Next Generation Sequencing Applications in Transcriptomic Studies

Master in Biomedical Technologies Management and Development

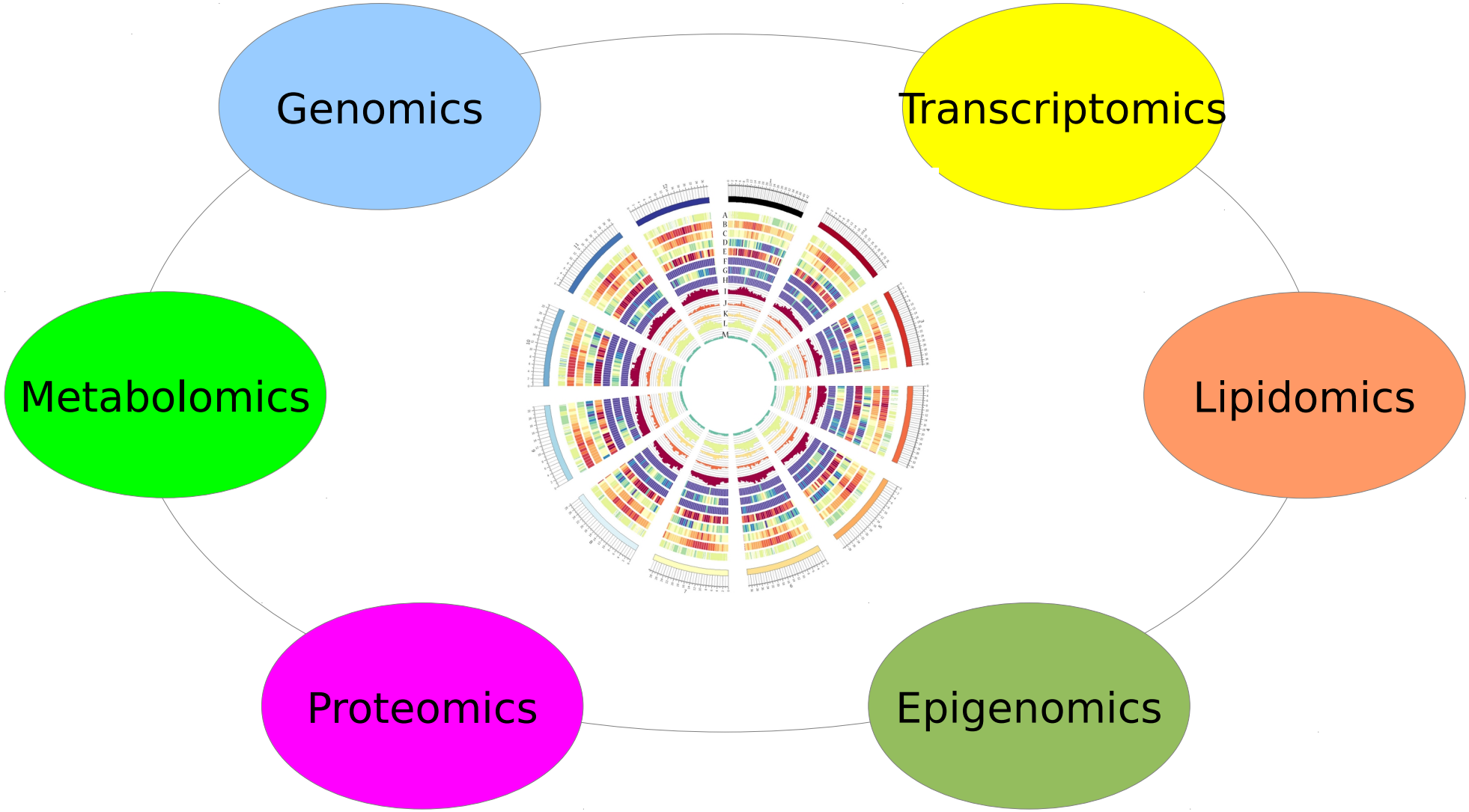
Madrid 24 Feb 2016

PRINCIPE FELIPE
CENTRO DE INVESTIGACION
Computational · Genomics

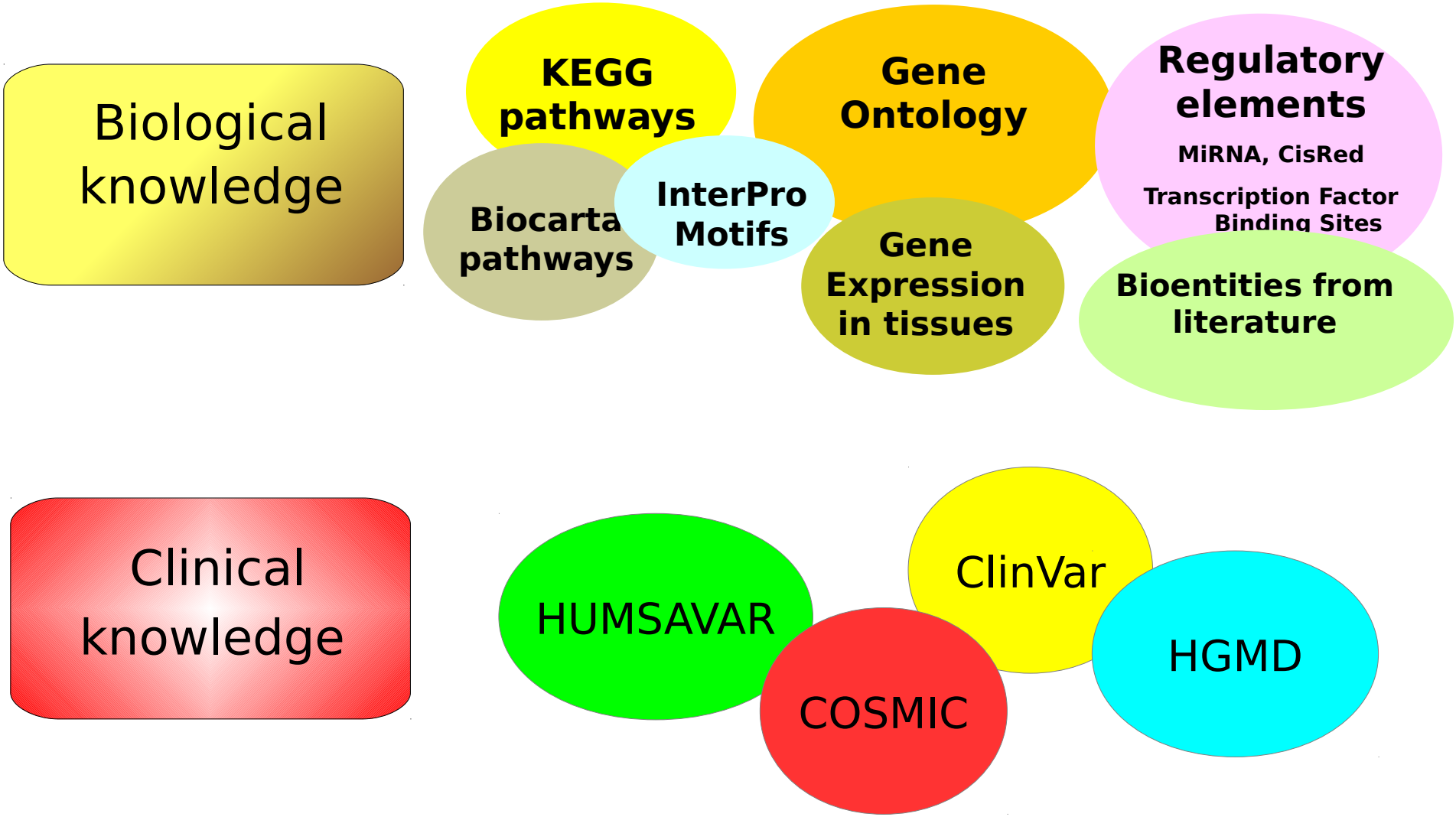
Francisco García
fgarcia@cipf.es

Genomic Computational Department . CIPF

Application of omic technologies



Application of omic technologies



Application of omic technologies

RNA-Seq is used to analyze the continually changing cellular transcriptome. Several applications to look:

- At alternative gene spliced transcripts
- Post-transcriptional modifications
- Gene fusion
- Mutations/SNPs
- **Changes in gene expression**.

RNA-Seq can look at different populations of RNA to include **total RNA**, small RNA, such as **miRNA**, tRNA, and ribosomal profiling.

Outline

1) **Introduction to NGS Data Analysis in Transcriptomic Studies**

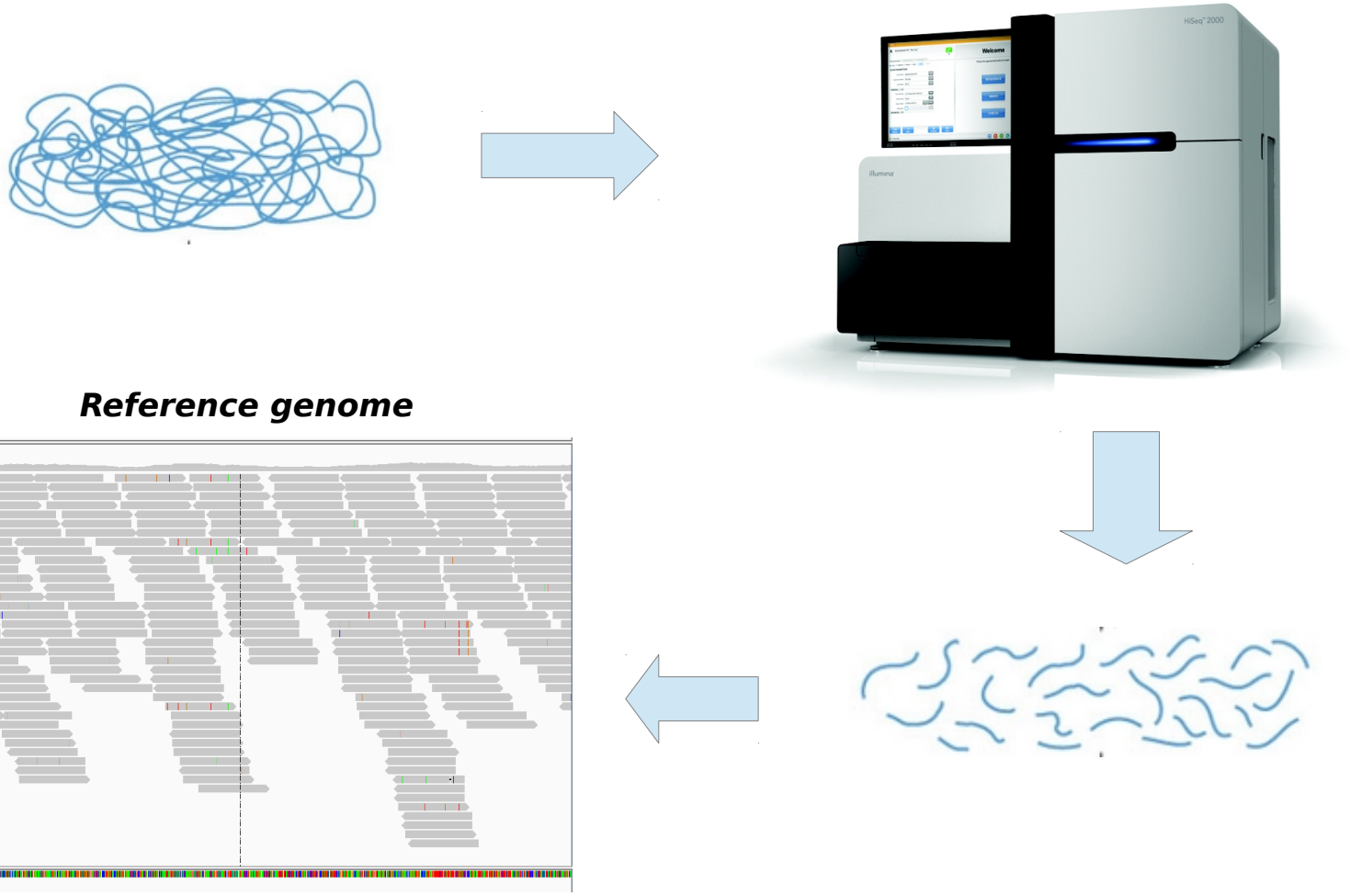
2) RNA-Seq and miRNA-Seq Data Analysis

3) Functional Profiling

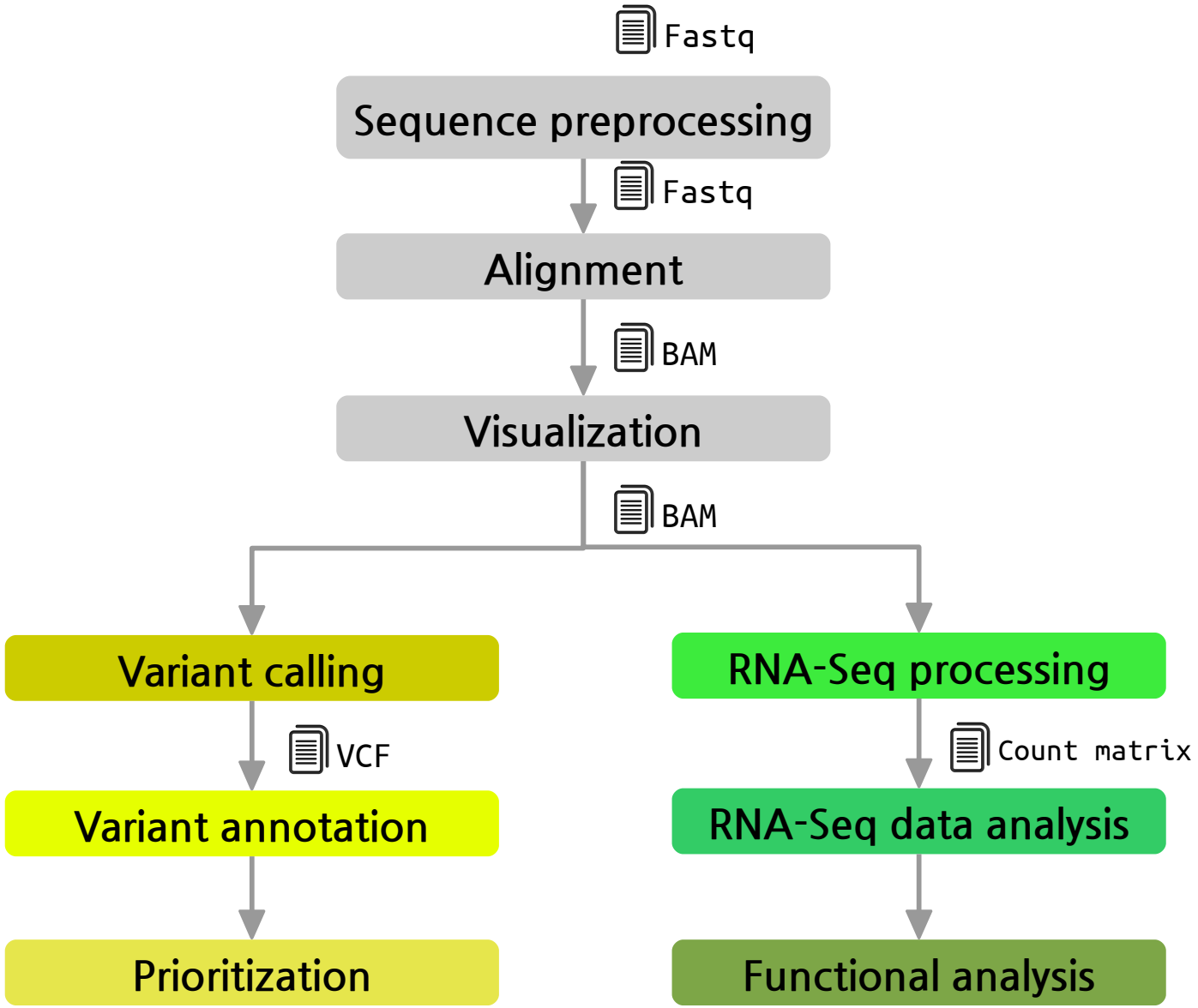
4) Omic Data Integration

NGS technologies

How do these technologies work ?



NGS Data Analysis Pipeline



Fastq format

- We could say “it is a fasta with **qualities**”:
  - 1. Header (like the fasta but starting with “@”)
  - 2. Sequence (string of nt)
  - 3. “+” and sequence ID (optional)
  - 4. Encoded quality of the sequence

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

BAM/SAM format

```
@PG     ID:HPG-Aligner  VN:1.0
@SQ     SN:20   LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1      16     20     76703   254    76=    *      0      0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1      16     20     76703   254    76=    *      0      0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEFGKDKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1      0      20     76655   254    76M    *      0      0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLFGEIIMHHIKL$BBGFFJIEHE    AS:i:254       NH:i:1  NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1     256    20     68253   254    4M1D72M *      0      0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTCAGGACACAGA
#4@ADEHFJFFEJDHJGKEFIHGHBGFHHFIICEIIFFKKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG    AS:i:254       NH:i:3  NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1      16     20     126103  246    53M4D23M       *      0      0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
   FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGIJKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIIB    AS:i:368       NH:i:1  NM:i:4
```

SAM Specification:
http://samtools.sourceforge.net/SAM1.pdf

Counts

Gene

Sample

Ensembl	Gene.Name	T1	T2	T3	T4	T5	WT1	WT2	WT3	WT4	WT5	WT6
ENSMUSG00000000134	Tfe3	312	295	333	258	392	257	344	223	423	277	389
ENSMUSG00000000142	Axin2	165	171	138	166	203	170	172	119	203	147	178
ENSMUSG00000000148	Brat1	213	196	207	224	350	204	268	143	300	177	288
ENSMUSG00000000149	Gna12	684	684	613	545	900	496	672	426	1023	583	797
ENSMUSG00000000154	Slc22a18	3	2	3	2	2	3	3	2	1	1	3
ENSMUSG00000000157	Itgb2l	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000159	Igsf5	0	0	0	0	0	0	0	0	0	0	0
ENSMUSG00000000167	Pih1d2	15	19	6	10	9	5	5	5	7	6	6
ENSMUSG00000000168	Dlat	899	777	967	756	1116	777	1047	614	1155	894	1126
ENSMUSG00000000171	Sdhd	1055	1003	1047	914	1430	939	1192	766	1390	916	1412
ENSMUSG00000000182	Fgf23	1	0	3	1	0	2	0	2	2	0	0
ENSMUSG00000000183	Fgf6	0	0	0	0	0	0	0	1	0	0	0
ENSMUSG00000000184	Ccnd2	1961	1978	1804	1779	2090	1655	2148	1585	2504	1895	2274
ENSMUSG00000000194	Gpr107	784	733	667	615	889	654	818	483	1034	627	1015
ENSMUSG00000000197	Nalcn	1120	1009	1047	917	1356	1129	1202	758	1625	1127	1044

Outline

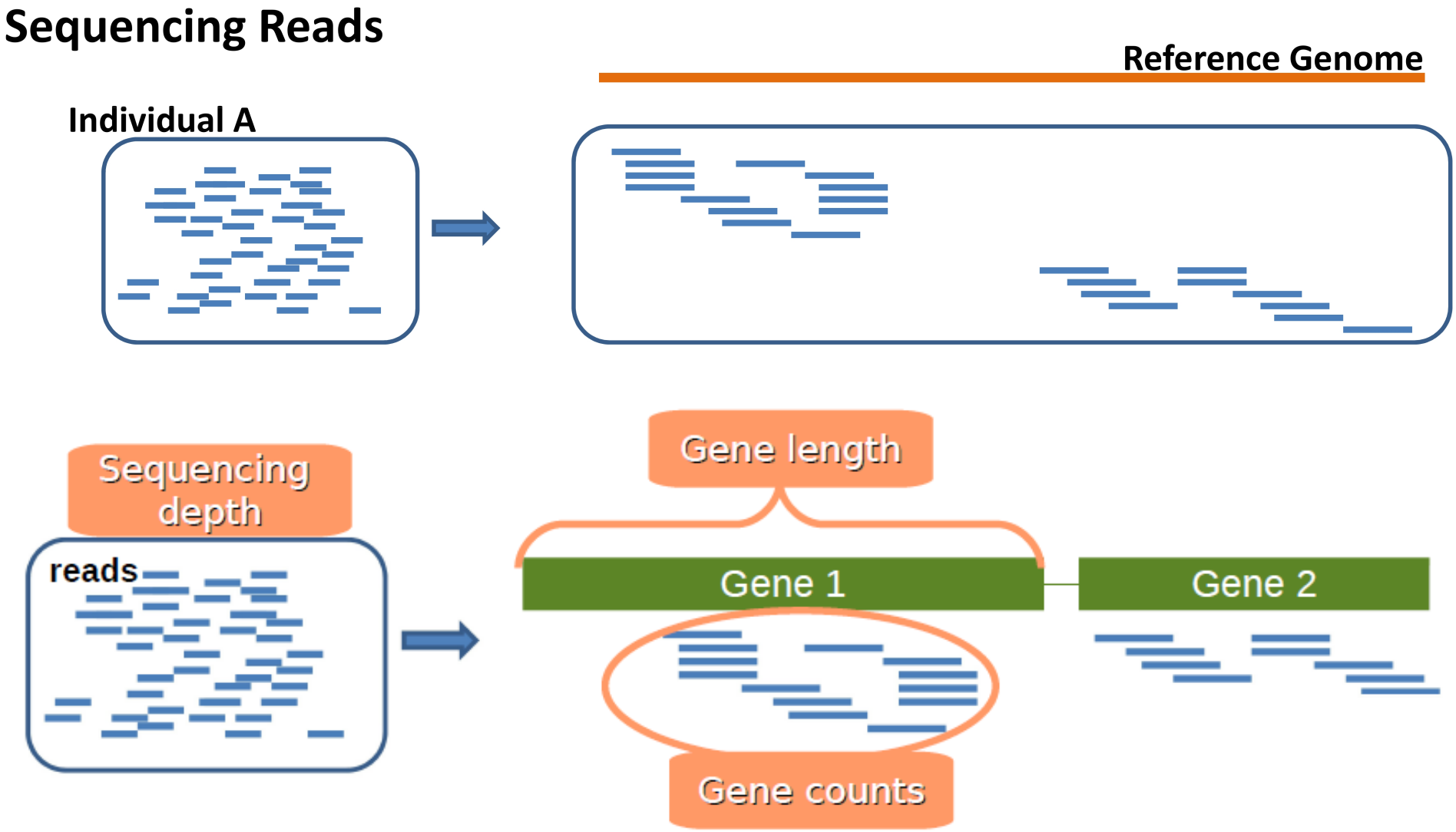
1) Introduction to NGS Data Analysis in Transcriptomic Studies

2) **RNA-Seq and miRNA-Seq Data Analysis**

3) Functional Profiling

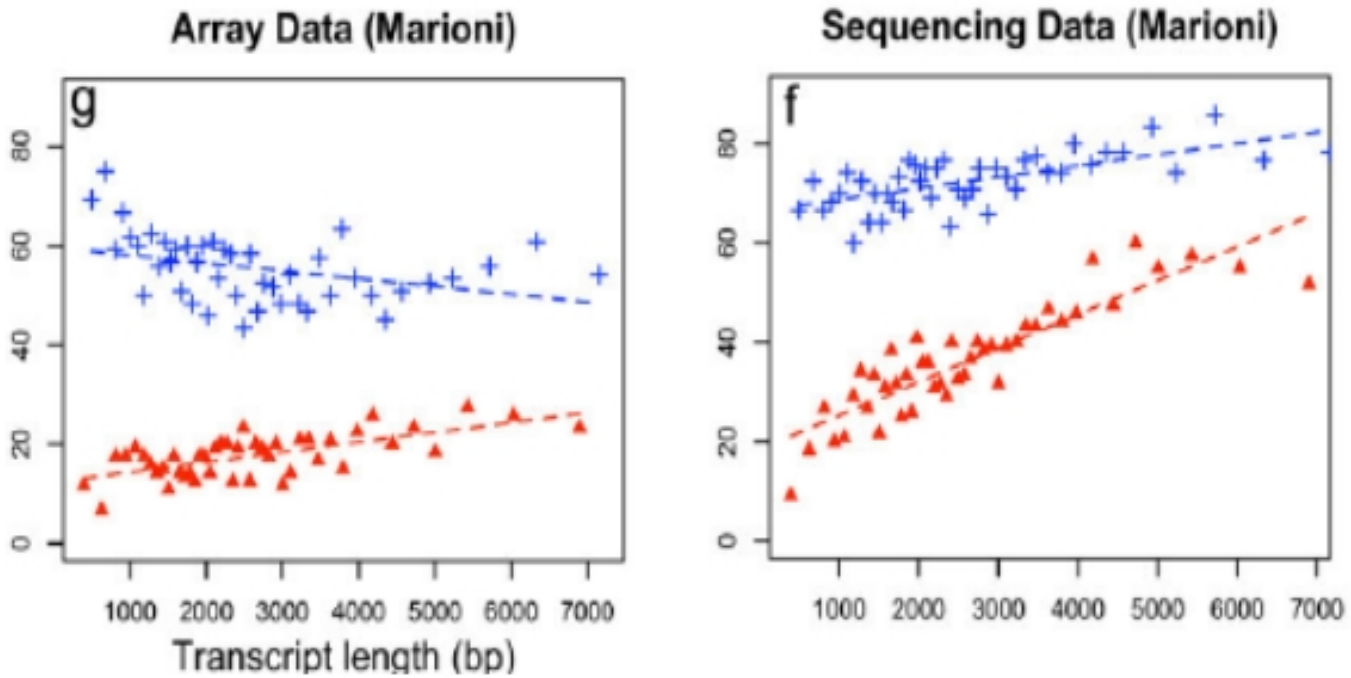
4) Omic Data Integration

General context



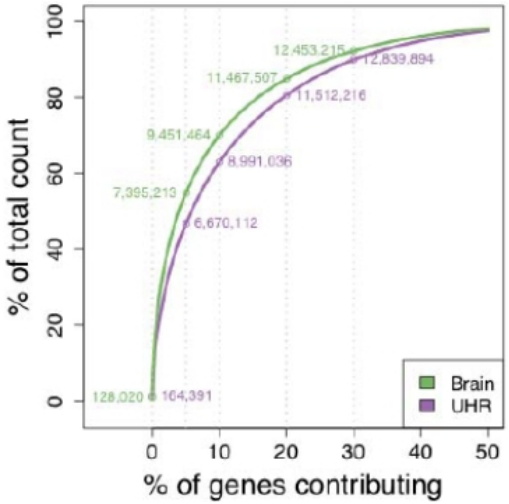
Gene/transcript length dependence

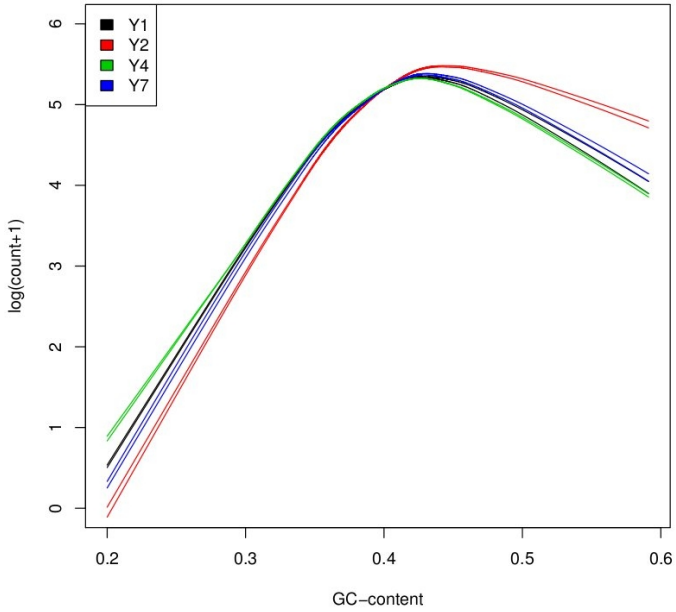
- Counts are proportional to...
  - the transcript length
  - the mRNA expression level.



Count Normalization

- **Transcript length**: *within* library
- **Library size**: *between* libraries
- Many **other biases** ...
  - Differences on the read count distribution among samples.
  - GC content of the gene affects the detection of that gene (Illumina)
  - sequence-specific bias is introduced during the library preparation





Count Normalization

- **RPKM**: Reads Per Kilobase of the transcript per Million mapped reads

$$RPKM = 10^9 \times \frac{C}{N*L}$$

- **C** is the number of mappable reads mapped onto the gene's exons.
- **N** is the total number of mappable reads in the experiment.
- **L** is the total length of the exons in base pairs.
- Fragments Per Kilobase of exon per Million fragments mapped (FPKM),

RNA-Seq Data Analysis Pipeline

Primary

1. Sequence preprocessing
2. Mapping
3. Quantification

Secondary

4. Normalization
5. Differential expression
6. Functional Profiling

Babelomics 5

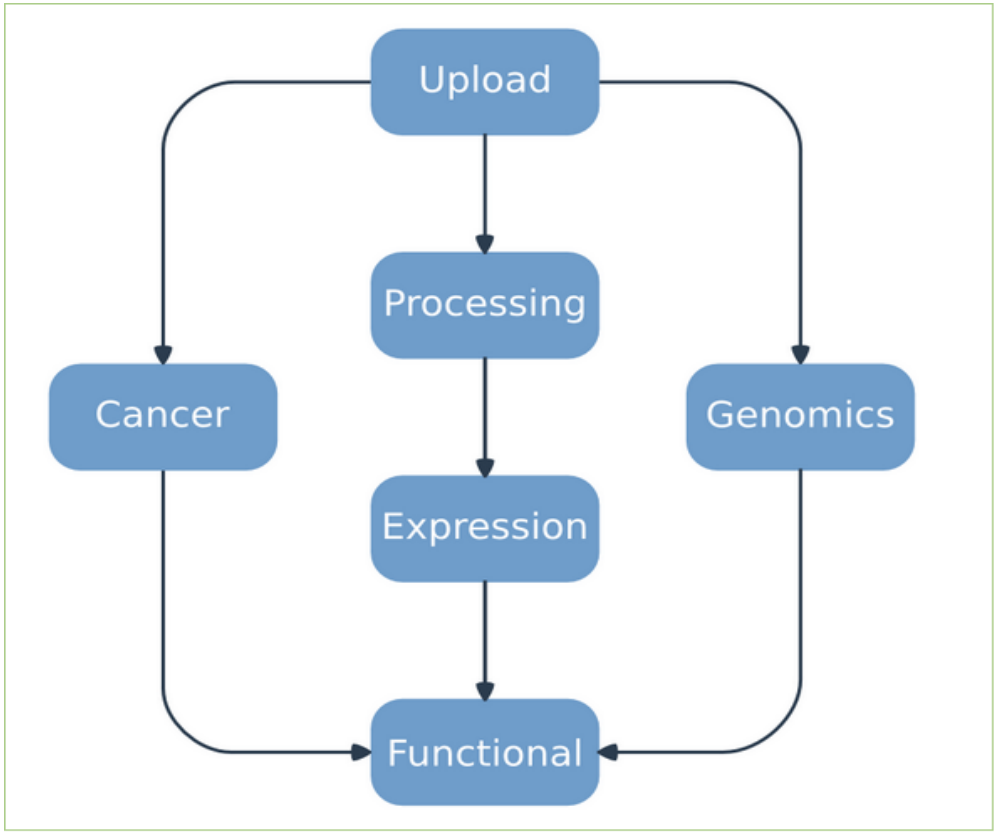
Gene Expression, Genome Variation and Functional Profiling Analysis Suite

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

Tool interface

Babelomics 5

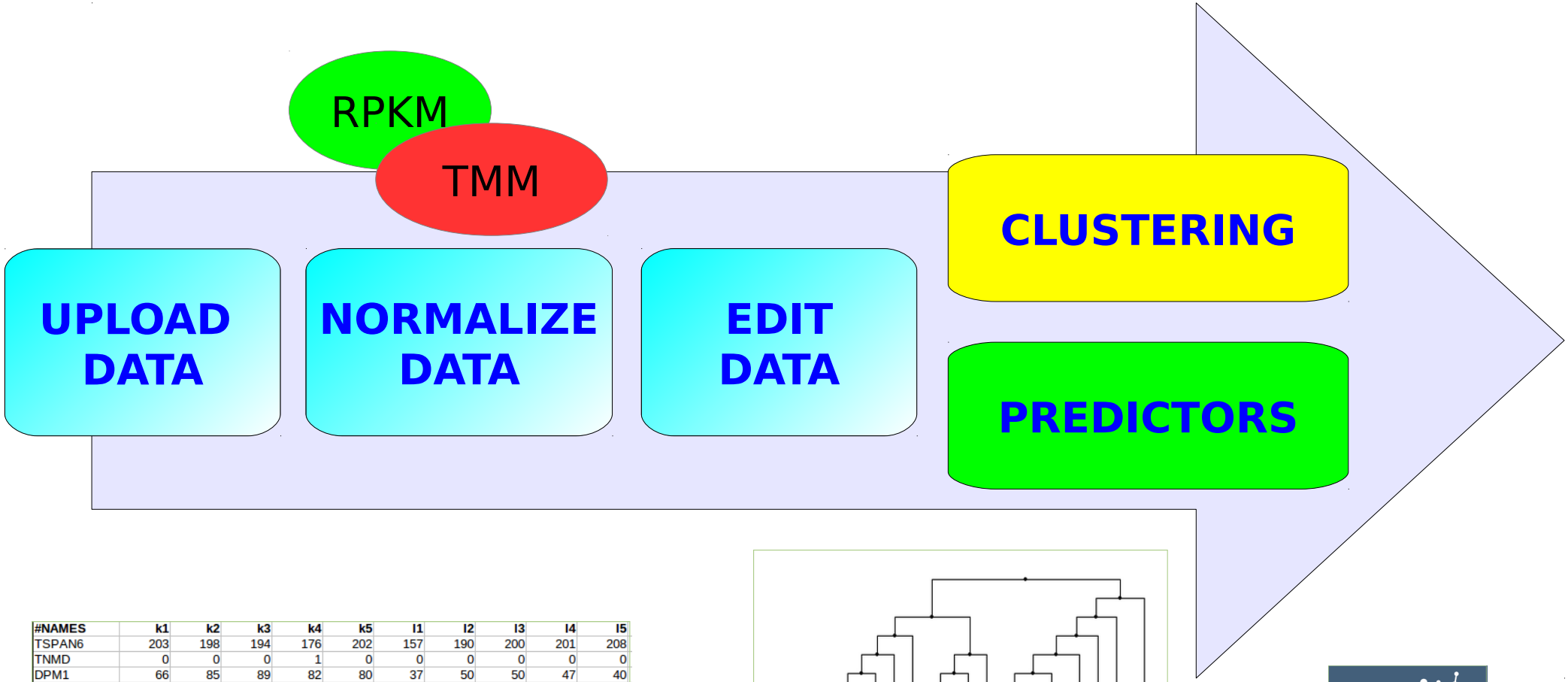
GENE EXPRESSION, GENOME VARIATION AND FUNCTIONAL PROFILING ANALYSIS SUITE



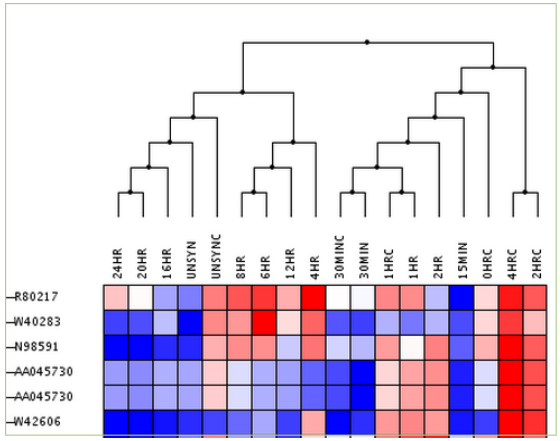
Babelomics 5

Analyzing omics data + functional profiling

Supervised and Unsupervised Classification



#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4	l5
TSPAN6	203	198	194	176	202	157	190	200	201	208
TNMD	0	0	0	1	0	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47	40
SCYL3	21	30	31	27	31	28	31	37	15	21
C1orf112	10	12	8	11	18	17	22	12	12	19
FGR	19	28	18	20	10	47	50	43	49	48
FUCA2	240	272	261	256	211	76	82	85	68	83
GCLC	98	100	84	94	86	354	362	373	369	326
NFYA	59	61	53	56	59	59	66	63	66	62
STPG1	34	43	41	31	46	6	7	7	8	7



Differential Expression

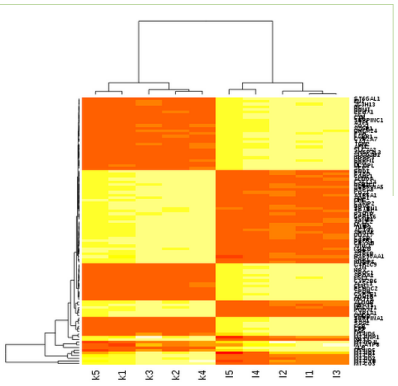
UPLOAD DATA

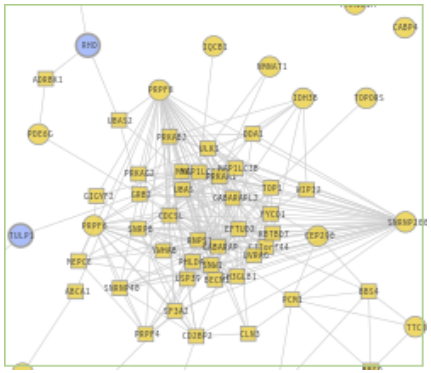
EDIT DATA

NORMALIZATION + DIFFERENTIAL EXPRESSION

FUNCTIONAL PROFILING

#NAMES	k1	k2	k3	k4	k5	l1	l2	l3	l4
TSPAN6	203	198	194	176	202	157	190	200	201
TNMD	0	0	0	1	0	0	0	0	0
DPM1	66	85	89	82	80	37	50	50	47
SCYL3	21	30	31	27	31	28	31	37	15
C1orf112	10	12	8	11	18	17	22	12	12
FGR	19	28	18	20	10	47	50	43	49
FUCA2	240	272	261	256	211	76	82	85	68
GCLC	98	100	84	94	86	354	362	373	369
NFYA	59	61	53	56	59	59	66	63	66
STPG1	34	43	41	31	46	6	7	7	8





Hands on

Babelomics 5

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

Processing / Normalization: RNA-Seq
Expression / Differential Expression: RNA-Seq

Online examples

Outline

1) Introduction to NGS Data Analysis in Transcriptomic Studies

2) RNA-Seq and miRNA-Seq Data Analysis

3) **Functional Profiling**

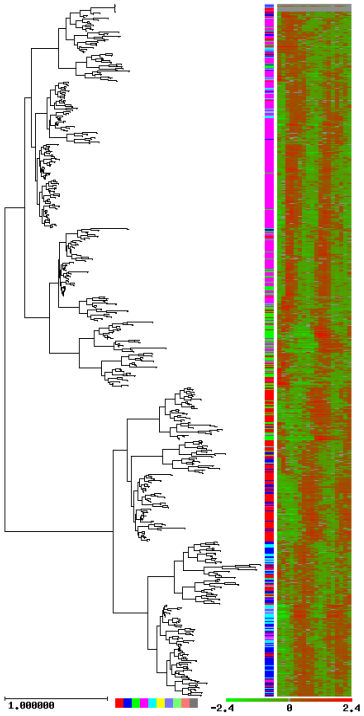
4) Omic Data Integration

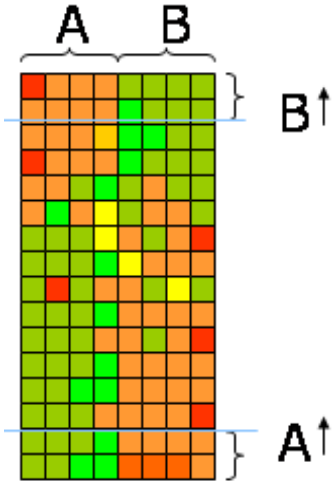
Functional Profiling from Babelomics (I)

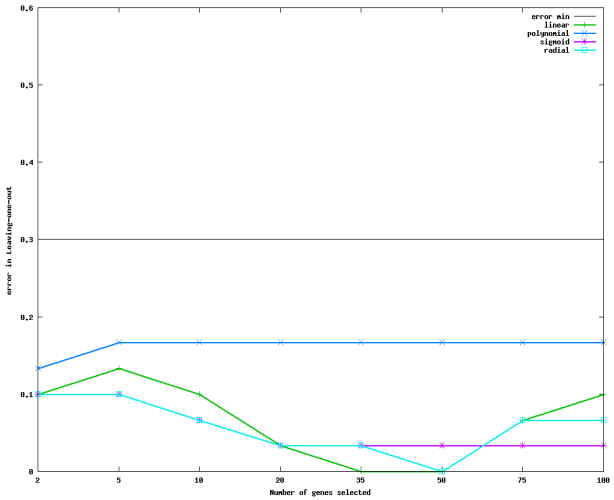
Francisco García
fgarcia@cipf.es

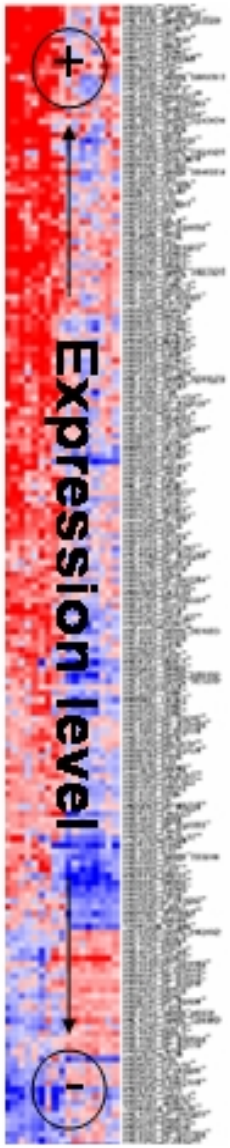
Over-representation and Gene Set Analysis

Genome-scale experiment output

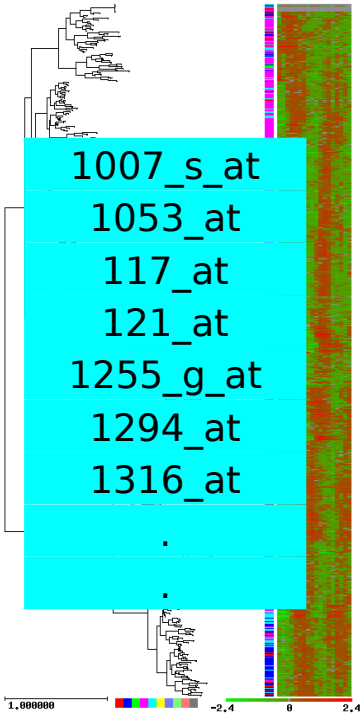


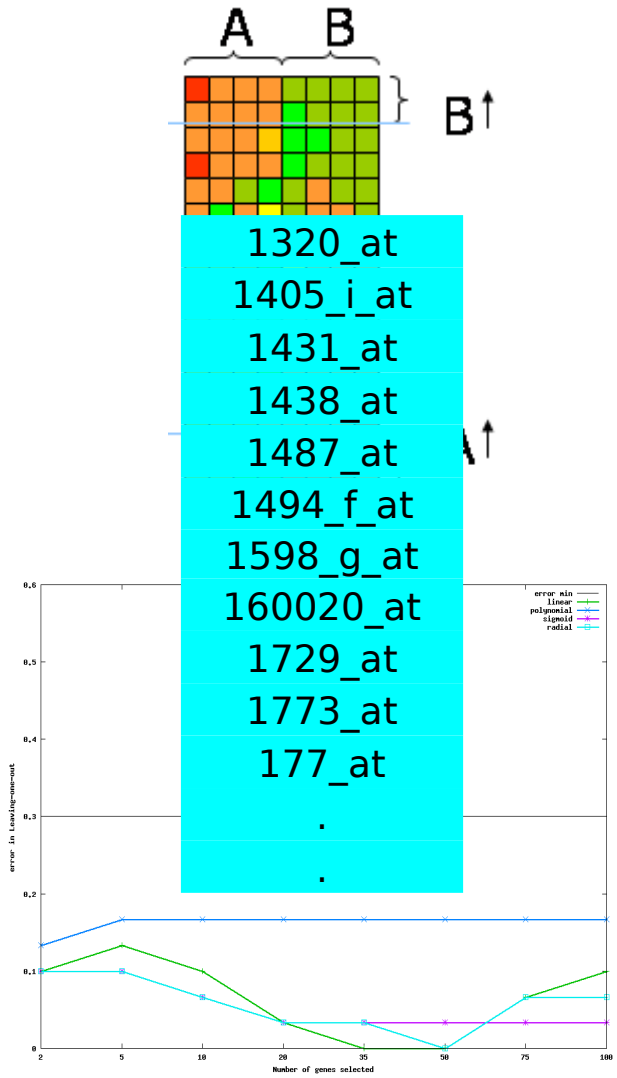


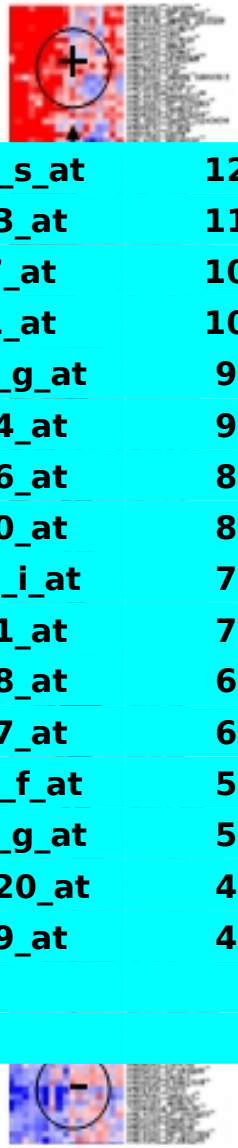




Genome-scale experiment output







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 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

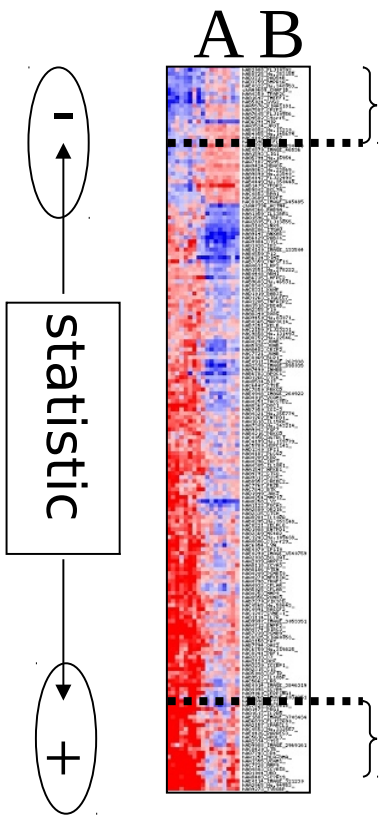
MiRNA, CisRed

Transcription Factor Binding Sites

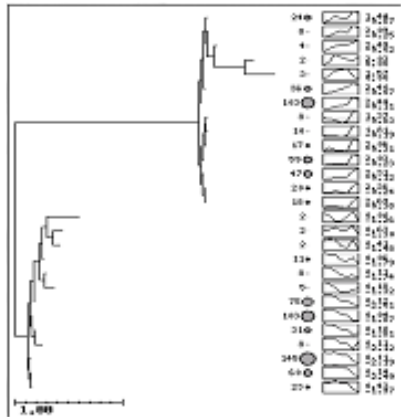
Bioentities from literature:

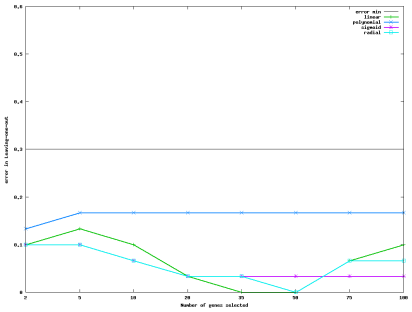
Diseases terms
Chemical terms

Over-representation analysis



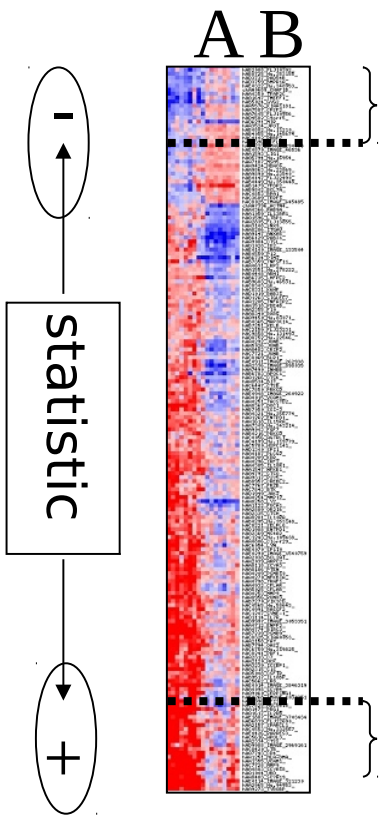
1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.





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
.
.

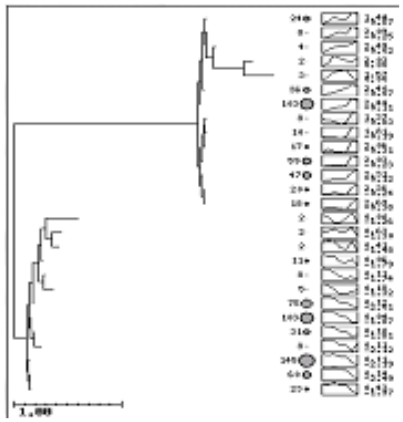
Over-representation analysis

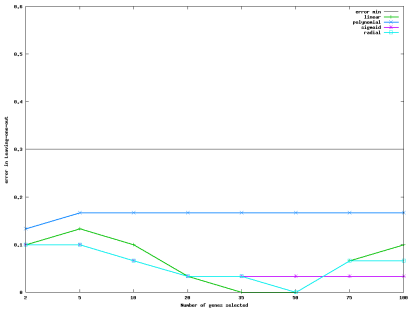


1007_s_at
1053_at
117_at
121_at
1255_g_at
1294_at
1316_at
.
.

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
.
.

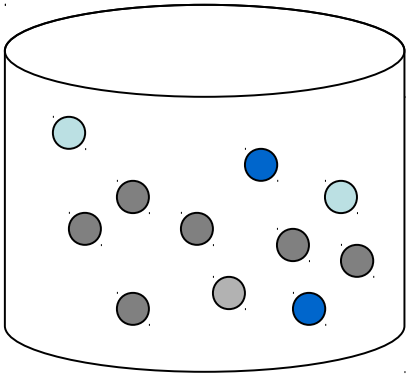
Function

2/11

Over-representation analysis

FatiGO test

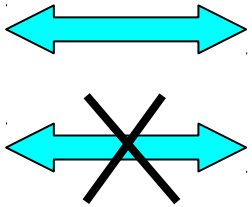
One Gene List (A)



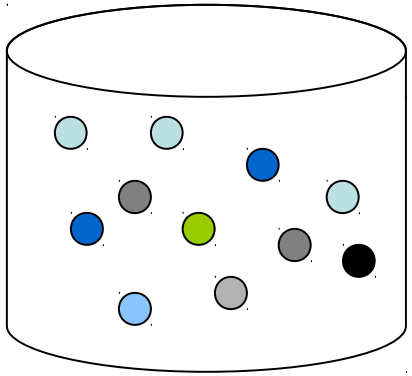
Biosynthesis 60%

Sporulation 20%

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



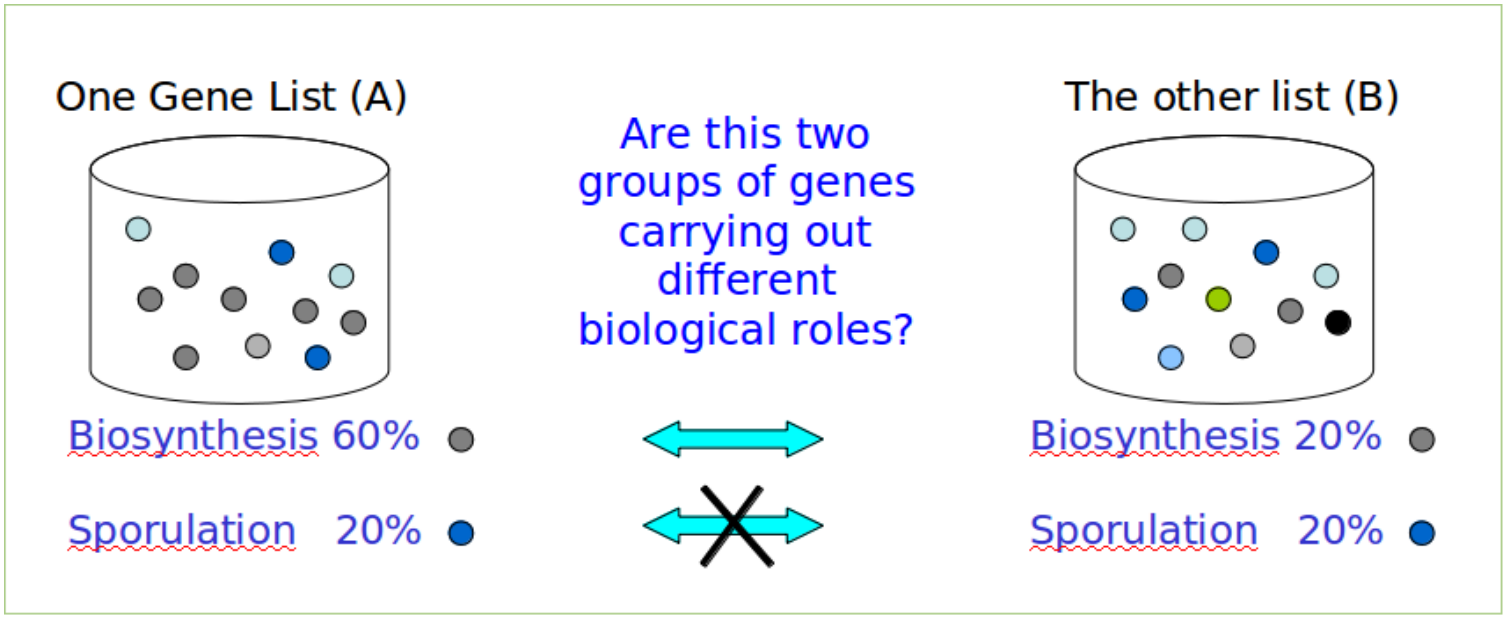
The other list (B)



Biosynthesis 20%

Sporulation 20%

Over-representation analysis

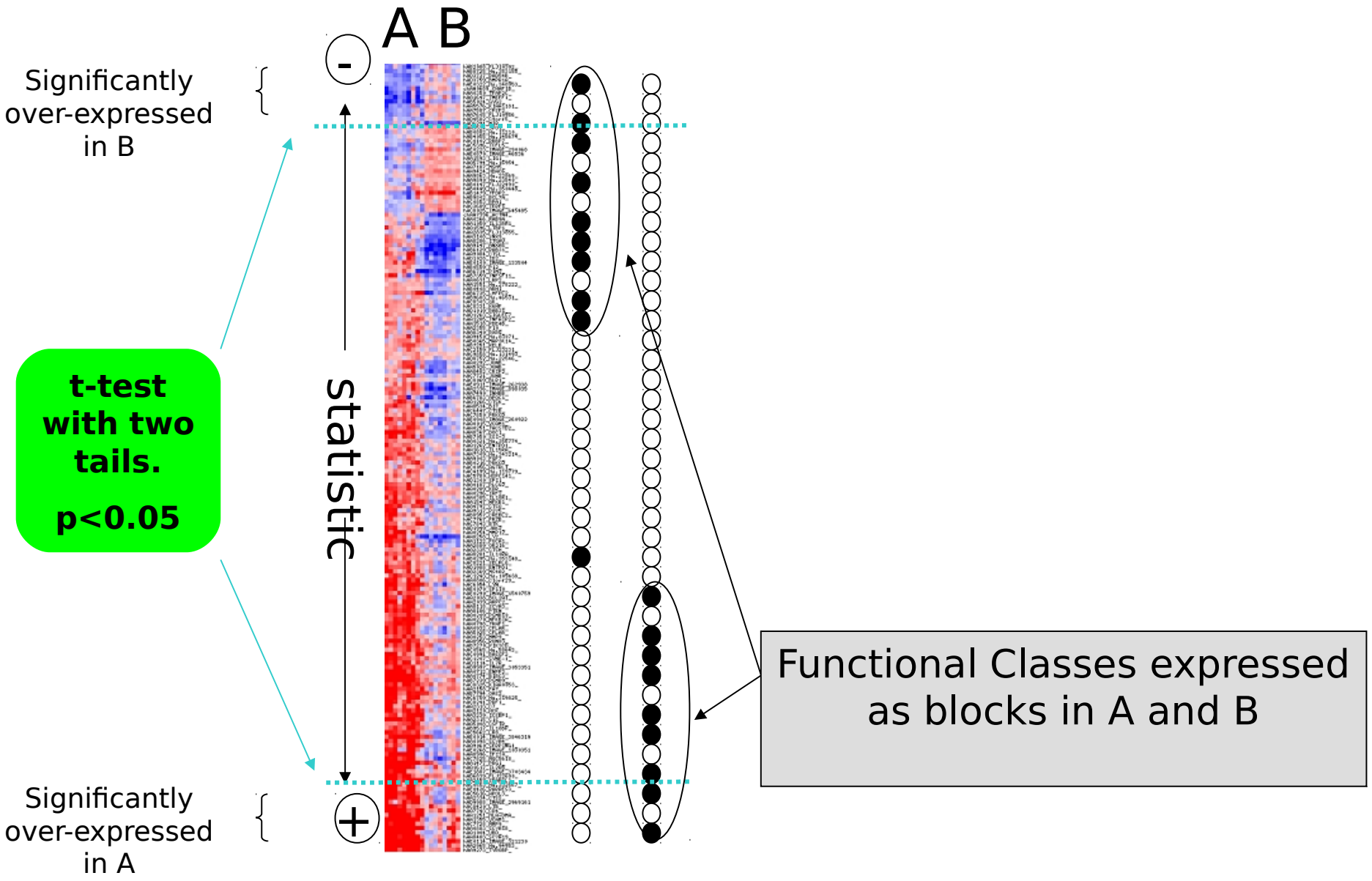


Genes in group A have significantly to do with biosynthesis, but not with sporulation.

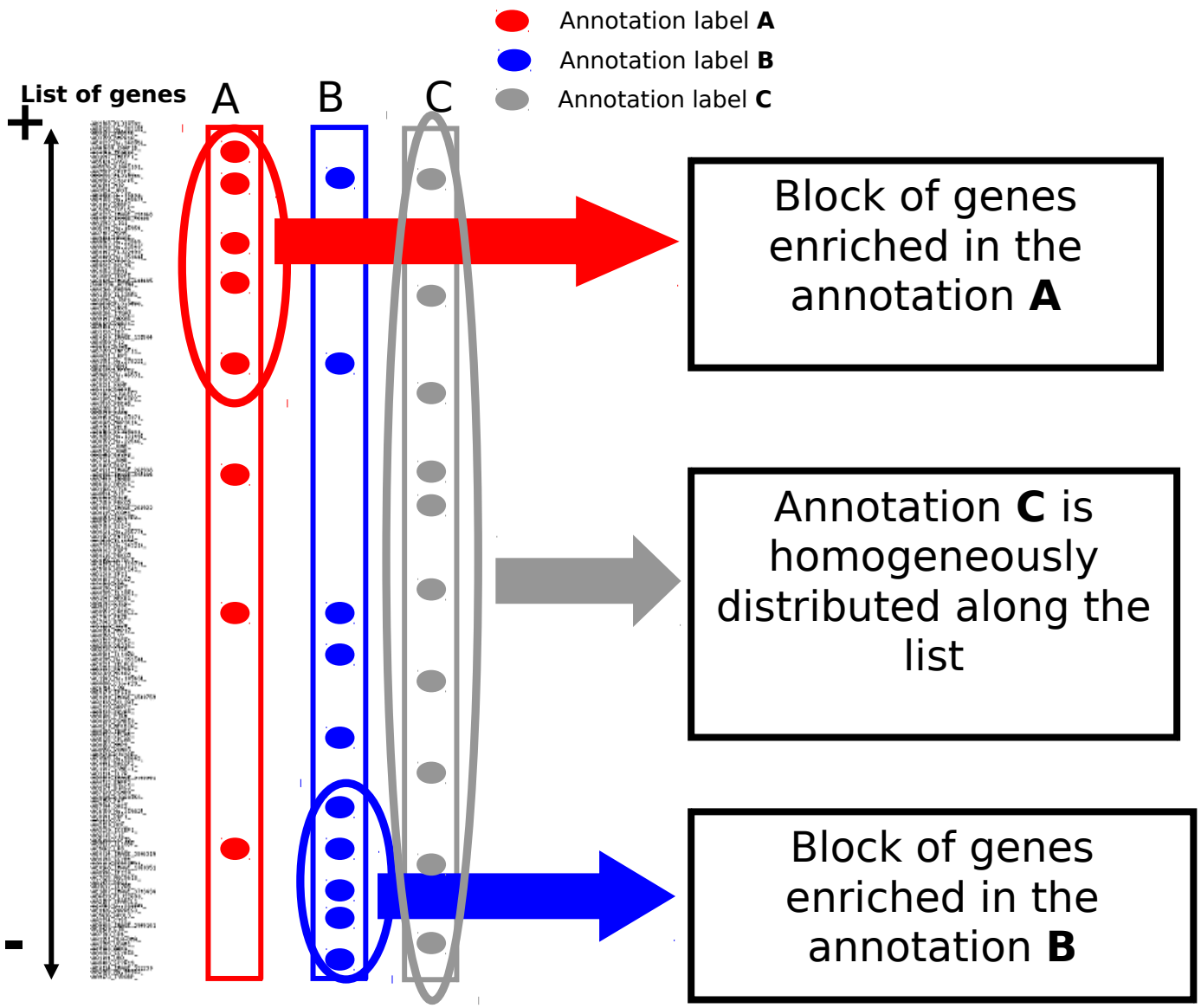
	A	B
Biosynthesis	**6**	**2**
No biosynthesis	**4**	**8**

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

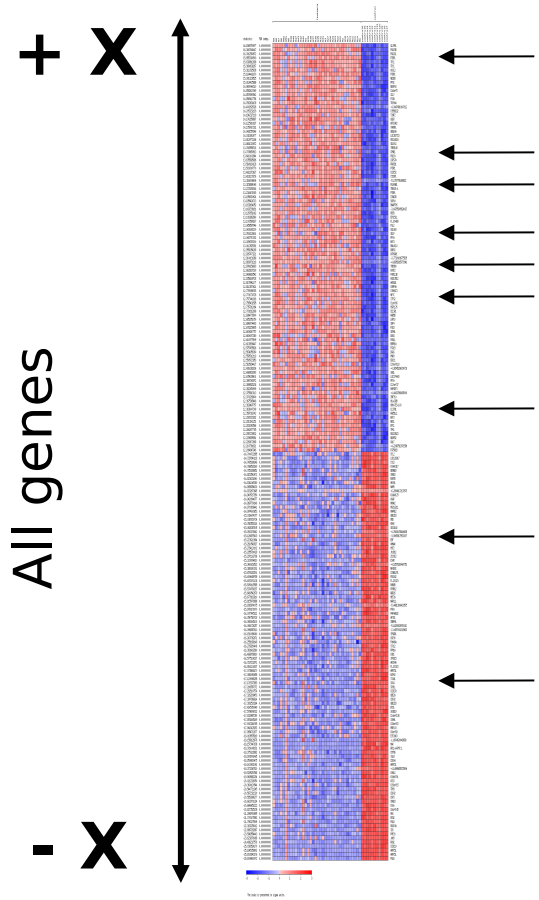
Gene Set Analysis



Gene Set Analysis



Gene Set Analysis

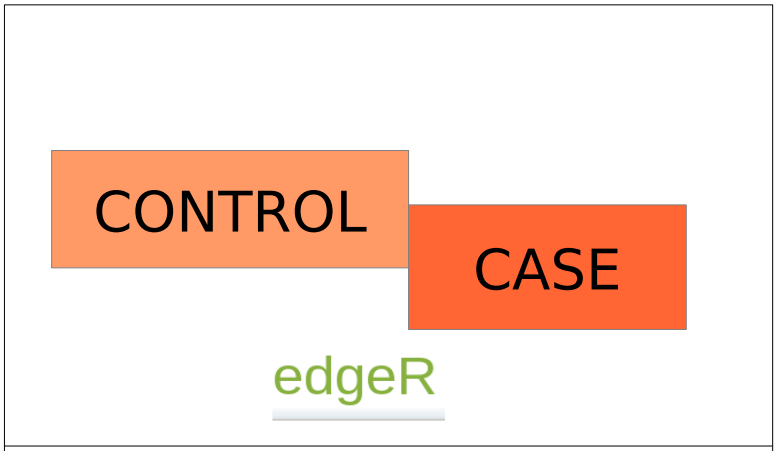


$$\ln\left(\frac{P(g\in F)}{P(g\notin F)}\right) = K + \alpha X$$

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

Gene Set Analysis for miRNAs



#TOP genes overexpressed in CANCER
#BOTTOM genes overexpressed in CONTROL

ID	statistic	fold	p.value	p.adjusted
hsa-miR-214	3.85975	0.97808	0.00013	0.00768
hsa-miR-24	3.80296	0.35442	0.00016	0.00904
hsa-miR-340	3.68651	0.65923	0.00026	0.01269
hsa-miR-186	3.68613	0.87417	0.00026	0.01269
hsa-miR-224	3.64089	0.89072	0.00030	0.01431
hsa-miR-130b	3.51088	0.70939	0.00049	0.02208
hsa-miR-31*	3.29490	1.11298	0.00107	0.04435
hsa-miR-181a*	3.28551	0.84347	0.00110	0.04435
hsa-miR-135b	3.27717	0.52963	0.00113	0.04435
hsa-miR-31	3.26077	1.19828	0.00120	0.04451
hsa-miR-181c	3.22127	0.93645	0.00137	0.04776
hsa-miR-659	3.17680	0.79468	0.00160	0.05354
hsa-miR-376a	3.09005	1.05265	0.00213	0.06671
hsa-miR-765	3.01381	0.93176	0.00273	0.07580
hsa-miR-601	3.01244	0.84301	0.00274	0.07580
hsa-miR-125a-3p	2.92902	1.01726	0.00358	0.09341
hsa-miR-1224-5p	2.88339	0.66179	0.00413	0.10486
hsa-miR-210	2.82385	0.66215	0.00497	0.11956

Step 1
Differential miRNA Expression Analysis

Gene Set Analysis for miRNAs

$$t = -sign(statistic) \cdot log(pvalue)$$

TargetScanHuman
Prediction of microRNA targets
Release 6.2: June 2012

$$\Delta g_i = \kappa \sum_{j \in R_i} t_j$$

where Δg_i represents the increment in the inhibition of gene i, t_j accounts for the differential expression of miRNA j, R_i is the set of microRNAs which target gene i and κ is just a proportionality constant.

Step 2
Transferring information from miRNA to mRNA

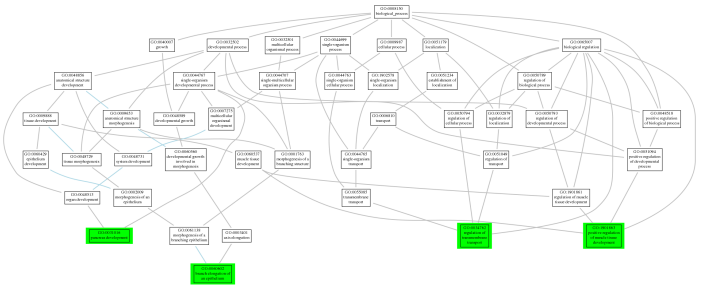
Gene Set Analysis for miRNAs

$$\log \frac{P(g \in F)}{P(g \notin F)} = \kappa + \alpha T$$

Esophageal carcinoma [ESCA]

ESCA : bp

	name	N	lor	pval	padj	pattern
GO:0060602	branch elongation of an epithelium	15	-1.1892876	3.011973e-08	0.001061682	-1
GO:1901863	positive regulation of muscle tissue development	15	-1.1171092	1.350276e-06	0.023797750	-1
GO:0034762	regulation of transmembrane transport	128	-0.4143974	3.324984e-06	0.033787909	-1
GO:0031016	pancreas development	62	-0.6092071	3.834228e-06	0.033787909	-1
GO:0034765	regulation of ion transmembrane transport	123	-0.4019287	9.932750e-06	0.070023351	0
GO:2000027	regulation of organ morphogenesis	135	-0.3792065	1.372379e-05	0.071409924	0
GO:0032409	regulation of transporter activity	103	-0.4302775	1.418121e-05	0.071409924	0
GO:0061138	morphogenesis of a branching epithelium	172	-0.3303407	2.033439e-05	0.083815362	0
GO:0022898	regulation of transmembrane transporter activity	94	-0.4409666	2.140044e-05	0.083815362	0
GO:0007440	foregut morphogenesis	11	-1.1498006	2.427932e-05	0.085581501	0



Step 3

GSA from logistic regression models

Hands on

Babelomics 5

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

Functional / FatiGO
Functional / Logistic Model

Online examples

Functional Profiling from Babelomics (II)

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Computational • Genomics

Francisco García
fgarcia@cipf.es

Network Enrichment

Protein-Protein Interactions (PPI)

- PPIs are a central point at almost every level of cell function:
  - Structure of subcellular organelles (structural proteins)
  - Packing the chromatin (histones)
  - Protein modifications (kinases)

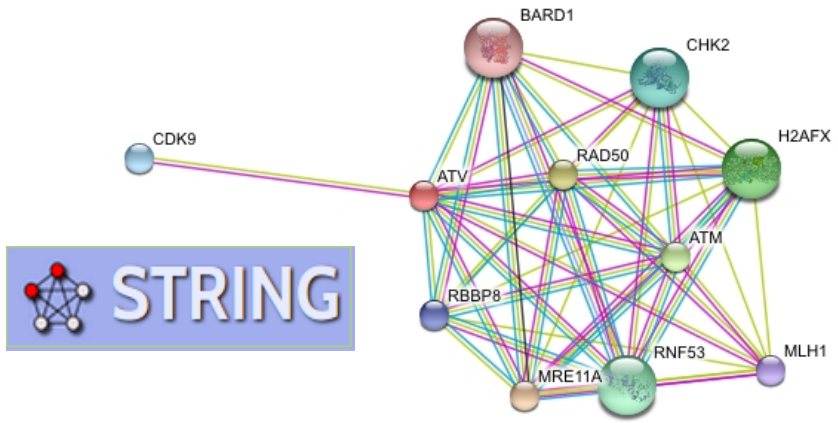
- Retrieving information about a **single protein**....

5/277 Interacting proteins for BRCA1 (ENSP00000350283)

Interactant		Interacti
GeneCard	External ID(s)	
NBN	ENSP00000265433	STRING (
TOPBP1	ENSP00000260810	STRING (score=.999)
UBA1	ENSP00000338413	STRING (score=.999)
UBE2D1	ENSP00000185885	STRING (score=.999)
GADD45A	ENSP00000360025	STRING (score=.998)

About this table

GeneCards®

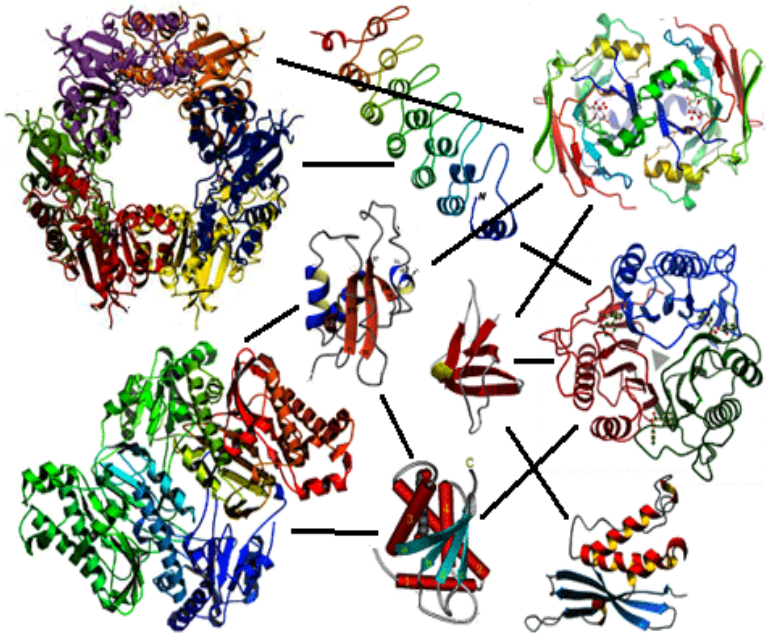


Protein-Protein Interactions (PPI)

- 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**?

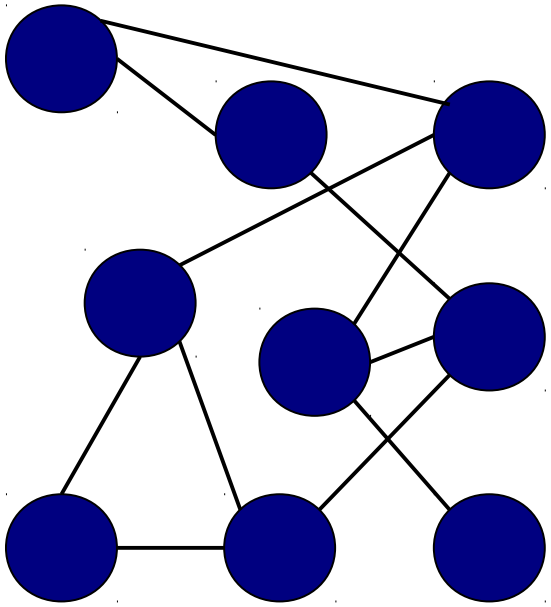
Graph Theory

Set of proteins interacting



Nodes = proteins
Edges = interaction events

Undirected graph



structured data

Graph Theory

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

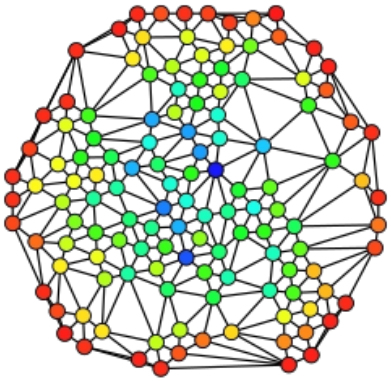
- **Degree (connectivity or connections)**: number of edges connected to a node. Nodes with high degree are called hubs.

- **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

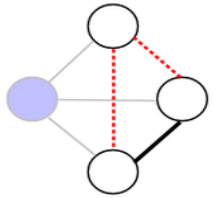


Graph Theory

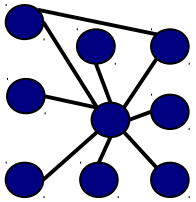
- Clustering coefficient (of a node): A measure of how interconnected the neighbours of that node are. Proportion of links between the nodes within its neighbourhood divided by the number of links that could possibly exist between them.

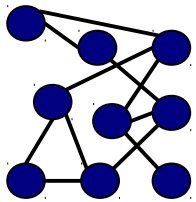
$$C_i = \frac{2e_i}{n_i(n_i - 1)}$$

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.

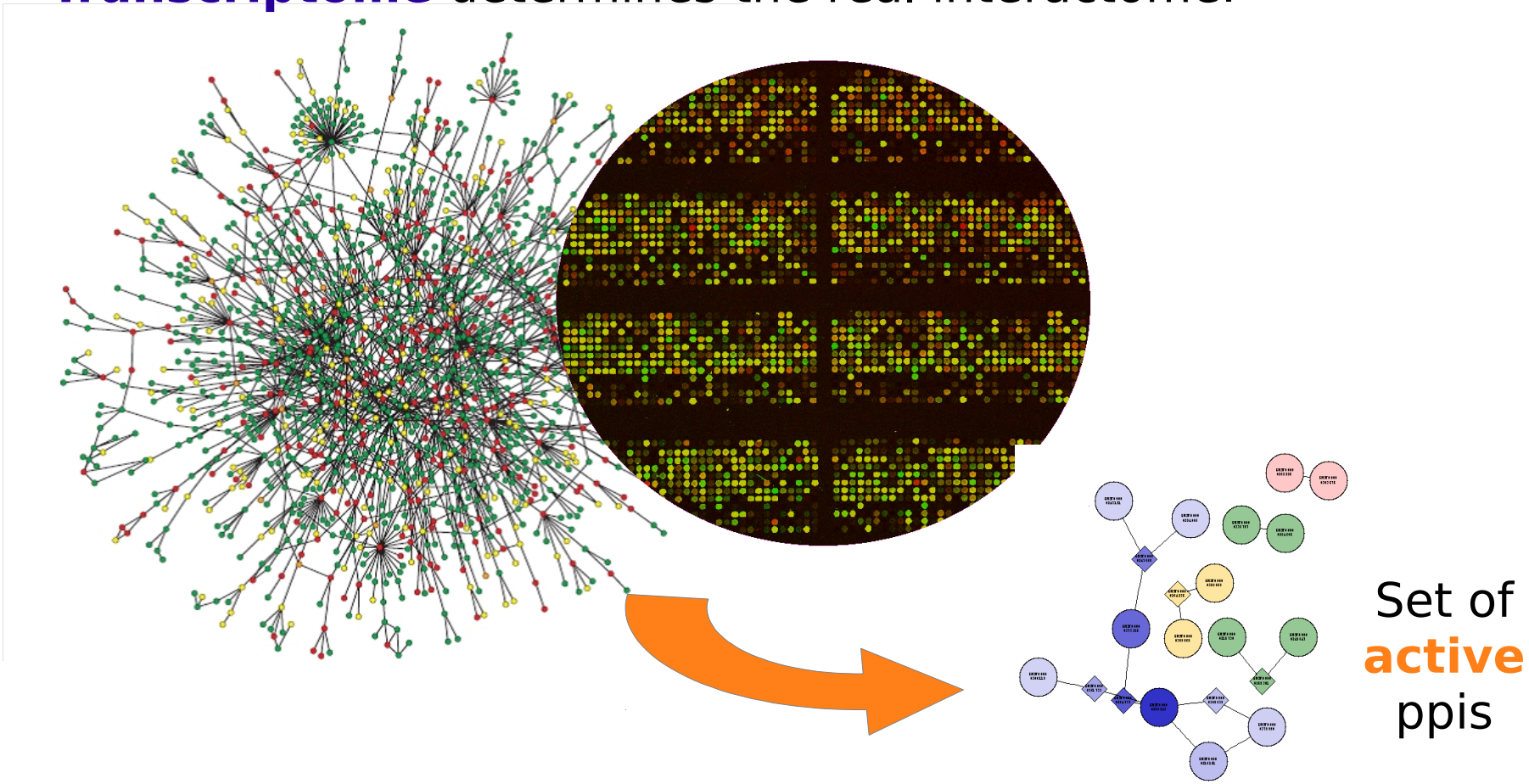
Component. A group of nodes connected among them.

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.



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**

Babelomics 5

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

Functional / Network Enrichment: SNOW

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

Outline

1) Introduction to NGS Data Analysis in Transcriptomic Studies

2) RNA-Seq and miRNA-Seq Data Analysis

3) Functional Profiling

4) **Omic Data Integration**

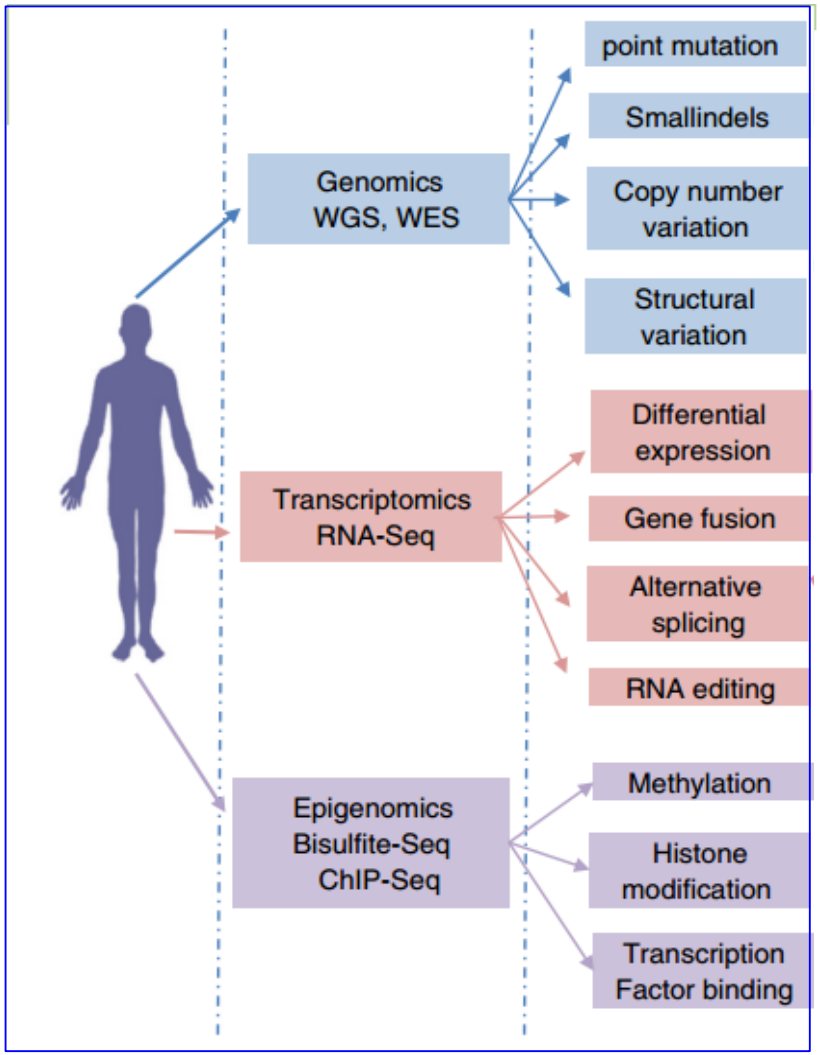
Omics Data Integration from a Systems Biology perspective

PRINCIPE FELIPE
CENTRO DE INVESTIGACION
Computational · Genomics

Francisco García
fgarcia@cipf.es

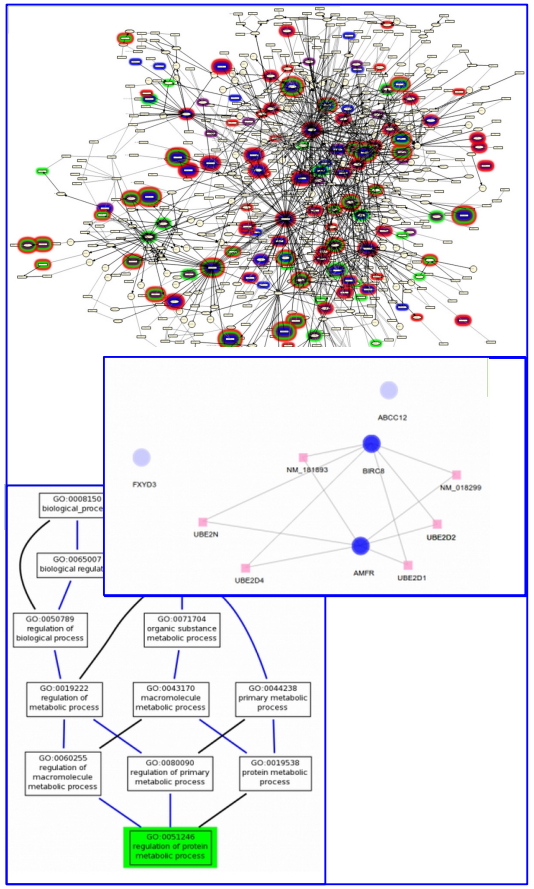
Omics Data Integration

Patient Technologies Data Analysis



Integration and interpretation

Molecular and clinical model



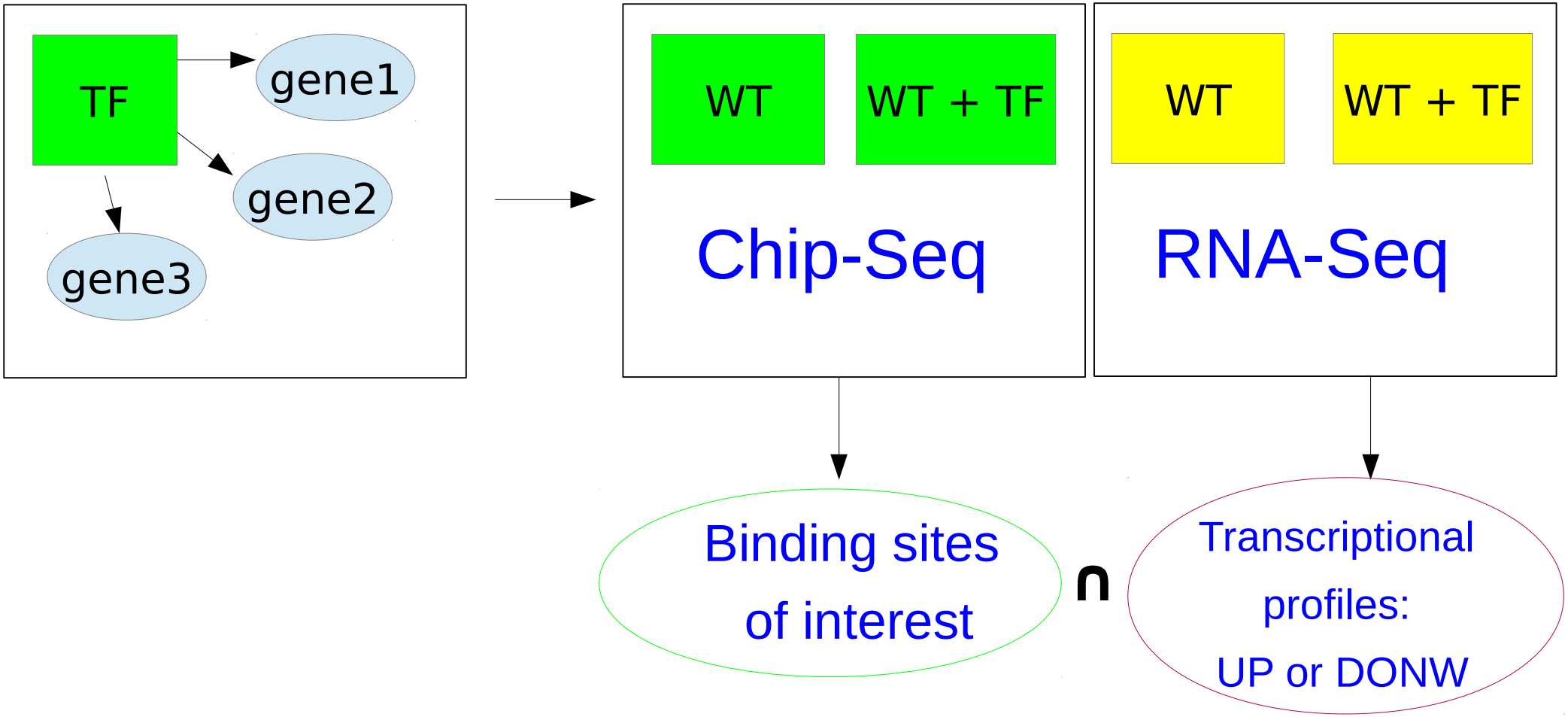
Omics Data Integration

Different strategies:

1) Ad-hoc approaches
2) Multidimensional Gene Set Analysis
3) Functional Meta-Analysis
4) PATHiVAR: a web tool to integrate transcriptomics and genomics results

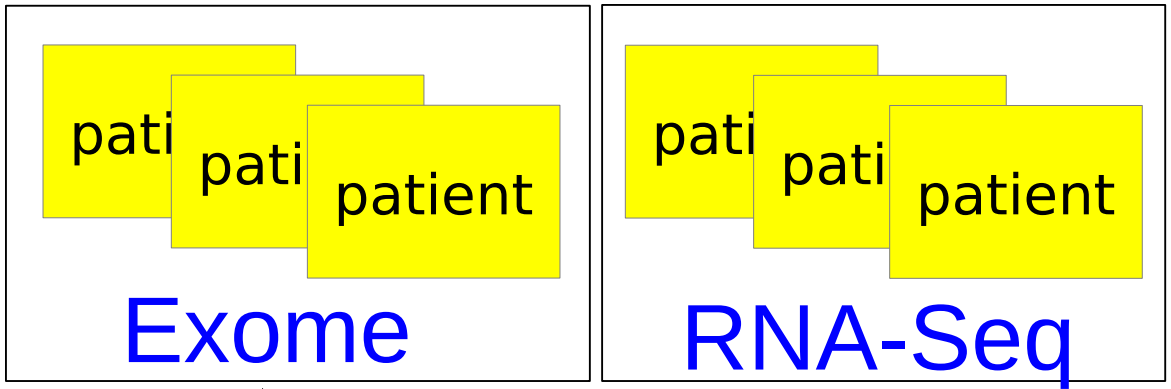
Ad-hoc approaches (1)

Chip-Seq & RNA-Seq



Ad-hoc approaches (2)

Exome & RNA-Seq



→ Intronic causative mutations

Exonic causative mutations

Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders

Marta Codina-Solà[123], Benjamín Rodríguez-Santiago[4], Aïda Homs[123], Javier Santoyo[5], Maria Rigau[1], Gemma Aznar-Laín[6], Miguel del Campo[137], Blanca Gener[8], Elisabeth Gabau[9], María Pilar Botella[10], Armand Gutiérrez-Arumí[123], Guillermo Antiñolo[1135], Luis Alberto Pérez-Jurado[123]* and Ivon Cuscó[123]*

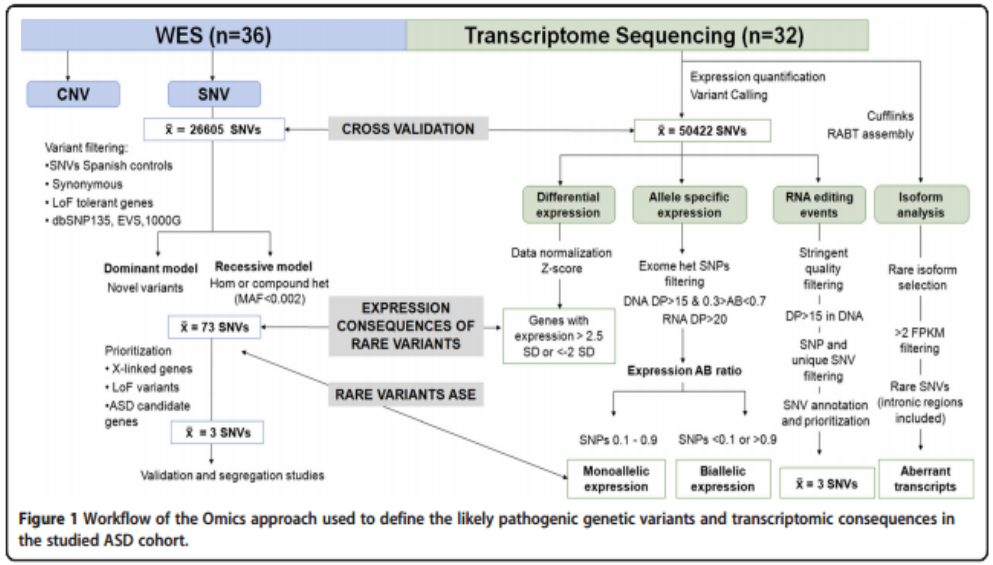


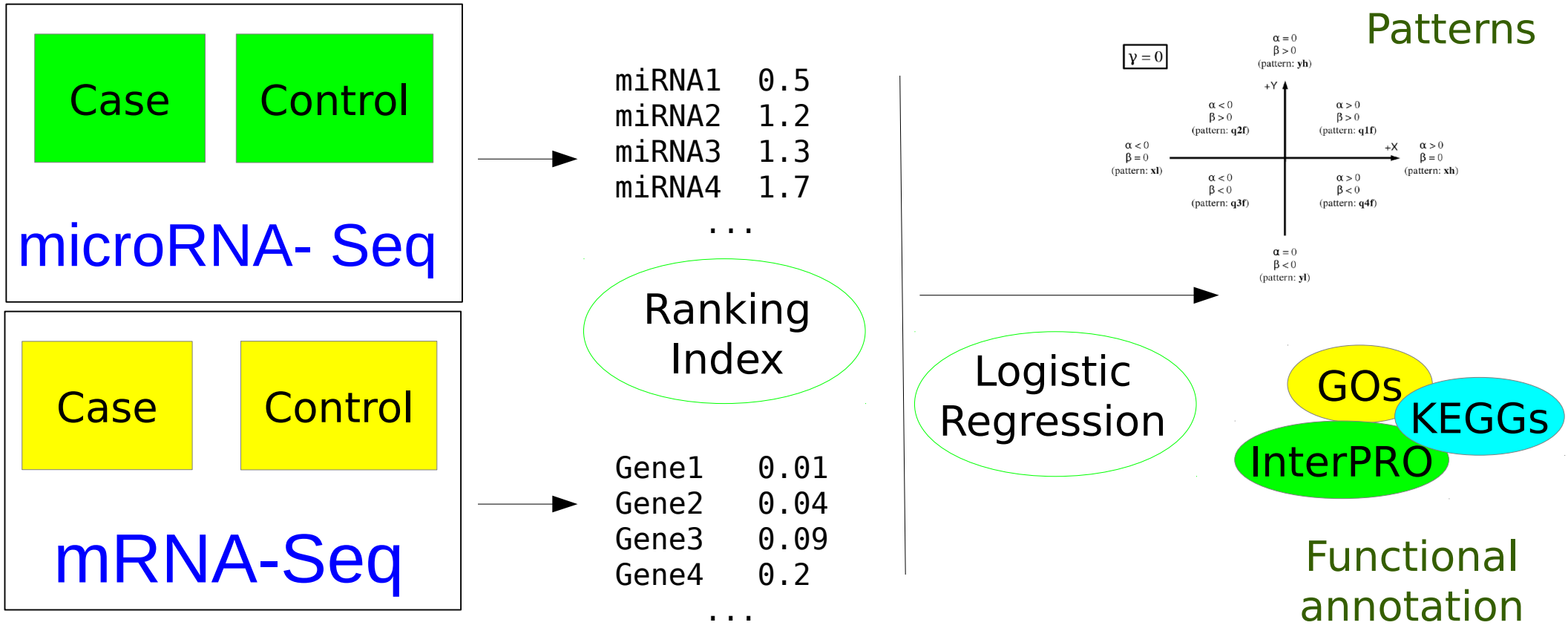
Figure 1 Workflow of the Omics approach used to define the likely pathogenic genetic variants and transcriptomic consequences in the studied ASD cohort.

Strategies

Omics Data Integration

Multidimensional Gene Set Analysis

MicroRNA-Seq & mRNA-Seq

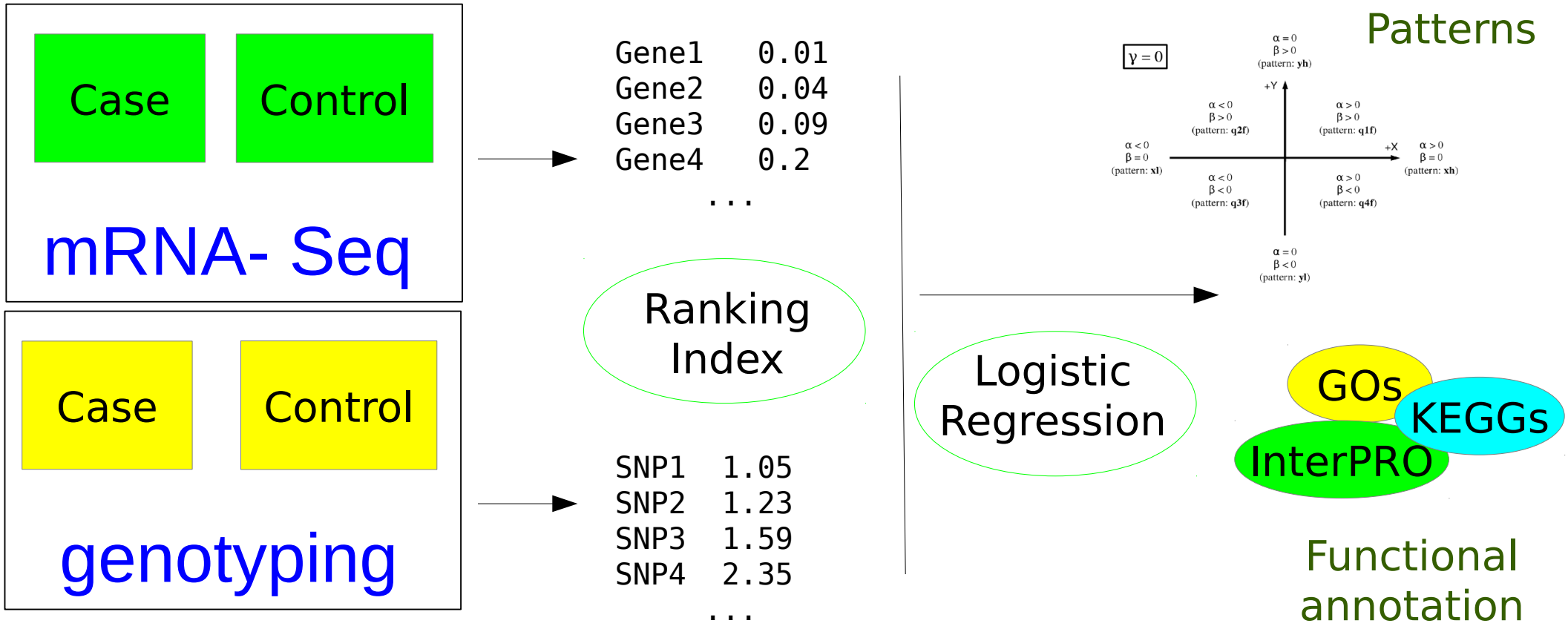


Strategies

Omics Data Integration

Multidimensional Gene Set Analysis

mRNA-Seq & genotyping association

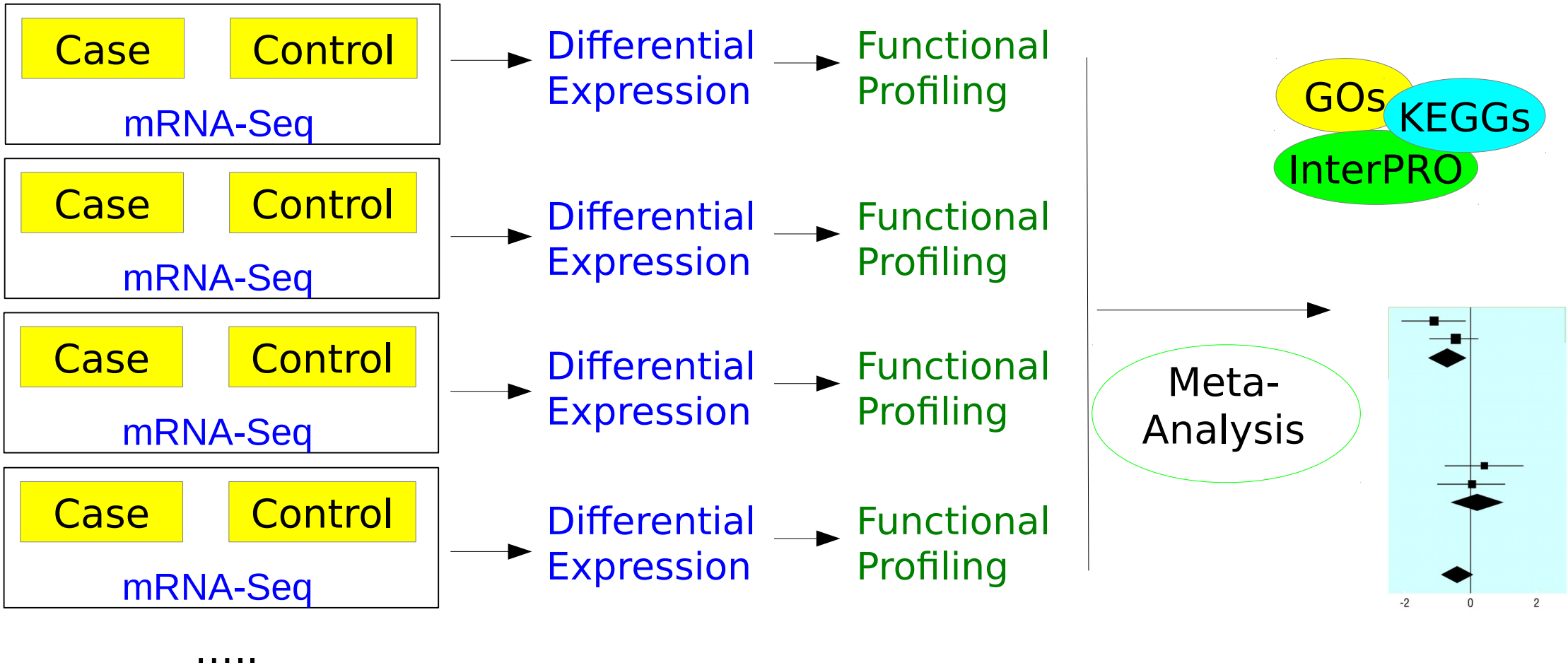


Strategies

Omics Data Integration

Functional Meta-Analysis

N mRNA-Seq studies



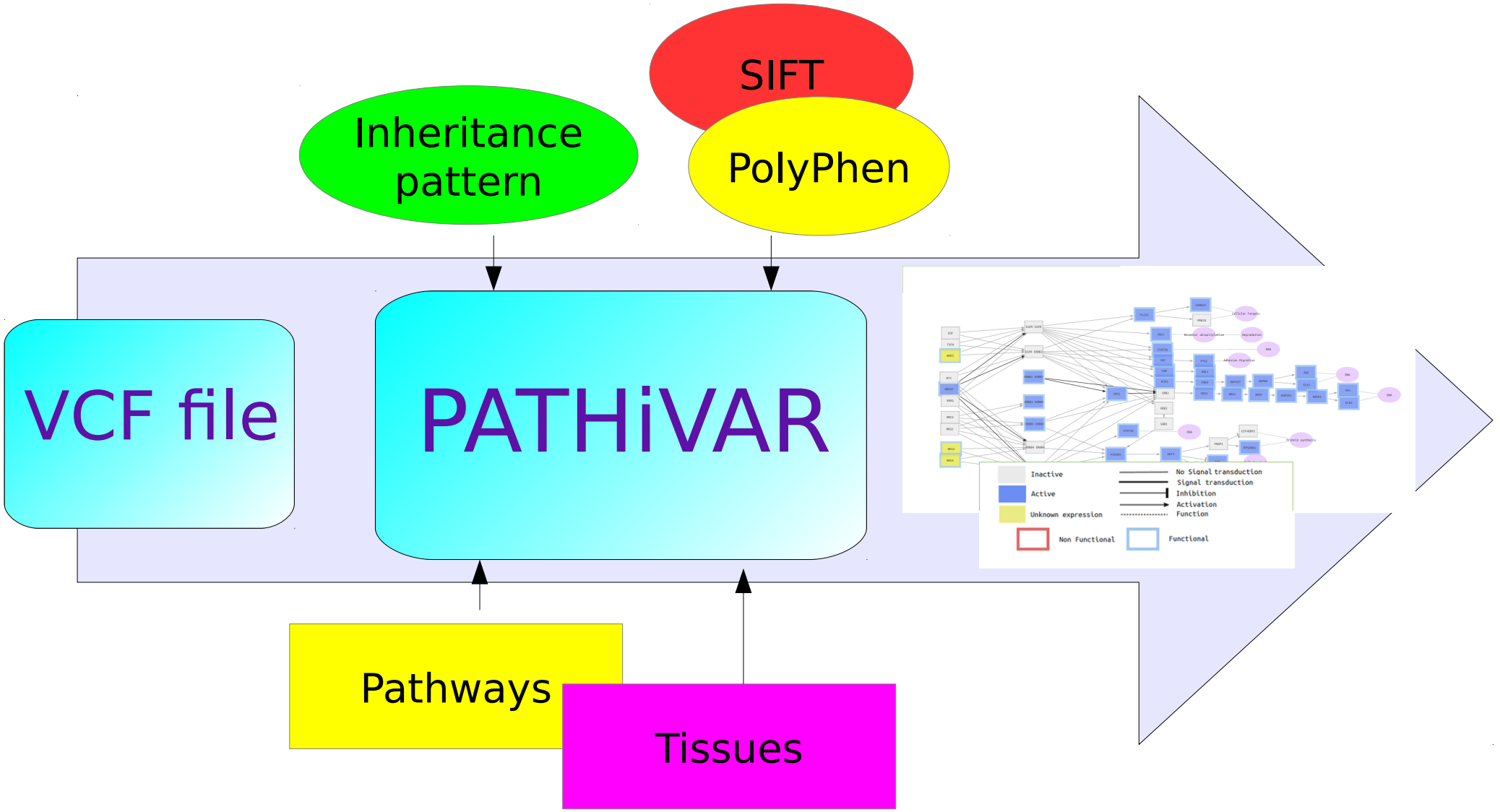
PATHiVAR: mutations and expression

- **PATHiVAR** estimates the functional impact that mutations have over the human signalling network.

- **PATHiVAR**:
  - Analyses VCF files
  - Extract the deleterious mutations
  - Locate them over the signalling pathways in the selected tissue (with the appropriate expression pattern)
  - Provide a comprehensive, graphic and interactive view of the predicted signal transduction probabilities across the different signalling pathways.

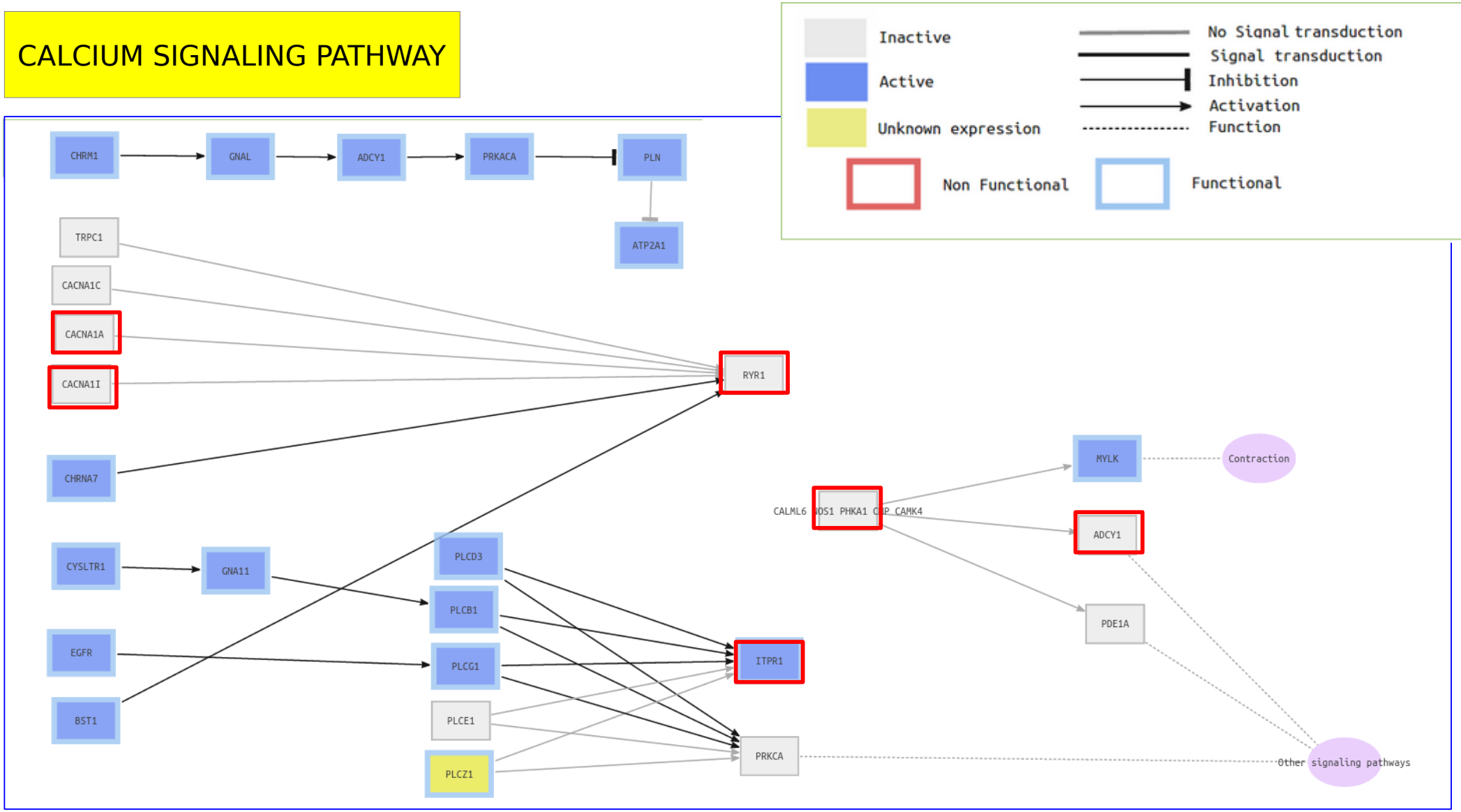
http://pathivar.babelomics.org/

How does PATHiVARK work?



PATHiVAR

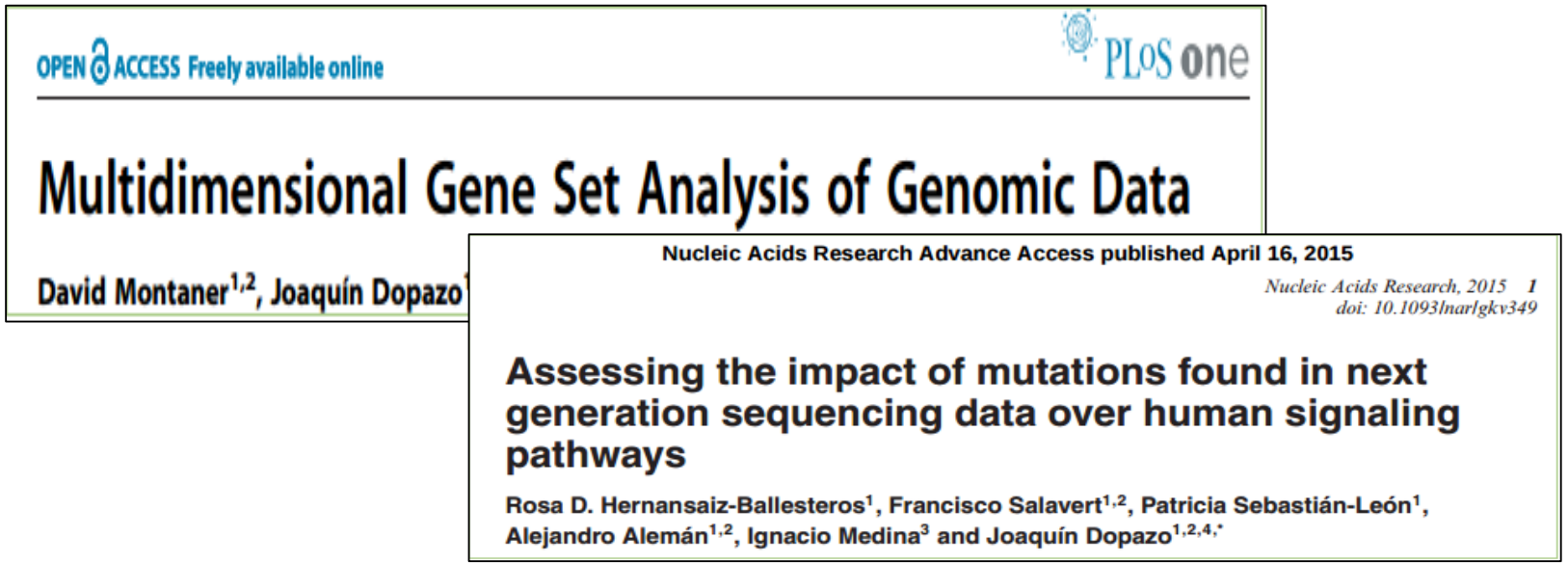
CALCIUM SIGNALING PATHWAY



Strategies

PATHiVAR

More information



PATHiVAR tutorial:

http://pathivar.babelomics.org/