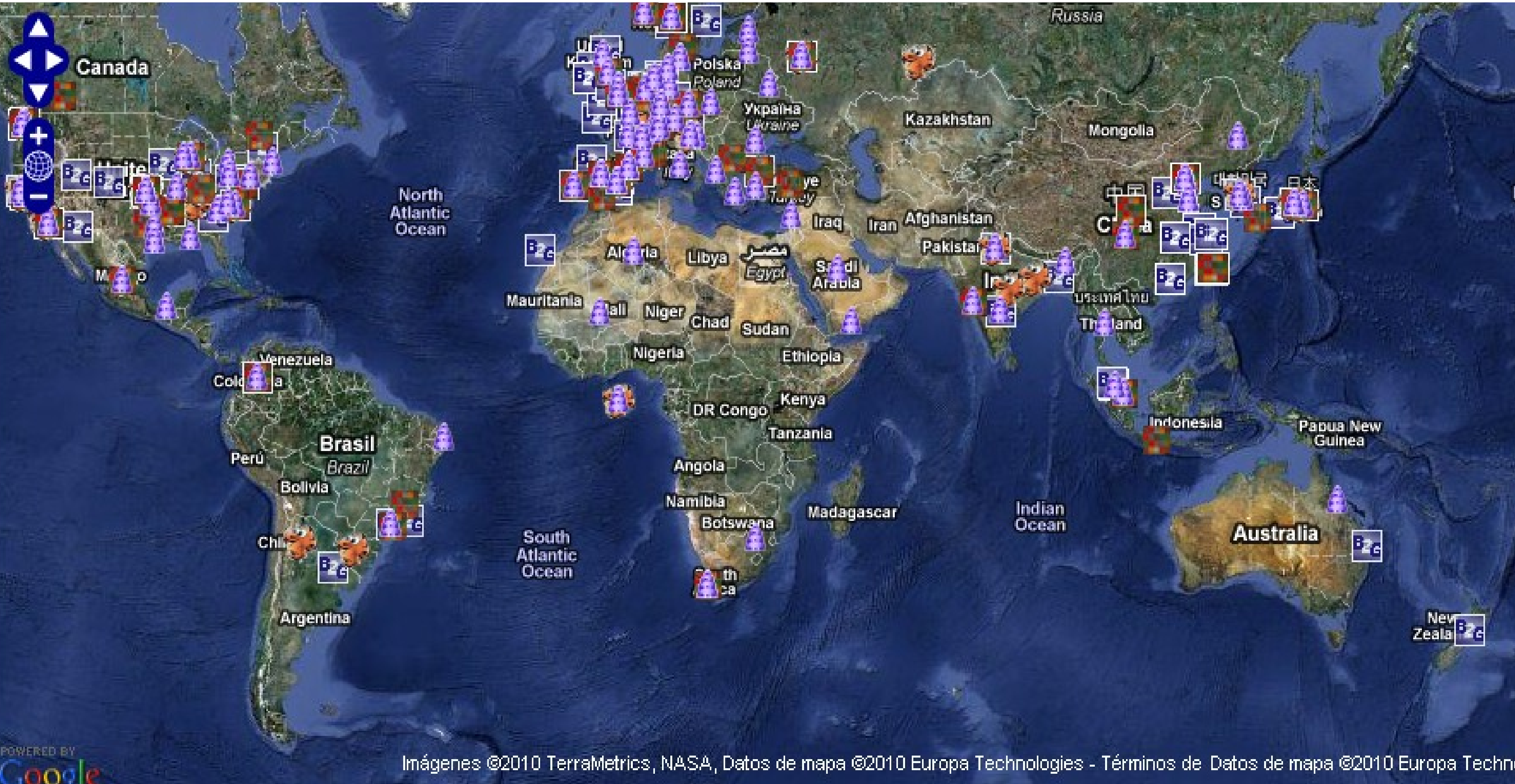
451 papers cite GEPAS (215 are SOTA cites)

632 papers cite Babelomics (442 are FatiGO cites)

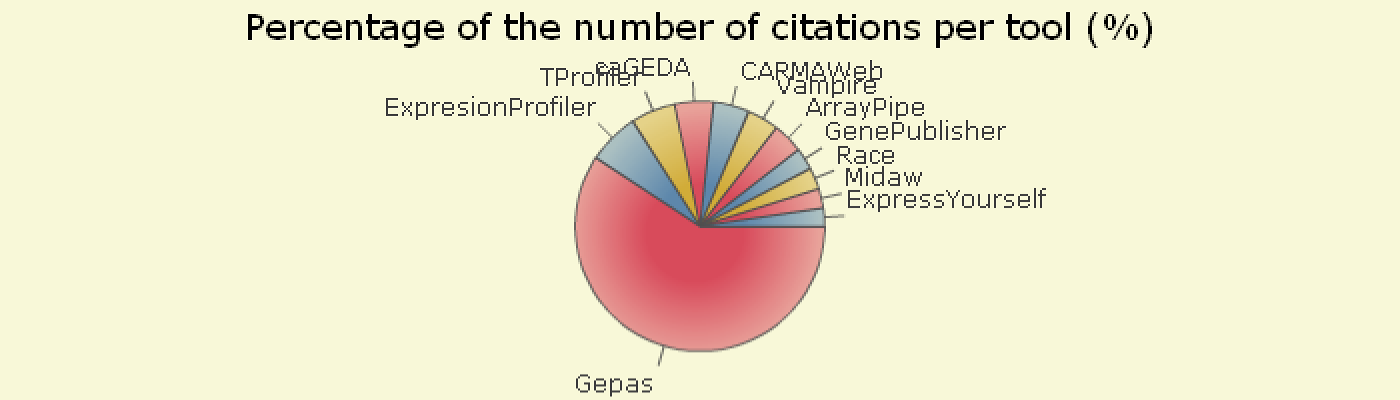
(source ISI Web of Knowledge, May 2010)

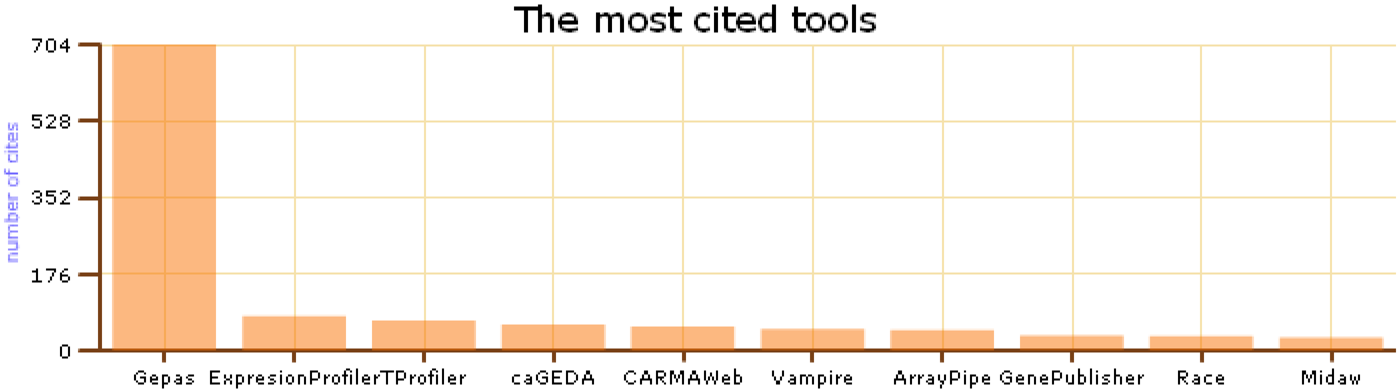
More than 150,000 experiments analysed during the last year.

More than 1000 experiments per day.

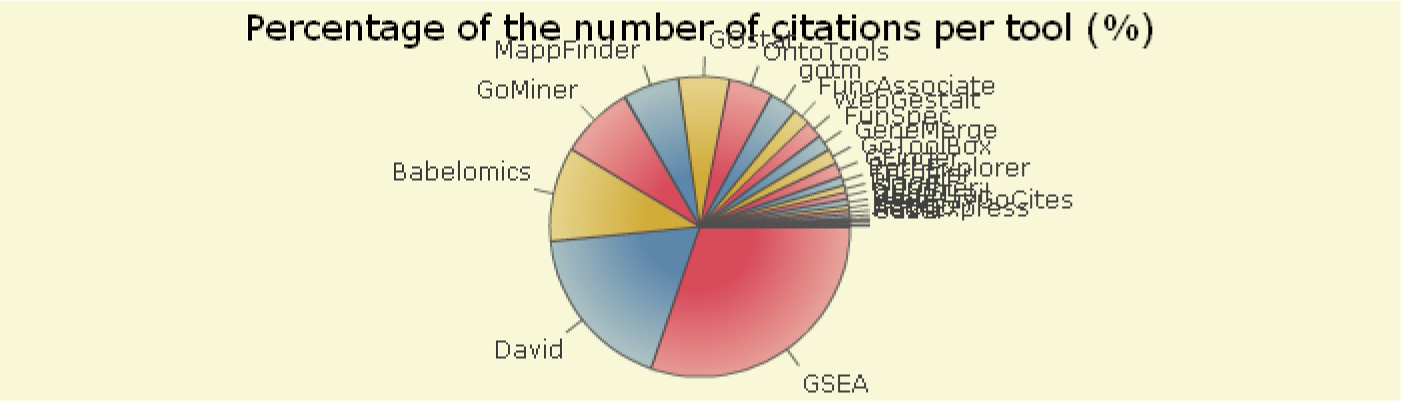


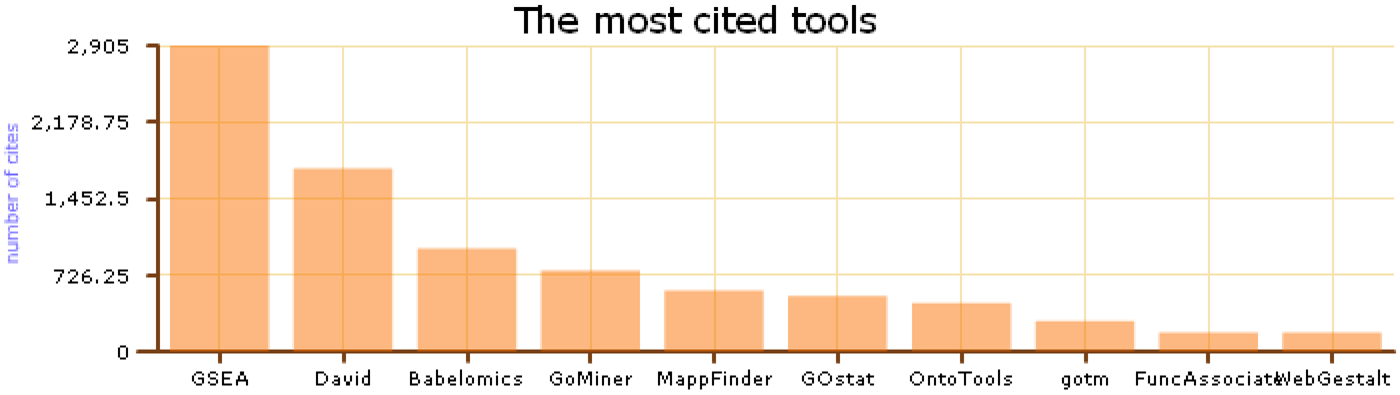
Tools for gene expression analysis



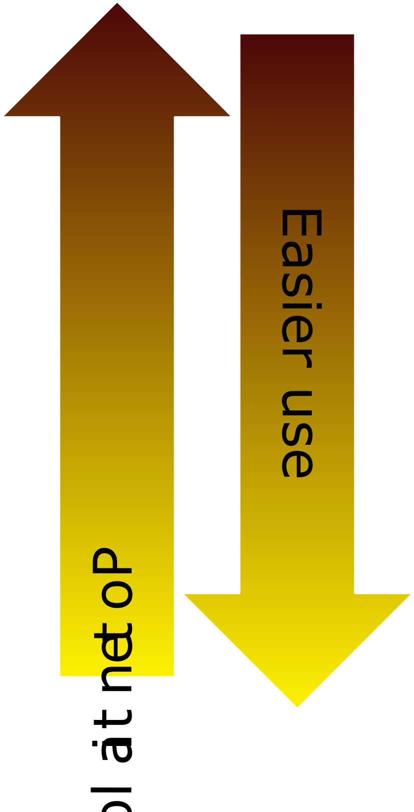


Tools for functional profiling





Other tools (non-commertial)



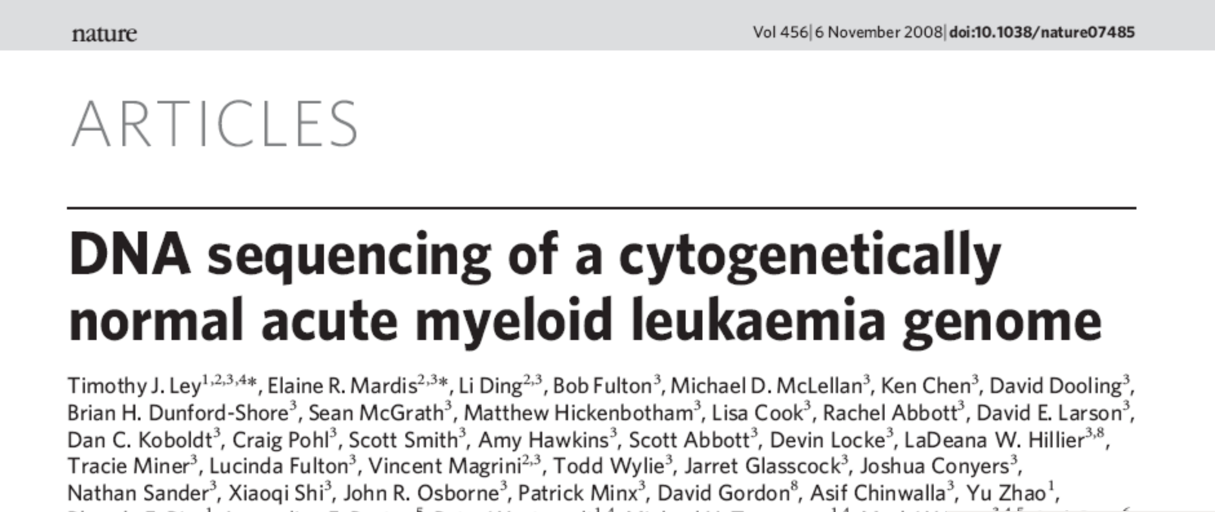
To cover more specific analysis requirements

Bioconductor: http://www.bioconductor.org

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

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

What is next?



nature

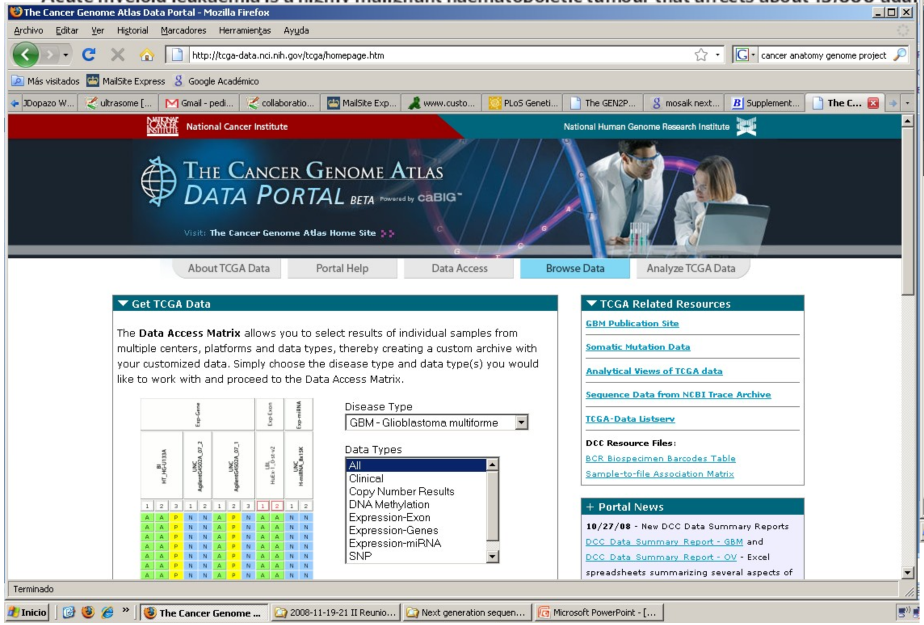
Vol 456| 6 November 2008| doi:10.1038/nature07485

ARTICLES

DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome

Timothy J. Ley[1,2,3,4]*, Elaine R. Mardis[2,3]*, Li Ding[2,3], Bob Fulton[3], Michael D. McLellan[3], Ken Chen[3], David Dooling[3], Brian H. Dunford-Shore[3], Sean McGrath[3], Matthew Hickenbotham[3], Lisa Cook[3], Rachel Abbott[3], David E. Larson[3], Dan C. Koboldt[3], Craig Pohl[3], Scott Smith[3], Amy Hawkins[3], Scott Abbott[3], Devin Locke[3], LaDeana W. Hillier[3,8], Tracie Miner[3], Lucinda Fulton[3], Vincent Magrini[2,3], Todd Wylie[3], Jarret Glasscock[3], Joshua Conyers[3], Nathan Sander[3], Xiaoqi Shi[3], John R. Osborne[3], Patrick Minx[3], David Gordon[8], Asif Chinwalla[3], Yu Zhao[1], Rhonda E. Ries[1], Jacqueline E. Payton[5], Peter Westervelt[1,4], Michael H. Tomasson[1,4], Mark Wa
Jennifer Ivanovich[4,7], Sharon Heath[1,4], William D. Shannon[1,4], Rakesh Nagarajan[4,5], Matthew
Daniel C. Link[1,4], Timothy A. Graubert[1,4], John F. DiPersio[1,4] & Richard K. Wilson[2,3,4]

Acute myeloid leukaemia is a highly malignant haematopoietic tumour that affects about 13,000 adul



▼ Get TCGA Data

The **Data Access Matrix** allows you to select results of individual samples from multiple centers, platforms and data types, thereby creating a custom archive with your customized data. Simply choose the disease type and data type(s) you would like to work with and proceed to the Data Access Matrix.

Exp-Gene								Exp-Exon		Exp-miRNA	
BI HT_HG-U133A			UNC AgilentG4502A_07_2		UNC AgilentG4502A_07_1			LBL HuEx-1_0-st-v2		UNC H-miRNA_8x15K	
1	2	3	1	2	1	2	3	1	2	1	2
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N
A	A	P	N	N	A	P	N	A	A	N	N

Disease Type

GBM - Glioblastoma multiforme

Data Types

All
Clinical
Copy Number Results
DNA Methylation
Expression-Exon
Expression-Genes
Expression-miRNA
SNP

▼ TCGA Related Resources

GBM Publication Site

Somatic Mutation Data

Analytical Views of TCGA data

Sequence Data from NCBI Trace Archive

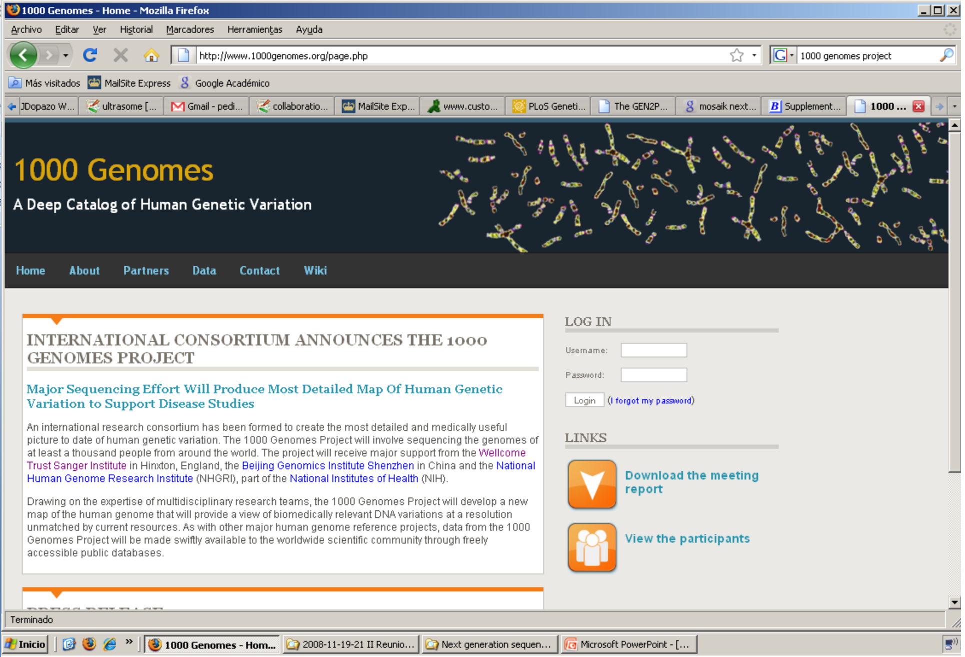
TCGA-Data Listserv

DCC Resource Files:

BCR Biospecimen Barcodes Table
Sample-to-file Association Matrix

+ Portal News

10/27/08 - New DCC Data Summary Reports DCC Data Summary Report - GBM and DCC Data Summary Report - OV - Excel spreadsheets summarizing several aspects of



INTERNATIONAL CONSORTIUM ANNOUNCES THE 1000 GENOMES PROJECT

Major Sequencing Effort Will Produce Most Detailed Map Of Human Genetic Variation to Support Disease Studies

An international research consortium has been formed to create the most detailed and medically useful picture to date of human genetic variation. The 1000 Genomes Project will involve sequencing the genomes of at least a thousand people from around the world. The project will receive major support from the Wellcome Trust Sanger Institute in Hinxton, England, the Beijing Genomics Institute Shenzhen in China and the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health (NIH).

Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

LOG IN

Username: ____

Password: ____

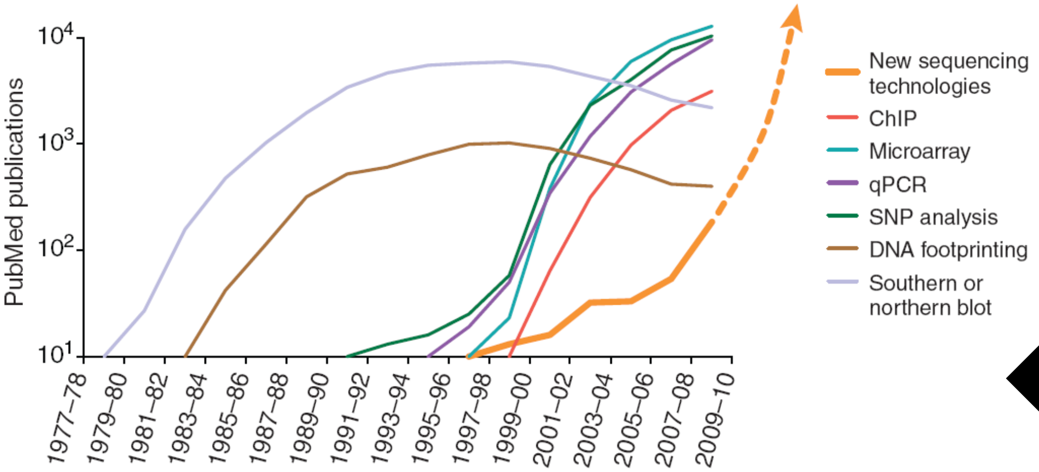
Login (I forgot my password)

LINKS

Download the meeting report

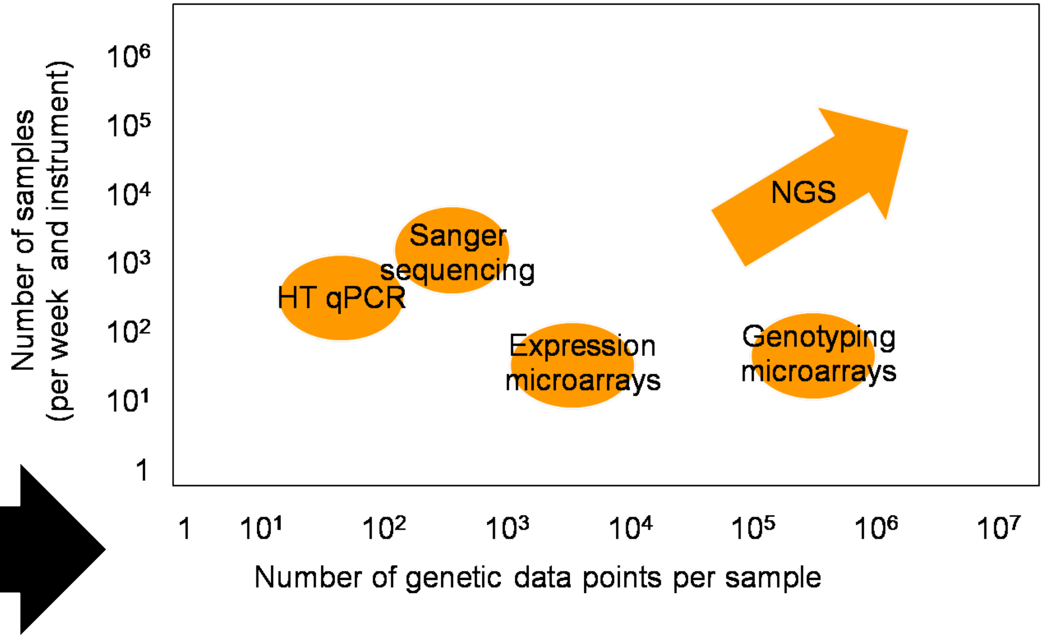
View the participants

Next generation technology is here

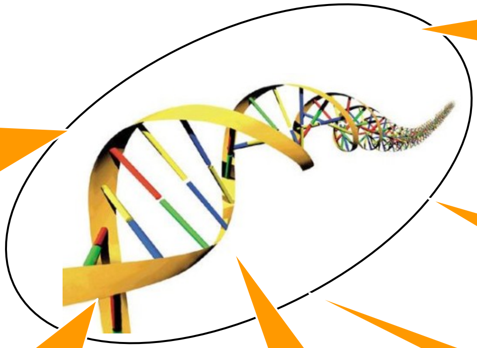


Observed and expected trend of publications in which NGS is being used.

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

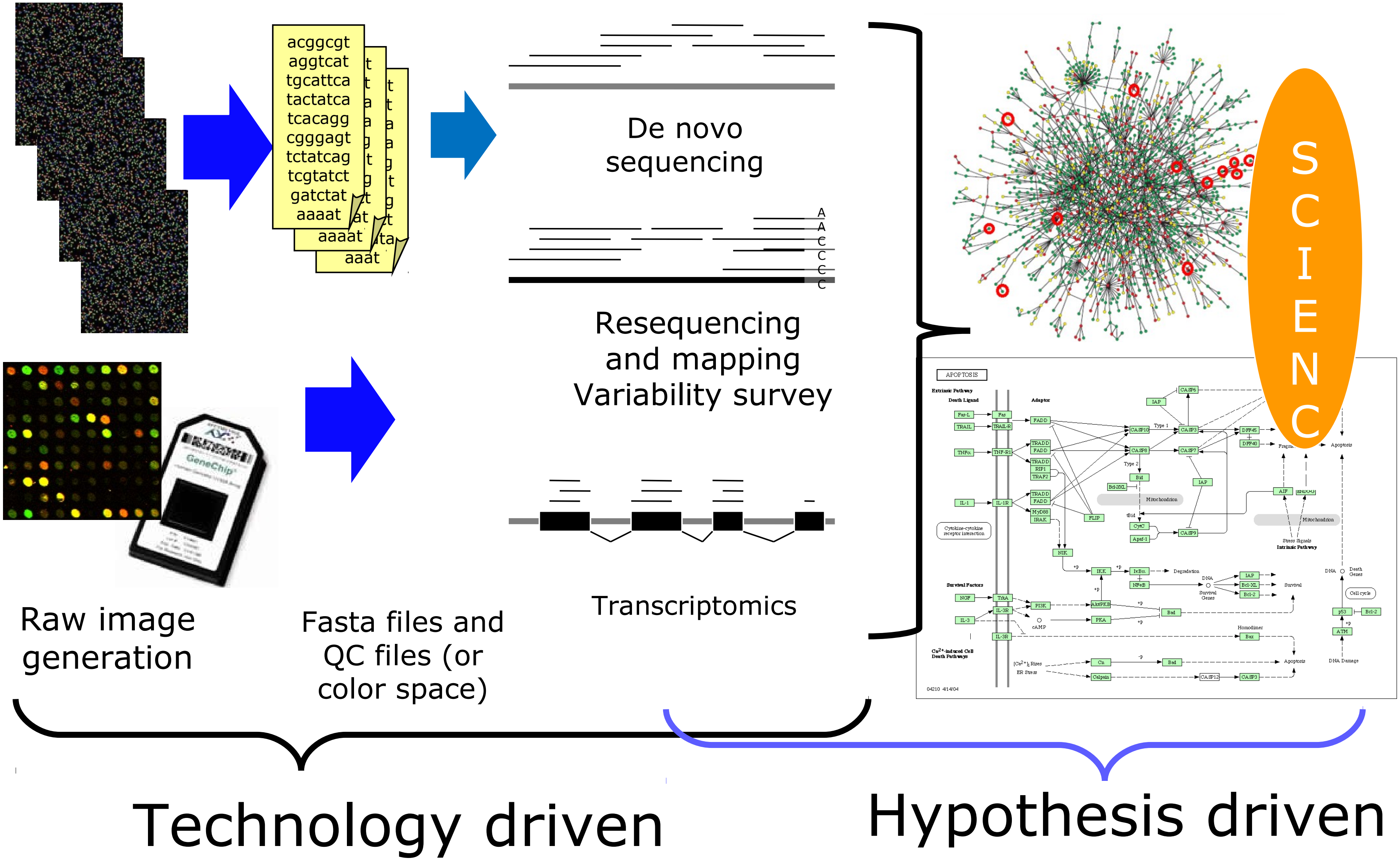
De novo
sequencing

Copy number
variation

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

Metagenomics
Metatranscriptomics

Pipeline general of analysis



De novo sequencing

Resequencing and mapping Variability survey

Transcriptomics

Raw image generation

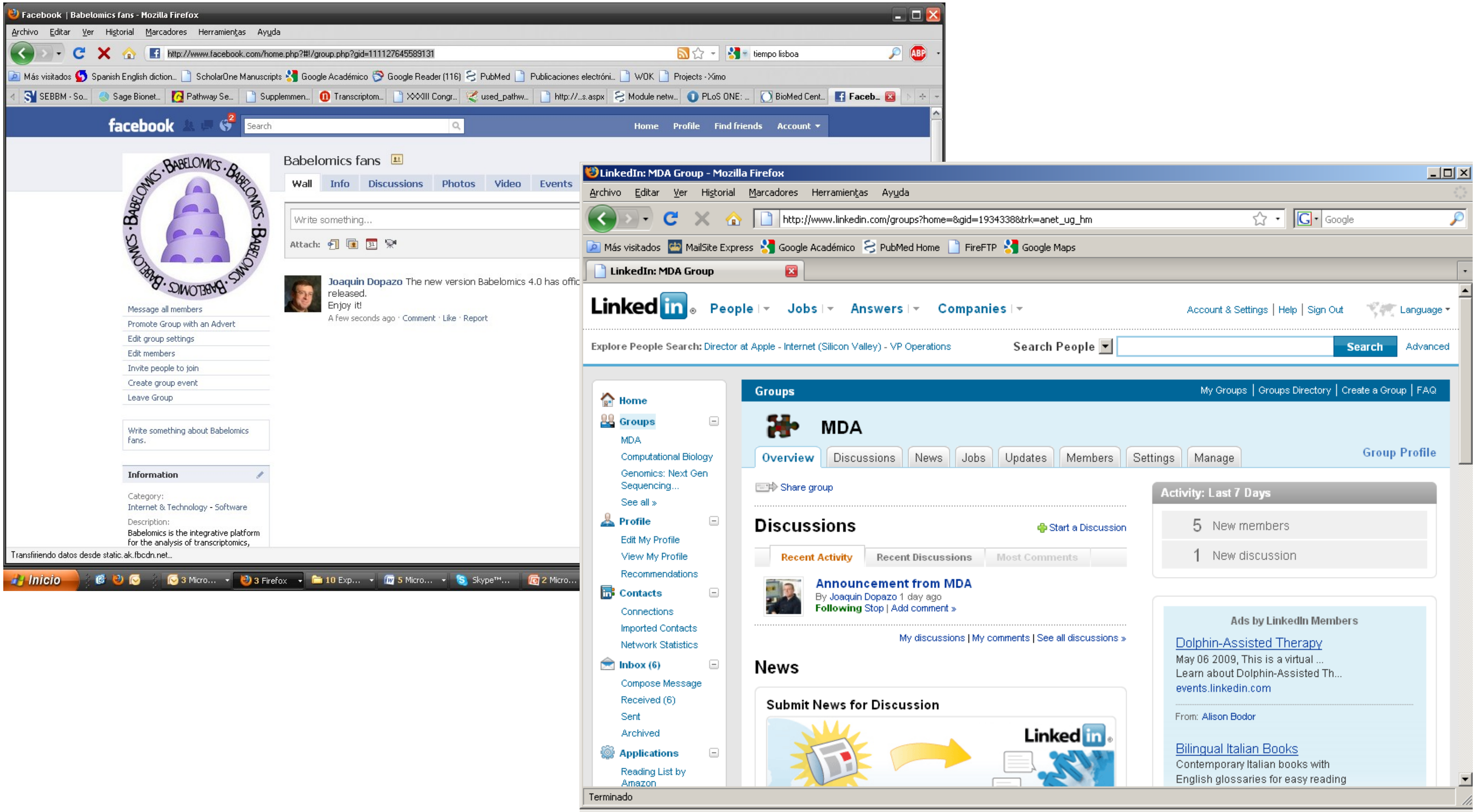
Fasta files and QC files (or color space)

Technology driven

Hypothesis driven

SOCIAL:

MDA group in Linked-in
Babelomics group in Facebook



Facebook | Babelomics fans - Mozilla Firefox

http://www.facebook.com/home.php?#!/group.php?gid=111127645589131

Babelomics fans

Write something...

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

Write something about Babelomics fans.

Information

Category:
Internet & Technology - Software

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

Transfiriendo datos desde static.ak.fbcdn.net...

LinkedIn: MDA Group - Mozilla Firefox

http://www.linkedin.com/groups?home=&gid=1934338&trk=anet_ug_hm

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

MDA

Share group

Discussions

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

My discussions | My comments | See all discussions »

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

Terminado

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

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

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

PRINCIPE FELIPE
CENTRO DE INVESTIGACION