Functional Profiling

Máster en Bioinformática y Biología Computacional

Francisco García García,   fgarcia@cipf.es
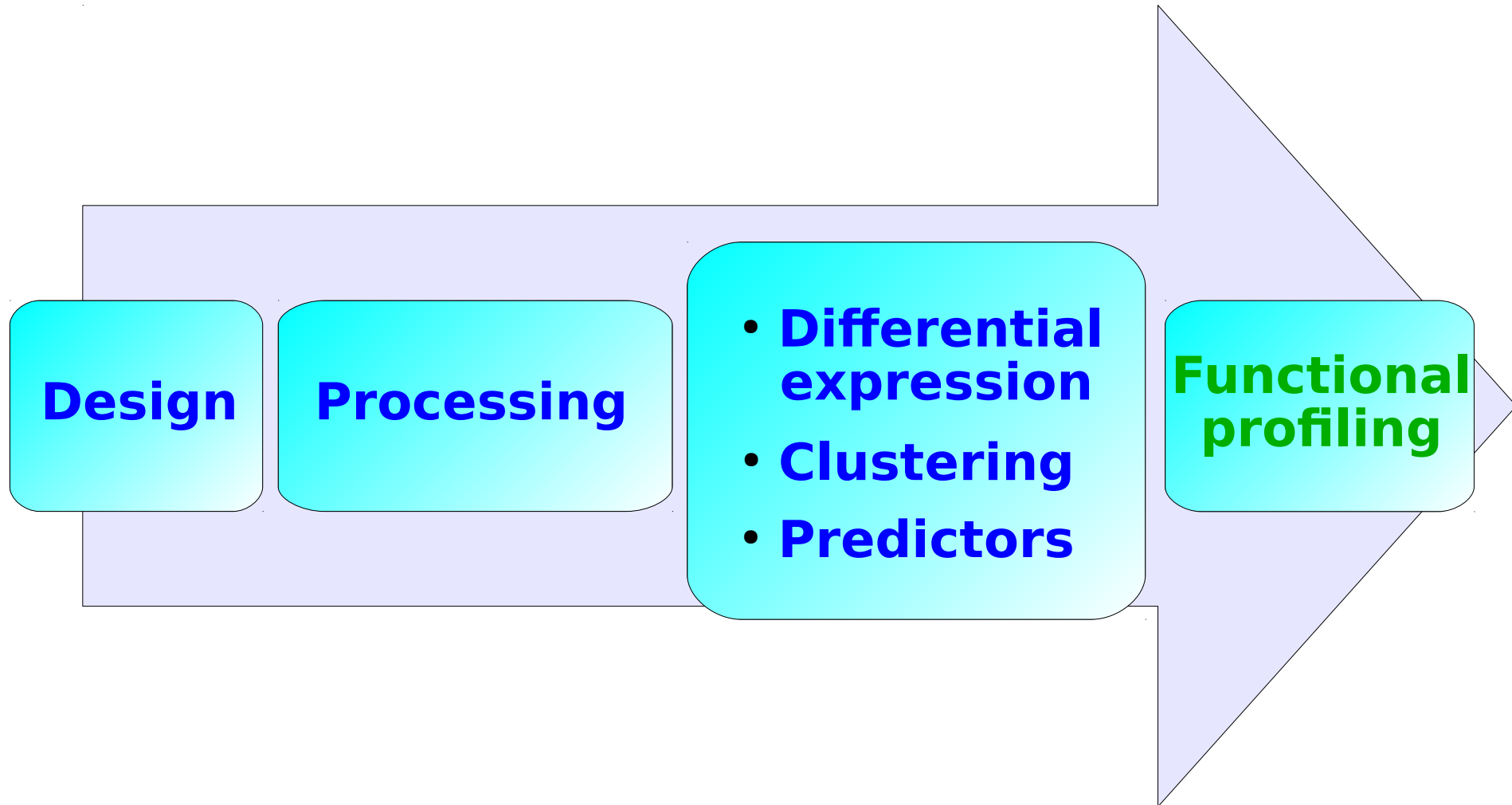
Outline

1. **Introduction**

2. **Over-Representation Analysis (ORA)**

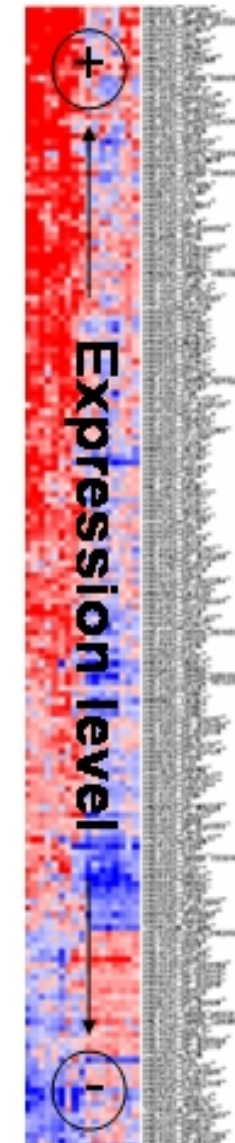3. **Gene Set Analysis (GSA)**
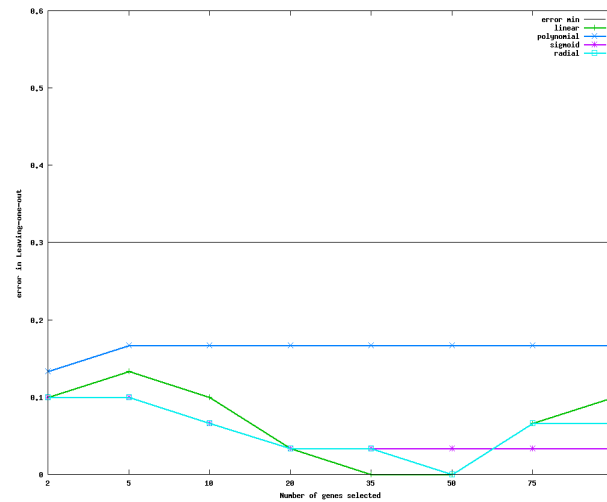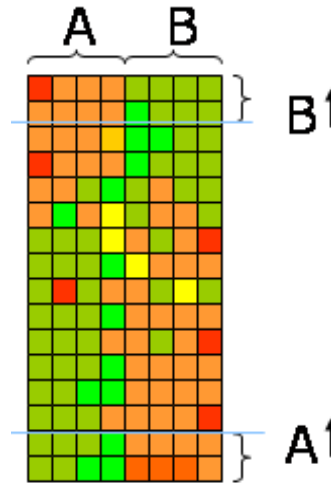
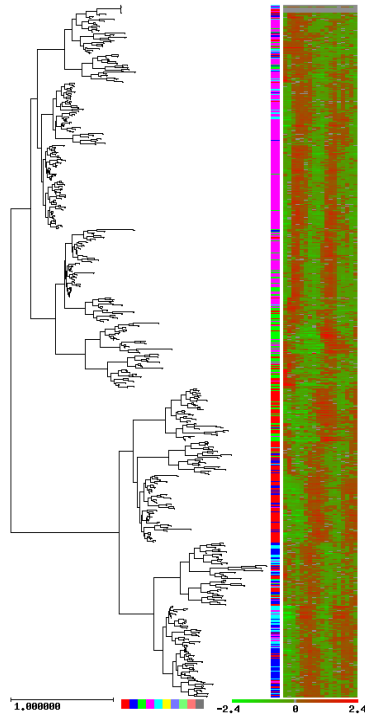4. **Network Analysis (NA)**

Outline

- Introduction

- Over-Representation Analysis (ORA)

- Gene Set Analysis (GSA)

- Network Analysis (NA)

Functional Profiling

Expression Data Analysis Pipeline

Design → **Processing** →
- **Differential expression**
- **Clustering**
- **Predictors**
→ **Functional profiling**

Genome-scale experiment output

Genome-scale experiment output

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

.

.

A B

B↑

1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at

.

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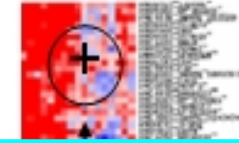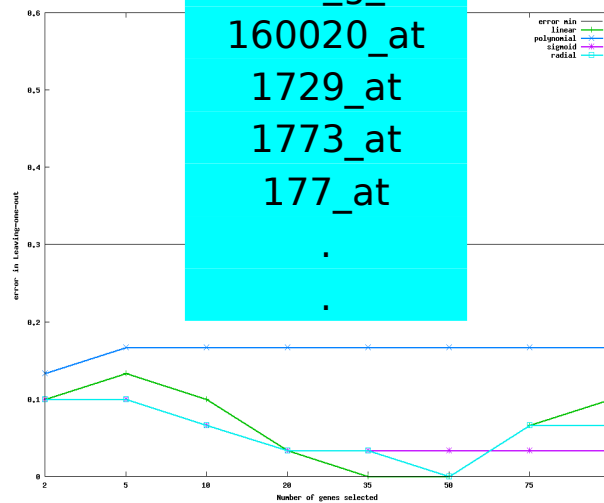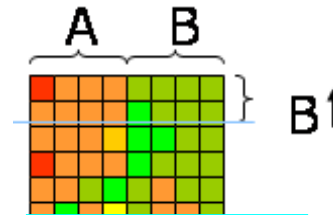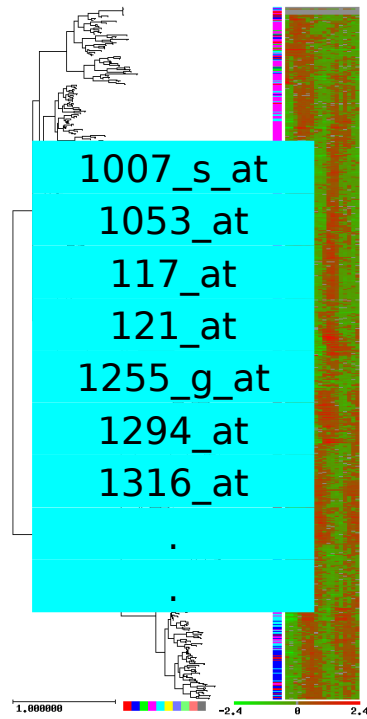
1007_s_at	12.4
1053_at	11.5
117_at	10.3
121_at	10.2
1255_g_at	9.9
1294_at	9.3
1316_at	8.2
1320_at	8.1
1405_i_at	7.7
1431_at	7.4
1438_at	6.5
1487_at	6.2
1494_f_at	5.9
1598_g_at	5.8
160020_at	4.8
1729_at	4.7
.	.
.	.

Introduction

Functional Profiling

Questions we try to answer

- Is there any significant functional enrichment in my gene list / gene sets?

- Are these genes involved in common pathways?

- Do they share specific regulation?

- Are they involved in the same disease?

Functional databases

Homo sapiens **Mus musculus** **Rattus norvegicus** **Gallus gallus** **Danio rerio** **Drosophila melanogaster** **C. elegans** **Saccharmoyces cerevisae** **Arabidopsis thaliana**

UniProt/Swiss-Prot

UniProtKB/TrEMBL

Ensembl IDs

EntrezGene

Affymetrix

Agilent

Genes IDs

HGNC symbol

EMBL acc

RefSeq

PDB

Protein Id

IPI....

Biological databases

KEGG pathways

Biocarta pathways

Keywords Swissprot

Gene Ontology

Biological Process Molecular Function Cellular Component

Gene Expression in tissues

Regulatory elements

MiRNA, CisRed

Transcription Factor Binding Sites

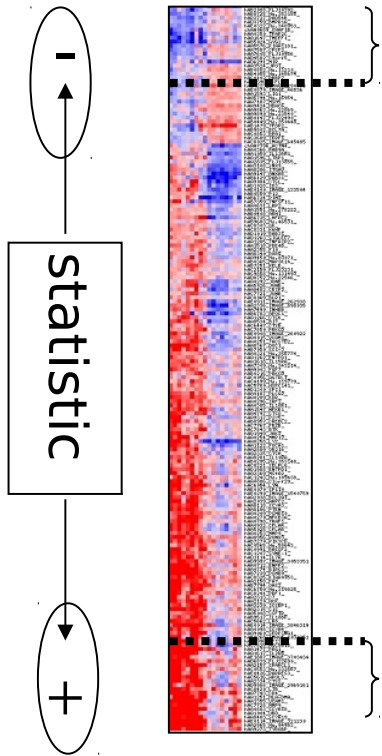**Bioentities from literature:**

Diseases terms
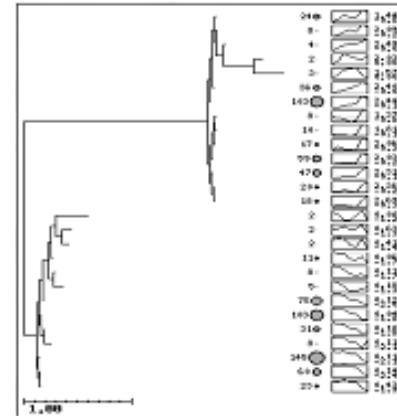**Chemical terms**

Introduction

Functional Profiling

Outline

- Introduction

- Over-Representation Analysis (ORA)

- Gene Set Analysis (GSA)
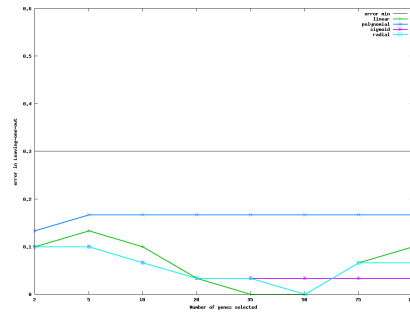
- Network Analysis (NA)

Functional Profiling

Over-Representation Analysis

ORA

Functional Profiling
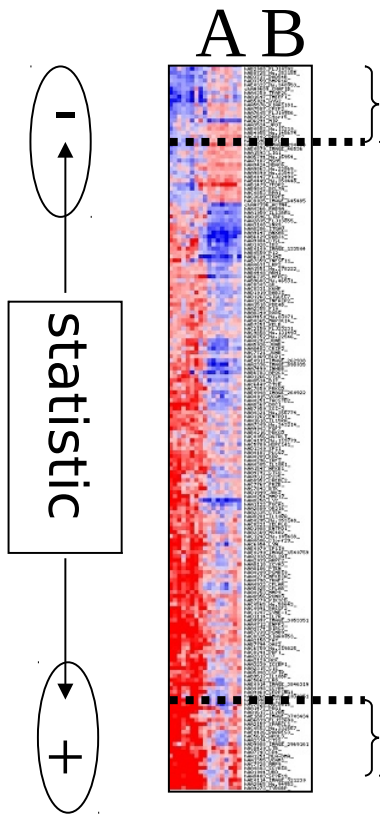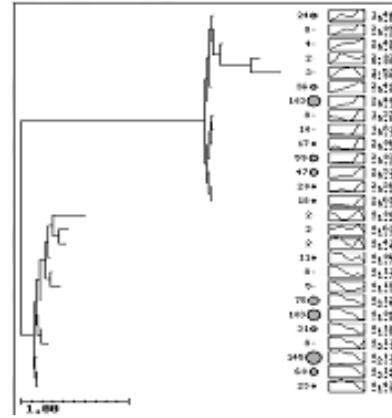
Over-Representation Analysis

A B

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

Function
4/7

1320_at
1405_i_at
1431_at
1438_at
1487_at
1494_f_at
1598_g_at
160020_at
1729_at
1773_at
177_at
.
.

Function
2/11

ORA

Functional Profiling
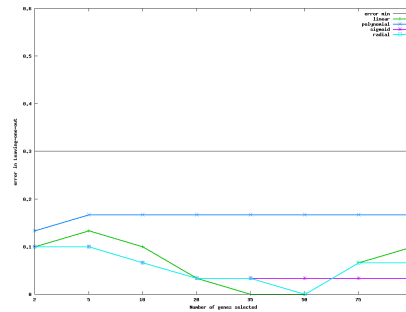
Over-Representation Analysis

FatiGO test

One Gene List (A)
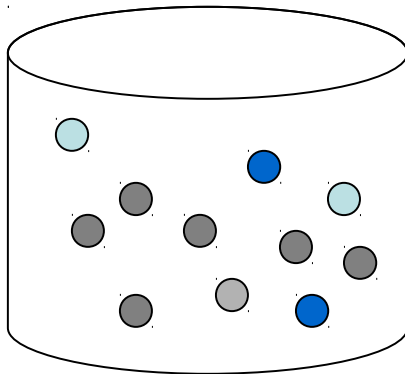
Are this two groups of genes carrying out different biological roles?

The other list (B)

Biosynthesis 60% ●

Sporulation 20% ●

Biosynthesis 20% ●

Sporulation 20% ●

ORA

Functional Profiling

Over-Representation Analysis

One Gene List (A)

Are this two groups of genes carrying out different biological roles?
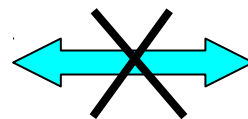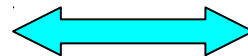
The other list (B)
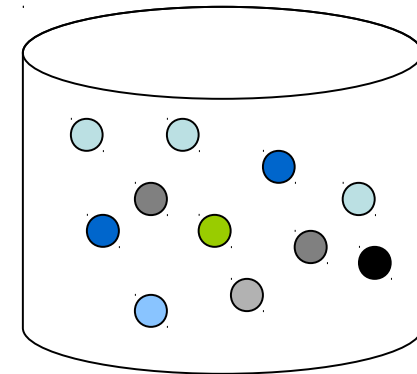
Biosynthesis 60%
Sporulation 20%

Biosynthesis 20%
Sporulation 20%
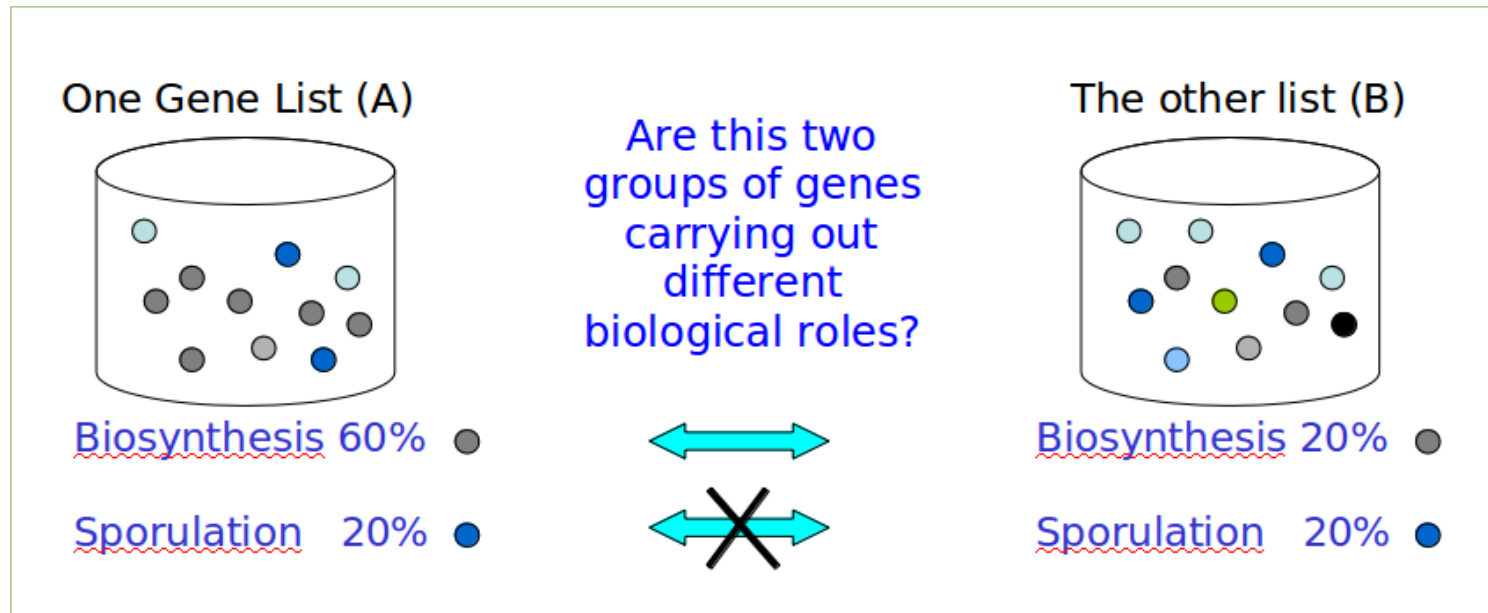
Genes in group A have significantly to do with biosynthesis, but not with sporulation.

	A	B
Biosynthesis	6	2
No biosynthesis	4	8

We do this for each term (GO, miRNA, Interpro , …) Thousand of terms, so Multiple Test Correction is needed!!!

ORA

Functional Profiling

Outline

- Introduction

- Over-Representation Analysis (ORA)

- **Gene Set Analysis (GSA)**

- Network Analysis (NA)

Functional Profiling

Gene Set Analysis

Significantly over-expressed in B

A B

t-test with two tails.

p<0.05

statistic

Significantly over-expressed in A

Functional Classes expressed as blocks in A and B

GSA

Functional Profiling

Gene Set Analysis

List of genes

A B C

Annotation label **A**
Annotation label **B**
Annotation label **C**

Block of genes enriched in the annotation **A**

Annotation **C** is homogeneously distributed along the list

Block of genes enriched in the annotation **B**

GSA

Functional Profiling
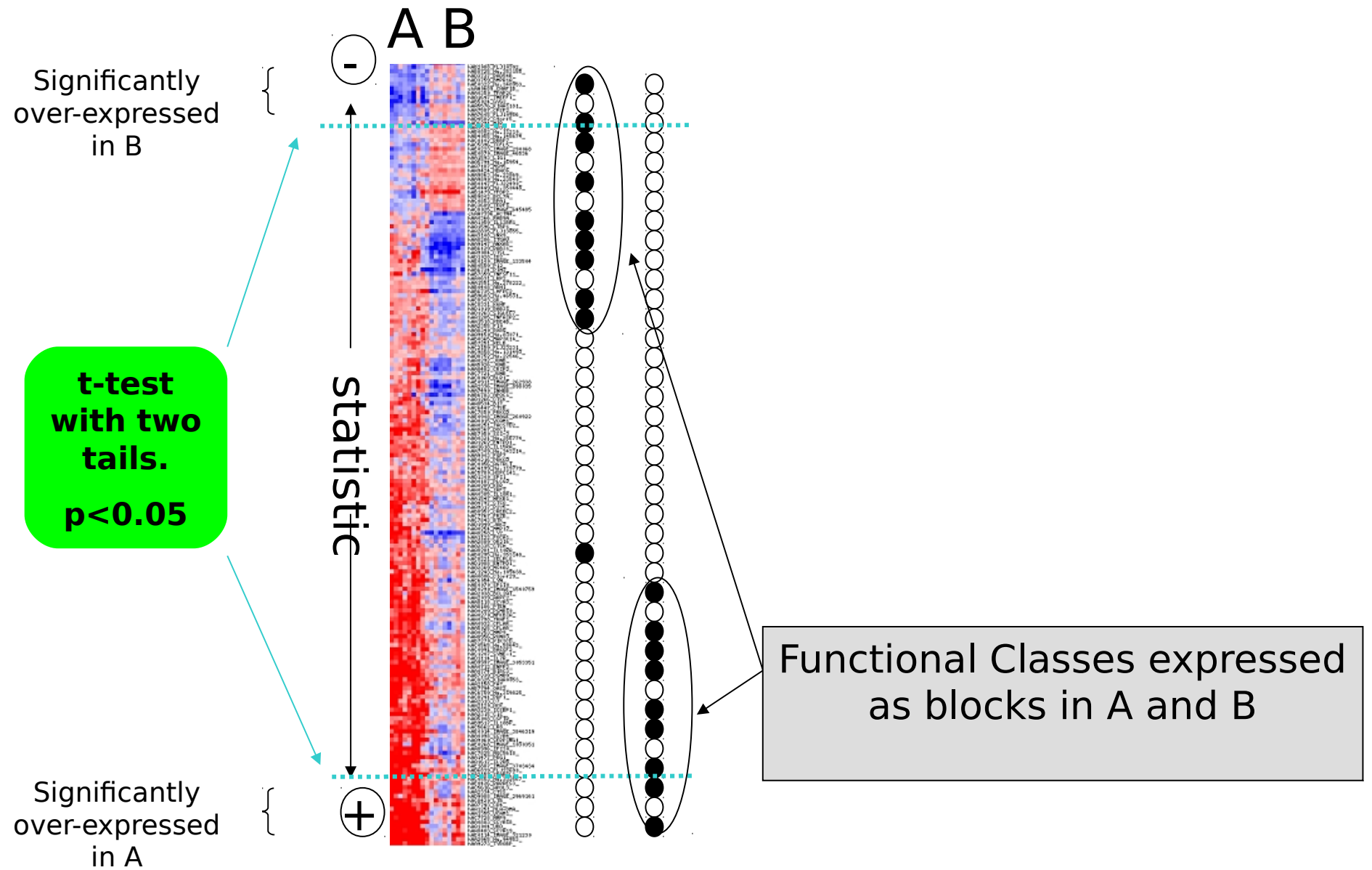
Gene Set Analysis

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

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

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

GSA

Functional Profiling

Hands on

Babelomics 5

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

Functional / FatiGO

Online examples

Hands on

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

Worked examples + exercises

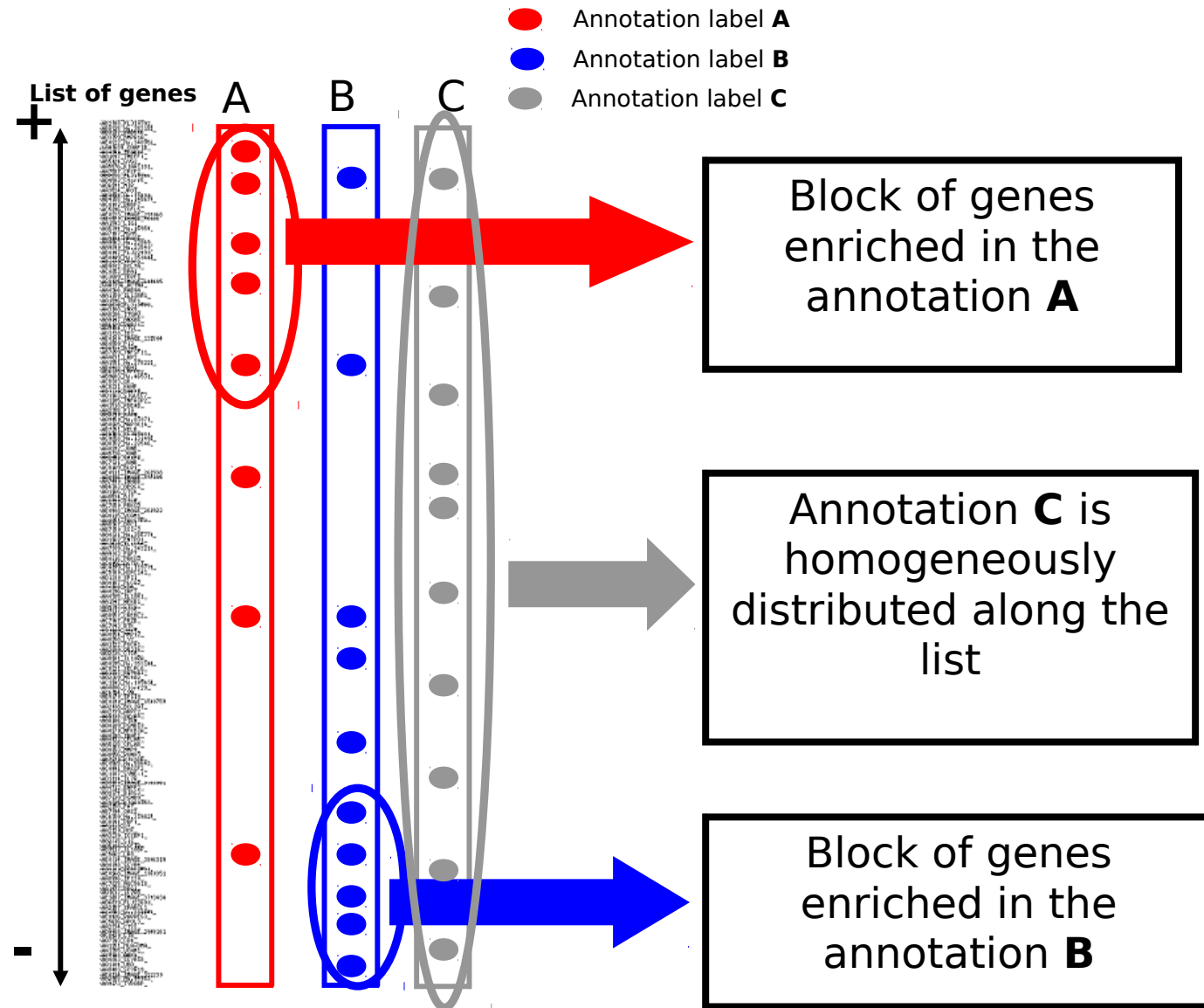# Outline

- Introduction

- Over-Representation Analysis (ORA)

- Gene Set Analysis (GSA)

- Network Analysis (NA)

Functional Profiling

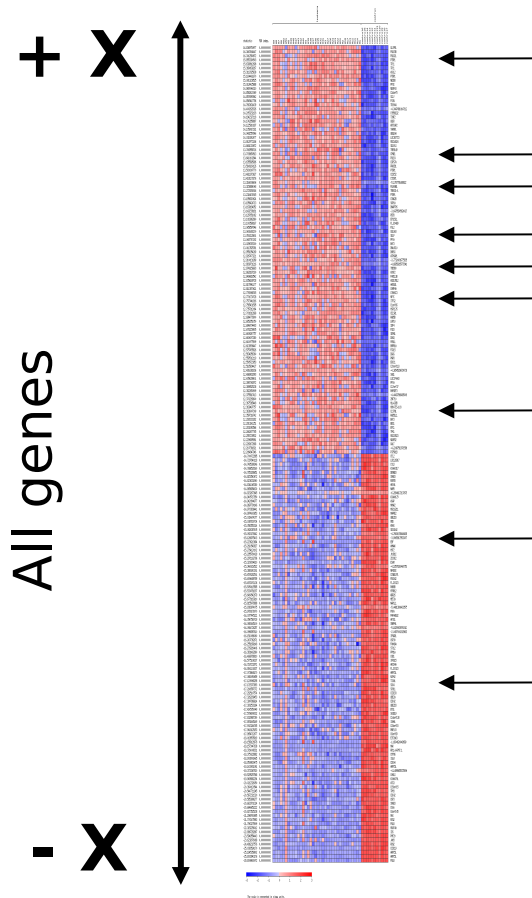# Protein-Protein Interactions (PPI)
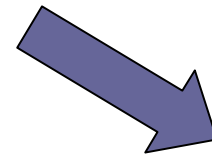
- How to extract information about **sets** of genes?

- How to perform **functional enrichment analysis** using protein-protein interactions as annotation source?

- How to **prioritize candidate genes**?

- How to get **new functional candidate genes**?
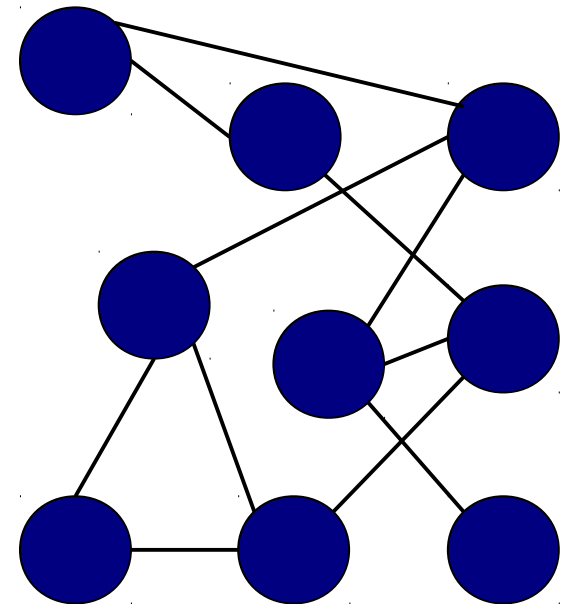
Graph Theory

Set of proteins interacting
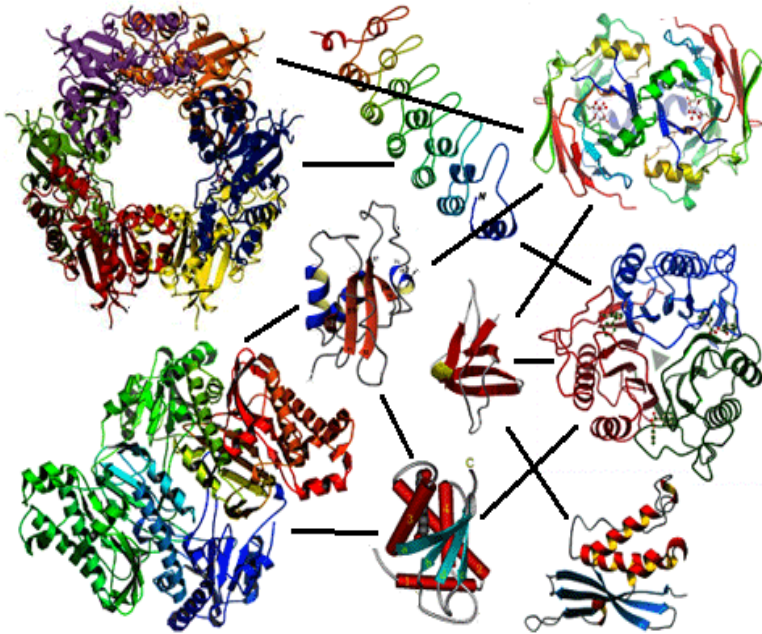
Undirected graph

Nodes = proteins
Edges = interaction events

structured data

Graph Theory

Graph theory may help us to study protein networks. Some interesting parameters:

- Degree (connectivity or connections): number of edges connected to a node. Nodes with high degree are called hubs.

- Betweenness: A measure of centrality of a node, it is defined by:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

σ_{st} is total number of shortest paths in the graph.

$\sigma_{st}(V)$ is the number of shortest paths that pass through node V

NA

Network Analysis in Babelomics 5
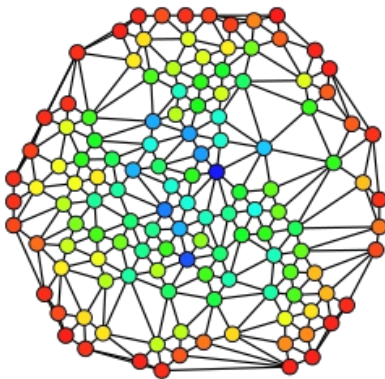
Graph Theory

- Clustering coefficient (of a node): A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.

$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$

e_i is the number of edges among the nodes connected to node 1
n_i is the number of neighbours of node i

c = 1/3
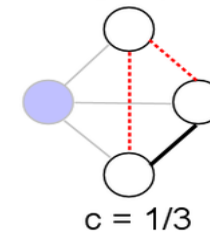
To differentiate between **star-shaped** nets and more **interconnected** nets.
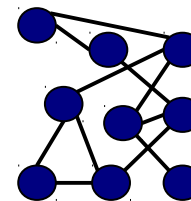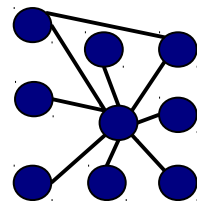
Graph Theory

Some Graph Theory concepts:

Shortest path. The path with less edges that connects two nodes.

Component. A group of nodes connected among them.

Bicomponent. A group of nodes connected to other group of nodes by only an edge. The edge that joins two bicomponents is called articulation point.

Interactome & Transcriptome

- **Interactome.** Complete collection of protein-protein interactions in the cell.

- **Transcriptome** determines the real interactome.

Set of **active** ppis
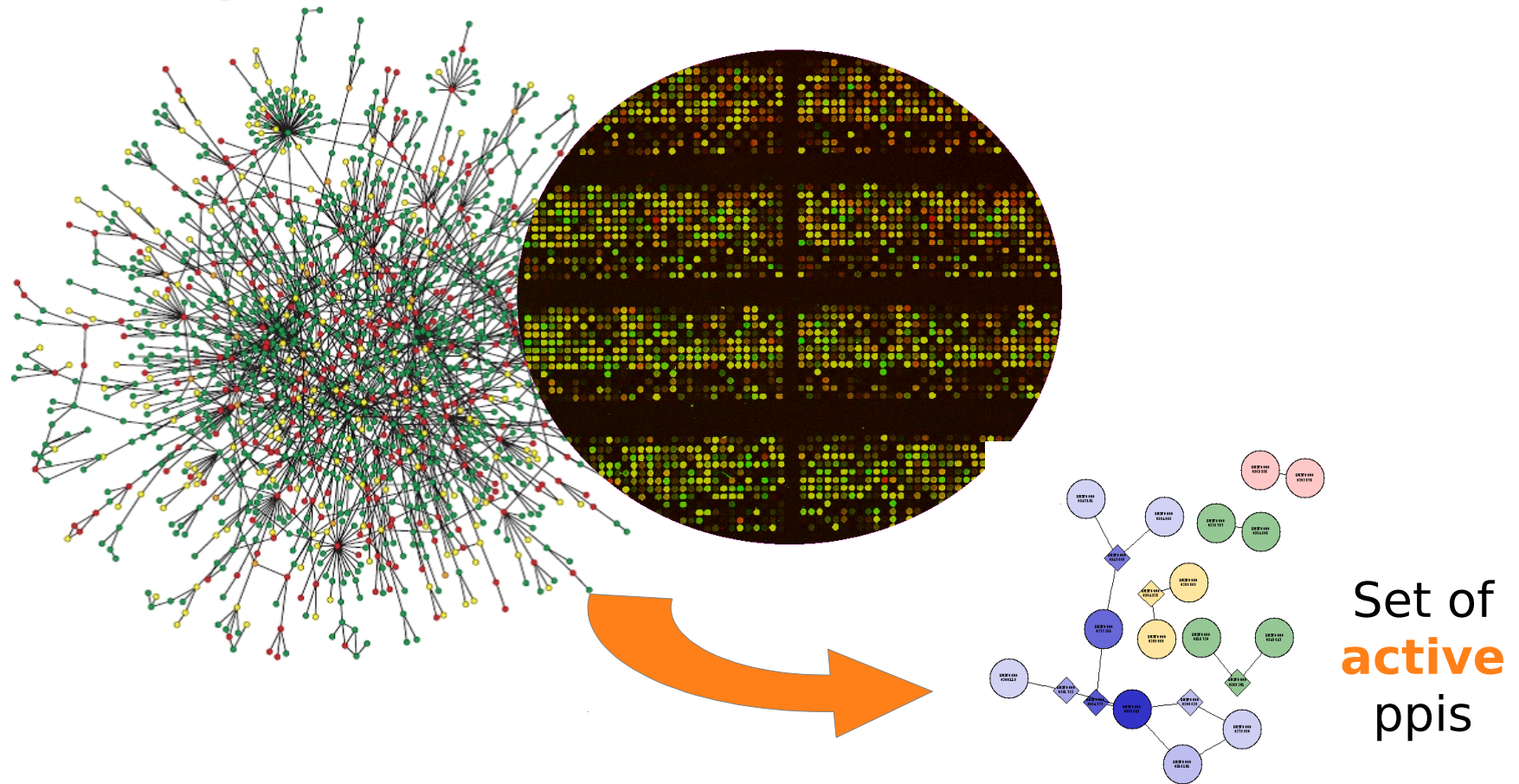
Interactome & Transcriptome

Goal

To develop a methodology that may extract from lists of proteins/genes the ppi networks acting and evaluates whether they have importance in the cooperative behaviour of the list.

How we evaluate the cooperative behaviour of a list of proteins/genes in terms of its ppi network parameters?

Two different approximations:

• Importance in **complete interactome**

• Cooperative behaviour - **Minimal Connected Network**

Network Analysis in Babelomics 5

Babelomics 5

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

Functional / Network Enrichment: SNOW

Online examples

Babelomics 5

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

Functional / Network Enrichment: SNOW

More exercises for you!

Hands on

There is a well-known list of 72 genes related to eye diseases (ABCA4, ABHD12,ADAMTS18,AIPL1,BBS1,BEST1,C2orf71,C8ORF37,CA4,CABP4, CEP290,CERKL,CHM,...)

1) Now we have a two new candidates: RHO and TULP1 . We would to know what is the relationship between all genes.
2) Also it would be interesting to explore new functional candidates.

Strategies from Babelomics?
- Single Enrichment
- **Network** Enrichment

Hands on

RHO	TULP1
ABCA4	MERTK
ABHD12	MPDZ
ADAMTS18	NMNAT1
AIPL1	NR2E3
BBS1	NRL
BEST1	OFD1
C2orf71	PDE6A
C8ORF37	PDE6B
CA4	PDE6G
CABP4	PRCD
CEP290	PROM1
CERKL	PRPF3
CHM	PRPF31
CLRN1	PRPF6
CNGA1	PRPF8
CNGB1	PRPH2
CRB1	RBP3
CRX	RD3
CYP4V2	RDH12
DHDDS	RGR
EYS	RLBP1
FAM161A	ROM1
FSCN2	RP1
GUCA1B	RP2
GUCY2D	RP9
IDH3B	RPE65
IMPDH1	RPGR
IMPG1	RPGRIP1
IMPG2	SAG
IQCB1	SEMA4A
KCNJ13	SNRNP200
KLHL7	SPATA7
LCA5	TOPORS
LRAT	TTC8
MAK	USH2A