Hacia un cambio de paradigma en el análisis de datos genómicos

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

INB

INSTITUTO NACIONAL DE BIOINFORMATICA

RTICC

ciberer

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Background

The road of excess leads to the palace of wisdom

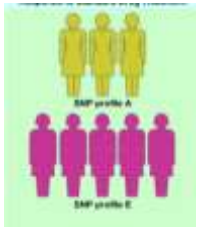
(*William Blake, 28 November 1757 – 12 August 1827, poet, painter, and printmaker*)

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...



...produces the final phenotype

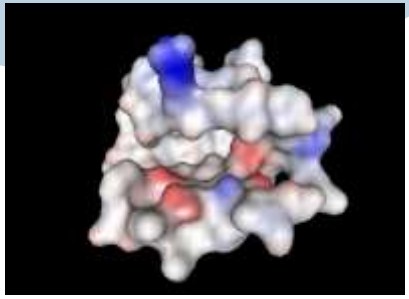
...code for proteins...

```
>protein kunase
acctgttgatggcgacagggactgtatgctg
atctatgctgatgcatgcatgctgactactga
tgtgggggctattgacttgatgtctatc....
```

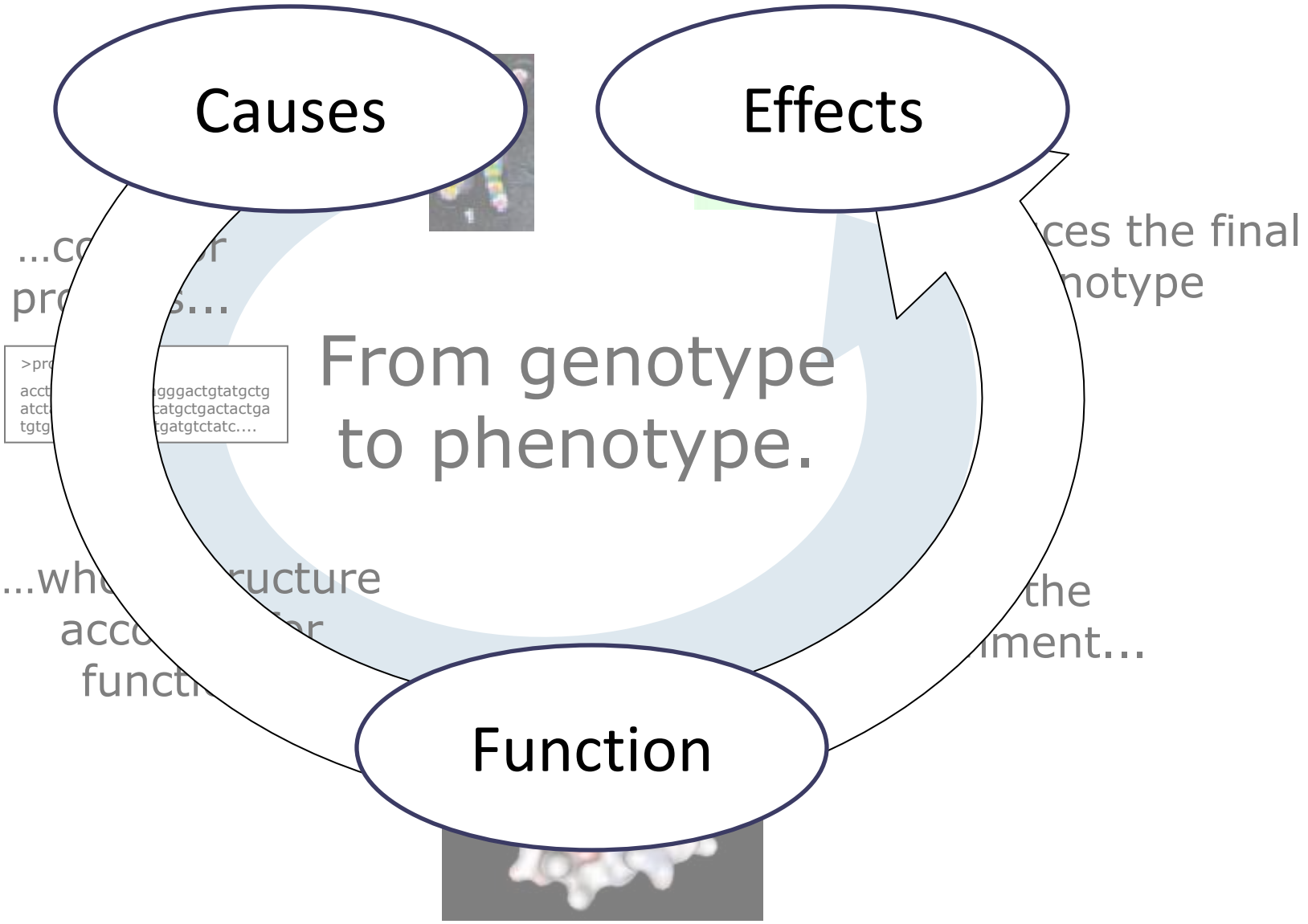
From genotype to phenotype.

...whose structure accounts for function...

...plus the environment...



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



From genotype to phenotype.

(in the functional post-genomics scenario)

Next Generation Sequencing
10^9bp per round

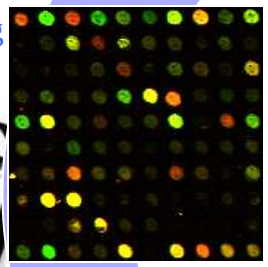
Genes in the DNA...

>protein kunase

acctgttgatggcgacagggactgtatgctgatct
atgctgatgcatgcatgctgactactgatgtgggg
gctattgacttgatgtctatc....

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

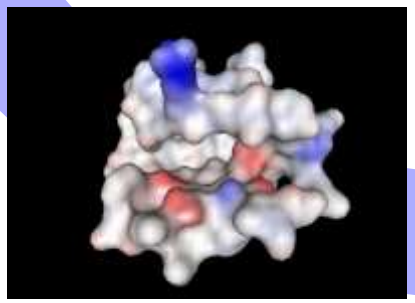
A typical tissue is expressing among 5000 and 10000 genes



...code for proteins...

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

100K-500K proteins

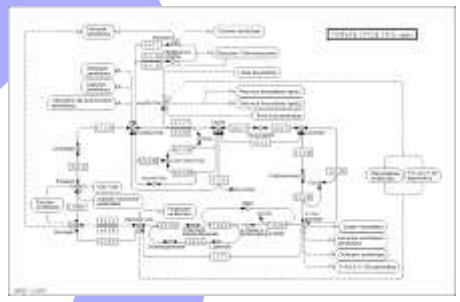


...that account for function if...

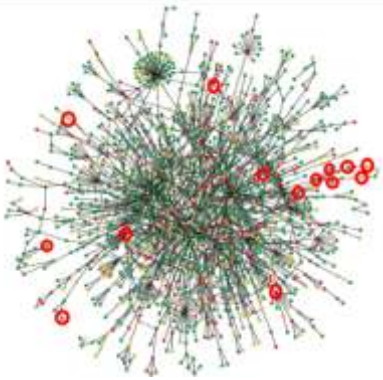
...which can be different because of the variability.

15 million SNPs

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



...conforming complex interaction networks...



...in cooperation with other proteins...

Each protein has an average of 8 interactions

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

>protein kunase

...when ex
proper
place

Causes

Effects

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

A typical tissue is expressing among 5000 and 10000 genes

...code proteins

From genotype to phenotyp

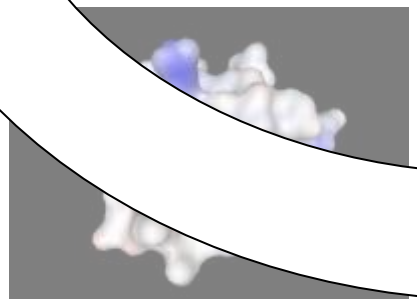
(in the functional genomics scen

Function (modules of proteins)

plex
ks...

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

100K-500K proteins



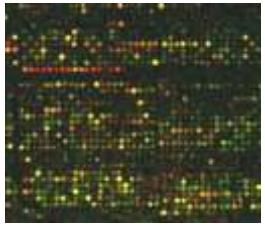
...whose structures account for function...

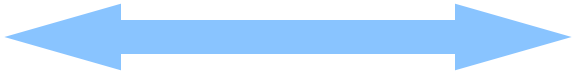
ration
other
proteins...

Each protein has an average of 8 interactions

Functional genomics

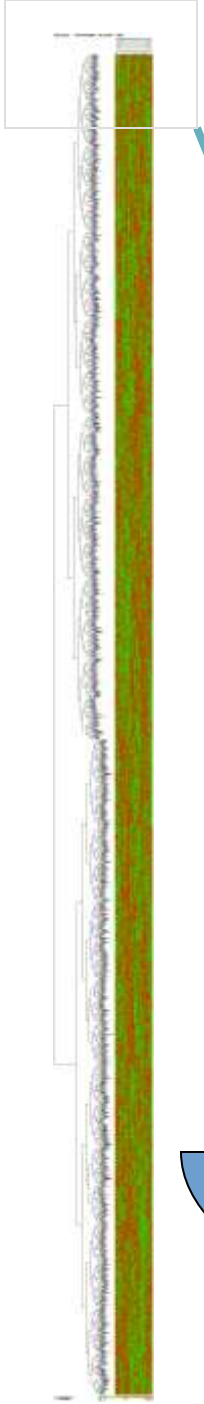
Differences at phenotype level are the visible cause of differences at molecular level which, in many cases, can be detected by measuring genomic mutations or the levels of gene expression, etc. The same holds for different experiments, treatments, strains, etc.





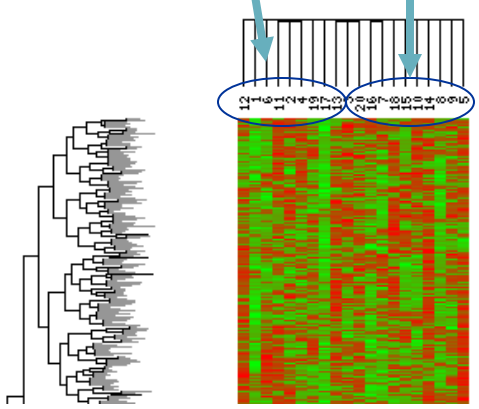
Lets have a closer look to the way we work in functional genomics with high-throughput data.

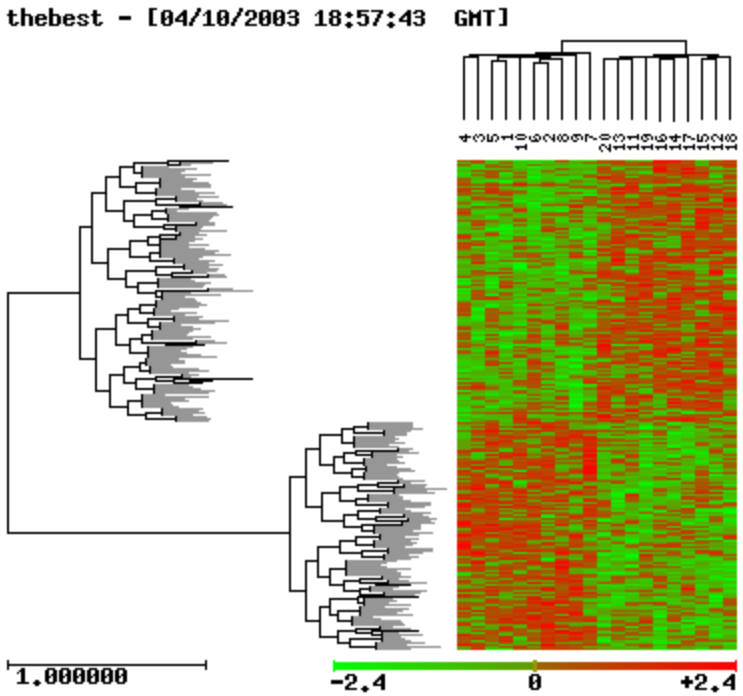
A closer look to a simple problem. Finding signatures , which implies gene selection for class discrimination



~25,000 genes

Case(10)/control(10)



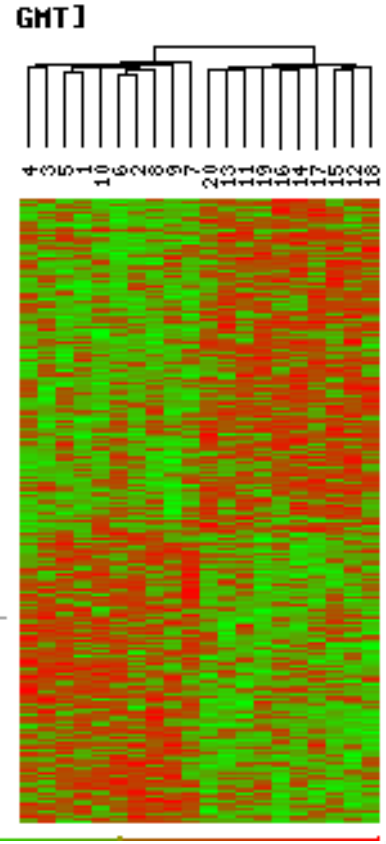


Genes differentially expressed among classes (t-test), with a p-value < 0.05

Signature

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

This is a multiple-testing statistic contrast.



Row	ID	unadj.p	adj_p	FDR_indep	FDR_dep	obs_stat
1630	1630	0.00019998	0.152685	0.49995	1	5.47044
4148	4148	0.00019998	0.746225	0.49995	1	4.49902
3348	3348	0.0009999	0.983002	0.861025	1	4.01726
3449	3449	0.00149985	0.986401	0.861025	1	3.99374
3443	3443	0.00129987	0.9959	0.861025	1	3.86046
4703	4703	0.00169983	0.9996	0.861025	1	3.7251
2731	2731					3.66628
14	14					3.62427
39	39					3.60596
1062	1062					3.58109
3738	3738					3.52935
1840	1840					3.43721
1007	1007					3.41937
1542	1542					3.41428
1360	1360					3.4025
844	844					3.40212
4631	4631					3.37412
11	11	0.00539946	1	0.8888	1	3.36813
4102	4102	0.00219978	1	0.861025	1	3.35909
4716	4716	0.00439956	1	0.8888	1	3.35235
3793	3793	0.00279972	1	0.861025	1	3.22175
3488	3488	0.0069993	1	0.8888	1	3.12957
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!

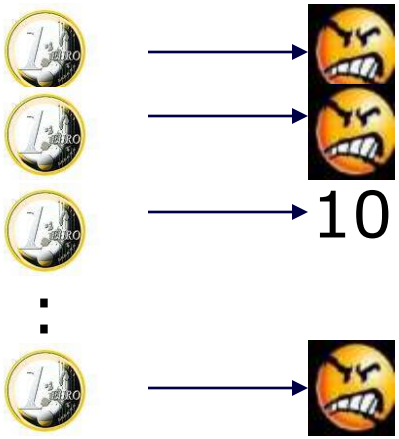
Some random sequences of values might look like if they were differentially distributed among two classes

The rationale for 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

Only a perturbing comment: do genes behave like coins?

The curse of dimensionality

The more we see the less we can be confirm

As we historically have moved from gene-targeted studies to GWAS or microarray-based gene expression studies we have gained a 1000x or even more in resolution but we have lost a lot in testing power.

I other words: our gene signatures, associated SNPs, etc., constitute an under-representation of the biologically relevant genes

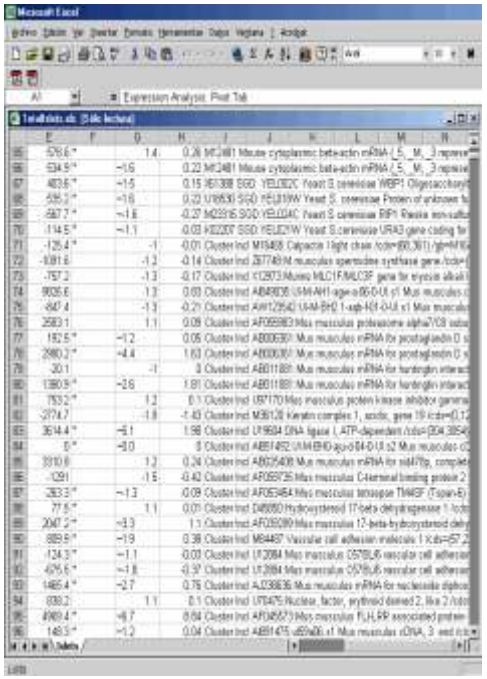
So far so good… we know the genes but, what are they doing in the cell?

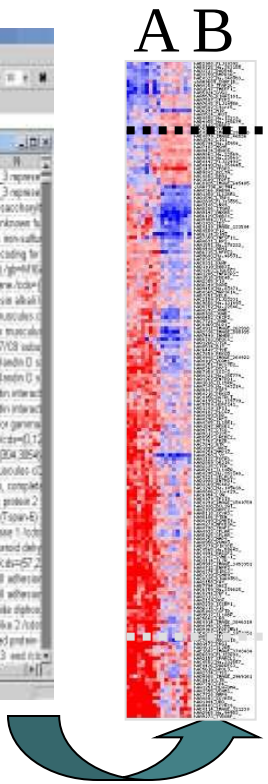
The data...

…how are structured?

What are these groups?

What is this gen?

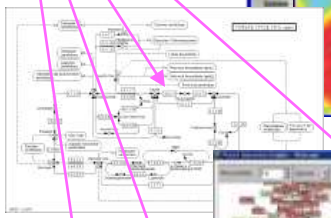


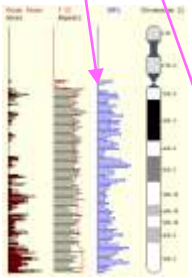


?

Cell cycle...

DBs Information



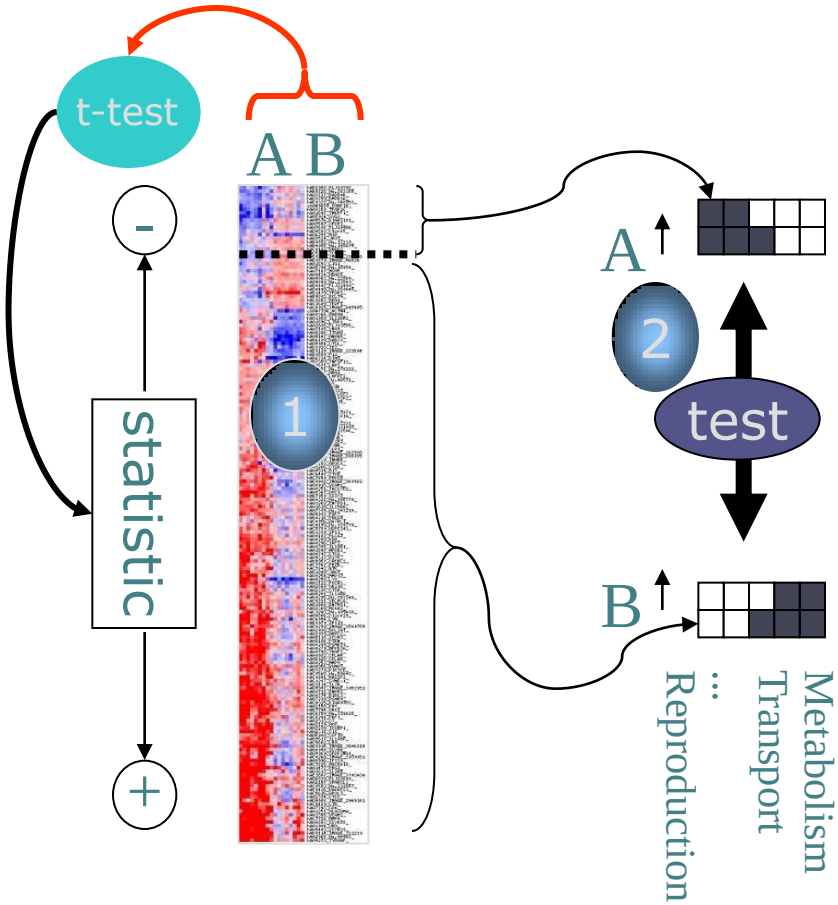


Analysis

Functional profiling

Links

Testing for functional enrichment

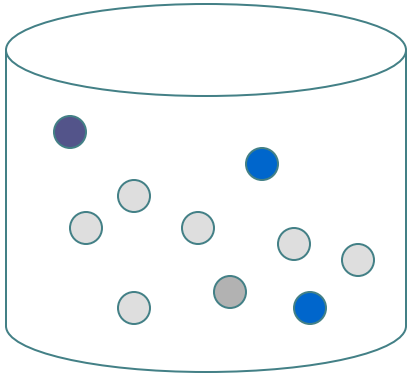


1 Genes are selected based on their experimental values and...

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

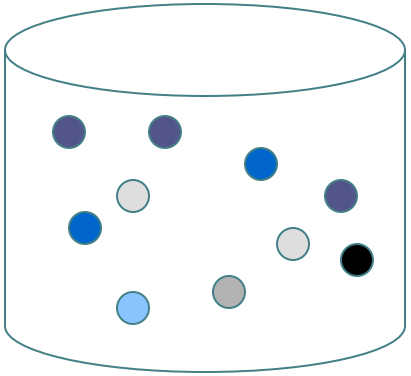
Testing for functional enrichment

Selected (A)



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

Background (B)



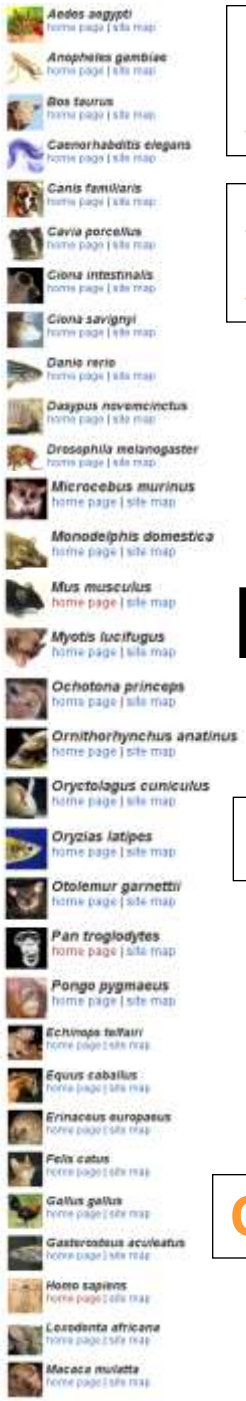
	Biosynthesis	Other	
	6	4	A
	2	8	B

The popular Fisher's test

Biosynthesis 60% ○ ⟷ Biosynthesis 20% ○

Metabolism 20% ● ⟷ Metabolism 20% ●

Genes in group A have significantly associated to biosynthesis, but not to metabolism.



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

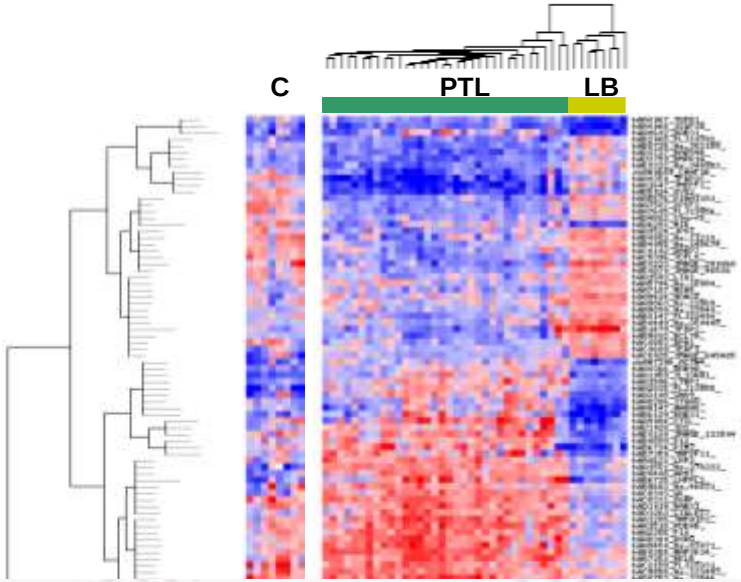
CisRed

Transcription Factor Binding Sites

mSigDB

Reactome

Understanding why genes differ in their expression between two different conditions



Limphomas from mature lymphocytes

hocyte

ed,

enes in the

ed among

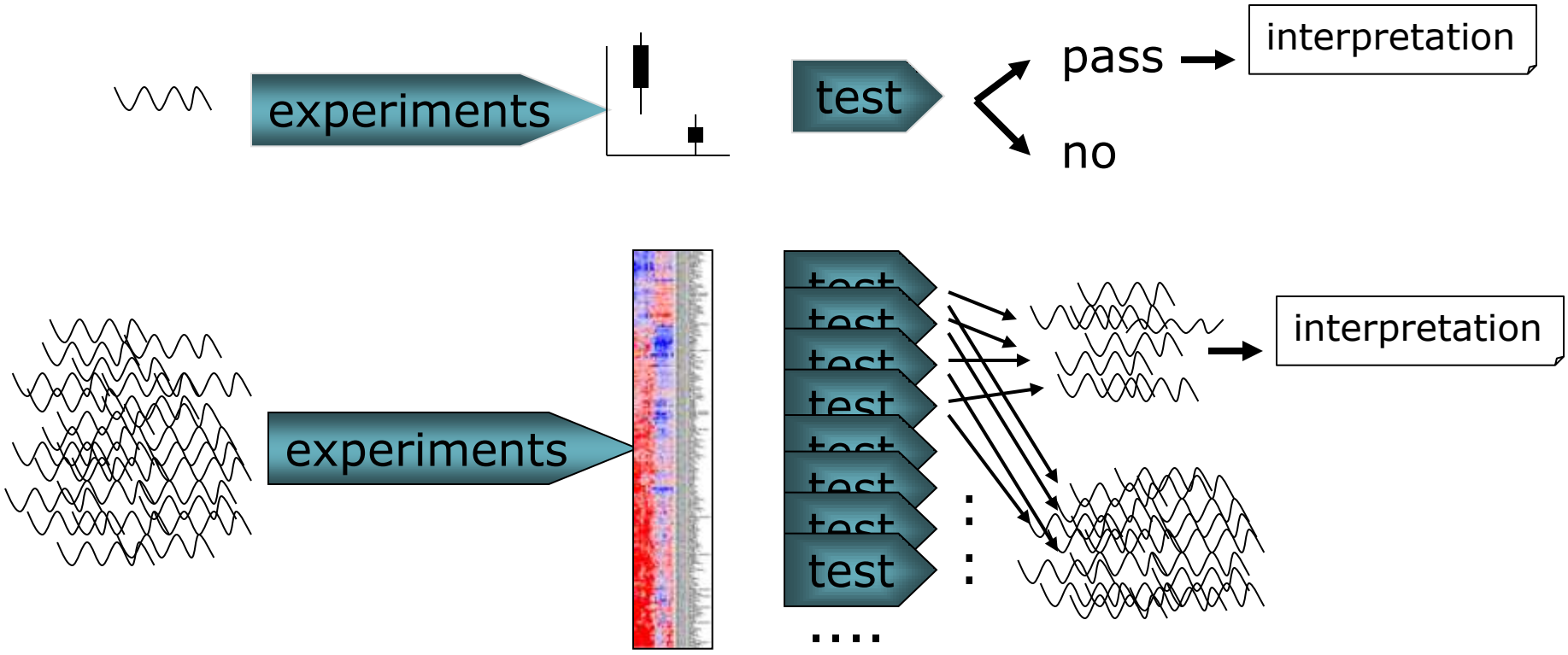
ated to

l in mature

Gene Ontology Term		p-values(*)			
response to external stimulus	36.36%				
	11.65%	0	0.0001	0	0
response to stress	21.59%				
	6.86%	0	0.0002	0.0004	0.0019
signal transduction	39.77%				
	26.05%	0.0065	0.1172	0.1492	0.7191
cell motility	9.09%				
	3.79%	0.0216	0.3464	0.3734	1
resistance to pathogenic bacteria	1.14%				
	0.04%	0.0642	0.8945	0.8857	1
viral replication	1.14%				
	0.15%	0.1529	0.9887	1	1
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

Rodriguez et al., Clinical Cancer Research Vol. 10, 6796-6806, 2004

Functional enrichment tests reproduce pre-genomics paradigms



Context and cooperation between genes is ignored

So, what is wrong with what we are doing?

Our aim:

We seek for the functions activated/deactivated in our experiment.

What we do:

We firstly seek for genes activated/deactivated one at a time (independently)

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

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)

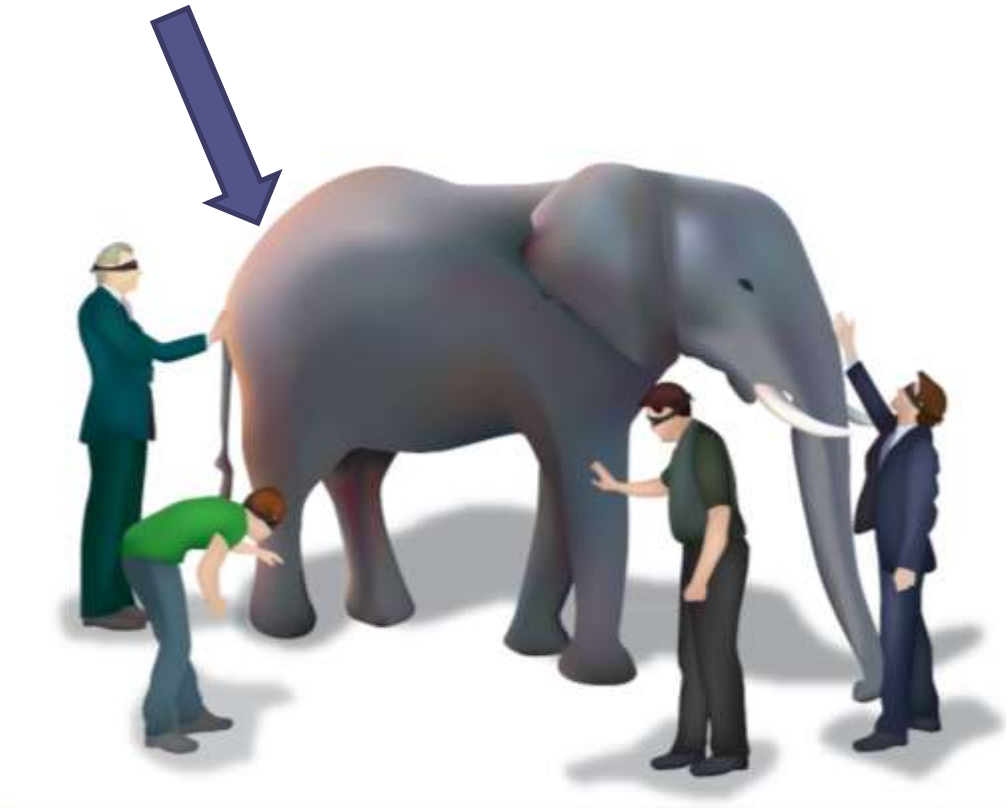
But, we forget about

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.

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

Experiment

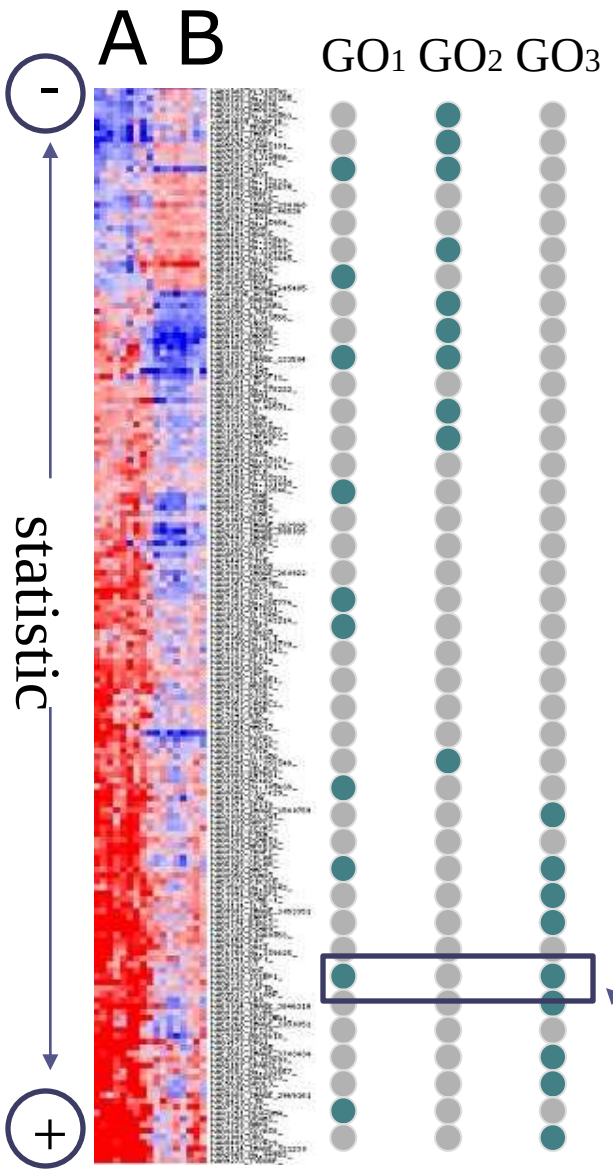


The wise but blindfolded men could not agree on a description of the elephant's phenotype

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?

Cooperative activity of genes (**modules**) can be detected and related to a macroscopic observation



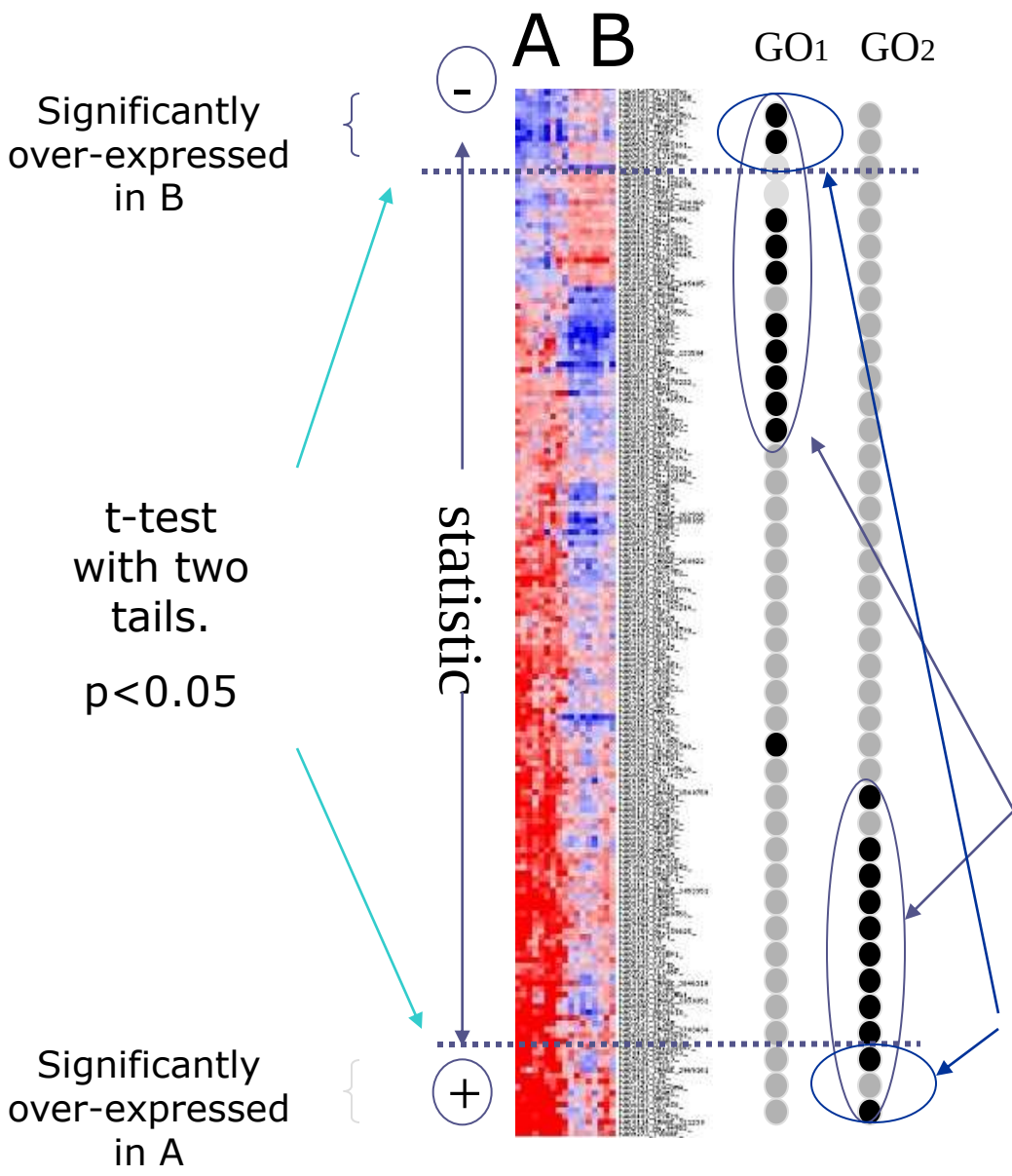
Ranking: A list of genes is ranked by their differential expression between two experimental conditions **A** and **B** (using fold change, a t-test, etc.)

Distribution of GO: Rows GO**1**, GO**2** and GO**3** represent the position of the genes belonging to three different GO terms (**modules**) across the ranking.

The first GO term is completely uncorrelated with the arrangement, while GOs **2** and **3** are clearly associated to high expression in the experimental conditions **B** and **A**, respectively.

Note that genes can be multi-functional

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



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

Modules expressed as blocks in A and B

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

Case study: functional differences in a class comparison experiment

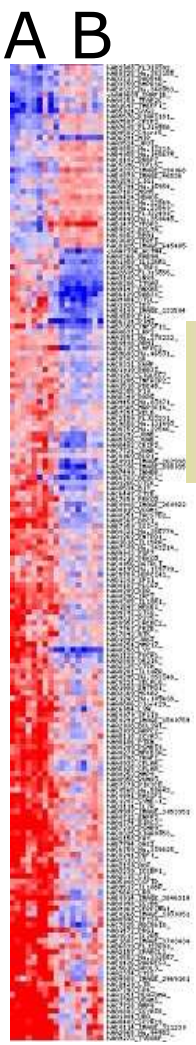
A

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

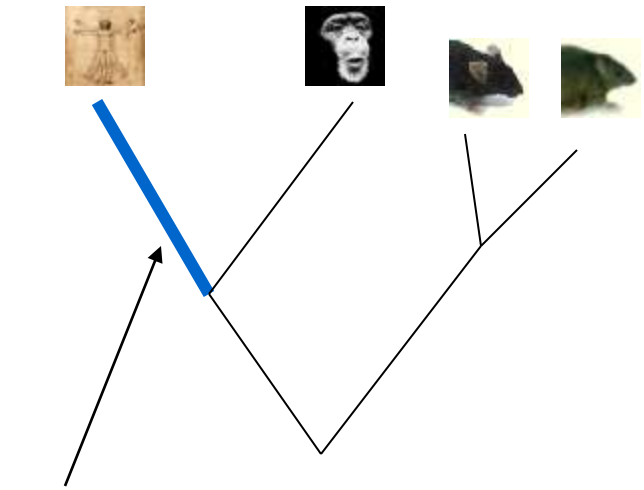
- Survival

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

Al-Shahrour et al., 2007 BMC Bioinformatics

Beyond arrays: evolutionary systems biology



Mutations occur on single genes but natural selection acts on phenotypes by operating on whole sub-cellular systems (represented by GO).

Comparison of the relative rates of synonymous (**Ks**) and non-synonymous (**Ka**) substitutions. The ratio of these values, the (ω=**Ka**/**Ks**) is a widely accepted measure of the selective pressure

20,469 known Ensembl human protein-coding genes from the Ensembl v.30.35h were used

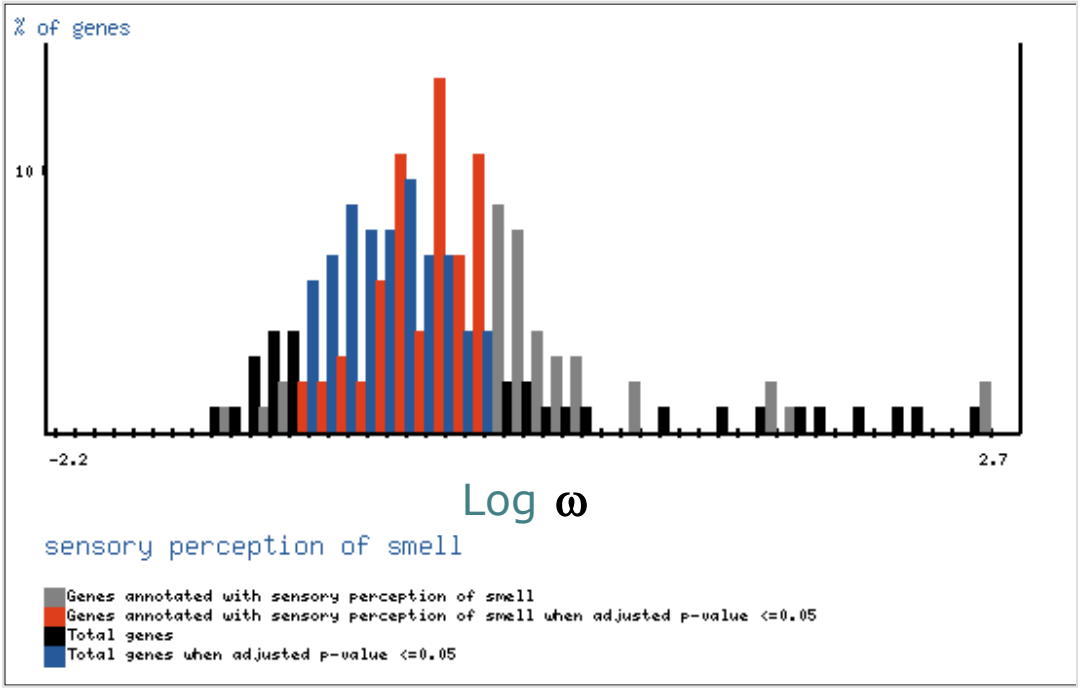
Al-Shahrour et al., 2007 BMC Bioinformatics

Gene-set analysis of GO terms positively selected in humans

GO term	p-value
sensory perception of smell (GO:0007608)	1.3×10^{-5}
sensory perception of chemical stimulus (GO:0007606)	0.0014
G-protein coupled receptor protein signalling pathway (GO:0007186)	0.0095

Gene-set enrichment is applied to the list of human genes ordered according ω values

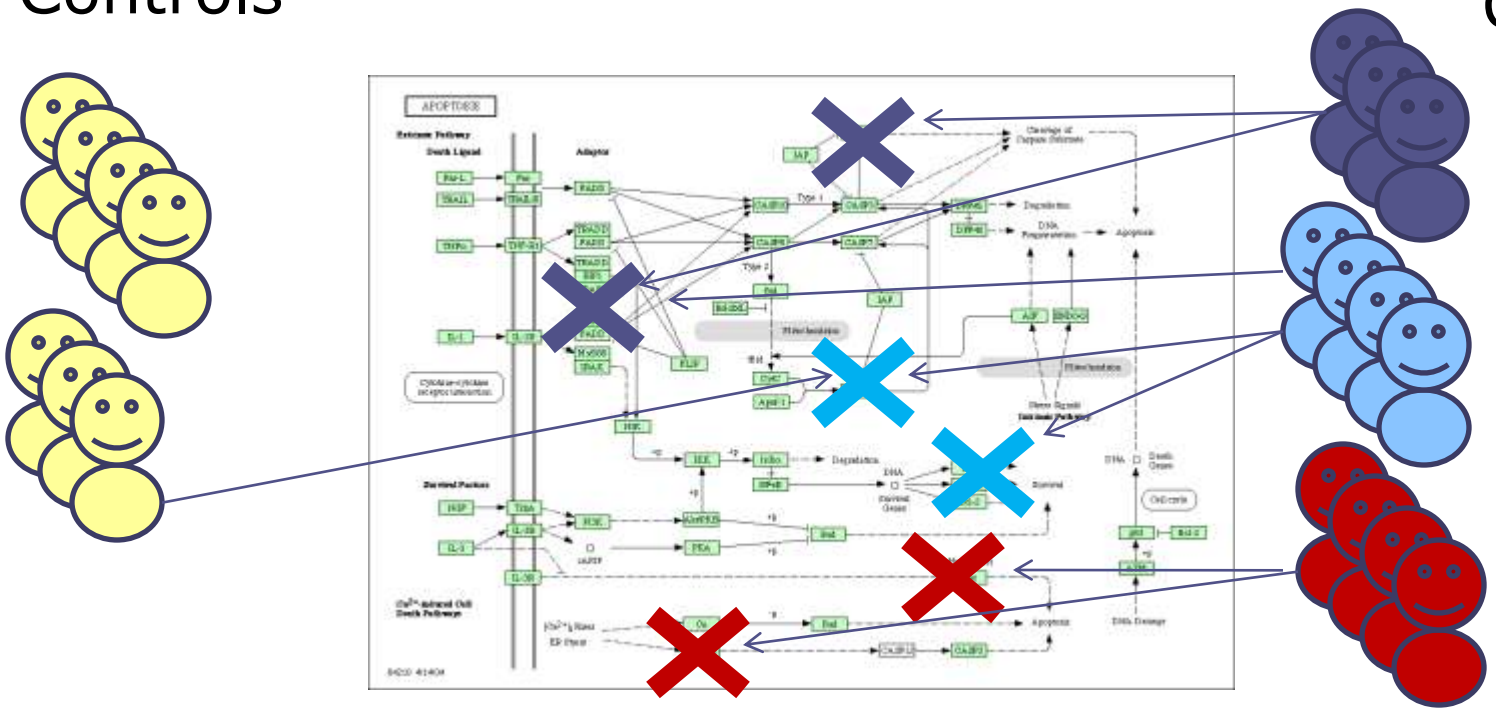
If genes positively selected are firstly tested (one at a time) and then analysed for significant enrichment of GO (functional enrichment), no results are found



Expanding the concept of gene-set analysis to GWAS

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 (unknow at this moment) affected.

Gene-set analysis of GWAS

SNPs are mapped to genes in LD.

Genes are arranged by the highest association value among the corresponding SNPs

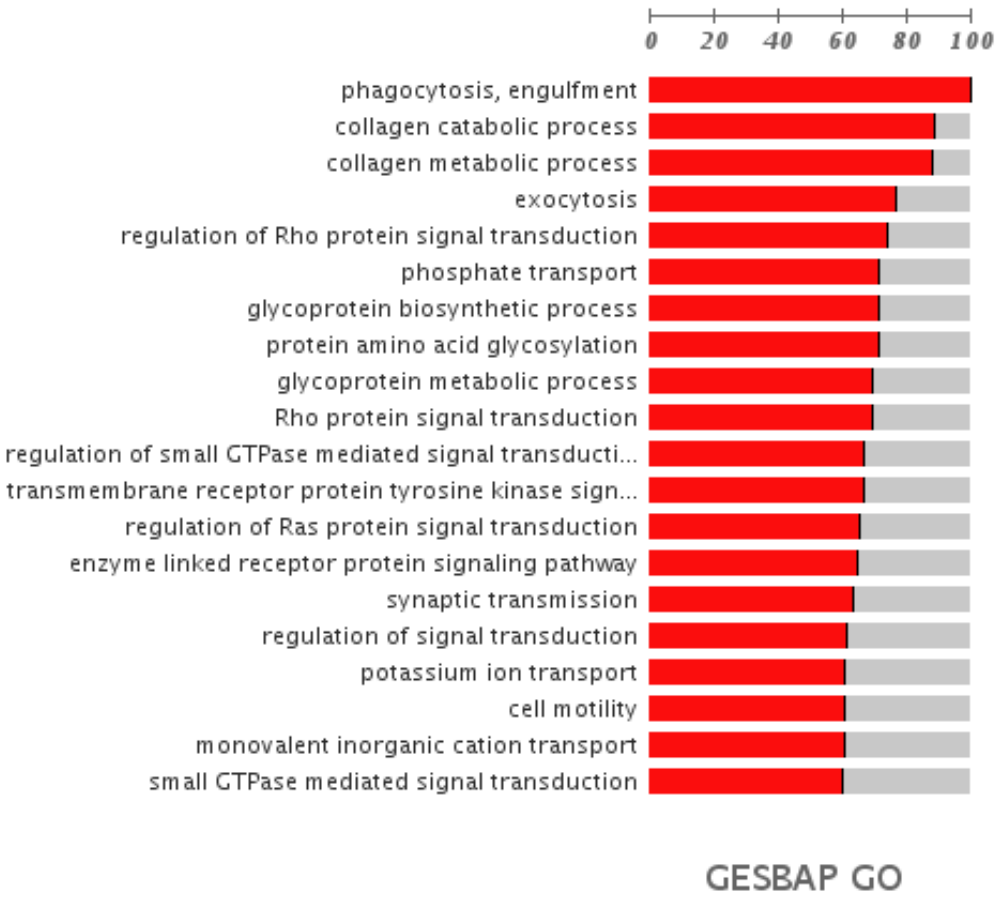
Gene-set analysis (or Pathway-Based analysis, PBA) is conducted on the gene ranked list.

GESBAP can do that

http://bioinfo.cipf.es/gesbap

Medina et al., 2009 NAR, In press

An example of GSA in GWAS



Breast Cancer

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

1145 cases 1142 controls. Affy 500K

Only 4 SNPs were significantly associated, mapping only in one gene: FGFR2

PBA reveals 19 GO categories including *regulation of signal transduction* (FDR-adjusted p-value=4.45×10^{-03}) in which FGFR2 is included.

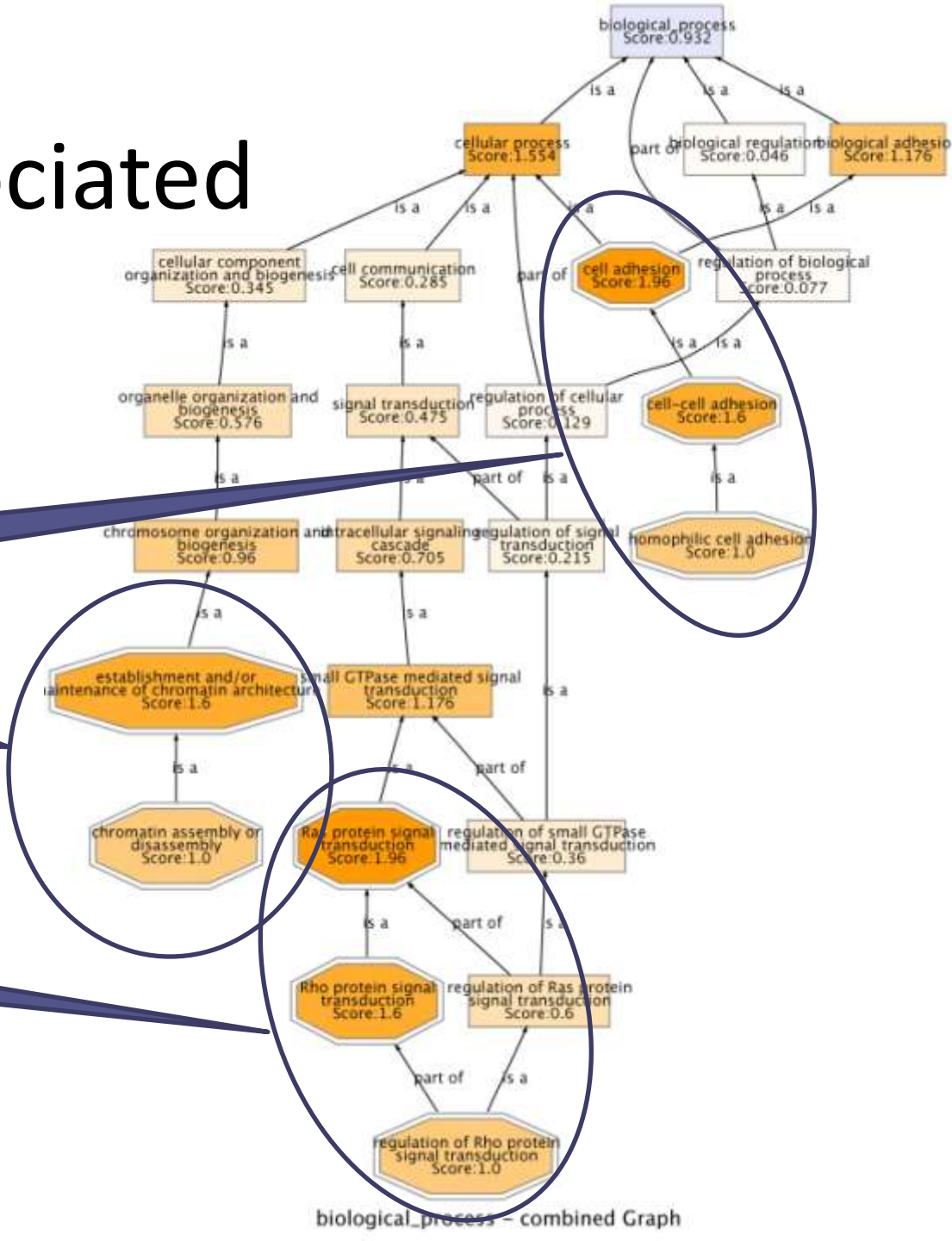
Bonifaci et al., BMC Medical Genomics 2008, ***1:62;*** *Medina et al., 2009 NAR, In press*

GO processes significantly associated to breast cancer

Metastasis

Chromosomal instability

Rho pathway



Using other gene modules: Protein-protein interaction networks

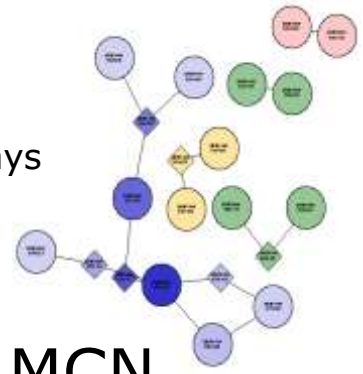
Evaluation of the cooperative behaviour of a list of genes

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

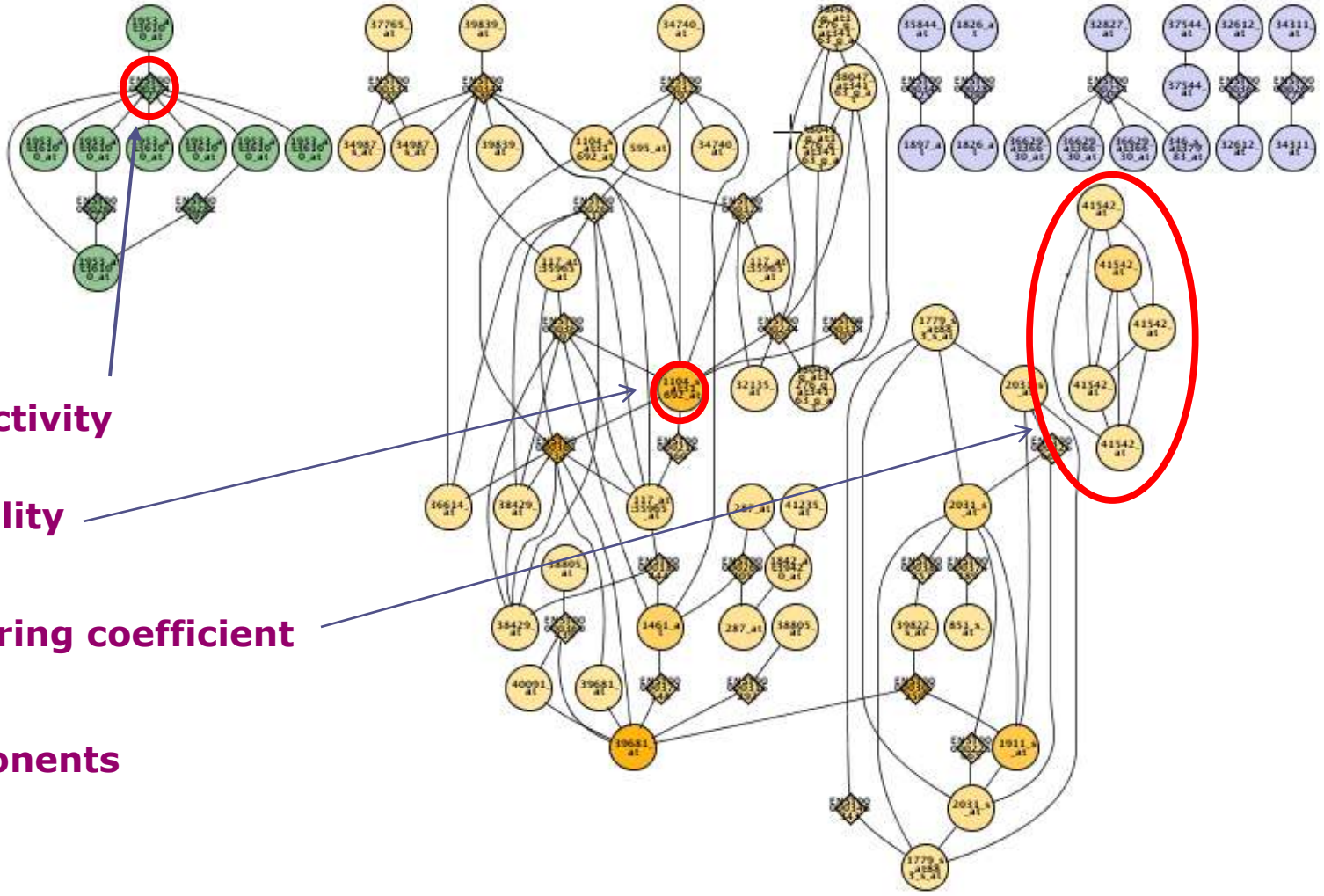


MCN

Prot 1 Nodes included in the list

Prot 2 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

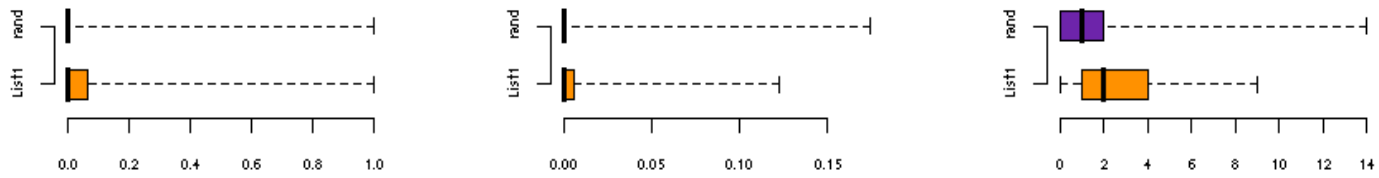
Distribution of the parameterrs' values versus distribution in random MCNs (compared through Kolmogorov-Smirnov tests)

Network parameters

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

Betweenness:
List1 > Random pval=2e-04

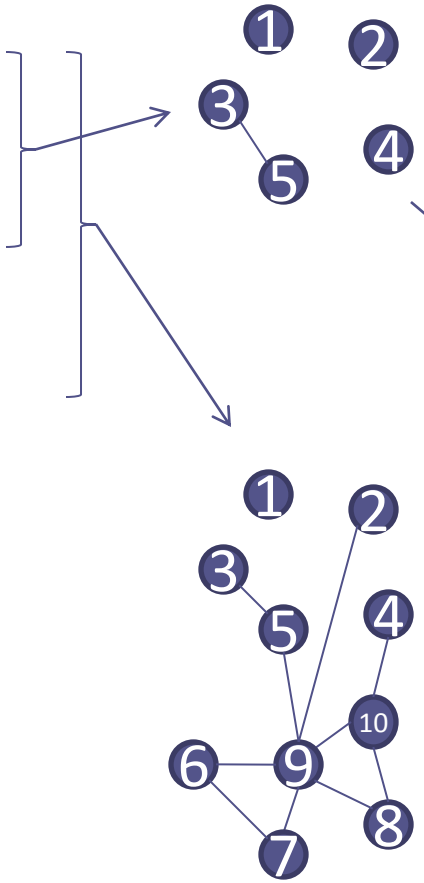
Connections:
List1 > Random pval=0



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**

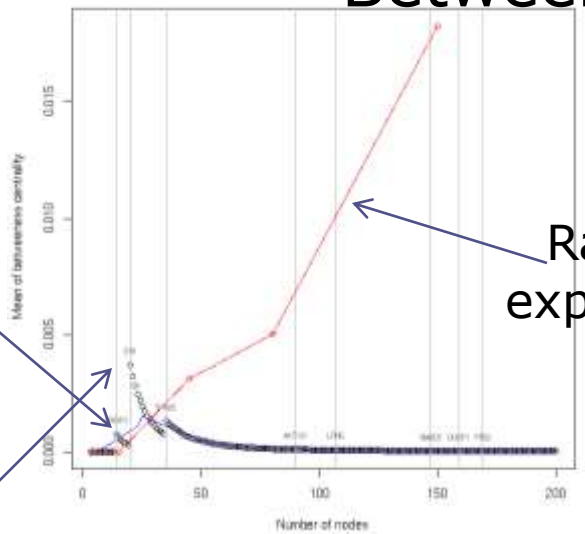
Study of relevant network parameters along the list of genes ranked by the most associated SNP

Gene	p-value
$Gene_1$	p_1
$Gene_2$	p_2
$Gene_3$	p_3
:	:
:	:
$Gene_n$	p_n



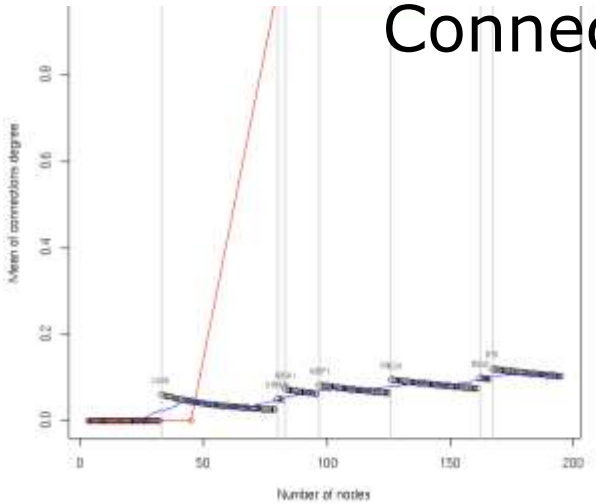
Threshold-free approach

Betweenness



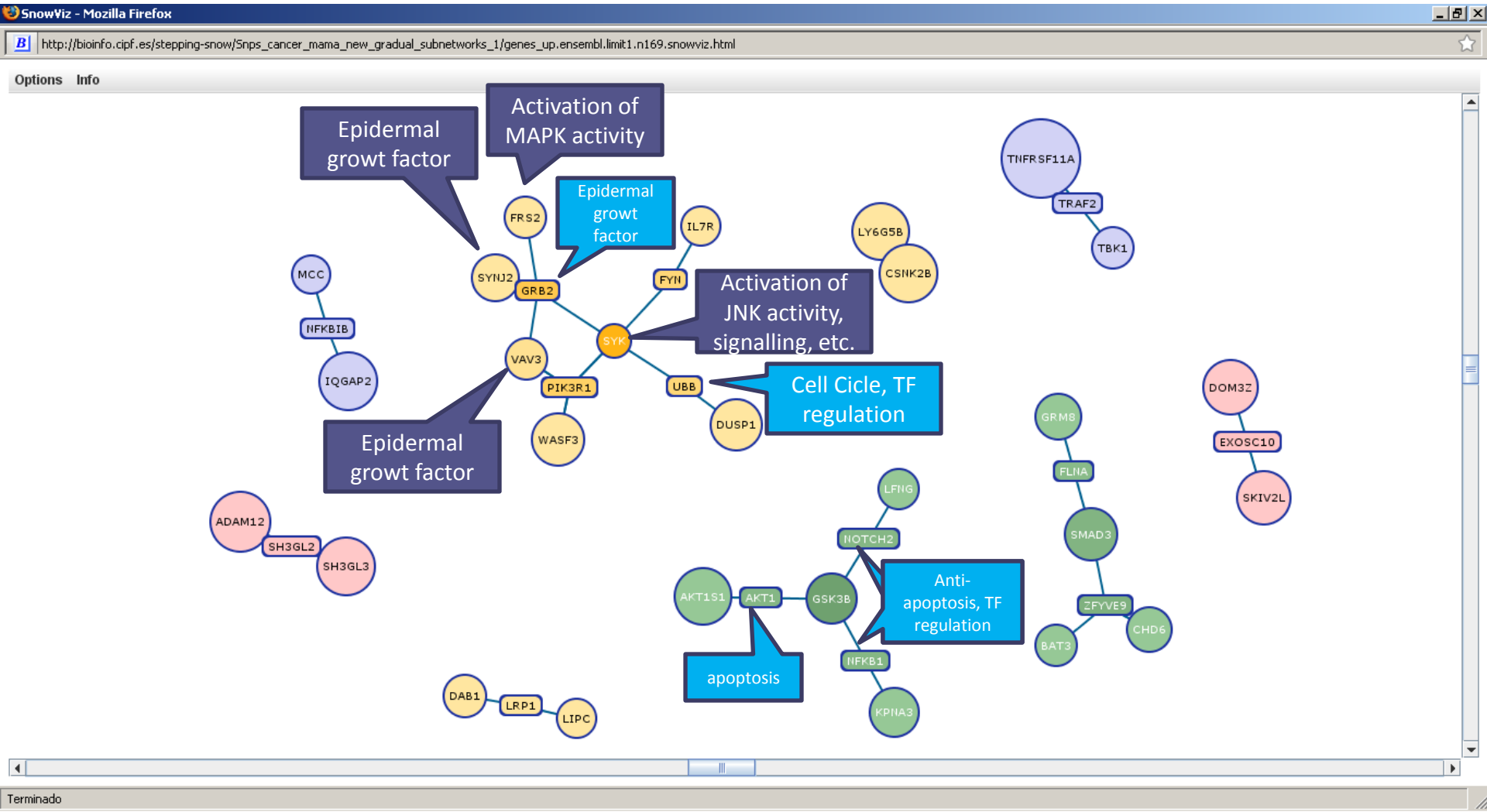
Random expectation

Connections



Significant connections

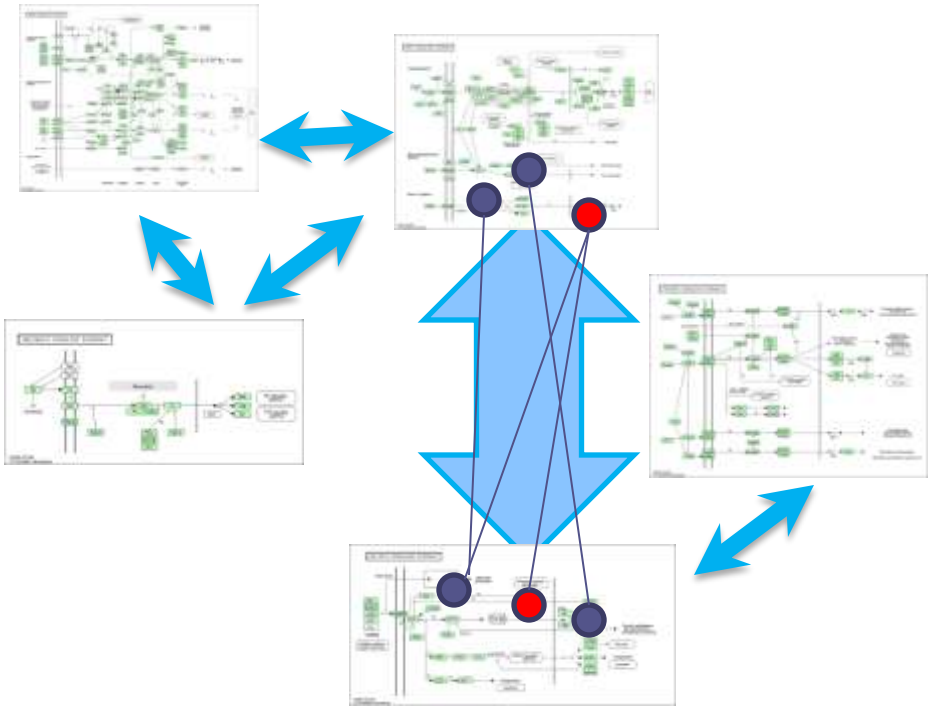
Breast Cancer CGEMS initiative. (Hunter et al. Nat Genet 2007)



http://bioinfo.cipf.es/snow

Minguez et al., 2009 NAR

Towards a higher level of organization: relationships between modules (supermodules)



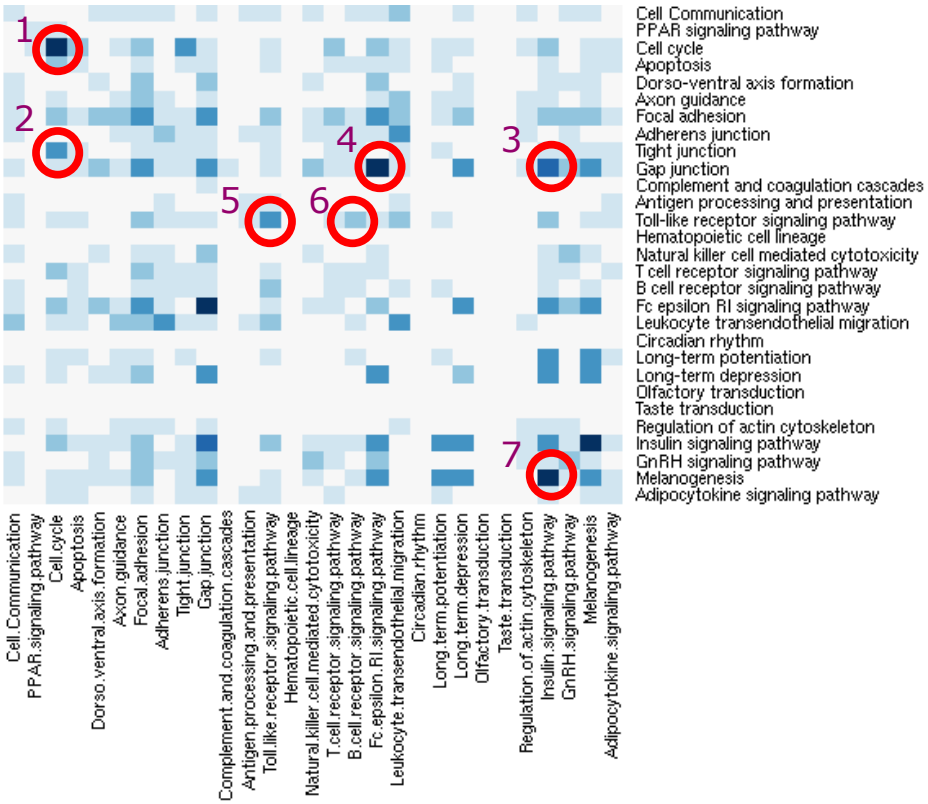
Rationale: inter-pathway PPIs implies physical proximity and suggests functional link

Proteins might be up or down in different experiments affecting the connectivity of the pathways.

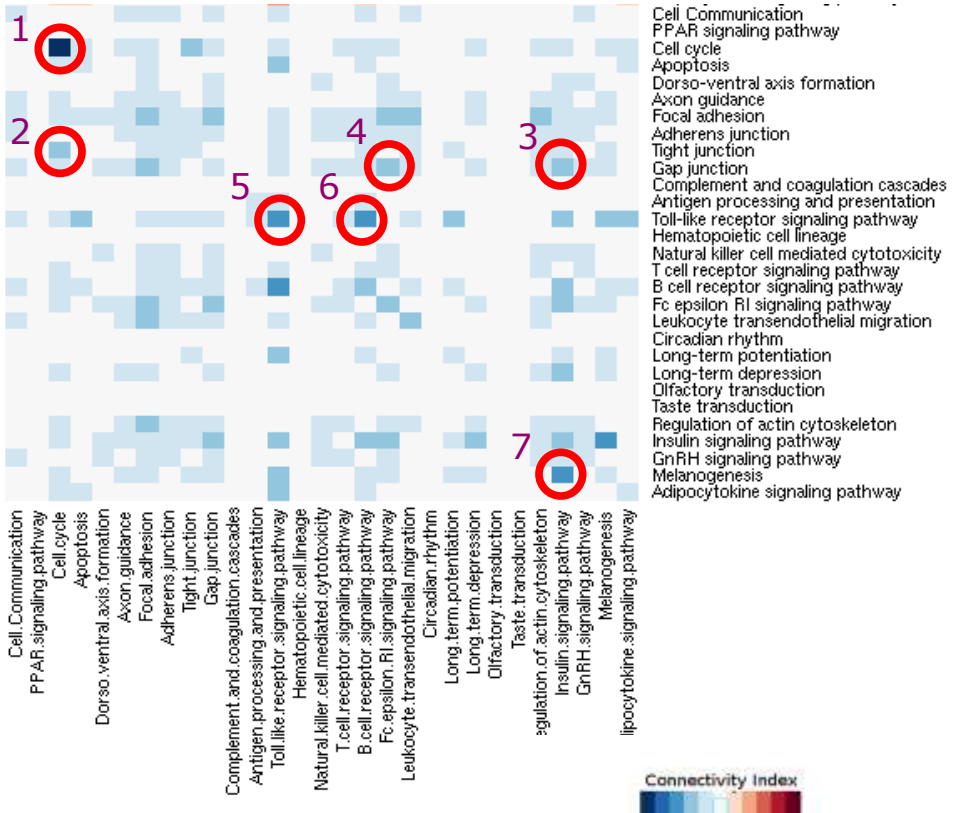
Towards a higher level of organization: relationships between modules (supermodules)

Changes in interactions between pathways in cancer
Cellular Processes (gains in cancer)

Prostate



Mammary Gland



1 - auto-connections in Cell cycle

2 - Cell cycle - Tight junction

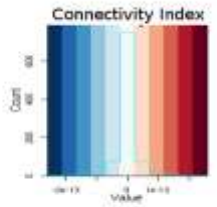
3 - Gap junction - Insulin signaling pathway

4 - Gap junction - Fc epsilon RI signaling pathway

5 - Toll like receptor signaling pathway auto-connections

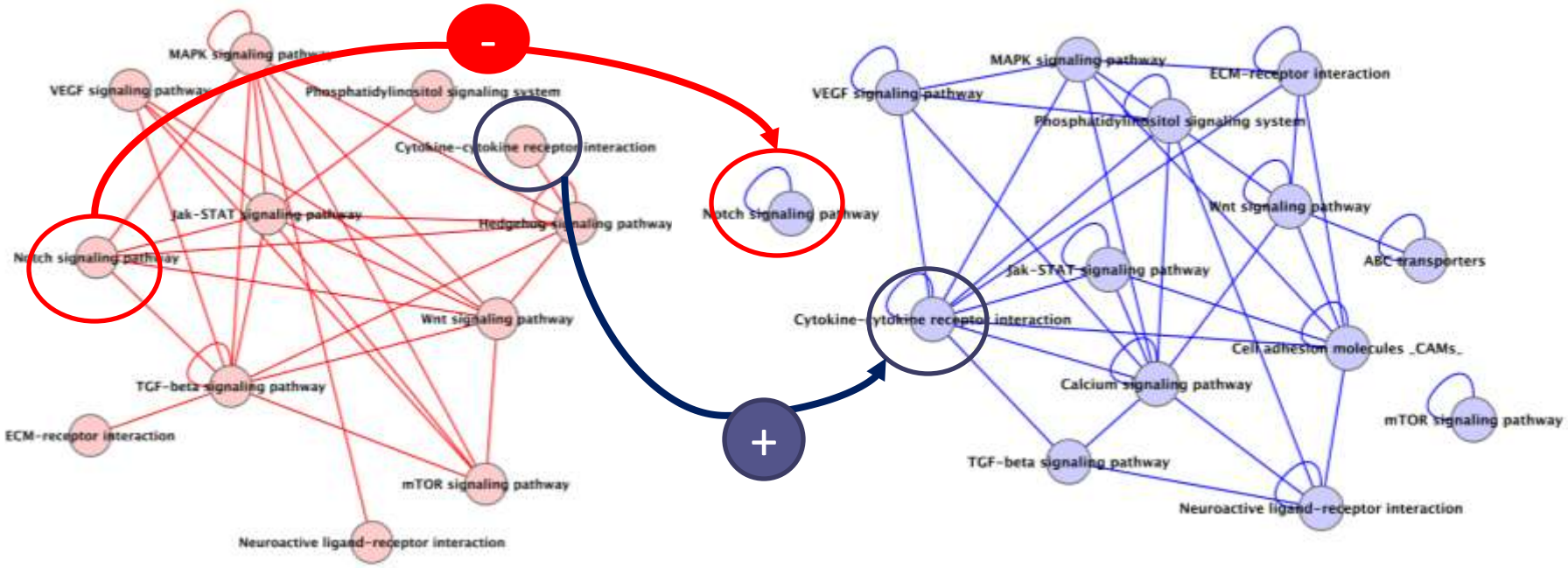
6 - Toll like receptor signaling pathway - B cell receptor signaling

7 - Insulin signaling pathway - Melanogenesis



Redish colors mean physical connections lost in cancer
Bluish colors mean physical connections gained in cancer

Relationships between pathways



Diseased state causes a rewiring both within and among pathways. Based on the presence/absence of transcripts, mapped on the interactome and then in the pathways context

The babelomics suite for functional profiling of genomic experiments

Over 3000 registered users. More than 1000 experiments analysed daily

http://www.babelomics.org

Al-Shahrour et al., 2005, 2006, 2007 , 2008 NAR; 2004, 2005 Bioinformatics, 2007 BMC Bioinformatics;

Biological information from:

- GO
- Interpro motifs
- KEGG pathways
- Biocarta pathways
- Swissprot keywords
- TFBSs (Transfac)
- Regulatory motifs (CisRED)
- miRNAs
- Protein interactions
- Tissues
- Text-mining
- Chromosomal location

For

Human, mouse, rat, chicken, cow, fly, worm, yeast, *A. thaliana* and bacteria

Tests for

- functional enrichment
- gene set enrichment
- network enrichment

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 Baù
Emidio Capriotti
Jose Carbonell
Ana Conesa
Hernán Dopazo
Pablo Escobar
Francisco García
Stefan Goetz
Martina Marbà
Marc Martí
Ignacio Medina
Pablo Minguez
David Montaner
Marina Naval
Javier Santoyo
Patricia Sebastian
François Serra
Sonia Tarazona
Joaquín Tárraga
Adriana Cucchi

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