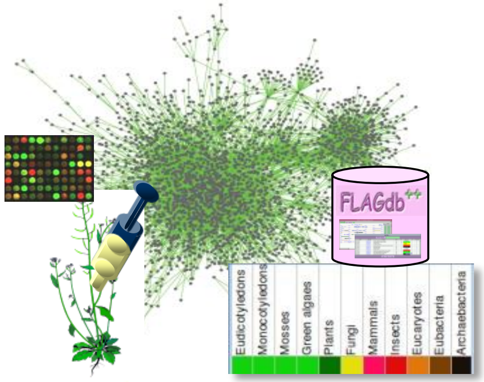



# From Gene Expression Modelling to Gene network to investigate Arabidopsis genes involved in stress response

Zaag* R, Tamby* JP, Guichard* C, Tariq Z, Rigaill G, Delannoy E, Renou JP, Balzergue S, Mary-Huard T, Aubourg S,
Martin-Magniette ML, Brunaud V.

IPS2
Institute of Plant Sciences
Paris - Saclay

POPS
Plateforme transcriptOmique
IPS2

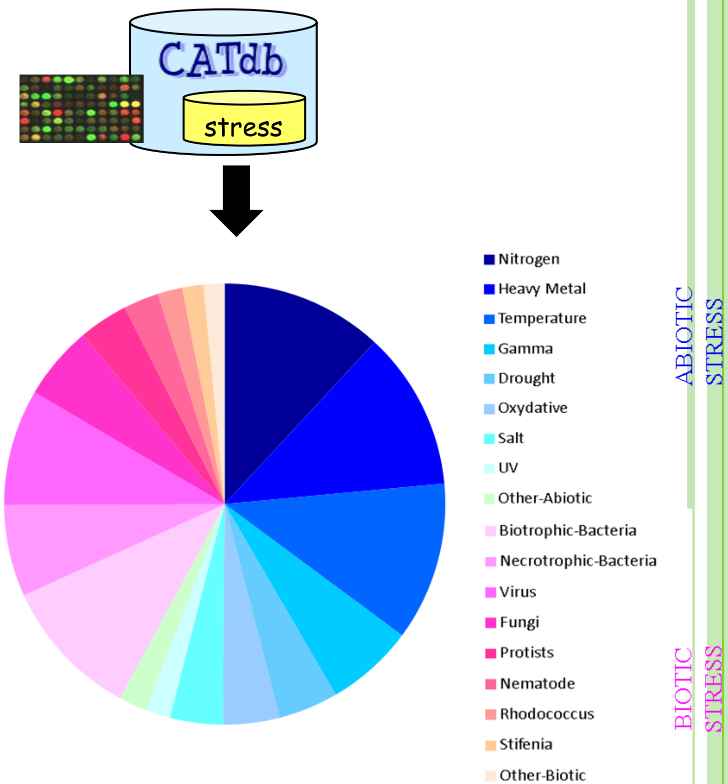
# Current challenge in genomics

- Between 20% and 40% of the predicted genes have no assigned function (Hanson et al., 2010)
- New challenge is the functional annotation to identify the function(s) of each gene
- Functional annotation must be expressed in a shared and controlled vocabularies -> use of the Gene Ontology
  - Molecular function
  - Biological process
  - Cellular component
- Functional annotation was first based on structural similarities. It is not enough to propose an exhaustive annotation
- New approaches are based on transcriptomic studies because co-expressed genes are often involved in a same biological process

# A dedicated transcriptomic dataset

➢ 387 transcriptomic comparisons in dye-swap dedicated to stress (2/3 abiotic stress and 1/3 biotic stress)

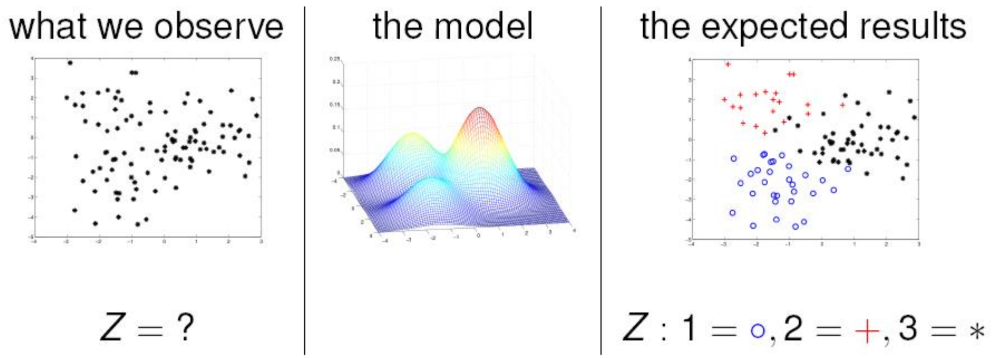
➢All the data generated by the platform POPS with the same protocol




Based on differential analyses**, 60% of the genes coding proteins** have their transcription impacted directly or not by a stress

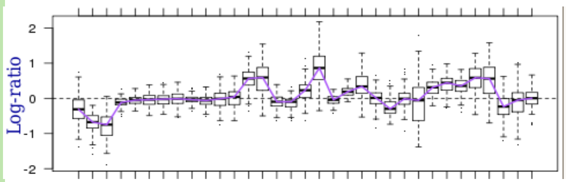
Large overlap of impacted genes between biotic and abiotic stresses

# Co-expression analysis by mixture models




Matrix by stress
{ genes x log-ratios}

Gaussian mixture




**Data-driven method**

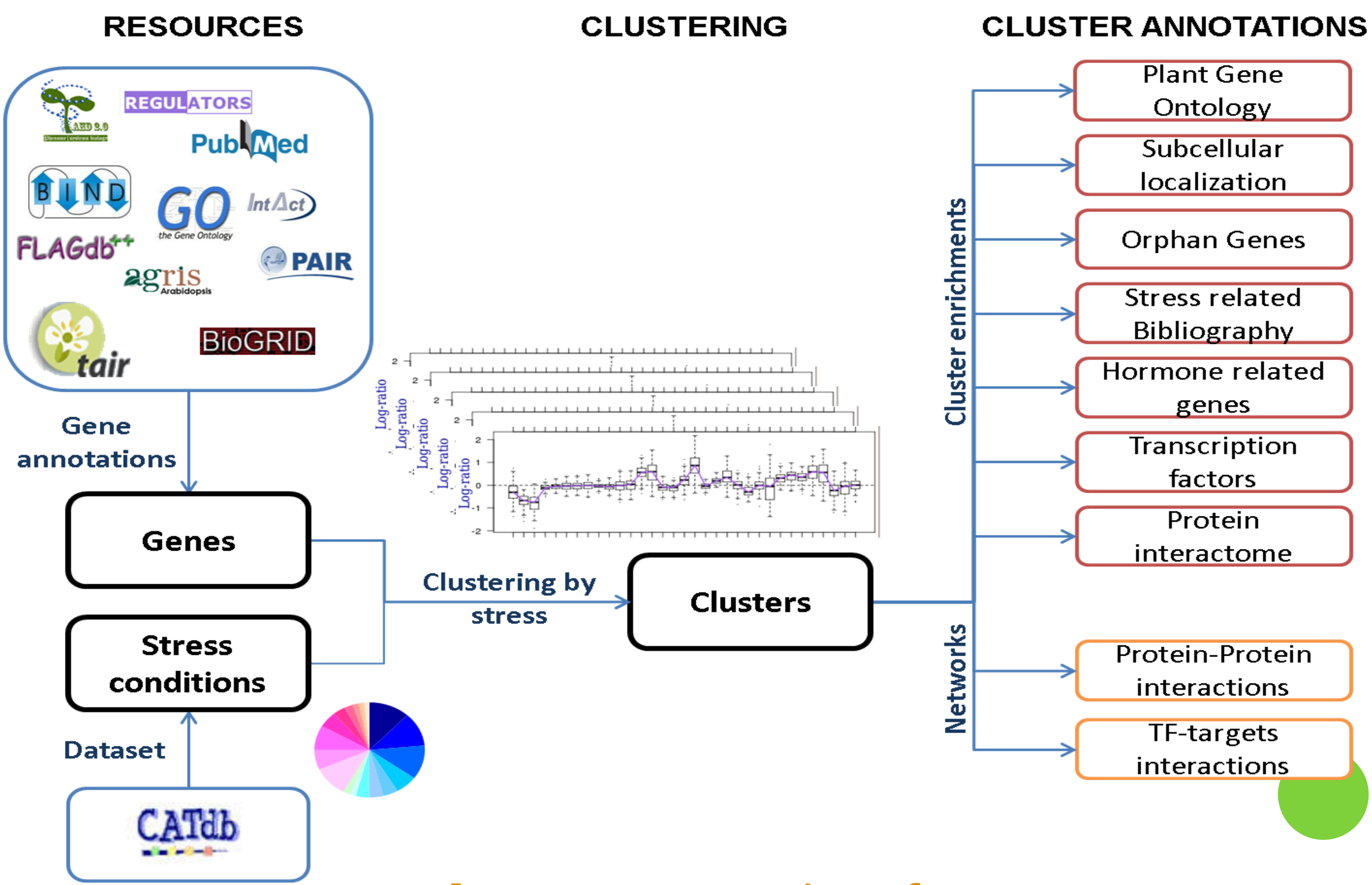
• number of cluster chosen by BIC

• 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

# GEM2Net approach




http://urgv.evry.inra.fr/GEM2NET

# Visualization by type of resource

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

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

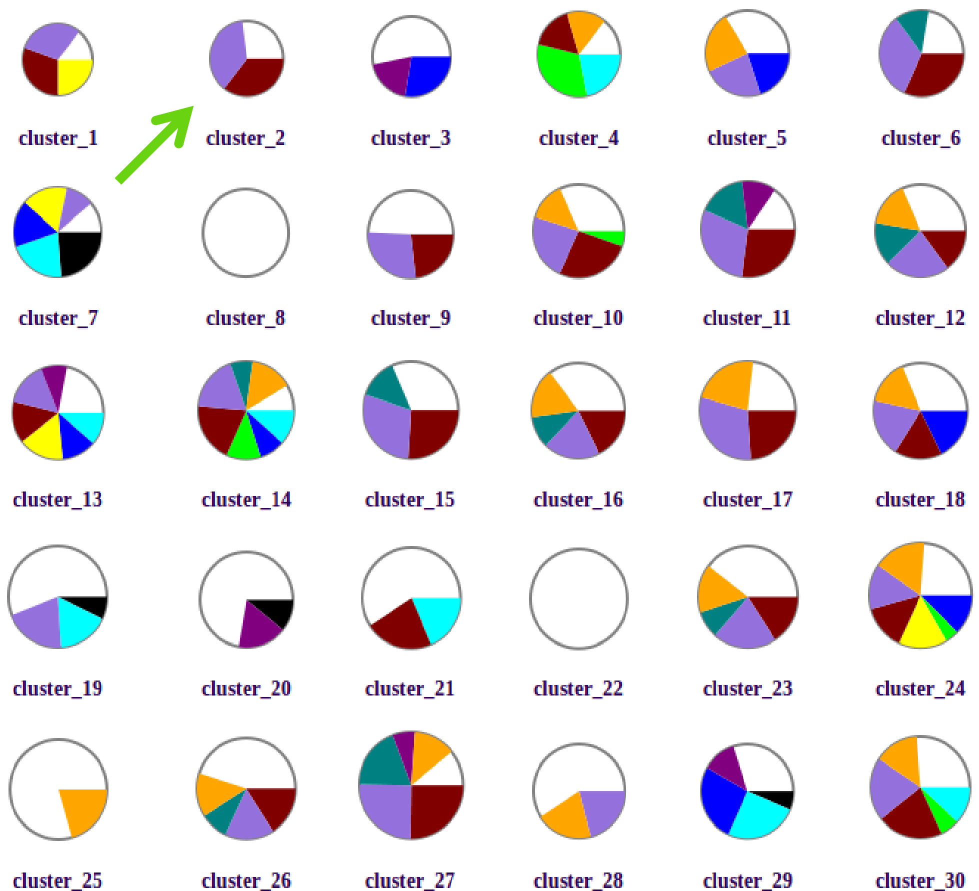
The GO Biological process was used to characterize the clusters for the stress category VIRUS. Results of gene set enrichment analyses are displayed as one pie chart per cluster, its size reflecting the total number of genes in the cluster.
While the mouse hovers over a pie chart, the total number of genes in cluster appears in a popup and in the 'Biological process' frame on the right side. As well, the number of genes annotated with a GO term is displayed and the hypergeometric test p-value is mentioned when statistical significance is achieved.

| Legends |
|---|
| DNA_or_RNA_metabolism |
| cell_organization_and_biogenesis |
| developmental_processes |
| electron_transport_or_energy_pathways |
| protein_metabolism |
| response_to_abiotic_or_biotic_stimulus |
| response_to_stress |
| signal_transduction |
| transcription_DNA_dependent |
| transport |

**Colors indicate a biological bias**

**Size of the pie proportional to the size of the cluster**

*click on a Cluster name to see all Functional Analyses for the cluster // click inside a circle to see the Clustered Gene list*




**Biological process**

31 genes in cluster_2

| term name | nb genes | p-value | Ref nb |
|---|---|---|---|
| response_to_abiotic_or_biotic_stimulus | 17 | 3.64e-9 | 3758 |
| response_to_stress | 18 | 1.57e-9 | 4117 |
| protein_metabolism | 1 | | |
| cell_organization_and_biogenesis | 2 | | |
| electron_transport_or_energy_pathways | 2 | | |
| developmental_processes | 4 | | |
| signal_transduction | 4 | | |
| transport | 6 | | |
| unknown_biological_processes | 6 | | |
| other_metabolic_processes | 15 | | |
| other_cellular_processes | 16 | | |
| other_biological_processes | 20 | | |

*Ref nb: number of genes annotated with the term in the reference set*
*(see documentation)*

# Visualization by type of resource

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

| cluster_49 | |
|---|---|
| # Protein-protein interactions | # TF-target interactions |
| 42 | 0 |

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

Networks of Protein-protein interactions or Target genes of Transcription factors (TFs) are shown for a selected cluster. By default, all protein interactions (experimental and predicted interactomes), as well as confirmed links of TFs to their targets are displayed for gene accessions inside the selected cluster. Out-cluster interactions can be seen on option. Functional annotation is available to charaterize nodes. On the right frame, Filters are provided to view only nodes of the selected term(s). Additional information is available on the bottom side by clicking on a node or an edge.

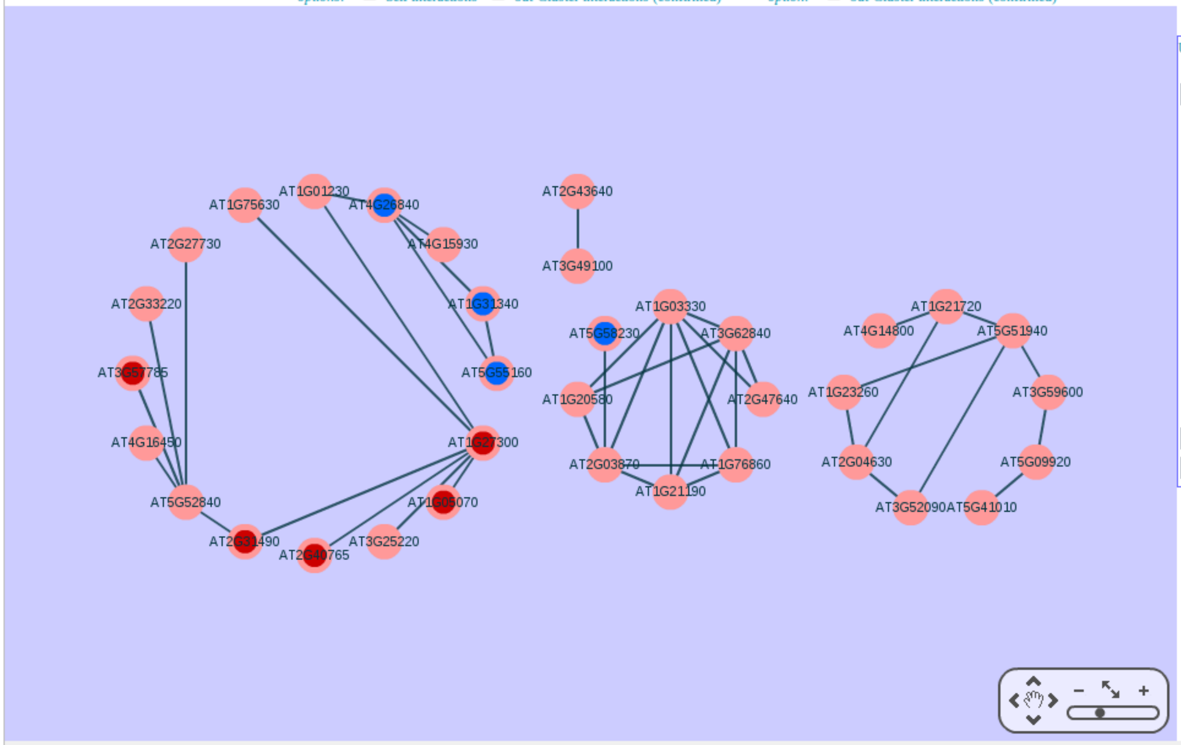
*Notice that this is a beta-test version*

Select a cluster: cluster_49

*FUNCTIONAL ANNOTATION:* ☐ Transcription Factor ☑ Hormone All Hormone ☑ Orphan

*PROTEIN INTERACTOMES*
*in-Cluster interactions:* ☑ All interactions ☐ Confirmed interactions
*options:* ☐ Self interactions ☐ out-Cluster interactions (confirmed)

*TARGETS of TRANSCRIPTION FACTORS*
*in-Cluster interactions:* ☐ Confirmed interactions
*option:* ☐ out-Cluster interactions (confirmed)




Filter | Search | Save

Use filters to view nodes of the selected item(s)
Filter by GO terms

*BIOLOGICAL PROCESS*

| *Terms* | *Pvalue* |
|---|---|
| response_to_stress | |
| other_cellular_processes | |
| other_metabolic_processes | |
| protein_metabolism | |
| response_to_abiotic_or_biotic_stimulus | |
| unknown_biological_processes | |
| cell_organization_and_biogenesis | |
| transcription_DNA_dependent | |
| developmental_processes | |
| electron_transport_or_energy_pathways | |
| other_biological_processes | |
| DNA_or_RNA_metabolism | |
| transport | |

*CELLULAR COMPONENT*

*MOLECULAR FUNCTION*

Layout: Circle Node Labels ☑

# Visualization by cluster

**Legends**

**Biological process**
- cell_organization_and_biogenesis
- response_to_stress

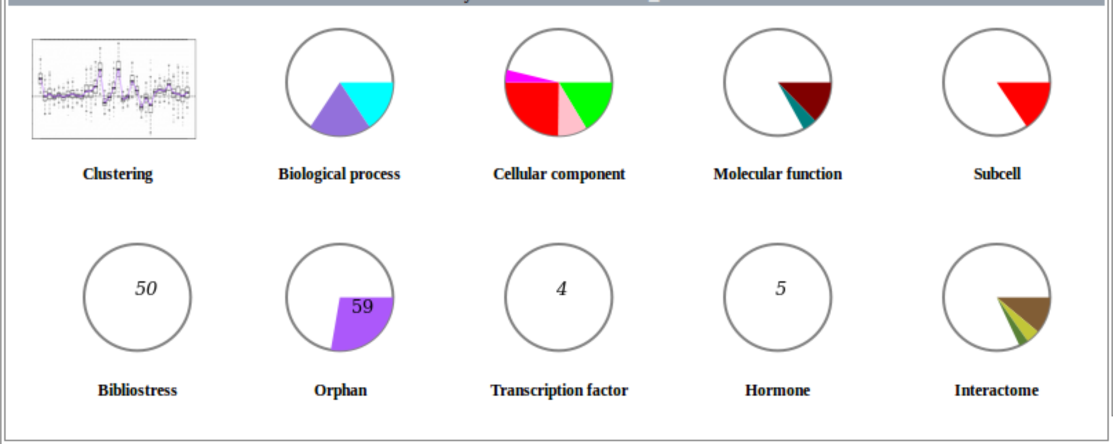
Cellular component
- chloroplast
- cytosol
- mitochondria
- ribosome

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

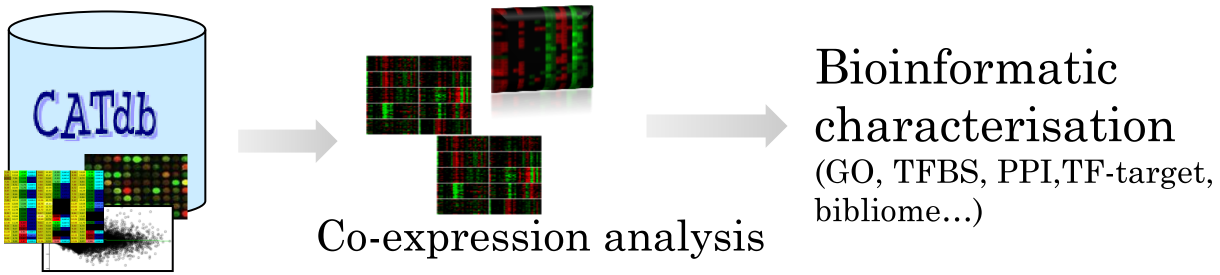
| Biological process | nb genes | pvalue | Ref nb |
|---|---|---|---|
| cell_organization_and_biogenesis | 36 | 1.17e-3 | 3367 |
| response_to_stress | 43 | 5.84e-4 | 4117 |
| **Cellular component** | **nb genes** | **pvalue** | **Ref nb** |
| chloroplast | 39 | 3.16e-3 | 3959 |
| cytosol | 21 | 2.50e-3 | 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**

other enrichments:

| gene_name | product | terms | 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 | ● ● | ● | ● | | | | ● | | | ● |

# First conclusions




When considering thousands of genes, Pearson correlation is not the best tool

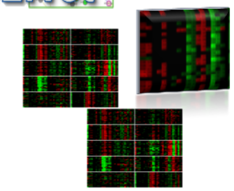
This large-scale co-expression study

- generates biologically meaningful clusters

-performs favorably as compared to those obtained with correlation-based approaches

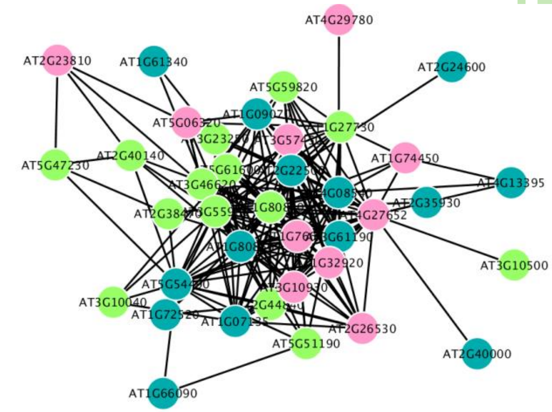
# Functional inference by coregulation

GEM2Net



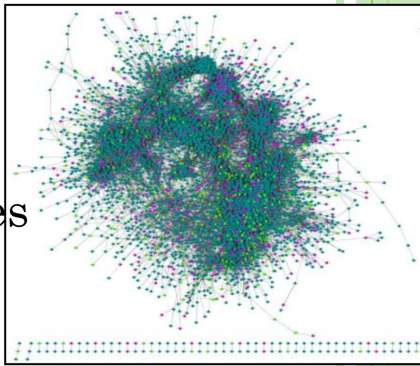
**Integration**

**across the 18 stress categories**



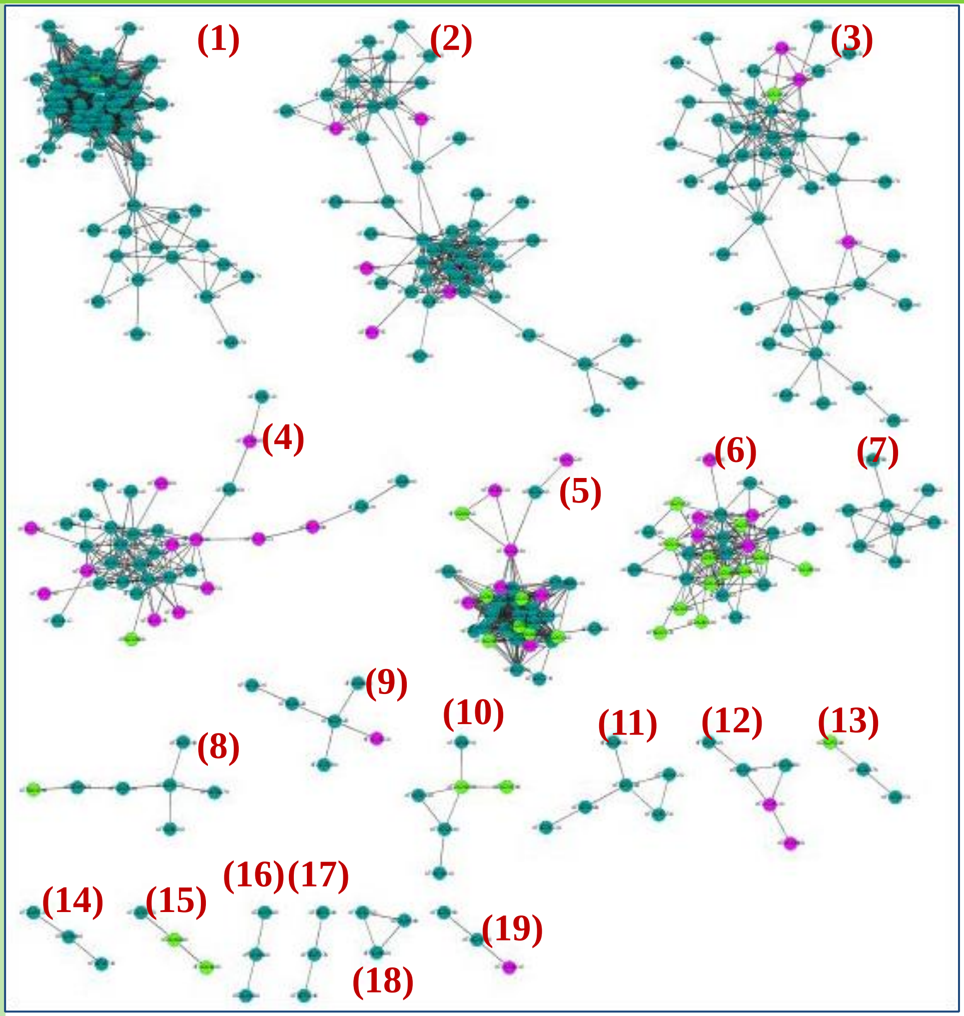
Comparing the coregulation network with a random network shows that a pair observed more than 3 times is statistically significant and has probably a biological meaning

Network with gene pairs conserved in at least 3 stresses:
5 626 genes with 713 orphans and 1682 partially annotated genes



**Identification of Coregulated genes** → **Describe groups of functional partners** → **Annotation of orphan genes**

# Coregulation network



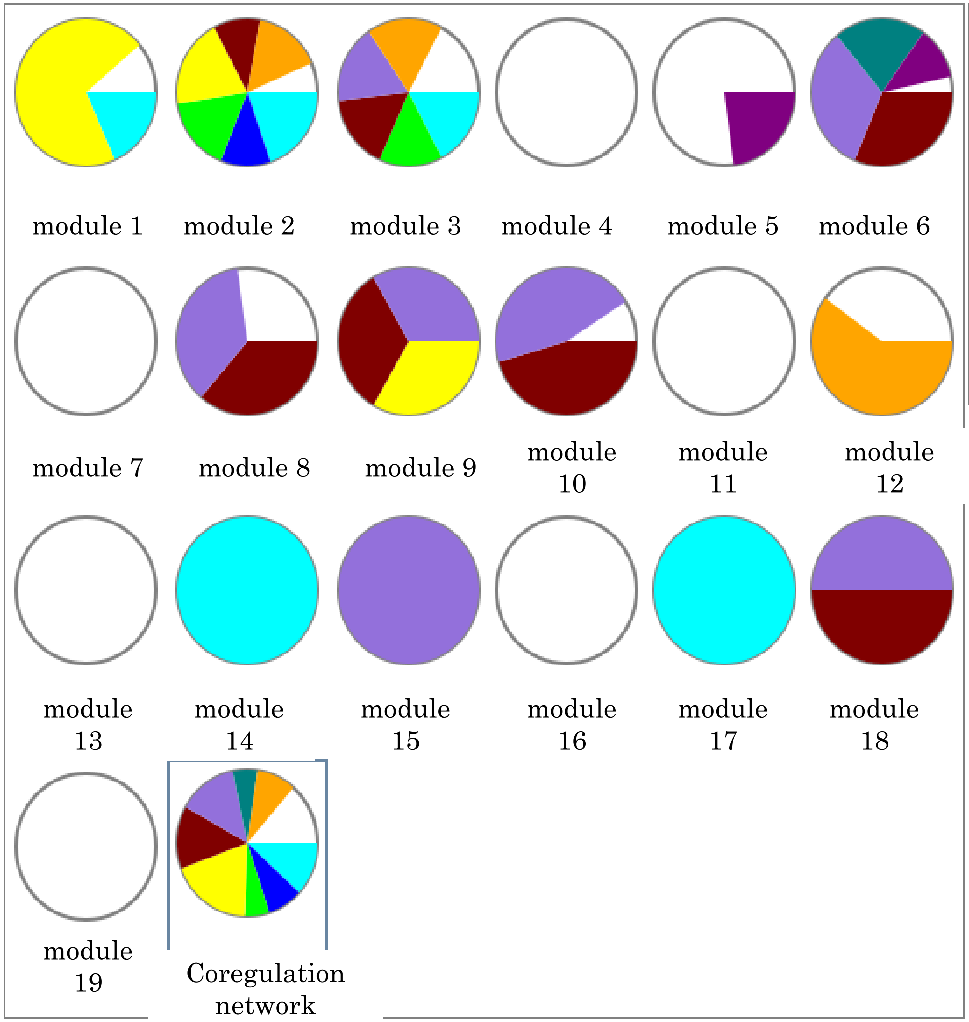


The network with gene pairs conserved in at least 7 stresses is the first network showing connected components

415 genes with 41 orphan genes, 1908 interactions

# Identification of functional modules

| Legends | |
|---|---|
| cell_organization_and_biogenesis | |
| developmental_processes | |
| electron_transport_or_energy_pathways | |
| protein_metabolism | |
| response_to_abiotic_or_biotic_stimulus | |
| response_to_stress | |
| signal_transduction | |
| transcription_DNA_dependent | |
| transport | |




**Biological_process**

6 genes in cluster_10

| term name | nb | p-value | Ref Term |
|---|---|---|---|
| response_to_abiotic_or_biotic_stimulus | 5 | 9.12e-5 | 3736 |
| response_to_stress | 5 | 1.51e-4 | 4139 |
| cell_organization_and_biogenesis | 1 | | |
| other_biological_processes | 4 | | |
| other_cellular_processes | 4 | | |
| other_metabolic_processes | 3 | | |
| signal_transduction | 2 | | |
| transport | 2 | | |
| unknown_biological_processes | 1 | | |

Coregulation modules are more specific and more homogeneous

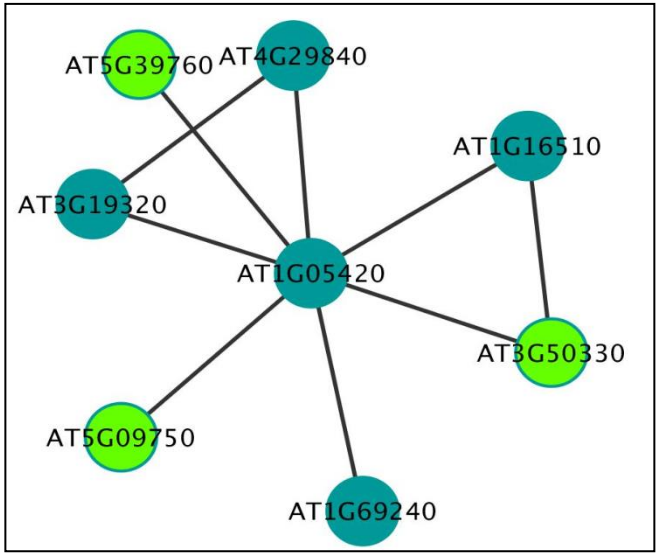
Cis-regulatory motifs are found in their promoters

**Topological analysis = a relevant approach to identify functional modules**

# Example of functional annotation

Network with gene pairs conserved in at least 13 stresses
( 8 genes, 9 interactions)

These genes are not known to be coregulated




BUT

6 genes share a same TFBS
indicating that they are under the control
of a same TF

This motif corresponds to EIN3, involved
in the regulation of important immune components

# Conclusions

- Model-based clustering allows to understand data better than pair-based methods

- Working with homogeneous data is really an ideal framework

- All the coexpression studies are available in and published in Zaag et al (2015) in NAR

- Modules are relevant to perform a functional annotation