Université de Rouen Normandie - UFR Sciences et Techniques
Master 1 mention Bioinformatique - Parcours BIMS
2020-2021

Rapport de stage

Knowledge transfer using multivariate gene expression projections onto a large-scale reference database

Présentée et soutenue par

Solène Pety

UMR Transfrontalière BioEcoAgro
INRAE
Pôle 1, équipe 1

Encadrant :
Dr. Andrea Rau, chargée de recherche

Université de Rouen Normandie - UFR Sciences et Techniques
Master 1 mention Bioinformatique - Parcours BIMS
2020-2021

Rapport de stage

Knowledge transfer using multivariate gene expression projections onto a large-scale reference database

Présentée et soutenue par

Solène Pety

UMR Transfrontalière BioEcoAgro
INRAE
Pôle 1, équipe 1

Encadrant :
Dr. Andrea Rau, chargée de recherche

REMERCIEMENTS

En premier lieu, je tiens à remercier avec une attention toute particulière Andrea Rau qui m'a encadrée au long de ces quatre mois. Je suis touchée par la confiance et l'autonomie qu'elle m'a laissées très rapidement. Ses conseils, nos discussions et les travaux qu'elle me demandait m'ont permis de progresser et d'acquérir de nouvelles compétences avec lesquelles je me sens aujourd'hui à l'aise. Je suis très reconnaissante d'avoir été choisie pour participer à ce projet et pour la bienveillance qu'elle a eue à mon égard. Cette expérience de stage aurait été différente sans elle.

Je remercie ensuite l'ensemble des équipes du site INRAE d'Estrées-Mons pour leur accueil chaleureux. Bien que le télétravail ne m'ait pas permis de rencontrer l'ensemble de l'équipe d'accueil, ni les agents du site j'ai pu découvrir les travaux de certains et les côtoyer lors de mes venues.

Je profite de cette occasion pour remercier l'ensemble de l'équipe pédagogique du master BioInformatique et Modélisation Statistiques de l'Université Rouen Normandie pour leur aide et leurs conseils dans la recherche de stage ainsi que dans l'encadrement de qualité qu'ils nous ont offert tout au long de notre formation. Je remercie spécialement Mme Caroline Bérard pour nous avoir transmis cette offre qui m'a tout de suite plu ; ses conseils lors de nos échanges au cours des derniers mois m'ont été précieux.

Finalement, je souhaite remercier l'ensemble des lecteurs de ce rapport, les personnes m'ayant aidé dans mes répétitions de soutenance. Ils m'ont apporté un éclairage qui m'a permis de retranscrire fidèlement l'expérience de ce stage et les résultats obtenus. Je remercie donc ma famille et mes amis pour le temps précieux qu'ils m'ont accordé et leur soutien.

Table des matières

Liste des figures

Liste des tableaux

TABLE DES ABREVIATIONS

UMRt : unité mixte de recherche transfrontalière

INRAE : Institut National de Recherche pour l'agriculture, l'alimentation et l'environnement

COOLKIT : Toolkit for translational study of chilling stress response in field-grown maize

RAM : Random Access Memory

Rmd : Rmarkdown

HTML : Hypertext Markup Language

GEM2Net : from Gene Expression Modeling to -omics Networks

CATMA : Complete Arabidopsis Transcriptome Micro Array

TAIR : The Arabidopsis Information Resource

ACP : Analyse en composantes principales

BCA : Between Class Analysis

PLIER : Pathway-Level Information ExtractoR

LV : Latent Variable

AUC : Area Under the ROC Curve

UV : UntraViolet

MAP : Maximum A Posteriori

LDA : Linear Discriminant Analysis

PLS-DA : Partial Least Squares Discriminant Analysis

GEO : Gene Expression Omnibus

1 Introduction

1.1 UMR Transfrontalière BioEcoAgro

L'Unité Mixte de Recherche Transfrontalière BioEcoAgro est une structure récente créée le 1er janvier 2020. Elle se place sous la tutelle de quatre grands acteurs distincts que sont INRAE, l'Université de Lille, l'Université de Picardie Jules Verne et l'Université de Liège. C'est une structure franco-belge rassemblant 300 chercheurs et techniciens sur les différents sites.

Le large panel de compétences que possède cette unité permet de découper ses objectifs de recherches en trois pôles principaux.

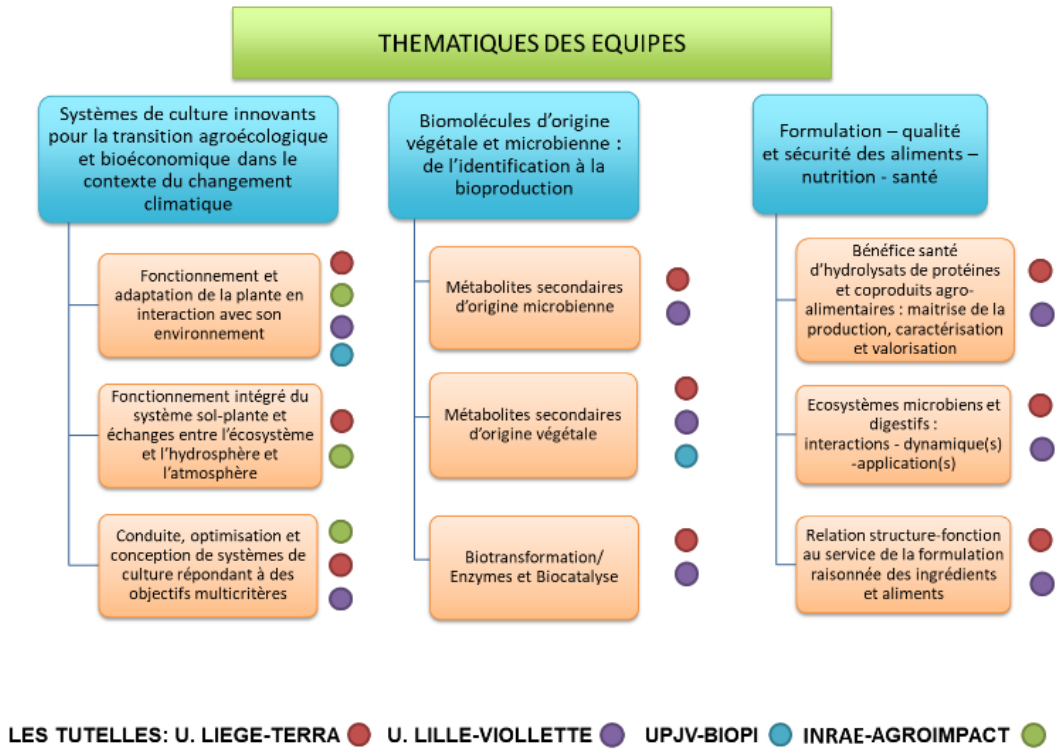


FIGURE 1 – **Organigramme thématique UMRt BioEcoAgro**, découpage en trois axes de recherches principaux.

Les plantes principalement étudiées sur le site d'Estrées-Mons sont le maïs, le pois et le miscanthus. Ce stage s'inscrit dans le projet **COOLKIT** encadré par l'équipe 1 du pôle 1 de l'UMRt. L'objectif principal de cette équipe est de mieux comprendre le comportement de la plante face à son environnement et les stress auxquels elle peut être confrontée, notamment dans le cadre du changement climatique.

Par exemple, parmi les projets en cours à INRAE, le site d'Estrées-Mons fait partie du projet Res0pest, conduit notamment par Sébastien DARRAS. L'objectif est de pouvoir réduire et améliorer l'utilisation des produits phytosanitaires. J'ai eu l'occasion de participer à une visite de présentation des parcelles expérimentales lors de ce stage. Quant à Catherine GIAUFFRET, avec qui nous avons pu collaborer lors de ce stage, elle participe au projet AMAIZING qui s'intéresse à la sélection et l'amélioration de la production de maïs issu de filières françaises.

1.2 Projet COOLKIT

Ce stage s'inscrit dans le cadre du projet COOLKIT, "Toolkit for translational study of chilling stress response in field-grown maize", porté par Catherine GIAUFFRET et Andrea RAU. Le plan de financement déposé pour ce projet n'a malheureusement pas été accepté. L'équipe continue néanmoins à le mener.

La culture du maïs en champ entraîne un certain nombre de contraintes et expose la plante à différents stress, lesquels sont amenés à s'accentuer dans les années à venir avec le réchauffement climatique. Parmi ceux-ci, on observe le stress hydrique en période de floraison, les attaques de champignons en automne. D'un point de vue agronomique et économique, il paraît donc intéressant de pouvoir semer le maïs plus tôt dans l'année mais ce choix expose la plante au froid du printemps. La culture de maïs résistant au froid a déjà réussi en parcelles contrôlées mais le passage en champ avec l'ajout de multiples stress supplémentaires n'a pas fonctionné. Il est donc essentiel de mieux comprendre les mécanismes moléculaires sous-jacents responsables de la réponse aux stress en champ pour le maïs.

Les plantes modèles comme *Arabidopsis thaliana* sont largement étudiées, bien caractérisées et maîtrisées, et elles possèdent des bases de données rassemblant un grand nombre d'informations les concernant, comme Genevestigator, GeneMania, MapMan, et ATTED-II. En revanche, l'étude des plantes "non-modèles" avec un fort intérêt agronomique, telles que le maïs, le blé, ou le riz, est rendue difficile par la complexité de leurs génomes, leur cycle de vie plus long, et leur grande taille. De plus, le passage en champ à une culture plus étendue est délicat car tous les paramètres ne peuvent pas être aussi bien contrôlés. La plante est exposée à une multitude de stress biotiques et abiotiques comme ceux évoqués précédemment. Il est donc important d'évaluer si les connaissances sur la réponse aux stress biotiques et abiotiques contenues dans les bases de données actuelles des plantes modèles peuvent être pertinentes et transférables vers d'autres espèces, et comment. Une approche prometteuse consisterait à projeter des données externes, issues d'autres expériences ou d'autres espèces, en se référant aux connaissances acquises sur Arabidopsis.

1.3 Objectifs de mon travail

Les objectifs du stage s'articulent de la manière suivante :

1. Le point de départ du stage était d'effectuer une analyse exploratoire des données transcriptomiques dans une base de données à grande échelle dédiée à la réponse aux stress de la plante modèle Arabidopsis. Il était particulièrement important de maîtriser et de comprendre le format, les caractéristiques des données et les catégories de stress attribuées aux échantillons dans la base de données. Cette exploration devait avoir lieu à l'échelle du transcriptome ainsi qu'à l'échelle de sous-groupes de gènes d'intérêt.

2. Ensuite, l'objectif était d'étudier plus formellement la structuration des données, et en particulier de tester la robustesse et l'homogénéité des groupes de stress via des analyses exploratoires multivariées et des classifications non-supervisées ou supervisées. Si les catégories de stress se révélaient homogènes, suffisamment spécifiques et bien séparées lors des étapes de classifications, il était alors envisageable de les prédire en projetant de nouvelles données. Une première étape de prédiction était de se servir des échantillons de la base de données initiale pour réaliser une validation interne de l'approche envisagée.

3. L'étape suivante consistait à utiliser les projections construites précédemment pour des données transcriptomiques d'Arabidopsis plus éloignées de la base de données de référence. Il s'agissait par exemple de choisir des échantillons appartenant à une autre base de données ou mesurés avec d'autres technologies (puces à ADN mono-couleur, RNA-seq). A partir du modèle construit, il était alors envisagé de prédire la classe de stress à laquelle l'expression de gènes de ces nouvelles données s'approche. Pour des données d'Arabidopsis, cette démarche aurait permis de valider notre modèle de manière plus approfondie. L'objectif était d'identifier les catégories de stress avec des profils transcriptomiques s'approchant le plus du cadre expérimental connu des données de validation.

2 Ressources utilisées

2.1 Environnement informatique

L'outil de travail utilisé dans le cadre du stage était mon ordinateur personnel ; la puissance de calcul était suffisante pour les fonctions et fichiers utilisés. L'ordinateur est un Aspire ES1-732 de chez Acer. Le processeur est un Intel(R) Pentium(R) CPU N4200 64 bits 1.10 Ghz et possède 8,00 Go de RAM. Il a été suffisant pendant la durée du stage bien que ses limites de calculs aient été atteintes.

Une grande partie du stage s'est réalisée à distance. INRAE ayant une licence Zoom, les points se faisaient en visioconférence sur cette plateforme. A cause notamment de problèmes techniques il est arrivé que les échanges soient téléphoniques.

Concernant les espaces partagés, la rédaction de ce rapport a été réalisé au format LaTeX sur OverLeaf en collaboration avec Andrea, ce qui a permis de pouvoir réfléchir ensemble au plan et au contenu des sous-parties. De plus, ce sont les rapports Rmarkdown qui nous servaient de support de discussions. La taille de ces fichiers étant assez grande, entre 1 Mo et 6 Mo, il était impossible de les partager par mail. Un DropBox a donc été créé pour le stage afin de déposer ces rapports. Il a permis également de pouvoir créer des dossiers contenant tous les fichiers nécessaires à la réalisation d'un exemple reproductible lorsqu'un problème dans le code n'arrivait pas à être résolu.

Enfin, la veille bibliographique de ce stage a été réalisée grâce à Zotero et importée sur Overleaf en tant que fichier .bib.

2.2 Pratique professionnelle

Le télétravail a été mis en place, suite à l'annonce du troisième confinement, dès la deuxième semaine de stage. Les points par visioconférence étaient réalisés grâce à Zoom. INRAE a depuis peu une licence Zoom permettant à chaque agent de bénéficier d'une salle privée. La fréquence des points dépendait de l'avancée dans les travaux et des contraintes de chacune. Un point quotidien a été fait dans les premières semaines puis avec l'acquisition d'autonomie et de confiance ces échanges ont pu s'espacer, entre une à trois fois par semaine. Le contact était toujours maintenu par mail, en particulier pour des échanges de liens, de problèmes rencontrés avec le code ou encore lorsqu'un rapport Rmarkdown était bien avancé.

Comme évoquée précédemment, la manière la plus simple, agréable et lisible de présenter les résultats au fur et à mesure de l'avancée du stage a été de construire des rapports Rmarkdown pour chaque sous-partie. Certains sont plus complets que d'autres en fonction de la problématique à laquelle ils devaient répondre. Les mêmes analyses devant se faire sur plusieurs ensembles de gènes, des boucles for ont été intégrées limitant la redondance du code et la multiplication de portions identiques. Les titres des sous-parties des rapports ont été générés automatiquement grâce à ces boucles. La sortie HTML choisie présente l'avantage de pouvoir afficher le code et une table des matières ce qui facilite la navigation et l'échange autour des figures produites. Les commentaires ajoutés ont permis également de comprendre ensemble les points faibles du code et comment il était possible de les améliorer ou de simplement vérifier que ce code correspondait bien à l'analyse attendue.

Le lundi matin, toutes les deux semaines, une réunion d'équipe est organisée rassemblant les agents d'Estrées-Mons et de Laon. Elle est l'occasion de rencontrer des personnes d'autres sites, de comprendre le fonctionnement de l'unité et ses problématiques. Il s'agit d'un temps d'échanges sur les travaux en cours et les avancées de chacun sur les différents projets en cours. A cette occasion, Andrea a pu me présenter aux agents que je n'avais pas encore rencontrés sur le site en présentiel et a pu évoquer le travail prévu pour le stage. Les résultats obtenus seront présentés lors d'une de ces réunions, constituant une répétition à la soutenance. En plus de ces contacts, d'autres échanges ont pu avoir lieu avec Catherine Giauffret notamment, biologiste/généticienne quantitative experte en maïs. Ils nous ont permis d'amener des interprétations biologiques lorsque cela était nécessaire et d'envisager d'autres pistes de réflexion.

Ce stage a généré un grand nombre de fichiers de données, de scripts R, Rmd et de rapports HTML. Afin que ce travail puisse se poursuivre par la suite, un GitHub a été créé pour pouvoir héberger l'ensemble des documents et rendre leur utilisation accessible et simple.

2.3 Outils informatiques et statistiques

Dans le cadre de ce stage, les différentes analyses ont été réalisées avec le langage R dans l'environnement RStudio. L'installation de nouveaux packages conçus pour les dernières versions de R m'ont permis de mettre à jour ma configuration et de travailler avec la version R 4.0.5 (2021-03-31).

```
R version 4.0.5 (2021-03-31)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19041)

Matrix products: default

locale:
[1] LC_COLLATE=French_France.1252  LC_CTYPE=French_France.1252    LC_MONETARY=French_France.1252 LC_NUMERIC=C
[5] LC_TIME=French_France.1252

attached base packages:
 [1] grid      parallel  stats4    stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] viridis_0.6.1           viridisLite_0.4.0       cowplot_1.1.1           gridExtra_2.3           ComplexHeatmap_2.7.10.9002
 [6] KEGGAPI_0.1.7.4         ParaMisc_0.0.1          Biostrings_2.58.0       XVector_0.30.0          AnnotationDbi_1.52.0
[11] IRanges_2.24.1          S4Vectors_0.28.1        Biobase_2.50.0          BiocGenerics_0.36.1     PLIER_0.99.0
[16] qvalue_2.22.0           rsvd_1.0.5              knitr_1.33              glmnet_4.1-1            Matrix_1.3-3
[21] pheatmap_1.0.12         gplots_3.1.1            RColorBrewer_1.1-2      kableExtra_1.3.4        forcats_0.5.1
[26] stringr_1.4.0           purrr_0.3.4             readr_1.4.0             tidyr_1.1.3             tibble_3.1.0
[31] ggplot2_3.3.3           tidyverse_1.3.1         dplyr_1.0.6

loaded via a namespace (and not attached):
 [1] colorspace_2.0-1    rjson_0.2.20        ellipsis_0.3.2      circlize_0.4.12     GlobalOptions_0.1.2 fs_1.5.0            clue_0.3-59
 [8] rstudioapi_0.13     farver_2.1.0        bit64_4.0.5         fansi_0.4.2         lubridate_1.7.10    xml2_1.3.2          codetools_0.2-18
[15] splines_4.0.5       doParallel_1.0.16   cachem_1.0.5        jsonlite_1.7.2      Cairo_1.5-12.2      broom_0.7.6         cluster_2.1.2
[22] dbplyr_2.1.1        png_0.1-7           compiler_4.0.5      httr_1.4.2          backports_1.2.1     assertthat_0.2.1    fastmap_1.1.0
[29] cli_2.5.0           htmltools_0.5.1.1   tools_4.0.5         gtable_0.3.0        glue_1.4.2          reshape2_1.4.4      Rcpp_1.0.6
[36] cellranger_1.1.0    vctrs_0.3.8         svglite_2.0.0       iterators_1.0.13    xfun_0.22           rvest_1.0.0         lifecycle_1.0.0
[43] gtools_3.8.2        XML_3.99-0.6        zlibbioc_1.36.0     scales_1.1.1        hms_1.1.0           yaml_2.2.1          memoise_2.0.0
[50] pander_0.6.3        stringi_1.6.2       RSQLite_2.2.7       foreach_1.5.1       caTools_1.18.2      shape_1.4.6         rlang_0.4.10
[57] pkgconfig_2.0.3     systemfonts_1.0.2   bitops_1.0-7        matrixStats_0.58.0  evaluate_0.14       lattice_0.20-44     labeling_0.4.2
[64] bit_4.0.4           tidyselect_1.1.1    plyr_1.8.6          magrittr_2.0.1      R6_2.5.0            generics_0.1.0      DBI_1.1.1
[71] pillar_1.6.1        haven_2.4.1         withr_2.4.2         survival_3.2-11     RCurl_1.98-1.3      modelr_0.1.8        crayon_1.4.1
[78] KernSmooth_2.23-20  utf8_1.2.1          rmarkdown_2.8       GetoptLong_1.0.5    readxl_1.3.1        blob_1.2.1          reprex_2.0.0
[85] digest_0.6.27       webshot_0.5.2       munsell_0.5.0
```

FIGURE 2 – **SessionInfo() RStudio**, rassemblant l'ensemble des informations liées à la version de R utilisée et des packages appelés.

Certains des packages utilisés provenaient de sources autre que CRAN comme Bioconductor ou GitHub. Un temps d'adaptation à la syntaxe propre à chaque source était nécessaire. J'ai rencontré également quelques problèmes lors de l'installation de certains packages, en particulier ceux de Bioconductor. Très souvent, la source des problèmes était le chemin vers lequel les packages étaient installés, à cause d'un espace dans le nom ou parce que le répertoire correspondait à un serveur en ligne. L'exécution de RStudio avec les droits d'administrateurs a résolu certains des problèmes rencontrés. Ils permettaient de modifier plus facilement les destinations des téléchargements. L'ensemble des erreurs apparues a pu être résolu grâce aux différentes recherches et aux forums StackOverflow et RStudio Community.

2.4 Données

Les données principales utilisées sont celles provenant de GEM2Net [1]. GEM2Net est une partie de la base de données transcriptomiques CATdb [2]. Elle a pour but de rassembler un ensemble de projets sur Arabidopsis avec une problématique liée à un stress biotique ou abiotique. Elle contient 18 catégories de stress, 9 biotiques et 9 abiotiques listées dans la Figure 3. Les effectifs détaillés pour chaque stress sont présentés dans l'Annexe 3. D'autres bases de données de stress pour Arabidopsis existent comme Stress-Responsive Transcription Factor Database ou Plant Stress Gene Database. Cependant, elles synthétisent les informations trouvées dans la littérature avec des listes de gènes impliqués ce qui ne permet pas d'analyses globales. Dans GEM2Net, les puces à ADN deux couleurs Complete Arabidopsis Transcriptome MicroArray (CATMA) modélisées pour Arabidopsis permettent de quantifier l'expression des genes. Aujourd'hui, il existe 7 versions de cette puce ;la dernière possède des sondes pour 35 656 gènes. L'extraction des données pour le stage a été faite le 11 mars 2021 par Cécile Guichard dans l'équipe Saclay Plants Omic (SPOmics) sur la plateforme transcriptomique POPS. Le tableau est donc composé de 17 341 gènes (en colonnes) et 387 échantillons (en lignes) correspondant aux expériences. Le nom des lignes est un identifiant unique (le SWAP ID), et le nom des colonnes correspond aux identifiants uniques des gènes au format The Arabidopsis Information Resource (TAIR), e.g. AT1G01010. Un des intérêts des valeurs de cette base de données est qu'elles ont été pré-processées et homogénéisées. Ces étapes préalables permettent de comparer les valeurs entre elles et de simplifier les analyses.

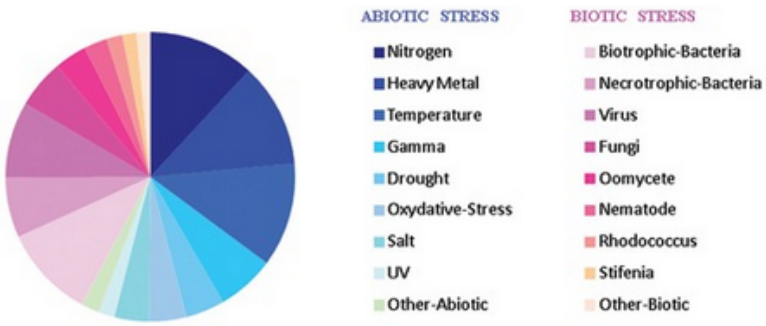


FIGURE 3 – **Répartition des catégories de stress dans GEM2Net**, extrait de Zaag et al. (2015) [1], Figure 1.

Dans un deuxième tableau relié à celui décrit ci-dessus par les SWAP ID se trouve un ensemble de métadonnées permettant de mieux comprendre et de remonter aux origines des données. On y retrouve notamment :

1. le project ID (qui correspond à celui renseigné dans CATdb),
2. la catégorie de stress à laquelle l'expérience a été attribuée,
3. le détail du plan d'hybridisation de l'échantillon (e.g., "NaCl_1 / Control_1" pour indiquer replicat 1 avec un stress salin hybridé avec replicat 1 d'un contrôle).

Pour réduire la taille du tableau, rendre plus facile les calculs et concentrer la recherche sur des ensembles de gènes liés par une même fonction ou voie métabolique, nous avons utilisé trois sets de gènes distincts :

1. Des sous-ensembles de termes d'ontologie de gènes (gene ontology ; GO), appélés les GO SLIM. A partir d'un fichier récupéré depuis TAIR [3], 10 termes GO SLIM en lien avec le stress ont été sélectionnés :
   (a) Rythme circadien (57 gènes)
   (b) Floraison (186 gènes)
   (c) Croissance (222 gènes)
   (d) Photosynthèse (75 gènes)
   (e) Stimulus abiotique (661 gènes)
   (f) Stimulus biotique (420 gènes)
   (g) Stimulus endogène (523 gènes)
   (h) Stimulus externe (552 gènes)
   (i) Lumière (292 gènes)
   (j) Stress (1177 gènes)

A partir de ce fichier, les listes de gènes sont récupérées pour chaque terme et des sous-fichiers pour chaque terme GO SLIM ont été créés.

2. Des bases de données de pathways AraCyc [4] et KEGG [5]. AraCyc (379 pathways) représente une base de données de pathways enzymatiques et métaboliques très organisée, complète, et validée manuellement chez Arabidopsis, hébérgée sur le Plant Metabolic Network (PMN). Bien que les données AraCyc soient très complètes, certains formats de noms posaient problème pour la délimitation des colonnes. Un travail de reformatage et d'expressions régulières a été nécessaire pour récupérer les noms des pathways et les gènes associés. Le Kyoto Encyclopedia of Genes and Genomes (KEGG) représente une grande collection de bases de données sur les fonctions de haut niveau du système biologique. Pour KEGG PATHWAY (136 pathways), le package KEGGAPI a été utilisé. Il permet de faire des requêtes et de récupérer la liste des gènes pour chaque pathway notamment.

3. Des signatures spécifiques aux stress, identifées par des analyses différentielles (voir description dans Methodes).

Enfin, pour exploiter des nouvelles de données transcriptomiques d'intérêt trouvées au NCBI, le package GEOQuery [6] a été employé. Les données brutes des projets sont disponibles sous forme d'une matrice d'expression avec les différents échantillons du projet en colonnes et les valeurs de puces en lignes. Il est possible de revenir aux gènes à partir de cette matrice grâce à la matrice de design propre à la version de la puce utilisée, disponible sur NCBI également.

3 Méthodes

Dans cette section, je présente de manière synthétique l'ensemble de méthodes statistiques utilisées pour réaliser les objectifs du stage, notamment : (1) l'identification de signatures de stress par une analyse différentielle ; (2) la classification non-supervisée des échantillons de la base de données GEM2Net ; (3) la classification supervisée des catégories de stress pour les échantillons dans GEM2Net ; et (4) la déconvolution non-supervisée de structure dans GEM2Net avec prise en compte de connaissances biologiques a priori. Dans l'ensemble des analyses sauf indication contraire, la différence a été faite entre les stress biotiques et abiotiques dans GEM2Net afin de limiter le nombre de données à manipuler et de visualiser et interpréter plus facilement les résultats. A partir du tableau de données de départ, deux sous-tableaux ont donc été créés pour les stress biotiques (163 échantillons) d'une part et abiotiques (224 échantillons) d'autre part, avec les 17 341 gènes en ligne et les échantillons en colonnes.

3.1 Identification de signatures de stress

Afin d'obtenir des listes de gènes caractéristiques et spécifiques de réponse pour chaque catégorie de stress (des "signatures de stress"), une analyse différentielle a été pensée. Cette analyse cherche à identifier des gènes avec une différence significative d'expression entre une catégorie de stress donnée comparée à la moyenne des autres. Pour cette réalisation, nous avons utilisé le package limma de Bioconductor [7]. Celui-ci permet l'ajustement d'un modèle linéaire pour chaque gène, en implémentant une méthode bayésienne empirique pour stabiliser l'estimation des variances et une correction pour tests multiples. Les fonctions de limma utilisées pour cette analyse sont `lmFit()`, `contrasts.fit()`, `eBayes()`, `topTable()` et `decideTests()`.

En entrée, le tableau de données est fourni comme décrit précédemment ainsi qu'une matrice de design. Ici, la matrice de design a en ligne les échantillons et en colonne les stress avec comme valeurs 1 et 0. Le nombre de 1 dans chaque colonne correspond donc à l'effectif du stress dans les données. `lmfit()` et `contrasts.fit()` permettent de calculer le modèle linéaire souhaité en accordant la même importance à chaque stress. `eBayes()` s'occupe ensuite d'appliquer la méthode bayésienne empirique et d'effectuer des tests statistiques sur ce modèle pour nous permettre ensuite d'ajuster avec `topTable()` les valeurs de p-value calculées pour contrôler le taux de fausse découvertes selon la méthode de Benjamini et Hochberg. Enfin, `decideTests()` identifie les gènes exprimés différentiellement en se basant sur les éléments statistiques calculés et un seuil de p-values ajustées, ici fixé à 0.05. La fonction renvoie une matrice binaire avec les stress en colonnes et les gènes en ligne. Une valeur de 1 ou -1 correspond donc à un gène différentiellement sur- ou sous-exprimé pour un stress donné comparé à la moyenne des autres, respectivement.

3.2 Classification non-supervisée des échantillons

Une fois les données brutes en main et connaissant la structure des données, la question se pose d'abord de la structuration des données GEM2Net par rapport aux catégories de stress. Cette structuration peut être explorée à l'échelle du transcriptome entier, ou pour des sous-ensembles des données définis par des listes de gènes (GO SLIM, signatures de stress). Nous nous sommes tournées vers une classification non-supervisée afin d'observer la séparation pré-existante éventuelle des stress. Plus celle-ci serait marquée, plus cela signifierait que les catégories de stress sont bien distinctes, suggérant qu'une prédiction de catégorie de stress pour d'autres données serait facilement réalisable. Par conséquent, une analyse en composantes principales (ACP) est un choix naturel qui consiste à calculer pour un grand jeu de données des composantes (i.e., des combinaisons linéaires de variables) décorrélées les unes des autres. Cela permet une réduction de dimension tout en conservant le plus d'informations possibles.

Le calcul de nouvelles coordonnées avec un nombre de dimensions limitées, dans un nouvel espace, favorise l'exploration a posteriori de la structuration des échantillons en corrélation avec les catégories de stress connues (non prises en compte dans l'ACP). Le choix du nombre de composantes conservées dépend ainsi du seuil de variance expliquée que nous souhaitons atteindre. Dans notre cas, le nombre de dimensions pour atteindre 80% était en général très important. De plus, dès la troisième dimension, cette variance expliquée descendait en-dessous de 10%. Pour nous focaliser sur l'exploration graphique de la structuration principale des données, nous avons donc fait le choix de n'exploiter que les 2 premières dimensions.

Pour l'ensemble des ACP réalisées, nous avons utilisé les packages FactoMineR [8] et FactoExtra qui permettent une manipulation des données, une visualisation et une personnalisation idéale. FactoExtra utilise des fonctions de packages déjà existants que sont FactoMineR, ggplot2, cluster et dendextend.

3.3 Classification supervisée des catégories de stress

Une fois les premières analyses exploratoires et non-supervisées réalisées, nous avons porté notre attention sur des analyses de classification supervisée afin d'établir la pertinence d'une approche prédictive pour les catégories de stress incluses dans GEM2Net.

3.3.1 Between class analysis

La Between Class Analysis (BCA), appelée également l'analyse de redondance (RDA), est un cas particulier d'ACP par rapport à une variable instrumentale qui permet d'étudier les relations entre deux tableaux, X et Y. Pour la BCA, un seul facteur est considéré comme variable réponse Y ; dans notre cas c'est la variable stress. Dans un premier temps, les composantes de la BCA sont construites à partir de X de manière à ce qu'elles soient autant que possible corrélées avec Y. Ensuite, les composantes de Y sont construites de manière à maximiser leurs corrélations avec les composantes extraites de X. Contrairement à l'ACP, la BCA est donc supervisée et a comme objectif d'augmenter la séparation qui pourrait exister entre ces catégories.

La fonction utilisée provient du package ade4 [9], qui contient un grand nombre de fonctions statistiques y compris des ACP et BCA.

3.3.2 Analyse discriminante linéaire

Après avoir pu observer le comportement des données et mieux cerner la séparabilité des catégories de stress avec la BCA, nous avons voulu tester des prédictions plus formelles de ces catégories. L'objectif était de voir s'il est possible de prédire la bonne catégorie de stress pour un ensemble de données de validation à partir d'un modèle linéaire construit sur un ensemble de données d'apprentissage issues de GEM2Net. Pour cette classification supervisée, nous avons choisi comme première stratégie une analyse discriminante linéaire (LDA). Comme l'ACP, la LDA est une technique de transformation linéaire, mais qui cherche à identifier un sous-espace de variables qui maximise la séparabilité entre les classes fournies en entrée tout en minimisant la variabilité intra-classe. Il est important à noter que la LDA ne peut être appliquée que dans des cas où le nombre de variables p est inférieur au nombre d'échantillons n. Pour les stress abiotiques (n=224), seuls les ensembles de gènes de moins de p=200 gènes ont donc été conservés pour cette analyse. Pour de plus grands ensembles, afin de réduire au préalable la dimensionnalité, nous avons appliqué la LDA sur les composantes principales et non pas sur les données brutes directement.

Pour la LDA, nous avons utilisé la fonction `lda()` du package ade4 avec la variable de stress indiquant les catégories à prédire, et les valeurs transcriptomiques servant de variables prédictives.

3.3.3 Régression des moindres carrés partiels

Dans la même idée de classification supervisée mais pour tester une autre méthode de calcul, nous nous sommes appuyées sur une régression des moindres carrés partiels (partial least squares discriminant analysis ; PLS-DA) implémentée dans le package mixOmics [10]. Le raisonnement est similaire à celui de la LDA avec une variable discrète contenant les classes à prédire par rapport à une combinaison linéaire de variables prédictives. La principale différence de la PLS-DA comparée à la LDA est qu'elle cherche plutôt à maximiser la covariance entre les variables indépendantes et la variable de réponse, recodée en interne comme variables indicatrices. A noter également que le package mixOmics propose une estimation parcimonieuse ("sparse") de la PLS-DA afin d'effectuer une sélection de variables en même temps que la classification.

3.3.4 Stratégies de classification multiclasse versus "one-versus-all"

De manière générale, lorsque le nombre de 2 classes à prédire est supérieur à 2 (ce qui est le cas pour les catégories de stress), il est possible d'appliquer différentes stratégies de classification, notamment les stratégies "multiclasse" ou "one-versus-all" (un contre tous). La stratégie multiclasse construit un modèle prédictif qui prend en compte l'ensemble des stress en même temps. La stratégie "one-versus-all" prend un stress à la fois et effectue une classification binaire de ce stress par rapport aux autres. On combine les résultats ainsi de toutes les classifications et on ne conserve que le score maximal pour prédire la classe finale. Pour ces deux stratégies, nous avons pu comparer des écarts entre les valeurs calculées, en s'appuyant sur les probabilités prédites d'appartenance à chaque catégorie de stress, pour chaque échantillon de validation.

3.3.5 Évaluation de la qualité de prédiction sur données de validation

Pour toute méthode de prédiction ou classification non-supervisée, il est essentiel de l'évaluer sur un échantillon de données indépendantes, dites "données de validation". Nous avons donc choisi de séparer les échantillons de GEM2Net en une partie apprentissage pour estimer les modèles et une partie validation pour l'évaluation. Afin de lisser la variabilité dû à un tirage aléatoire d'échantillons de validation, nous avons généré 10 ensembles aléatoires et stratifiés de données de validation, composés d'un échantillon choisi au hasard pour chaque catégorie de stress. Bien que ces données de validation ne soient pas représentatives des effectifs par stress dans les données d'apprentissage, elles permettent une évaluation du pouvoir discriminant du modèle pour l'ensemble des catégories de stress. Ces dix ensembles d'échantillons de validation ont été construits aléatoirement, avec leur reproductibilité garantie en fixant la graine du générateur aléatoire grâce à un `set.seed()`. A noter que les données d'apprentissage et de validation proviennent toutes les deux de la base de données GEM2Net, ce qui assure qu'elles soient formatées et pré-traitées de la même manière.

Pour tester la robustesse de nos modèles, 10 ensembles de 9 échantillons stratifiés ont été construits. Ainsi, une boucle a été construite permettant de calculer les modèles décrits ci-dessus à chaque tour pour chaque ensemble. Les résultats présentés sont donc des moyennes ou des proportions sur 10 groupes. Cette démarche permet de réduire un potentiel effet des projets d'origine des échantillons, ou la probabilité de n'avoir que des échantillons avec des valeurs extrêmes et potentiellement non représentatifs de la catégorie de stress dans sa globalité.

3.4 Déconvolution non-supervisée de structure incorporant connaissances biologiques

La première partie de notre démarche était de partir des catégories de stress pour effectuer des classifications et prédictions. Une toute autre stratégie est d'effectuer une déconvolution non-supervisée de structure dans les données GEM2Net en lien avec des connaissances biologiques (e.g., pathways). Une telle méthode, Pathway-Level Information ExtractoR (PLIER), a été implémentée dans un package R éponyme [11]. Il présente un fort intérêt de calculs en se basant sur un ensemble de pathways et de gènes pour construire un modèle composé de variables latentes reliées de manière forte aux pathways fournis. Ce modèle permet de mettre en avant des pathways importants pour le jeu de données fourni en entrée, et de cibler des variables latentes les plus pertinentes. Cette méthode a été récemment étendue (dite MultiPLIER) dans le cadre d'un transfert d'apprentissage ("transfer learning") entre une base de données transcriptomiques à grand-échelle et des données plus petites et ciblées liées aux maladies rares chez l'humain [12]. L'idée de cette stratégie avec nos données est donc d'obtenir des variables latentes liées à des

pathways impliqués dans des réactions de la plante face aux stress, et ensuite de projeter de nouvelles données dans cette représentation à faible dimension.

3.4.1 Sélection de voies de signalisation non-redondantes

Pour l'analyse PLIER, nous avons identifié plusieurs base de données différentes de pathways (KEGG, AraCyc), en nous interrogeant sur la redondance potentielle des voies extraites. N'ayant aucun intérêt à avoir plusieurs fois la même information il a fallu trouver un indice pour calculer d'éventuels chevauchements. Dans ce contexte, une distance euclidienne n'est pas adaptée car elle dépend du nombre de gènes impliqués. Nous nous sommes donc tournées vers l'indice de Jaccard, qui consiste à calculer le rapport de l'union et l'intersection de deux listes de gènes. En calculant cet indice pour toutes les voies deux à deux, nous avons pu ensuite regarder la distribution des valeurs de cet indice pour fixer le seuil à partir duquel deux voies sont considérées comme trop proches. Nous avons choisi un seuil de 0.5. Cette méthode nous a permis de retirer environ 30 voies redondantes du modèle avant l'application des filtres.

3.4.2 Estimation du modèle

Pour la construction du modèle PLIER, le tableau de données transcriptomiques dans son intégralité est fourni avec les gènes en lignes (17 341) et les échantillons en colonnes (387). Il est également nécessaire de fournir une matrice binaire (C) indiquant l'appartenance des gènes (en lignes) pour chacun des pathways d'intérêt (en colonnes ; ici pathways AraCyc, KEGG, et GO SLIMs ciblés). Dès qu'un gène est impliqué dans une voie, sa valeur à cette position est de 1, sinon elle est de 0. PLIER effectue de longs calculs, environ 20 min avec la configuration de mon ordinateur. Il applique un certain nombre de filtres dans les données. Lors de ces calculs, afin de mettre en avant les signaux biologiques, PLIER induit de la sparsité avec plusieurs paramètres de régularisation. Un grand nombre de 0 est donc imposé dans le modèle. Pour accélérer les calculs, nous avons fait le choix de filtrer nos données (e.g., enlever l'ensemble de gènes non-impliqués dans des pathways, et pathways possédant < 10 gènes). Ainsi, le modèle a été construit avec 210 pathways et 5026 gènes. De plus, en nous intéressant au package, nous avons pu tester les codes de la vignette associée afin de mieux appréhender les préconisations pour le paramétrage de la fonction `PLIER()`. La fonction `num.pc()` nous a permis d'obtenir une première estimation du nombre de variables latentes. Selon les auteurs de PLIER ou MultiPLIER, cette valeur doit être ensuite multipliée par 2 ou 1,3. Suivant en général, dans cette partie, le raisonnement de MultiPLIER, nous avons gardé leur indicateur pour notre modèle.

3.4.3 Identification de variables latentes pertinentes

Une fois le modèle PLIER estimé, un grand nombre de variables latentes (latent variable ; LV) est obtenu, 55 dans notre cas. Toutes les LV ne présentent pas un intérêt pour les pathways rentrés. Il faut mettre en avant les LV significativement associées avec un ou plusieurs pathways. En sortie de la fonction `PLIER()`, nous obtenons notamment un résumé de l'ensemble des LVs avec quelques indicateurs statistiques. Celui qui nous a permis de mettre en évidence les LV significatives est l'aire en dessous de la courbe ROC (AUC), calculée avec une approche de validation croisée pour les gènes appartenant à chaque pathway. Cet indicateur est compris entre 0 et 1 ; il permet ici de mesurer à quel point une LV est significativement associée avec chaque pathway. Pour cette validation croisée, PLIER met de côté 1/5 des gènes appartenant à un pathway comme ensemble de validation, et construit son modèle avec les données d'apprentissage restantes pour ensuite évaluer si les composantes PLIER parviennent à les reconstituer. Plus les gènes de validation pour un pathway donné sont bien prédites avec cette LV, plus la valeur de l'AUC sera grande et plus l'association est forte. Nous avons pris un seuil d'AUC > 0.75 pour sélectionner les LV d'intérêt.

4 Résultats

4.1 Analyse exploratoire des réponses transcriptomiques au stress dans GEM2Net

Afin de constituer des ensembles de gènes représentant des signatures de réponse à chaque stress et ainsi compléter des sets de gènes utilisés dans nos analyses, nous avons réalisé des analyses différentielles. Le Tableau 1 présente les effectifs des gènes différentiellement exprimés pour les stress abiotiques en fonction du nombre d'occurrences. Peu de gènes, 8 sur les 17341 sont différentiellement exprimés sur 7 des 9 stress au moins. Au contraire, la majorité ne présente pas d'expression différentielle. Ce sont les 4636 gènes qui se distinguent pour un et un seul stress qui nous ont permis de construire nos ensembles signatures.

Nb occurences	0	1	2	3	4	5	6	7	8	9
Nb gènes	9556	4636	2159	661	224	86	11	5	2	1

TABLE 1 – Effectif des DEG par nombre d'occurrences pour l'analyse différentielle effectuée sur l'ensemble des gènes pour les 9 stress abiotiques.

Les relations entre une sous-sélection des signatures identifiées sont illustrées dans la Figure 4. Nous avons tracé cet Upset plot avec le package UpSetR [13]. Les effectifs des signatures de stress sont indiqués sur la gauche du plot, et les intersections entre signatures par les barres verticales. Nous n'observons aucun chevauchement entre les signatures représentées dans la figure, ce qui révèle une forte spécificité des réponses au stress.

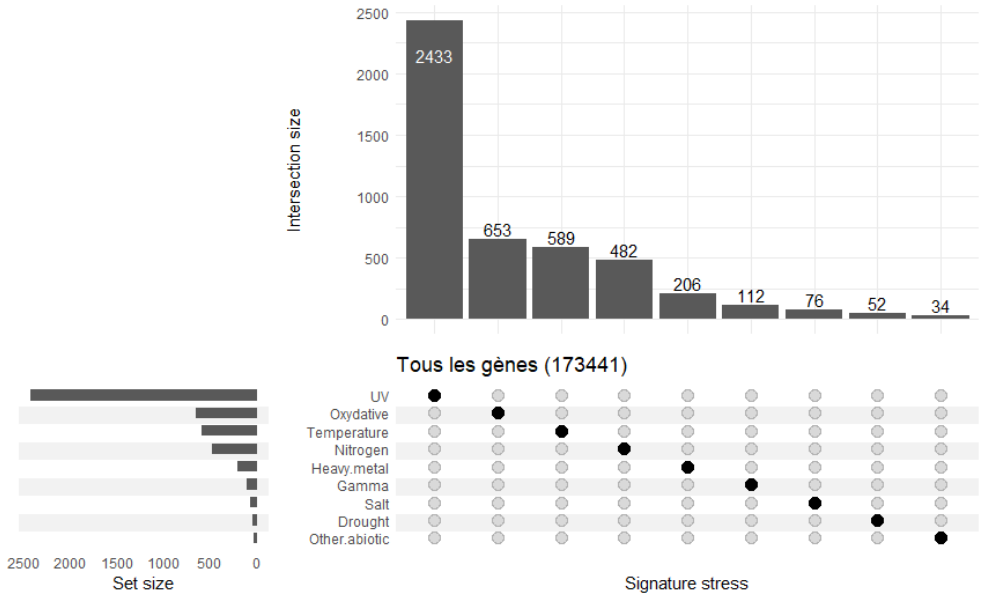
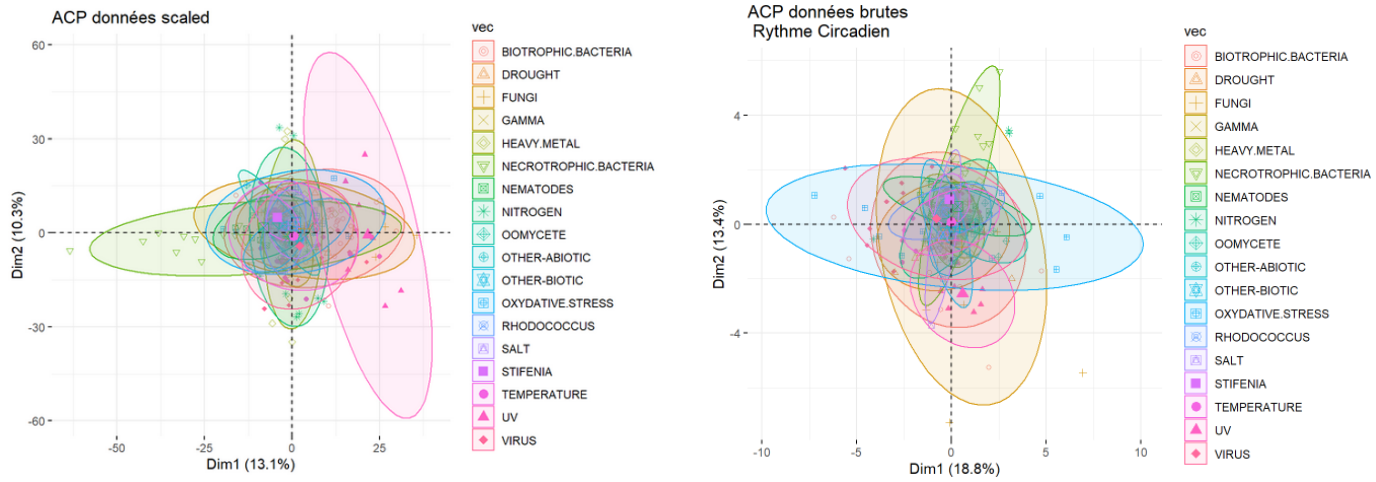


FIGURE 4 – **Upset plot des signatures de stress identifiées par une analyse différentielle**. Chaque signature de stress a été créée grâce à l'utilisation du package limma [7], avec une valeur de seuil de p-value ajustée de 0.05.

Ensuite, une visualisation des données transcriptomiques dans un espace de dimensionnalité réduite a mis en évidence la séparation plus ou moins forte des catégories de stress, ainsi que les stress particuliers qui se démarquent des autres. Sur une première ACP à l'échelle du transcriptome (Figure 5A), on observe un chevauchement important de l'ensemble des stress. Même si les stress *UV* et *bactérie nécrotrophique* semblent se démarquer par la taille des ellipses, aucune séparation réelle n'est observée. De la même manière, sur un ensemble réduit et plus ciblé de gènes associés au GO SLIM rythme circadien (Figure 5B), on retrouve un chevauchement entre toutes les catégories de stress, même si pour cet ensemble, en particulier, ce sont les stress *oxydatif* et *champignon* qui ressortent. De plus, dans un cas comme dans l'autre, le pourcentage de variance expliqué reste relativement faible (23.6% et 32.2% pour

les deux dimensions cumulées). Des résultats similaires ont été obtenus sur l'ensemble des gene sets étudiés (voir Annexe 1 pour d'autres exemples).



A. ACP globale

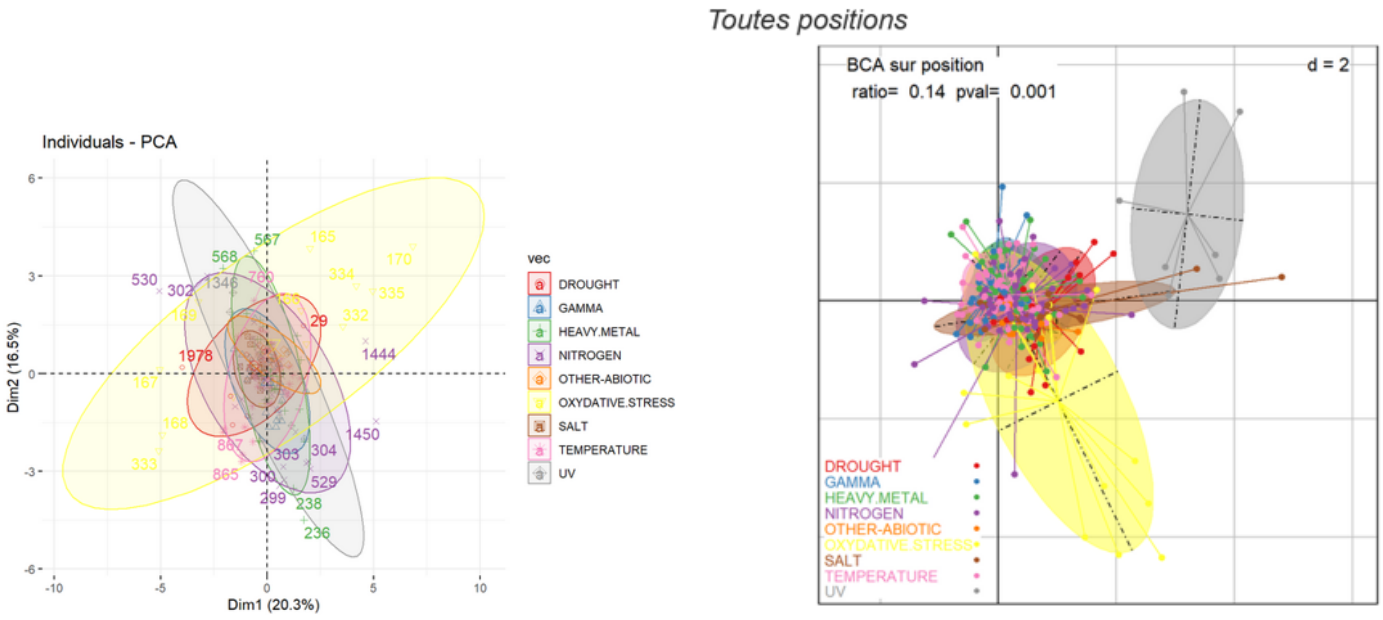
B. ACP GO SLIM rythme circadien

FIGURE 5 – **Visualisation des deux premières composantes d'une ACP à l'échelle globale du transcriptome et pour les gènes du rythme circadien**, coloration et ellipses de confiance par catégorie de stress (18 stress), A. pour l'ensemble des données GEM2Net (17 341 gènes), B. uniquement pour les gènes GO SLIM rythme circadien (p=57).

Cette première visualisation nous a permis de nous rendre compte de la complexité de la séparation de ces catégories de stress dans la base de données GEM2Net. Nous nous sommes rendues compte également du poids de certaines catégories de stress, comme *UV*. Cette catégorie en particulier ne présente pas de valeurs particulièrement extrêmes. Même dans un ensemble de 50 gènes choisis au hasard, l'allure de l'ACP est la même que pour les analyses globales ou ciblées. Pour alléger ces figures et nous focaliser sur des ensembles moins hétérogènes de GEM2Net, nous avons dès ce moment séparé les stress biotiques (163 échantillons) et abiotiques (224 échantillons), chacun composé de 9 catégories. Pour correspondre aux objectifs du projet basés sur une problématique de température, nous avons privilégié les stress abiotiques.

Avant de nous focaliser sur les analyses de classification supervisée, nous avons voulu comparer les résultats observés pour l'ACP avec une BCA. Par cette démarche, nous avons voulu évaluer à quel point certaines catégories de stress pouvaient se démarquer grâce à cette analyse multivariée supervisée (Figure 6B) comparée à une analyse multivariée non-supervisée (Figure 6A). Les BCA ont été réalisées avec le package ade4 [9]. Comme attendu, les stress *UV* et *oxydatif* se retrouvent éloignés des autres catégories. On remarque davantage l'ellipse pour le stress *sel* avec certains échantillons vraiment démarqués, mais le chevauchement entre catégories reste trop important pour pouvoir distinguer des groupes précis.

Au vue de ces résultats, nous nous sommes interrogées sur les catégories de stress et la manière dont elles ont été construites. Ayant à notre disposition d'autres métadonnées, nous avons reconstruit les ACP avec les identifiants des projets comme noms des points (Annexe 2). Certains projets se retrouvent en effet regroupés sans généralisation possible malgré une homogénéisation et un pré-traitement des données. Par ailleurs, les SWAP ID nous ont également servi comme noms de points pour identifier les échantillons spécifiques qui entraînaient un allongement des ellipses, en particulier pour la catégorie de stress *UV*.



A. ACP Swap ID B. BCA

FIGURE 6 – **ACP et BCA pour les stress abiotiques ($n = 224$) et les gènes du rythme circadien ($p = 57$).** Coloration et ellipses de confiance par catégorie de stress abiotique (9 stress), A. pour une ACP, avec quelques Swap IDs soulignés, B. pour une BCA.

4.2 Classification supervisée des réponses transcriptomiques au stress dans GEM2Net

Les modèles de classification supervisée calculés grâce à la LDA ont permis de faire des prédictions sur nos données de validation acquises par échantillonnages stratifiés. Pour évaluer la précision de ces classifications, nous avons calculé les probabilités d'attribution de chaque stress pour les données de validation, ainsi que leur classification finale obtenue par la règle de maximum a posteriori (MAP). Ces valeurs sont représentées dans un heatmap (Figure 7A et B), tracé avec le package ComplexHeatmap [14]. Nous pouvons y voir la probabilité d'affectation correcte (moyennée sur 10 ensembles de validation) des catégories de stress (en colonnes) par rapport aux ensembles de gènes fournis en entrée (lignes). D'autre part, les heatmaps mettent en évidence la proportion d'échantillons correctement classée sur les 10.

Le clustering des colonnes du heatmap des probabilités d'attribution (Figure 7, gauche) met en avant les stress dont les résultats de prédiction sont les meilleurs. Comme attendu, le stress *UV* a la plus grande confiance de prédiction, avec au moins 0.8 de probabilité. Le stress *métal lourd* a de fortes probabilités également. Nous pensions obtenir une corrélation entre la signature d'un stress et les valeurs obtenues pour celui-ci. Pour l'ensemble de stress *abiotique "autre"*, la confiance de prédiction est assez forte (0.7) mais pas aussi marqué que pour la catégorie UV. Le heatmap des attributions par la règle de MAP (Figure 7, droite) permet d'apporter une information complémentaire. Pour les 10 ensembles de données de validation, on peut constater que malgré des probabilités de prédiction relativement faibles, en dessous d'un seuil de 0.7 par exemple, certains stress se retrouvent tout de même bien classés par la règle du MAP, notamment *UV* et *métal lourd* qui sont quasiment systématiquement classés correctement. Certains stress se retrouvent avec des scores intéressants pour certains ensembles de gènes seulement, comme le stress *oxydatif* qui est globalement mal prédit sauf pour les gènes appartenant à la signature de stress *abiotique "autre"*. En ce qui concerne les signatures de stress, nous constatons que celle pour le stress *Gamma* semble être souvent liée à de mauvaises classifications pour l'ensemble des catégories de stress. En revanche, les gènes dans le terme GO SLIM rythme circadien classe correctement tous les stress pour la majorité des données de validation.

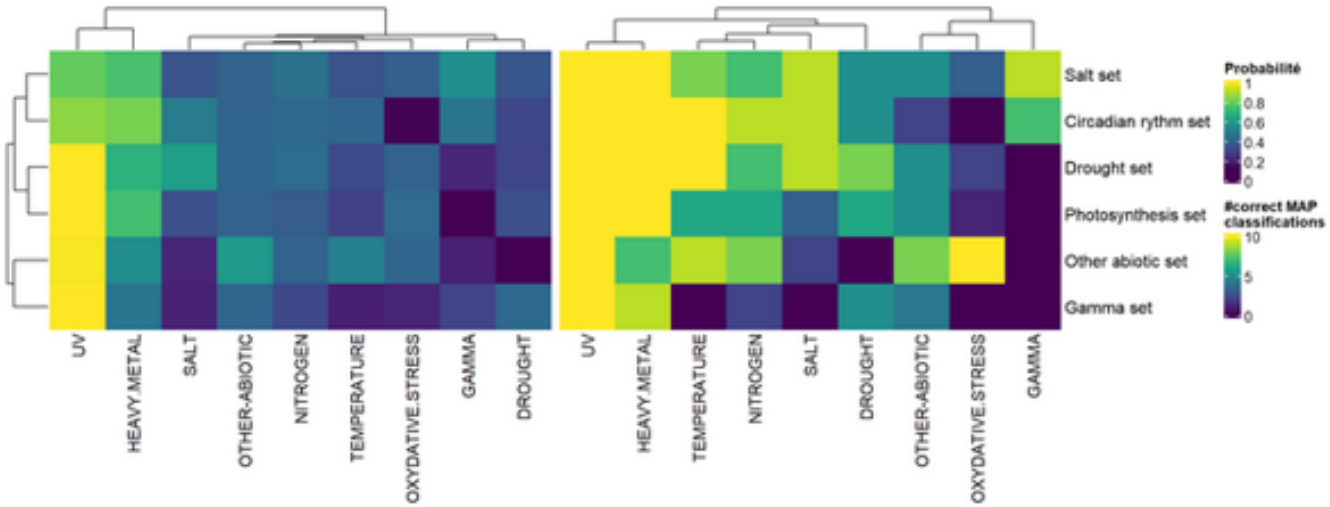


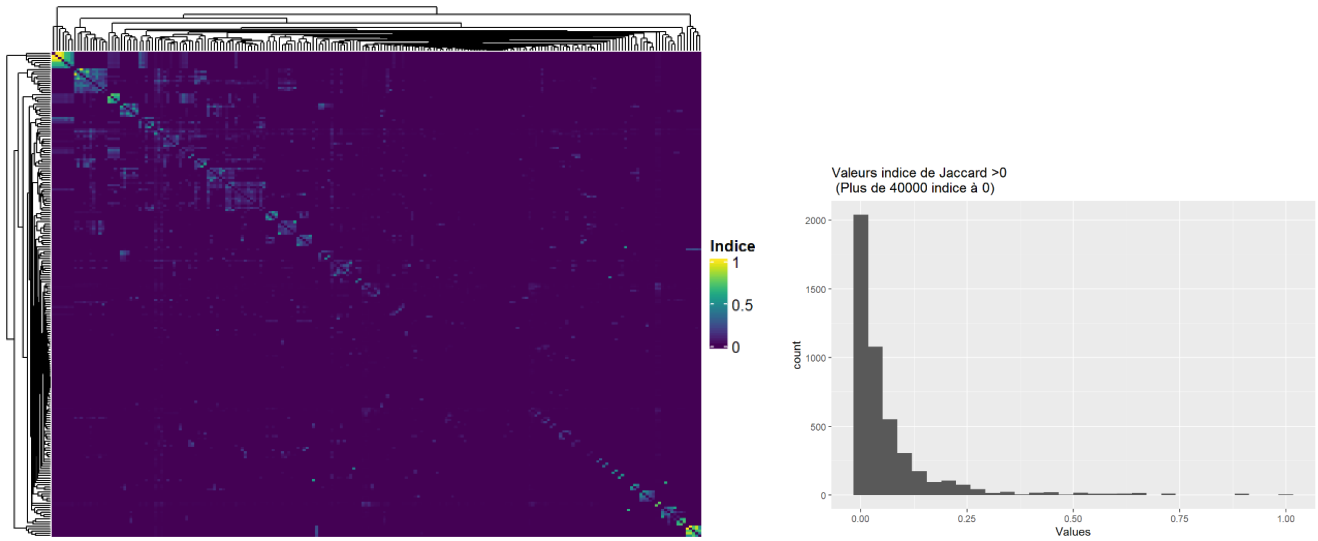
FIGURE 7 – **Heatmap représentant la qualité des prédictions LDA par catégorie de stress sur un sous-ensemble d'ensembles de gènes ($p < 150$)**, les deux heatmaps ont été construits avec le package ComplexHeatmap [14] et la palette viridis. (A) Matrice de confiance de prédiction, correspondant à la probabilité d'attribution correcte pour les échantillons de validation (en colonnes, moyenné sur 10 échantillonnages) par rapport aux ensembles de gènes (en lignes) ; (B) matrice de proportions d'échantillons bien classés (sur 10 échantillonnages stratifiés) par la règle de MAP.

La comparaison des probabilités de prédiction pour les différentes méthodes utilisées (multiclasse, ACP, one-versus-all) a montré que la stratégie multiclasse reste la plus performante pour la majorité des stress et des ensembles de gènes, que ce soit pour la LDA ou la PLS-DA (Annexe 3). La PLS-DA permettait de pouvoir calculer plusieurs types de distances pour la détermination des classes. Nous avons testé celle des centroïdes et la distance maximale (Annexe 4). La proportion d'échantillons correctement classés est très faible même pour un stress comme UV qui se démarquait dans d'autres analyses. En utilisant la distance maximale, le stress *azote* était très souvent correctement classifié.

4.3 Déconvolution non-supervisée de GEM2Net avec AraCyc, KEGG et GOSLIM par PLIER

La classification précise des catégories de stress dans GEM2Net peu convaincante, nous avons privilégié une autre approche : la déconvolution non-supervisée de structure dans les données GEM2Net avec la méthode PLIER. Une étape importante est donc la sélection de voies de signalisation à retenir par les termes GO SLIM et les bases de données KEGG et AraCyc.

Afin d'identifier et d'éliminer des voies particulièrement redondantes, nous avons calculé un indice de Jaccard entre chaque paire de pathways. La Figure 8A permet de visualiser la répartition des indices de Jaccard entre chaque paire de pathways (colonnes et lignes). Plus l'indice est proche de 0, bleu, plus les pathways sont composés de gènes distincts et inversement, plus la couleur est claire ici plus le chevauchement entre les pathways est grand. La valeur de la diagonale a été fixée à 0 pour faciliter la visualisation. Grâce à cette visualisation, on peut identifier des clusters de pathways très redondants, situés en particulier autour de la diagonale grâce à la clusterisation hiérarchique des lignes et des colonnes. La Figure 8B montre la répartition des valeurs de l'indice de Jaccard à travers l'ensemble des paires de pathways. Pour rendre le graphique plus lisible, toutes les valeurs égales à 0 ont été supprimées de cette représentation. On observe que la majorité des valeurs se situe entre 0 et 0.25, ce qui n'est pas surprenant compte tenu de l'implication de certains gènes dans plusieurs pathways. Nous avons décidé de ne garder qu'un pathway représentatif (le premier en ordre alphabétique) dans les cas où l'indice de Jaccard dépassait le seuil de 0.5. Ce filtrage a mené à la suppression de 28 pathways redondants.



A. Valeurs de l'indice de Jaccard

B. Distribution indices de Jaccard

FIGURE 8 – **Identification de pathways redondants avec l'indice de Jaccard**, (A) Heatmap des indices de Jaccard entre chaque paire de pathways (210 voies), tracé par ComplexHeatmap [14] et la palette viridis. Plus la couleur tend vers jaune, plus l'indice se rapproche de 1 et plus les pathways concernés sont redondants. Les valeurs de diagonales ont été fixées à 0. (B) Distribution des valeurs non-nulles de l'indice de Jaccard, tracée avec ggplot2.

Une fois la sélection de pathways stabilisée, nous avons pu appliquer la méthode PLIER aux données GEM2Net. Parmi les sorties de PLIER, la matrice de coefficients (U) liant les variables latentes (Z) à la matrice indiquant l'appartenance aux pathways des gènes (C) est particulièrement intéressante à visualiser. Elle a en colonnes les LV et en lignes les différents pathways du modèle. Ses valeurs sont normalisées par la valeur maximale observée pour chaque composante et sont généralement représentées en valeur absolue. Elles varient donc de 0 à 1. Comme PLIER impose une sparsité pour U grâce à un paramètre de rétrécissement, la majorité des éléments de cette matrice prend une valeur 0. Cela permet de se focaliser en particulier sur les LV fortement associées à un ou plusieurs pathways. Une sous-partie de la matrice U pour les LVs significativement associées à au moins un pathway est représentée dans la Figure 9 (AUC > 0.75). Les cases violettes correspondent à des indicateurs d'association entre la LV (en colonnes) et le pathway (en lignes).

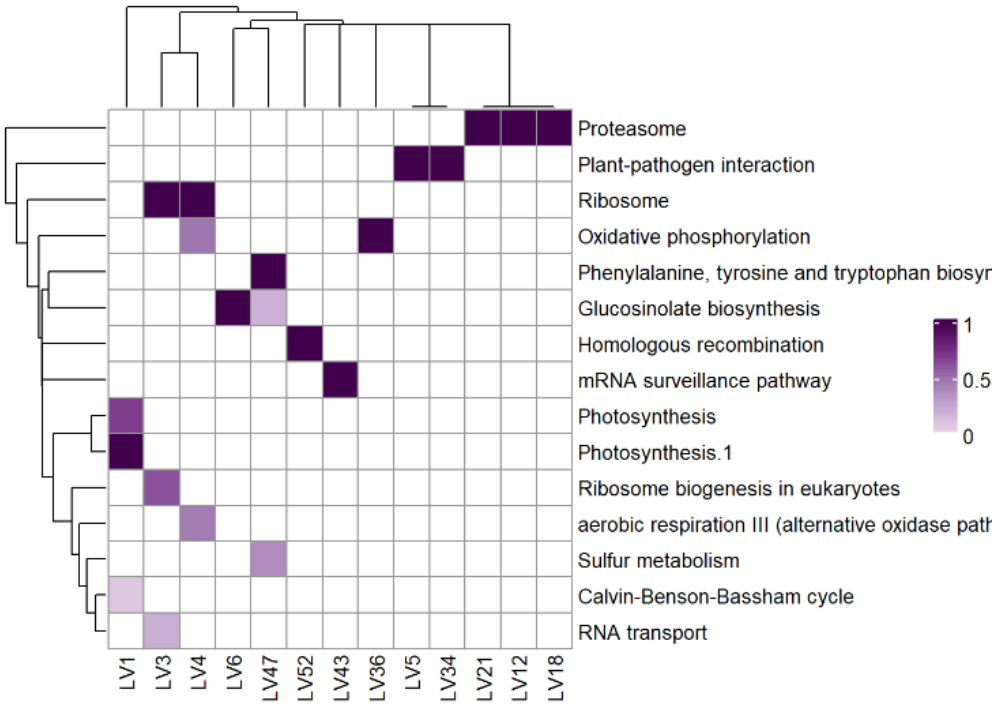


FIGURE 9 – **Visualisation par heatmap d'une sous-partie de la matrice U du modèle PLIER**, reliant les LVs et pathways. La matrice U fournit en colonnes les LVs calculées et en colonnes les pathways. Ne sont tracées ici que les LVs fortement associées à au moins un pathway (AUC > 0.75), les valeurs indiquées sont celles de U en valeur absolue et normalisées par la valeur maximale correspondante.

Nous pouvons récupérer plus d'informations concernant les LV significatives grâce au résumé que propose PLIER. On y retrouve le nom du pathway et l'indice de la LV pour laquelle une forte association a été observée (Table 2). Pour chaque élément de ce tableau, un certain nombre d'indicateurs statistiques permet d'évaluer la force d'association entre les LVs et les pathways : AUC, p-value (calculée avec un test de permutation) et FDR. Plusieurs pathways étant liés à une même LV, un seul est retenu après avoir pris en compte les trois indices (AUC la plus importante et p-value et FDR les plus petits possibles).

Pathway	LV index	AUC	p-value	FDR
Photosynthesis.1	1	0.870	9.00e-09	1.49e-07
Ribosome	3	0.958	4.25e-28	4.21e-26
Ribosome	4	0.917	4.29e-24	2.12e-22
Biotic	5	0.627	2.19e-06	1.97e-05
Glucosinolate biosynthesis	6	1.000	5.25e-04	2.17e-03
Proteasome	12	0.956	5.48e-07	6.78e-06
Proteasome	18	0.870	3.74e-05	2.18e-04
Endocytosis	21	0.701	3.63e-04	1.63e-03
Biotic	34	0.592	6.73e-04	2.67e-03
Oxidative phosphorylation	36	0.844	7.57e-07	8.32e-06
Spliceosome	43	0.651	1.97e-03	5.91e-03
Phenylalanine, tyrosine and tryptophan biosynthesis	47	0.838	9.50e-04	3.48e-03
Homologous recombination	52	0.752	3.76e-03	9.80e-03

TABLE 2 – **LVs significatives de PLIER et leurs pathways associés**. Les statistiques pour chaque paire LV-pathway correspondent à l'aire sous la courbe (AUC), p-value et false discovery rate (FDR).

Il est ainsi possible de repérer les pathways liés à une LV précise à l'image du protéasome pour les LV21, LV12 et LV18, ainsi que le terme GO SLIM Photosynthesis.1 pour la LV1. Cet ensemble de gènes faisait partie (avec celui du rythme circadien) des deux GO SLIM les plus précis des 10 que nous avions étudiés précédemment. La présence d'un autre pathway Photosynthesis, cette fois provenant de KEGG, nous a poussées à confirmer notre sélection de pathways via l'indice de Jaccard. Le pathway de KEGG contient 33 gènes tandis que le GO SLIM en possède 75. En faisant un recoupement entre ces gènes, nous avons observé que seulement 5 gènes étaient communs à ces deux listes. Nous avons donc conclu que, malgré le fait que les deux sets partagent un même nom, il s'agit bien de deux pathways différents. Aussi, malgré une valeur d'association plus faible, le cycle de Calvin-Benson se retrouve associé à la LV1. Ces trois pathways, tous les trois liés par la photosynthèse sont donc regroupés avec la même LV (Annexe 4). Il est donc intéressant de retenir cette variable latente pour aller plus loin dans l'interprétation du modèle PLIER calculé.

Une autre sortie pertinente du modèle PLIER est la matrice des LVs (B), qui contient les composantes en lignes et les échantillons en colonnes. Pour revenir aux catégories de stress étudiées dans la première partie du stage, nous avons observé le profil de ces LVs significatives en découpant par stress abiotiques (Figure 10 pour la LV1). A l'image de nos premières conclusions concernant l'interprétabilité de ces catégories, seuls les stress *UV* et *oxydatif* semblent se démarquer des autres. La même tendance a été observée sur l'ensemble des LV ; aucune ne peut servir de référence pour un stress ou ne semble corrélée à l'expression d'un stress.

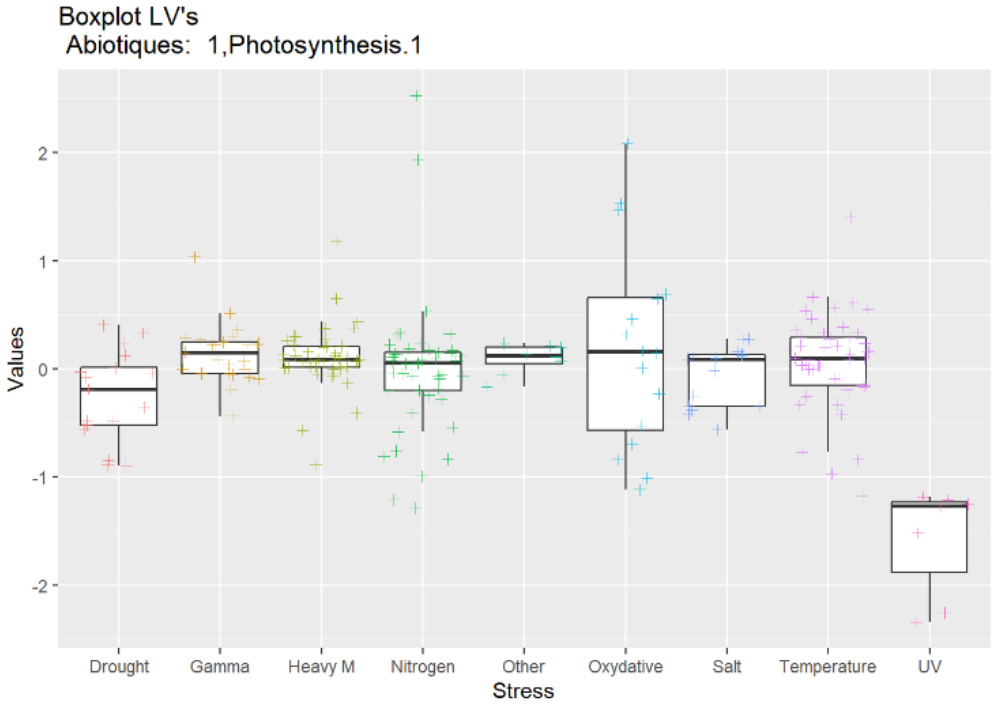


FIGURE 10 – **Boxplot des valeurs de la LV1 (significativement associée avec le pathway Photosynthesis.1) par catégorie de stress abiotiques, tracé avec ggplot.**

4.4 Vers un transfert de connaissances depuis GEM2Net : de PLIER à MultiPLIER

En calculant un modèle PLIER, nous souhaitions utiliser la démarche décrite pour MultiPLIER [12] afin d'effectuer un transfert du modèle appris sur GEM2Net vers de nouvelles données. La fonction `GetNewDataB()` permet de calculer, à partir d'un modèle PLIER et d'une nouvelle matrice d'expression (au même format que les données d'expression de base), une nouvelle matrice de LVs B.

Ayant repéré la LV1 qui semble fortement associée à la photosynthèse, nous avons donc cherché un autre jeu de données transcriptomiques correspondant à une problématique liée à la photosynthèse. L'étude choisie [15] s'intéresse à la chronologie d'expression des gènes impliqués dans la sénescence des feuilles d'Arabidopsis. Ce sont des données mésurées sur la même puce (CATMA V3) que celle utilisée pour la base de données GEM2Net, mais qui ne sont pas incluses dans cette dernière. Cette étude contient 22 échantillons, correspondant à 2 prélèvements (matin et après-midi) sur 11 jours d'expérience. Les données pré-processées sont librement disponibles au NCBI (GSE22982) et ont été téléchargées grâce au package GEOquery [6]. Chaque échantillon dans la matrice d'expression correspond à la moyenne de 4 réplicats.

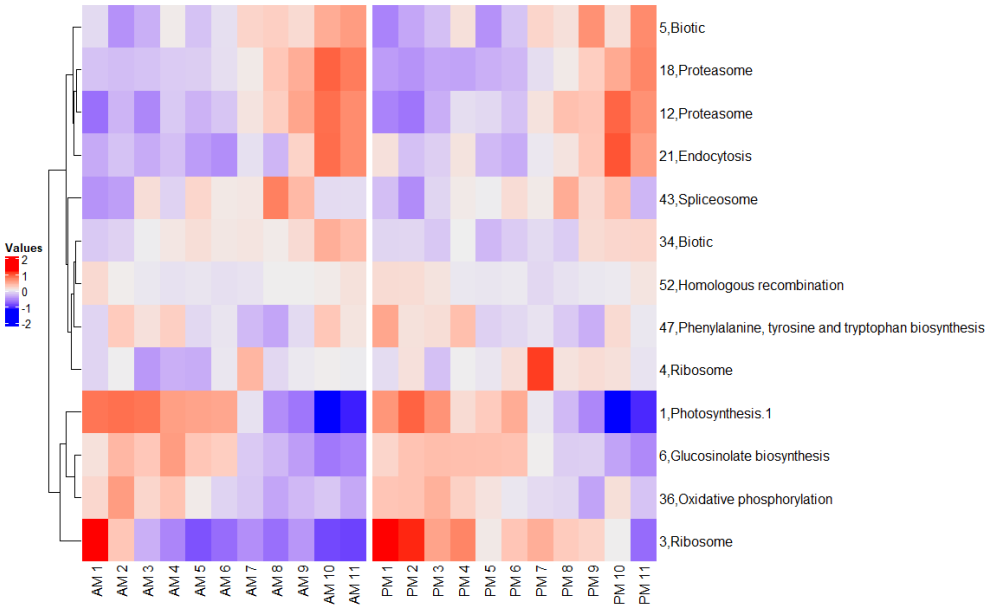


FIGURE 11 – **Heatmap des valeurs projetées des LVs significatives** pour les nouvelles données liées à la photosynthèse [15]. Les échantillons du matin ont été séparés de ceux de l'après-midi. L'ordre chronologique a été conservé ; une colonne représente donc une journée de l'étude. Seules les valeurs pour les LV significatives sont tracées.

Nous visualisons les valeurs projetées des LVs significatives pour ce nouveau jeu de données dans l'espace du modèle construit sur GEM2Net (Figure 11). Le heatmap sépare les données du matin de celles de l'après-midi afin de faciliter la lecture. Les colonnes respectent la chronologie des jours de l'étude en se déplaçant de gauche à droite. Le clustering hiérarchique des lignes pour les LV permet d'identifier les 4 LV significatives qui ressortent le plus pour ce jeu de données : *Photosynthesis.1, Glucosinolate biosynthesis, Oxidative phosphorylation et Ribosome*. Pour la LV liée à la photosynthèse, nous constatons que les valeurs diminuent au cours de l'expérience. La temporalité de cette étude nous permet également une autre visualisation des valeurs projetées de ces LVs. En séparant de nouveau les tendances du matin de celles de l'après-midi, nous traçons les courbes de tendance à travers les jours de l'étude pour chaque LV significative (Figure 12). Nous constatons que certaines LV présentent des variations plus marquées que d'autres, en particulier la LV1, associée avec le pathway Photosynthesis.1. Celle-ci diminue fortement tout au long des 11 jours de l'étude. Cela semble suggérer que le signal biologique porté par cette LV, construite sur les données GEM2Net, est retrouvé également dans un jeu de données externe, non-utilisé dans la construction du modèle PLIER.

FIGURE 12 – **Tendances temporelles des valeurs projetées de LV significatives pour les nouvelles données liées à la photosynthèse**, pour les échantillons du matin (rouge) et de l'après-midi (bleu) pour chaque LV significative. Nous retrouvons les 11 jours de l'étude en abscisse. La même échelle a été choisie pour l'ensemble des LV pour faciliter les comparaisons entre LVs.

5 DISCUSSION

L'enjeu du stage était d'observer et de caractériser les mécanismes moléculaires liés à la réponse au stress de plantes non-modèles présentant un intérêt agronomique. Pour celles-ci, le nombre de données disponibles est limité et les conditions expérimentales en champs sont plus complexes qu'en chambres contrôlées. Dans le même temps, les plantes modèles comme Arabidopsis sont très largement étudiées et ces connaissances sont disponibles dans des bases de données publiques. Dans ce stage, nous avons voulu étudier la faisabilité d'un transfert de ces connaissances acquises vers d'autres données de moindres dimensions, que ce soit pour d'autres expériences Arabidopsis, ou de manière plus ambitieuse. pour d'autres espèces.

5.1 Conclusions

Une des premières questions que nous nous sommes posée concernait la manière de découper notre jeu de données. Les GO SLIM choisis correspondaient aux termes les plus parlants et spécifiques aux plantes. La grande taille de certains de ces ensembles semblait être un avantage dans un premier temps mais elle s'est révélée trop large et pas assez spécifique comme l'a prouvé par exemple le GO SLIM Photosynthèse (75 gènes) qui ne possède que quelques gènes en commun avec le pathway Photosynthèse de KEGG. Ce terme GO SLIM fait pourtant parti avec celui du rythme circadien des deux gene sets les plus petits et les plus précis. Pour ces deux termes, les résultats étaient plus intéressants en révélant une plus grande variance expliquée lors des classifications.

Le coeur de notre travail a été de s'appuyer sur GEM2Net, une base de données transcriptomiques publique à grand-échelle qui répertorie un nombre important d'expériences en lien avec les stress biotiques et abiotiques chez la plante modèle Arabidopsis. GEM2Net représente une grande source de données de puces à ADN ayant subi des protocoles techniques et bioinformatiques homogènes, permettant des analyses globales intéressantes. Les opérations effectuées sur l'ensemble des expériences permettent donc de pouvoir les considérer sur le même plan et de les comparer entres elles. Les expériences dans GEM2Net ont été classifiées au préalable en différentes catégories de stress, définissant ainsi une variable potentiellement utilisable pour des stratégies de classification et de prédiction. Ce fut tout l'enjeu de notre première démarche.

Certains stress se sont retrouvés démarqués lors des différentes classifications sans que cela soit lié à des valeurs extrêmes différentes des autres groupes. Quelque soit la méthode de calculs utilisée, les stress se chevauchaient et ne permettaient pas la construction d'un modèle de projection et de prédiction fiable et robuste. La LDA nous a permis d'obtenir des résultats de prédiction fiables pour seulement deux catégories de stress (*UV* et *métal lourd*). Des résultats similaires ont été observés quelque soit la méthode (BCA, LDA, PLS-DA), la stratégie (multiclasse, "one-versus-all") ou l'ensemble de gènes utilisés. Nous avons remarqué que ces analyses ne considéraient pas l'intégralité des données.

La stratégie de déconvolution non-supervisée par PLIER présentait l'avantage de prendre en compte l'ensemble des données, ainsi que toute une collection de pathways. Nous avons pu passer outre les catégories de stress et aboutir à un modèle contenant des variables latentes associées pour certaines à un ou plusieurs pathways fournis sur la base des valeurs d'expression. In fine, 13 variables latentes significatives ressortent du modèle. Grâce à la stratégie étendue de PLIER, MultiPLIER, nous avons pu commencer à exploiter la LV1, fortement associée à la photosynthèse, avec un nouveau jeu de données transcriptomiques [15]. La tendance des valeurs calculées à partir de notre modèle se rapproche des observations faites de la baisse de l'activité photosynthétique pendant la sénescence de la feuille dans l'étude.

5.2 Limites

La complexité de la manipulation de l'intégralité des données, ainsi que l'interprétabilité de nos résultats, nous ont conduites à identifier plusieurs ensembles de gènes plus ciblés, y compris les signatures de stress identifiées par des analyses différentielles et les termes GO SLIM. Pour certaines de nos analyses, les catégories de stress biotiques et abiotiques ont également été analysées séparément. Il serait intéressant d'étudier si d'autres ensembles de gènes pourraient mener à une discrimination plus importante entre les catégories de stress.

Notre analyse a permis de réaliser que les échantillons qui composaient les catégories de stress n'étaient pas complètement homogènes, malgré le pré-traitement homogénéisé dans GEM2Net. Pour donner un exemple concret, le projet 87 de CATdb fait partie des échantillons du stress *température*. Les données pour ce stress étaient extrêmement variables. Cependant, en regardant de plus près la fiche descriptive de ce projet, nous nous sommes rendues compte que la problématique de cette étude semblait porter plutôt sur la caractérisation d'un variant. La température n'intervient pas comme stratégie particulière dans la partie expérimentale, mais se retrouve citée comme condition de culture de ce mutant. A partir de cette observation, suite à un entretien avec Catherine Giauffret, nous nous sommes interrogées sur l'impact de la structuration des données par projet. Afin d'approfondir cette observation, les analyses précédentes de classification ont été reprises en utilisant comme noms de points les identifiants des projets. Sur les ACP obtenues pour chaque gene set, les échantillons d'un même projet se sont souvent retrouvés regroupés ensemble (Annexe 2A). Un clustering par K-means sur ces données a confirmé qu'une structuration semble plus forte pour l'effet projet que pour les catégories de stress (Annexe 2B).

Concernant PLIER, bien que les pathways utilisés semblent constituer une liste exhaustive des informations présentes sur les bases de données, il est possible que certains signaux biologiques dans nos données ne puissent pas être distingués avec le modèle actuel. Une perte d'informations est possible à ce niveau. Par manque de temps aussi, une seule variable latente a pu être exploitée davantage.

5.3 Perspectives

Le questionnement sur la construction des catégories de stress nous a permis de réaliser que la classe de stress attribuée à un échantillon ne correspondait pas forcément à un mot clé centrale de l'étude correspondante au projet. Un travail important bibliographique pour redéfinir les termes importants de la cinquantaine de projets composant actuellement GEM2Net serait à prévoir afin d'obtenir une variable plus précise et peut être de meilleurs résultats de classification. Il est probable qu'un certain nombre d'échantillons soit dans une catégorie de stress inappropriée. Le temps ayant été précieux pendant ce stage nous n'avons pas pu approfondir cette recherche.

Les 50 gènes aléatoires tirés ont montré des résultats similaires à ces termes GO SLIM trop larges. La qualité et la précision des ensembles sélectionnés est donc importante dans cette analyse. Par ailleurs, l'analyse différentielle effectuée a permis de récupérer des petits groupes de gènes bien plus spécifiques nous permettant d'exécuter nos scripts sur au moins 8 ensembles de gènes à chaque nouvelle analyse. Il serait donc pertinent de reprendre la liste des GO SLIM et de sélectionner de nouveaux termes aussi précis que ceux de la Photosynthèse et du rythme circadien. Par ailleurs, GEM2Net propose un certain nombre de clusters de gènes prêts à être exploités. Il peut être intéressant pour la suite de les étudier et d'évaluer leurs pertinences.

Actuellement, notre modèle PLIER est construit avec les paramètres par défaut, excepté celui du nombre de LV dont la méthode a été décrite précédemment. Pour conserver un plus grand nombre de pathways intéressants pour nous, c'est-à-dire, impliqués dans la réponse au stress comme la biosynthèse de l'éthylène par exemple, il serait peut être judicieux de reconstruire un nouveau modèle PLIER avec un seuil plus stringent du nombre de gènes impliqués. Suite à l'analyse PLIER, le modèle obtenu semble avoir une dizaine de variables latentes d'intérêts. Par manque de temps, nous n'avons pas pu exploiter longuement les LV obtenues ou trouver de test quantitatif intéressant pour évaluer le lien entre cette LV et le pathway associé. La variable latente Photosynthèse serait particulièrement appropriée pour notre étude si le lien s'avère effectif.

6 REFERENCES BIBLIOGRAPHIQUES

Références

[1] Rim Zaag et al. « GEM2Net : from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response ». eng. In : *Nucleic Acids Research* 43.Database issue (jan. 2015), p. D1010-1017. ISSN : 1362-4962. DOI : 10.1093/nar/gku1155.

[2] Séverine Gagnot et al. « CATdb : a public access to Arabidopsis transcriptome data from the URGV-CATMA platform ». In : *Nucleic Acids Research* 36.suppl_1 (jan. 2008), p. D986-D990. ISSN : 0305-1048. DOI : 10.1093/nar/gkm757. URL : https://doi.org/10.1093/nar/gkm757 (visité le 12/04/2021).

[3] E. Huala et al. « The Arabidopsis Information Resource (TAIR) : a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant ». eng. In : *Nucleic Acids Research* 29.1 (jan. 2001), p. 102-105. ISSN : 1362-4962. DOI : 10.1093/nar/29.1.102.

[4] Lukas A. Mueller, Peifen Zhang et Seung Y. Rhee. « AraCyc : a biochemical pathway database for Arabidopsis ». eng. In : *Plant Physiology* 132.2 (juin 2003), p. 453-460. ISSN : 0032-0889. DOI : 10.1104/pp.102.017236.

[5] M. Kanehisa et S. Goto. « KEGG : kyoto encyclopedia of genes and genomes ». eng. In : *Nucleic Acids Research* 28.1 (jan. 2000), p. 27-30. ISSN : 0305-1048. DOI : 10.1093/nar/28.1.27.

[6] Sean Davis et Paul S. Meltzer. « GEOquery : a bridge between the Gene Expression Omnibus (GEO) and BioConductor ». eng. In : *Bioinformatics (Oxford, England)* 23.14 (juil. 2007), p. 1846-1847. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btm254.

[7] Matthew E. Ritchie et al. « limma powers differential expression analyses for RNA-sequencing and microarray studies ». In : *Nucleic Acids Research* 43.7 (avr. 2015), e47-e47. ISSN : 0305-1048. DOI : 10.1093/nar/gkv007. URL : https://doi.org/10.1093/nar/gkv007 (visité le 17/06/2021).

[8] Sébastien Lê, Julie Josse et François Husson. « FactoMineR : An R Package for Multivariate Analysis ». en. In : *Journal of Statistical Software* 025.i01 (2008). Publisher : Foundation for Open Access Statistics. URL : https://ideas.repec.org/a/jss/jstsof/v025i01.html (visité le 17/06/2021).

[9] Stéphane Dray et Anne-Béatrice Dufour. « The ade4 Package : Implementing the Duality Diagram for Ecologists ». en. In : *Journal of Statistical Software* 22.1 (sept. 2007). Number : 1, p. 1-20. ISSN : 1548-7660. DOI : 10.18637/jss.v022.i04. URL : https://www.jstatsoft.org/index.php/jss/article/view/v022i04 (visité le 17/06/2021).

[10] Florian Rohart et al. « mixOmics : An R package for 'omics feature selection and multiple data integration ». en. In : *PLOS Computational Biology* 13.11 (nov. 2017). Publisher : Public Library of Science, e1005752. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1005752. URL : https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752 (visité le 17/06/2021).

[11] Weiguang Mao et al. « Pathway-level information extractor (PLIER) for gene expression data ». en. In : *Nature Methods* 16.7 (juil. 2019). Number : 7 Publisher : Nature Publishing Group, p. 607-610. ISSN : 1548-7105. DOI : 10.1038/s41592-019-0456-1. URL : https://www.nature.com/articles/s41592-019-0456-1 (visité le 17/03/2021).

[12] Jaclyn N. Taroni et al. « MultiPLIER : A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease ». en. In : *Cell Systems* 8.5 (mai 2019), 380-394.e4. ISSN : 24054712. DOI : 10.1016/j.cels.2019.04.003. URL : https://linkinghub.elsevier.com/retrieve/pii/S240547121930119X (visité le 28/02/2021).

[13] Jake R. Conway, Alexander Lex et Nils Gehlenborg. « UpSetR : an R package for the visualization of intersecting sets and their properties ». eng. In : *Bioinformatics (Oxford, England)* 33.18 (sept. 2017), p. 2938-2940. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btx364.

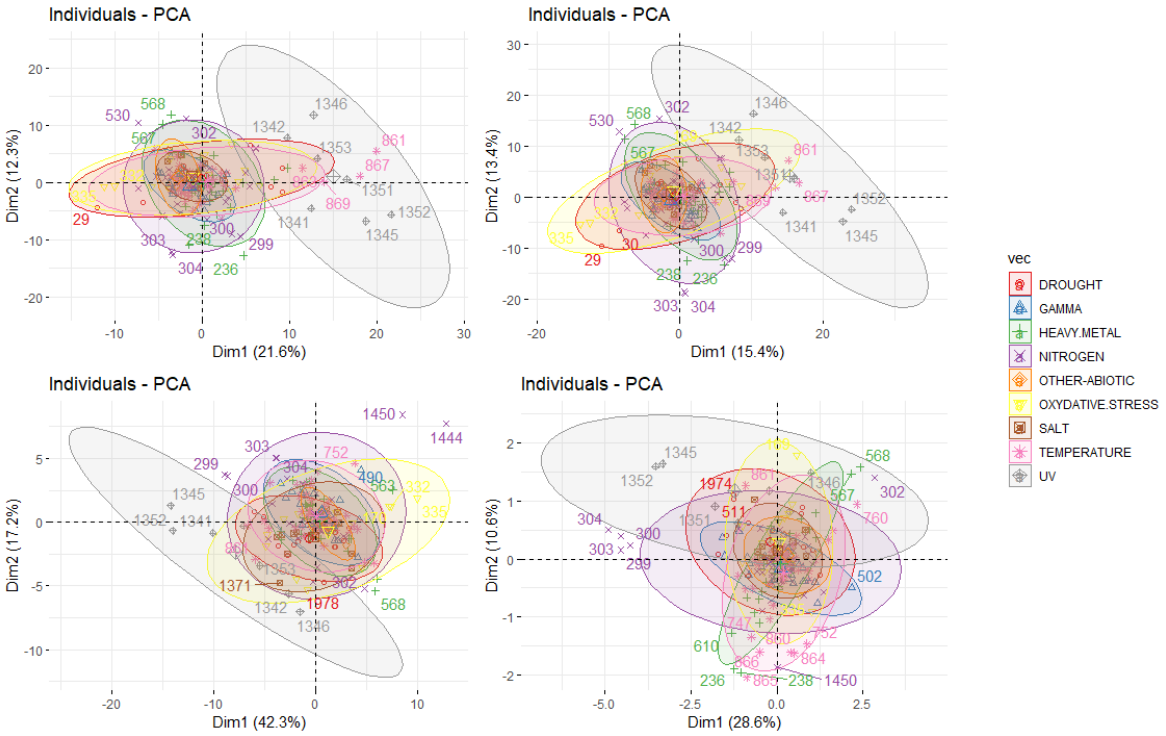
[14] Zuguang Gu, Roland Eils et Matthias Schlesner. « Complex heatmaps reveal patterns and correlations in multidimensional genomic data ». eng. In : *Bioinformatics (Oxford, England)* 32.18 (sept. 2016), p. 2847-2849. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btw313.

[15] Emily Breeze et al. « High-Resolution Temporal Profiling of Transcripts during Arabidopsis Leaf Senescence Reveals a Distinct Chronology of Processes and Regulation ». In : *The Plant Cell* 23.3 (mar. 2011), p. 873-894. ISSN : 1040-4651. DOI : 10.1105/tpc.111.083345. URL : https://doi.org/10.1105/tpc.111.083345 (visité le 18/06/2021).

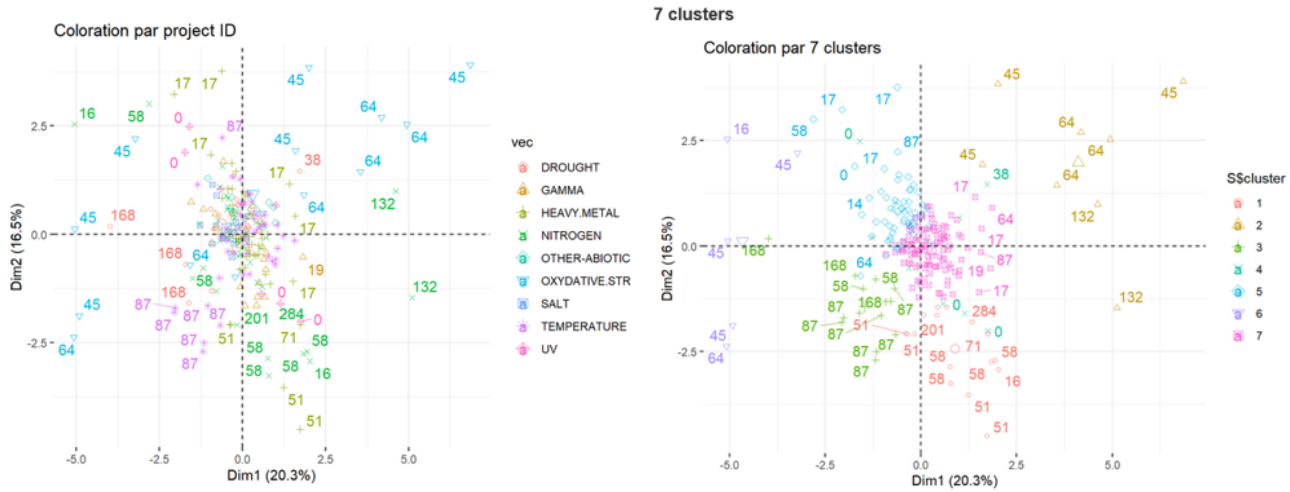
7 ANNEXES

Stress abiotiques	Drought	Gamma	Heavy metal	Nitrogen	Other abiotic	Oxydative stress	Salt	Temperature	UV
Nb gènes	17	25	45	46	8	16	15	45	7
Stress biotiques	Biotrophic bacteria	Fungi	Necrotrophic bacteria	Nematode	Oomycete	Other biotic	Rhodococcus	Stifenia	Virus
Nb gènes	40	21	26	10	14	6	7	6	33

TABLE 3 – Effectif des stress GEM2Net, pour les 9 stress abiotiques (*n=224*) et pour les 9 stress biotiques (*n=163*) lors de l'extraction des données (11 mars 2021).



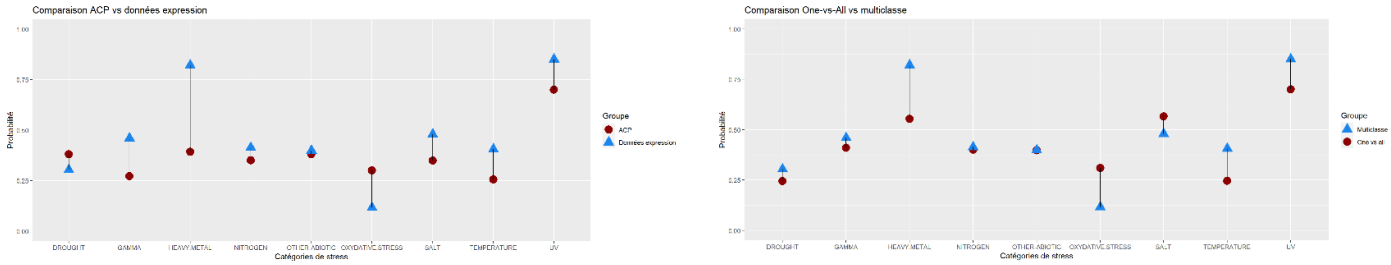
Annexe 1 – **ACP sur les données d'expression pour les stress abiotiques de plusieurs termes GO SLIM**, (A) GO SLIM biotique (420 gènes), (B) GO SLIM abiotique (661 gènes), (C) GO SLIM photosynthèse (75 gènes), qui donne le pourcentage de variance expliquée le plus élevé, et (D) un ensemble de 50 gènes choisis aléatoirement.



A. Coloration par stress

B. Résultats kmeans

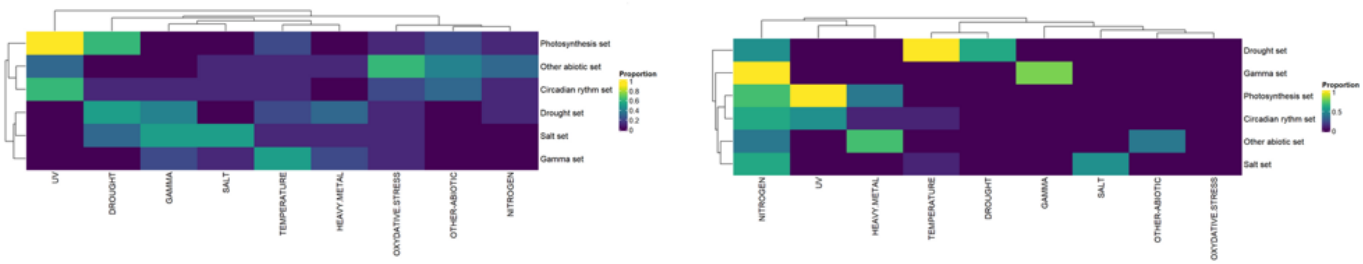
Annexe 2 – **Etude de l'effet du projet**, (A) ACP de l'ensemble rythme circadien (p=57 gènes) pour les stress abiotiques seulement avec l'identifiant du projet en nom des points et coloration par stress. (B) ACP avec les mêmes données mais coloration en fonction de la classe attribuée par K-means à 7 clusters. Les noms des points sont les identifiants de projets.



A. ACP vs données d'expression

B. One-vs-all vs multiclasse

Annexe 3 – **Comparaison des stratégies LDA**, les valeurs tracées sur ces graphiques sont les probabilités par la LDA. (A) Comparaison entre les résultats de la LDA réalisée sur les données d'expression brute (bleu) ou les composantes principales (rouge) expliquant 90% de la variance. (B) Comparaison entre les résultats de la LDA avec la stratégie multiclasse (bleu) ou "one-versus-all" (rouge).



A. Distance centroïdes

B. Distance maximale

Annexe 4 – **Qualité des prédictions PLS-DA pour deux types de calculs de distance**. Heatmap montrant les confiances de prédiction des stress abiotiques pour l'ensemble des ensembles de gènes ($p < 150$) pour (A) la distance des centroïdes, (B) la distance maximale. Le package utilisé pour cette visualisation est complexHeatmap [14] et la palette viridis.

Pathway	LV index	AUC	p-value	FDR	Nb genes	Database
Calvin-Benson-Bassham cycle	1	0.837	1.49e-03	4.77e-03	27	AraCyc
Photosynthesis	1	0.766	5.13e-03	1.24e-02	33	KEGG
Photosynthesis.1	1	0.870	9.00e-09	1.49e-07	75	GOSLIM
Ribosome biogenesis in eukaryotes	3	0.789	2.48e-05	1.54e-04	68	KEGG
Ribosome	3	0.958	4.25e-28	4.21e-26	193	KEGG
RNA transport	3	0.813	2.76e-09	5.47e-08	115	KEGG
RNA polymerase	3	0.918	1.37e-04	7.51e-04	27	KEGG
Aerobic respiration III (alternative oxidase pathway)	4	0.760	3.41e-03	9.12e-03	36	AraCyc
Oxidative phosphorylation	4	0.806	3.36e-06	2.77e-05	75	KEGG
Ribosome	4	0.917	4.29e-24	2.12e-22	193	KEGG
Plant-pathogen interaction	5	0.868	1.47e-14	4.84e-13	139	KEGG
Glucosinolate biosynthesis	6	1.000	5.25e-04	2.17e-03	17	KEGG
Proteasome	12	0.956	5.48e-07	6.78e-06	40	KEGG
Proteasome	18	0.870	3.74e-05	2.18e-04	40	KEGG
Proteasome	21	0.929	1.30e-06	1.28e-05	40	KEGG
Plant-pathogen interaction	34	0.782	1.07e-08	1.51e-07	139	KEGG
Oxidative phosphorylation	36	0.844	7.57e-07	8.32e-06	75	KEGG
mRNA surveillance pathway	43	0.762	2.49e-05	1.54e-04	85	KEGG
Phenylalanine, tyrosine and tryptophan biosynthesis	47	0.838	9.50e-04	3.48e-03	33	KEGG
Sulfur metabolism	47	0.781	9.38e-03	2.06e-02	25	KEGG
Glucosinolate biosynthesis	47	0.859	9.86e-03	2.12e-02	17	KEGG
Homologous recombination	52	0.752	3.76e-03	9.80e-03	40	KEGG

TABLE 4 – Summary PLIER de l'ensemble des pathways associés aux LV significatives, avec les paramètres statistiques liés ainsi que la base de données d'origines et le nombre de gènes impliqués.

RESUME

Il existe un fort intérêt agronomique à sélectionner des plantes non-modèles résistantes aux stress divers et potentiellement multiples, en particulier dans un contexte de changement climatique. Pour ce faire, il est essentiel de mieux comprendre leur réponse transcriptomique aux stress. Cependant, les données disponibles pour ces plantes sont bien plus limitées que celles des plantes modèles comme *Arabidopsis thaliana*. Notamment, la base de données GEM2Net repertorie la réponse transcriptomique d'Arabidopsis aux stress biotique et abiotique dans plusieurs centaines d'expériences. C'est sur cet ensemble de données que nos analyses ont été réalisées au cours de ce stage. L'enjeu ici est donc de tester la mise en place d'un transfert de connaissances depuis ces données. Après un découpage en plusieurs ensembles de gènes par utilisation des GO SLIM et la réalisation d'analyses différentielles, nous avons pu poursuivre nos analyses en réalisant des classifications non supervisées (ACP) et supervisées (BCA, LDA, PLS-DA). Ces études ont révélé que les catégories de stress définies dans GEM2Net ne sont pas totalement homogènes et présentent un recoupement important. Seul un petit sous-ensemble de 3 catégories stress se démarquait suffisamment pour avoir une prédiction satisfaisante par la LDA. Une deuxième démarche, moins centrée sur les catégories de stress, consistait à appliquer une déconvolution non supervisée sur les données avec incorporation de connaissances biologiques (PLIER). Le modèle obtenu met en avant 13 variables latentes significativement associées à des pathways, porteuses d'une information biologique. Le signal identifié a pu être qualitativement validé dans un jeu de données transcriptomiques externe chez Arabidopsis.

Mots-clés : Arabidopsis, données transcriptomiques, réponse au stress, projections multivariées, transfer learning

Mot-clés des acquis : R, Rmarkdown, classification, déconvolution, visualisations