Transfert de connaissances par **projections multivariées** transcriptomiques sur une base de données de **référence multi-stress** chez ***Arabidopsis thaliana***

Encadrée par Dr. Andrea Rau, chargée de recherche

Contexte

Contexte changement climatique :
Sélectionner plantes résistantes aux stress
- ➢ intérêt agronomique fort

Souche maïs résistante au froid

Plantes modèles extrêmement bien connues, contrôlées
Genevestigator, GeneMania, MapMan, ATTED-II

Transfert de connaissances possible réponse transcriptomique au stress d'A. thaliana ?

GEM2Net

Construction d'un modèle permettant projections de nouvelles données (Arabidopsis ou autres espèces)

Contexte :
GEM2Net : base de données de stress de référence pour A. thaliana

CATdb

Publiée en 2008
231 projets en 2015 pour Arabidopsis et **51** pour les autres espèces
371 en 2021 pour Arabidopsis et **97** pour les autres espèces *(06/2021)*

Caractéristiques GEM2Net :
- Données trancriptomiques homogènes
- Regrouper les gènes en fonction unités co-expression
- Outils de visualisation clusters et métadonnées

GEM2Net

Contexte :
GEM2Net : base de données de stress de référence pour A. thaliana

Publiée en 2008
231 projets en 2015 pour Arabidopsis et **51** pour les autres espèces
371 en 2021 pour Arabidopsis et **97** pour les autres espèces *(06/2021)*

Caractéristiques GEM2Net :
- Données trancriptomiques homogènes
- Regrouper les gènes en fonction unités co-expression
- outils de visualisation clusters et métadonnées

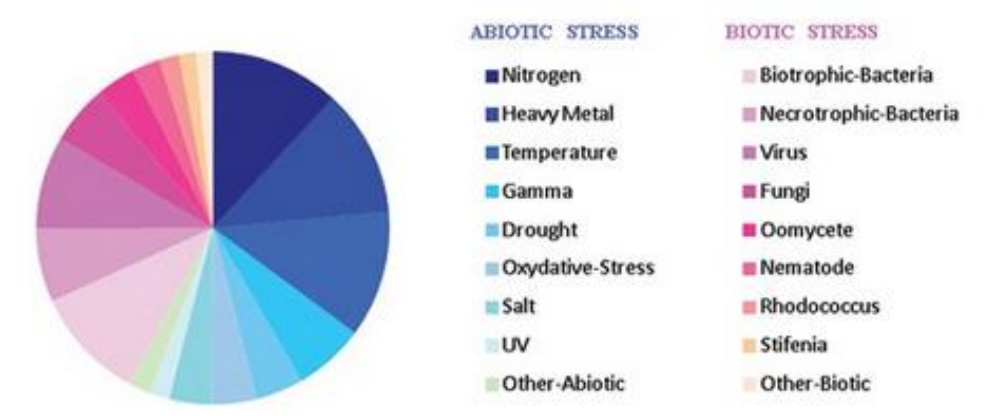


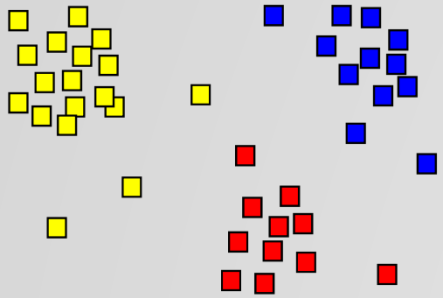
Fig. 1 Zaag et al.

Drought	Gamma	Heavy Metal	Nitrogen	Other abiotic	Oxydative stress	Salt	Temperature	UV
17	25	45	46	8	16	15	45	7

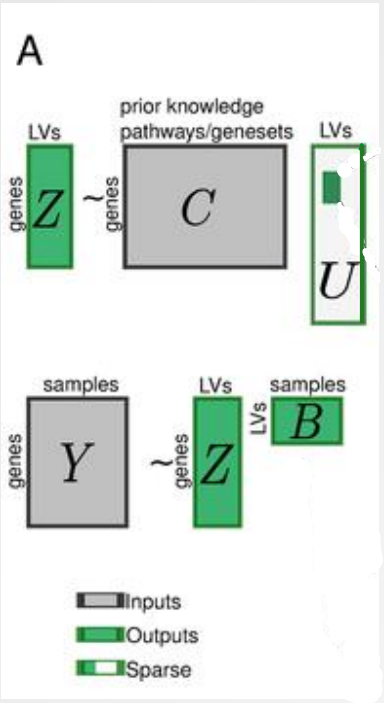
Biotrophic Bacteria	Fungi	Necrotrophic Bacteria	Nematode	Oomycète	Other biotic	Rhodococcus	Stifenia	Virus
40	21	26	10	14	6	7	6	33

Un transfert de connaissances à partir des données GEM2Net est-il possible ?

En utilisant les catégories de stress de la base de données ?

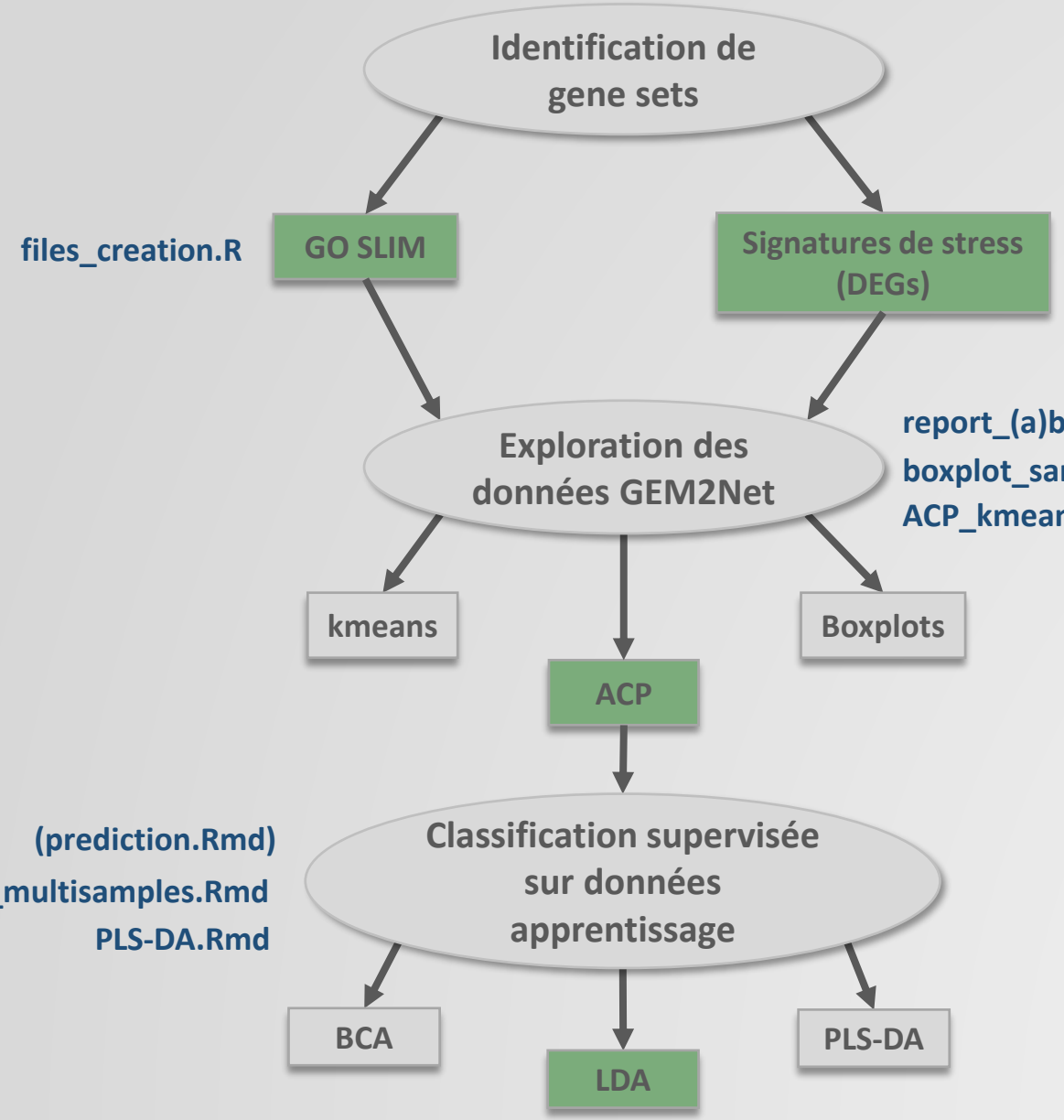


En passant par une déconvolution non supervisée pour faire ressortir un signal biologique ?



Issue article de MAO W. et al.

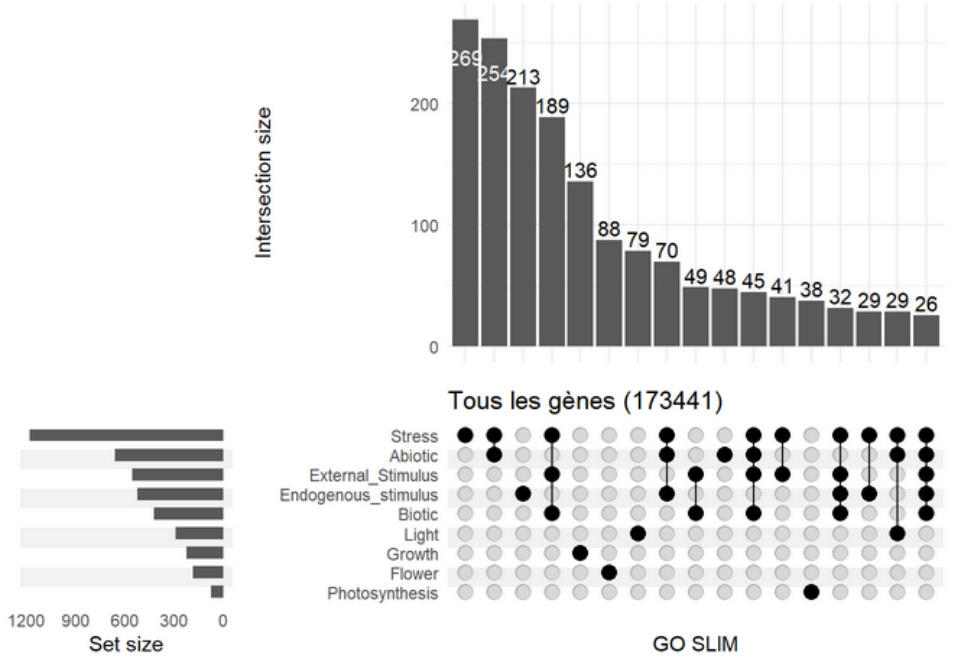
Classification (non)-supervisée des stress



Identification des gene sets

report_abiotic.html
upset_signature.R

GO SLIM : Gene Ontology



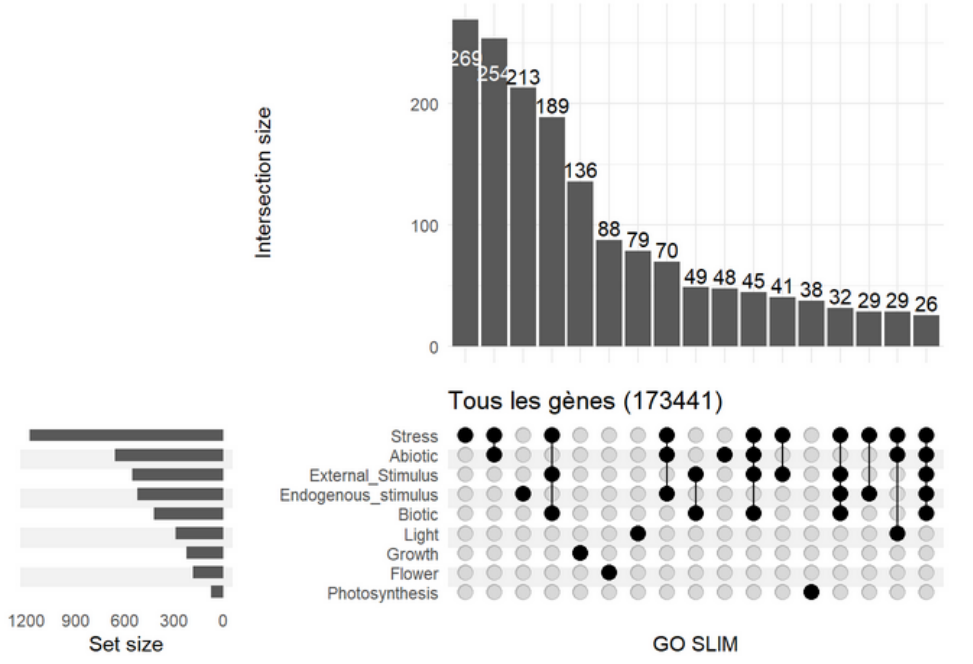
- **Rythme circadien** (57 gènes)
- **Floraison** (186 gènes)
- **Croissance** (222 gènes)
- **Photosynthèse** (75 gènes)
- **Stimulus abiotique** (661 gènes)
- **Stimulus biotique** (420 gènes)
- **Stimulus endogène** (523 gènes)
- **Stimulus externe** (552 gènes)
- **Lumière** (292 gènes)
- **Stress** (1177 gènes)

+ Set aléatoire (50 gènes)

Identification des gene sets

report_abiotic.html
upset_signature.R

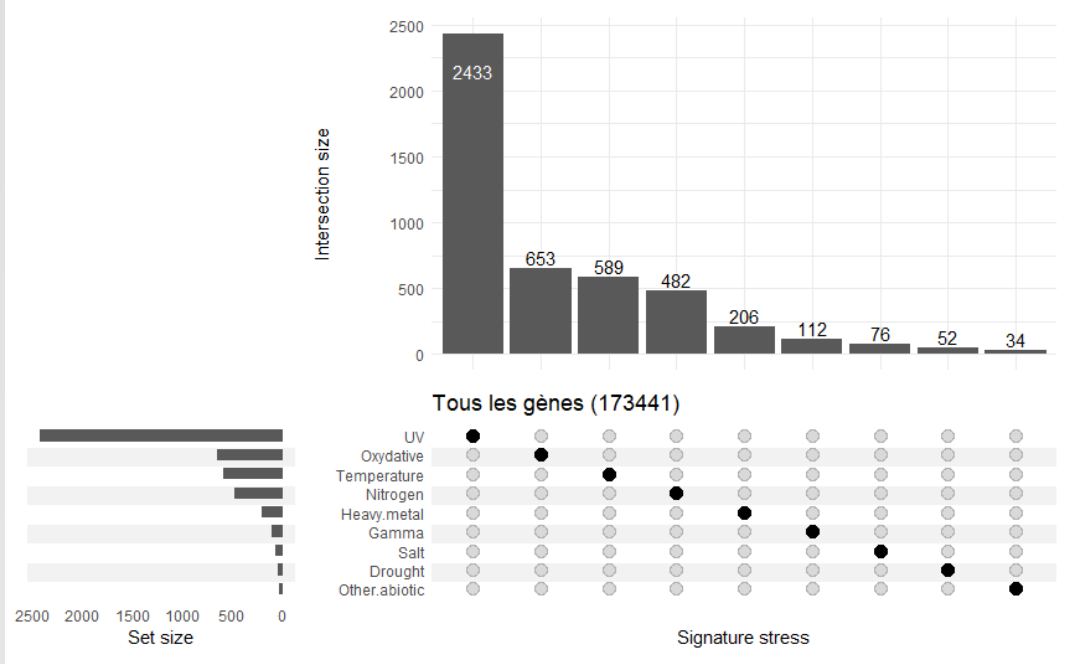
GO SLIM : Gene Ontology



- **Rythme circadien** (57 gènes)
- **Floraison** (186 gènes)
- **Croissance** (222 gènes)
- **Photosynthèse** (75 gènes)
- **Stimulus abiotique** (661 gènes)
- **Stimulus biotique** (420 gènes)
- **Stimulus endogène** (523 gènes)
- **Stimulus externe** (552 gènes)
- **Lumière** (292 gènes)
- **Stress** (1177 gènes)

+ Set aléatoire (50 gènes)

Analyse différentielle



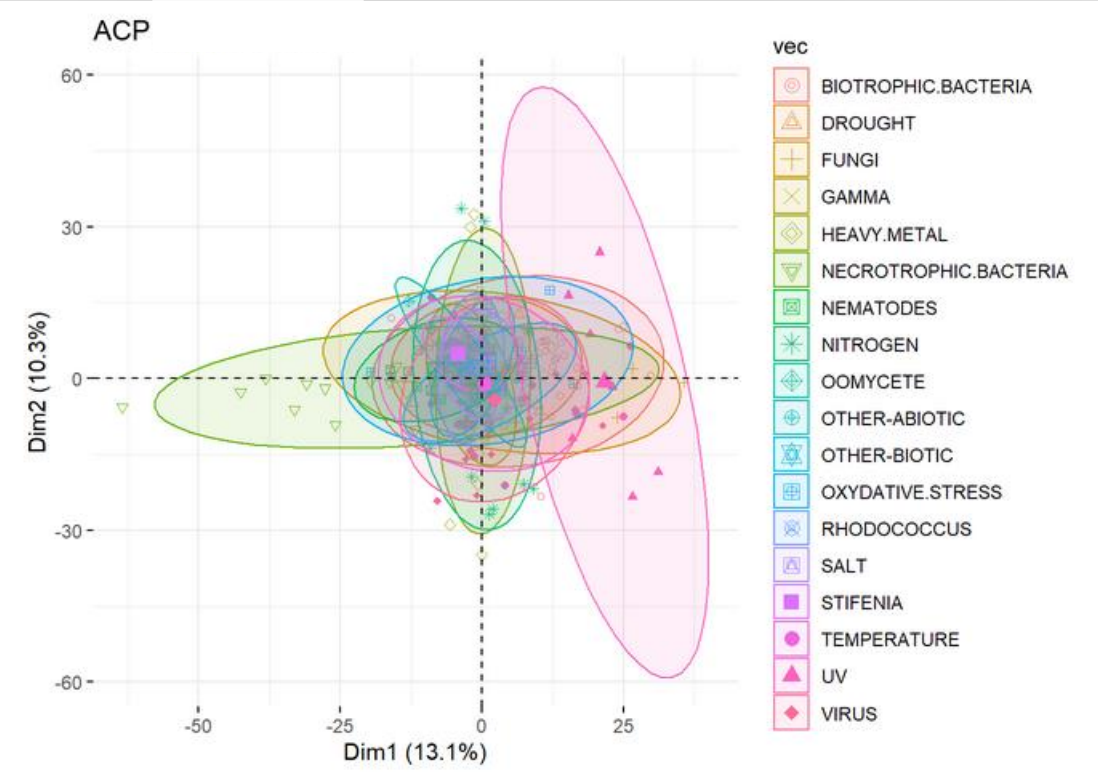
Package limma (Linear Models for MicroArray data) :
p-value ajustée méthode de Benjamini et Hochberg
seuil *p-value* 0,05,

Occurence	0	1	2	3	4	5	6	7	8	9
Nombre DEG	9556	4636	2159	661	224	86	11	5	2	1

Résultats

Premières observations :
Séparation entre les différents stress possible ?

abiotic.html



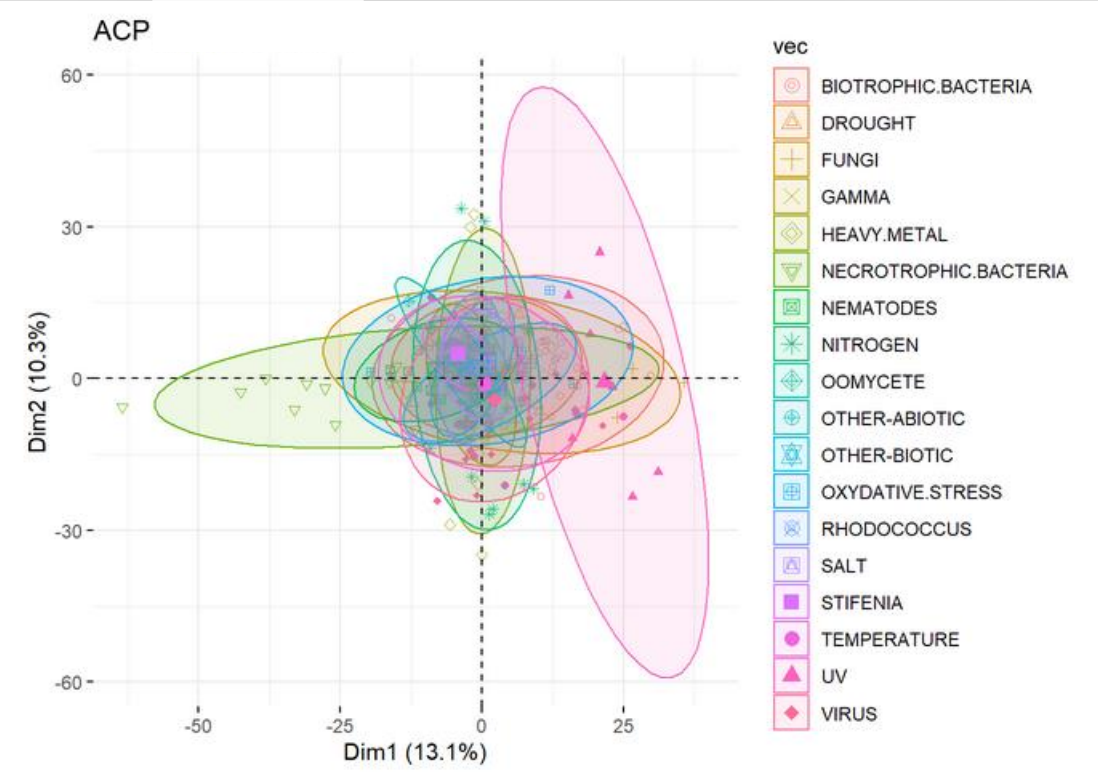
ACP globale sur les données d'expressions, pour l'ensemble des catégories de stress (387 échantillons, 17341 gènes)

FACTOMINER

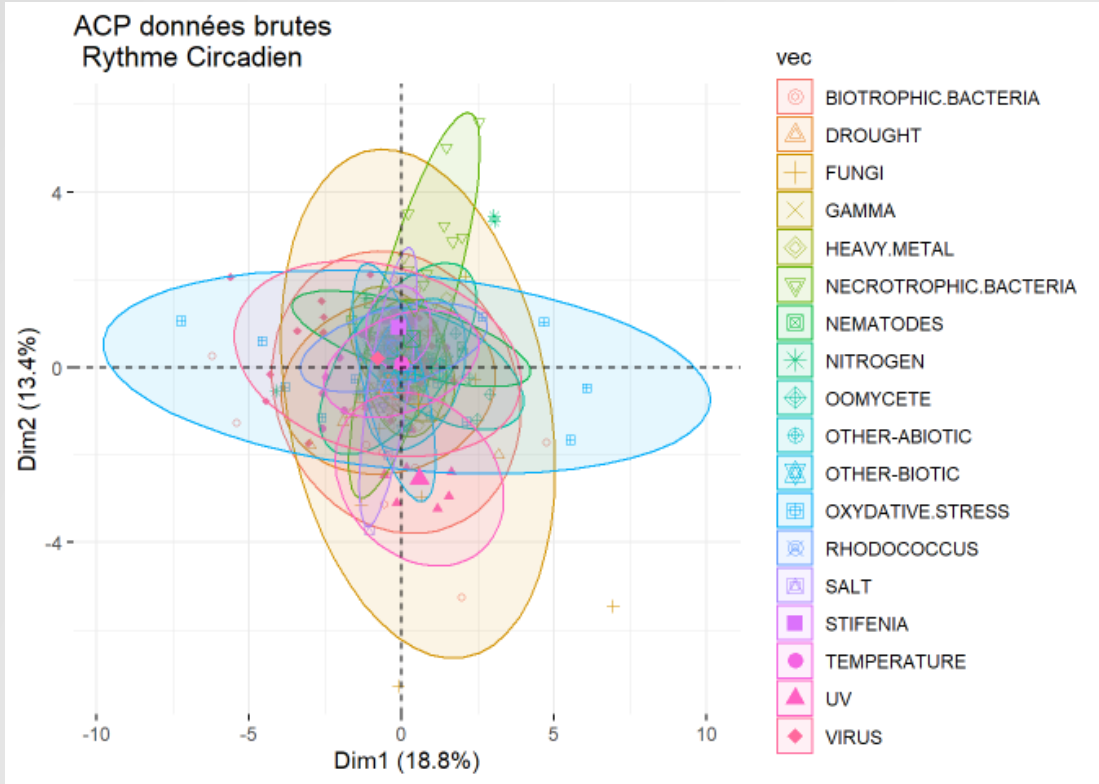
Résultats

Premières observations :
Séparation entre les différents stress possible ?

abiotic.html



ACP globale sur les données d'expressions, pour l'ensemble des catégories de stress (387 échantillons, 17341 gènes)

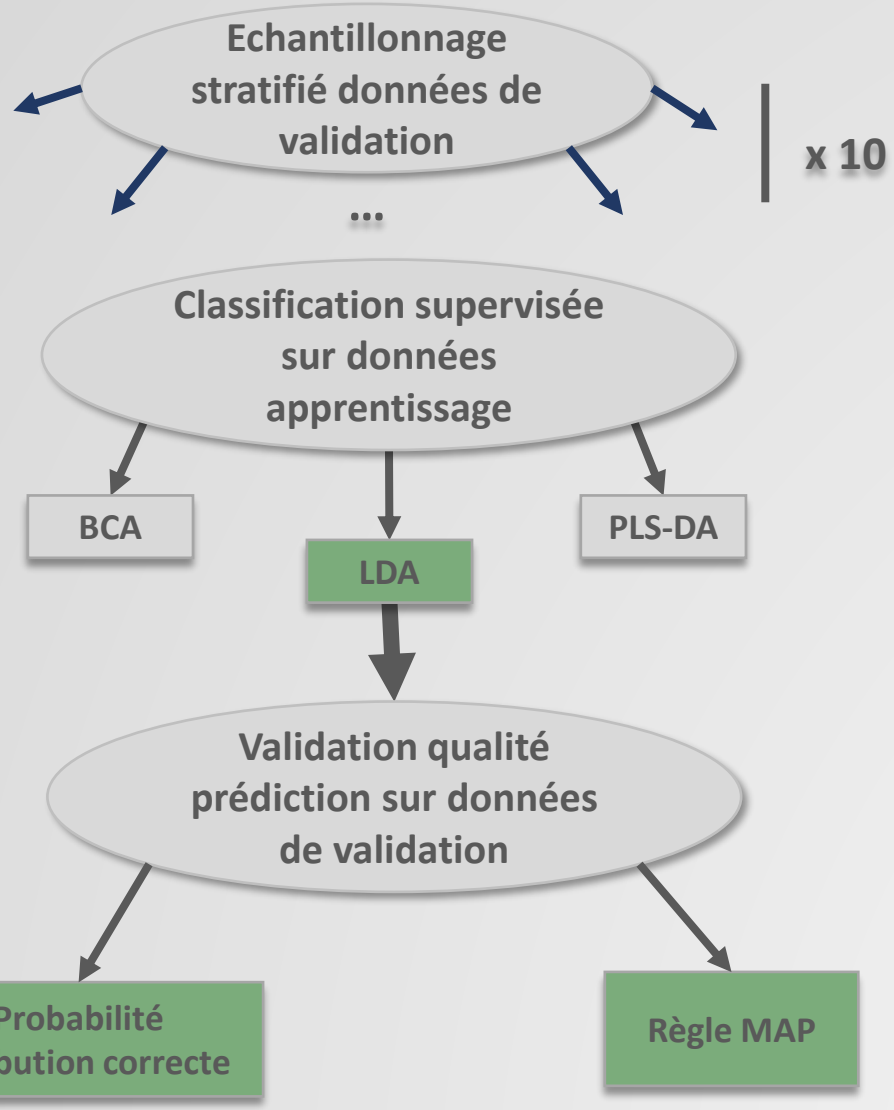


ACP globale sur les données d'expression, pour les 57 gènes du terme GO SLIM rythme circadien

Classification supervisée : BCA, kmeans, boxplots

FACTOMINER

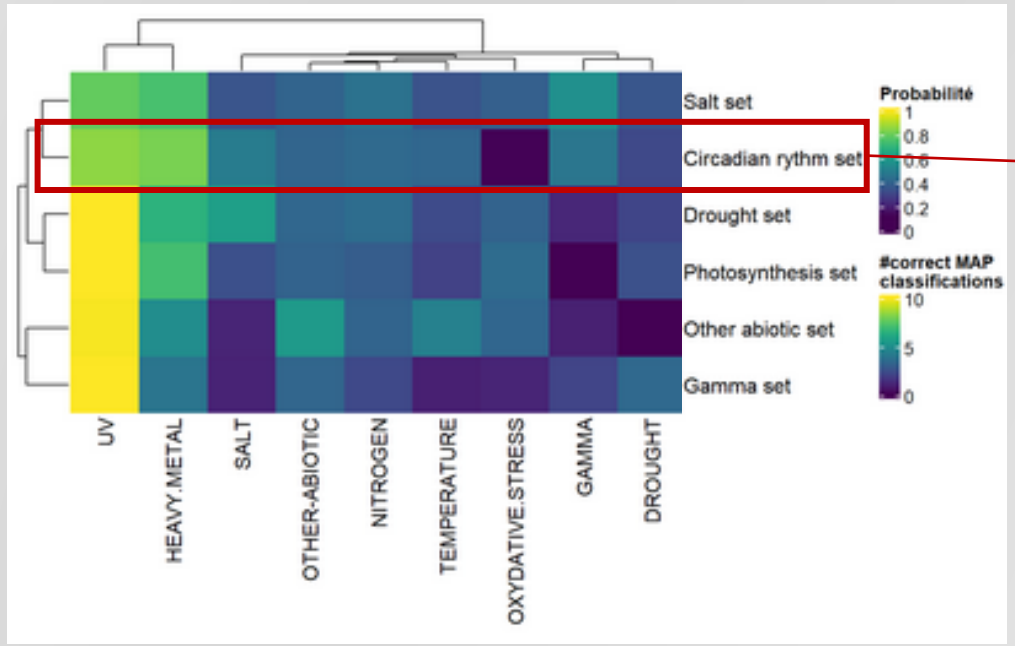
Linear Discriminant Analysis
Construction du modèle



Résultats

Linear Discriminant Analysis
Prédiction sur données d'expression

LDA_multisamples.html



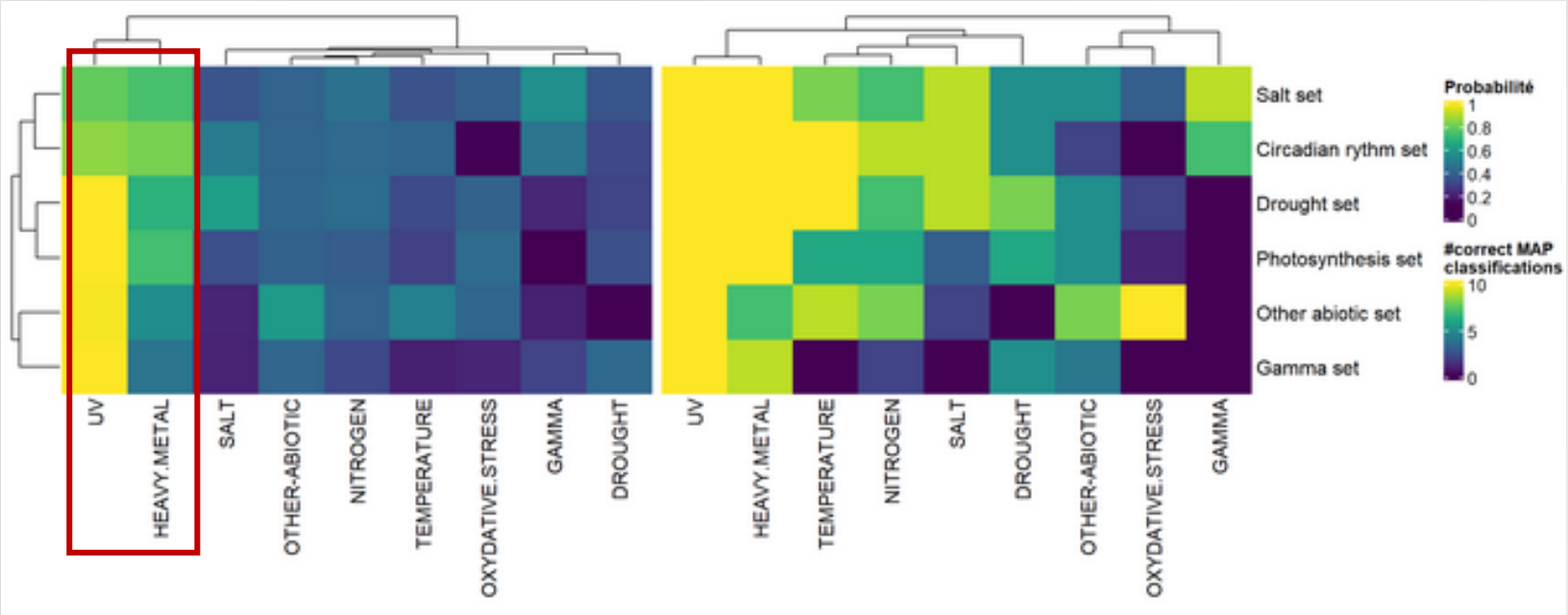
	Catégories de stress								
	DROUGHT	GAMMA	HEAVY.METAL	NITROGEN	OTHER-ABIOTIC	OXYDATIVE.STRESS	SALT	TEMPERATURE	UV
DROUGHT	0.3038	0.1895	0.0528	0.0248	0.0022	0.0132	0.1067	0.3070	0.0000
GAMMA	0.0003	0.4592	0.1191	0.0171	0.0006	0.0000	0.0004	0.4029	0.0000
HEAVY.METAL	0.0348	0.0212	0.8210	0.0459	0.0005	0.0007	0.0026	0.0727	0.0000
NITROGEN	0.0122	0.0062	0.1531	0.4137	0.0000	0.1526	0.0037	0.2584	0.0000
OTHER-ABIOTIC	0.0155	0.0018	0.0062	0.0036	0.3986	0.0000	0.2729	0.3014	0.0000
OXYDATIVE.STRESS	0.0044	0.0235	0.3675	0.2704	0.1994	0.1163	0.0001	0.0183	0.0000
SALT	0.0219	0.0672	0.1489	0.1556	0.0150	0.0007	0.4790	0.1115	0.0000
TEMPERATURE	0.0053	0.2188	0.1188	0.2338	0.0024	0.0001	0.0144	0.4062	0.0000
UV	0.0000	0.0030	0.0036	0.0000	0.0000	0.0000	0.0000	0.1431	0.8505

Matrice probabilité rythme circadien (moyenne sur 10 échantillons)

Résultats

Linear Discriminant Analysis
Prédiction sur données d'expression

LDA_multisamples.html



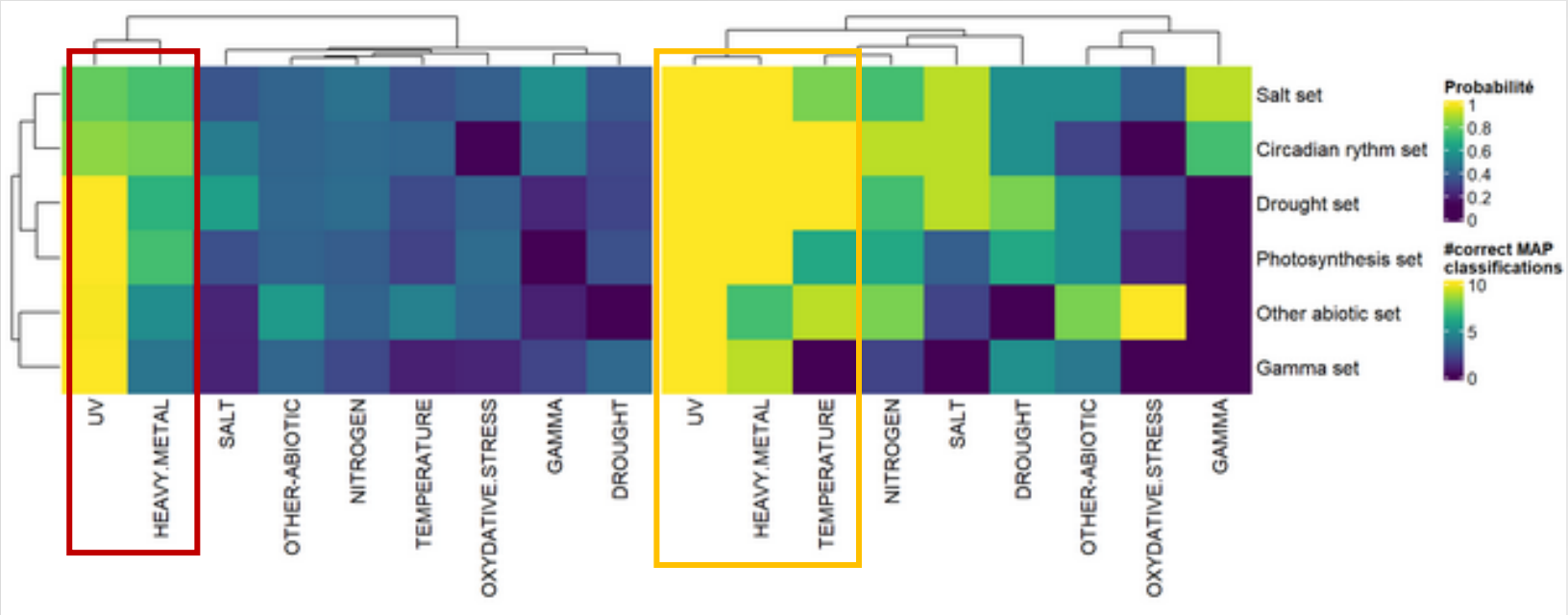
Heatmap des résultats de LDA :
- **Confiance de prédiction (à gauche),**
- **Proportion d'échantillons bien classés (/10) (à droite)**

Stress avec plus grande confiance de prédiction

Résultats

Linear Discriminant Analysis
Prédiction sur données d'expression

LDA_multisamples.html



Heatmap des résultats de LDA :
- **Confiance de prédiction (à gauche),**
- **Proportion d'échantillons bien classés (/10) (à droite)**

Stress avec plus grande confiance de prédiction

Stress les mieux classés sur les 10 échantillons

PLS-DA

Déconvolution non supervisée structure avec incorporation connaissances biologiques : PLIER

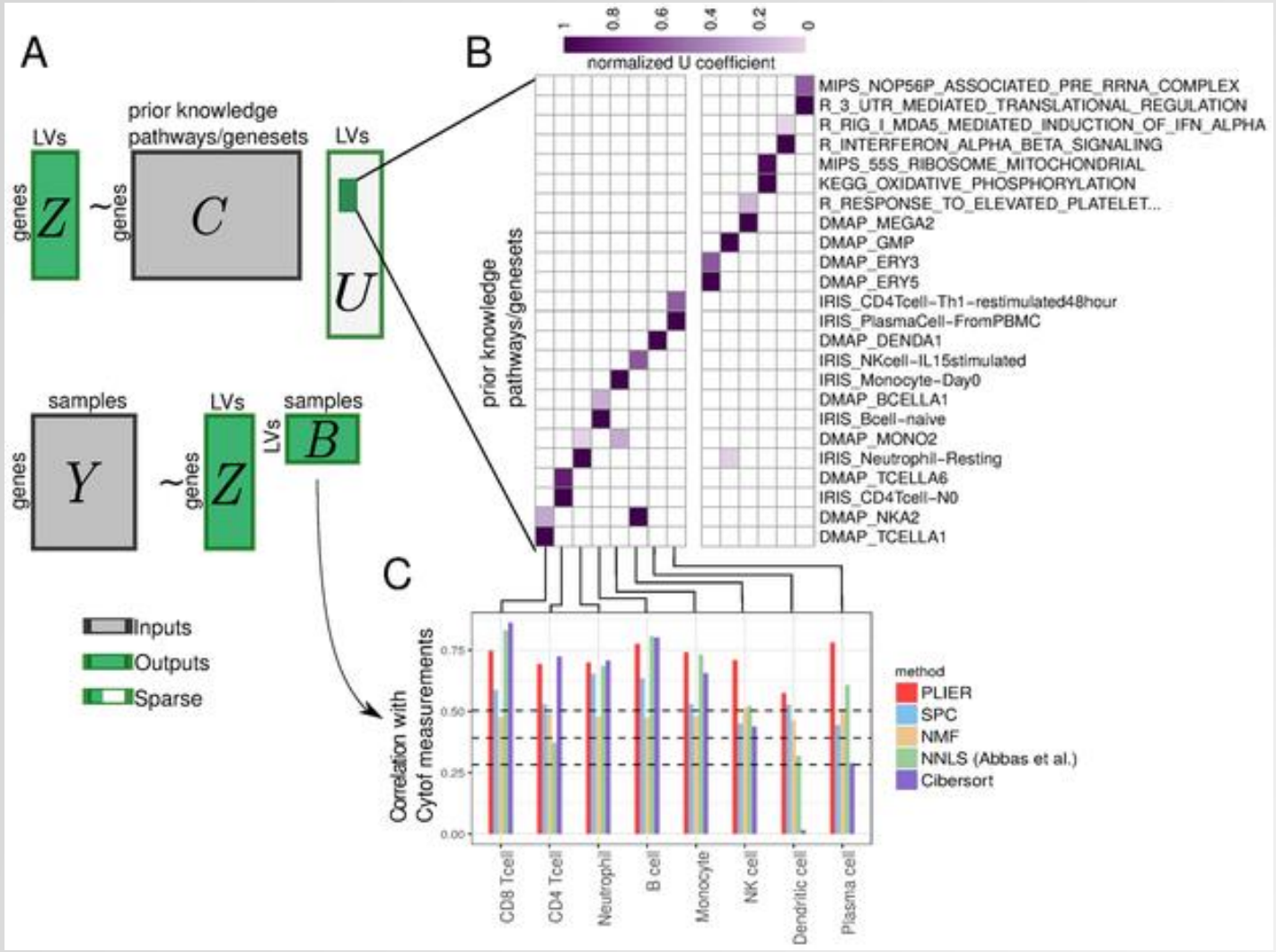
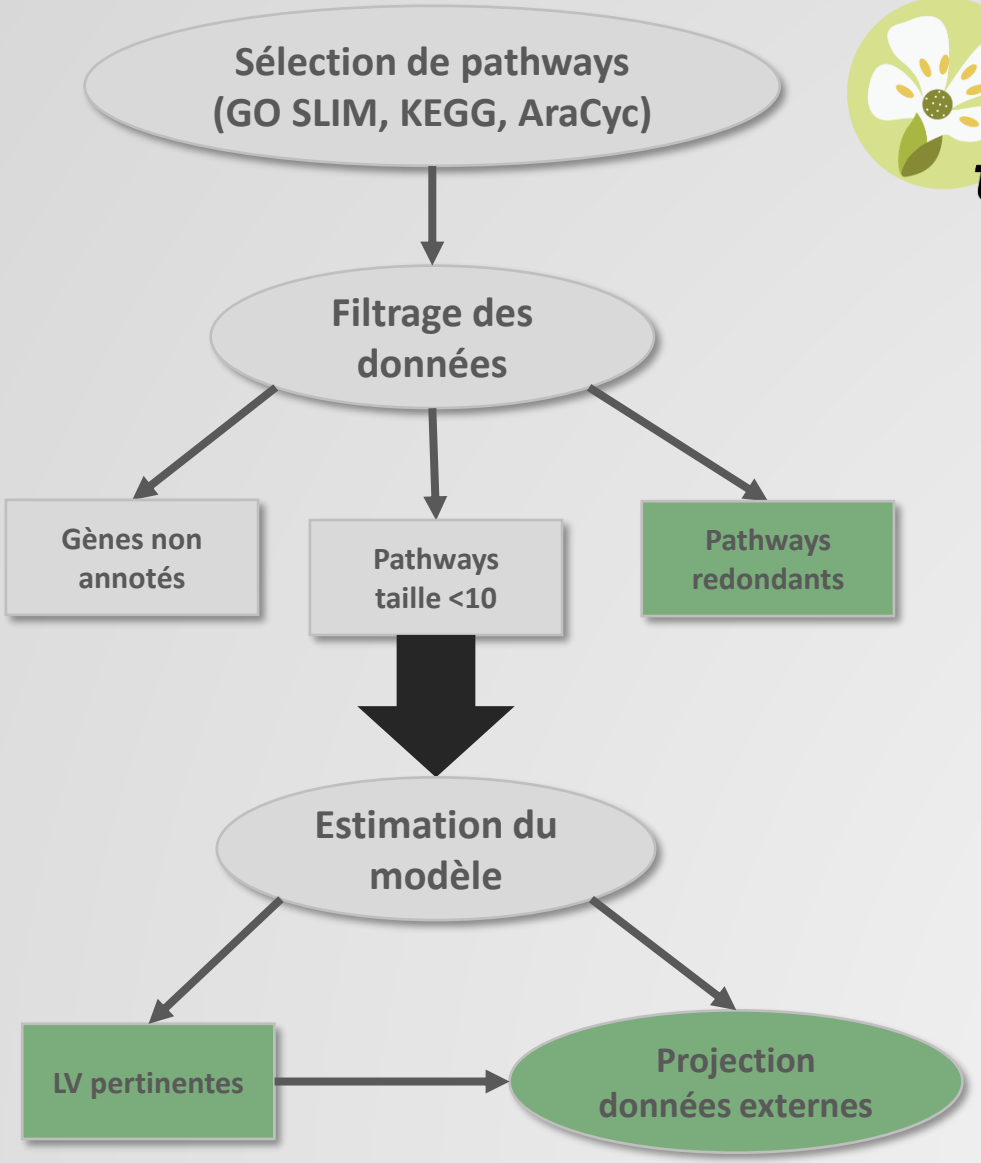


Figure issue de l'article Mao W. et al.

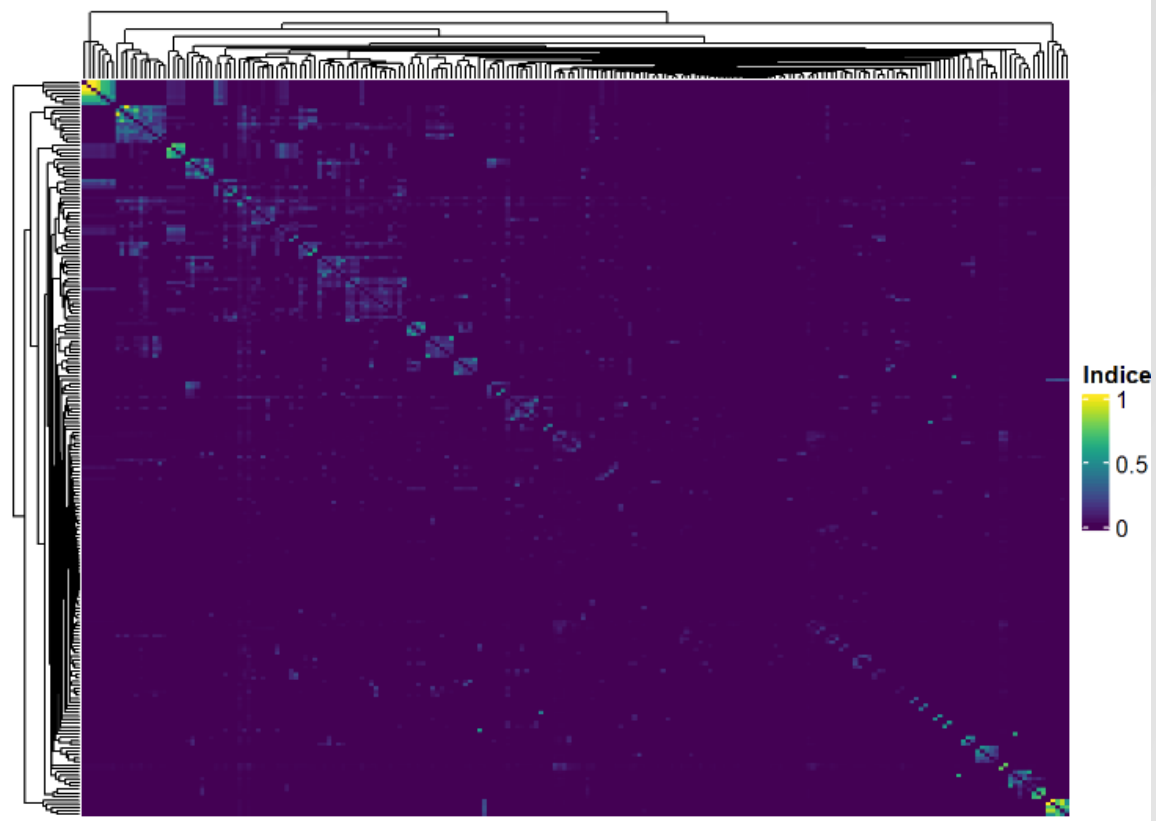
Mise en place stratégie PLIER

PLIER.html

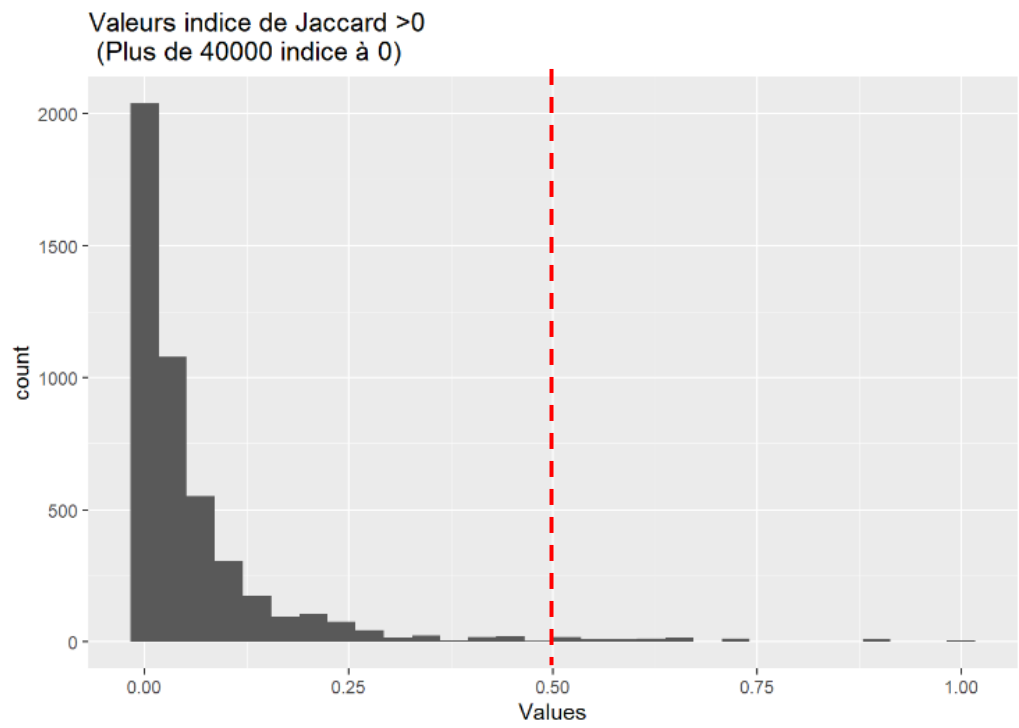


MultiPLIER

Construction du modèle
Choix des pathways et filtrage

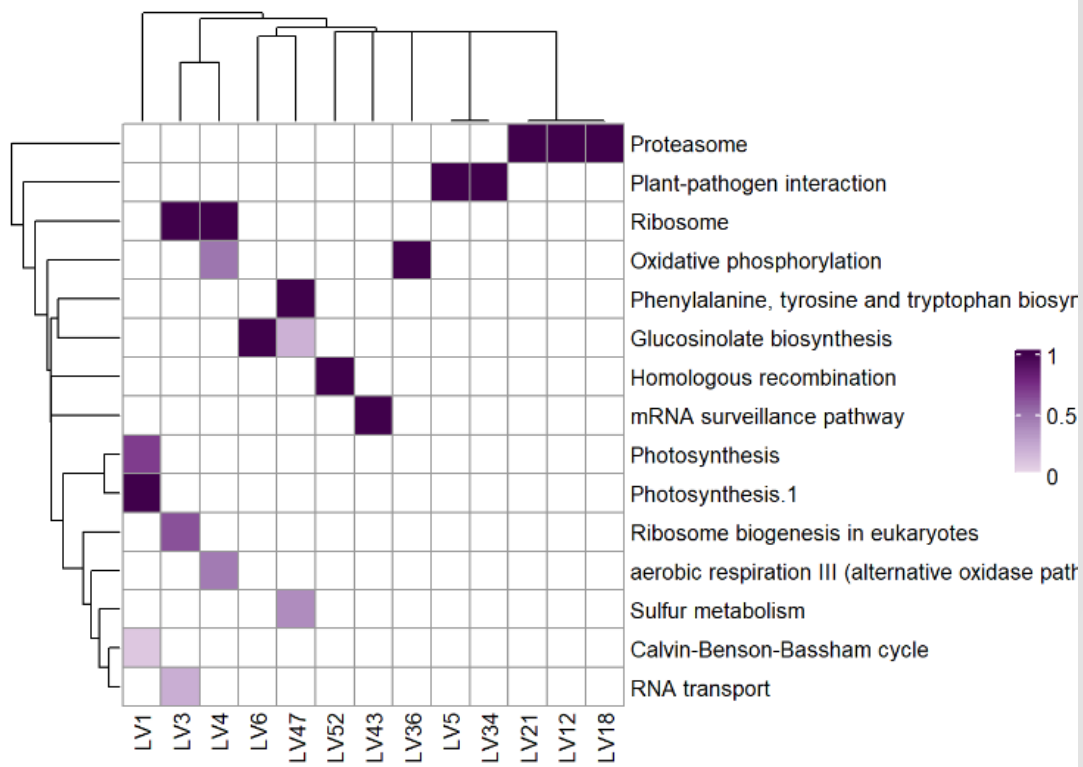


Heatmap indice de Jacard pour les 210 pathways choisis



Répartition des valeurs de l'indice de Jaccard >0.

Modèle PLIER
Variables latentes significativement reliées à des pathways



Visualisation Matrice U, pour les variables latentes significatives (AUC > 0,75)

	pathway	LV index	AUC	p-value	FDR	Nb genes	Database
1	Calvin-Benson-Bassham cycle	1	0.8369377	0.0014948	0.0047737	27	AraCyc
2	Photosynthesis	1	0.7659314	0.0051340	0.0123968	33	KEGG
6	Photosynthesis.1	1	0.8695934	0.0000000	0.0000001	75	GOSLIM
10	Ribosome biogenesis in eukaryotes	3	0.7893658	0.0000248	0.0001540	68	KEGG
11	Ribosome	3	0.9583288	0.0000000	0.0000000	193	KEGG
12	RNA transport	3	0.8132661	0.0000000	0.0000001	115	KEGG
14	RNA polymerase	3	0.9179986	0.0001366	0.0007511	27	KEGG
17	aerobic respiration III (alternative oxidase pathway)	4	0.7598522	0.0034087	0.0091207	36	AraCyc
19	Oxidative phosphorylation	4	0.8062127	0.0000034	0.0000277	75	KEGG
20	Ribosome	4	0.9172438	0.0000000	0.0000000	193	KEGG

Reprise summary du modèle PLIER

Gene Expression Omnibus

High-resolution temporal profiling of transcripts during Arabidopsis leaf senescence reveals a distinct chronology of processes and regulation.

Breeze E, et al.

Plant Cell 2011 Mar;23(3):873-94.

CATMA V3

11 jours d'étude

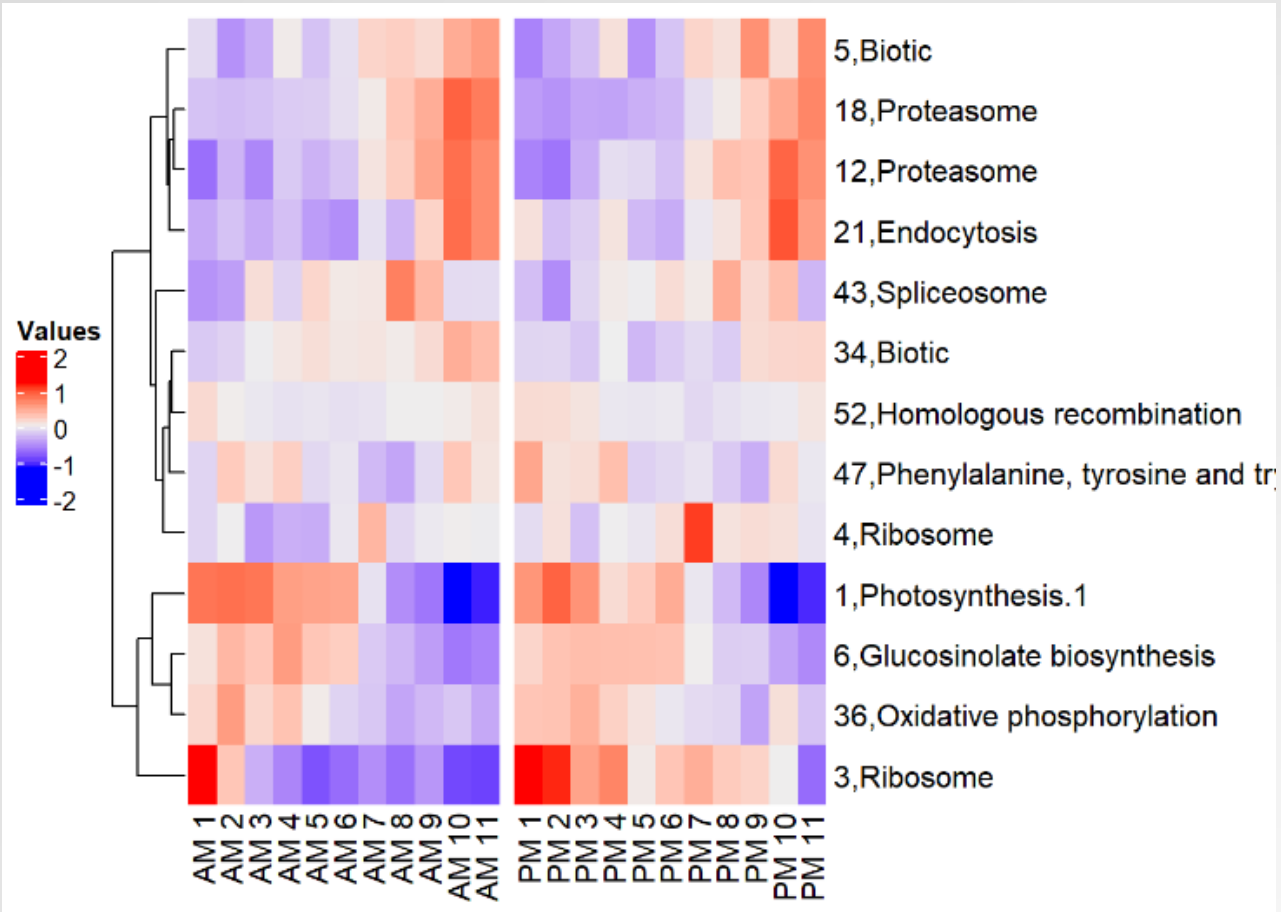
2 valeurs par jour :
- matin : 7h d'exposition lumière
- après-midi : 14h d'exposition

Raw data :
Moyenne sur 4 réplicats

Modèle PLIER
Intégration de nouvelles données

Calcul de la nouvelle matrice : GetNewDataB()

$$(Z^t.Z + \lambda_2.\mathrm{D})^{-1}.Z^t.Y$$



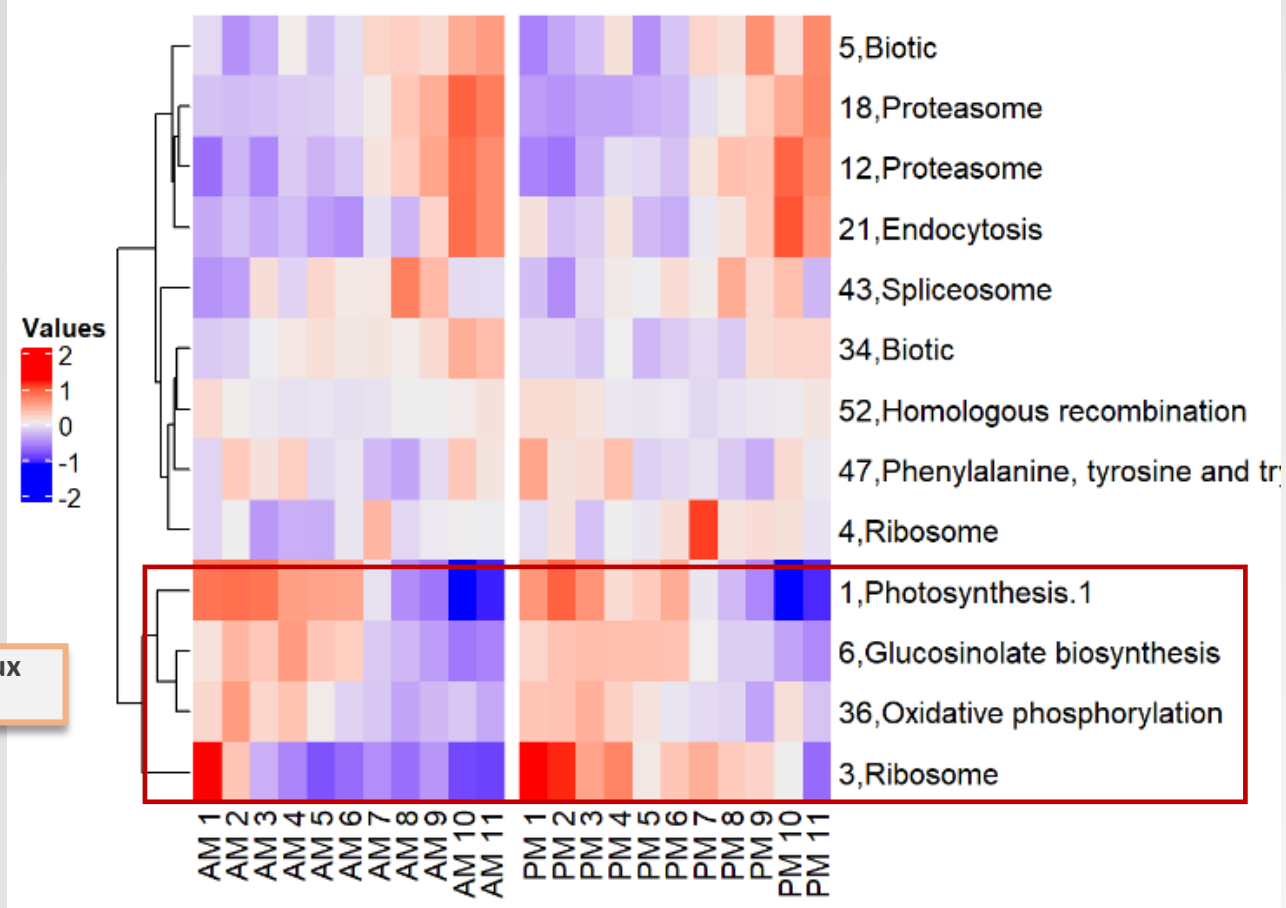
Heatmap nouvelle matrice B pour les variables latentes significatives

Modèle PLIER
Intégration de nouvelles données

Calcul de la nouvelle matrice : GetNewDataB()

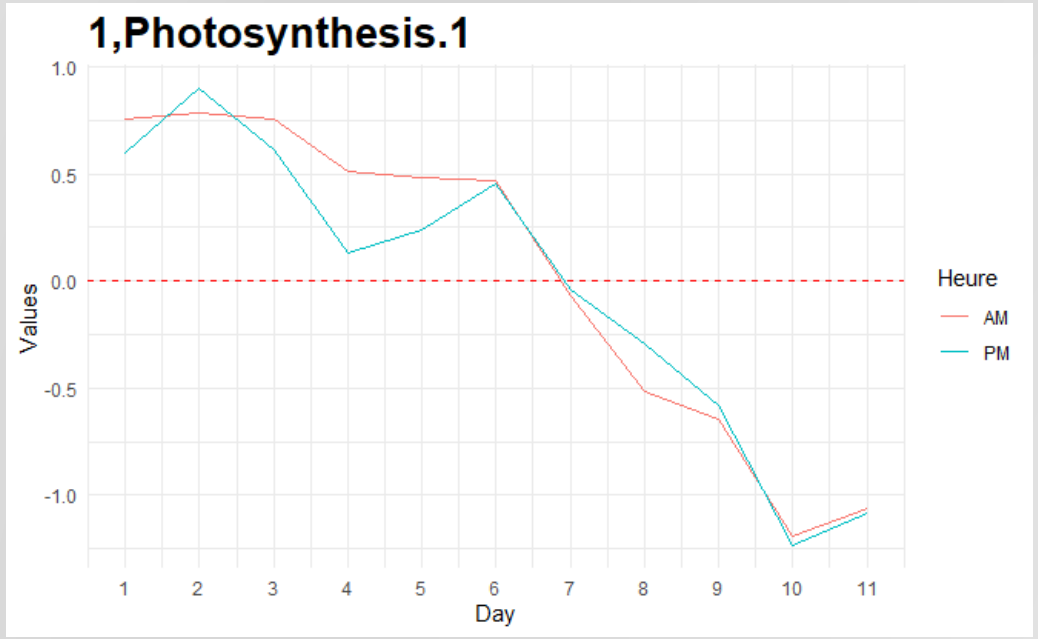
$$(Z^t.Z + \lambda_2.\mathrm{D})^{-1}.Z^t.Y$$

Pathways les plus variables dans les nouveaux échantillons

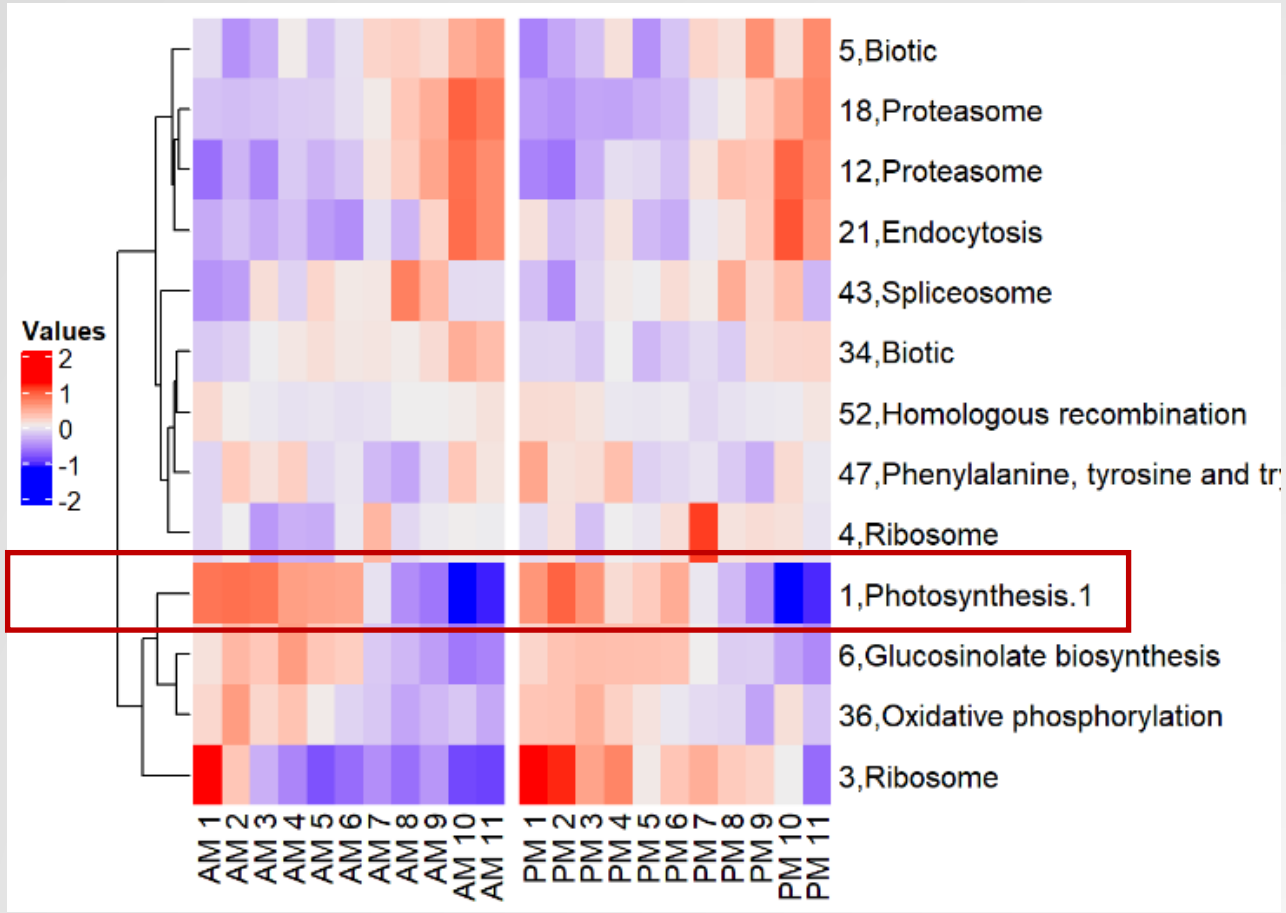


Heatmap nouvelle matrice B pour les variables latentes significatives

Modèle PLIER
Intégration de nouvelles données



Plotline des valeurs de la matrice B pour la variable latente 1



Heatmap nouvelle matrice B pour les variables latentes significatives

Conclusions et perspectives

Bilan première stratégie :

- Utilisation difficile des catégories de stress : manque homogénéité, recoupement
- La LDA pertinente pour 3 catégories de stress seulement
- Gene set pas assez spécifiques
- Effet projet important

Bilan deuxième stratégie :

- Modèle a 13 variables latentes significatives
- Dimensions données input similaires avec données présentées vignettes
- Signal LV photosynthèse retrouvé avec autres données hors GEM2Net (MultiPLIER)

Perspectives :

- Repenser les catégories de stress, nouvelle variable ?
- Etude bibliographique, fiches projets
- Approfondir l'étude des variables latentes de PLIER : autres données, indicateur quantitatif
- Utilisation données autres espèces ? Méthode(s) à utiliser ?
- Transmission scripts et fichiers (GitHub)

Merci de votre attention

(1) Zaag R et al. , **GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response.** *Nucleic Acids Res*. 2015 Jan doi: 10.1093/nar/gku1155. Epub 2014 Nov 11.

(2) Séverine Gagnot et al., **CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform,** *Nucleic Acids Research*, Volume 36, Issue suppl_1, 1 January, https://doi.org/10.1093/nar/gkm757

(3) Mao W. *et al.* **Pathway-level information extractor (PLIER) for gene expression data**. *Nat Methods* **16,** 607–610 (2019). https://doi.org/10.1038/s41592-019-0456-1

(4) Taroni JN et al. **MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease.** *Cell Syst*. 2019 doi:10.1016/j.cels.2019.04.003

Remerciements

L'UMRt BioEcoAgro, en particulier :

Andrea RAU

Joël Léonard

Isabelle Lejeune

Catherine Giauffret

Aline Waquet

Marjolaine Becquerelle

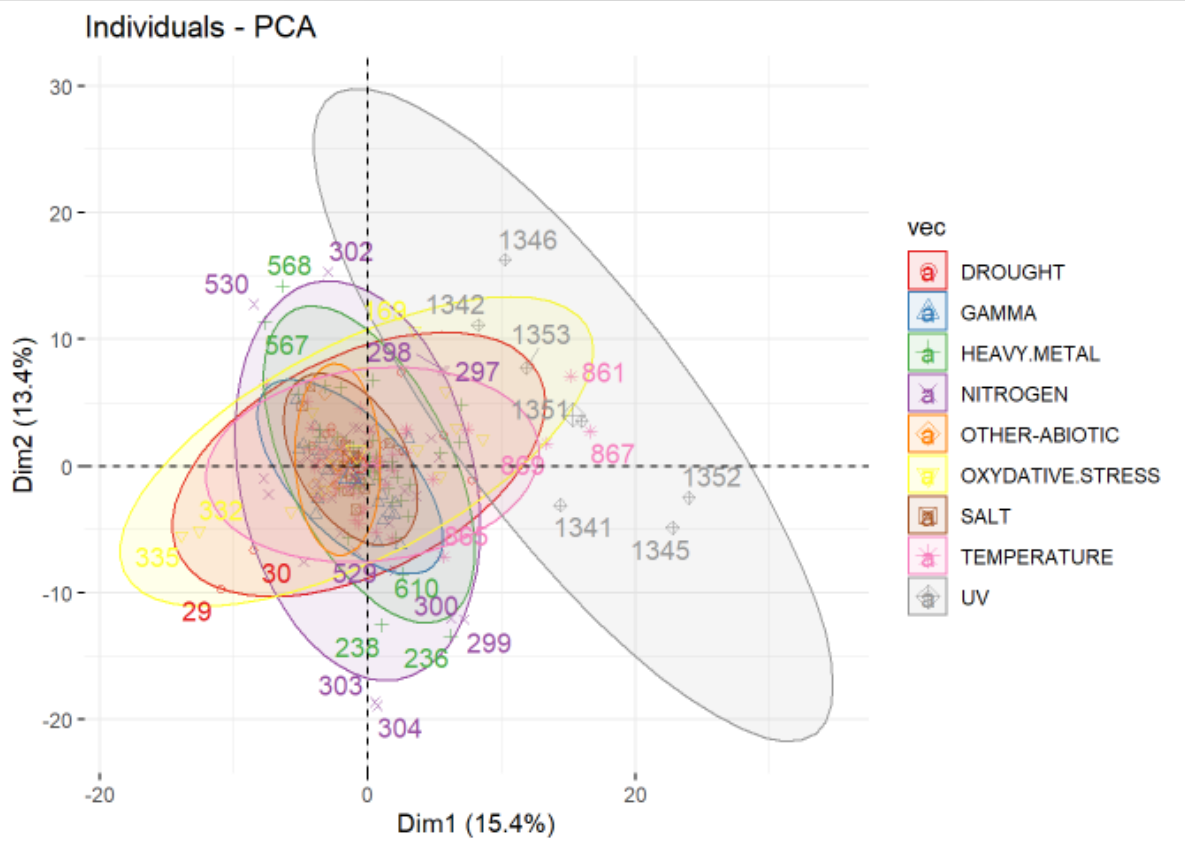
Célestin Valentin

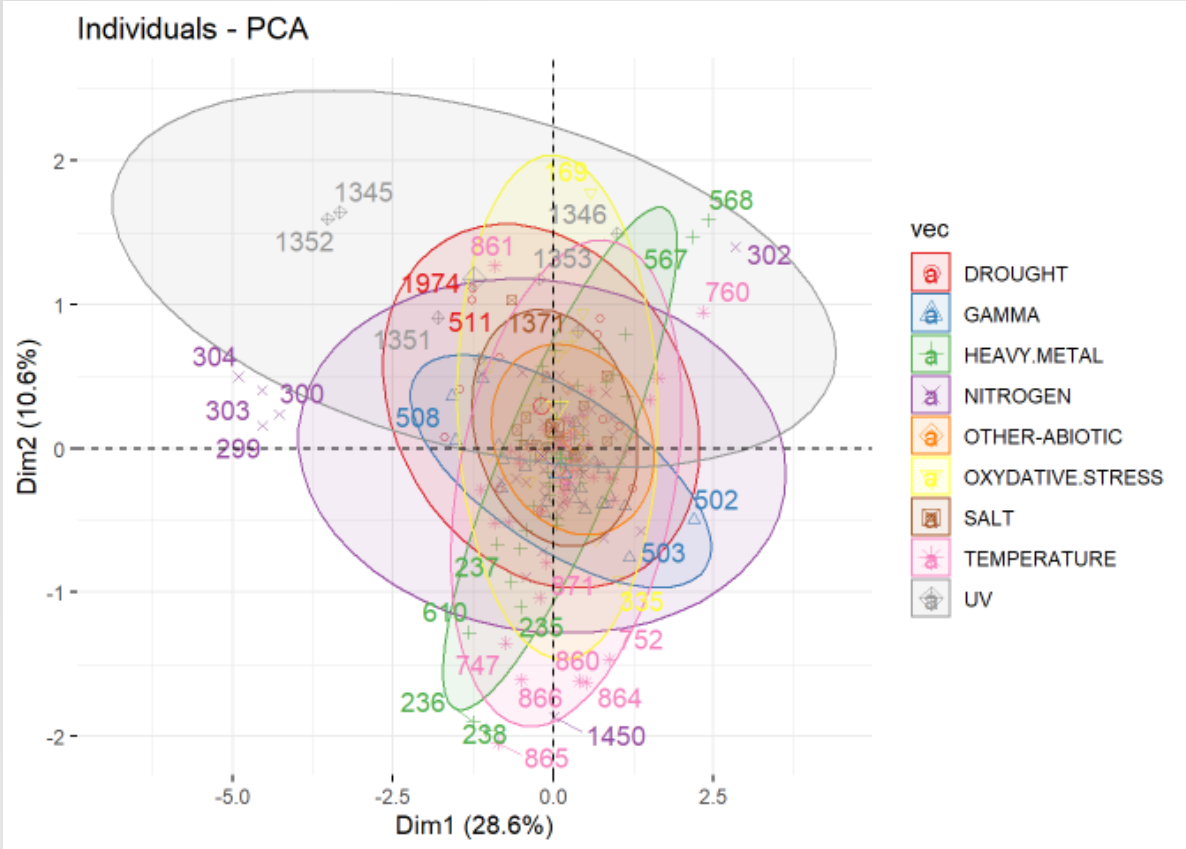
Arthur Carlin

Maryse Hulmel

Annexes

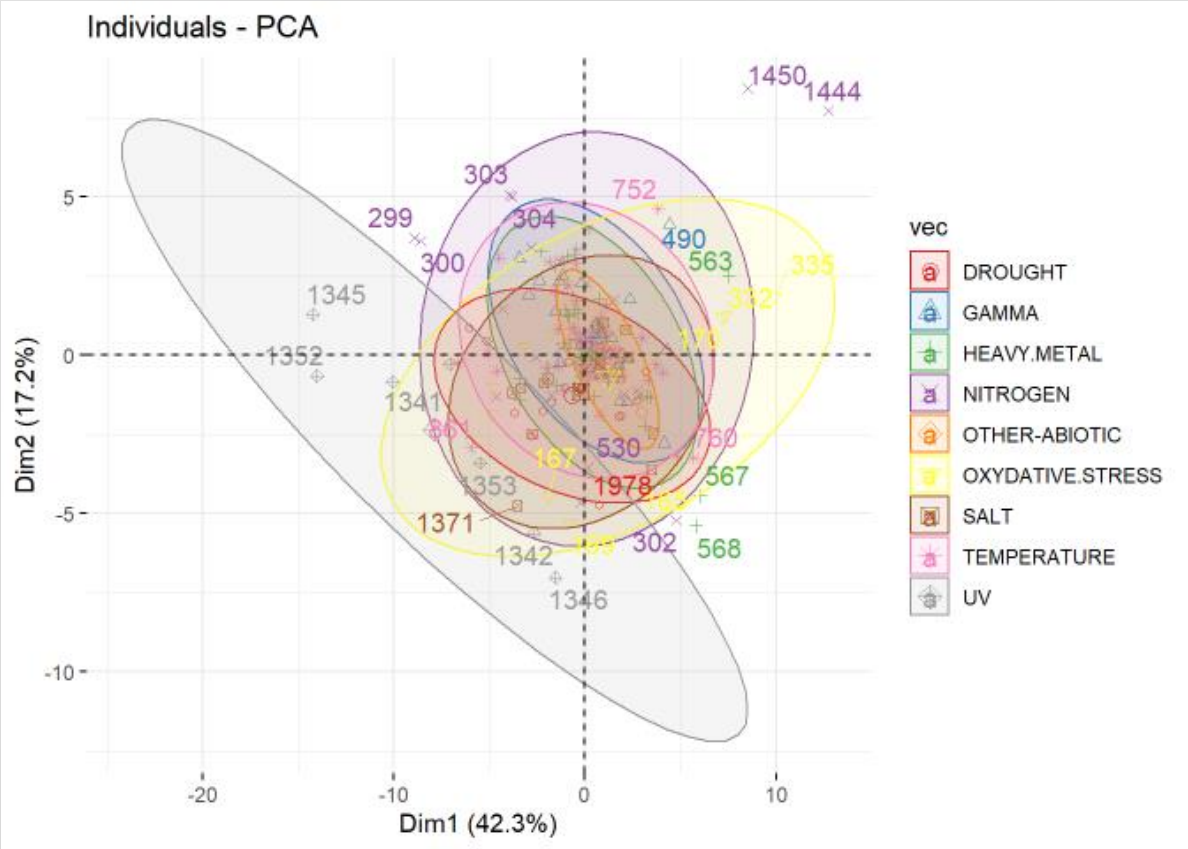
Autres ACP

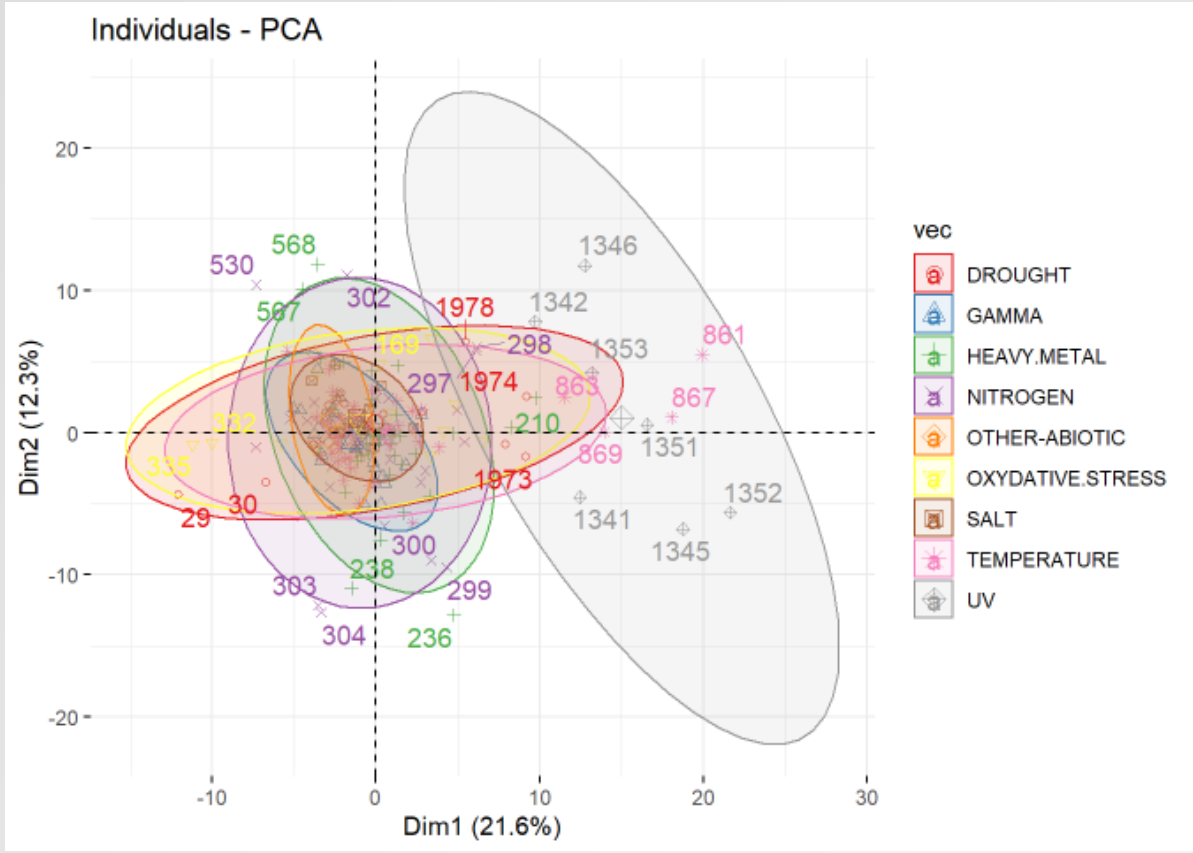




ACP complémentaires (abiotic/random)

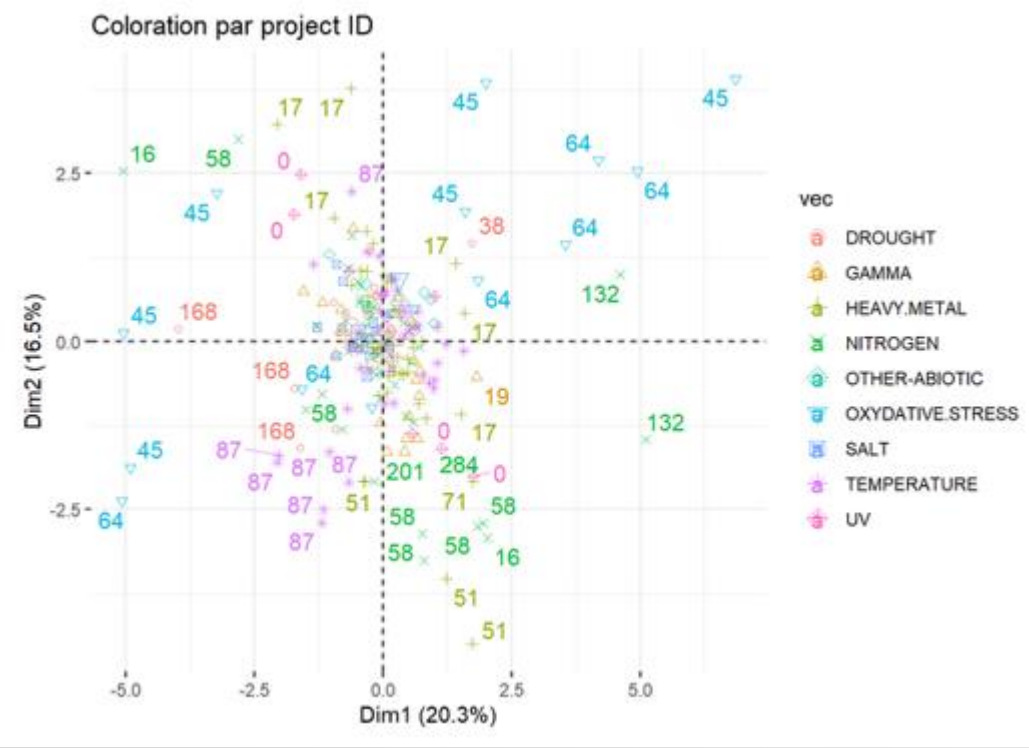
Autres ACP





ACP complémentaires (photosynthesis/abiotic)

Project ID



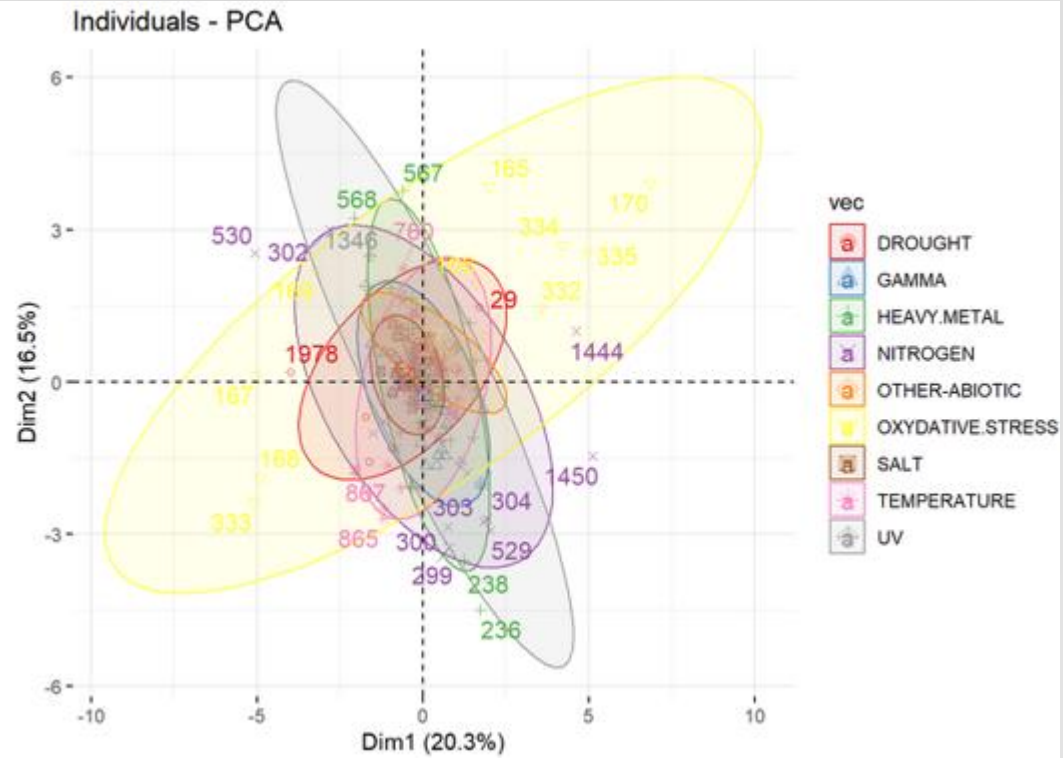
ACP par stress, label project ID

ACP par cluster, label project ID

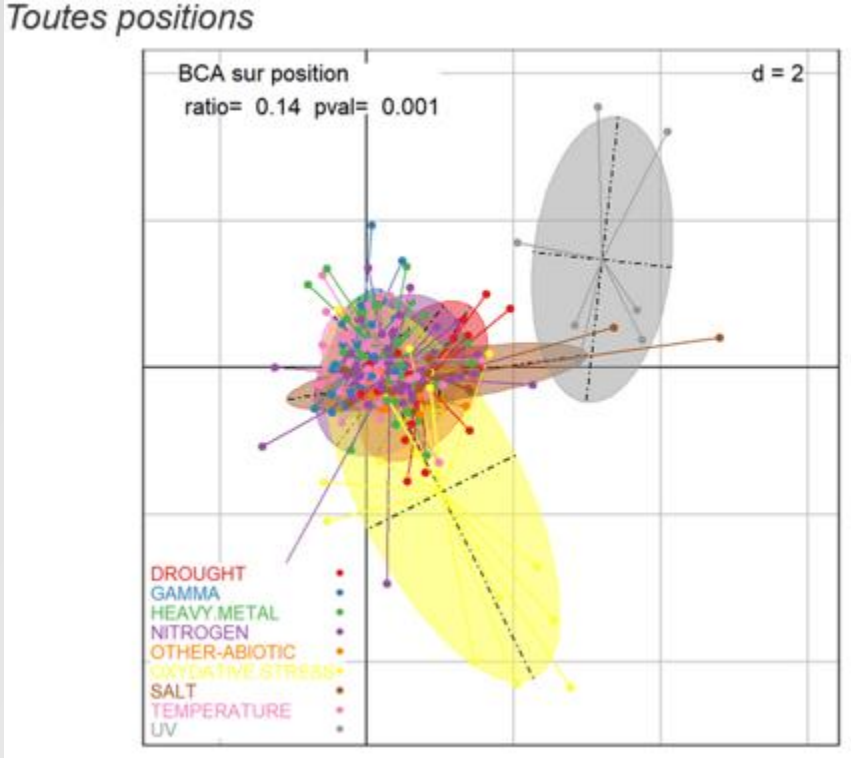
Boxplots

Boxplot Rythme circadien, regroupement par stress

ACP vs BCA

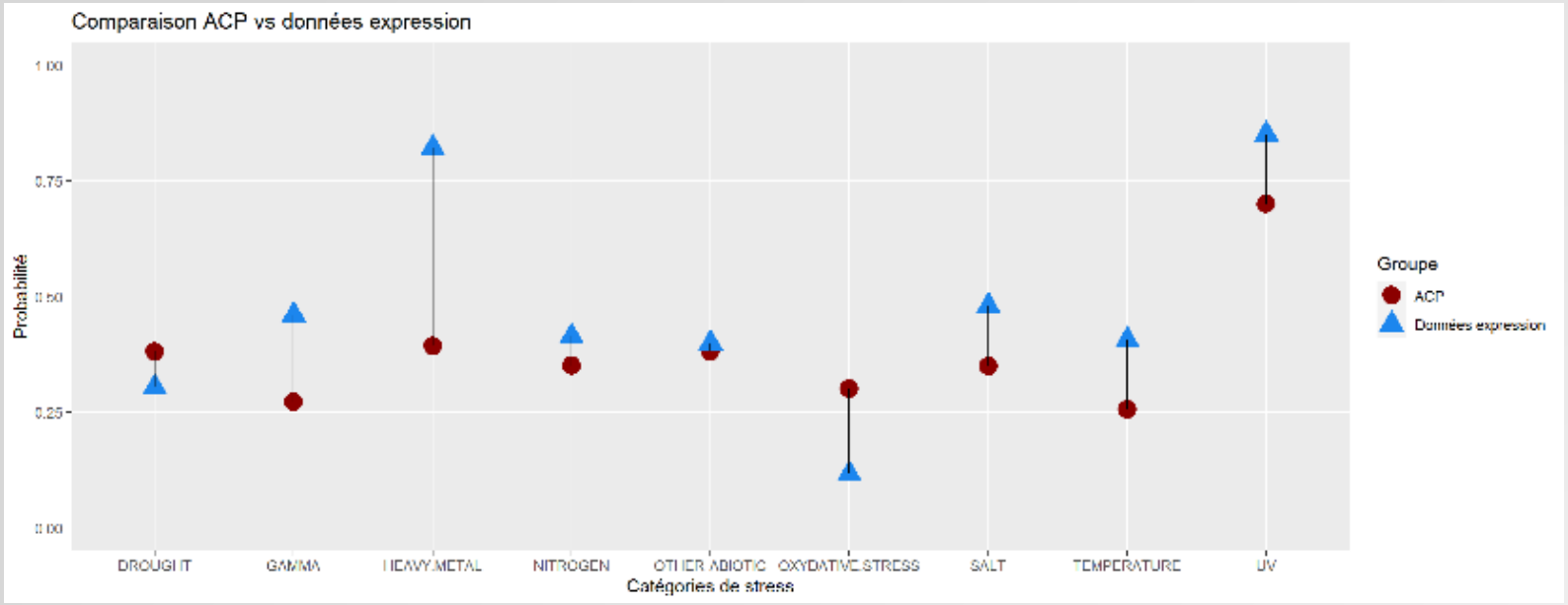


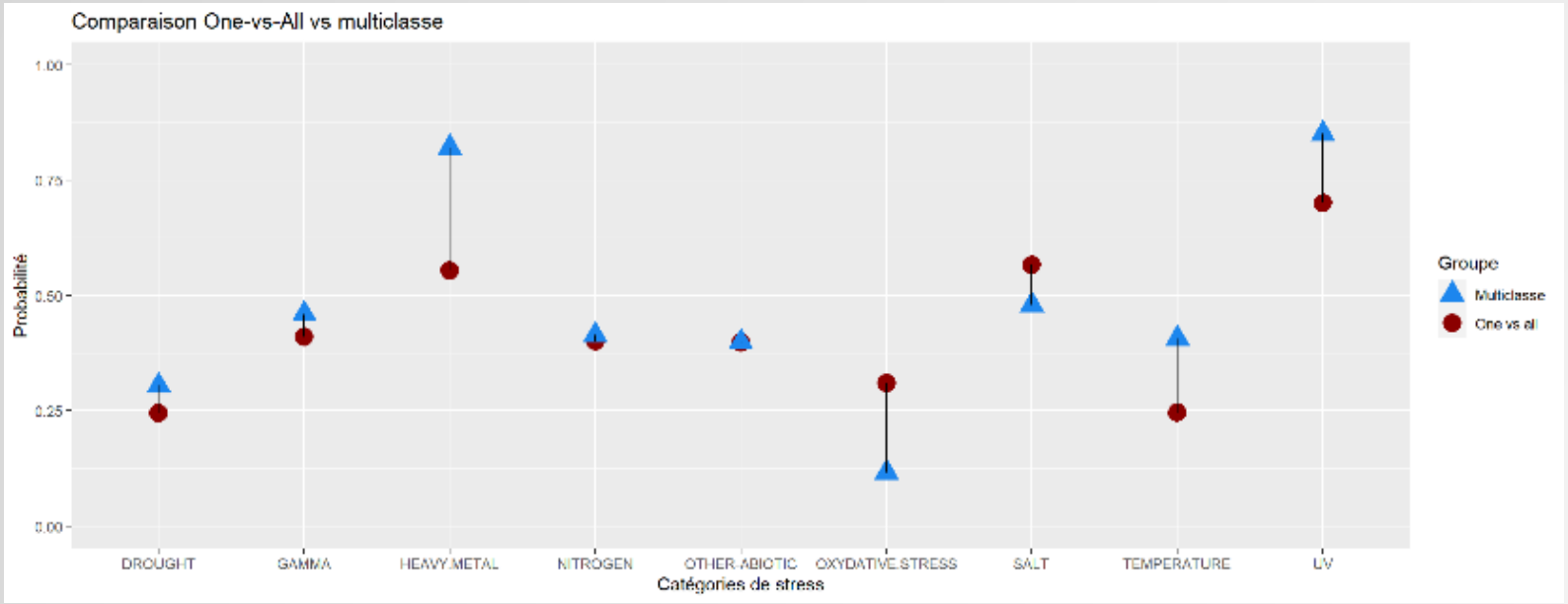
ACP par stress, label SWAP ID
(rythme circadien, ade4)



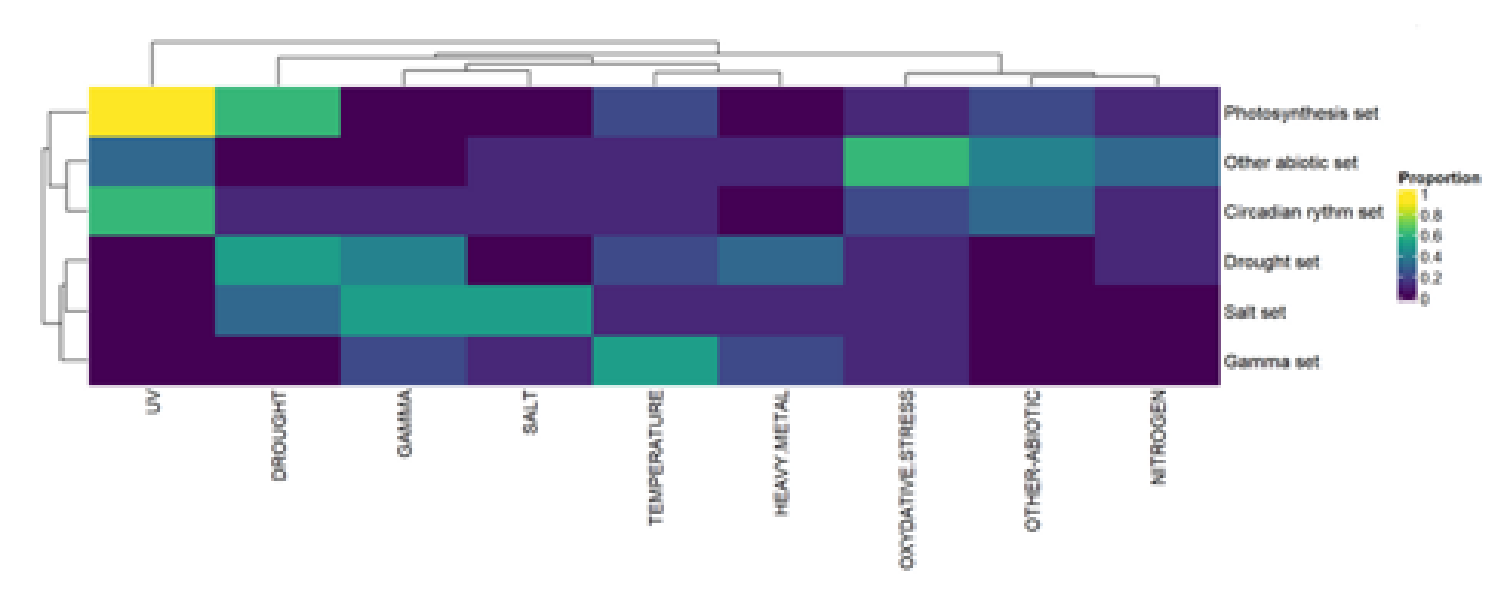
BCA par stress
(rythme circadien, ade4)

LDA : multiclasse/one-vs-all/ACP



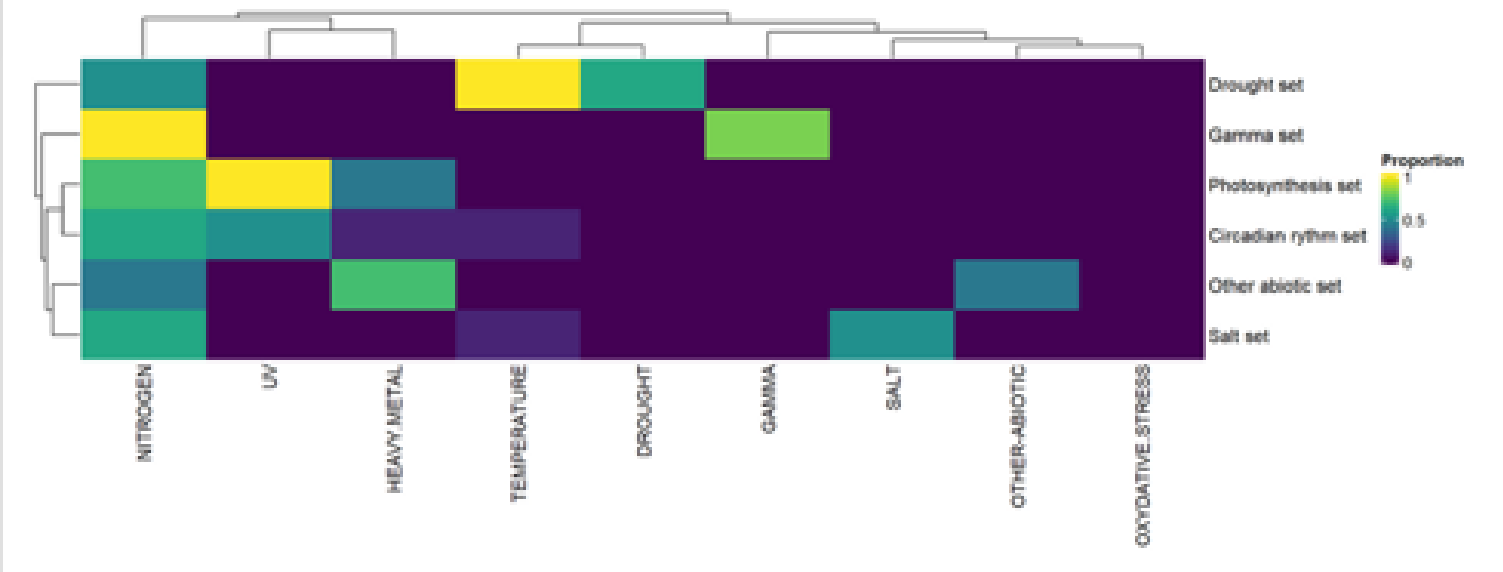


PLS-DA

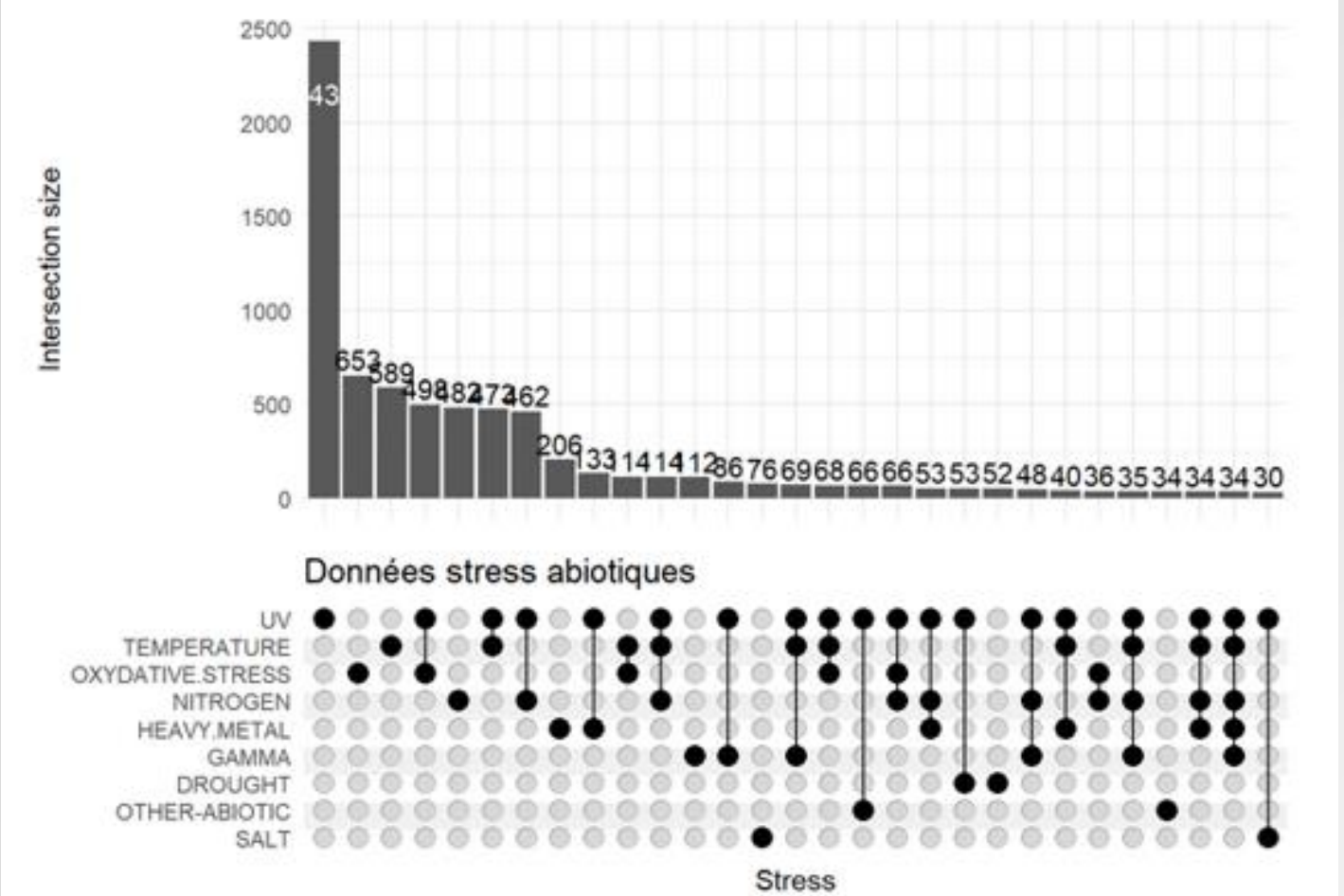


PLS-DA (centroïds distance)

PLS-DA (max distance)

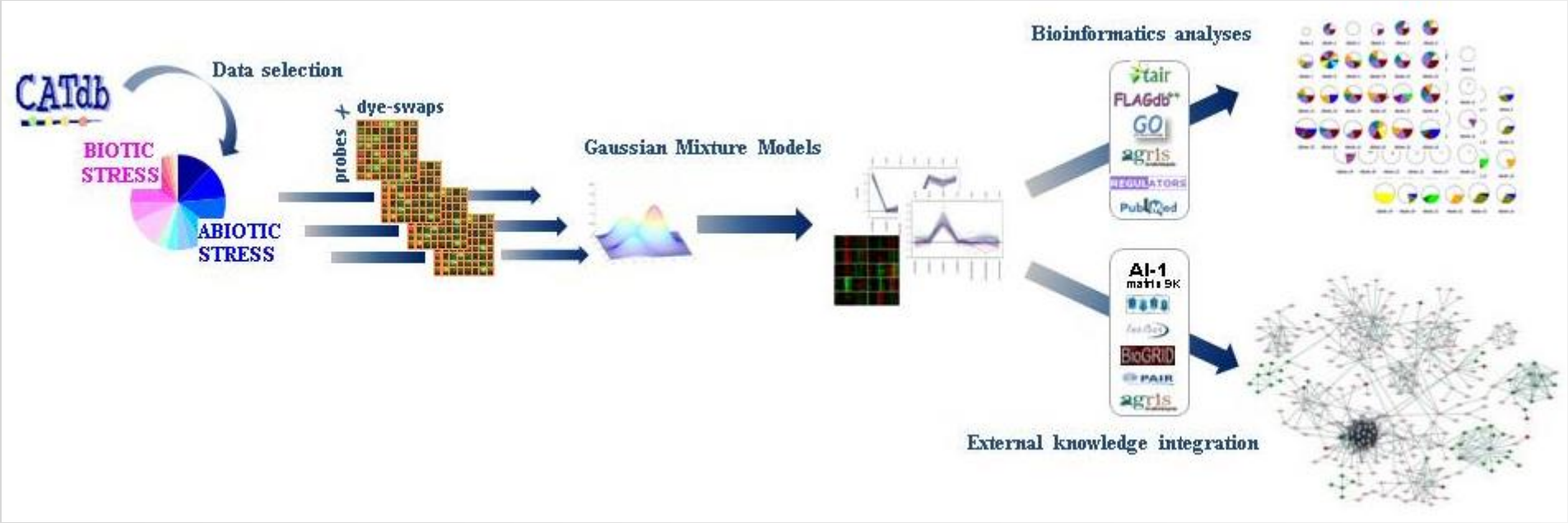


Autre upset plot



Upset plot catégories de stress abotiques

Analyse co-expression



http://tools.ips2.u-psud.fr/GEM2NET/

Capture écran Rmarkdown

- Gene set : Circadian rythm (57 gènes)
  - ACP (SWAP_ID)
  - BCA
- Gene set : Abiotic (661 gènes)
- Gene set : Biotic (420 gènes)
- Gene set : Endogenous stimulus (523 gènes)
- Gene set : External stimulus (552 gènes)
- Gene set : Flower (186 gènes)
- Gene set : Growth (222 gènes)
- Gene set : Light (292 gènes)
- Gene set : Photosynthesis (75 gènes)
- Gene set : Stress (1177 gènes)
- Gene set : Random (50 gènes)

ACP focus stress abiotic

Code

Solène Pety

25/03/2021

Code

Dans tout le jeu de données, on retrouve :
- 9 stress biotiques (**Biotrophic bacteria, Fungi, Nématodes, Oomycète, Rhodococcus, Stifenia, Necrotrophic bacteria, Virus et Other biotic**).
- 9 stress abiotiques (**Heavy metal, UV, Drought, Gamma, Nitrogen, Oxydative stress, Salt, Temperature et Other abiotic**).

Seuls les stress biotiques sont étudiés ici. Les fichiers utilisés ont deux colonnes informatives avec le stress appliqué sur l'échantillon en première colonne et le SWAP_ID pour pouvoir remonter aux informations de l'expérience précise.

Le fichier de chaque Gene Set est chargé à partir de la liste fournie dans le script. Les échantillons pour ce set sont séparés enuite pour les stress biotiques et abiotiques.

Code