IX International Course of Massive Data Analysis FOR GENOMICS

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED DE ENFERMEDADES RARAS

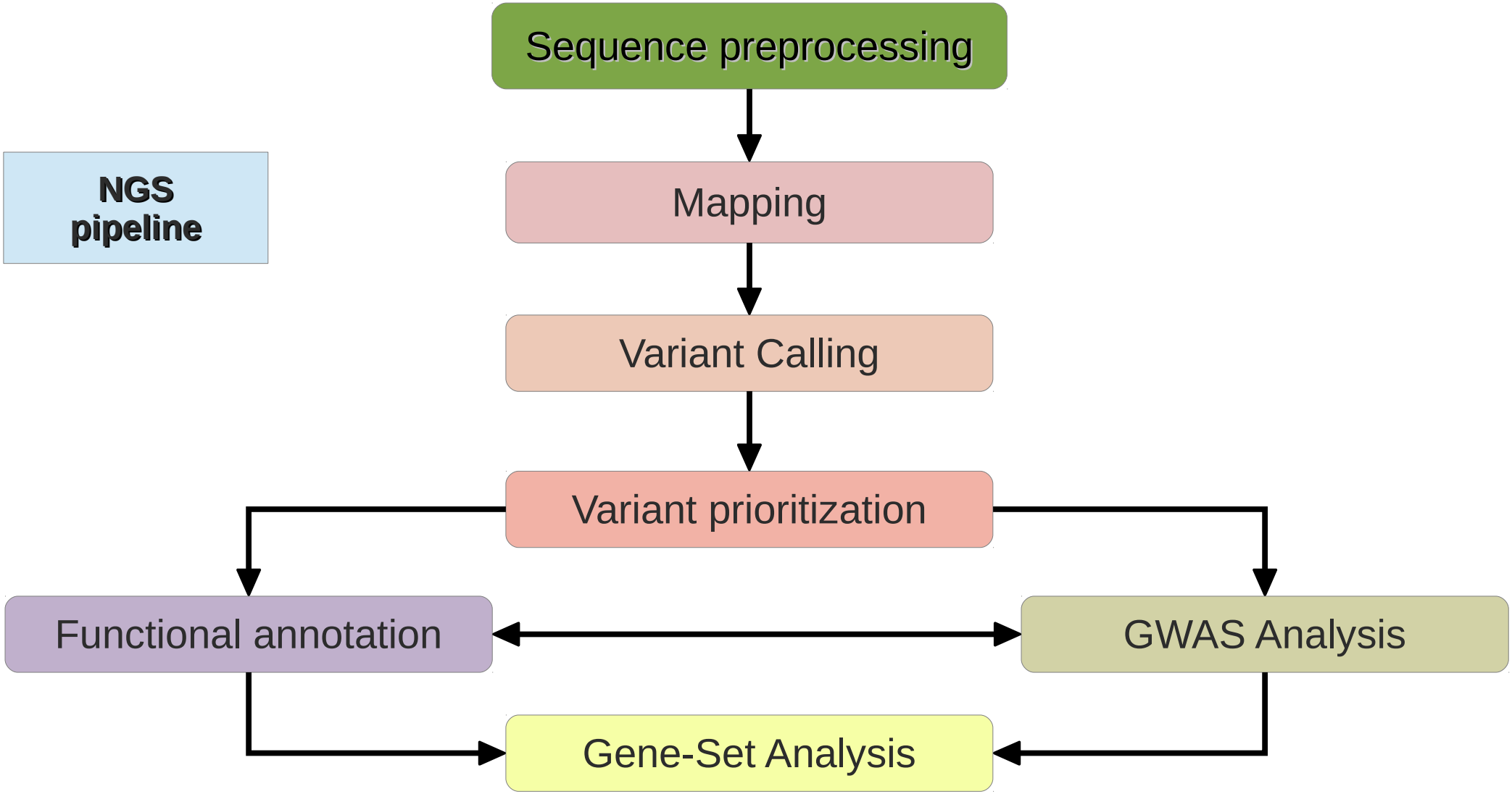
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

David Montaner

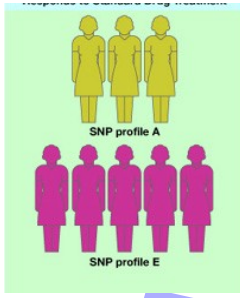
Gene Set Analysis - NGS genomics

Where are we?


NGS pipeline
Sequence preprocessing
Mapping
Variant Calling
Variant prioritization
Functional annotation
GWAS Analysis
Gene-Set Analysis

Genetic Research

Genes in the DNA...



…produces the final phenotype

…code for proteins...

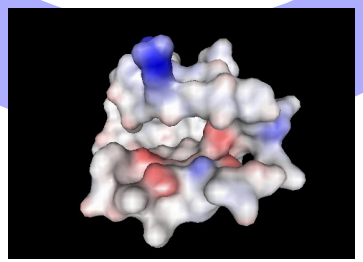
>protein kinase

acctgttgatggcgacagggactgtatgc
tgatctatgctgatgcatgcatgctgacta
ctgatgtgggggctattgacttgatgtcta
tc....

From genotype to phenotype.

…whose structure accounts for function...

…plus the environment...

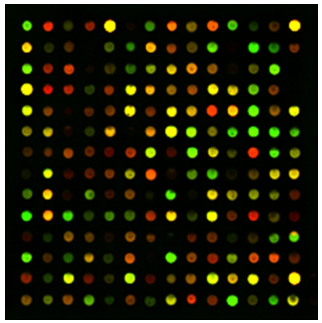


Data is information

High Throughput Technologies

The road of excess leads to the palace of wisdom.

William Blake
Proverbs of Hell
(1790–1793)



Genomic Research

Genes in the DNA...

Next Generation Sequencing
SOLID 6Gbp per round

>protein kinase

acctgttgatggcgacagggactgtatgct
gatctatgctgatgcatgcatgctgactactg
atgtgggggctattgacttgatgtctatc....

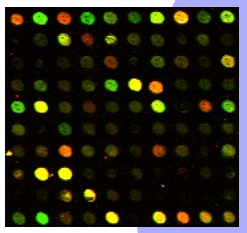
…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype...

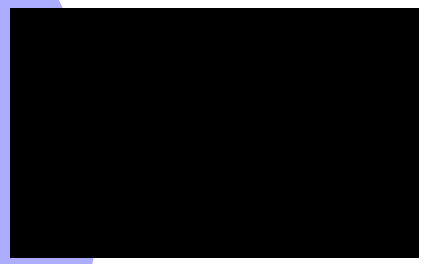
…when expressed in the proper moment and place...

A typical tissue is expressing among 5000 and 10000 genes

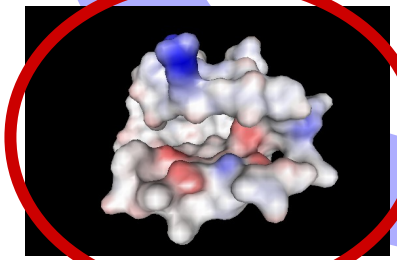


From genotype to phenotype.

(in the functional post-genomic scenario)



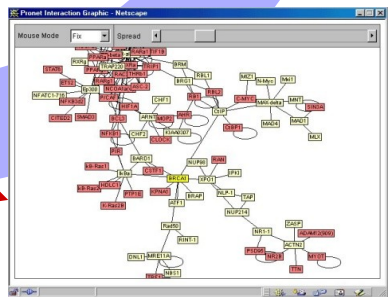
…code for proteins...



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

100K-500K proteins

…whose structures account for function...



…in cooperation with other proteins…

Each protein has an average of 8 interactions

Genomic Research

Genes in the DNA...

Next Generation Sequencing
SOLID 6Gbp per round

>protein kinase

acctgttgatggcgacagggactgtatgct
gatctatgctgatgcatgcatgctgactactg
atgtgggggctattgacttgatgtctatc....

…which can be different because of the variability.

10 million SNPs

…whose final effect configures the phenotype...

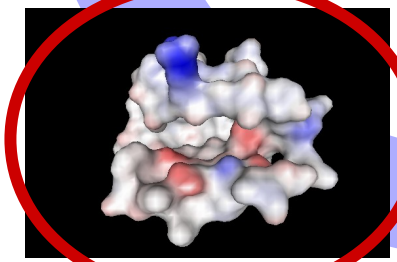
…when expressed in the proper moment and place...

A typical tissue
expressing amo
5000 and 10000
genes

DATA not equal to INFORMATION

(in the functional post-genomic scenario)

…code for proteins...

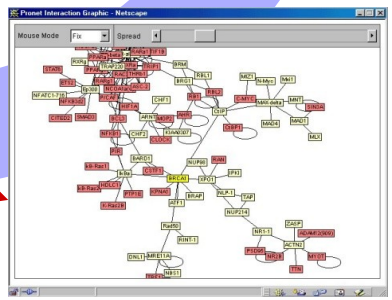


That undergo post-translational modifications, somatic recombination...

100K-500K proteins

…whose structures account for function...

…conforming complex interaction networks...



…in cooperation with other proteins…

Each protein has an average of 8 interactions

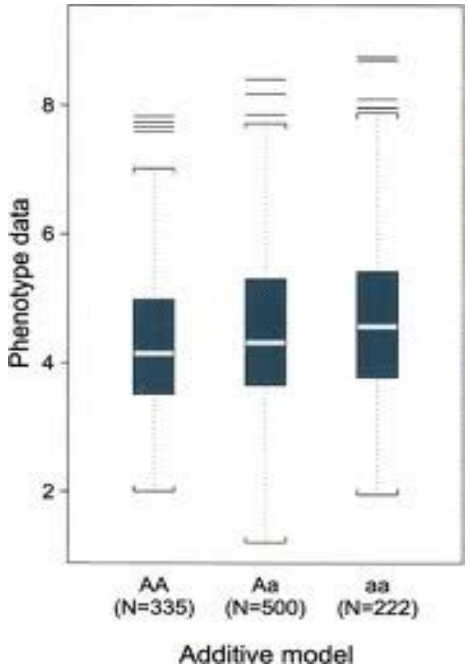
Experimental Data Analysis Results

SNP 2	SNP 1: A	SNP 1: a	
B	f_{AB}	f_{aB}	f_B
b	f_{Ab}	f_{ab}	f_b
	f_A	f_a	

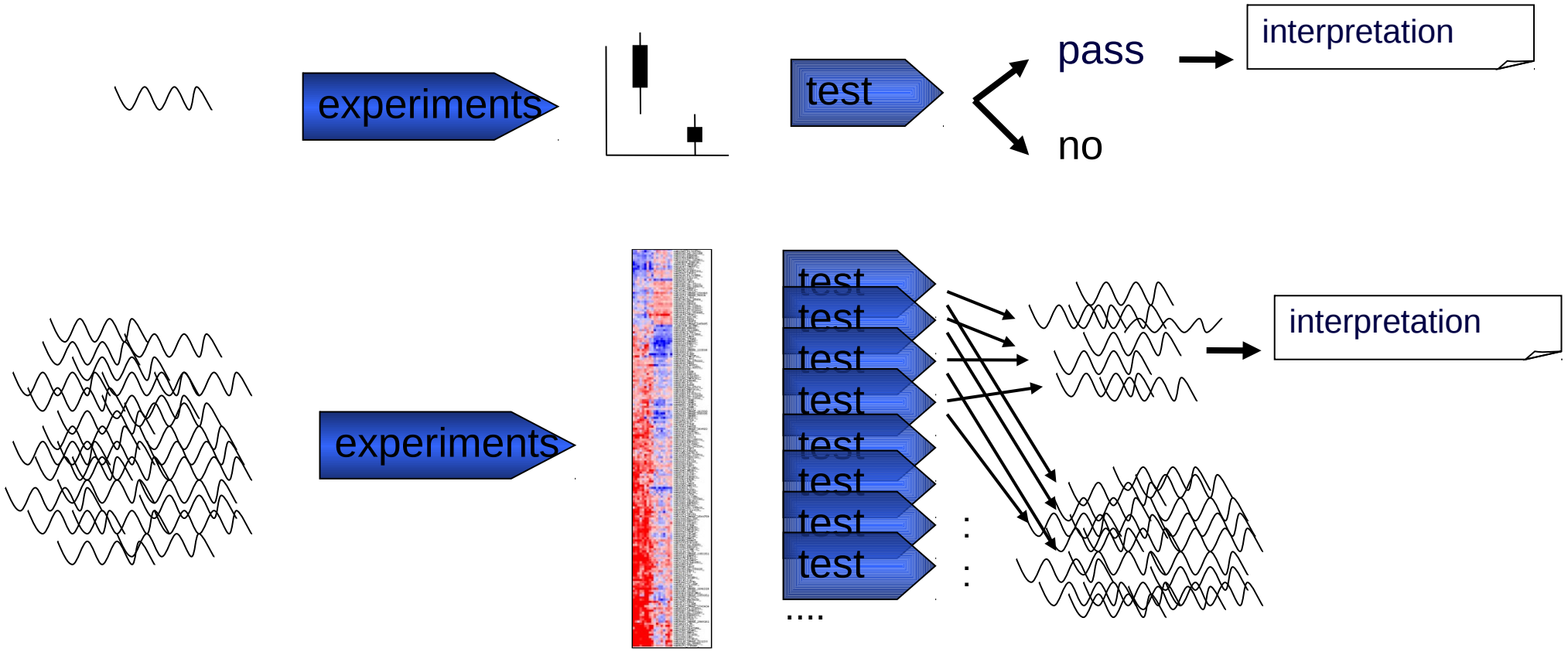
Phenotype	SNP 1: A	SNP 1: a	
B	f_{AB}	f_{aB}	f_B
b	f_{Ab}	f_{ab}	f_b
	f_A	f_a	

- Linkage Disequilibrium (LD)
- Association Studies
- ...

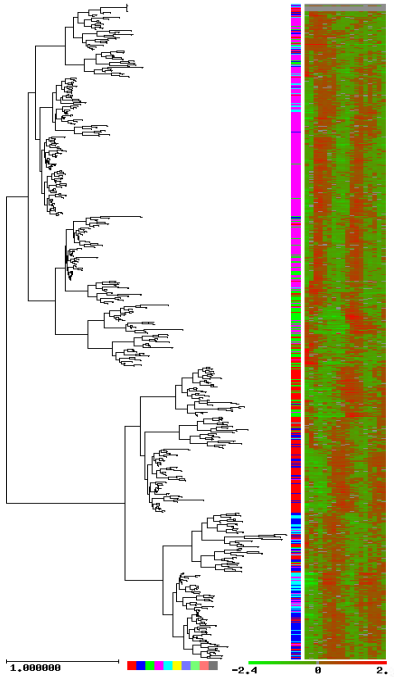
- Statistics
- P-values
- Posterior probabilities

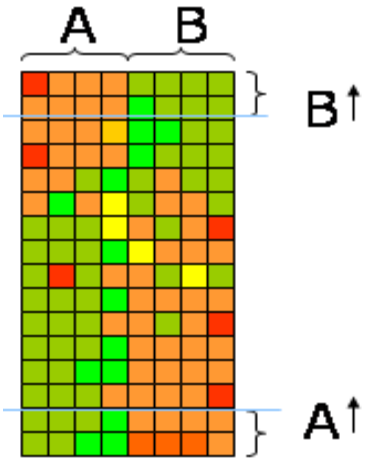


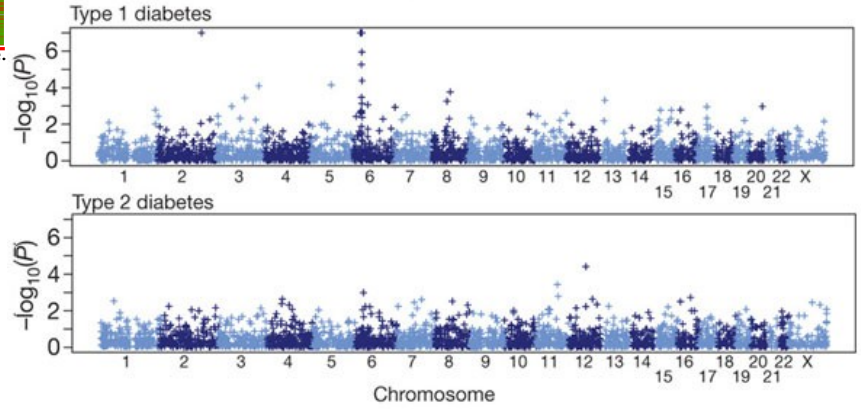
Hypothesis Testing (and Interpretation)

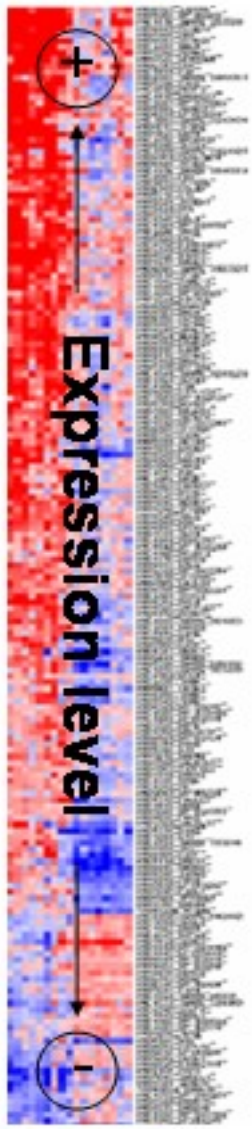


Genome-scale experiment output

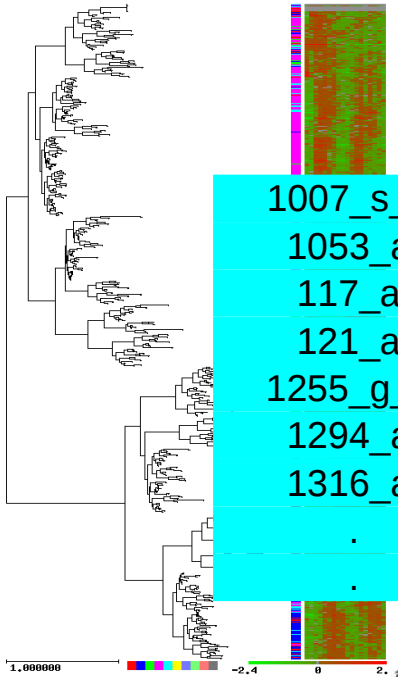




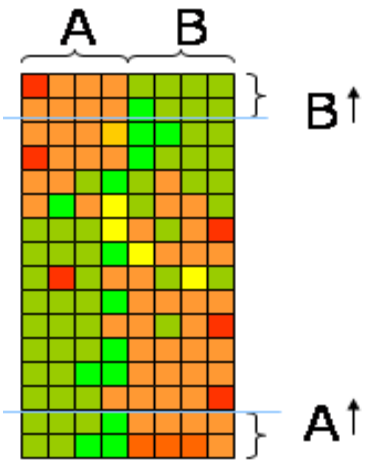


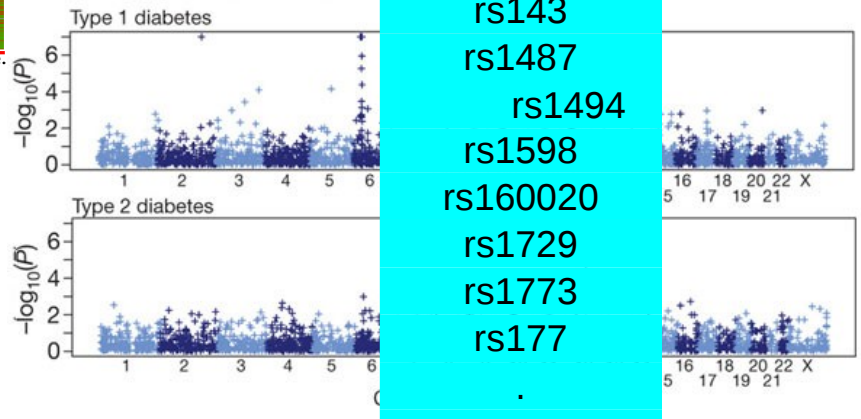


Genome-scale experiment output



1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.





rs1320
rs1405t
rs1431
rs143
rs1487
rs1494
rs1598
rs160020
rs1729
rs1773
rs177
.
.

1007_s_at	12.4
1053_at	11.5
117_at	10.3
121_at	10.2
1255_g_at	9.9
1294_at	9.3
1316_at	8.2
1320_at	8.1
1405_i_at	7.7
1431_at	7.4
1438_at	6.5
1487_at	6.2
1494_f_at	5.9
1598_g_at	5.8
160020_at	4.8
1729_at	4.7
.	.
.	.

Genome-scale experiment output

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

rs1320
rs1405t
rs1431
rs143
rs1487
rs1494
rs1598
rs160020
rs1729
rs1773
rs177
.
.

1007_s_at	**12.4**
1053_at	**11.5**
117_at	**10.3**
121_at	**10.2**
1255_g_at	**9.9**
1294_at	**9.3**
1316_at	**8.2**
1320_at	**8.1**
1405_i_at	**7.7**
1431_at	**7.4**
1438_at	**6.5**
1487_at	**6.2**
1494_f_at	**5.9**
1598_g_at	**5.8**
160020_at	**4.8**
1729_at	**4.7**
.	.
.	.

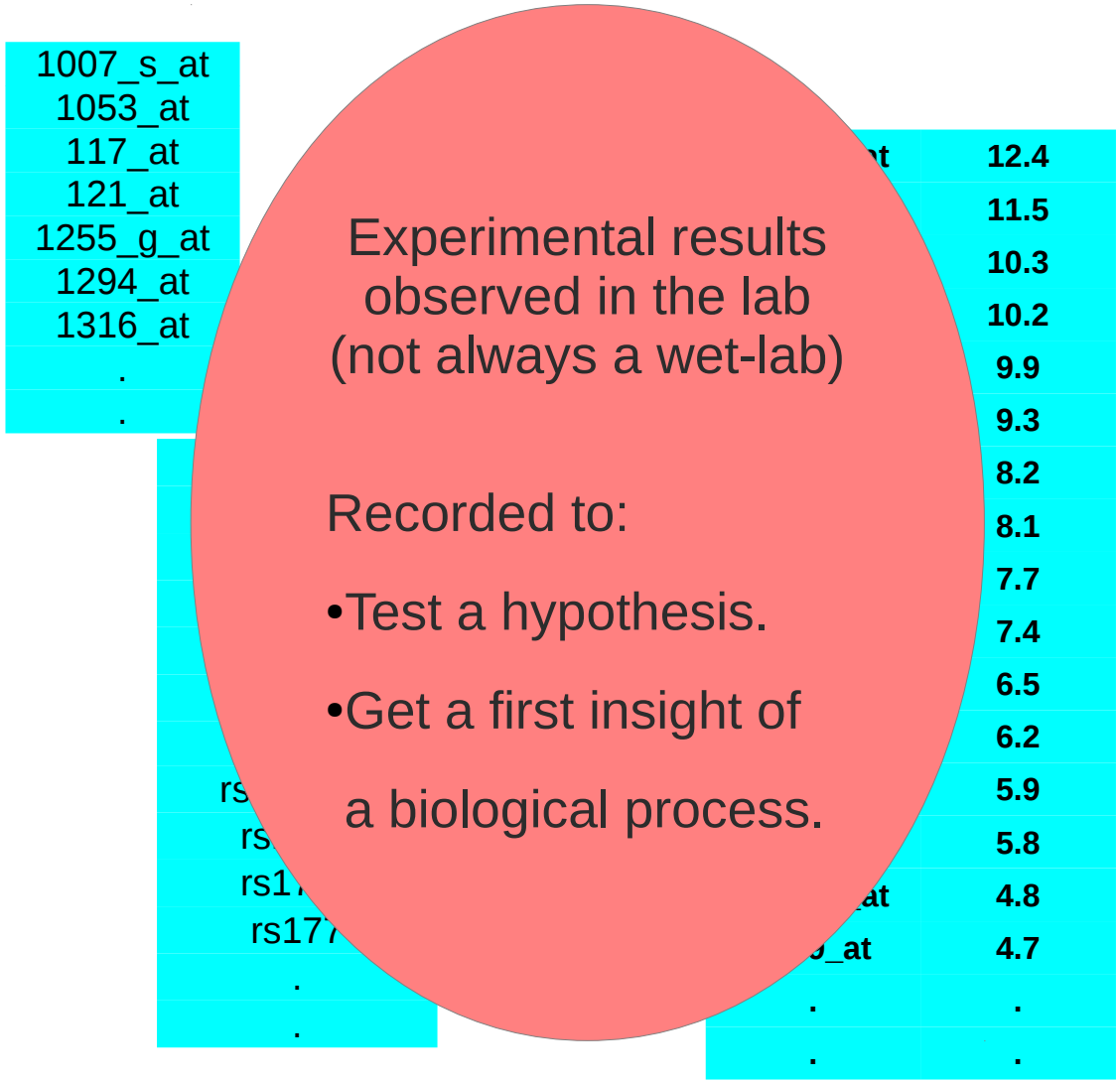
Functional interpretation

1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

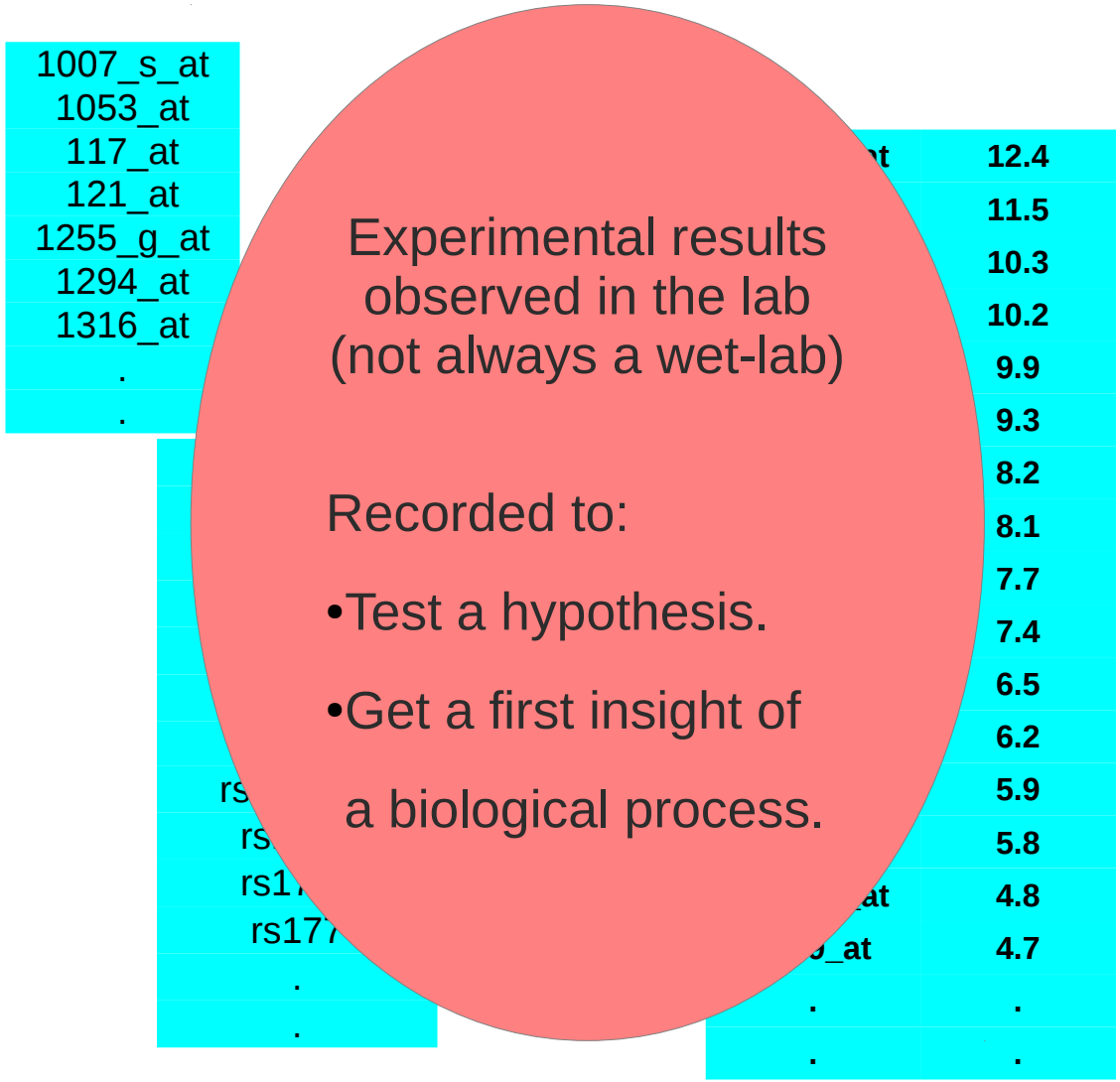
rs1320
rs1405t
rs1431
rs143
rs1487
rs1494
rs1598
rs160020
rs1729
rs1773
rs177
.
.

1007_s_at	**12.4**
1053_at	**11.5**
117_at	**10.3**
121_at	**10.2**
1255_g_at	**9.9**
1294_at	**9.3**
1316_at	**8.2**
1320_at	**8.1**
1405_i_at	**7.7**
1431_at	**7.4**
1438_at	**6.5**
1487_at	**6.2**
1494_f_at	**5.9**
1598_g_at	**5.8**
160020_at	**4.8**
1729_at	**4.7**
.	**.**
.	**.**

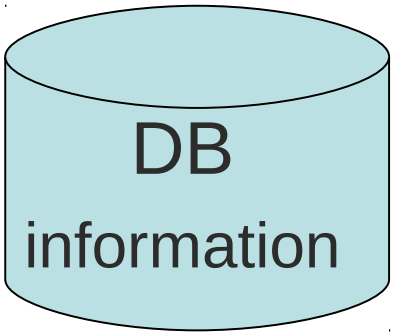
Functional interpretation



Functional interpretation

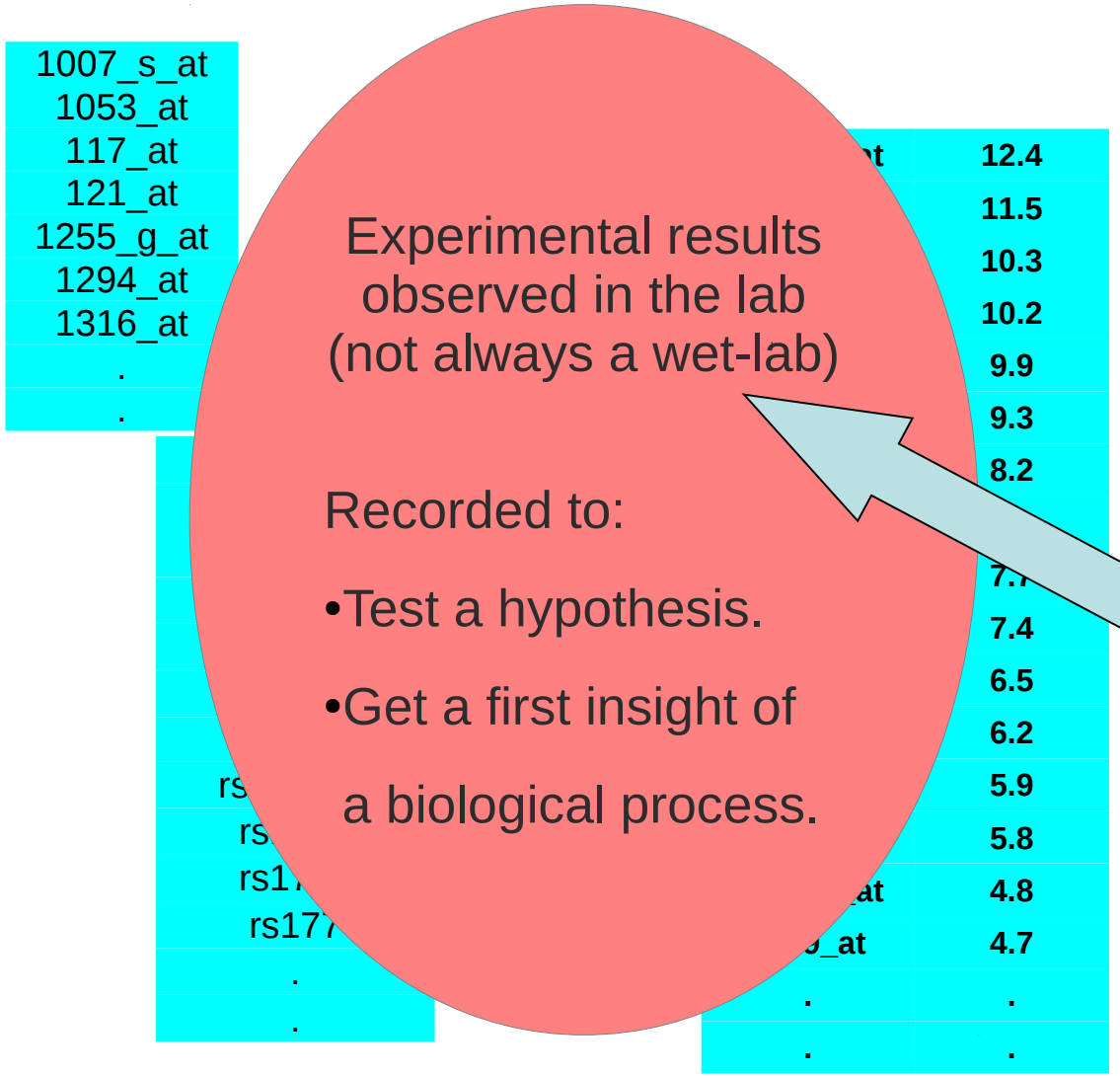


To “interpret” experimental results is to use **current knowledge** to rearrange them in a meaningful way.

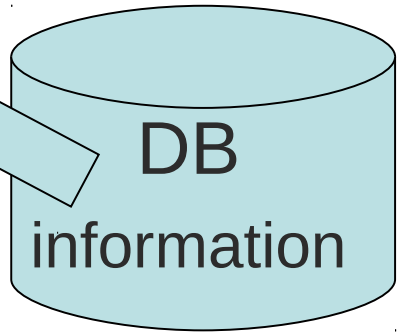


Already tested and stored

Functional interpretation



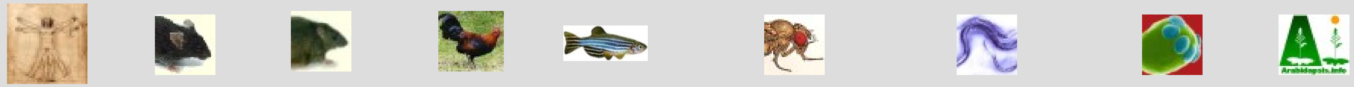
To “interpret” experimental results is to use **current knowledge** to rearrange them in a meaningful way.



Already tested and stored

Biological Databases

Some of the biological databases contains ***Functional Information*** of the genes and sequences



UniProt/Swiss-Prot **EntrezGene** **HGNC symbol** **PDB**

UniProtKB/TrEMBL **Affymetrix** **EMBL acc** **Protein Id**

Ensembl IDs **Agilent** **RefSeq** **IPI….**

Gene IDs

Functional databases

KEGG pathways

Reactome

Gene Ontology

ss Molecular Function Ce

Regulatory elements

miRNA

CisRed

Transcription Factor

Binding Sites

Keywords Swissprot

Biocarta pathwa

Gene Expression in tissues

InterPro Moti

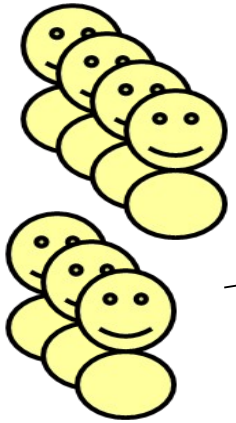
Bioentities from literature:

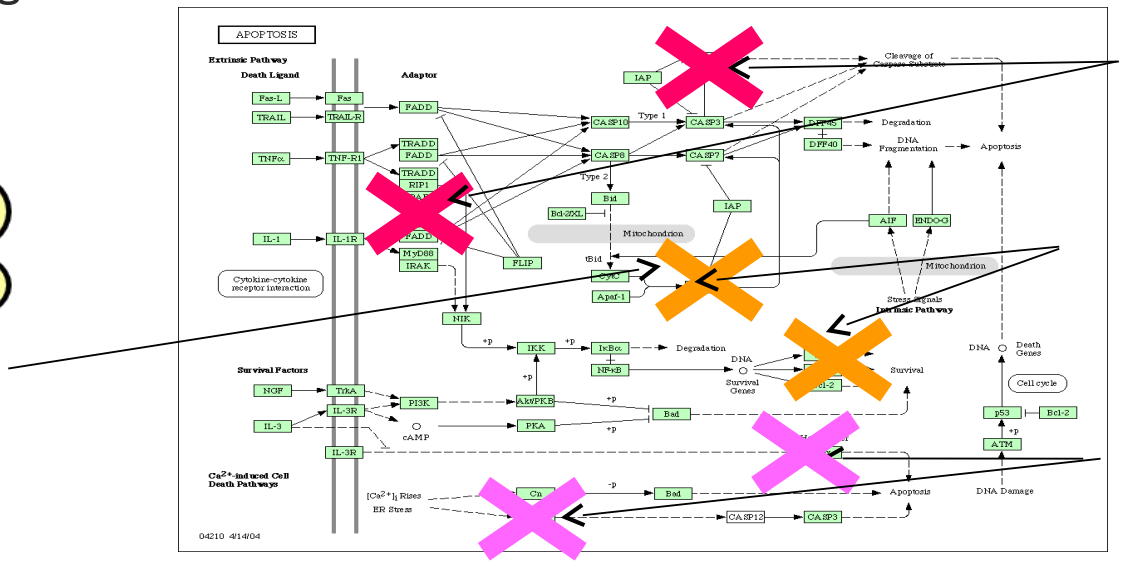
es terms **Chemical terms**

David Montaner

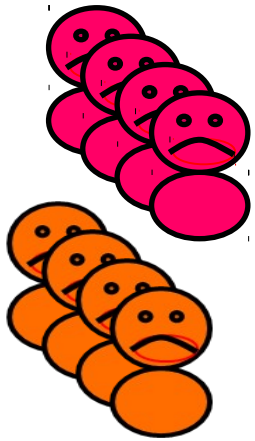
Introduction
Drawbacks

Controls



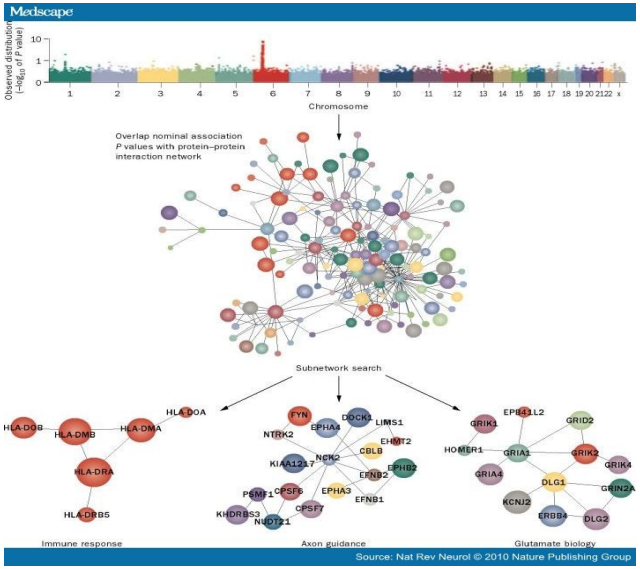


Cases



Variants are considered independently. However, complex phenotypes are expected to be induced by different genes in the same functional module. A different strategy must be taken: Methodologies based on ***Gene-Set Analysis or networks*** permit the study of functional modules (group of genes that cooperate to carry out a biological function)

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.

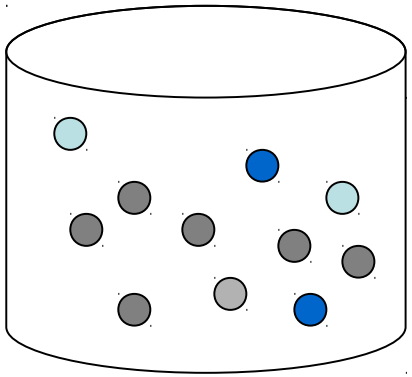


Ignacio Medina
Cristina Y. González

Genomic-Wide Association Studies

FatiGO test (Fisher)

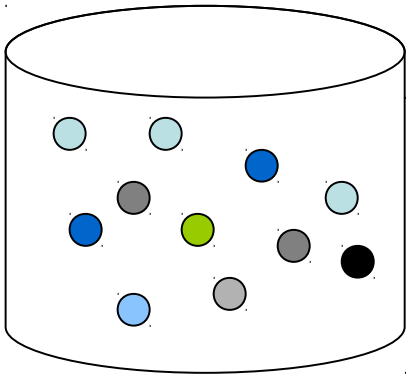
One Gene List (A)



Biosynthesis 60%

Sporulation 20%

The other list (B)

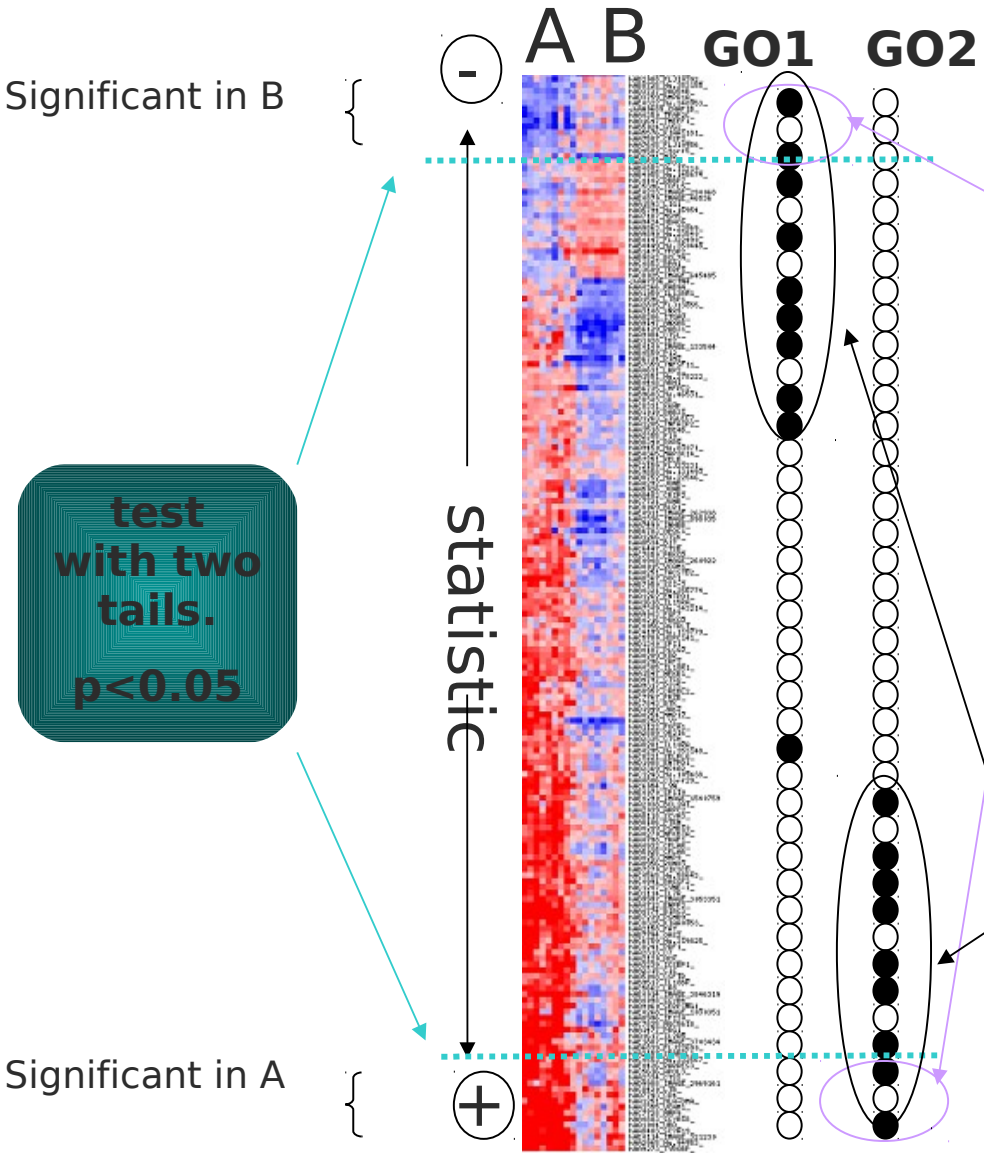


Biosynthesis 20%

Sporulation 20%

	A	B
Biosynthesis	**6**	**2**
No biosynthesis	**4**	**8**

FatiGO approach may not be very powerful



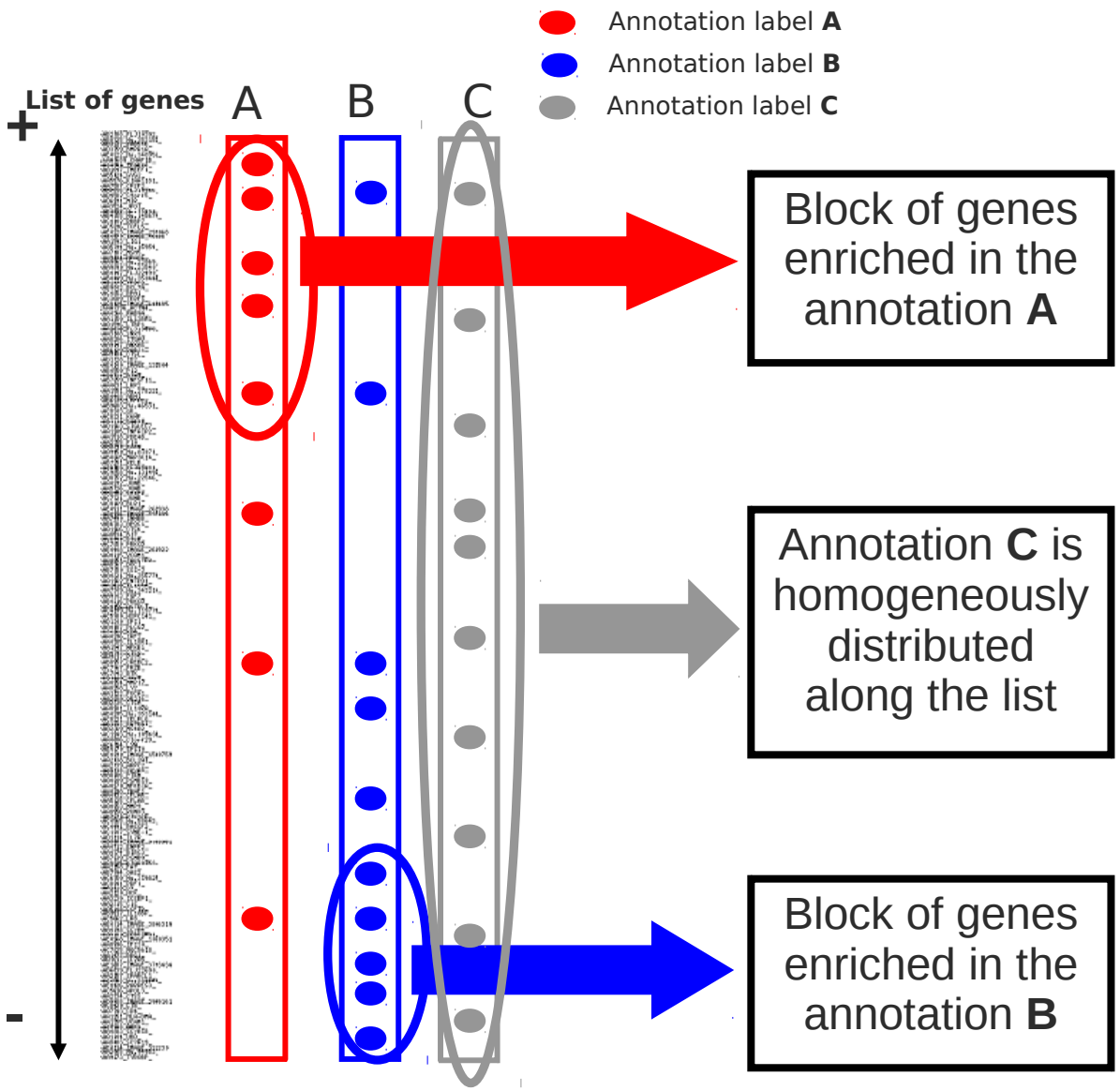
Very few genes selected to arrive to a significant conclusion on GO1 and GO2

If a threshold based on the experimental values is applied, and the resulting selection of genes compared for enrichment of a functional term, this might not be found

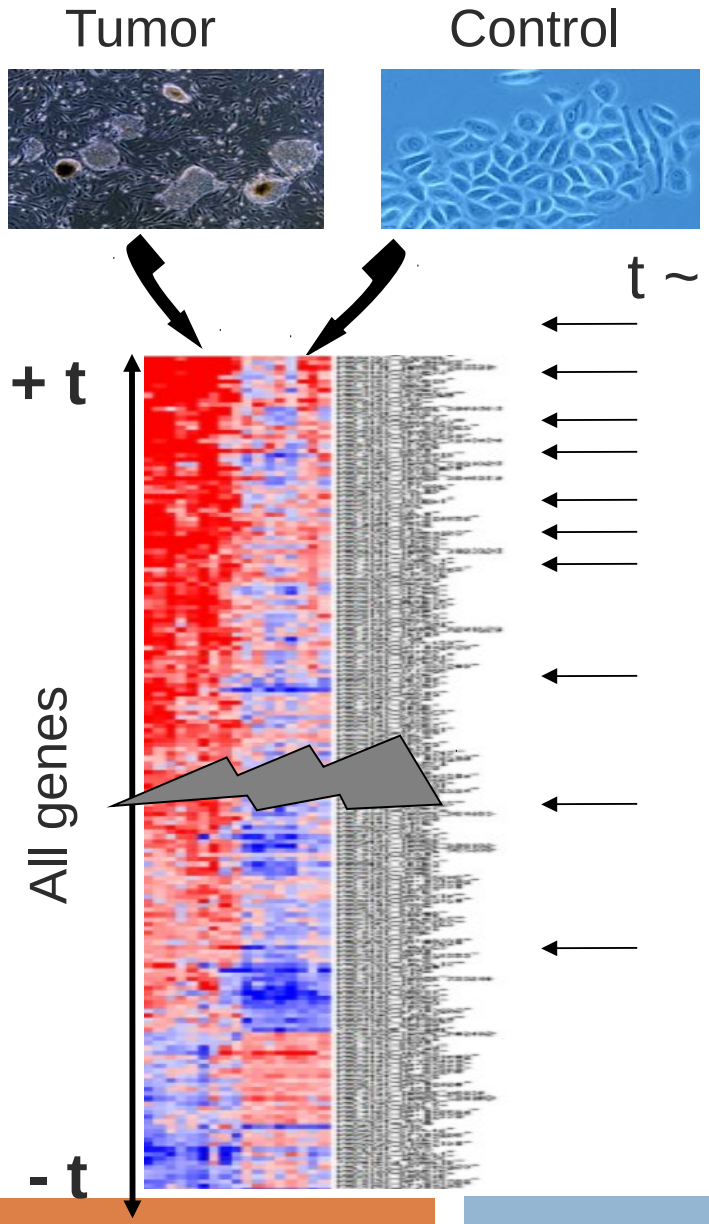
Functional Classes expressed as blocks in A and B

Testing along an ordered list

- Index ranking genes according to some biological aspect under study.
- Database that stores gene class membership information.
- **FatiScan** searches over the whole ordered list, trying to find runs of functionally related genes.



Example – two classes?



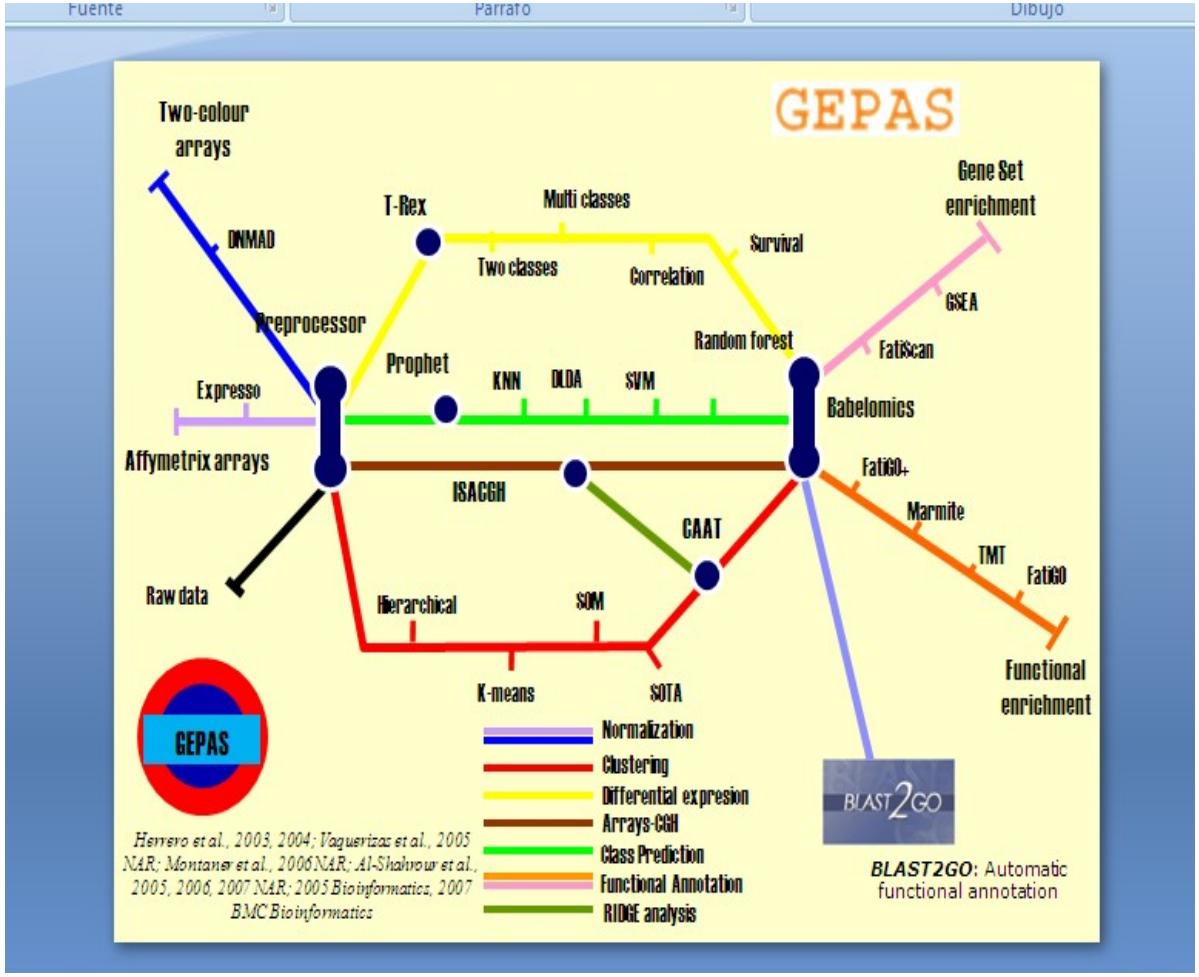
t ~ Tumor mean expression – Control mean expression

Proliferation

Is more associated with the genes on the top of the list

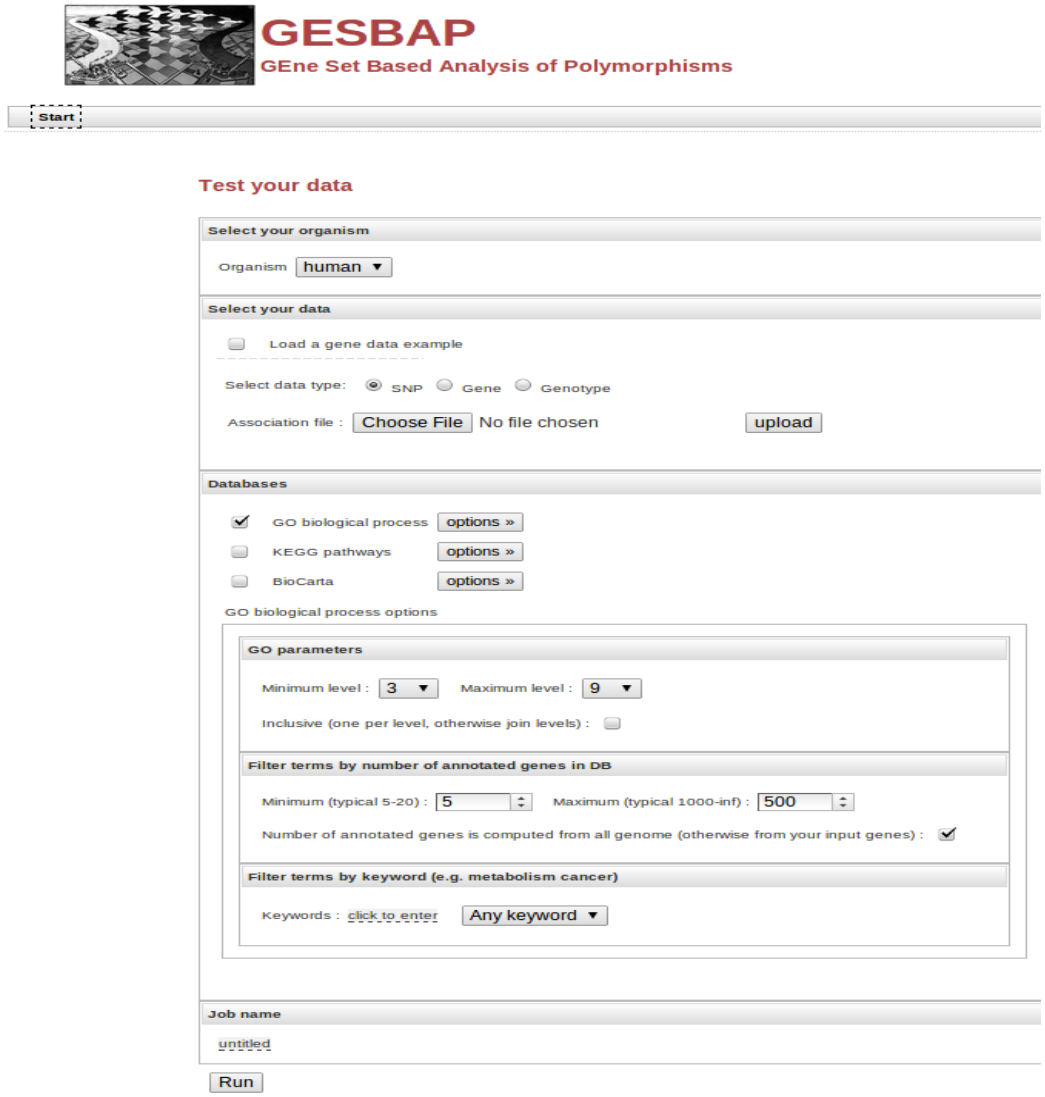
Is more associated with the genes that show higher expression in Tumors

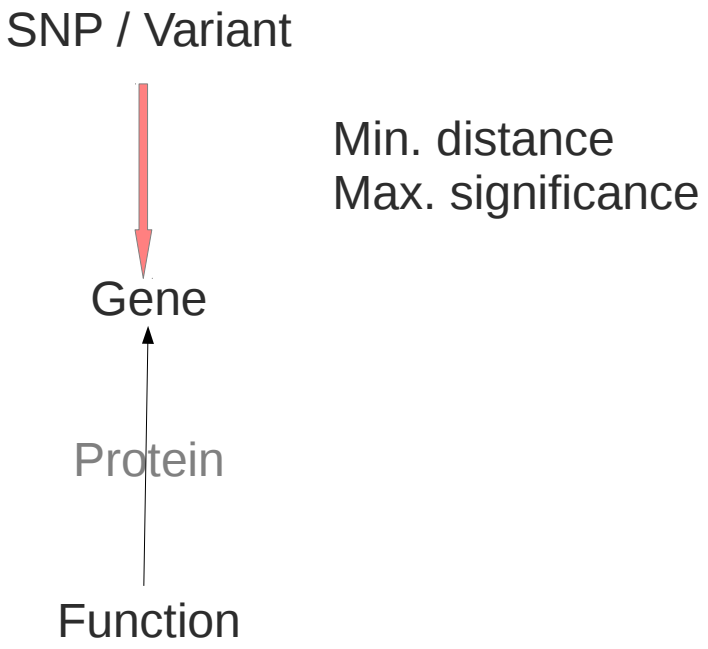
Babelomics - Gepas



www.babelomics.org

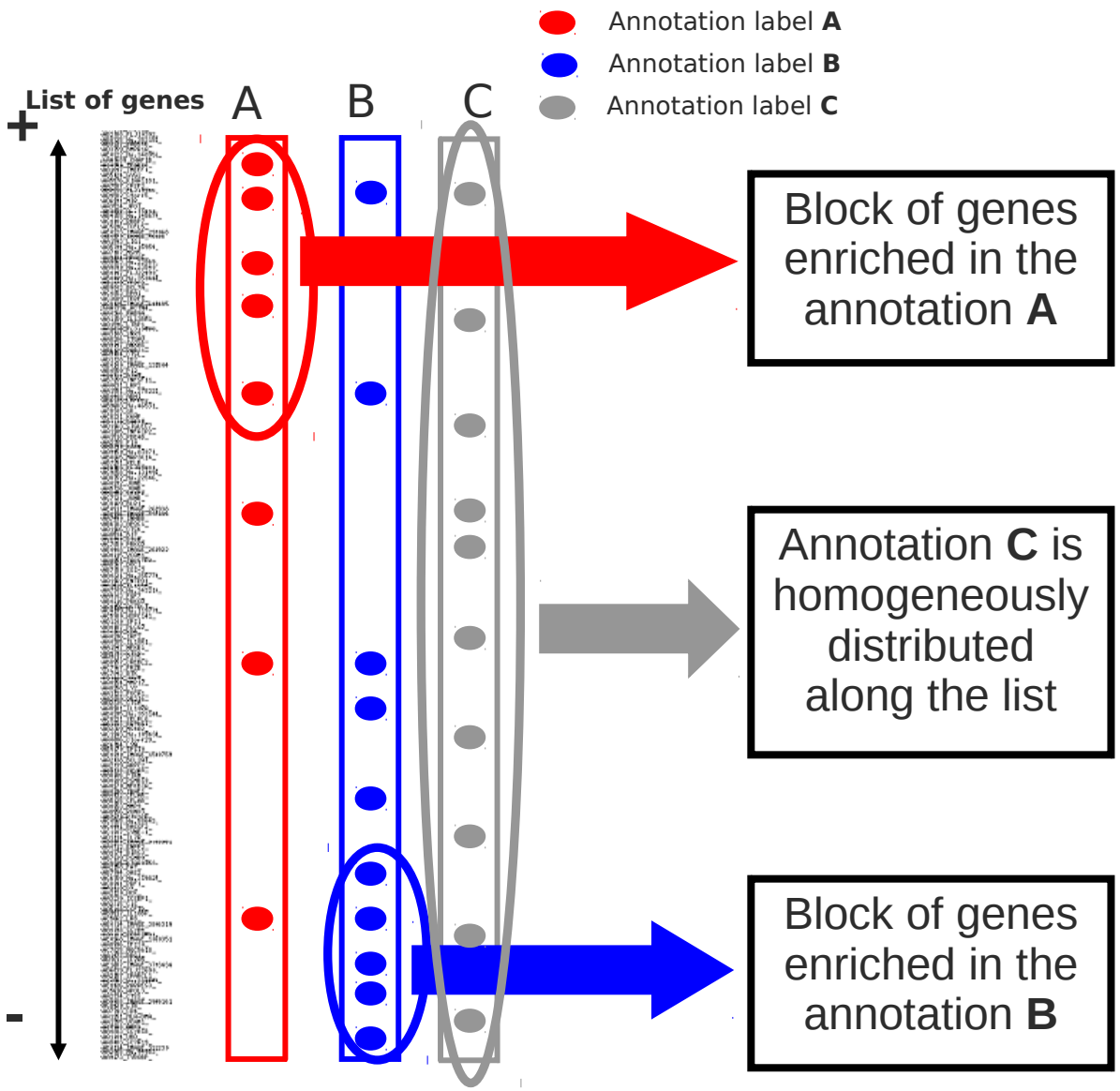
Functional Profiling for Genomic Studies



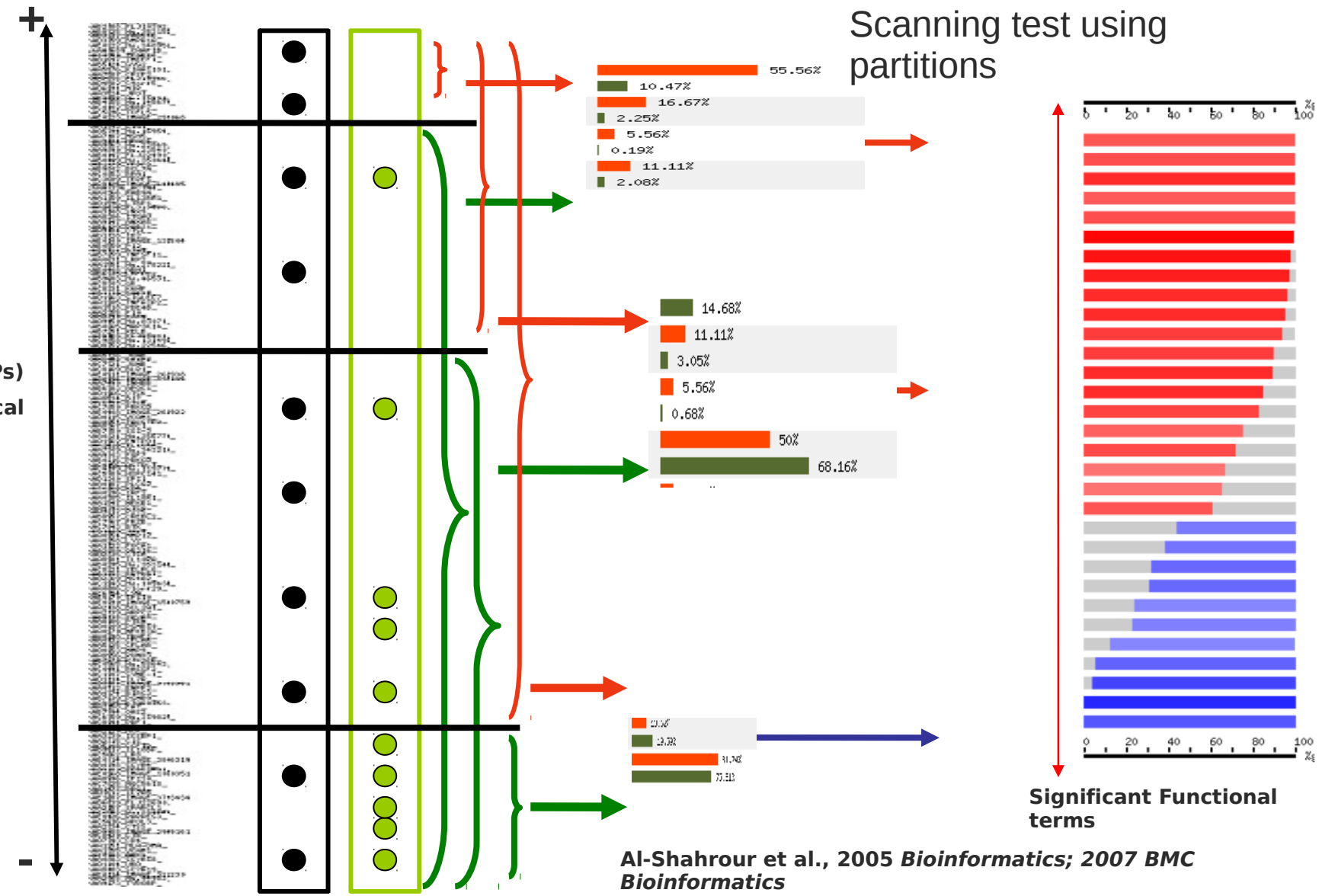


Testing along an ordered list

- Index ranking genes according to some biological aspect under study.
- Database that stores gene class membership information.
- **FatiScan** searches over the whole ordered list, trying to find runs of functionally related genes.



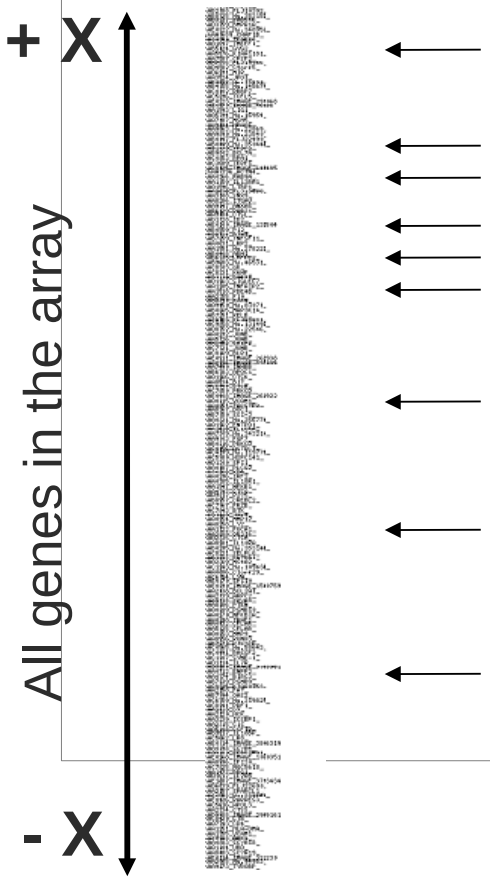
FatiScan Methodology



Al-Shahrour et al., 2005 ***Bioinformatics; 2007 BMC Bioinformatics***

Logistic regression test

- Not using partitions
- But logistic regression model



$$\ln\left(\frac{P(g\in F)}{P(g\notin F)}\right) = K + \alpha X$$

alpha > 0 : **increasing** X increases the probability of the gen to be annotated

alpha < 0 : **decreasing** X increases the probability of the gen to be annotated

Remarks

- The unit of information over which we test is shifted from genes to functional blocks (multiple testing again)
- We do one statistical test for each block
- All genes in the block are treated equally
- Annotation information is 0, 1
- Genes independently may not show a strong pattern of association but the block coordinately does.
- Only ranking genes according to a unique condition

Gene Set Methods - 2 general Approaches

- Competitive Hypothesis – Babelomics / Gesbap
  - Each functional block is compared to the remaining genes of the genome (or the second list).
  - Independent of the test used to derive the ranking.
- Self Contained Hypothesis – Goeman 2004
  - Checks that the block it self is differentially expressed, correlated with phenotype, associated to survival...
  - Has to be developed with the test creating the ranking of the genes.

?