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

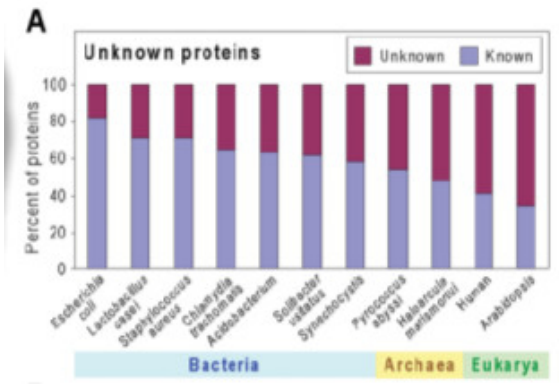
M.-L. Martin-Magniette[1,2] & E. Delannoy[1]

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

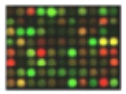
Nevertheless

- 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

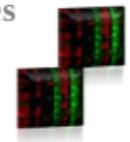
Quantitative data



Contextual data

coexpression

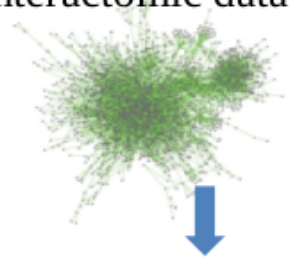
Groups of genes having a similar expression profile



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%

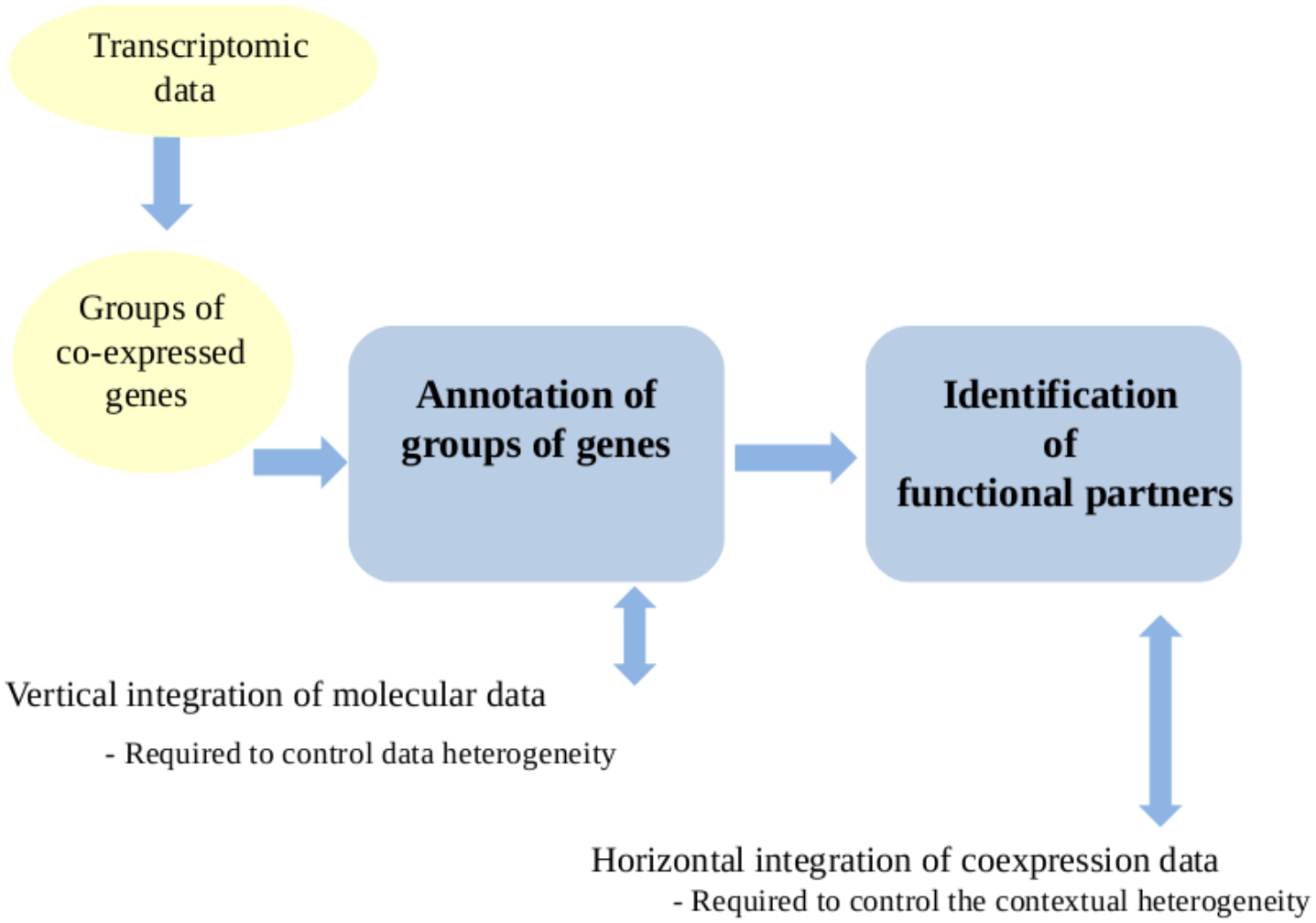
by omics integration

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

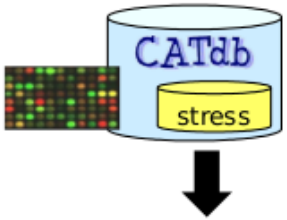
But various sources of heterogeneity exist

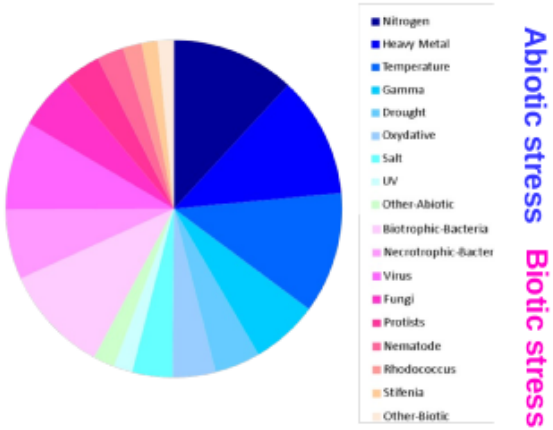
- Data are qualitative or quantitative
- Available information describes the biological entities or their relationships
- Observations are obtained with various techniques
- Various semantic frameworks are used

From Gene Expression Modeling to Networks



A dedicated transcriptomic dataset



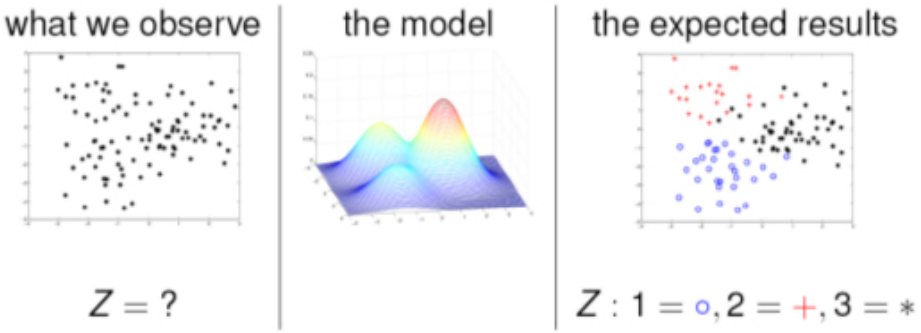


- 387 transcriptomic comparisons in dye-swap dedicated to stress
- 2/3 describe abiotic stresses and 1/3 biotic stresses
- All the data were generated on the same transcriptomic platform with the same protocol

First results

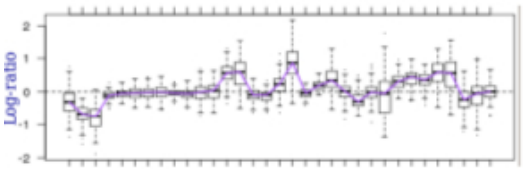
- Based on differential analyses, 60% of the genes coding proteins have their transcription impacted directly or not by a stress
- Large overlap of impacted genes between biotic and abiotic stress

Coexpression study using mixture model



Matrix by stress
{ genes x log-ratios}

Gaussian mixture



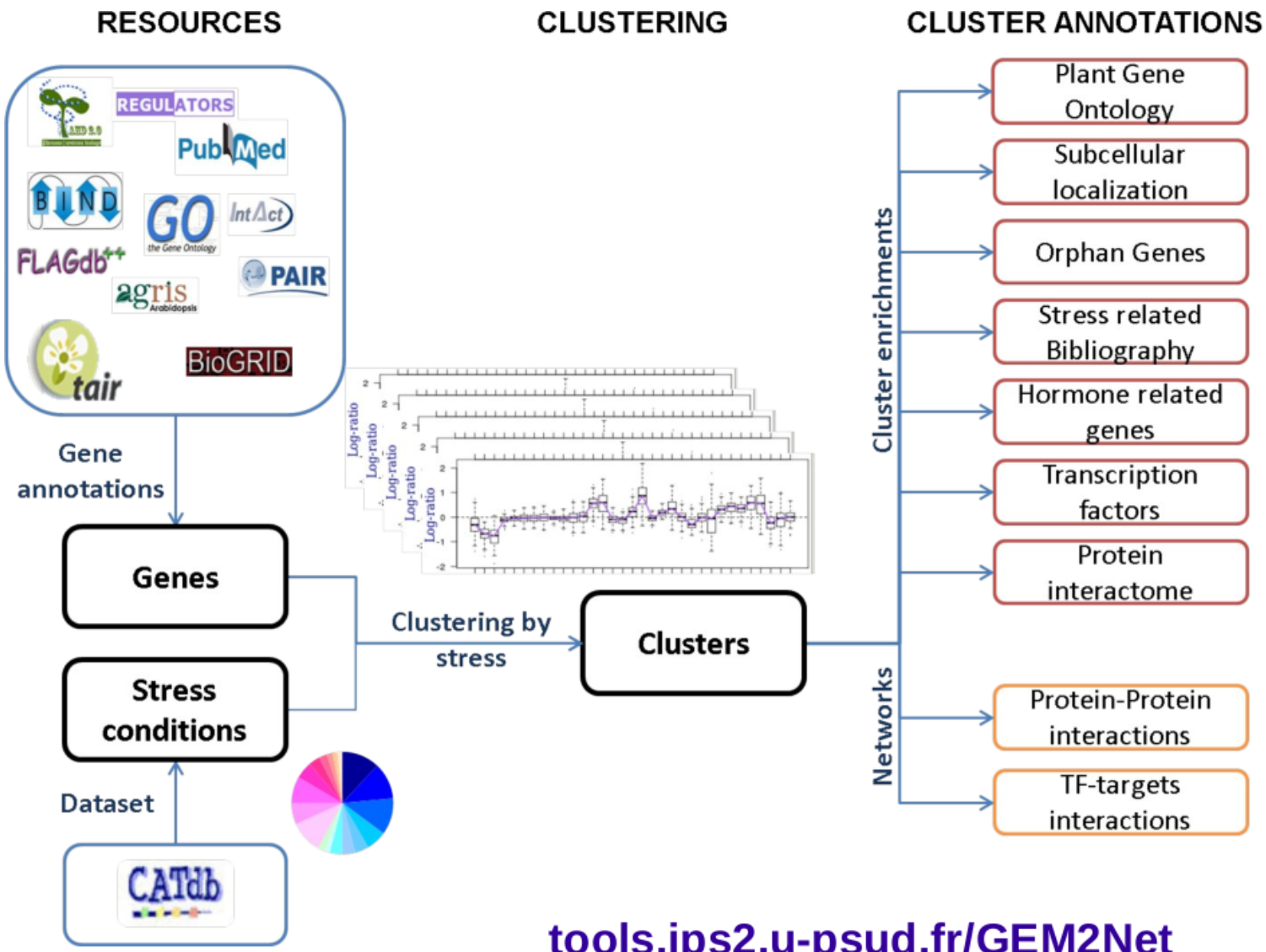
Data-driven method

- 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

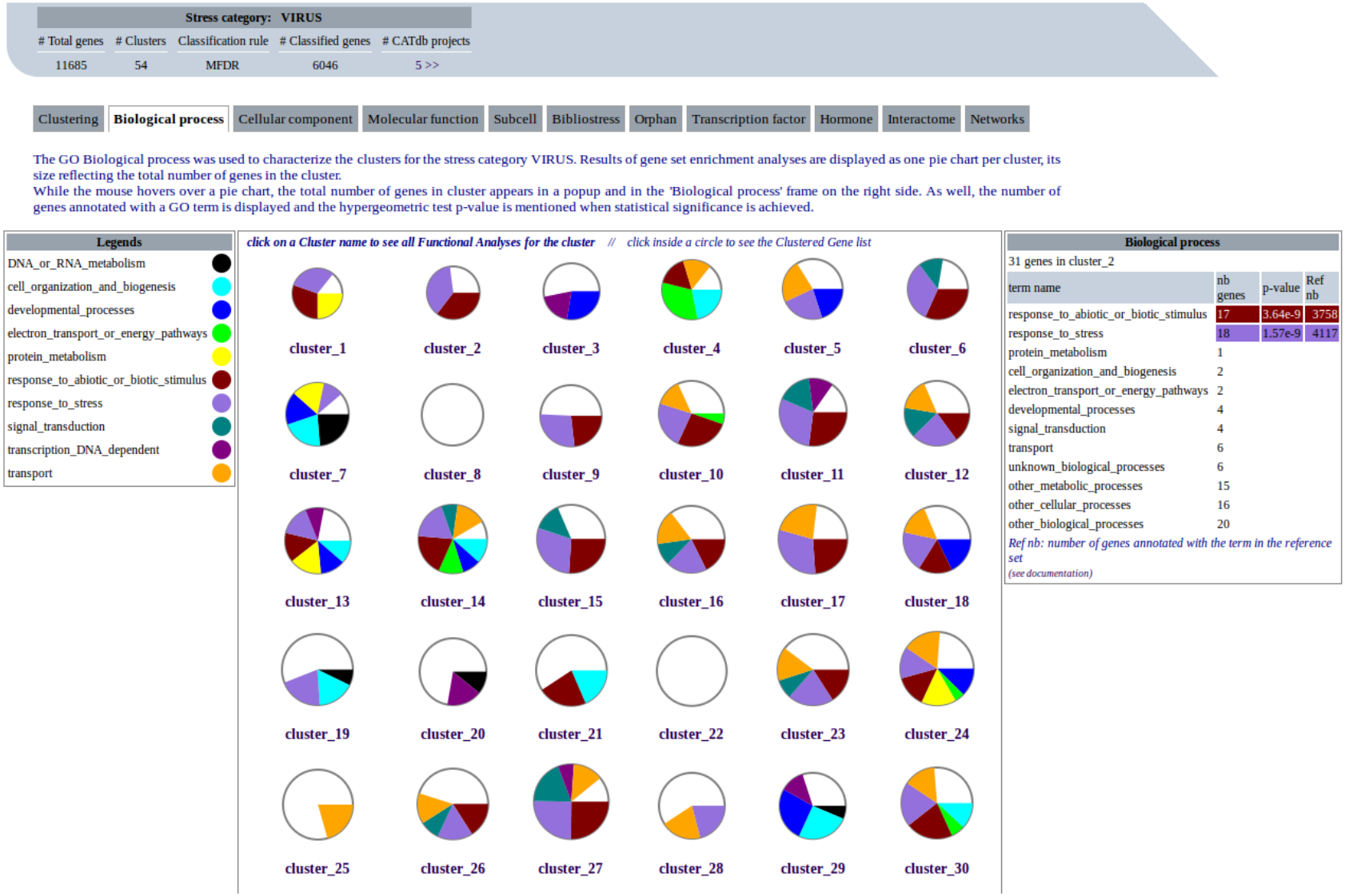
~ 700 clusters of co-expression

Annotation of coexpressed genes



tools.ips2.u-psud.fr/GEM2Net

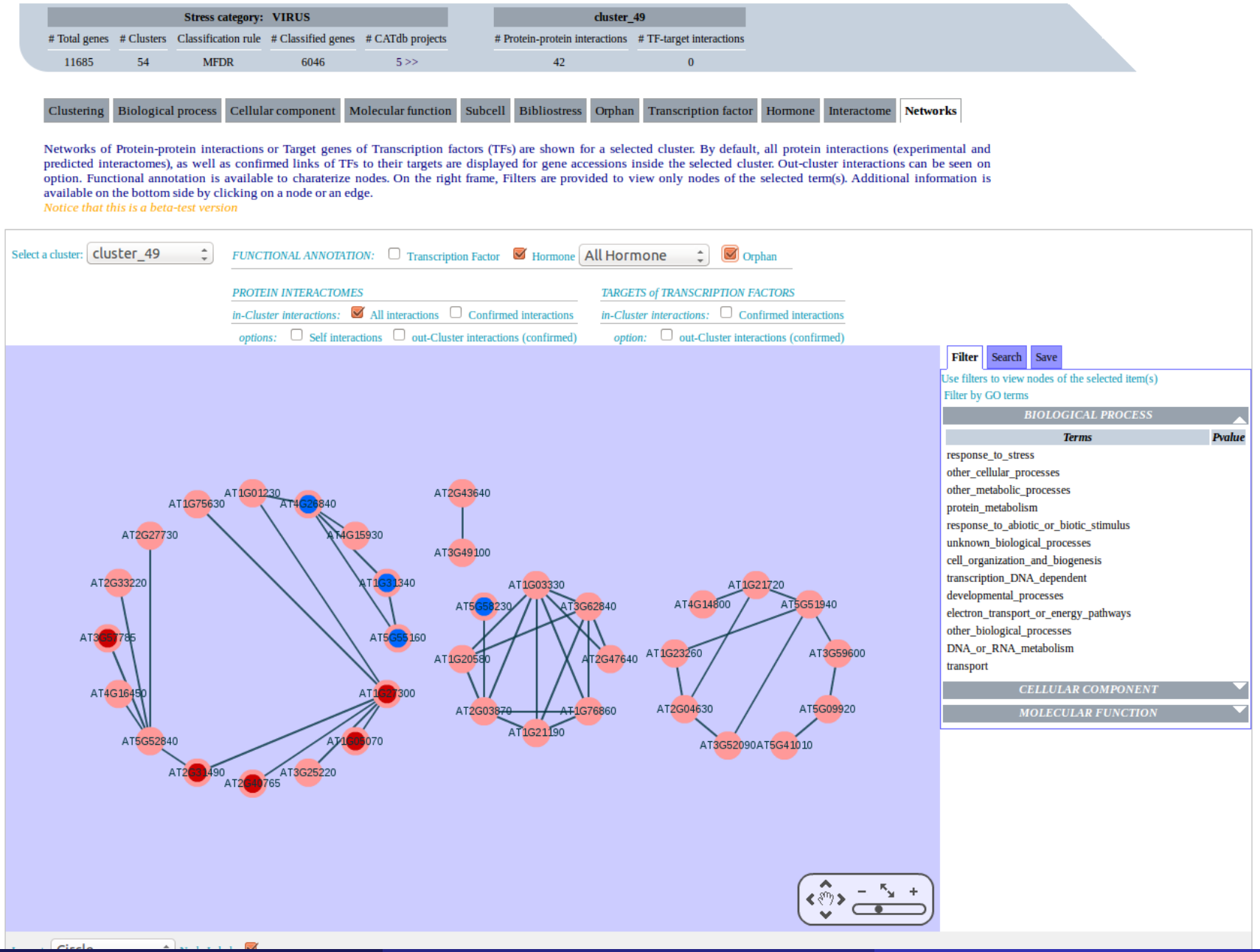
Visualisation by type of resource



Pie size proportional to cluster size

Colors indicate biological biases

Visualisation of interactions



Overview of a cluster

Stress category: VIRUS				
# Total genes	# Clusters	Classification rule	# Classified genes	# CATdb projects
11685	54	MFDR	6046	5 >>

cluster_49		
# Genes in cluster	# TAIR genes	# Other genes
213	212	1

Download 'Biological process' Gene list

Download functional annotation table

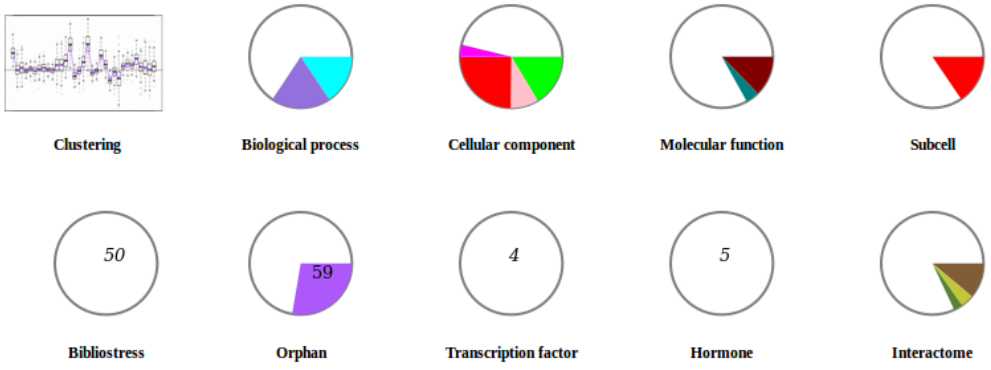
Clustering | Biological process | Cellular component | Molecular function | Subcell | Bibliostress | Orphan | Transcription factor | Hormone | Interactome | Networks

Overview of all functional annotation analyses made for the cluster 'cluster_49'. While clicking on a circle, a gene set enrichment list for the concerning annotation appears below the main panel. Results of analyses are recapped in the 'Functional Annotation' table on the right side and are downloadable using the link above.

Legends

- **Biological process**
  - cell_organization_and_biogenesis
  - response_to_stress
- **Cellular component**
  - chloroplast
  - cytosol
  - mitochondria
  - ribosome
- **Molecular function**
  - protein_binding
  - structural_molecule_activity
- **Subcell**
  - Mitochondria
- **Interactome**
  - Interactome PAIR
  - Interactome AI1
  - Interactome LCI

Analyses overview for cluster_49



Functional Annotation for cluster_49

213 genes in cluster_49

Biological process	nb genes	pvalue	Ref nb
cell_organization_and_biogenesis	36	1.17e-3	3367
response_to_stress	43	5.84e-4	4117
Cellular component	**nb genes**	**pvalue**	**Ref nb**
chloroplast	39	3.16e-3	3959
cytosol	21	2.50e-3	1692
mitochondria	59	2.45e-12	3571
ribosome	9	3.82e-3	484
Molecular function	**nb genes**	**pvalue**	**Ref nb**
protein_binding	27	6.77e-3	2584
structural_molecule_activity	9	8.47e-3	549
Subcell	**nb genes**	**pvalue**	**Ref nb**
mitochondria	33	9.27e-6	2313

cluster_49: gene set enrichment list for Biological process annotation
62 genes

gene_name	product	terms	other enrichments: Bp	Cc	Mf	Su	Bs	Or	Tf	Ho	In
AT1G01160	GRF1-INTERACTING FACTOR 2		●		●						●
AT1G01230	ORMDL FAMILY PROTEIN		●	●							●
AT1G04070	TRANSLOCASE OF OUTER MEMBRANE 22-I		●	●							
AT1G05070	PROTEIN OF UNKNOWN FUNCTION (DUF1068)		●					●			●
AT1G21720	PROTEASOME BETA SUBUNIT C1		●	●							●
AT1G23260	MMS ZWEI HOMOLOGUE 1		●		●						●
AT1G24450	RIBONUCLEASE III FAMILY PROTEIN		●	●							●
AT1G27310	NUCLEAR TRANSPORT FACTOR 2A		●		●						●
AT1G31170	SULFIREDOXIN		●	●		●					●
AT1G32310	NOT DEFINED		●		●			●			
AT1G52740	HISTONE H2A PROTEIN 9		●		●						●
AT1G61570	TRANSLOCASE OF THE INNER MITOCHONDRIAL MEMBRANE 13		●	●							●
AT1G65290	MITOCHONDRIAL ACYL CARRIER PROTEIN 2		●	●	●	●					●
AT1G67350	NOT DEFINED		●	●				●			●

Vertical integration

Results

- Numerous enrichments
- Overlap with TF regulations and PPI

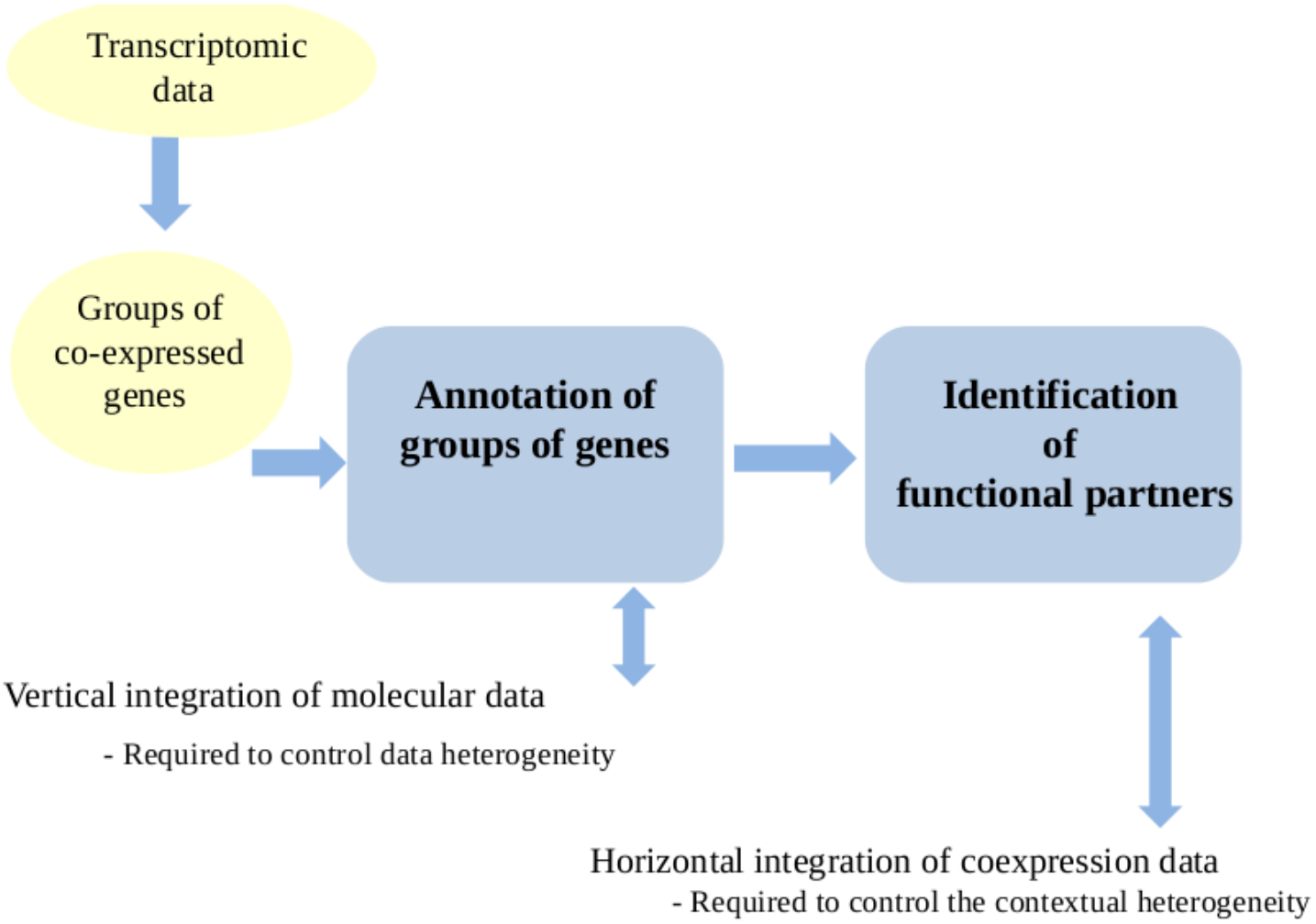
Conclusions on this large-scale co-expression study

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

Nevertheless

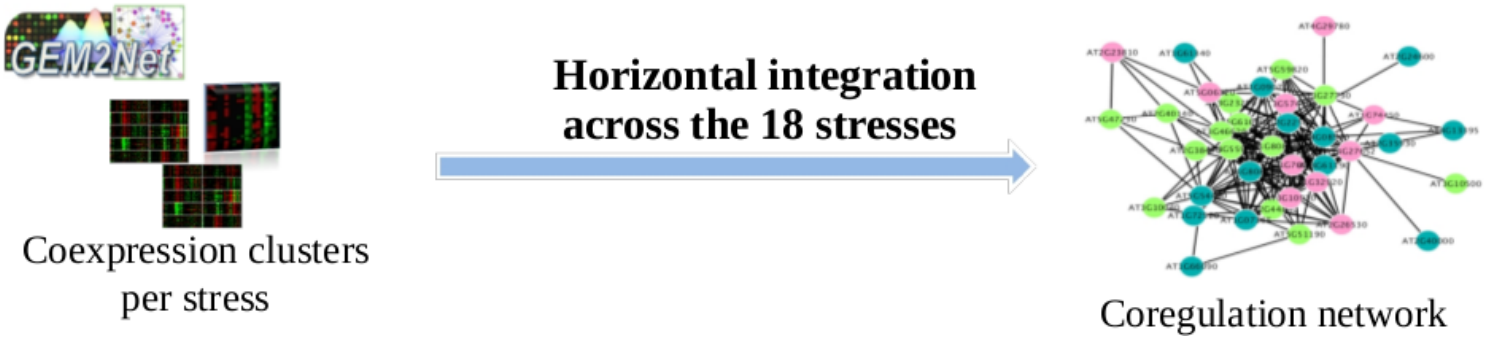
- 18 co-expression studies were generated
- Interpretation and use are not straightforward
- Co-expression is not enough to suggest co-regulation and to be used in a guilt by association approach (Dhaeseleer *et al.*, 2000)

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

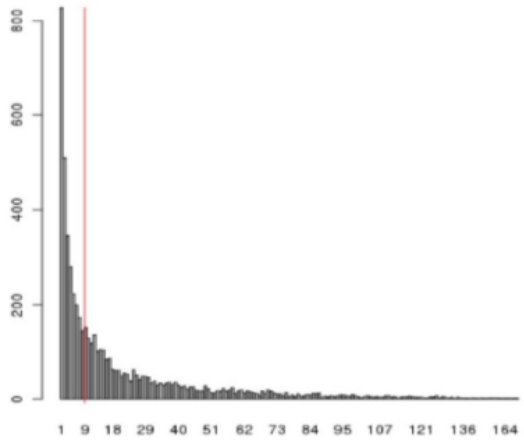


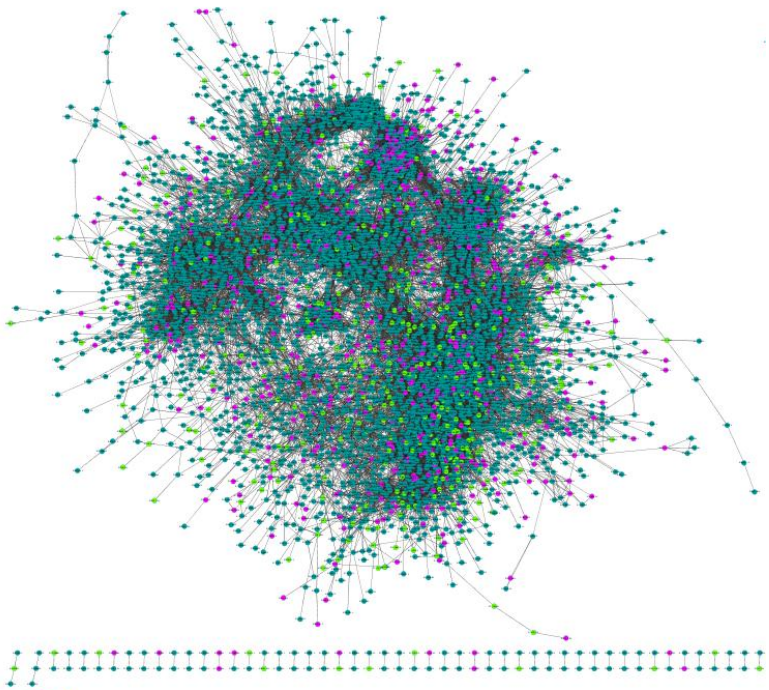
Method

- 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

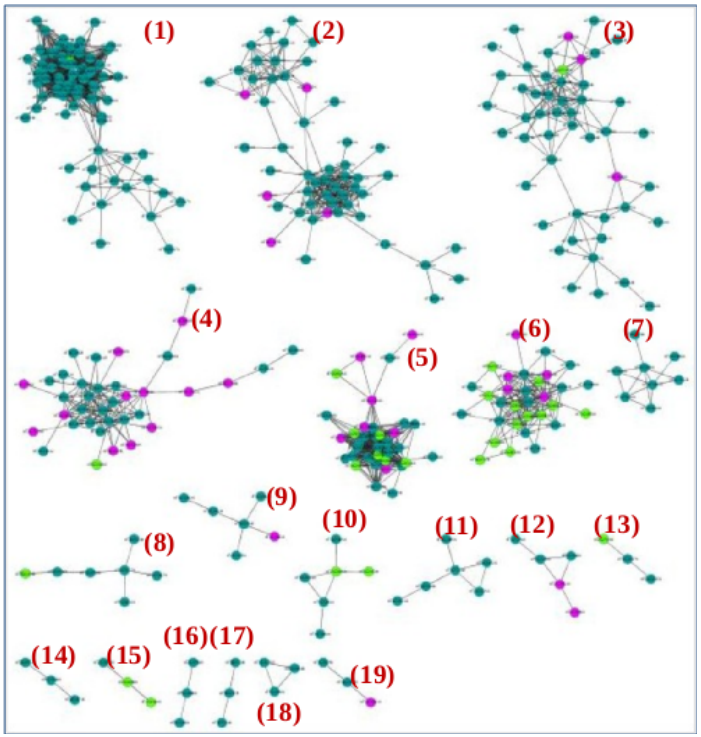




- Degree distribution is a power law
- Considered as an important quality criterion (Gillis et Pavlidis, 2012)

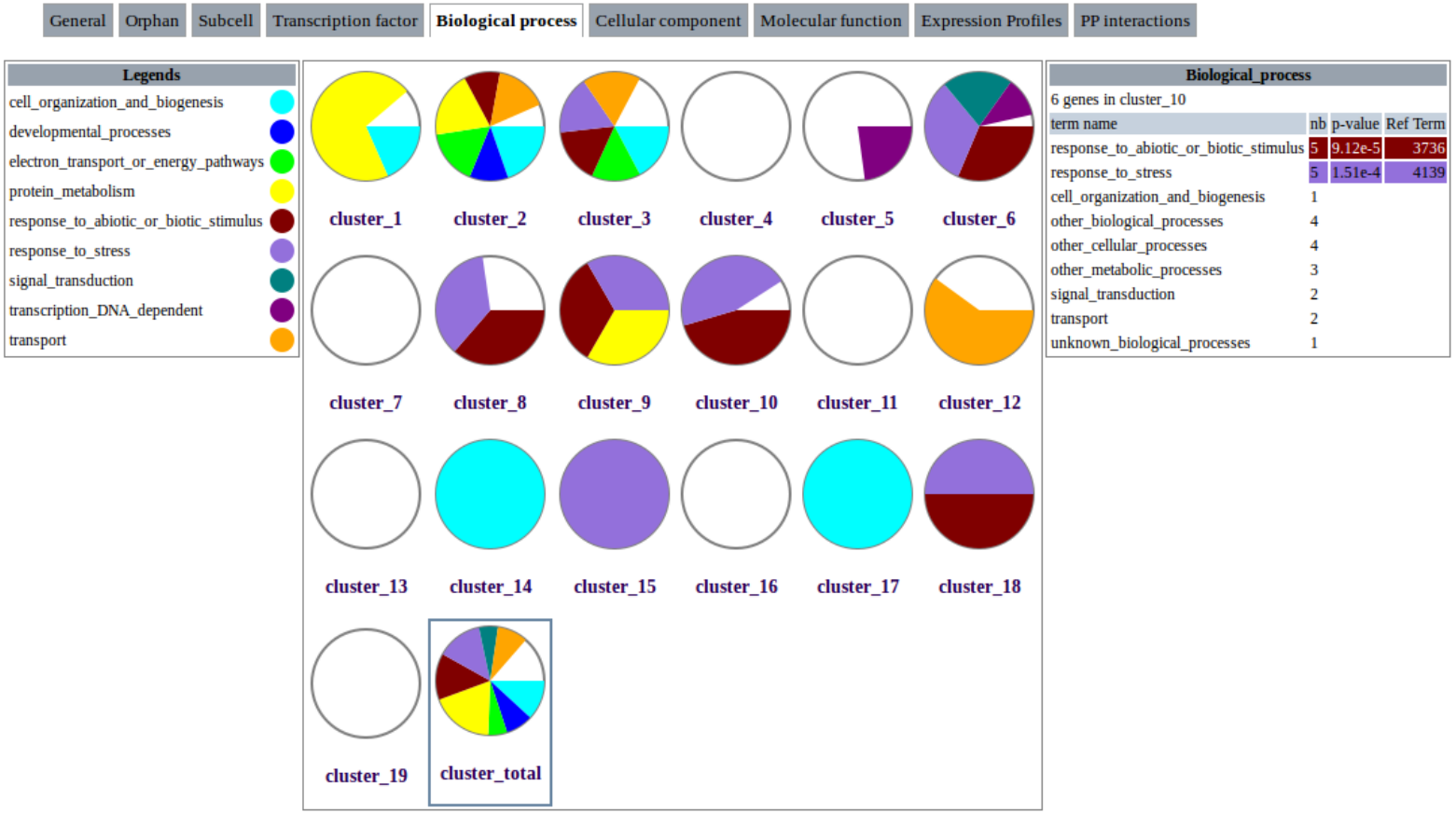
Topological properties

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