

# Rapport de Stage

Master II SSD - Biostatistique

Effectué à l'Institut des Sciences des Plantes de Montpellier  
Dans l'équipe SIRENE

## INFERENCE STATISTIQUE DES RESEAUX DE REGULATION DE GENES CHEZ LA MICRO TOMATE EN REPONSE AU $CO_2$ ELEVE.



*Réalisé par :*  
**Paul BURGAT**

*Enseignant référent :*  
**Élodie BRUNEL-PICCININI**

*Tuteurs de stage :*  
**Antoine MARTIN  
Sophie LEBRE  
Océane CASSAN**

Année Universitaire 2021 - 2022  
Université de Montpellier

# Résumé

La régulation de l'expression des gènes est considérée comme un des principaux leviers de la réponse des plantes à leur environnement. Afin d'identifier les mécanismes régulateurs qui influent sur l'expression des gènes, l'inférence de réseaux de régulation est aujourd'hui considérée comme une des approches les plus efficaces.

Plusieurs modèles statistiques existent pour reconstruire les réseaux de régulation de gènes. Cependant, la comparaison de ces modèles et de leur pertinence reste encore peu étudiée. A partir de données transcriptomiques générées sous  $CO_2$  élevé, les objectifs de mon stage étaient d'inférer des réseaux de régulation de gènes chez la micro tomate à l'aide de deux méthodes statistiques, les Random Forests et le LASSO, puis de comparer ces approches et les réseaux de régulation obtenus, et enfin d'identifier des gènes qui ont une grande importance dans ces réseaux. La comparaison des deux modèles en terme de qualité de prédiction nous a permis de constater que le LASSO obtient de meilleures performances, tout en étant cependant moins interprétable que les Random Forests. Nous avons identifié 7 gènes régulateurs appartenant aux gènes les plus importants des deux réseaux inférés, dont le gène *MYB15*, aussi identifié comme un gène important du réseau inféré par mon équipe d'accueil avec la méthode des Random Forests sur la plante *Arabidopsis* dans un contexte de réponse au  $CO_2$  élevé.

## Remerciements

Tout d'abord, je tiens à remercier Mme.**Brunel-Piccinini**, mon encadrante académique, pour nous avoir informé, moi et mes camarades de Master, des différentes offres de stage ainsi que des différentes dates pour des salons étudiants-chercheurs durant notre recherche de stage, cela m'a permis de prendre contact avec Mr.**Martin**, chef d'équipe et chercheur de l'équipe SIRENE au laboratoire IPSIM, et **Océane Cassan**, doctorante en biostatistique dans la même équipe, et par la suite d'être pris en tant que stagiaire dans cette même équipe.

Ensuite, je remercie tous les membres de l'équipe SIRENE pour leur accueil des plus chaleureux, l'ambiance de travail était très agréable, sérieuse et dynamique et leur expertise en biologie m'a permis de bien comprendre le contexte dans lequel j'effectuais mes travaux. De plus, ils m'ont permis d'effectuer des manipulations biologiques tout au long de mon stage, ce qui était très intéressant car j'ai pu voir comment étaient générées les données que j'ai traitées. Je remercie aussi Mme.**Cleynen**, chercheuse en Statistique à l'*Institut Montpelliérain Alexander Grothendieck* (IMAG), ainsi que **Valérian Sobolak**, un camarade de Master, pour leur collaboration sur une partie de mon travail.

Je souhaite aussi exprimer une grande reconnaissance envers mes tuteurs Mr.**Martin**, **Océane Cassan** et Mme.**Lèbre**, chercheuse en statistique à l'IMAG, pour leurs conseils avisés et leur expertise dans les domaines de la biologie végétale et des statistiques respectivement, ainsi que leur aide précieuse à la relecture et à la correction de ce rapport.

Et enfin, je souhaite remercier particulièrement **Océane Cassan** qui m'a suivi tout le long du stage et qui m'a aidé à mener à bien mes différentes missions ainsi qu'à trouver des solutions aux problèmes rencontrés, le tout avec bienveillance et pédagogie.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Le $CO_2$ élevé, une menace pour la qualité des plantes cultivées . . . .	4
1.2	Introduction aux réseaux de régulation de l'expression des gènes . . .	6
1.2.1	Régulation et expression des gènes . . . . .	6
1.2.2	Construction d'un réseau de régulation de l'expression des gènes	7
1.3	Modélisation et inférence statistique pour les réseaux . . . . .	8
1.4	Problématiques . . . . .	9
<b>2</b>	<b>Méthodologie</b>	<b>9</b>
2.1	Bio informatique . . . . .	9
2.1.1	Définition des données RNA-seq . . . . .	9
2.1.2	Acquisition des données . . . . .	10
2.2	Statistique . . . . .	11
2.2.1	Régression linéaire multiple avec interactions . . . . .	11
2.2.2	Random Forests . . . . .	12
2.2.3	LASSO . . . . .	15
2.3	Biostatistique . . . . .	17
2.3.1	Normalisation des données avec la méthode TMM . . . . .	17
2.3.2	Analyse de l'expression différentielle de gènes . . . . .	17
2.3.3	Analyse d'enrichissement de l'ontologie des gènes . . . . .	22
2.3.4	Inférence de réseaux de régulation de gènes . . . . .	24
<b>3</b>	<b>Résultats</b>	<b>26</b>
3.1	Présentation des jeux de données . . . . .	26
3.1.1	Données phénotypiques . . . . .	26
3.1.2	Données transcriptomiques . . . . .	28
3.2	Impact du fort $CO_2$ sur les caractéristiques phénotypiques de la micro tomate . . . . .	30
3.2.1	Modèles linéaires avec interactions appliqués aux données phé- notypiques . . . . .	30
3.2.2	Analyse des résultats sur les différentes caractéristiques phé- notypiques de la micro tomate . . . . .	30
3.3	Analyse des données transcriptomiques . . . . .	33
3.3.1	Normalisation et filtration des données . . . . .	33
3.3.2	Analyse en composantes principales des données . . . . .	34
3.3.3	Analyse de l'expression différentielle des gènes . . . . .	35
3.3.4	Enrichissement ontologique des gènes : GO terms . . . . .	36
3.4	Inférence de réseaux . . . . .	37
3.4.1	Inférence de réseaux par Random Forests . . . . .	37
3.4.2	Inférence de réseaux par LASSO . . . . .	41
3.4.3	Comparaison des 2 méthodes . . . . .	42
<b>4</b>	<b>Conclusions et perspectives</b>	<b>50</b>
<b>5</b>	<b>Annexe</b>	<b>53</b>

# 1 Introduction

## 1.1 Le $CO_2$ élevé, une menace pour la qualité des plantes cultivées

L'équipe *Sirène* dans laquelle j'ai réalisé mon stage à l'*Institut des Sciences des Plantes de Montpellier (IPSIM)* s'intéresse aux effets du  $CO_2$  sur la biologie végétale dus à l'augmentation de sa concentration dans l'atmosphère.

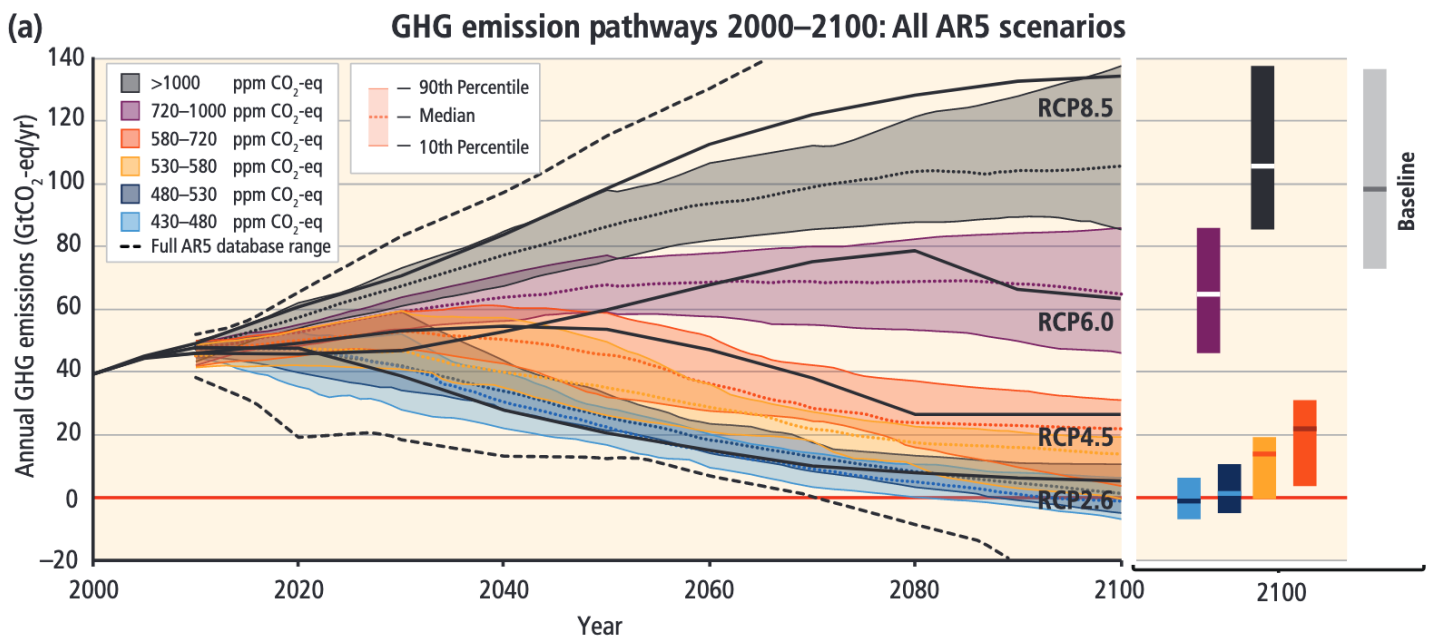


FIGURE 1 – Evolution des concentrations en  $CO_2$  (ppm=parties par million) dans l'atmosphère de nos jours jusqu'en 2100 selon plusieurs scénarios. GHG correspond au gaz à effet de serre dont le  $CO_2$  fait partie. L'axe des ordonnées représente nos émissions par an, tandis que les couleurs des courbes représentent des scénarios, caractérisés par la concentration totale de  $CO_2$  dans l'atmosphère en 2100. Sur l'axe des ordonnées, un "Annual GHG emissions" de 0 correspondrait à aucune émission de gaz à effet de serre sur l'année. [15] (Source figure : [IPCC\\_AR5\\_REPORT](#), 2014)

Il est montré sur ce graphique, datant de 2014, 6 scénarios possibles d'émission de  $CO_2$  dans l'air de nos jours jusqu'en 2100. On peut remarquer que 430ppm se situe dans la limite inférieure des scénarios et nous pourrions, dans le scénario évalué comme assez probable et dans la continuité de la tendance actuelle qui est le scénario "baseline", dépasser les 1000ppm en 2100.

Il a été montré dans la littérature mais aussi dans mon équipe d'accueil que des conditions de fort  $CO_2$  ont des effets sur la physiologie des plantes.

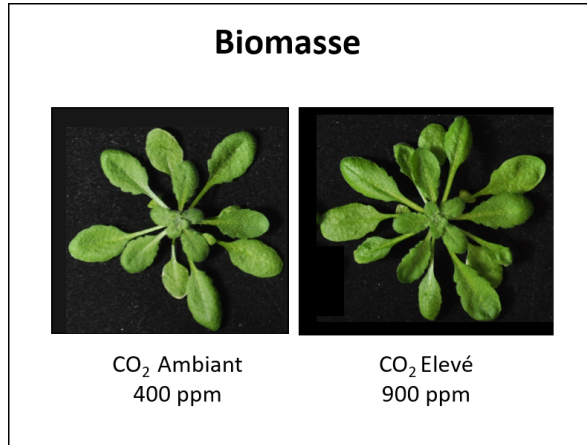


FIGURE 2 – Biomasse d'*Arabidopsis thaliana* dans 2 conditions de  $CO_2$  atmosphérique

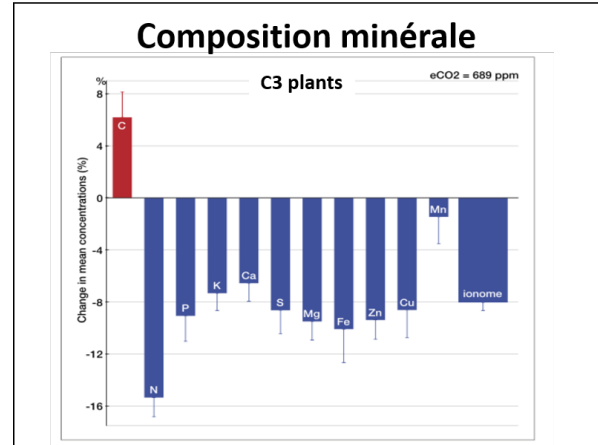


FIGURE 3 – Teneur en minéraux des plantes C3 dans des conditions de fort  $CO_2$ . ([18])

Sur la FIGURE 2 réalisée par mon équipe d'accueil, la biomasse d'une plante modèle *Arabidopsis thaliana* a été testée dans des conditions de  $CO_2$  ambiant et  $CO_2$  élevé. On voit que la biomasse de la plante est plus grande en  $CO_2$  élevé qu'ambiant (on remarque notamment un plus grand nombre de feuilles).

Cet effet peut à priori être considéré comme positif pour l'agriculture, mais on voit sur la FIGURE 3 que ce stress environnemental en fort  $CO_2$  a des effets très négatifs comme la diminution de la teneur en minéraux et une perte de la qualité nutritionnelle des plantes. Ces effets négatifs pourraient poser problème d'un point de vue des carences alimentaires, surtout pour les pays en voie de développement.

Cette diminution de la teneur en minéraux est observée chez les plantes à photosynthèse dites C3 (qui représente la majorité des plantes, par exemple le riz, le blé, la tomate ...) et n'est pas encore expliquée, mais on suspecte que des régulations transcriptomiques sont associées à cette diminution. Dans mon équipe d'accueil, une étude a déjà été faite sur les racines de la plante modèle *Arabidopsis thaliana*, ce qui a permis de générer des hypothèses pour expliquer cette diminution, comme des régulations négatives (ou inhibitions) des systèmes de prélèvement des nutriments. Cependant une étude d'une plante d'intérêt agronomique est pertinente, d'une part pour comparer cette étude avec celle de la plante *Arabidopsis*, et d'autre part pour des raisons d'alimentation déjà mentionnées dans le paragraphe précédent. D'où l'intérêt de mon stage sur la *micro tomate* [10].

La micro tomate est une variété cultivée de tomate naine miniature, destinée à l'origine au jardinage domestique. Elle partage avec *Arabidopsis thaliana* certaines

caractéristiques qui en font un bon système modèle pour la biologie, comme sa petite taille, son cycle de vie court et son petit génome totalement séquencé. C'est donc sur cette variété cultivée de tomate que nous allons réaliser des analyses. Parmi ces analyses, nous allons inférer des réseaux de régulation de gènes dont nous allons introduire le principe, nous allons aussi parler de l'état de l'art dans ce domaine.

## 1.2 Introduction aux réseaux de régulation de l'expression des gènes

### 1.2.1 Régulation et expression des gènes

Chaque cellule d'une plante contient la même information génétique, le même ensemble de gènes (un gène est une unité fonctionnelle de l'ADN qui contient des informations de base pour le développement des caractéristiques d'un individu). Pourtant, différents ensembles de gènes sont nécessaires pour les diverses fonctions des différentes cellules ou tissus, ainsi que pour les réponses des plantes aux stimuli ou stress environnementaux. Pour ce faire, l'activité des gènes est régulée en fonction des exigences physiologiques d'un type de cellule, d'un stade de développement ou de conditions environnementales particulières. Cette régulation de l'activité est connue sous le nom d'expression génétique. Le terme "expression" de gène peut être utilisé de différentes manières qui prêtent parfois à confusion, mais dans ce rapport, si un produit génique (ARN ou protéine) est produit, le gène est considéré comme "exprimé". En fonction du nombre de molécules d'ARN transcrites à partir du gène, le gène sera plus ou moins fortement exprimé. Il peut arriver, lors de circonstances particulières, que des gènes ne soient pas exprimés en permanence. Lorsqu'un gène présente différents niveaux d'expression dans différentes circonstances, on parle d'expression différentielle. Les circonstances qui peuvent s'appliquer incluent, mais ne sont pas limitées à, différents tissus végétaux (racine vs feuille), différents stades de développement (germination vs développement reproductif), ou en réponse à différents stimuli environnementaux, ce dernier étant mon objet d'étude.

La première étape de l'expression génétique est la transcription au cours de laquelle un segment particulier d'ADN est "copié" en ARNm (ARN messenger) par une enzyme appelée ARN polymérase en collaboration avec d'autres protéines connues sous le nom de facteurs de transcription. C'est l'interaction de ces facteurs de transcription avec des séquences d'ADN spécifiques qui régule le processus de transcription des gènes.

Une fois la première étape de l'expression génétique (la transcription) effectuée, une étape phare de l'expression génétique est réalisée qui est la traduction de l'ARN messenger en protéine. Les protéines vont opérer tout un ensemble de fonction dans la cellule, une de ces fonctions est le rôle de facteur de transcription pour permettre l'activation ou la répression de la transcription d'autres gènes. Voici la [FIGURE 4](#) qui illustre mes propos :

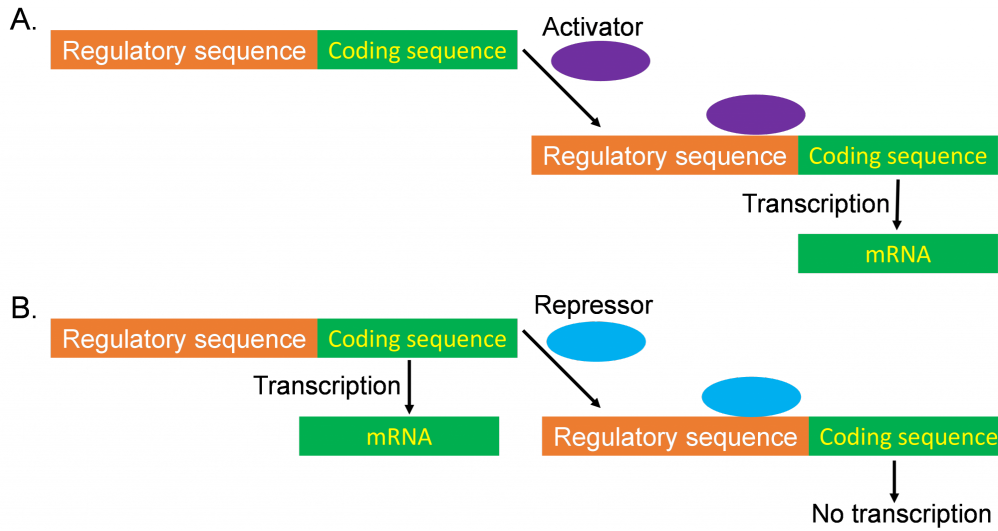


FIGURE 4 – La régulation positive de la transcription des gènes (A) se produit lorsqu’un facteur de transcription (activateur) active la transcription des gènes. La régulation négative de la transcription des gènes (B) se produit lorsqu’un facteur de transcription (répresseur) arrête la transcription des gènes. Ce graphique provient de la publication [gene-expression-and-regulation](#)

### 1.2.2 Construction d’un réseau de régulation de l’expression des gènes

Regardons maintenant comment construire, à partir de la régulation de l’expression des gènes, un réseau de connexions entre gènes régulateurs (facteurs de transcription) et gènes cibles (gènes dont la transcription est activée ou réprimée par ces gènes régulateurs) à l’aide de la FIGURE 5 ci-dessous.

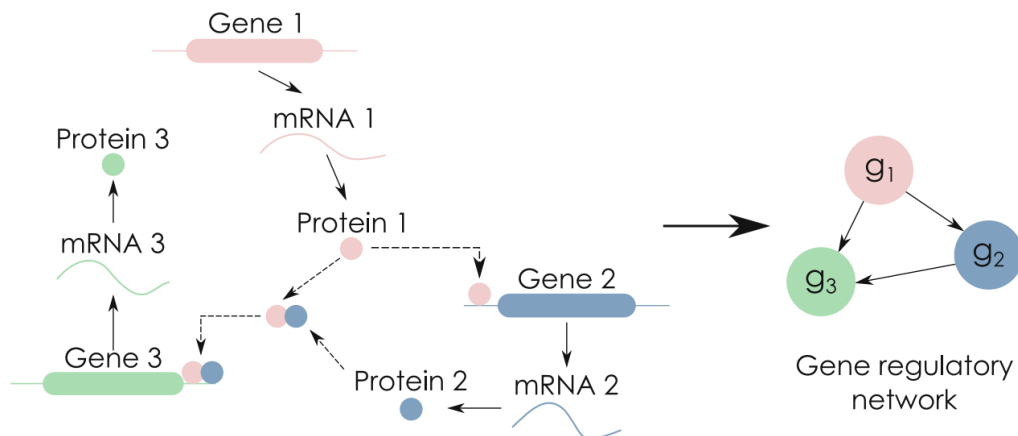


FIGURE 5 – Représentation d’un réseau de régulation de 3 gènes. [25]



Sur la partie gauche de cette FIGURE 5, le gène 1 est traduit en protéine 1, qui va activer la transcription du gène 2 en ARNm 2 qui va ensuite être traduit en protéine 2, et ces 2 protéines vont activer la transcription du gène 3 en ARNm 3 qui va ensuite être traduit en protéine 3. Sur la partie droite de la figure, ce mécanisme est représenté dans un réseau de 3 gènes avec des flèches orientées qui partent du gène régulateur en direction du gène cible. On a donc une flèche qui part de  $g_1$  vers  $g_2$  car  $g_1$  régule la transcription de  $g_2$ , ainsi qu’une flèche orientée de  $g_2$  vers  $g_3$  et une flèche orientée de  $g_1$  vers  $g_3$ .

### 1.3 Modélisation et inférence statistique pour les réseaux

L’inférence des réseaux de régulation de l’expression de gènes est confrontée au fléau de la dimension. En effet, le nombre d’échantillons biologiques disponibles pour réaliser de tels réseaux est généralement bien inférieur au nombre de variables prédictives qui représentent les dimensions d’un modèle. Pour palier à ce problème de dimensions, plusieurs domaines sont applicables notamment l’apprentissage automatique. Il existe une grande variété d’outils disponibles pour l’inférence de réseaux. Beaucoup d’entre eux ont été comparés les uns aux autres à l’occasion des challenges DREAM [20]. Ces défis visent à comparer les méthodes d’inférence de réseaux les plus récentes sur des données biologiques simulées et validées. Il existe deux grandes catégories de réseaux : non-orientées (co-expression) ou orientée (régulation). Ces défis fournissent des mesures de performance pour 27 méthodes basées sur des techniques de régression, des mesures d’information mutuelle, la corrélation ou le cadre bayésien, entre autres méthodes. Regardons quelques exemples de méthodes appartenant aux différentes catégories de réseaux :

- Réseaux non-orientés : méthode d’information mutuelle : ARACNE [21] ; méthode de corrélations partielles : WCGNA [17].
- Réseaux orientés : méthodes de régression : TIGRESS [12] (Least Angle Regression) et GENIE3 [14] (Forêts aléatoires).

Les mesures de performance recueillies par DREAM (c’est-à-dire l’aire sous les courbes de précision et de rappel ou les scores globaux), ainsi que les efforts plus récents pour comparer les nouvelles méthodes à ces normes (c’est-à-dire les mesures F, les courbes ROC) sont des ressources utiles pour aider à faire un choix.

Dans ce rapport, nous utilisons deux méthodes qui ont été parmi les plus performantes des challenges DREAM (cf Figure 2a [20]) pour modéliser des réseaux de régulation de gènes :

- La méthode du package DIANE [9] qui s’appuie sur le package GENIE3 [14], une procédure d’apprentissage automatique qui utilise les Random Forests [6] qui est une méthode basée sur l’inférence d’une collection d’arbres de régression.

- La méthode LASSO [27], une régression linéaire pénalisée qui permet la sélection de variables (cf section 2.2.3).

## 1.4 Problématiques

Nous répondons dans ce rapport à 2 problématiques qui sont :

- Identifier des gènes régulateurs chez la micro tomate qui contrôlent la réponse des transcriptomes racinaires sous  $CO_2$  élevé et qui permettraient potentiellement d'expliquer les effets négatifs du  $CO_2$  sur la nutrition minérale de cette plante.
- Réaliser une étude comparative de 2 méthodes statistiques pour inférer des réseaux de régulation de gènes : les Random Forests et le LASSO.

# 2 Méthodologie

## 2.1 Bio informatique

### 2.1.1 Définition des données RNA-seq

Les données transcriptomiques traitées dans ce rapport sont des données dites "RNA-seq", leur analyse a un grand intérêt biologique car elle a pour but de démontrer que certains gènes ont un certain nombre de transcrits dans une condition mais un nombre bien différent dans une autre par exemple, de plus, il s'agit d'une technologie à haut débit qui permet de mesurer l'expression de tous les gènes du génome avec précision (ce qui n'était pas le cas des technologies ayant précédé le RNA-Seq comme les puces à ARN). Définissons dans un premier temps ce type de données.

Pour obtenir des données RNA-seq, on récolte les ARN de nombreuses cellules à la fois, on les coupe en fragments qu'on amplifie par PCR (méthode qui permet de dupliquer en grand nombre, avec un facteur de multiplication de l'ordre du milliard, une séquence d'ADN ou d'ARN connue) et auxquels on ajoute des adaptateurs, qui sont des courtes séquences de nucléotides permettant le séquençage et puis on utilise un séquenceur qui, à partir de ces fragments amplifiés, nous retourne de petits bouts de séquence génomique qu'on appelle des "lectures" (communément appelées "reads"), sous forme de texte avec des A, T, G, C. Nous obtenons des dizaines de millions de ces "reads". Le nombre de reads est représentatif de l'abondance des ARN correspondants dans la cellule, le but étant d'estimer cette abondance. Les fichiers transmis par le séquenceur sont au format FASTQ qui est un format de fichier texte permettant de stocker à la fois des séquences biologiques (uniquement des séquences nucléiques) et les scores de qualité associés. Voici un exemple type de la présentation d'un fichier FASTQ :

```
@HWI-ST865 :166 :D0C4KACXX :2 :1101 :1241 :1970 1 :N :0 :
CCAGCGACACTTGCAGCTTAGGGGCAAGAGGCTCCCACAACACCCTGTGCGA
+
GFFFIGIIIFGEHHIJJIIGGGHIIID=BFG ?EDECC@FGCHC ?BCCBB)53(;;B;?8299?
```

Chaque bloc de 4 lignes, commençant par un "@", représente un "read" : nous avons en première ligne l'identifiant du "read", en deuxième ligne la séquence du fragment d'ARN de longueur fixe correspondant puis la qualité (fiabilité) du séquençage de ce morceau. Le séquenceur produit des reads courts, d'environ 30 à 150 nucléotides.

### 2.1.2 Acquisition des données

Comment passer de ces données RNA-seq (au format FASTQ) aux données transcriptomiques (cf FIGURE 10) que l'on traite dans la partie 3.3. Pour ce faire, toute une procédure est mise en oeuvre.

#### La préparation :

- Copier les données à un emplacement qui a beaucoup d'espace disponible et le plus possible de mémoire vive, l'accès à un serveur de stockage et à un cluster de calcul est idéal.
- Se procurer la séquence génomique de l'espèce concernée (pour nous la micro tomate) en format fasta (.fa), par exemple depuis [NCBI](#). Elle sert à différents programmes pour construire des "index" qui servent au mapping (voir ci-dessous).

#### Le Contrôle de Qualité :

Il faut s'assurer tout d'abord que le séquenceur a bien fait son travail. Les premiers nucléotides d'un read sont séquencés avec beaucoup de fiabilité, mais plus on avance dans la séquence, moins c'est précis. On utilise le programme [fastp](#) qui permet dans un premier temps de retirer les adaptateurs ajoutés sur les fragments d'ARN pour le séquençage mais aussi de savoir s'il faut rétrécir les reads, et de quelle longueur, ce qu'on appelle le "trimming". Cette étape est très importante puisque des reads de mauvaise qualité s'aligneront (voir paragraphe "mapping" plus bas) difficilement sur le génome de l'espèce concernée (pour nous la micro tomate), et rendront toutes les analyses ultérieures inutiles.


#### L'alignement ("mapping") :

A priori, on ne sait pas de quelle partie du génome provient un read, mais uniquement sa séquence. On va donc "aligner" chaque read obtenu après le Contrôle de Qualité ("QC") sur le génome de la micro tomate, c'est-à-dire rechercher dans le génome la position d'une sous-séquence similaire à celle du read. Celle-ci devrait en théorie être unique pour un read suffisamment long (>30 nucléotides), et donc on obtient la position d'origine du transcrit d'où il vient. On imagine le génome

comme une grande ligne droite, longue de quelques milliards de caractères, et les reads comme de petits segments venant s'aligner le long de cette ligne, là où la séquence est similaire.

Le nombre moyen de reads par position mappée s'appelle "profondeur de séquençage". C'est pour garantir une profondeur suffisante qu'on doit séquençer des centaines de millions de reads, et s'assurer de leur qualité, car c'est ce qui donne la puissance statistique au moment de déterminer le taux d'expression d'un transcrit. Cette étape d'alignement des reads se fait avec [STAR](#). Les fichiers de sortie sont des fichiers BAM qui donnent les résultats de l'alignement pour chaque séquence, dont le nom de la séquence de référence, la position de l'alignement sur la référence, la séquence du read, la qualité de l'alignement etc.

### Le comptage :

La dernière étape est de compter combien de reads ont été alignés dans nos régions d'intérêt du génome : les transcrits d'ARN messenger pour les gènes codants. Pour cela, on utilise [htseq-count](#). Les étapes précédentes demandant de grandes ressources en terme de calcul, elles sont réalisées sur serveur adapté. Après le comptage, les analyses peuvent être faites sur un poste de travail standard. On récupère donc ces comptages en local qu'on exporte ensuite sur  pour obtenir le tableau [10](#), où à chaque gène correspond maintenant une variable quantitative représentant son expression dans une condition donnée.

## 2.2 Statistique

### 2.2.1 Régression linéaire multiple avec interactions

Lors de nos analyses des données phénotypiques partie [3.2](#), nous utilisons des modèles linéaires à plusieurs variables explicatives, qui sont donc des régressions linéaires multiples. Les objectifs de ce modèle sont :

- Estimer et interpréter les paramètres d'une régression linéaire incluant plusieurs variables catégorielles et/ou numériques.
- Expliquer la signification d'une interaction entre deux variables et interpréter son coefficient.

Le modèle de régression linéaire multiple représente la relation entre une variable réponse  $y$  et  $m$  prédicteurs  $x_1, x_2, \dots, x_m$ .

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \epsilon = \beta_0 + \sum_{i=1}^m \beta_i x_i + \epsilon \quad (1)$$

avec  $\beta_0$  l'intercept du modèle,  $\beta_1, \beta_2, \dots, \beta_m$  les coefficients des variables prédictives associées et  $\epsilon$  un bruit gaussien centré de variance  $\sigma$  qui représente l'erreur du modèle.

Comme dans le cas de la régression linéaire simple, les coefficients  $\beta_1, \dots, \beta_m$  peuvent être calculés à partir de la méthode des moindres carrés.

### Modèle avec interactions :

Le modèle précédent suppose que les effets des variables prédictives sur la variable réponse sont additifs. Pour considérer la possibilité que l'effet d'un prédicteur sur la réponse dépende de la valeur d'un autre prédicteur, nous devons spécifier une interaction entre ces deux prédicteurs.

Soit le modèle définit par une variable réponse  $y$  et 2 variables prédictives  $x_1$  et  $x_2$  qui sont catégorielles binaires ( $x_i \in \{0, 1\}$ ), l'interaction entre les variables  $x_1$  et  $x_2$  est définie dans l'expression mathématique du modèle comme suit :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (2)$$

L'interaction est donc équivalente à l'ajout d'un nouveau prédicteur au modèle, égal au produit des deux variables qui interagissent. Séparons le modèle en deux équations selon la valeur prise par  $x_1$  :

Si  $x_1 = 0$  :

$$y = \beta_0 + \beta_2 x_2 + \epsilon$$

Si  $x_1 = 1$  :

$$y = \beta_0 + \beta_1 + (\beta_2 + \beta_{12}) x_2 + \epsilon$$

Pour ce modèle avec interactions, l'interprétation des coefficients change un peu car le coefficient devant  $x_2$  change en fonction de la valeur prise par  $x_1$  ainsi que l'intercept. Le modèle avec interactions est donc équivalent à estimer séparément la droite de régression (ordonnée à l'origine et pente) pour chacune des valeurs de  $x_1$ .

### 2.2.2 Random Forests

Nous verrons plus tard dans ce rapport (cf partie 2.3.4) qu'on utilise 2 méthodes de régression pour inférer des réseaux de régulation de gènes. La première est la méthode des Random Forests [6] qui utilise une agrégation d'arbres de régression. Cette méthode utilise donc un modèle de régression non-linéaire et non-paramétrique. Ainsi l'avantage de cette méthode par rapport à une régression linéaire paramétrique est d'une part qu'elle peut modéliser des non linéarités dans l'influence d'une variable prédictive sur une variable réponse et d'autre part, aucune contrainte n'est imposée a priori sur la distribution des données à modéliser. Définissons notre modèle à partir de cette méthode.

Soient  $Y_i$  la variable réponse dans l'échantillon  $i$  (cette variable réponse correspond pour nous au nombre de comptages de l'expression d'un gène cible dans un échantillon  $i$ ) et  $X_i = (X_{i1}, \dots, X_{im})$   $m$  variables prédictives (qui correspond pour nous aux nombres de comptages de l'expression de  $m$  gènes régulateurs dans le

même échantillon i). Dorénavant, nous employons le terme de niveau d'expression d'un gène pour parler du nombre de comptages de l'expression de ce gène. La variable  $Y_i$  est modélisée par une combinaison non linéaire des  $X_i$  à l'aide d'arbres de régression :

$$Y_i = \text{RandomForest}(X_i) + \epsilon_i \quad (3)$$

avec  $\epsilon_i$  l'erreur du modèle. Il n'existe pas de formule mathématique explicite pour le modèle d'une random forest.

Un arbre de régression est construit en choisissant des conditions et des seuils sur les variables prédictives.

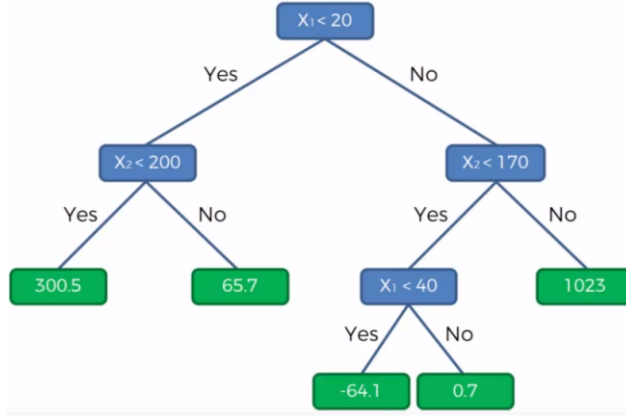


FIGURE 6 – Arbre de régression

Les feuilles de l'arbre (les rectangles verts sur la FIGURE 6) sont les prédictions du modèle pour la variable réponse.

Soit le  $k^{ieme}$  noeud de notre arbre. Le coût d'un noeud, qu'on cherche à minimiser pour obtenir la meilleure condition et la meilleure variable, a pour formule :

$$J(k) = \frac{n_{gauche}}{n_k} MSE_{gauche} + \frac{n_{droite}}{n_k} MSE_{droite} \quad (4)$$

avec  $n_{gauche}$  le nombre d'observations dans le noeud enfant de gauche provenant du noeud k,  $n_{droite}$  le nombre d'observations dans le noeud enfant de droite,  $n_k$  le nombre d'observations dans le noeud k, et  $MSE_{gauche}$  et  $MSE_{droite}$  correspondant aux puretés des noeuds enfants de gauche et de droite.

La pureté d'un noeud a pour formule :

$$MSE_{node} = \frac{1}{n_{node}} \sum_{i=1}^{n_{node}} (\bar{y}_{node} - y^{(i)})^2 \quad (5)$$

avec  $\bar{y}_{node}$  la moyenne des variables réponses appartenant au noeud,  $y^{(i)}$  une des

Ajustement d'un arbre de régression :

- Choisir la variable et la condition sur cette variable qui permettent au mieux de diminuer le coût d'un noeud.
- Répéter en créant des nouvelles branches, jusqu'à épuisement des variables, ou atteinte de la profondeur d'arbre maximale.

variables réponses parmi celles qui appartiennent au noeud.

On peut constater que les arbres de régression sont faciles à entraîner, faciles à utiliser et interprétables. Néanmoins on peut leur reprocher de ne pas être assez précis ni assez généralisables. En effet, l'arbre est performant avec les données d'entraînement ou d'apprentissage (données utilisées pour créer l'arbre) mais il n'est pas assez robuste pour donner de bonnes performances de prédiction sur des données de test (données non utilisées pour créer l'arbre). Or, le but d'un algorithme de machine learning est de pouvoir prédire des données dont on ne connaît pas encore la prédiction, c'est à dire des données autres que celles d'apprentissage.

Pour palier à ce problème de généralisation, les Random Forests ont été proposées par Léo Breiman en 2001 [6]. Cet algorithme permet d'utiliser plusieurs arbres de régression afin de créer une forêt et d'améliorer la généralisation de l'ensemble du modèle. On emprunte la théorie de l'ensemble learning afin de faire fonctionner ces arbres ensemble. Les Random Forests combinent donc la simplicité des arbres de régression afin de gagner de la précision et de la généralisation. Ce qui donne des modèles de bien meilleure qualité. Décrivons les principales étapes de cet algorithme.

1. **Le bagging (bootstrap aggregation) :** On crée aléatoirement un jeu de données à partir du jeu de données de base sur lequel nous allons effectuer un arbre de régression. Un échantillon bootstrap peut être obtenu en tirant aléatoirement  $n$  observations avec remise dans le jeu de données d'apprentissage, chaque observation ayant une probabilité  $1/n$  d'être tirée.
2. **Entraînement de l'arbre de régression sur le jeu de données créé par bagging :** Lors de l'apprentissage de chaque noeud des arbres de régression, on ne peut choisir que parmi un sous ensemble des variables prédictives échantillonnées aléatoirement, ici de taille la racine du nombre de variables prédictives.
3. **Répéter les étapes 1 et 2 :** On répète les 2 étapes pour créer autant d'arbres que l'on souhaite. On a bien une forêt d'arbres créés à partir de jeux de données créés aléatoirement, d'où le nom de forêts aléatoires (Random Forests).

Comme chaque jeu de données pour chaque arbre sera échantillonné aléatoirement du jeu de données de base, on maximise la variété des arbres, ce qui va permettre aux arbres forts dans un domaine de compenser les arbres plus faibles dans d'autres domaines, ce qui fait la force de l'ensemble learning. L'utilisation du bagging fait qu'on a laissé une partie des données de côté pour l'entraînement (apprentissage) de chaque arbre. Ces données non-utilisées sont appelées les données out of bag (OOB).

Une fois notre forêt aléatoire construite, nous pouvons maintenant évaluer le pouvoir prédictif de celle-ci avec le critère du MSE (Mean Squared Error) sur l'out of bag :

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (6)$$

avec  $n$  qui est égal au nombre d'échantillons,  $y_i$  le niveau d'expression observé d'un gène cible dans l'échantillon  $i$  et  $\hat{y}_i$  est la moyenne des prédictions out-of-bag du niveau d'expression du même gène cible, c'est à dire des prédictions obtenues par les arbres construits sans l'échantillon  $i$ .

Introduisons maintenant la notion de poids d'influence des variables prédictives (niveaux d'expression des gènes régulateurs) sur les variables réponses (niveaux d'expression des gènes cibles) que nous utilisons dans la partie 3.4.1. Une fois la régression (non linéaire et non paramétrique) réalisée par la méthode des Random Forests sur le niveau d'expression d'un gène cible (variable réponse), chaque niveau d'expression de gènes régulateurs (variables prédictives) a un pouvoir prédictif, ou une influence, plus ou moins important sur la variable réponse. On définit le  $MSE_{gene_R}$  qui représente le MSE (voir équation (6)) du modèle en ayant randomisé à travers les échantillons les valeurs de l'expression du gène régulateur dont nous voulons connaître l'influence. Nous allons évaluer cette importance grâce à un poids d'influence défini par la formule suivante, qui représente l'augmentation relative de MSE induite par la randomisation d'une variable prédictive, exprimée en %.

$$gene_{influence} = \frac{MSE_{gene_R} - MSE}{MSE} \times 100 \quad (7)$$

le  $MSE$  correspond au  $MSE$  défini par l'équation (6), c'est à dire les erreurs de prédiction de notre modèle sans toucher les valeurs des variables prédictives. Nous multiplions par 100 pour obtenir un nombre, et plus un pourcentage, qui correspond au poids d'influence du gène régulateur sur le gène cible

Plus la valeur de  $gene_{influence}$  sera grande, plus le niveau d'expression du gène régulateur observé aura un pouvoir prédictif important sur le niveau d'expression du gène cible.

### 2.2.3 LASSO

La seconde méthode que nous allons utiliser pour inférer des réseaux de régulation de gènes est une régression linéaire pénalisée, la régression LASSO développée par Robert Tibshirani en 1996 [27]. Nous appliquons cette régression car l'estimation classique d'un modèle linéaire se fait dans les cas où le nombre d'observations (pour nous les échantillons) est plus grand que le nombre de variables prédictives (pour nous les gènes régulateurs retenus comme différentiellement exprimés), ce qui, nous le verrons dans la partie 3.4, n'est pas notre cas. La régression LASSO est pratiquée dans un contexte où il y a un très grand nombre de variables explicatives  $p$  (les gènes régulateurs) et où le nombre d'observations  $n$  est largement inférieur à  $p$ . Les objectifs de cette méthode sont :

1. Trouver un estimateur biaisé ayant un bon pouvoir prédictif
2. Faire de la sélection de variables



Notre but est de sélectionner les gènes régulateurs les plus importants pour chaque gène cible grâce à cette sélection de variables.

Tout d'abord, on suppose que  $y_i$ , le niveau d'expression d'un gène cible dans un échantillon  $i$ , suit une loi de Poisson (car ce sont des données de comptage) de paramètre  $\mu_i$  qui est l'espérance théorique de cette loi. La théorie des modèles linéaires généralisés nous dit que ce paramètre de la loi du niveau d'expression du gène cible est modélisé par :

$$\log(\mu_i) = \beta_0 + \beta_{i,1}X_{i,1} + \dots + \beta_{i,p}X_{i,p} \quad (8)$$

avec  $\beta_0$  l'intercept du modèle,  $X_i = (X_{i,1}, \dots, X_{i,p})$  le vecteur des variables explicatives (les gènes régulateurs) associées à l'échantillon  $i \in \{1, \dots, n\}$  ( $n$  = le nombre d'échantillons dans mes données),  $\beta_i = (\beta_{i,1}, \dots, \beta_{i,p})$  les coefficients des variables à estimer avec  $p$  le nombre de variables explicatives au total.

Dans le cadre d'un modèle linéaire standard, le coefficient  $\hat{\beta}_i$  est obtenu en minimisant la somme des carrés des résidus. Avec la méthode LASSO,  $\hat{\beta}_\lambda^L$  est aussi obtenu en minimisant cette somme mais sous une contrainte supplémentaire :

$$\min_{\beta} \sum_{i=1}^n (\log(y_i) - X_i \beta^T)^2 \quad \text{sous contrainte} \quad \sum_{j=1}^p |\beta_j| \leq \gamma \quad (9)$$

avec  $\beta = (\beta_1, \dots, \beta_p)$ ,  $i$  allant de 1 à  $n$  le nombre d'échantillons,  $j$  allant de 1 à  $p$  le nombre de gènes régulateurs et  $\gamma$  le paramètre qui contrôle le niveau de régularisation des coefficients estimés.

En écrivant ce problème sous la forme Lagrangienne, on obtient :

$$\min_{\beta} \sum_{i=1}^n (\log(y_i) - X_i \beta^T)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (10)$$

pour un certain  $\lambda > 0$  qui dépend de  $\gamma$ .

L'estimateur LASSO est tel que  $\hat{\beta}_\lambda^L \in \arg \min_{\beta} \{ \sum_{i=1}^n (\log(y_i) - X_i \beta^T)^2 + \lambda \sum_{j=1}^p |\beta_j| \}$ .

La contrainte va contracter la valeur des coefficients et la forme de la pénalité  $L_1$  va permettre à certains coefficients de valoir exactement zéro. Voici la FIGURE 7 en 2 dimensions qui illustre mes propos :

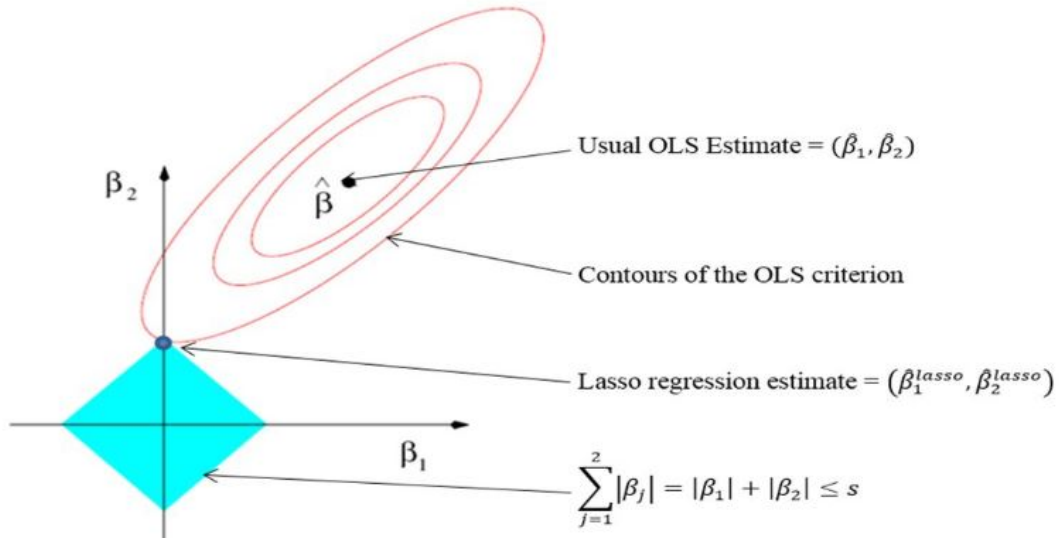


FIGURE 7 – Illustration d’une régression LASSO en 2 dimensions. Sur cette figure est représenté un graphe en 2 dimensions, qui sont les 2 coefficients à estimer. Lorsque qu’on réalise un modèle linéaire classique, le meilleur estimateur  $\hat{\beta}$  est l’estimateur des moindres carrés (OLS en Anglais). Lors d’une régression LASSO, on veut que la somme des valeurs absolues des coefficients soit inférieure à un certain  $s > 0$ , du fait de la pénalité  $L_1$ , ce sous-ensemble a la forme d’un losange représenté en bleu sur la figure. En projetant l’estimateur des moindres carrés sur ce sous-ensemble, on obtient l’estimateur LASSO. De par l’irrégularité de la pénalité, des coefficients sont mis à 0, c’est le cas sur la figure de  $\beta_1$ , la régression LASSO sélectionne donc dans cet exemple seulement la variable dont le coefficient est non nul, donc la variable dont le coefficient est  $\beta_2$ .

## 2.3 Biostatistique

### 2.3.1 Normalisation des données avec la méthode TMM

Après l’acquisition des données décrites dans la partie 2.1.2, on obtient le tableau 10 qui représente les comptages de l’expression des gènes de la micro tomate dans 24 échantillons. Les données récupérées sont cependant brutes, soumises à des biais de profondeur de séquençage. La normalisation des données est essentielle pour les analyses que nous allons effectuer comme l’analyse en composantes principales (ACP, cf partie 3.3.2) ou encore l’analyse de l’expression différentielle des gènes (cf partie 3.3.3) car on veut rendre les comptages comparables entre les différents échantillons dans un premier temps, mais cette normalisation permet aussi de corriger les biais techniques de la récolte de données. Pour se faire, on utilise la méthode de normalisation TMM (Trimmed Mean of M-values) décrite dans l’article suivant [23] .

### 2.3.2 Analyse de l’expression différentielle de gènes

Pour réaliser une analyse de l’expression différentielle de gènes, on utilise la méthode appliquée avec le papier edgeR [22] qui est très utilisée mais qui repose sur un

certain nombre d'hypothèses que nous avons souhaité analyser en détail ci-dessous. Nous rappelons ici que nous avons 8 conditions expérimentales avec 3 réplicats par condition, ce qui fait 24 échantillons.

**Ajustement du modèle linéaire généralisé :** Soit  $n$  le nombre d'échantillons d'ARN, on considère  $\pi_{gi}$  la fraction de tous les fragments d'ARN de l'échantillon  $i$  provenant du gène  $g$ . Posons  $G$  le nombre total de gènes, alors  $\sum_{g=1}^G \pi_{gi} = 1$ . Soit  $\sqrt{\phi_g}$  le coefficient de variation CV (écart-type divisé par la moyenne) de  $\pi_{gi}$  entre les réplicats  $i$ . Soit  $N_i$  le nombre total de reads dans l'échantillon  $i$ . Ainsi, on peut considérer que  $y_{gi}$ , le nombre de reads correspondant au gène  $g$  de l'échantillon  $i$ , suit une loi de Poisson tel que :

$$y_{gi} | \pi_{gi} \sim \mathcal{P}(\mu_{gi}) \text{ avec } \mathbb{E}(y_{gi} | \pi_{gi}) = \mu_{gi} = N_i \pi_{gi}$$

On suppose également que  $\pi_{gi} \sim \Gamma(\gamma, m)$  avec  $\gamma$  représentant la dispersion. On peut alors montrer que  $y_{gi}$ , par les lois mélange Gamma-Poisson, suit une loi négative binomiale de paramètres  $p = \frac{\gamma}{\mu + \gamma}$  et  $\gamma$  dont la démonstration a été vérifiée en collaboration avec Alice Cleynen et Valérian Sobolak et laissée en annexe 5.

La vraisemblance du modèle est donc :

$$\mathbf{V}(y, p, \gamma) = \frac{\Gamma(y + \gamma)}{\Gamma(\gamma)y!} (1 - p)^y p^\gamma = \exp\{y \log(1 - p) + \gamma \log p\} \nu(y) \quad (11)$$

où  $\nu(y) = \frac{\Gamma(y + \gamma)}{\Gamma(\gamma)y!}$ .

Ce modèle appartient à la famille exponentielle avec les paramètres suivant :

- $\theta = \log(1 - p) \Leftrightarrow p = 1 - e^\theta$
- $b(\theta) = -\gamma \log(p) = -\gamma \log(1 - e^\theta)$

Et on sait que :  $\mathbb{E}(y) = b'(\theta) = \gamma \frac{e^\theta}{1 - e^\theta} = \gamma \frac{1 - p}{p}$

La théorie GLM (pour Modèles Linéaires Généralisés) peut être utilisée pour ajuster un modèle log-linéaire :

$$\begin{aligned} \log \mu &= \mathbf{x}^T \beta + \log N \\ \Leftrightarrow \mu &= e^{\mathbf{x}^T \beta} \times N \end{aligned} \quad (12)$$

On écrit maintenant la log-vraisemblance du modèle afin de retrouver les coefficients  $\beta_j$  du vecteur  $\beta$  :

$$\mathbf{LV}(y, \gamma, \mu) = y \log \left( \frac{\mu}{\mu + \gamma} \right) + \gamma \log \left( \frac{\gamma}{\gamma + \mu} \right) + \log(\nu(dy)) \quad (13)$$

$$\begin{aligned}
&= y \log \left( \frac{e^{x^T \beta} N}{e^{x^T \beta} N + \gamma} \right) + \gamma \log \left( \frac{\gamma}{\gamma + e^{x^T \beta} N} \right) + \log(\nu(dy)) \\
&= y(\log(e^{x^T \beta} N) - \log(e^{x^T \beta} N + \gamma)) + \gamma(\log(\gamma) - \log(e^{x^T \beta} N + \gamma)) + \log(\nu(dy))
\end{aligned}$$

On dérive cette log-vraisemblance partiellement par  $\beta_j$  :

$$\begin{aligned}
\frac{\partial \mathbf{LV}}{\partial \beta_j} &= y \left( x_j - \frac{x_j e^{x^T \beta} N}{e^{x^T \beta} N + \gamma} \right) - \gamma \frac{x_j e^{x^T \beta} N}{\gamma + e^{x^T \beta} N} \\
&= x_j \left( y - y \frac{\mu}{\mu + \gamma} - \frac{\gamma \mu}{\gamma + \mu} \right) \\
&= x_j \left( \frac{y\gamma - \gamma\mu}{\mu + \gamma} \right) = x_j \left( \frac{y - \mu}{\frac{\mu}{\gamma} + 1} \right)
\end{aligned} \tag{14}$$

Et on retrouve bien l'équation de l'article edgeR [22] avec  $\gamma = \phi^{-1}$

On obtient le modèle final :

- $y|\pi \sim \mathcal{P}(\mu)$  avec  $\mu = N\pi$
- $\pi \sim \Gamma(\phi^{-1}, \frac{\mu\phi}{N})$ , la démonstration de  $m = \frac{\mu\phi}{N}$  se trouve aussi dans l'annexe 5.
- $y \sim \mathbf{NB}(\mu, \phi^{-1})$  en écrivant la négative binomiale avec son paramètre d'espérance  $\mu \left( \mathbb{E}(y) = \gamma \frac{1-p}{p} = \mu \right)$

Ainsi la dérivée de la log-vraisemblance, par rapport aux coefficients  $\beta_g$  peut s'écrire  $X^T z_g$  où  $X$  est la matrice de design de colonne  $\mathbf{x}_i$  et  $z_{gi} = (y_{gi} - \mu_{gi})/(1 + \phi_g \mu_{gi})$ . La matrice d'information de Fisher, pour les coefficients, peut s'écrire  $\mathcal{I}_g = X^T W_g X$  avec  $W_g$  est la matrice diagonale des poids provenant de la théorie GLM.

Avant de pouvoir estimer la dispersion  $\phi_g$ , on doit d'abord estimer les coefficients  $\hat{\beta}_g$ . N'ayant pas d'expression explicite, on peut estimer ces coefficients par l'intermédiaire de l'algorithme de Newton-Raphson. Pour éviter que les  $\beta_g$  ne dépendent de  $\phi_g$ , on approxime l'information de Fisher en considérant que  $\mu_{gi} = \mu_g$ .

Introduisons maintenant une nouvelle définition, celle de l'**APL : Adjusted Profile Likelihood**. L'APL de  $\phi_g$  est la log-vraisemblance pénalisée tel que :

$$APL_g(\phi_g) = \ell(\phi_g, \mathbf{y}_g, \hat{\beta}_g) - \frac{1}{2} \log \det \mathcal{I}_g \tag{15}$$

avec :

- $\mathbf{y}_g$  le vecteur des reads du gène  $g$ ,
- $\hat{\beta}_g$  le vecteur des coefficients estimés,

- $\det \mathcal{I}_g$  le déterminant de la matrice d'information de Fisher calculable grâce à la décomposition de Cholesky.

### Estimation de la dispersion $\phi_g$ :

Il existe différentes manières d'estimer la dispersion suivant le postulat de départ.

**Dispersion commune :** L'estimation de la dispersion pour chaque gène individuellement ne devrait pas être considérée sauf si on dispose d'un nombre conséquent de réplicats. Le mieux est donc d'obtenir le partage d'informations entre les gènes. La méthode la plus simple, pour avoir ce partage d'informations, est de supposer que tous les gènes ont la même dispersion ( $\phi_g = \phi$ ).

La dispersion commune peut être estimée en maximisant la fonction de vraisemblance partagée :

$$APL_s(\phi) = \frac{1}{G} \sum_{g=1}^G APL_g(\phi) \quad (16)$$

Cette maximisation peut se faire numériquement grâce à l'algorithme de Newton approximatif sans dérivation [1].

**Dispersion tendancielle :** Une généralisation de la dispersion commune consiste à modéliser  $\phi_g$  comme une fonction lisse du nombre moyen de reads pour chaque gène. EdgeR propose plusieurs méthodes pour réaliser ça dont la méthode APL pondérée localement : les  $\phi_g$  sont estimés en faisant une log-vraisemblance partagée, cela consiste à faire une moyenne pondérée des APL pour le gène  $g$  et ses gènes voisins par le nombre moyen de reads.

**Dispersions génétiques :** Dans les applications scientifiques réelles, il est plus probable que chaque gène dispose d'une dispersion différente qui dépend de nombreux facteurs (séquence génomique/taille du génome/niveau d'expression ...). On cherche donc un compromis entre une dispersion commune et une dispersion individuelle pour chaque gène  $\phi_g$  en utilisant l'approche empirique de Bayes à vraisemblance pondérée, proposée par Robinson et Smyth [24].

Dans cette approche,  $\phi_g$  est estimé en maximisant :

$$APL_g(\phi_g) + G_0 APL_{Sg}(\phi_g) \quad (17)$$

avec  $G_0$  le poids donné à la vraisemblance partagée et  $APL_{Sg}$  la log-vraisemblance locale partagée.

Le choix optimal de  $G_0$  dépend de la variabilité de la dispersion entre les gènes. Ici les auteurs ont choisi de prendre  $G_0 = 20/df$  car ils ont obtenu de bons résultats en essayant une grande variété de jeux de données, où  $df$  est le degré de liberté pour l'estimation de la dispersion. Pour les expériences multigroupes,  $df$  est le nombre d'échantillons moins le nombre de groupes de traitement distincts, ce qui est notre cas avec 24 échantillons pour 8 groupes distincts, ce qui nous donne  $df = 16$ .

### Tests sur les gènes différentiellement exprimés :

Une fois les estimations de dispersions obtenues et les modèles linéaires généralisés binomiaux négatifs ajustés, nous pouvons passer aux procédures de test pour déterminer l'expression différentielle en utilisant soit le test de quasi-vraisemblance (QL pour Quasi-Likelihood) [19], soit le test du rapport de vraisemblance.

Bien que le test du rapport de vraisemblance soit un choix plus évident pour les inférences avec les GLM, le test QL est préféré car il reflète l'incertitude de l'estimation de la dispersion pour chaque gène. Il permet un contrôle du taux d'erreur plus robuste et plus fiable lorsque le nombre de réplicats est faible.

### La méthode QL :

L'ajustement d'un modèle de quasi-vraisemblance nécessite de spécifier la variance des valeurs observées, jusqu'à une constante de proportionnalité, en fonction des moyennes modélisées. On reprend notre modèle  $y_{gi}|\pi_{gi} \sim \mathcal{P}(\mu_{gi})$  d'espérance  $\mathbb{E}(y_{gi}|\pi_{gi}) = \mu_{gi} = N_i\pi_{gi}$ .

On a montré que si  $\pi_{gi} \sim \Gamma(\phi_g^{-1}, \frac{\mu_{gi}\phi_g}{N_i})$ , alors  $y_{gi} \sim \mathbf{NB}(\mu_{gi}, \phi_g^{-1})$  d'espérance  $\mu_{gi}$  et de variance :

$$\mathbb{V}(y_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2 \quad (18)$$

Le modèle quasi-nétagif binomial introduit une nouvelle dispersion  $\Phi_g$  utilisée pour modéliser la variance des observations du gène  $g$ , c'est à dire :

Soit  $y_{gi}$  qui suit une loi quasi-négative binomiale, on a :

$$\mathbb{V}(y_{gi}) = \Phi_g(\mu_{gi} + \phi_g \mu_{gi}^2) \quad (19)$$

Pour chacun des  $g$  gènes, les paramètres des moyennes modélisées sont estimés en maximisant :

$$\sum_i \mathbf{QL}_g(\hat{\mu}_{gi}|y_{gi}) \quad (20)$$

avec  $y_g = (y_{1g}, \dots, y_{Ig})'$  est le vecteur des observations du gène  $g$  à travers les échantillons,  $\mu_g = (\mu_{1g}, \dots, \mu_{Ig})'$  est le vecteur des moyennes correspondantes et  $\mathbf{QL}_g(\hat{\mu}|y)$  est la fonction de quasi-vraisemblance correspondant à la fonction de variance choisie pour le gène  $g$ .

La réalisation d'un test d'hypothèse pour l'expression différentielle en utilisant l'approche de la quasi-vraisemblance implique le calcul d'une statistique de test du rapport de quasi-vraisemblance et l'estimation du paramètre de dispersion  $\Phi_g$  (la constante de proportionnalité de la relation moyenne-variance spécifiée).

Le test du ratio de quasi-vraisemblance est :

$$LRT_g = 2 \left( \sum_i \mathbf{QL}_g(\hat{\mu}_{gi}|y_{gi}) - \sum_i \mathbf{QL}_g(\tilde{\mu}_{gi}|y_{gi}) \right) \quad (21)$$

avec  $\tilde{\mu}_{gi}$  et  $\hat{\mu}_{gi}$  sont les estimations du maximum de quasi-vraisemblance pour  $\mu_{gi}$  sous les hypothèses nulle et alternative respectivement. Les hypothèses étant :

- L'hypothèse nulle  $\mathcal{H}_0$  où on suppose que, pour un  $g$  donné, les  $\mu_{gi}$  sont égaux :  $\mu_{g1} = \mu_{g2} = \dots = \mu_{gI}$
- L'hypothèse alternative  $\mathcal{H}_1$  où l'on suppose, pour un  $g$  donné, qu'il existe au moins un  $\mu_{gi}$  différent des autres pour  $i$  allant de 1 à  $I$ .

Le paramètre de dispersion,  $\Phi_g$ , peut être estimé par :

$$\hat{\Phi}_g = \frac{2 \left( \sum_i \mathbf{QL}_g(y_{gi}|y_{gi}) - \sum_i \mathbf{QL}_g(\hat{\mu}_{gi}|y_{gi}) \right)}{n - p} \quad (22)$$

avec  $p$  la dimension de l'espace de paramètres des moyennes complet.  $\mathbf{QL}_g(y_{gi}|y_{gi})$  est la fonction de quasi-vraisemblance du modèle saturé, c'est à dire le modèle dans lequel il existe autant de paramètres que d'observations.

On utilise le QL-test pour déterminer si la variabilité entre les moyennes de groupe est plus grande que la variabilité des observations à l'intérieur des groupes. Si ce rapport est suffisamment élevé, on peut conclure que les moyennes ne sont pas égales, et donc que le gène  $g$  est différentiellement exprimé entre les 2 groupes.

On définit le QL-test comme :

$$F_{QL} = \frac{LRT_g/q}{\hat{\Phi}_g} \quad (23)$$

Le numérateur représentant deux fois la différence entre la quasi-vraisemblance du modèle complet et du modèle nul divisé par  $q$  qui désigne la différence des dimensions entre les espaces de paramètres de moyennes complet et nul. Le dénominateur étant l'estimation de la dispersion vu précédemment. Ce test statistique est semblable au [F-test](#) que l'on peut retrouver, notamment, dans les test ANOVA.

### 2.3.3 Analyse d'enrichissement de l'ontologie des gènes

Parmi une liste de gènes différentiellement exprimés (DEG), nous aimerions savoir si des fonctions particulières (GO terms) dans les descriptions des gènes ressortent de manière récurrente pour ces gènes, c'est ce qu'on appelle faire l'ontologie

des gènes. Cette fonctionnalité est apportée par le package DIANE [9] dans mes résultats partie 3.3.4 qui s'appuie sur le package R clusterProfiler [28] et qui utilise des tests exacts de Fisher sur la distribution hypergéométrique pour savoir quel GO terms sont le plus représentés. Voici une explication plus détaillée de la méthode.

Soient "DEG" une liste de gènes différentiellement exprimés, "GO term" une fonction particulière présente dans les descriptions des gènes. On observe la table de contingence suivante :

	∈ GO term	∉ GO term	Total ligne
∈ DEG	a	b	a+b
∉ DEG	c	d	c+d
Total colonne	a+c	b+d	a+b+c+d (=n)

TABLE 1 – Exemple de table de contingence. Ici a, b, c et d sont des nombres de gènes.

Dans cet exemple, nous voulons tester si la différence entre la proportion de gènes ayant le GO term dans la liste DEG est significativement différente de la proportion de gènes ayant le GO term dans l'ensemble total de gènes. Fisher a montré que si les totaux marginaux sont fixés, la probabilité de cette configuration est donnée, sous l'hypothèse nulle d'absence d'association, par la loi hypergéométrique :

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \quad (24)$$

De manière à calculer si des données observées sont significativement éloignées de l'indépendance, c'est-à-dire la probabilité d'observer des données aussi ou plus éloignées que celles observées si l'hypothèse nulle (indépendance entre l'appartenance à la liste de DEG et avoir le GO term particulier) est satisfaite, il faut calculer la valeur de p.



### 2.3.4 Inférence de réseaux de régulation de gènes

Une des motivations principales pour inférer un réseau est de faire des prédictions sur les gènes les plus influents, c'est à dire les gènes qui régulent le plus de gènes cibles dans le réseau, sur la base d'une liste de gènes différentiellement exprimés. Nous présentons ici la méthodologie des inférences de réseaux de régulation de gènes avec les méthodes Random Forests et LASSO. Ces 2 méthodes sont retenues pour plusieurs raisons :

- Elles permettent toutes 2 la sélection de variables.
- Nous utilisons une méthode paramétrique et linéaire (LASSO) contre une autre non-paramétrique et non linéaire (Random Forests).
- Elles ont obtenu de bonnes performances dans le challenge DREAM [20] qui est un article qui compare les performances de plusieurs méthodes dans la construction de réseaux de régulation de gènes.

#### Construction d'un réseau de gènes avec la méthode des Random Forests :

Une fois les régressions effectuées sur tous les gènes cibles par la méthode des Random Forests avec les gènes régulateurs comme variables prédictives et après avoir récupéré l'influence de chaque gène régulateur sur chaque gène cible (cf partie 2.2.2), on obtient une matrice avec les gènes cibles en colonne, les gènes régulateurs en ligne ainsi que le poids d'influence du gène régulateur sur le gène cible. Plus le poids d'influence d'un régulateur sur la cible est élevé, plus l'importance de l'arête orientée de ce régulateur vers le gène cible sera grande. La construction d'un réseau de régulation de gènes s'effectue à partir de cette matrice en procédant de la manière suivante :

1. On fixe un nombre d'arêtes à retenir dans le réseau en fonction de la densité du réseau. Le nombre maximal d'arêtes possibles  $E_{max}$  du réseau est  $E_{max} = N_{regulateurs}(N_{cibles} - 1)$ . Des études telles que [13] ont montré que les valeurs typiques de la densité dans les réseaux biologiques se situent approximativement entre 0.1 et 0.001. Le nombre d'arêtes retenues par un réseau se calcule comme suit :  
Soit  $E$  le nombre d'arêtes sélectionné en fonction de la densité  $d$  choisie et  $E_{max}$  le nombre d'arêtes maximal, la densité  $d$  a pour équation  $\frac{E}{E_{max}}$ , et on peut alors trouver  $E$  :  $E = d \times E_{max} = d \times N_{regulators}(N_{genes} - 1)$ . On sélectionne donc les  $E$  arêtes avec la valeur d'importance la plus grande
2. La présence de chacune des arêtes retenues dans le réseau est ensuite testée de façon non-paramétrique, par une approche par permutation.

Ces tests statistiques non-paramétriques sont effectués à l'aide du package R `rfPermute` [3] et permettent de tester si les arêtes retenues dans le réseau sont significatives. Pour ce faire, on permute de manière aléatoire le niveau d'expression du gène cible dans les échantillons dont on dispose et on ajuste une forêt aléatoire

aux données perturbées toujours avec les mêmes régulateurs, cela permet de simuler l'hypothèse nulle du test : que l'expression des régulateurs n'a pas de lien avec celle du gène cible. On obtient un niveau d'importance pour chaque arête sur ces données perturbées. On réitère la même procédure (permutation + estimation) 1000 fois, ce qui nous permet d'estimer la distribution nulle des importances des arêtes. La proportion de valeurs d'importance générées sous l'hypothèse nulle étant supérieures à celle observée sans permutation donne la p-valeur du test.

On applique ensuite la correction FDR [5] pour les tests multiples, et seules les arêtes avec une p-valeur ajustée inférieure à un seuil de FDR choisi sont conservées pour former le réseau final.

### **Construction d'un réseau de gènes avec la méthode LASSO :**

Une fois les régressions effectuées sur tous les gènes cibles par la méthode LASSO avec les gènes régulateurs comme variables prédictives (cf partie 2.2.3), on sélectionne pour chaque gène cible les gènes régulateurs dont le coefficient estimé est non nul. On obtient donc, pour chaque gène cible, une liste de régulateurs qui régulent ce gène cible, et on peut construire un réseau de régulation de gènes comme précédemment avec des arêtes orientées qui partent du régulateur et qui vont vers le gène cible régulé par ce régulateur.

## 3 Résultats

Tous les résultats présentés dans cette partie sont reproductibles avec les codes Rmarkdown disponibles sur les dépôts github suivants :

- Pour les analyses phénotypiques de la micro tomate : [Analyse\\_phénotypique](#)
- Pour les analyses transcriptomiques de la micro tomate : [Analyse\\_transcriptomique](#)

Les liens sont disponibles dans l'annexe 5.

### 3.1 Présentation des jeux de données

Pour réaliser mes travaux, j'ai utilisé plusieurs jeux de données. Tout d'abord des données phénotypiques de la *micro tomate* pour confirmer les hypothèses d'augmentation de la biomasse et de diminution de la teneur en minéraux en fort  $CO_2$  dans le fruit et les racines de la plante, puis des données transcriptomiques pour l'inférence statistique des réseaux de régulation.

Les conditions de culture de la micro tomate sont pour le  $CO_2$  de 400ppm ou 900ppm, ce qui correspond au  $CO_2$  atmosphérique actuel et au  $CO_2$  atmosphérique qu'on pourrait atteindre dans des dizaines d'années (voir section 1.1), pour le Fer de 10 (qui correspond à un approvisionnement standard) ou de 0 (qui correspond à une carence)  $\mu M$  (micromolaire), et pour le nitrate de 10 (approvisionnement standard) ou de 0.5 (faible approvisionnement)  $mM$  (millimolaire). Les cultures ont été réalisées en hydroponie, ce qui permet un accès plus facile aux racines de la plante. Notre plan d'expérience est combinatoire et il permet d'étudier l'effet du fort  $CO_2$  et ses interactions avec des niveaux de nutriments apportés à la plante.

#### 3.1.1 Données phénotypiques

Tout d'abord, présentons 12 des 24 échantillons au total, qui correspondent à 3 réplicats pour chacune des 8 conditions, du jeu de données de la teneur en minéraux dans les fruits de la micro tomate.

Cond. No	CO2 ppm	Fe $\mu$ M	N mM	Zn 213.857	Fe 259.940	Ca 393.366	Cu 324.754	Mg 285.213	Mn 403.076	K 766.491	P 213.618
1	400	10	10.0	70.39375	126.70875	1106.8063	27.07452	6983.060	54.14904	150036.16	18443.16
1	400	10	10.0	59.78887	148.44133	1139.0811	29.89444	7623.081	60.81971	151698.80	16534.72
1	400	10	10.0	60.08714	101.19940	970.8817	24.24569	6082.505	43.22058	143913.97	18521.60
2	400	10	0.5	34.75023	89.50818	2911.6484	23.16682	7848.288	101.09159	93730.86	14468.73
2	400	10	0.5	33.45297	76.94183	2777.7116	21.18688	7316.165	83.63243	92084.88	14161.76
2	400	10	0.5	33.85411	78.62889	2839.3767	21.84136	7602.977	90.64164	95282.93	15605.65
3	400	0	10.0	58.42349	24.43164	1192.9015	27.61838	7194.587	91.35309	108115.32	18525.56
3	400	0	10.0	67.87176	18.61000	1068.4329	21.89412	6165.384	74.44000	93991.45	19244.93
3	400	0	10.0	59.57291	18.08463	1278.6899	25.53125	6768.971	78.72134	117571.38	20137.77
4	400	0	0.5	39.90214	18.24098	1026.0551	28.50153	7062.679	61.56331	80545.33	15151.41
4	400	0	0.5	43.42940	23.38506	1946.5278	30.06650	8813.940	77.95020	91346.50	15912.98
4	400	0	0.5	50.28206	31.69956	1590.4434	31.69956	10031.271	112.58809	82998.19	15871.64

FIGURE 8 – Tableau présentant la teneur en minéraux en  $\mu g/g$  de masse sèche du fruit de la micro tomate dans plusieurs conditions

On regarde la quantité de minéraux pour 8 minéraux distincts : le zinc(Zn), le fer(Fe), le calcium(Ca), le cuivre(Cu), le magnésium(Mg), le manganèse(Mn), le potassium(K) et le phosphate(P), les numéros suivant le minéral dans le tableau correspond au spectre d'émission de ce minéral. Les 12 lignes manquantes du tableau sont les mêmes conditions de fer et de nitrate que pour 400ppm de  $CO_2$  mais en 900ppm de  $CO_2$ .

Regardons maintenant le jeu de données FIGURE 9 pour la biomasse des racines et des feuilles. Nous disposons cette fois de 37 échantillons, donc 5 réplicats par condition ou un peu moins selon la qualité de la mesure effectuée expérimentalement pour récolter les données, les conditions étant les mêmes que précédemment, et les masses des racines et des feuilles données en grammes. 15 des 37 échantillons du tableau sont présentés :

No	CO2ppm	Iron_supply	N_supply	masse_racines_g	masse_feuilles_g
1	400	10	10.0	0.81	2.87
2	400	10	10.0	0.87	3.71
3	400	10	10.0	0.65	3.08
4	400	10	10.0	0.97	3.76
5	400	10	10.0	1.08	4.30
6	400	10	0.5	0.63	1.40
7	400	10	0.5	1.23	1.79
8	400	10	0.5	1.05	1.45
9	400	10	0.5	0.36	1.21
10	400	10	0.5	0.83	1.37
11	400	0	10.0	0.98	2.84
12	400	0	10.0	1.25	3.76
13	400	0	10.0	1.02	3.29
14	400	0	10.0	0.59	2.01
15	400	0	10.0	0.89	4.39

FIGURE 9 – Tableau présentant la biomasse des racines et des feuilles de la micro tomate dans plusieurs conditions

### 3.1.2 Données transcriptomiques

Regardons maintenant un aperçu du tableau de données généré grâce à l'analyse bio-informatique des données RNA-seq (cf FIGURE 10), c'est à partir de ce tableau de données que toute l'analyse transcriptomique va être faite.

	aCO2.HighN.Fesupply_1	aCO2.HighN.Fesupply_2	aCO2.HighN.Fesupply_3	aCO2.LowN.Fesupply_1	aCO2.LowN.Fesupply_2	aCO2.LowN.Fesupply_3
Sly00g0382731	0	0	0	0	0	0
Sly00g0382741	0	0	0	0	0	0
Sly00g0382751	228	302	282	281	464	371
Sly00g0382761	664	882	773	1024	1427	1302
Sly00g0382771	0	0	0	0	0	0
Sly00g0382781	0	0	0	0	0	0
Sly00g0382791	0	0	0	0	0	0
Sly00g0382801	0	1	1	0	0	0
Sly00g0382811	0	0	0	0	0	0
Sly00g0382821	0	0	0	0	0	0

FIGURE 10 – Tableau représentant le nombre de comptages de l'expression de 10 gènes de la micro tomate dans plusieurs conditions

Les lignes de ce jeu de données représentent les gènes de la micro tomate (39110 gènes) et les colonnes représentent les 8 conditions vues précédemment avec 3 réplicats par condition, ce qui fait 24 échantillons. Les notations pour les conditions environnementales sont les suivantes : "*aCO<sub>2</sub>*" correspond au *CO<sub>2</sub>* atmosphérique actuel (400ppm) et "*eCO<sub>2</sub>*" à une concentration forte de *CO<sub>2</sub>* atmosphérique (900ppm) ; "LowN" correspond à un faible approvisionnement en nitrate (0.5mM) et "HighN" à un approvisionnement en nitrate standard (10mM) ; "Fesupply" correspond à un approvisionnement standard en Fer (10μM) et "Festar" à une carence en Fer (0μM).

Les nombres dans le tableau représentent le niveau d'expression de chaque gène dans les différents échantillons. Ce nombre correspond au nombre de "reads" de ce gène (voir 2.1.1), c'est à dire le nombre de fragments d'ARN messagers qui proviennent de ce gène.

On remarque qu'il y a beaucoup de gènes qui ne sont pas exprimés ou très peu quelles que soient les conditions d'expérimentation. C'est aussi le cas sur la totalité du génome. Aussi, nous allons traiter ces faibles expressions de gène dans l'analyse transcriptomique.

## 3.2 Impact du fort $CO_2$ sur les caractéristiques phénotypiques de la micro tomate

Passons à l'analyse de nos données phénotypiques pour confirmer ou non la diminution de la teneur en minéraux du fruit de la micro tomate ainsi que l'augmentation de la biomasse de la plante au niveau des feuilles et des racines. Pour ce faire, nous utilisons des modèles linéaires multiples (cf partie 2.2.1) car ils sont adaptés aux modèles multivariés comme le notre composé de 3 variables explicatives (le  $CO_2$ , le nitrate et le fer). De plus, les modèles linéaires sont interprétables d'un point de vue biologique car ils permettent de savoir quelles variables ont un effet significatif sur la réponse.

### 3.2.1 Modèles linéaires avec interactions appliqués aux données phénotypiques

Sur le tableau de données vu précédemment (cf FIGURE 8), on regarde la quantité dans le fruit de la micro tomate de 8 minéraux dans 24 échantillons. Le but est ici de regarder l'effet du passage de  $CO_2$  ambiant à élevé ainsi que ses interactions avec les limitations en nitrate N et fer Fe sur la quantité des minéraux dans le fruit. On effectue alors un modèle linéaire pour chaque minéral pour vérifier si le fort  $CO_2$  a un effet significatif sur la quantité de celui-ci dans le fruit.


Soient  $\mathbf{X}=(X_1, X_2, X_3)$  avec  $X_1 = 0$  si on est en  $CO_2$  ambiant, 1 sinon.  $X_2 = 0$  si on est en approvisionnement abondant en nitrate, 1 sinon.  $X_3 = 0$  si on est en approvisionnement en fer, 1 sinon. Soit  $Y$  la quantité du minéral dans le fruit ou bien la biomasse des racines ou des feuilles dans les analyses suivantes.

On a donc 3 variables explicatives catégorielles  $X_1, X_2$  et  $X_3$  qui prennent les valeurs 1 ou 0, et une variable réponse quantitative. Les conditions de référence sont les conditions où  $X_i = 0$  pour  $i \in \{1, 2, 3\}$ , donc  $CO_2$  ambiant, approvisionnement en nitrate (10mM) et approvisionnement en fer (10 $\mu$ M) qui correspondent aux niveaux "contrôles", non perturbés, de ces 3 facteurs environnementaux. Pour un minéral, l'équation est de la forme :

$$Y = \beta_0 + \sum_{i=1}^3 \beta_i X_i + \sum_{i=1}^2 \sum_{j=i+1}^3 \gamma_{i,j} (X_i * X_j) + \delta (X_1 * X_2 * X_3) + \epsilon \quad (25)$$

avec  $\beta_0 \in \mathbb{R}$  l'intercept lorsqu'on est dans des conditions de référence, les  $\beta_i \in \mathbb{R}$  pour  $i \in \{1, 2, 3\}$  qui sont les coefficients des variables explicatives avec effets simples, les  $\gamma_{i,j}$  les coefficients des interactions d'ordre 2 entre variables explicatives et le  $\delta$  le coefficient de l'interaction d'ordre 3 entre ces mêmes variables. De plus  $\epsilon$  est un bruit gaussien centré de variance  $\sigma$ .

### 3.2.2 Analyse des résultats sur les différentes caractéristiques phénotypiques de la micro tomate

On réitère cette équation (25) pour les 8 minéraux et on obtient les résultats présentés en FIGURE 11 à l'aide du package `ggplot2` sur .

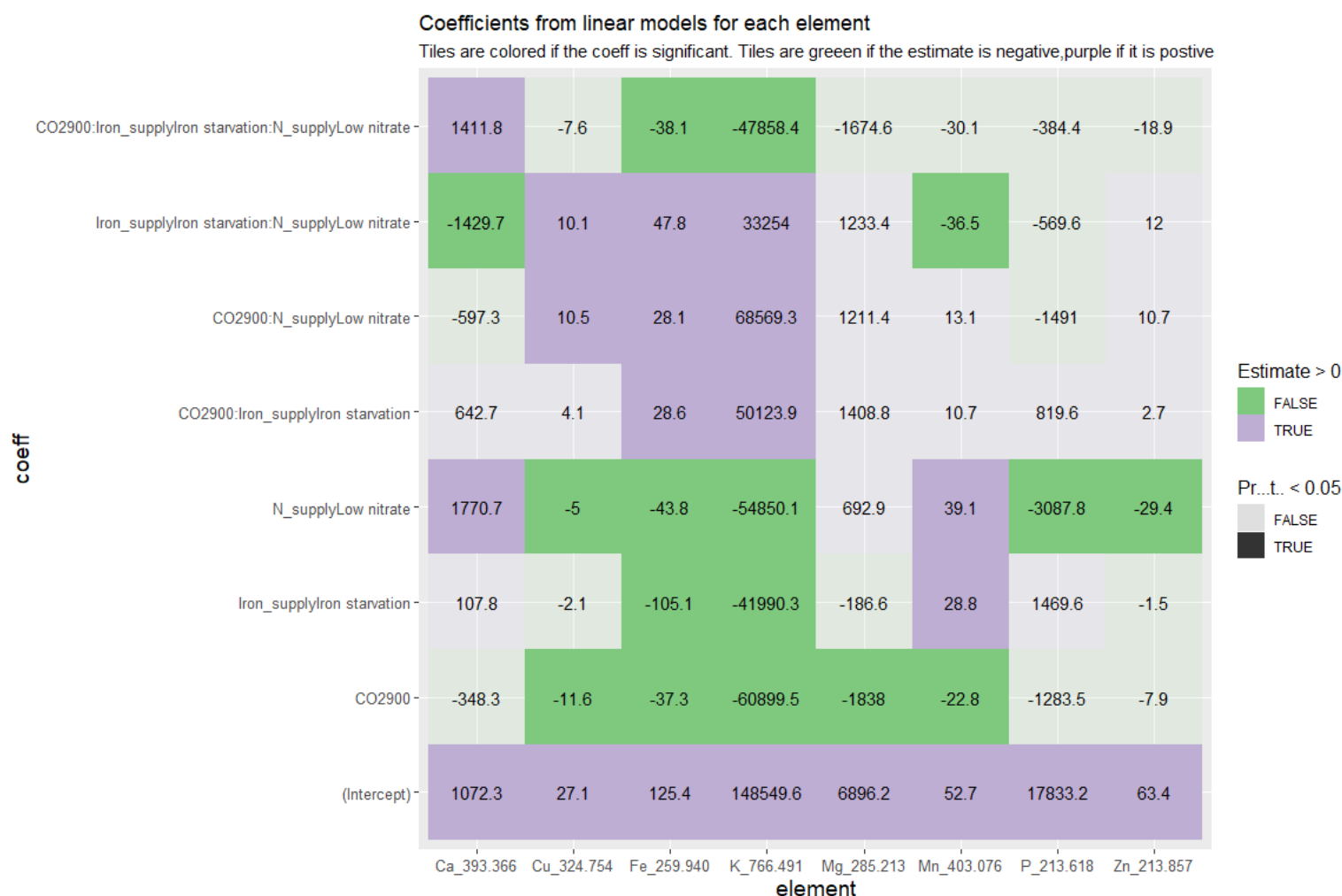


FIGURE 11 – Coefficients estimés sur la quantité des minéraux dans le fruit d'après l'équation (25). Les lignes de ce tableau correspondent aux différentes variables explicatives catégorielles et les interactions entre elles, et les colonnes aux 8 minéraux considérés, "Iron\_supplyIron starvation" signifie qu'on passe de 10  $\mu M$  de fer à 0, et "N\_supplyLow nitrate" signifie qu'on passe de 10 mM de nitrate à 0.5. Les valeurs estimées des coefficients des modèles linéaires pour chaque élément sont représentées dans ce tableau. Un coefficient coloré en vert veut dire que le coefficient est significativement différent de 0 et négatif, il est significativement différent de 0 et positif s'il est coloré en violet, et il n'est pas significatif s'il est gris.

On peut donc constater à l'aide de ces modèles linéaires effectués que lorsqu'on passe en fort  $CO_2$  (CO2900 sur la FIGURE 11, deuxième ligne en partant du bas) dans des conditions de référence pour le nitrate et le fer, la quantité de tous les minéraux baissent, dont 5 significativement. Ce qui rejoint l'hypothèse de départ sur la diminution de la teneur en minéraux dans le fruit et l'appauvrissement de la qualité nutritionnelle des plantes en fort  $CO_2$ .



On cherche maintenant à évaluer l'impact des différents stress environnementaux ainsi que leurs interactions sur la biomasse des feuilles de la micro tomate (équation (25) avec  $Y$ =biomasse des feuilles) à partir de la FIGURE 9.

Coefficients estimés sur la masse de feuilles

	masse_feuilles_g	Estimates	std.error	t.value	p.value
(Intercept)		3.544	0.2900666	12.2178846	0.0000000
CO2900		2.336	0.4736767	4.9316333	0.0000307
Iron_supplyIron starvation		-0.286	0.4102161	-0.6971935	0.4912335
N_supplyLow nitrate		-2.100	0.4102161	-5.1192531	0.0000182
CO2900:Iron_supplyIron starvation		-0.479	0.6431808	-0.7447362	0.4624242
CO2900:N_supplyLow nitrate		-1.902	0.6266154	-3.0353546	0.0050339
Iron_supplyIron starvation:N_supplyLow nitrate		0.520	0.5801332	0.8963459	0.3774487
CO2900:Iron_supplyIron starvation:N_supplyLow nitrate		0.533	0.8661617	0.6153586	0.5431173

FIGURE 12 – Coefficients estimés sur la masse des feuilles d'après l'équation (25)

Les notations sont les mêmes que pour les modèles linéaires effectués précédemment. Les coefficients surlignés en jaune sont significatifs. On remarque que l'effet du fort  $CO_2$  dans des conditions de référence pour le nitrate et le fer est significativement différent de 0 et positif avec une p-valeur très faible. Le stress environnemental en fort  $CO_2$  augmente donc la biomasse des feuilles de la micro tomate significativement. On peut aussi voir que l'effet d'un faible approvisionnement en nitrate dans des conditions de référence pour le  $CO_2$  et le fer est significativement différent de 0 et négatif, on a donc une perte de biomasse des feuilles en faible approvisionnement en nitrate. Ceci est attendu car il s'agit de plantes qui ne bénéficient pas de l'effet fertilisant d'un apport abondant en nitrate. On remarque un dernier effet significatif avec ce modèle linéaire qui est négatif, il s'agit de l'interaction entre un stress en fort  $CO_2$  et un faible approvisionnement en nitrate. Cet effet a un coefficient estimé de -1.902 qui annule pratiquement l'effet estimé du fort  $CO_2$  de 2.336. Un environnement en fort  $CO_2$  et en faible approvisionnement de nitrate augmente donc légèrement la biomasse des feuilles ( $2.336 - 1.902 = 0.434$ ) mais nettement moins que lors d'un stress en fort  $CO_2$  dans des conditions de référence. Un faible apport en nitrate annule donc l'effet stimulant du fort  $CO_2$  sur la biomasse des parties aériennes.

Regardons maintenant la biomasse des racines de la micro tomate en FIGURE 13.

De manière équivalente, on voit que le passage en fort  $CO_2$  dans des conditions d'approvisionnement en nitrate et en fer augmente la biomasse des racines significativement (coefficient estimé de 0.774) et que l'interaction entre un stress en fort

	masse_racines_g	Estimates	std.error	t.value	p.value
(Intercept)		0.876	0.1028558	8.5167800	0.0000000
CO2900		0.774	0.1679628	4.6081639	0.0000753
Iron_supplyIron starvation		0.070	0.1454600	0.4812319	0.6339630
N_supplyLow nitrate		-0.056	0.1454600	-0.3849855	0.7030573
CO2900:Iron_supplyIron starvation		-0.250	0.2280678	-1.0961650	0.2820264
CO2900:N_supplyLow nitrate		-0.730	0.2221939	-3.2854192	0.0026653
Iron_supplyIron starvation:N_supplyLow nitrate		0.006	0.2057115	0.0291671	0.9769312
CO2900:Iron_supplyIron starvation:N_supplyLow nitrate		0.224	0.3071354	0.7293199	0.4716558

FIGURE 13 – Coefficients estimés sur la masse des racines d’après l’équation (25)


$CO_2$  et un faible approvisionnement en nitrate annule pratiquement l’effet du fort  $CO_2$  (coefficient estimé de -0.730).

### 3.3 Analyse des données transcriptomiques

Comme observé dans la littérature et dans d’autres espèces végétales, les analyses des phénotypes de la micro tomate ont révélé que dans des conditions de fort  $CO_2$ , la biomasse de ses racines et de ses feuilles augmentent, mais la teneur en minéraux du fruit diminue. On suppose que cette diminution est due à des régulations transcriptomiques, d’où le choix de réaliser une analyse du transcriptome, sous le même plan d’expérience combinatoire.

#### 3.3.1 Normalisation et filtration des données

Repartons du tableau de données vu précédemment (cf FIGURE 10), nous avons donc 39110 lignes qui sont les gènes et 24 colonnes qui sont les échantillons, composés de 8 conditions avec toujours 3 réplicats par condition.

Le problème de ce tableau est que les données sont brutes, soumises à des biais de profondeur de séquençage (voir partie 2.3.1), on va donc procéder à la normalisation de nos données grâce à la fonction `normalize` du package DIANE [9] sur . Cette fonction utilise la méthode TMM (Trimmed Mean of M values) [23].

Comme cela a été fait dans l’équipe [9], on choisit aussi de retirer les gènes faiblement exprimés de l’étude car ils pourraient réduire la sensibilité des détections des gènes différentiellement exprimés dans les analyses ultérieures [26], on conserve

seulement les gènes dont la somme des lectures sur tous les échantillons (24) dépasse 240, soit une moyenne d'au moins 10 lectures du gène par échantillon.

On obtient un nouveau jeu de données composé de 22704 lignes (donc 22704 gènes conservés) avec les données normalisées à l'aide de la méthode TMM.

### 3.3.2 Analyse en composantes principales des données

On réalise une ACP sur les données normalisées et filtrées qui consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres, résumant l'information des variables initiales avec une plus faible dimension. Ces nouvelles variables sont nommées « composantes principales » ou axes principaux. On regarde ensuite quels axes principaux indépendants expliquent au mieux la variabilité des données. Dans notre cas, les variables sont les stress environnementaux.

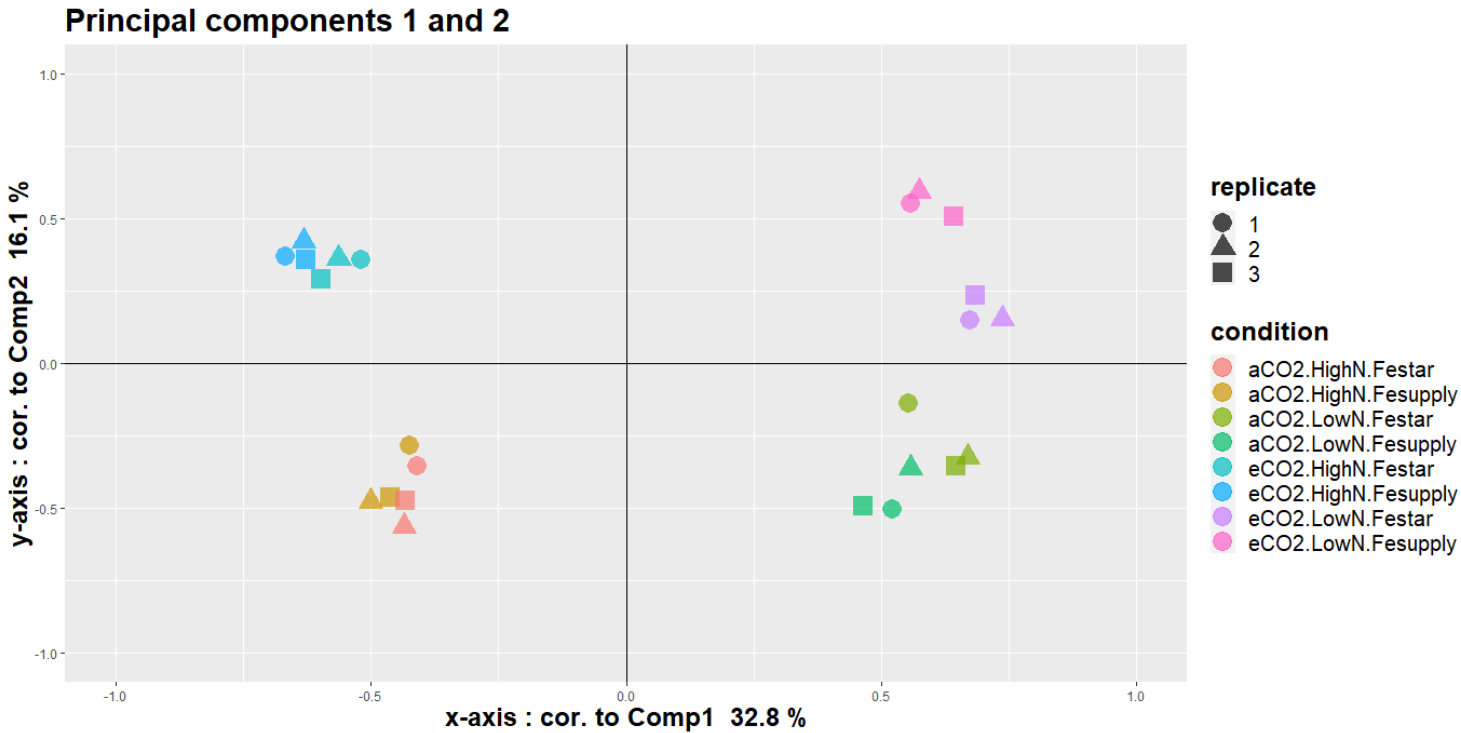


FIGURE 14 – Contribution des variables aux composantes principales 1 et 2. Les couleurs des variables représentent la condition expérimentale.

D'après la FIGURE 14, on constate que c'est l'effet du faible nitrate qui explique le plus la variabilité des données avec une variance expliquée de 32.8%. La condition de faible nitrate est notée "LowN", et de fort nitrate "HighN", on voit bien que toutes les conditions placées à gauche de l'axe des x ont comme notation "HighN" et celles à droite "LowN", d'où notre interprétation. Par la même procédure, on constate ensuite que c'est l'effet du  $CO_2$  élevé qui contribue à la deuxième composante principale avec une variance expliquée de 16.1%.

D'après la FIGURE 15, on constate que l'effet de privation de fer est la troisième composante principale avec une variance expliquée de 9.85%.

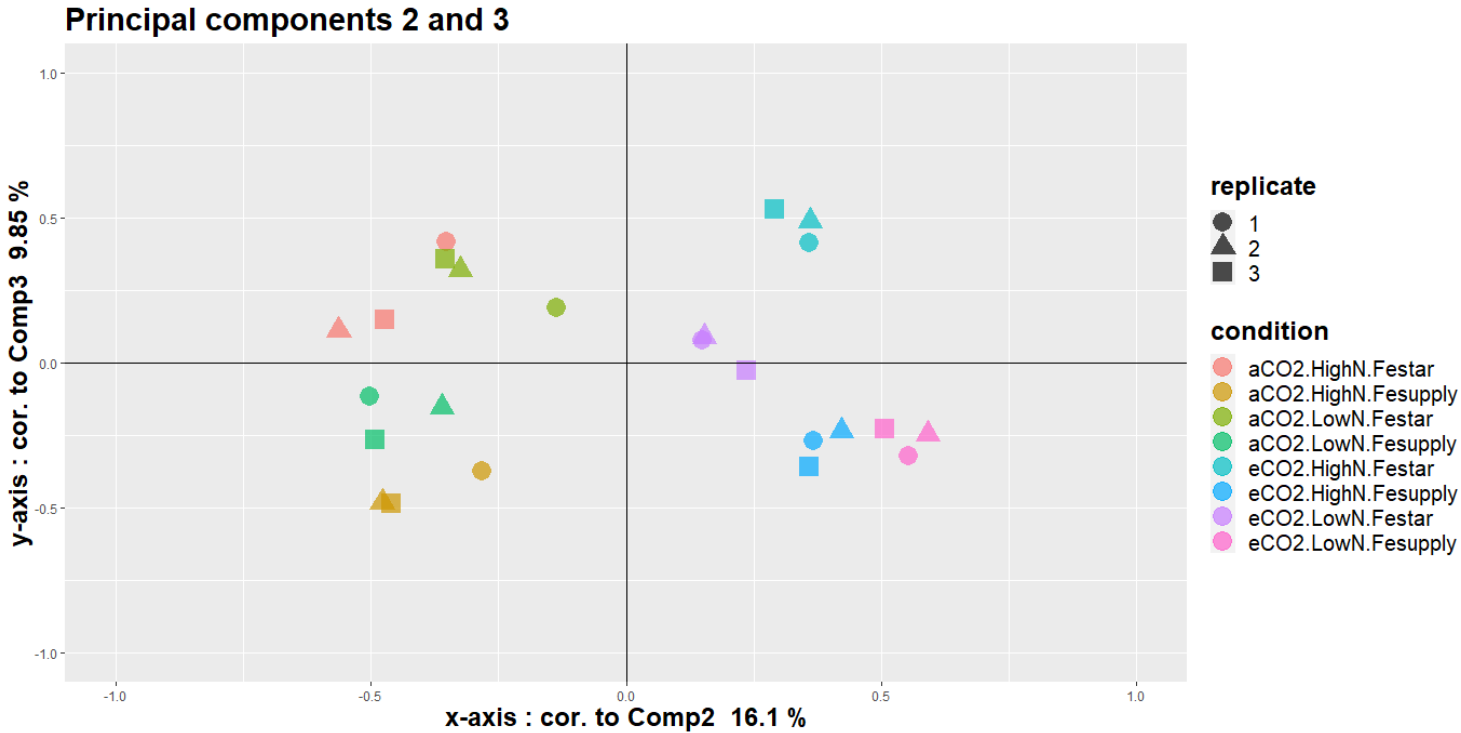


FIGURE 15 – Contribution des variables aux composantes principales 2 et 3. Les couleurs des variables représentent la condition expérimentale.

Après cette analyse en composantes principales, on peut constater que c'est l'effet du faible nitrate qui explique le plus la variance des données, ceci suivi de l'effet du fort  $CO_2$  puis moins de 10% de la variance est expliquée par la privation de fer. Ce qui diffère de l'étude réalisée chez *Arabidopsis thaliana* dans mon équipe d'accueil, organisme chez lequel l'effet  $CO_2$  avait une amplitude moins importante que les carences en Fer et en Nitrate. Cela met en avant des différences d'intérêt liées à la réponse de chaque organisme à ces stress environnementaux.

### 3.3.3 Analyse de l'expression différentielle des gènes

Pour l'analyse de l'expression différentielle des gènes (DEA), on choisit de s'intéresser aux gènes différentiellement exprimés lors d'un stress en fort  $CO_2$  dans des conditions de faible approvisionnement en nitrate et d'approvisionnement en fer. Nous faisons ce choix car des analyses ont été faites sur la plante *Arabidopsis thaliana* dans mon équipe d'accueil avant mon arrivée et nous voudrions comparer les réseaux de régulation composés de ces gènes obtenus entre les 2 organismes. Ces gènes sont également intéressants car, comme vu lors des analyses phénotypiques, le faible nitrate semble interagir négativement avec le fort  $CO_2$ , en annulant notamment l'augmentation de biomasse induite par le fort  $CO_2$ .

Afin de réaliser ces analyses, j'utilise les fonctions `estimateDispersion` et `estimateDEGs` du package DIANE [9] qui utilise le package edgeR [22] pour tester si

les gènes sont différentiellement exprimés entre 2 conditions en utilisant un test de quasi-vraisemblance expliqué dans la partie 2.3.2. Dans le package DIANE, l'utilisateur de ces fonctions peut aussi ajuster un seuil de p-valeur ajustée (FDR) à ne pas dépasser pour les tests effectués ainsi qu'un Log Fold Change (LFC) absolu minimal. Pour ma part, après discussion avec mes tuteurs, nous avons choisi un seuil de FDR à 1 % pour avoir suffisamment de gènes différentiellement exprimés pour la suite des analyses en ayant un taux de faux positifs faible (ici donc 1 %). Et on choisit de prendre un LFC égal à 1 car on veut prendre des gènes suffisamment différentiellement exprimés pour que les chances qu'ils aient des répercussions biologiques plus grandes soient accrues.

On obtient alors une liste de 2314 gènes différentiellement exprimés pour ce changement de conditions sur lesquels nous allons poursuivre nos analyses.

### 3.3.4 Enrichissement ontologique des gènes : GO terms

Reprenons cette liste de 2314 gènes différentiellement exprimés, nous aimerions savoir si des fonctions particulières dans les descriptions des gènes ressortent de manière récurrente pour ces gènes. On procède donc à l'enrichissement ontologique de ces gènes dont la procédure est détaillée dans la partie 2.3.3. On utilise la fonction `enrich_go_custom` du package DIANE qui s'appuie sur le package R `clusterProfiler` [28] et qui emploie des tests exact de Fisher sur la distribution hypergéométrique pour déterminer quels termes d'ontologie sont significativement plus représentés parmi cette liste de 2314 gènes. Nous effectuons des tests multiples donc nous devons ajuster la p-valeur de chaque test à l'aide de la correction FDR [5] (False Discovery Rate).

On obtient une seule caractéristique significativement plus représentée (avec un seuil fixé à 5%) parmi 1900 testées lors de ces enrichissement ontologiques qui est "regulation of transcription". Les enrichissements ontologiques pour les organismes non modèles comme la micro tomate peuvent donner moins de résultats car ces organismes sont moins bien annotés.

### 3.4 Inférence de réseaux

Passons maintenant à l'inférence de réseaux de régulation de gènes. Comme il a été détaillé et expliqué dans les sections 1.2 et 2.3.4, l'une des motivations pour inférer un réseau est de faire des prédictions sur les gènes les plus influents, c'est à dire les gènes qui régulent le plus de gènes cibles dans le réseau, sur la base de notre liste de 2314 gènes différentiellement exprimés. Nous allons donc dans cette section montrer deux réseaux inférés, un en utilisant la méthode des Random Forests (cf section 2.2.2) et un autre en utilisant la méthode LASSO (cf section 2.2.3).

Avant de réaliser l'inférence de réseaux via ces 2 approches, nous sélectionnons dans un premier temps les gènes régulateurs parmi la liste des 2314 gènes différentiellement exprimés. Les gènes de la micro tomate étant étudiés depuis moins longtemps [16] (premier séquençage du génome en 2011) que d'autres plantes comme *Arabidopsis thaliana* dont le génome est étudié depuis 2000, les gènes régulateurs n'ont pas encore été aussi clairement identifiés. Nous allons donc utiliser les annotations des gènes de la micro tomate récupérées sur le site [GBF-Genome](#), les données sur le génome de la micro tomate contenues dans ce site sont mises à disposition par le Laboratoire de Génomique et de Biotechnologie des Fruits (GBF). A partir de ces données, nous sélectionnons les gènes régulateurs : ceux qui ont une description de *transcription factor* dans l'annotation des gènes sont considérés pour la suite de l'étude comme gènes régulateurs.

Les méthodes de régression étant très sensibles à la corrélation entre les variables prédictives, qui doit donc être traitée pour des raisons de stabilité, nous regroupons les régulateurs fortement corrélés en un seul pour éviter qu'on ne perde des arêtes potentielles entre gènes régulateurs et gènes cibles [9]. Pour effectuer ce regroupement, on résume l'expression des régulateurs qui sont corrélés au-dessus d'un certain seuil (90%) comme de nouvelles variables, étant la moyenne de ces régulateurs corrélés. Cela facilite l'estimation et l'interprétation des modèles. Ceci étant fait, on obtient 94 gènes (ou groupes de gènes fortement corrélés) considérés comme régulateurs parmi les 2275 gènes (ou groupes de gènes fortement corrélés), nous pouvons maintenant commencer l'inférence de réseaux.

#### 3.4.1 Inférence de réseaux par Random Forests

Pour inférer un réseau avec la méthode des Random Forests (cf partie 2.3.4), j'utilise la fonction `network_inference` du package DIANE [9] qui utilise le package R GENIE3 [14]. Cette fonction prend en entrée une liste de gènes, pour nous 2275 gènes après regroupement des gènes régulateurs fortement corrélés, qui seront les noeuds du réseau inféré, ainsi que la matrice du niveau d'expression de ces gènes dans les conditions étudiées. Il faut aussi spécifier quels gènes sont des régulateurs parmi cette liste de gènes. Cette fonction permet alors de déterminer un poids d'influence de chaque régulateur sur chaque gène d'entrée en utilisant leurs niveaux d'expression respectifs (cf partie 2.2.2). Pour un gène cible, la méthode utilise une forêt aléatoire qui fait une régression avec l'expression du gène cible en variable réponse et l'expression des gènes régulateurs en variables prédictives, le poids d'in-

fluence des gènes régulateurs sur le gène cible est extrait des arbres une fois ajustés.

On réitère la méthode pour tous les gènes cibles, on a donc pour chaque gène cible un classement des arêtes régulateur-cible, et on effectue un classement global de ces arêtes par rapport à la valeur de leur poids d'influence. On obtient en sortie de cette fonction `network_inference` une matrice avec les gènes cibles en colonne (pour nous 2275), et les gènes régulateurs en ligne (pour nous 94) ainsi que le poids du lien entre le régulateur et le gène cible (cf équation (7) partie 2.2.2). Un détail qu'il est utile de mentionner est qu'un gène régulateur peut aussi être un gène cible d'autres gènes régulateurs, c'est pour cela que les gènes régulateurs sont aussi dans la liste des gènes cibles.

J'utilise ensuite la fonction `test_edges` du package DIANE, l'objectif de cette fonction est de réaliser un test non paramétrique de la présence ou absence d'une arête, via une approche par permutation. L'idée est de construire un premier réseau biologiquement pertinent avec les poids des liens les plus importants donnés par la fonction `network_inference` et qui est ensuite affiné par des tests statistiques, la procédure est détaillée dans la partie 2.3.4.

On choisit une densité d'arêtes présentes dans le réseau de 0.01 dans notre cas car cela donne un bon compromis entre un sous-ensemble suffisamment petit d'arêtes à tester et une valeur de densité cohérente et pas trop restrictive. On calcule maintenant le nombre d'arêtes  $E$  retenues dans le réseau  $E = d \times E_{max}$  avec  $E_{max} = N_{regulators}(N_{cibles} - 1) = 94(2275 - 1) = 213756$ , on obtient donc  $E = 0.01 \times 213756 \approx 2138$ . On sélectionne donc les 2138 arêtes avec les poids les plus importants du réseau pour la suite de nos analyses.

On effectue ensuite les tests statistiques non-paramétriques à l'aide du package R `rfPermute` [3] sur les 2138 arêtes retenues dans le réseau avec nos 24 échantillons, et on obtient une p-valeur pour chaque arête. On applique ensuite la correction FDR [5] pour les tests multiples, et seules les arêtes avec une p-valeur ajustée inférieure à un seuil de FDR de 0.025 sont conservées pour former le réseau final. Ce seuil a été choisi après discussion avec mes tuteurs, cela représente un bon compromis entre le nombre d'arêtes conservées dans le réseau (2006 arêtes) tout en ayant un seuil assez faible pour avoir des arêtes le plus significatives possible. Cette étape est réalisée à l'aide de la fonction `network_from_tests` du package DIANE. On conserve donc 2006 arêtes sur les 2138 arêtes retenues de base.

Passons maintenant à la visualisation du réseau inféré. Dans un premier temps, à l'aide de la fonction `draw_discarded_edges` du package DIANE, on représente le réseau inféré avec en rouge les arêtes qui ne sont pas significatives selon les tests de permutation sur la FIGURE 16.



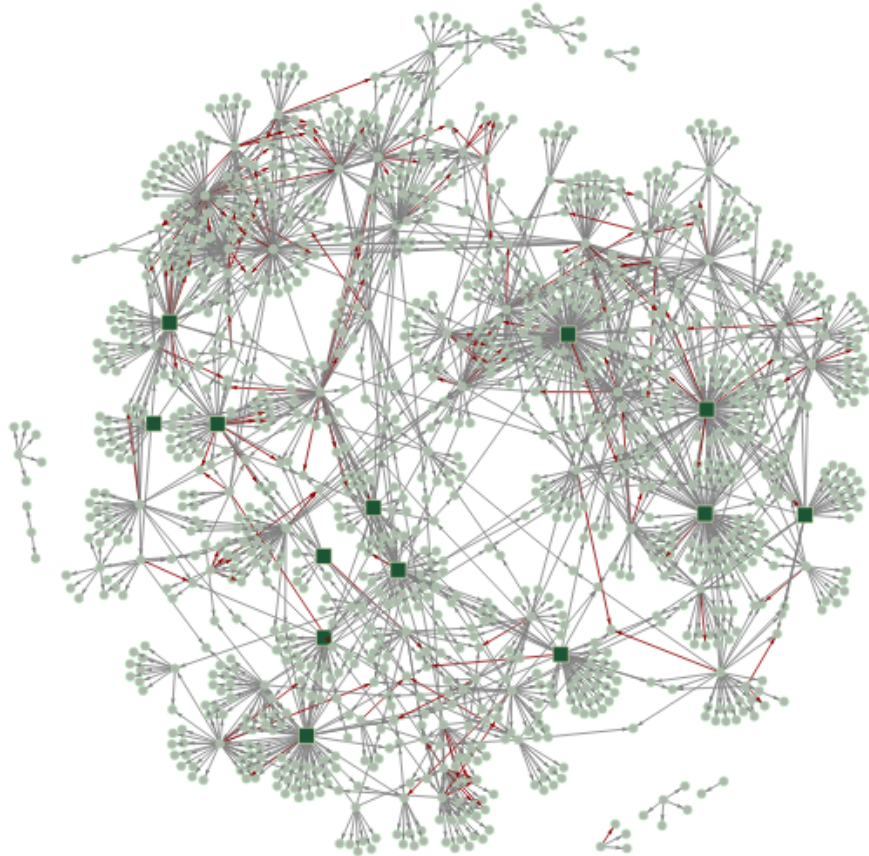


FIGURE 16 – Réseau de gènes de régulation en réponse à un stress environnemental en fort  $CO_2$  en condition de faible nitrate et d’approvisionnement standard en fer. Les noeuds en forme de carrés verts foncés représentent les gènes régulateurs fortement corrélés regroupés en un seul gène régulateur, les noeuds ronds blancs représentent soit des gènes régulateurs soit des gènes cibles. Les arêtes grises orientées sont significatives selon nos tests de permutation et partent du gène régulateur et vont vers le gène cible, les arêtes en rouge sont celles qui ne sont pas significatives selon nos tests de permutation. Il y a au total 2138 arêtes dans ce réseau, 2006 sont en grises et 132 en rouge.



Puis on représente le réseau final après avoir testé les arêtes de notre réseau et retiré celles qui n'étaient pas significatives avec la fonction `draw_network` du package DIANE sur la FIGURE 17 :

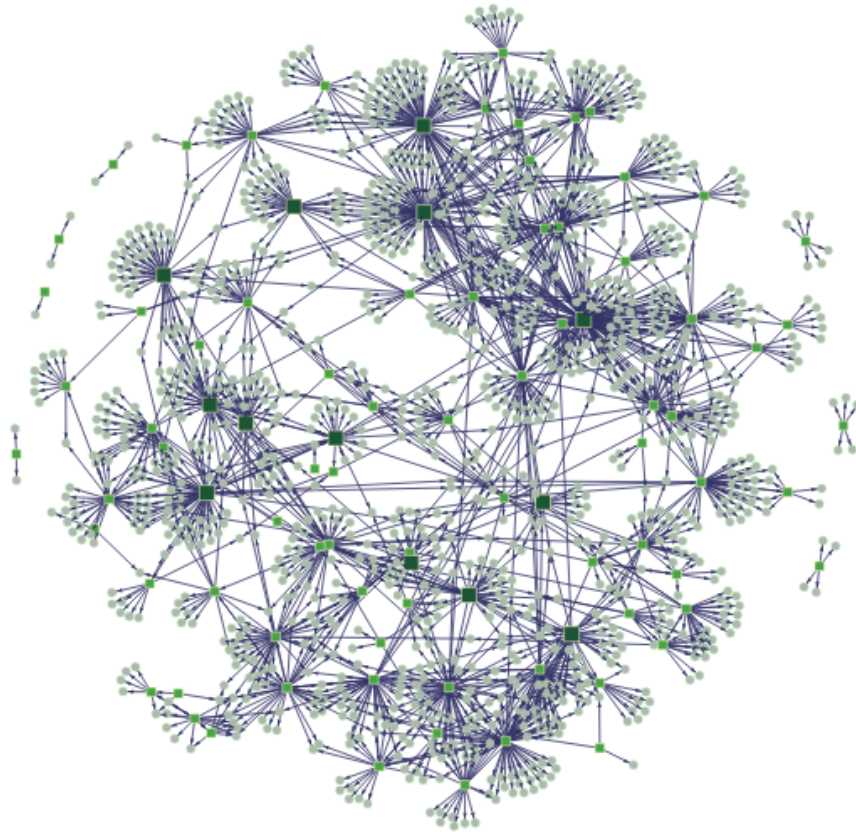


FIGURE 17 – Réseau de gènes de régulation avec tests réalisés sur les arêtes en réponse à un stress environnemental en fort  $CO_2$  en condition de faible nitrate et d'approvisionnement standard en fer. Comme sur la FIGURE 16, les noeuds en forme de carrés verts foncés représentent les gènes régulateurs regroupés, la différence ici est que les noeuds en forme de carrés verts clairs représentent les gènes régulateurs, et les noeuds ronds représentent les gènes cibles. Ce réseau final est composé de 2006 arêtes bleues orientées.

### 3.4.2 Inférence de réseaux par LASSO

On veut maintenant effectuer un réseau de régulation en utilisant les modèles linéaires avec pénalités : la régression LASSO (cf 2.2.3). Nous appliquons cette régression car l'estimation classique d'un modèle linéaire se fait dans le cas  $n > p$ , ce qui n'est pas le cas ici, car nous avons un nombre d'échantillons  $n=24$  et un nombre de variables explicatives  $p=94$ .

Dans notre cas  $n=24 < 94=p$ , le LASSO donnera un coefficient non nul à au plus  $n$  gènes régulateurs pour chaque régression sur un gène cible. On réalise une validation croisée pour choisir le paramètre  $\lambda$ , c'est à dire qu'on sélectionne le  $\lambda$  qui minimise l'erreur de prédiction sur les données tests (données que l'on a pas utilisées lors de l'apprentissage).

Il y a 2 possibilités pour le choix du  $\lambda$  que l'on sélectionne lors de la validation croisée :

1. Le  $\lambda_{min}$  qui est le  $\lambda$  qui minimise l'erreur moyenne de prédiction de validation croisée sur les données tests.
2. Le  $\lambda_{1se}$  qui est la plus grande valeur de  $\lambda$  telle que l'erreur de prédiction se situe à au plus 1 écart type des erreurs de prédiction de  $\lambda_{min}$

Après discussions avec mes tuteurs, nous avons choisi de sélectionner le  $\lambda_{1se}$  pour inférer le réseau de régulation de gènes. L'avantage de ce paramètre est qu'il permet une sélection de variables plus restrictive que le  $\lambda_{min}$ , son inconvénient est que cette sélection se fait aux dépens d'une erreur de prédiction un peu plus grande. Cependant, le but de l'inférence de réseau de régulation étant de trouver des gènes importants qui pourraient être la cause de la diminution de la teneur en minéraux chez les plantes en condition de fort  $CO_2$ , une sélection plus réduite de gènes importants est biologiquement plus intéressante car effectuer des mutations sur un gène coûtent en temps de manipulation et en argent, il est donc nécessaire de prendre en compte ces paramètres pour respecter le budget ainsi que le temps de travail des biologistes.

Pour inférer un réseau de régulation en utilisant la régression LASSO, j'utilise la fonction [cv.glmnet](#) du package R `glmnet` [11]. Cette fonction prend en entrée la matrice du niveau d'expression des 94 gènes régulateurs qui sont les variables explicatives dans les 24 échantillons, les échantillons sont en ligne et les gènes en colonne, ainsi que les niveaux d'expression d'un gène cible qui est la variable réponse dans les 24 échantillons. On précise aussi dans les paramètres de la fonction à quel type de loi la variable réponse appartient avec la commande `"family="poisson"` car les niveaux d'expressions du gène cible sont des comptages et suivent une loi de poisson ainsi que le nombre de folds utilisés pour la validation croisée avec la commande `"nfolds=6"` car disposant de 24 échantillons, on effectue la validation croisée sur 6 parties égales (folds) de 4 échantillons.

On applique cette fonction sur les 2275 gènes différentiellement exprimés. On obtient donc, pour chaque gène cible, une liste de gènes régulateurs qui régulent ce gène cible, et on peut donc construire un réseau de régulation comme précédemment composé de 16670 arêtes, qui est bien plus conséquent que le réseau de régulation inféré avec la méthode des Random Forests. Cette différence s'explique par une sélection distincte des gènes régulateurs pour un gène cible, nous verrons dans la partie 3.4.3 qu'avec la méthode du LASSO, on sélectionne en moyenne plus de régulateurs pour un gène cible, d'où un réseau de régulation final composé de plus d'arêtes.

### 3.4.3 Comparaison des 2 méthodes

Nous voulons maintenant comparer statistiquement et biologiquement les inférences de réseaux réalisées avec ces 2 méthodes. La sélection de gènes régulateurs pour expliquer l'expression d'un gène cible étant différente selon la méthode, nous n'obtenons pas les mêmes réseaux de régulation. Il peut arriver qu'aucun gène régulateur soit sélectionné pour un gène cible, dans ce cas là, ce gène cible n'apparaîtra pas dans le réseau. Nous obtenons pour la méthode des Random Forests un réseau composé de 1298 gènes, et un réseau composé de 2243 gènes pour la méthode LASSO parmi les 2275 gènes différentiellement exprimés de l'étude.

#### Comparaison statistique :

Les modèles d'inférence de réseaux étant basés sur des régressions, ils ont un pouvoir prédictif qui peut être évalué. La capacité d'un modèle à prédire correctement l'expression d'un gène cible au moyen de l'expression des régulateurs retenus dans le réseau nous a semblé un critère de validation méthodologique pertinent.

Pour la comparaison statistique entre le LASSO et les Random Forests, on regarde les erreurs de prédiction de l'expression des gènes cibles des 2 méthodes sur l'out of bag pour les Random Forests et sur le fold out en validation croisée pour le LASSO avec le critère du MSE (voir équation(6)).

On applique ce même critère sur les gènes cibles des 2 méthodes et on obtient les MSE de ces gènes cibles pour chaque méthode. Nous pouvons alors comparer ces MSE obtenus.

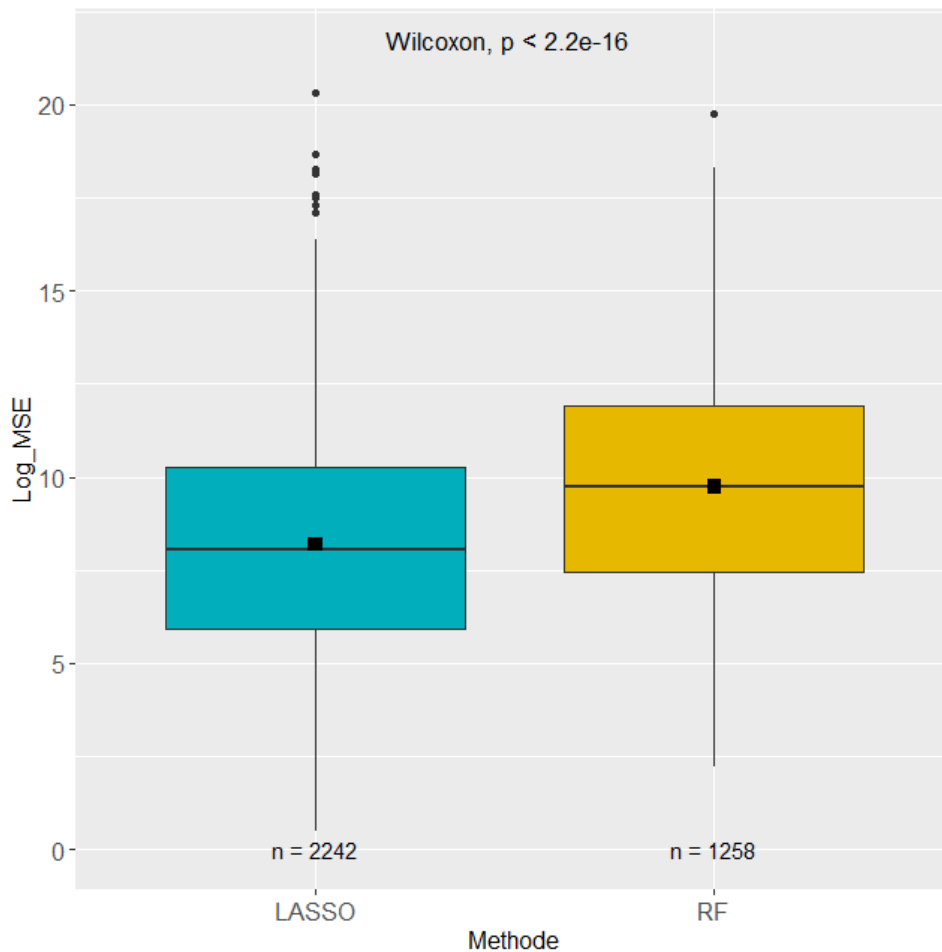


FIGURE 18 – Boxplot représentant le log du MSE des gènes cibles en fonction de la méthode utilisée, le nombre de gènes cibles pour le LASSO étant de 2242 et de 1258 pour les Random Forests. Un test de Wilcoxon a été réalisé pour comparer les moyennes des log des MSE, représentées en carré sur la figure, des 2 méthodes.

On s'aperçoit sur la FIGURE 18, où on a représenté la moyenne des log du MSE pour chaque méthode par un carré, que la moyenne de la méthode Random Forests est significativement plus grande que celle du LASSO, ce à quoi on pouvait s'attendre car le modèle LASSO sélectionne un plus grand nombre de régulateurs par gène cible que le modèle des Random Forests. Un test non-paramétrique a été effectué pour comparer ces moyennes, le test de Wilcoxon. On rejette l'hypothèse nulle de ce test qui est que les moyennes sont égales car nous obtenons une p-valeur inférieure à  $2.2e-16$ . Ceci nous permet donc de dire que les prédictions des niveaux d'expression des gènes cibles sont meilleures avec la méthode du LASSO. On effectue ces analyses sur les log des MSE pour éviter les valeurs extrêmes.

Il est aussi intéressant de regarder combien de gènes régulateurs sélectionne chaque méthode pour un gène cible en moyenne. En effet, d'un point de vue biologique, il est intéressant d'avoir peu de régulateurs par gène cible tout en gardant un bon pouvoir de prédiction pour d'éventuelles manipulations à réaliser sur ces gènes régulateurs. Le nombre moyen de gènes régulateurs sélectionnés pour un gène cible

est de 1.6 pour les Random Forests et de 7.3 pour le LASSO.

### Comparaison biologique :

Le deuxième type de validation envisagé est une validation qui a pour objectif de quantifier la pertinence biologique des relations de régulation prédites.

Afin de comparer les réseaux de régulation obtenus pour les 2 méthodes, il faut dans un premier temps rechercher des réseaux de régulation connus dans des bases de données expérimentales. Ne disposant pas de données suffisantes sur la micro tomate qui n'a pas été beaucoup étudiée, nous devons faire correspondre les gènes de la micro tomate appartenant aux réseaux inférés par les 2 méthodes avec les gènes d'une autre plante sur laquelle plus d'études ont été réalisées. On veut donc chercher des gènes orthologues à ceux de la micro tomate, c'est à dire des gènes qui ont un lien évolutif avec ceux de la micro tomate mais qui proviennent d'une autre espèce. On réalise cette orthologie avec la plante modèle *Arabidopsis thaliana* dont beaucoup de réseaux de régulation sont connus dans des bases de données, notamment la base de données ConneCTF [7] qui contient du DAP-Seq, CHIP-Seq et TARGET. Ces expériences contiennent des interactions établies expérimentalement, soit relatives à la fixation d'un régulateur à son gène cible (CHIP-Seq, DAP-Seq), ou relatives au changement d'expression induit par un régulateur sur ses cibles (TARGET).

Nous effectuons la recherche d'orthologues entre les gènes de la micro tomate et ceux d'*Arabidopsis thaliana* à l'aide d'un outil bio-informatique [MicroTomHomolog](#) développé dans mon équipe d'accueil. Cet outil permet, à partir d'une liste d'accessions de gènes de la micro tomate, d'extraire les séquences protéiques correspondantes pour chaque gène grâce au programme [extract\\_fasta\\_seq](#) puis de comparer ces séquences protéiques avec les séquences protéiques d'*Arabidopsis thaliana* grâce au programme blast [2]. On utilise un seuil de signification statistique pour signaler les correspondances entre les séquences protéiques de ces gènes qui est de  $1e-05$ , ce qui signifie qu'on s'attend à ce que la probabilité que ces correspondances soient trouvées simplement par hasard est de  $1e-05$ . Si la signification statistique attribuée à une correspondance est supérieure à ce seuil, cette correspondance ne sera pas signalée.

### Validation du réseau de gènes inféré avec la méthode des Random Forests :

On réalise donc cette recherche d'orthologues dans un premier temps sur tous les gènes du réseau inféré avec la méthode des Random Forests, que cela soit des gènes régulateurs, des éléments d'un groupe de gènes regroupés car fortement corrélés (cf 3.4) ou des gènes cibles, ce qui fait en tout 1337 gènes. On obtient 1194 gènes orthologues chez *Arabidopsis*, il y a donc 143 gènes du réseau dont on n'a pas trouvé d'orthologue. On obtient finalement un réseau de gènes composé de 1787 arêtes chez *Arabidopsis* au lieu de 2006 chez la micro tomate car on retire du réseau les 143 gènes sans orthologue, et de ce fait toutes les arêtes les concernant. Nous pouvons

maintenant vérifier si les 1787 arêtes du réseau inféré avec la méthode des Random Forests sont retrouvées dans la base de données ConnecTF [7].

Pour ce faire, nous appliquons la fonction `evaluate_network` du package AraNetBench [8]. Cette fonction nous renvoie le nombre d'arêtes du réseau validées, c'est à dire retrouvées dans la base de données ConnecTF, le nombre d'arêtes non validées par ConnecTF, le nombre d'arêtes sans donnée de validation, le pourcentage d'arêtes validées ainsi que le pourcentage d'arêtes non validées du réseau. La différence entre une arête non validée par ConnecTF et une arête sans donnée de validation est la suivante : une arête est dite non validée si le gène régulateur d'où part l'arête a été étudié dans au moins une expérience de ConnecTF, mais que le gène auquel il est relié ne figure pas parmi les cibles attribuées à ce régulateur dans ConnecTF. On compte donc cette arête comme une erreur. Une arête est dite sans donnée de validation si le gène régulateur d'où part l'arête n'a été étudié dans aucune expérience de ConnecTF, on n'a donc aucune information sur quelles sont ses cibles dans notre validation, donc on ne compte pas cette arête comme une erreur. Les résultats obtenus sont les suivants :

- 328 arêtes validées par ConnecTF.
- 867 arêtes non validées par ConnecTF.
- 592 arêtes sans donnée de validation.
- Un pourcentage d'arêtes validées de 27.4%.
- Un pourcentage d'arêtes non validées de 72.6%.
- Un pourcentage d'arêtes sans données de validation d'environ 33.1%.

Nous nous sommes ensuite demandés si, finalement, le réseau inféré par cette méthode possédait plus d'arêtes validées par ConnecTF qu'un réseau inféré avec les mêmes gènes régulateurs et cibles mais dont les connexions sont faites aléatoirement. Nous avons donc testé si le réseau inféré avec les Random Forests est meilleur en terme d'arêtes validées par ConnecTF qu'un ensemble de réseaux inférés aléatoirement à l'aide de la fonction `test_validation_rate` du package AraNetBench. Le résultat de ce test est présenté sur la FIGURE 19, ainsi que le Z-score, qui correspond au nombre d'écarts types séparant le pourcentage d'arêtes validées par le réseau inféré avec les Random Forests de la moyenne des pourcentages d'arêtes validées par un ensemble de réseaux inférés aléatoirement.

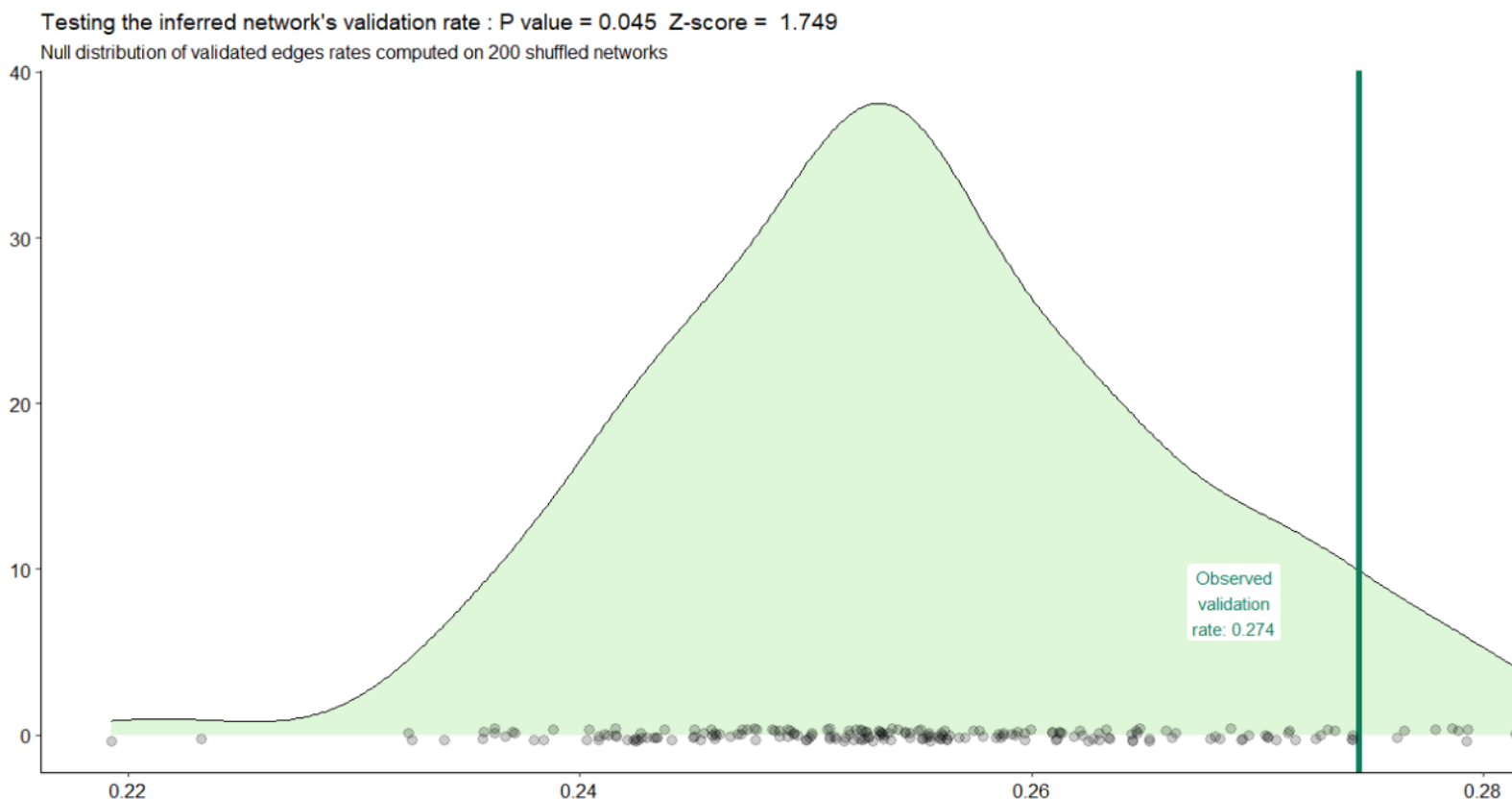


FIGURE 19 – Graphique représentant la distribution nulle des pourcentages d’arêtes validées sur 200 réseaux de régulation inférés aléatoirement ainsi que le pourcentage d’arêtes validées du réseau inféré par la méthode des Random Forests.

On voit sur la FIGURE 19 que le pourcentage d’arêtes validées du réseau inféré par les Random Forests se situe dans la queue de la distribution nulle des pourcentages d’arêtes validées sur les 200 réseaux de régulation inférés aléatoirement. On peut donc rejeter l’hypothèse nulle qui est que le pourcentage d’arêtes validées par ConectTF est le même pour le réseau inféré par la méthode des Random Forests que pour un réseau inféré aléatoirement car la p-valeur du test effectué est de 4.5%, le réseau inféré par les Random Forests est donc significativement meilleur qu’un réseau inféré aléatoirement et il permet de capturer des relations de régulation qui sont conservées et validées par des données expérimentales chez *Arabidopsis*.

### Validation du réseau de gènes inféré avec la méthode LASSO :

On réitère exactement le même processus que pour les Random Forests, on rappelle que le nombre d’arêtes du réseau inféré par la méthode LASSO chez la microtomate est de 16670.

On obtient un nombre total de gènes dissociés qui participent au réseau inféré de 2282. Parmi ces gènes, 277 n’ont pas d’orthologues chez *Arabidopsis*. On obtient donc un réseau de gènes composé de 14759 chez *Arabidopsis* après avoir retiré les

gènes sans orthologues ainsi que les arêtes les concernant. Nous regardons maintenant quel pourcentage de ces arêtes est retrouvé dans ConnecTF par le même processus que précédemment. Les résultats obtenus sont les suivants :

- 1841 arêtes validées par ConnecTF.
- 5199 arêtes non validées par ConnecTF.
- 7719 arêtes sans donnée de validation.
- Un pourcentage d'arêtes validées de 26.2%.
- Un pourcentage d'arêtes non validées de 73.8%.
- Un pourcentage d'arêtes sans données de validation d'environ 52.3%.

Comme précédemment, nous testons si le réseau inféré avec LASSO est meilleur en terme d'arêtes validées par ConnecTF qu'un réseau inféré aléatoirement, c'est à dire un réseau inféré avec les mêmes gènes régulateurs et cibles que le réseau inféré avec LASSO, mais dont les connexions sont faites aléatoirement. On présente les résultats de ce test sur la FIGURE 20.

Testing the inferred network's validation rate : P value = 0 Z-score = 3.616  
Null distribution of validated edges rates computed on 200 shuffled networks

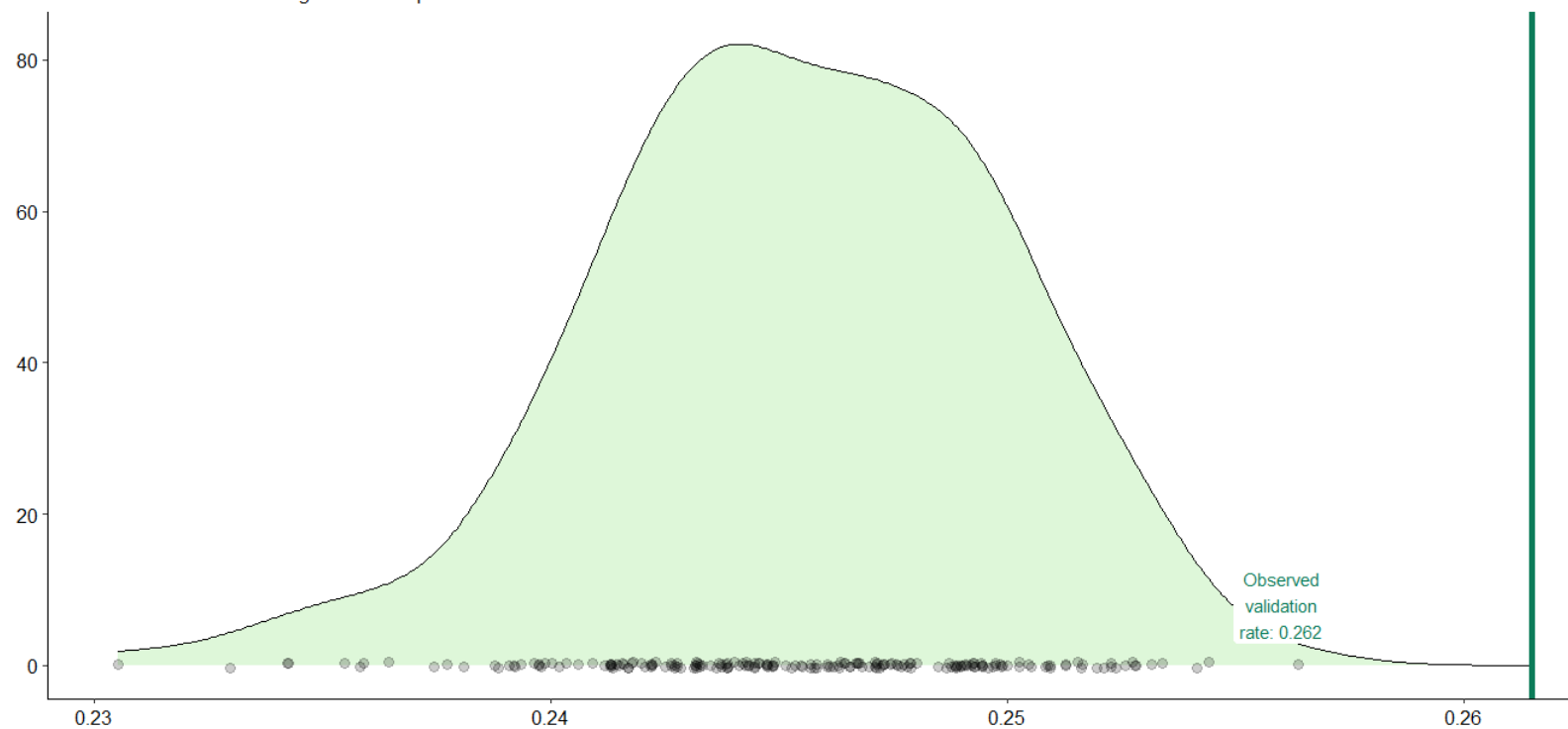


FIGURE 20 – Graphique représentant la distribution nulle des pourcentages d'arêtes validées sur 200 réseaux de régulation inférés aléatoirement ainsi que le pourcentage d'arêtes validées du réseau inféré par la méthode LASSO.



On voit sur la FIGURE 20 que le test nous renvoie une p-valeur égale à 0, et que le pourcentage d'arêtes validées du réseau inféré par LASSO est au-delà de la queue de la distribution nulle des pourcentages d'arêtes validées sur les 200 réseaux de régulation inférés aléatoirement. On peut donc rejeter l'hypothèse nulle qui est que le pourcentage d'arêtes validées par ConneCTF est le même pour le réseau inféré par la méthode LASSO que pour un réseau inféré aléatoirement de manière certaine, le réseau inféré par LASSO est donc significativement meilleur qu'un réseau inféré aléatoirement.

Voici le tableau 2 récapitulatif des comparaisons statistiques et biologiques effectuées entre ces 2 méthodes.

**Tableau récapitulatif de la comparaison des 2 méthodes :**

Méthode	Nombre d'arêtes chez la micro tomate	Nombre d'arêtes chez <i>Arabidopsis</i>	Taux d'arêtes validées par ConneCTF	P-valeur	Z-score	Moyenne des log(MSE) pour les gènes cibles	Nombre de régulateurs par gène cible en moyenne
Random forest	2006	1787	27.4%	0.045	1.749	9.74	1.6
LASSO	16670	14759	26.2%	0	3.616	8.20	7.3

TABLE 2 – Tableau comparatif des réseaux de régulation inférés par les 2 méthodes.

On remarque que même si le taux d'arêtes validées par ConneCTF est légèrement meilleur avec la méthode des Random Forests, la méthode du LASSO possède un Z-score bien plus grand, ce qui veut dire que le pourcentage d'arêtes validées avec cette méthode est bien plus grand que ce à quoi on s'attendrait avec un réseau inféré aléatoirement.

On peut ensuite regarder les gènes régulateurs qui sont susceptibles d’avoir un grand intérêt biologique, ce qu’on appellera les régulateurs d’intérêts. Un régulateur est dit d’intérêt s’il a un grand nombre de connexions dans le réseau (c’est à dire le nombre de gènes cibles qu’il régule mais aussi le nombre de régulateurs dont il est la cible).

Pour finir la comparaison entre les 2 méthodes, on cherche donc à identifier les régulateurs d’intérêts des 2 réseaux inférés et comparer ces 2 listes. On souhaiterait retrouver à peu près les mêmes gènes régulateurs dans ces 2 listes car nous voulons sélectionner ceux qui permettraient potentiellement d’expliquer les effets négatifs du  $CO_2$  élevé sur la nutrition minérale des plantes. Si un régulateur se retrouve dans les 2 réseaux inférés, il y a des chances qu’il fasse partie de cette catégorie, et des expériences biologiques pourront ensuite être réalisée sur lui afin de voir si cela a un impact sur la nutrition minérale des plantes dans des conditions de fort  $CO_2$ .

Nous sélectionnons donc les 20 régulateurs les plus importants pour chaque méthode sur la base du nombre de connexions qu’ils possèdent dans le réseau et faisons ensuite l’intersection de ces listes. Nous retrouvons alors 7 régulateurs présents dans ces 2 listes. Parmi ces 7 gènes, on y retrouve un régulateur qui était aussi apparu comme un des gènes les plus connectés du réseau inféré par la méthode des Random Forests sur des données transcriptomiques réalisé chez *Arabidopsis* par mon équipe d’accueil dans les mêmes conditions avant mon arrivée, ce gène est *MYB15*. Ce résultat est intéressant parce que cela montre qu’il peut y avoir des gènes qui orchestrent la réponse au  $CO_2$  élevé qui sont conservés entre espèces, ce qui est une piste de recherche très prometteuse. Les listes de ces gènes sont données en annexe [5](#).

## 4 Conclusions et perspectives

### Conclusions sur les résultats :

L'analyse de l'impact d'une forte concentration en  $CO_2$  atmosphérique sur les caractéristiques phénotypiques de la micro tomate est conforme à ce qui avait déjà été observé dans mon équipe d'accueil sur la plante modèle *Arabidopsis* ainsi qu'aux études de biologie végétale sur le sujet, c'est à dire l'augmentation de la biomasse des racines et des parties aériennes en approvisionnement standard de nitrate, et une diminution de la teneur en minéraux dans le fruit de la plante toujours en approvisionnement standard de nitrate. Ces résultats ont cependant encore plus de poids que ceux obtenus chez *Arabidopsis* car, comme il a été dit dans l'introduction partie 1.1, la micro tomate est une plante d'intérêt agronomique.

Une fois les hypothèses sur l'analyse phénotypique confirmées, nous sommes donc passés à l'analyse du transcriptome de la micro tomate pour répondre aux problématiques (cf 1.4) de mon stage. Le fait d'avoir inféré des réseaux, avec les méthodes Random Forests et LASSO, significativement meilleurs que le hasard est très encourageant car les gènes régulateurs identifiés comme les plus connectés pour chaque réseau ont potentiellement une forte chance de contrôler la réponse des transcriptomes racinaires dans des conditions de  $CO_2$  élevé. Parmi ces gènes d'intérêt, 7 gènes ont été identifiés de manière robuste par les listes des 2 méthodes. Ils ont donc encore plus de chances d'être d'intérêt biologique et pourraient être étudiés plus en détail dans des analyses ultérieures. De plus, un autre fait encourageant est que parmi ces 7 gènes, le gène "MYB15" a déjà été identifié comme un gène régulateur d'intérêt dans le réseau inféré chez la plante *Arabidopsis* par mon équipe d'accueil, des études peuvent donc être faites en priorité sur ce gène.

L'étude comparative de 2 méthodes statistiques, qui sont les Random Forests et le LASSO, pour inférer des réseaux de régulation de gènes a révélé que la méthode des Random Forests commet plus d'erreurs dans les prédictions des niveaux d'expression des gènes cibles que le LASSO. D'un point de vue statistique, la méthode du LASSO est donc meilleure que celle des Random Forests. Cependant, le gros avantage de la méthode des Random Forests est qu'elle est plus sparse, avec un réseau de régulation obtenu composé de beaucoup moins d'arêtes et beaucoup moins dense. Cette méthode possède notamment un pourcentage d'arêtes sans donnée de validation beaucoup plus faible que pour le LASSO (33.1% d'arêtes sans données de validation contre 52.3%), ce qui est plus interprétable et plus intéressant pour le biologiste. Pour aider à choisir laquelle de ces 2 méthodes est la meilleure, nous avons aussi regardé si elles permettaient de prédire des arêtes biologiquement pertinentes. Les tests effectués sur les 2 réseaux inférés ont révélé qu'ils étaient tous deux meilleurs en terme d'arêtes validées dans les bases de données expérimentales que des réseaux construits aléatoirement. Les tests renvoyaient des p-valeurs de 4.5% pour le réseau inféré avec la méthode des Random Forests contre 0% pour celui inféré avec LASSO, avec un Z-score plus élevé pour la méthode LASSO, le LASSO semble donc meilleur en terme de validation biologique.

En conclusion, on pourrait choisir la méthode LASSO pour inférer des réseaux de régulation en s'appuyant sur ces comparaisons. Mais le problème de cette méthode est qu'elle sélectionne trop de variables prédictives (gènes régulateurs) pour un gène cible, on obtient donc des réseaux trop denses et difficilement interprétables.

### **Limitations :**

Le critère de validation biologique est imparfait car en regardant les arêtes de nos 2 réseaux inférés, on remarque que beaucoup n'ont pas de données de validation, ce qui veut dire que beaucoup des arêtes présentes dans le réseau n'ont jamais été étudiées dans les bases de données expérimentales. On ne peut donc pas valider un grand nombre d'interactions du réseau, mais ce n'est pas pour autant que ces arêtes sont "fausses", elles peuvent éventuellement exister mais on a aucun moyen de le vérifier. De plus, trouver des gènes orthologues de la micro tomate chez *Arabidopsis* pour effectuer la validation biologique rajoute des hypothèses et une couche d'incertitude supplémentaire.

### **Perspectives :**

Les perspectives de mon travail peuvent être de réaliser des expériences et manipulations biologiques sur les 7 gènes régulateurs appartenants aux listes des gènes régulateurs d'intérêt des 2 méthodes, et plus particulièrement le gène *MYB15* qui était aussi apparu comme gène régulateur important du réseau inféré par mon équipe d'accueil sur la plante modèle *Arabidopsis*. Nous pourrions alors constater si des mutations sur ces gènes entraîneraient des réponses phénotypiques des plantes en  $CO_2$  élevé différentes de celles que nous avons obtenues. Pour aller plus loin, ces gènes candidats jouant un rôle potentiel dans la réponse phénotypique au  $CO_2$  pourraient être étudiés fonctionnellement par différentes techniques expérimentales dans mon équipe d'accueil. Cela permettrait d'élucider des mécanismes de régulation encore inconnus dans le cadre de la réponse au  $CO_2$  chez les plantes.

Nous pourrions aussi travailler la méthode du LASSO pour avoir un niveau de sparsité équivalent à celui des Random Forests, et donc de pouvoir ensuite comparer les méthodes à nombre moyen semblable de gènes régulateurs sélectionnés pour expliquer l'expression d'un gène cible. On pourrait dans ce cas, à partir des critères de validation statistiques et biologiques, retenir une méthode plus performante qu'une autre. Nous pourrions aussi considérer d'autres mesures d'influence des variables pour la méthode des Random Forests, comme celles présentées par Erwan Scornet avec SIRUS [4].

Une autre perspective de mon travail est de faire une analyse plus en détail des gènes d'intérêt, notamment l'enrichissement ontologique de ces gènes les plus connectés dans les réseaux prédits pour regarder si des fonctions particulières de gènes sont significativement représentées. Nous pourrions aussi regarder si ces gènes d'intérêt sont retrouvés dans un réseau inféré à partir de gènes différentiellement

exprimés entre des conditions de référence pour le  $CO_2$ , le nitrate et le fer contre des conditions de fort  $CO_2$  et de référence pour les apports en nitrate et fer. On obtiendrait alors un réseau de régulation de gènes uniquement lié à l'effet du fort  $CO_2$  et il serait intéressant de retrouver des gènes d'intérêt communs avec ceux que nous avons déjà étudiés.

## 5 Annexe

Liens vers les dépôts Github contenant mes codes réalisés sur Rmarkdown :

Lien pour les analyses phénotypiques : [https://github.com/Paul30hub/Analyse\\_phenotypique](https://github.com/Paul30hub/Analyse_phenotypique)

Analyses transcriptomiques : [https://github.com/Paul30hub/Analyse\\_transcriptomique](https://github.com/Paul30hub/Analyse_transcriptomique)

### Demonstration :

On dispose du modèle suivant :

- $y|\pi \sim \mathcal{P}(\mu)$  avec  $\mu = N\pi$
- $\pi \sim \Gamma(\gamma, m)$

On veut trouver la loi de y.

$$\mathbf{P}(Y = y) = \int_0^{+\infty} e^{-Nu} \frac{(Nu)^y}{y!} e^{-\frac{u}{m}} \frac{u^{\gamma-1}}{\Gamma(\gamma)m^\gamma} du$$

$$= \frac{N^y}{y!} \frac{1}{\Gamma(\gamma)m^\gamma} \int_0^{+\infty} e^{-u(N+\frac{1}{m})} u^{y+\gamma-1} du$$

On effectue le changement de variables suivant :

$$x = u(N + \frac{1}{m}), \text{ on a donc : } dx = (N + \frac{1}{m})du \text{ et } u = x(N + \frac{1}{m})^{-1}$$

L'intégrale se réécrit donc :

$$\int_0^{+\infty} e^{-x} x^{y+\gamma-1} \left(N + \frac{1}{m}\right)^{-(y+\gamma-1)} \left(N + \frac{1}{m}\right)^{-1} dx$$

$$= \int_0^{+\infty} e^{-x} x^{y+\gamma-1} \left(N + \frac{1}{m}\right)^{-(y+\gamma)} dx$$

$$= \frac{\Gamma(y + \gamma)}{\left(N + \frac{1}{m}\right)^{y+\gamma}} \left( \text{car } \int_0^{+\infty} e^{-ax} x^b dx = \frac{\Gamma(b+1)}{a^{b+1}} \right)$$

On revient au calcul de  $\mathbf{P}(Y = y)$  :

$$\begin{aligned} \mathbf{P}(Y = y) &= \frac{N^y}{y!} \times \frac{1}{\Gamma(\gamma)m^\gamma} \times \frac{\Gamma(y + \gamma)}{\left(\frac{Nm+1}{m}\right)^{y+\gamma}} \\ &= \frac{\Gamma(y + \gamma)}{\Gamma(\gamma)y!} \left( \frac{Nm}{Nm+1} \right)^y \left( \frac{1}{Nm+1} \right)^\gamma \end{aligned}$$

En prenant  $Nm = \frac{\mu}{\gamma} \Leftrightarrow m = \frac{\mu}{\gamma N}$ , on obtient,

$$\begin{aligned} \mathbf{P}(Y = y) &= \frac{\Gamma(y + \gamma)}{\Gamma(\gamma)y!} \left( \frac{\mu}{\mu + \gamma} \right)^y \left( \frac{\gamma}{\mu + \gamma} \right)^\gamma \\ &= \frac{\Gamma(y + \gamma)}{\Gamma(\gamma)y!} (1 - p)^y p^\gamma \text{ avec } p = \frac{\gamma}{\mu + \gamma} \\ &= \mathbf{NB}(p, \gamma) \end{aligned}$$

Listes des 20 gènes orthologues chez *Arabidopsis* les plus connectés des 2 réseaux construits par les 2 méthodes statistiques :

**Random Forests :**

AT2G40140  
AT4G37260  
AT5G06839  
AT5G48150  
AT4G30410  
mean\_AT4G38900-AT1G71450-AT5G52020-AT1G01260  
AT1G08010  
AT3G23250  
AT2G47810  
AT4G34530  
mean\_AT5G04840-AT1G08320-AT1G69560-AT1G69310  
mean\_AT1G80840-AT5G24110-AT2G31180-AT3G56400-AT4G11070-AT1G16490  
mean\_AT4G37850-AT5G57150-AT5G43650-AT4G20970-AT2G44840-AT2G38300-AT1G19210-  
AT1G01260-AT5G52260-AT3G10040  
AT2G31180  
AT5G19790  
AT5G08520  
mean\_AT2G42300-AT3G15510-AT3G49950  
mean\_AT5G51790-AT4G37850-AT1G71520-AT5G15130-AT2G44940  
mean\_AT5G13330-AT5G56960  
mean\_AT5G65210-AT4G08250

**LASSO :**

AT5G52020  
AT4G34530  
AT1G64380  
mean\_AT4G38900-AT1G71450-AT5G52020-AT1G01260  
AT2G44940  
AT3G12720  
mean\_AT5G13330-AT5G56960  
mean\_AT5G62470-AT4G18170  
mean\_AT1G76880-AT1G76890  
AT3G23250  
AT2G20180  
AT1G68552  
mean\_AT1G80840-AT5G24110-AT2G31180-AT3G56400-AT4G11070-AT1G16490  
AT1G05805  
AT4G30410  
AT1G75390  
AT2G47810  
mean\_AT1G48000-AT4G11070  
AT4G18690  
AT3G28857

7 gènes sont présents dans les 2 listes, dont AT3G23250 qui correspond au gène *MYB15*.



## Références

- [1] *Algorithms for Minimization Without Derivatives*. 1972.
- [2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3) :403–410, oct 1990.
- [3] E. Archer. *rfPermute : Estimate Permutation p-Values for Random Forest Importance Metrics*, 2022. R package version 2.5.1.
- [4] C. Bénard, G. Biau, S. D. Veiga, and E. Scornet. SIRUS : Stable and interpretable RRule set for classification. *Electronic Journal of Statistics*, 15(1), Jan. 2021.
- [5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1) :289–300, jan 1995.
- [6] L. Breiman. *Machine Learning*, 45(1) :5–32, 2001.
- [7] M. D. Brooks, C.-L. Juang, M. S. Katari, J. M. Alvarez, A. Pasquino, H.-J. Shih, J. Huang, C. Shanks, J. Cirrone, and G. M. Coruzzi. ConnecTF : A platform to integrate transcription factor–gene interactions and validate regulatory networks. *Plant Physiology*, 185(1) :49–66, nov 2020.
- [8] O. Cassan. *AraNetBench : Evaluates an inferred regulatory network against state of the art interaction databases in Arabidopsis thaliana*, 2022. R package version 0.0.0.9000.
- [9] O. Cassan, S. Lèbre, and A. Martin. Inferring and analyzing gene regulatory networks from multi-factorial expression data : a complete and interactive suite. *BMC Genomics*, 22(1), may 2021.
- [10] Y. Dan, Z. Fei, and C. Rothan. Microtom - a new model system for plant genomics. 2007.
- [11] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 2010.
- [12] A.-C. Haury, F. Mordélet, P. Vera-Licona, and J.-P. Vert. TIGRESS : Trustful inference of gene REgulation using stability selection. *BMC Systems Biology*, 6(1), Nov. 2012.
- [13] W. Hayes, K. Sun, and N. Przulj. Graphlet-based measures are suitable for biological network comparison. *Bioinformatics*, 29(4) :483–491, jan 2013.
- [14] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9) :e12776, Sept. 2010.
- [15] IPCC. *Index*, book section Index, page 1523–1535. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.
- [16] J. J. Bancifra, Tarlac State University, Tarlac City, Philippines, and <https://orcid.org/0000-0003-0641-1305>. Supervisory practices of department heads and teachers’ performance : Towards a proposed enhancement program. *APJAET - Journal ay Asia Pacific Journal of Advanced Education and Technology*, pages 25–33, Sept. 2022.
- [17] P. Langfelder and S. Horvath. WGCNA : an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), Dec. 2008.
- [18] I. Loladze. Hidden shift of the ionome of plants exposed to elevated CO2 depletes minerals at the base of human nutrition. *eLife*, 3, may 2014.

- [19] S. P. Lund, D. Nettleton, D. J. McCarthy, and G. K. Smyth. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunk dispersion estimates. *Statistical Applications in Genetics and Molecular Biology*, 11(5), jan 2012.
- [20] D. Marbach, , J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, and G. Stolovitzky. Wisdom of crowds for robust gene network inference. *Nature Methods*, 9(8) :796–804, jul 2012.
- [21] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera, and A. Califano. ARACNE : An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7(S1), Mar. 2006.
- [22] D. J. McCarthy, Y. Chen, and G. K. Smyth. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10) :4288–4297, jan 2012.
- [23] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3) :R25, 2010.
- [24] M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21) :2881–2887, sep 2007.
- [25] G. Sanguinetti and V. A. Huynh-Thu. Gene regulatory networks. 2019.
- [26] Y. Sha, J. H. Phan, and M. D. Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, aug 2015.
- [27] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288, 1996.
- [28] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. clusterProfiler : an r package for comparing biological themes among gene clusters. *OMICS : A Journal of Integrative Biology*, 16(5) :284–287, may 2012.