From gene expression modeling to gene network to investigate Arabidopsis thaliana stress response

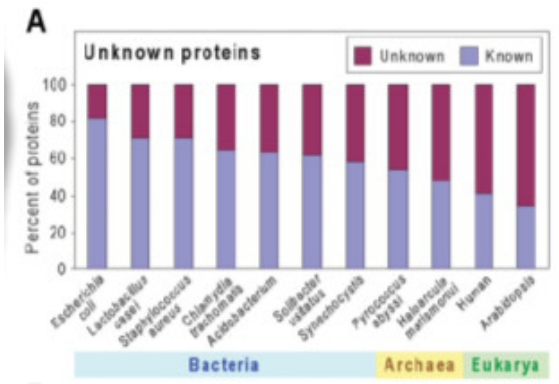
M.-L. Martin-Magniette[1,2] & E. Delannoy[1]

1- Plant Science Institut of Paris-Saclay (IPS2)

2- Applied Mathematics and Informatics Unit at AgroParisTech

Functional annotation

Definition or prediction of the gene functions and of the relationship between them



- Between 20% and 40% of the predicted genes have no assigned function (Hanson *et al*, 2009)
- For *Arabidopsis thaliana*, only 16% of the genes have a validated function

Orphan genes

- Defined has genes without homologs with a known function (Fukushi and Nishikawa, 2003)
- Usually discarded of the published studies
- 5015 orphan genes in *A. thaliana* (Zaag *et al*, 2015)

by sequence similarity

Based on a comparison of protein sequences
to identify structural similarities

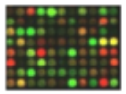
Nevertheless

- A high similarity does not guarantee a functional similarity (Tian *et al*, 2003)
- Some sequences with a low similarity may share a same function (Galperin *et al*, 1998)
- Protein sequence comparison gives information about the biochimical function (Nehrt *et al*, 2011)

by omics analysis

Based on guilt by association studies
by identification of genes having similar features at the molecular level

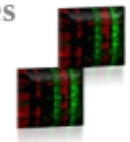
Quantitative data



Contextual data

coexpression

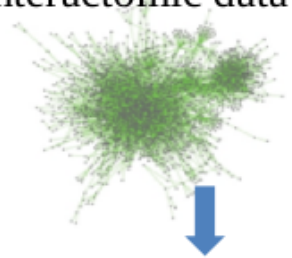
Groups of genes having a similar expression profile



The most extensive resources

Coexpression usually measured a correlation value

Interactomic data



Partners of interactions

Absolute data

10% of the interactome has been studied

Arabidopsis Interactome Mapping Consortium (2011)

High rate of false negative 84%

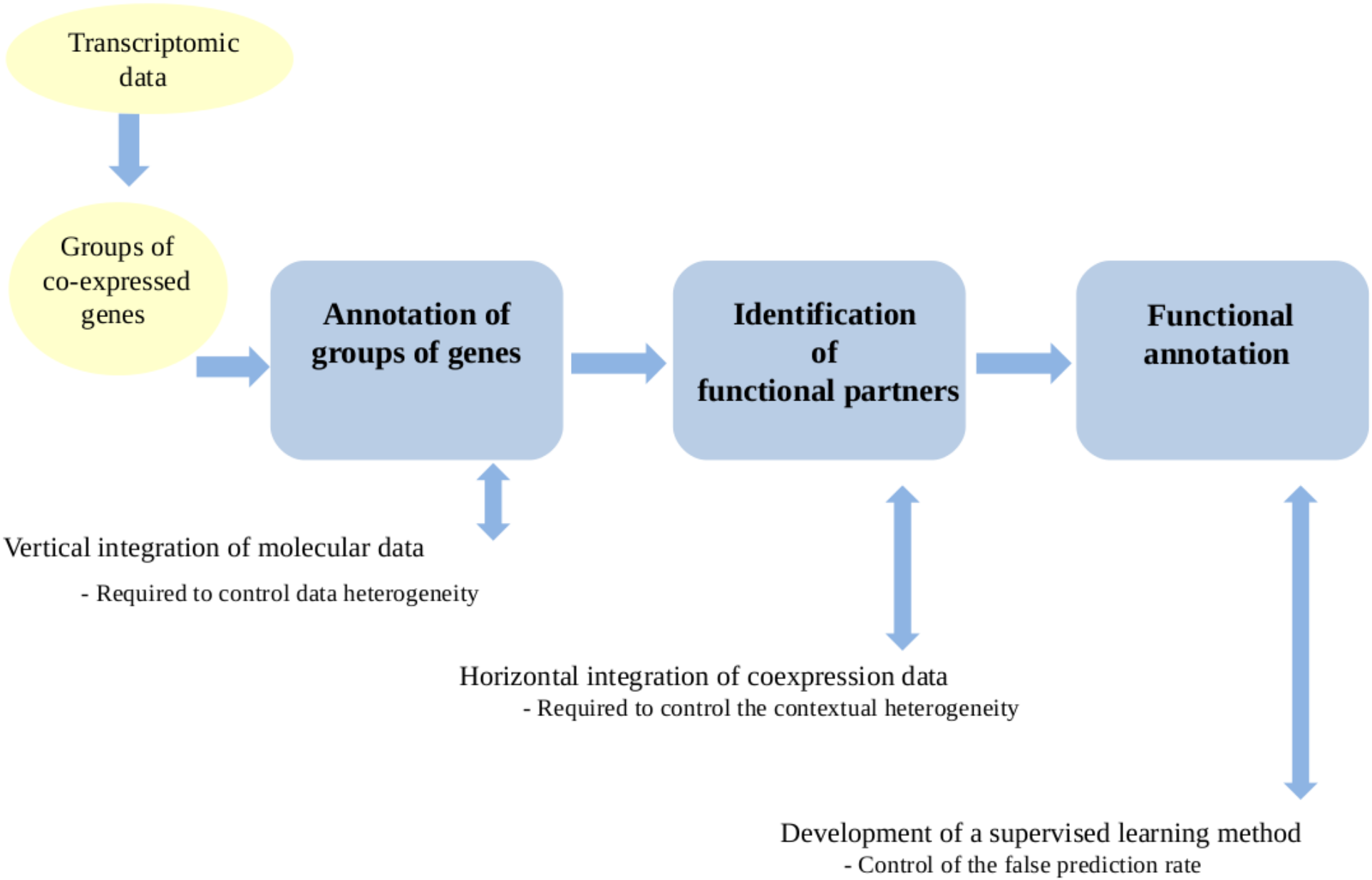
by omics integration

Integrating various resources of omics data
improves the success of prediction (Radiovojac *et al*, 2013)

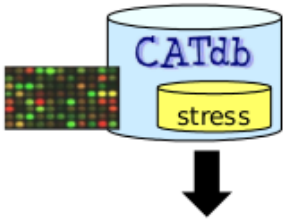
But various sources of heterogeneity exist

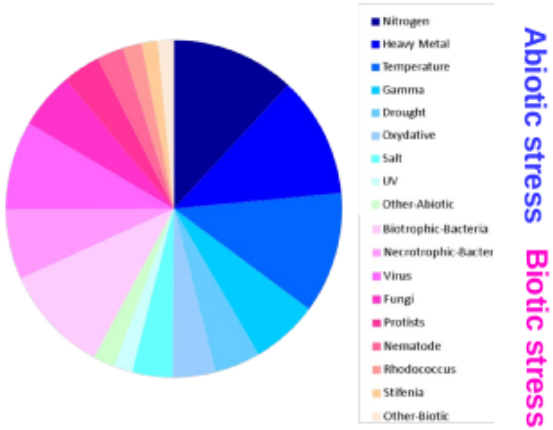
- Data are qualitative or quantitative
- Available information describes the biological entities or their relationships
- Observations are obtained with various techniques
- Various semantic frameworks are used

From Gene Expression Modeling to Networks



A dedicated transcriptomic dataset



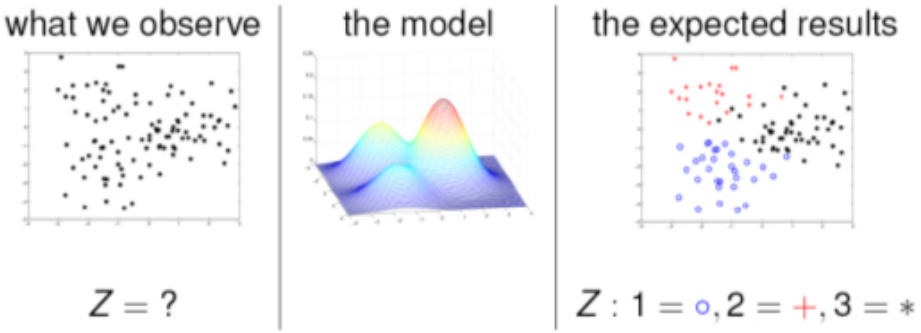


- 387 transcriptomic comparisons in dye-swap dedicated to stress
- 2/3 describe abiotic stresses and 1/3 biotic stresses
- All the data were generated on the same transcriptomic platform with the same protocol

First results

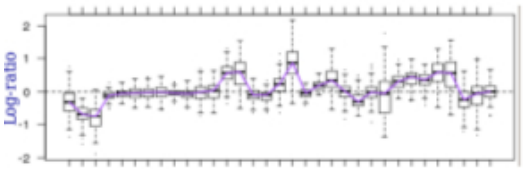
- Based on differential analyses, 60% of the genes coding proteins have their transcription impacted directly or not by a stress
- Large overlap of impacted genes between biotic and abiotic stress

Coexpression study using mixture model



Matrix by stress
{ genes x log-ratios}

Gaussian mixture



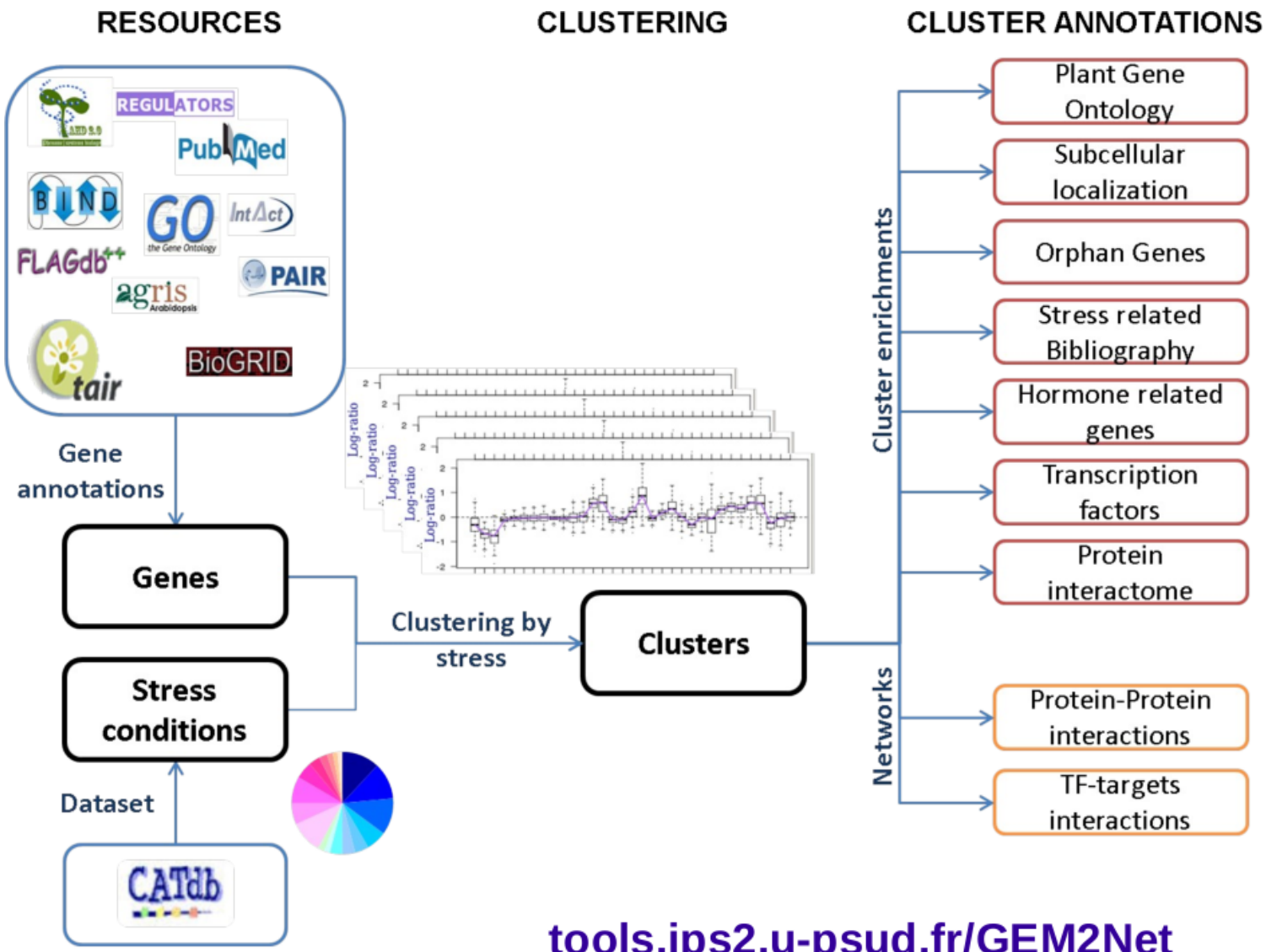
Data-driven method

- number of cluster chosen by BIC
- gene classification based on the conditional probabilities

Stress category	Gene_nb	Cluster_nb
Nitrogen	13 495	59
Temperature	11 365	34
Drought	8 143	34
Salt	5 729	30
Heavy metal	10 617	57
UV	7 894	37
Gamma	5 350	32
Oxydative stress	10 127	52
Nectrophic bacteria	11 220	50
Biotrophic bacteria	12 023	56
Fungi	9 773	51
Rhodococcus	1 900	13
Oomycete	5 508	31
Nematode	7 413	27
Stifenia	1 525	17
Virus	11 832	54

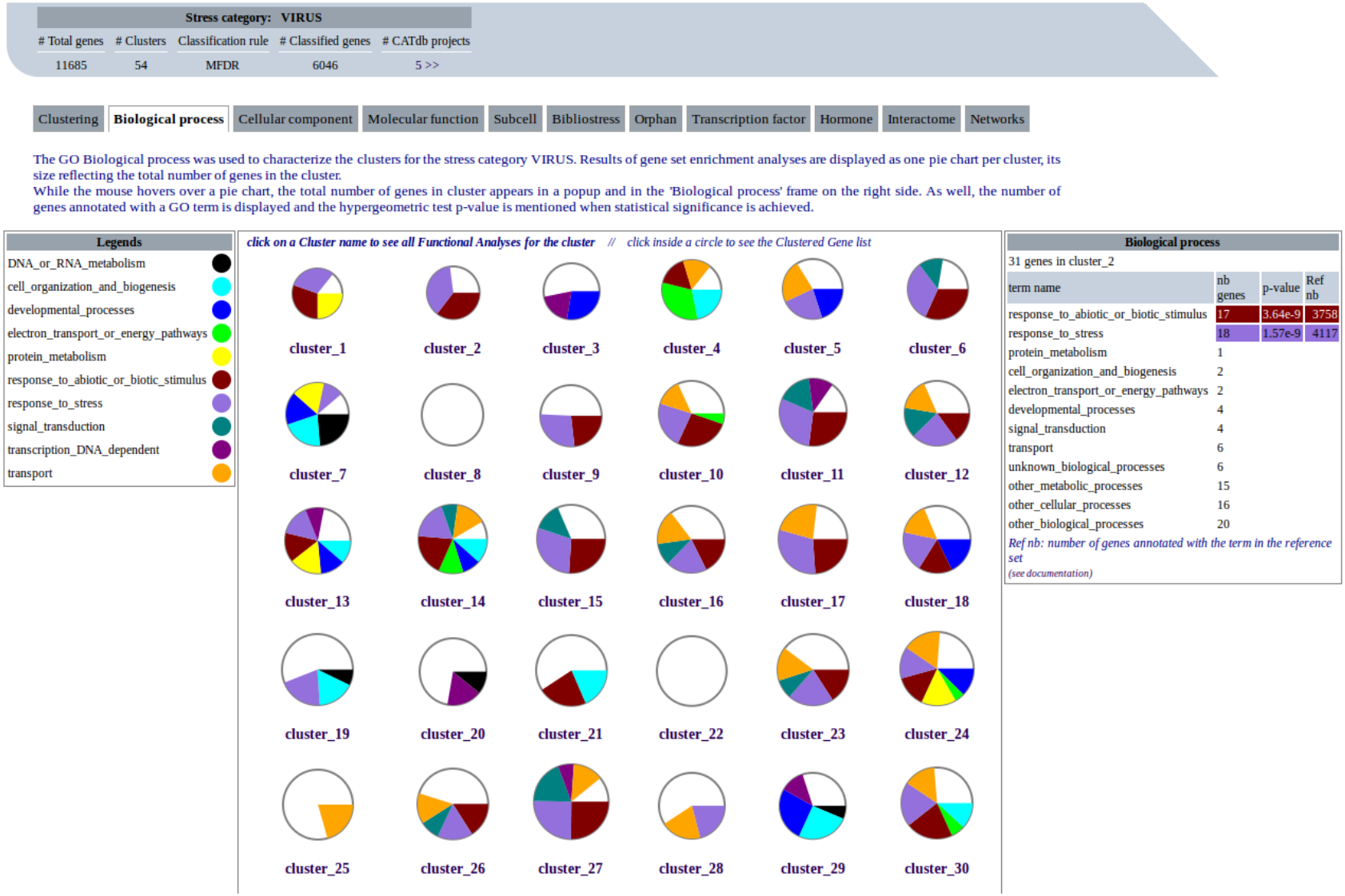
~ 700 clusters of co-expression

Annotation of coexpressed genes



tools.ips2.u-psud.fr/GEM2Net

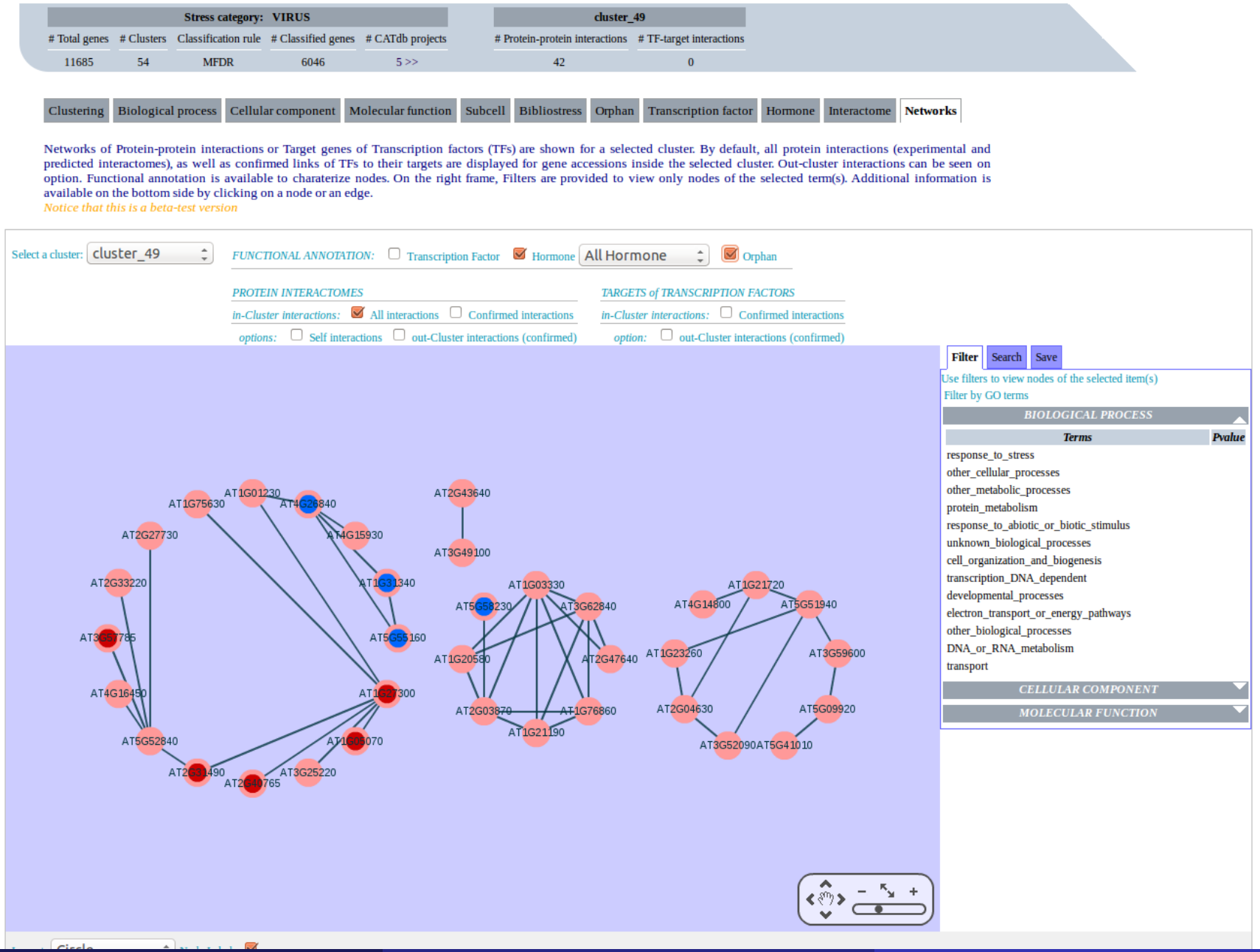
Visualisation by type of resource



Pie size proportional to cluster size

Colors indicate biological biases

Visualisation of interactions



Vertical integration

Results

- Numerous enrichments
- Overlap with TF regulations and PPI

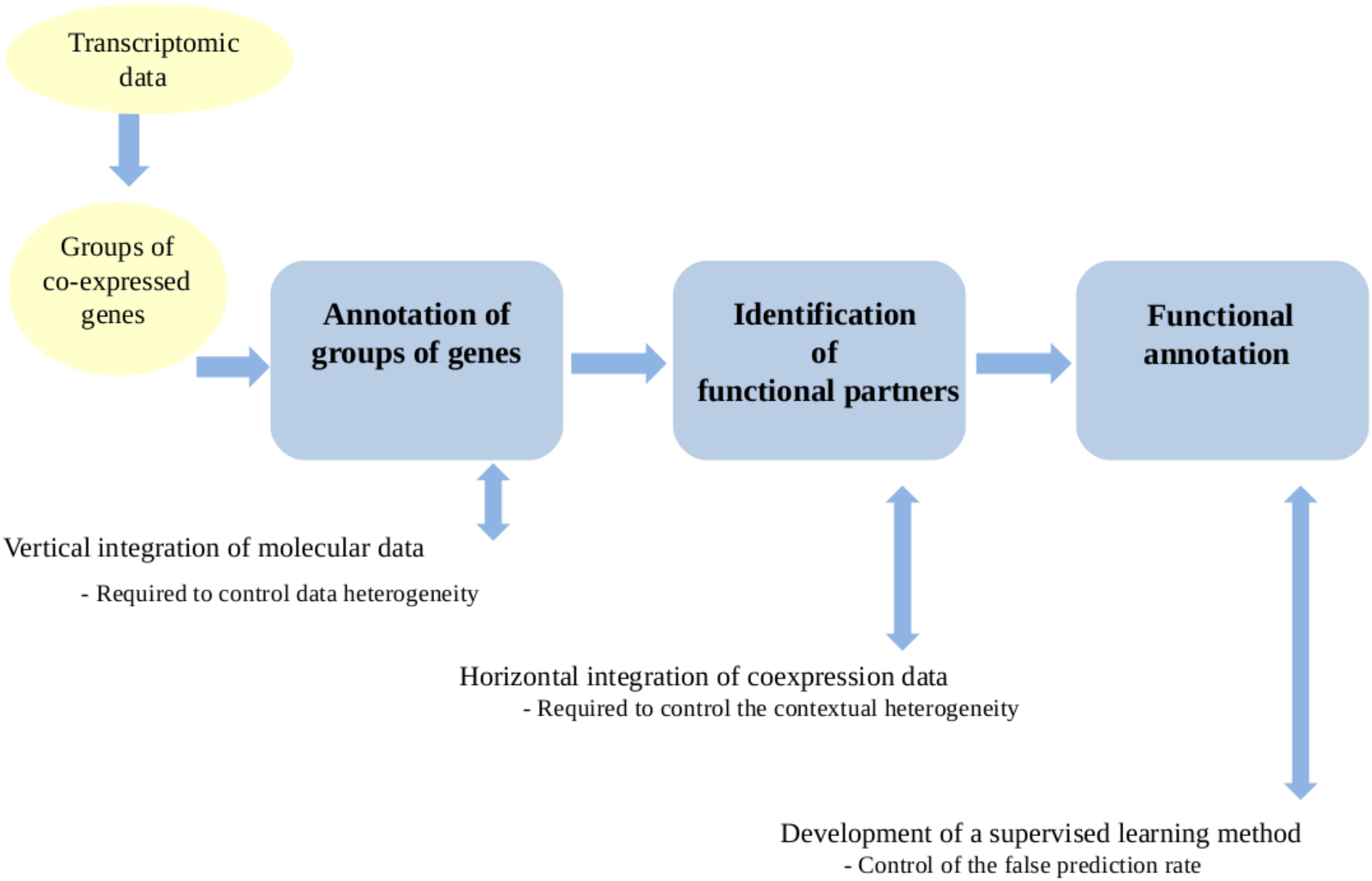
Conclusions on this large-scale co-expression study

- It generates meaningful groups of genes
- It performs favorably as compared to those obtained with correlation-based approaches (higher % of enrichments)

Nevertheless

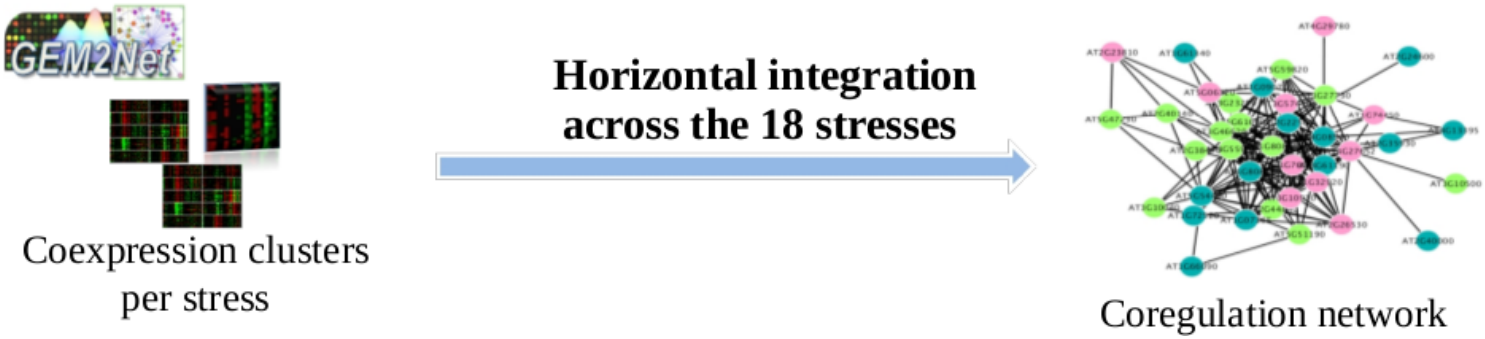
- 18 co-expression studies were generated
- Interpretation and use are not straightforward
- Co-expression is not enough to suggest co-regulation and to be used in a guilt by association approach (Dhaeseleer *et al.*, 2000)

Horizontal integration



From coexpression to coregulation

- Small overlap between two clusters of two different stresses
- Horizontal integration done at the level of the gene pairs

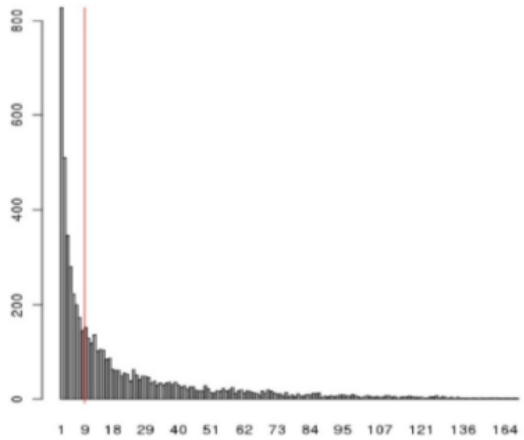


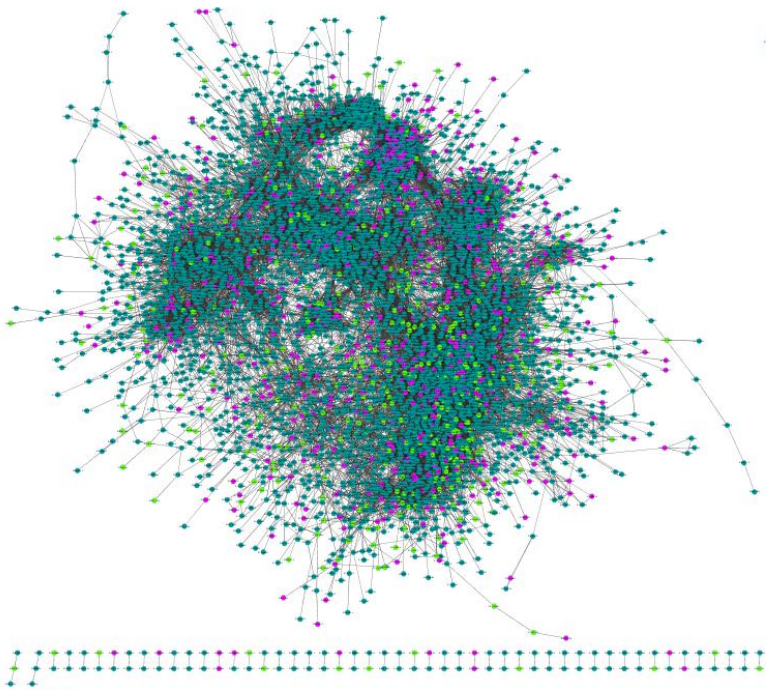
Method

- For each pair of genes, calculation to be in a same cluster of co-expression
- Comparison with a random network: a pair observed more than 3 times is statistically significant (resampling test)

Coregulation network
5 626 genes and 57 833 interactions

713 orphans and 1 682 with a missing GOSlim annotation

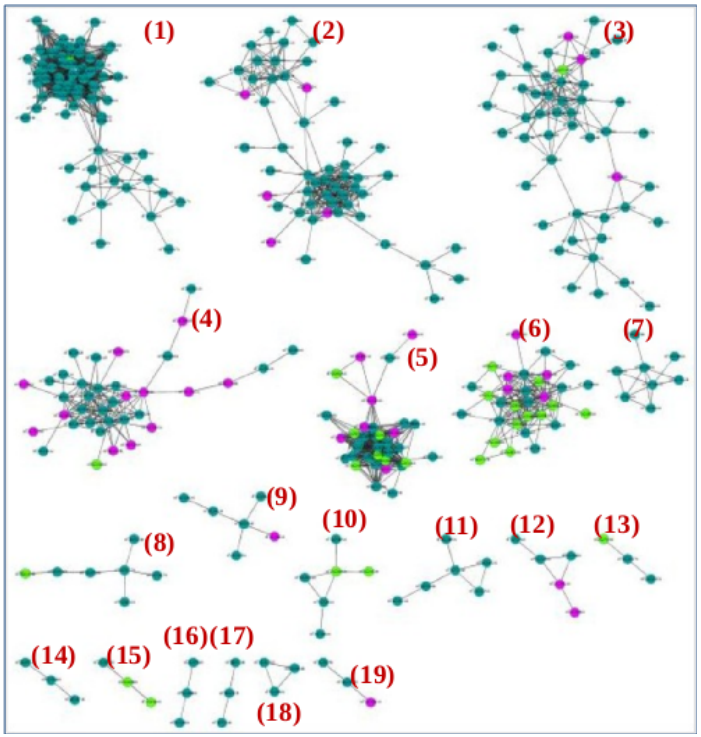




- Degree distribution is a power law
- Considered as an important quality criterion (Gillis et Pavlidis, 2012)

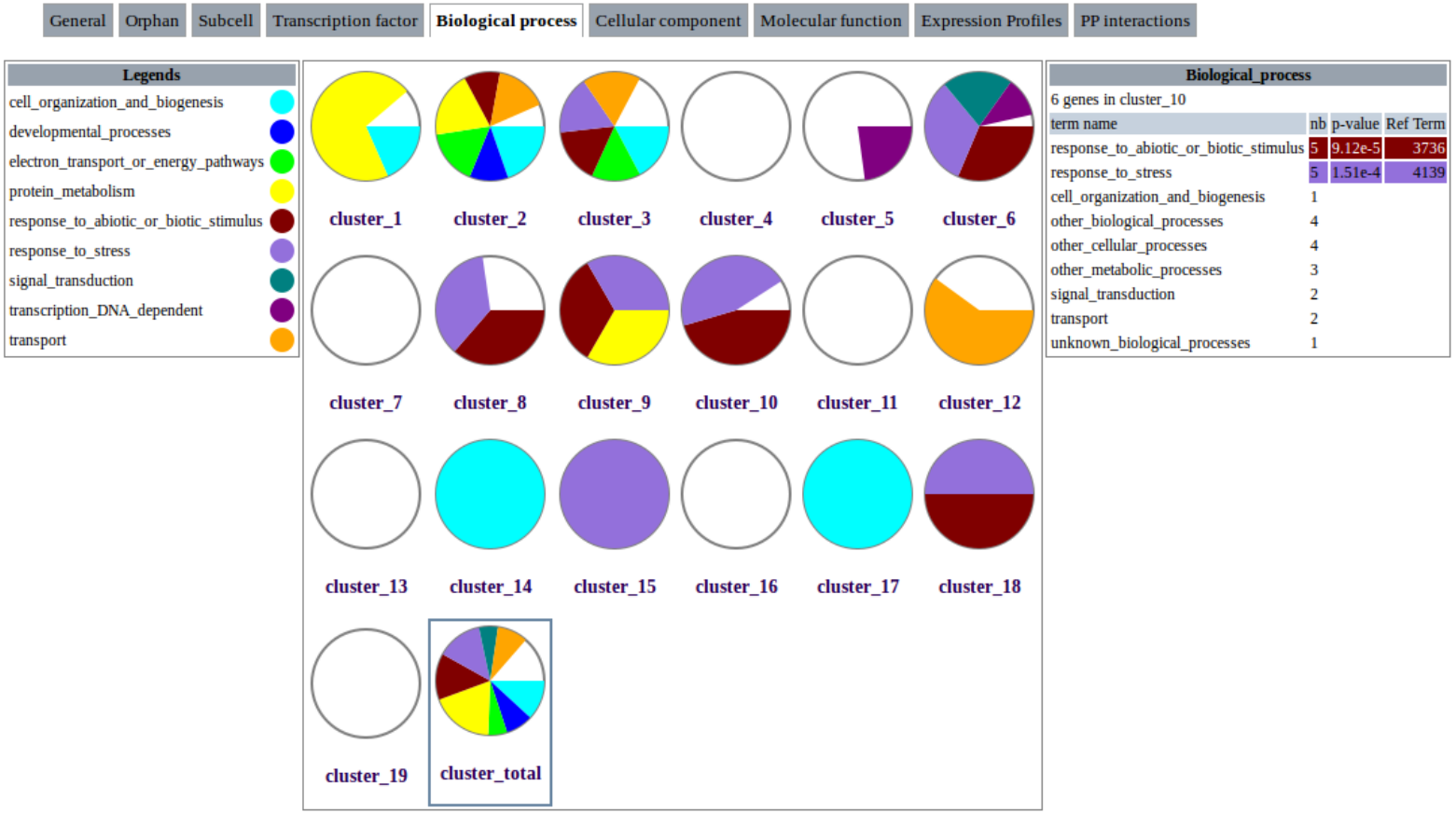
Topological properties

The network with gene pairs conserved in at least 7 stresses
415 genes with 41 orphans, 1 908 interactions



Cis-regulatory motifs found with PLMDetect (Bernard *et al.*, 2010)

- 10 components are enriched in motifs
- For 4 components, the motif is present in over 80% of the gene promotors
- Component 2 has 5 motifs related to the light regulation, present at most in 50% of gene promoters



Conclusions

- Coregulation modules are more specific and more homogeneous
- Cis-regulatory motifs are found in their promoters
- Topological analysis = an approach to identify functional modules

Horizontal integration

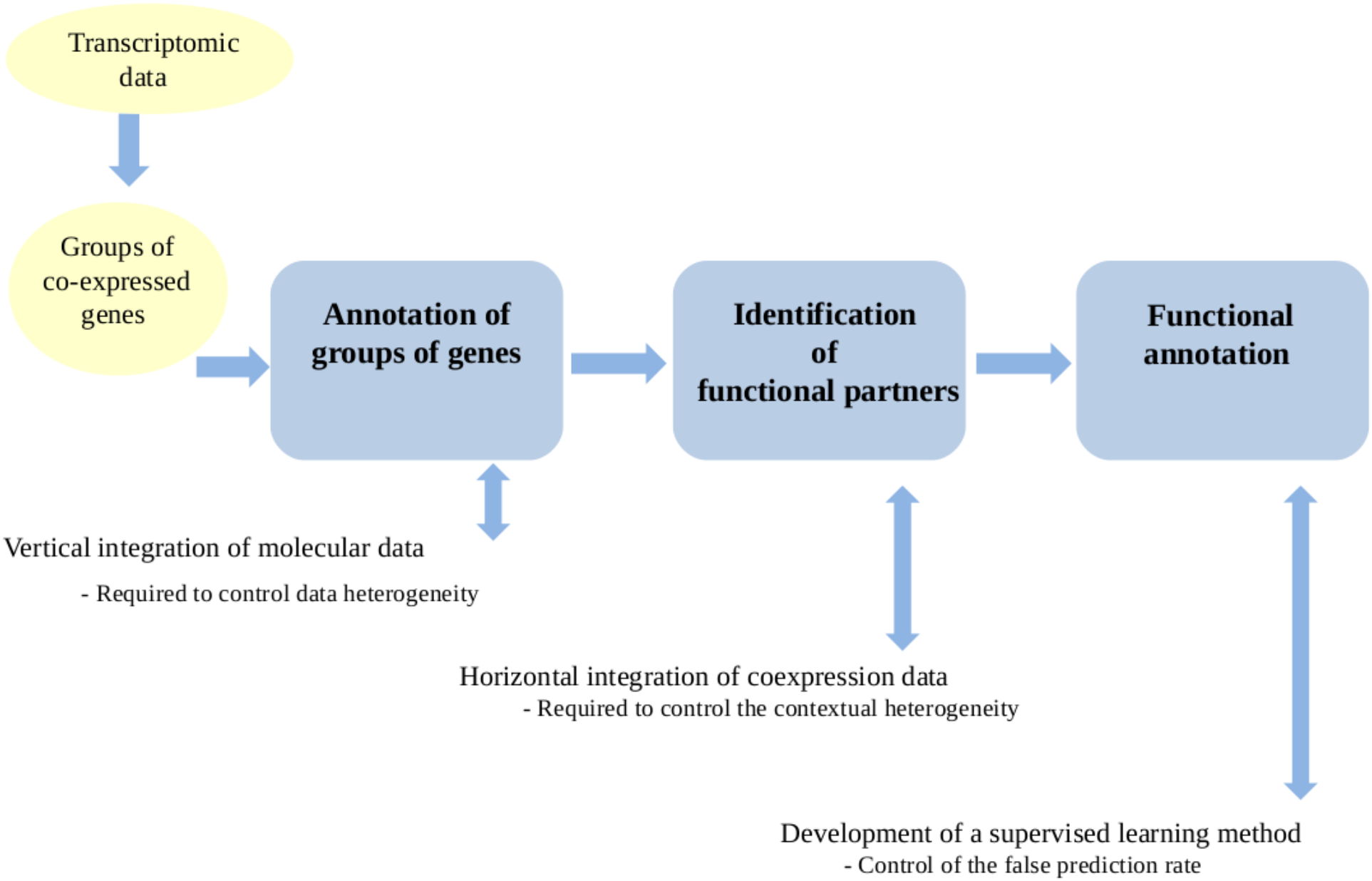
Results

- A comparaison with a random network allows us to transform an integration of coexpression results into a coregulation network
- Functional modules are identified by a topological analysis

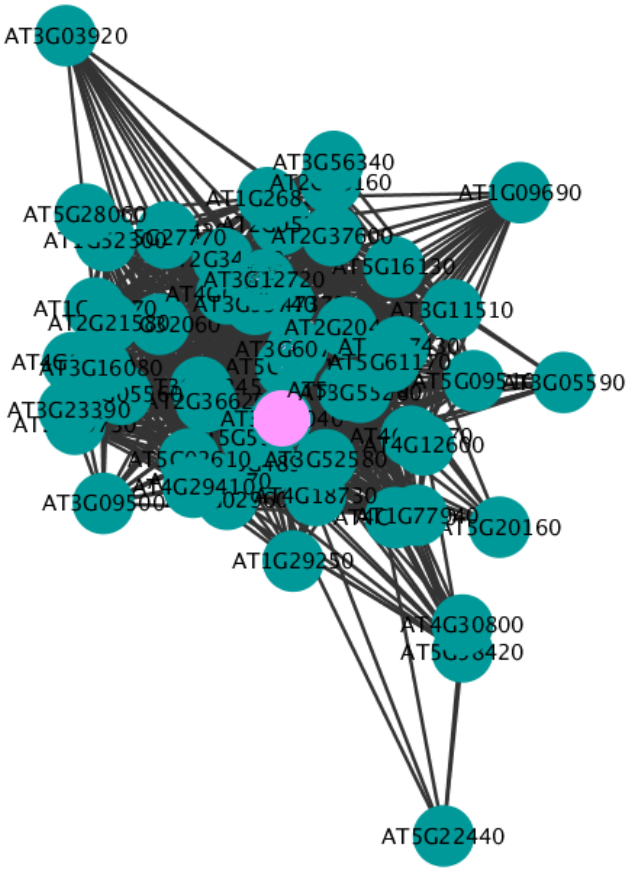
In progress

- Identification of Determinants of Expression Regulation to explain the coregulation (TF, small RNA, SMAR, chromatin marks, ...)
- Integration with a metabolomic network (coll. with V. Fromion and A. Goelzer, dpt MIA)

Annotation based on networks



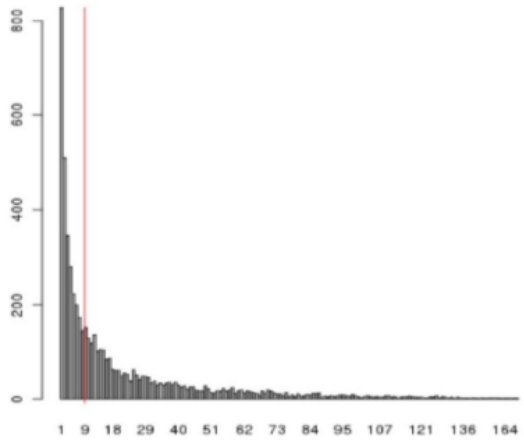
based on topological features

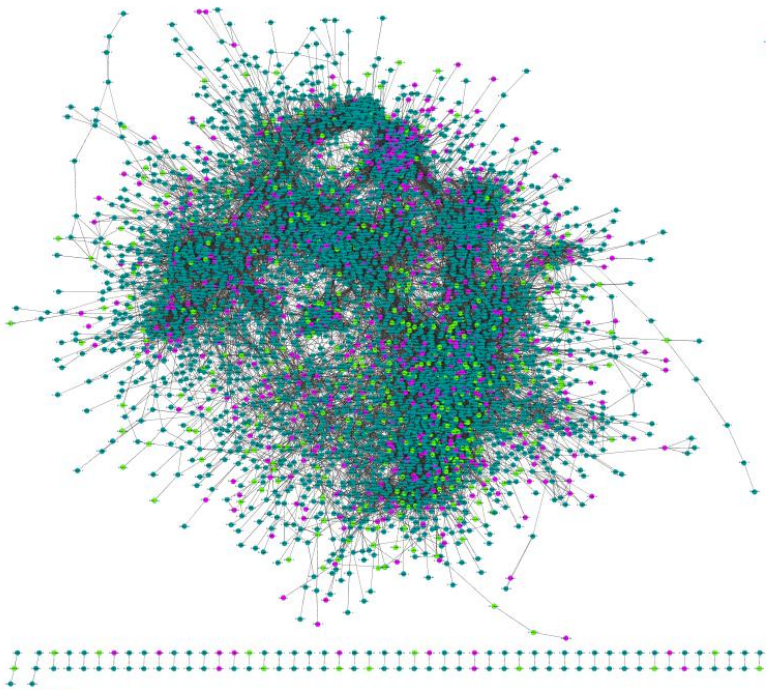


- Available information: presence of a F-box, a conserved domain present in numerous protein with a bipartite structure
- 48 of the 55 first neighbors of this gene are annotated as *Structural constituent of ribosome*

Coregulation network
5 626 genes and 57 833 interactions

713 orphans and 1 682 with a missing GOSlim annotation





- Degree distribution is a power law
- Considered as an important quality criterion (Gillis et Pavlidis, 2012)

Functional annotation per gene

Most methods of annotation are based on PPI network by using their neighborhood

The majority vote: for each gene, it predicts the 3 most frequent terms of its neighbourhood (Schwikowski *et al.* 2000)

Network	FDR	Fmeas
PPI	0.610	0.190
Transcriptome	0.866	0.081

Some comments

- To get validations, FDR must be controlled
- The question can be recast as a specific method per GO term (binary supervised classification)

Statistical framework

Working set = {all the genes with an annotation for a given ontology}

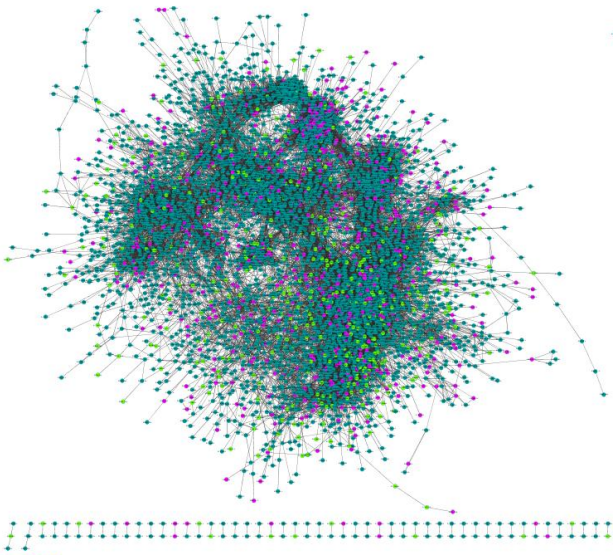
In the training set

- For each gene, calculate a score representing the presence of each term in its neighborhood
- For each term, define a threshold by comparing the scores and the right presence of the term so that FDR was less or equal to 20%

In the test set

- Calculate the score for each term
- Predict the presence/absence of the term (score $\geq$ threshold =presence)
- Estimate the FDR

Method parameters



- network (transcriptome, PPI, Union)
- semantic framework
- for extracting the information describing the network
- for explaining how the neighbors contribute to the score calculation
- for describing the dependance (or not) between terms

Ontology	Transcriptome	PPI	Union
BP	638	495	703
MF	234	166	266
CC	161	139	187
All	1033	800	1156

Application with DAVID ontology

Number of analyzed terms

Ontology	Transcriptome	PPI	Union
BP	32	44	150
MF	9	8	32
CC	39	13	70
All	80	65	252

Ontology	Trscript.	PPI	Union	Trscript.	PPI	Union
	FDR			Fmeas		
BP	0.164	0.155	0.136	0.234	0.368	0.258
MF	0.156	0.186	0.162	0.310	0.529	0.373
CC	0.153	0.139	0.150	0.424	0.628	0.319
All	0.157	0.159	0.149	0.322	0.508	0.316

Comparison with the majority vote

The majority vote (Schwikowski *et al.* 2000)

For each gene, it predicts the 3 most frequent terms of its neighbourhood

Ontology	Trscript.	PPI	Union	Trscript.	PPI	Union
	FDR			Fmeas		
BP	0.597	0.345	0.614	0.327	0.294	0.214
MF	0.626	0.306	0.621	0.458	0.629	0.380
CC	0.635	0.254	0.562	0.234	0.583	0.193
All	0.618	0.322	0.600	0.296	0.393	0.228

Our method reduces the false positives among the genes predicted for having the term without an important decrease of the Fmeas

Conclusions about the functional annotation

Results

- Annotation per term is important
- Annotation depends on input data
- FDR can be controlled

Results

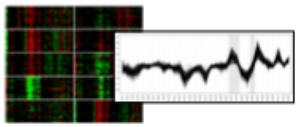
- First results are promising
- Analysis with coexpression and interactome data give more predictions
- most sophisticated classifiers
- Think about some validations ...

Actors of this project

Bioinformatics

R. Zaag
V. Brunaud
J.-P. Tamby
C.Guichard
Z.Tariq
S. Aubourg

Statistics



G. Celeux
C. Maugis
T. Mary-Huard
G. Rigaill

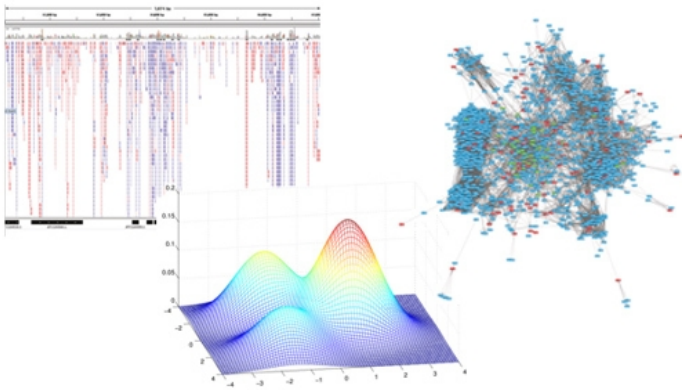
Biology

J-P Renou
S. Balzergue

And thank you for your attention !

Funding :

Advertissement



- Le 16 mai : journee de la transcriptomique vegetale
- mai-decembre 2017 : Ecole-Chercheurs “De l’expression des genes aux reseaux”