Master 2 mathématiques Appliquées pour l'ingénierie, l'industrie et l'innovation à l'université Paul Sabatier (Toulouse III)

Rapport de stage de fin d'études

# Étude de l'influence de la prostaglandine E2 péri-conceptionnelle sur le développement de l'embryon bovin

Laboratoire d'accueil : INRAE Jouy-en-Josas, BREED UMR 1198

Aden ALI KAMIL

*Responsables du stage :* CHARPIGNY Gilles
NUTTINCK Fabienne
*Tuteur de l'université :* Jean-Michel LOUBES

1er Février 2020 — 15 Juillet 2020

# Remerciements

Avant toute autre personne, je tiens à remercier mes deux maîtres de stage, Gilles Charpigny et Fabienne Nuttinck, qui m'ont donné la chance d'intégrer l'unité BREED de l'INRAE de Jouy-en-Josas, ce n'est pas pour des raisons d'usage, mais bien pour tout ce que vous m'avez apporté. Merci à Gilles d'avoir toujours pris le temps de m'expliquer les choses, d'avoir pris le temps de discuter de choses diverses et variées pour transmettre ses connaissances. Merci à Fabienne pour sa gentillesse, sa bonne humeur et son grand soutien.

Ensuite je tiens à remercier le reste de l'équipe (Mariam, Corinne, Olivier, Laurent) pour tous les moments de partage, les discussions, l'entraide.

Je souhaite adresser mes remerciements les plus sincères aux personnes ayant contribué au bon déroulement de mon stage.

Enfin je veux aussi remercier chaleureusement Véronique Duranthon de m'avoir accueilli au sein de son équipe pour mon stage de M2.

# Table des matières

**Remerciements** **1**

**Présentation du stage** **3**

**1 Contexte biologique** **5**

1.1 Notions de biologie . . . . . 5

1.2 La technologie RNA-seq . . . . . 7

1.3 Modèle expérimental . . . . . 7

**2 Traitement de données bio-informatique** **9**

2.1 Description de données brutes . . . . . 9

2.2 Fiabilité de données . . . . . 10

2.3 Alignement . . . . . 11

**3 L'analyse différentielle de données RNA-seq** **13**

3.1 Statistique descriptive . . . . . 13

3.2 Filtrage . . . . . 15

3.3 Normalisation . . . . . 16

3.3.1 RLE . . . . . 16

3.3.2 TMM . . . . . 17

3.4 Détection des gènes différentiellement exprimés . . . . . 18

3.4.1 Modélisation . . . . . 18

3.4.2 Test statistique . . . . . 19

3.4.3 DESeq2 . . . . . 19

3.4.4 edgeR . . . . . 21

3.4.5 Comparaison de deux packages . . . . . 23

**4 Classification Ascendante Hiérarchique** **25**

**5 Analyse en Composantes Principales** **27**

**6 Intégration des données omiques** **32**

6.1 PLS . . . . . 32

6.2 PLS-DA . . . . . 35

**7 Inférence de réseaux** **38**

7.1 Principe . . . . . 38

7.2 Modèle graphique gaussien . . . . . 39

**Discussion** **42**

**Conclusion** **43**

**Bibliographie** **44**

**Annexe** **45**

# Présentation du stage

J'ai effectué mon stage de fin d'études à l'institut national de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE) de Jouy-en-Josas, j'ai intégré l'unité mixte de recherche 1198 « Biologie de la Reproduction, épigénétique, Environnement, Développement » (BREED). Au cours de ce stage, j'ai pu m'intéresser aux analyses de données du type « omiques » pour mieux comprendre l'impact des prostaglandines péri-conceptionnelles présentes dans le micro-environnement ovocytaire sur le développement embryonnaire chez la vache.

Chez les mammifères, le micro-environnement de l'ovocyte s'enrichit en prostaglandine E2 (PGE2) au moment de l'ovulation et la fécondation. La PGE2 est un médiateur lipidique bien connu en cancérologie pour sa capacité à stimuler la survie et la multiplication des cellules tumorales. Chez la vache la mortalité embryonnaire se produit principalement durant les deux semaines qui suivent la fécondation. La qualité du follicule pré-ovulatoire a été identifiée comme étant une des causes majeures des échecs précoces de gestation. Des études réalisées nous montrent que le niveau de PGE2 présent dans le micro-environnement ovocytaire influence durablement la cinétique de division et la capacité de développement des cellules des embryons tant dans les compartiments embryonnaires qu'extra-embryonnaires.

Le projet sur lequel j'ai travaillé durant mon stage de master 2 s'intitule « Embryomimétisme ». Ce projet dont les partenaires sont les unités BREED et PRC (INRAE de Jouy-en-Josas et de Nouzilly), ainsi qu'Allice est financé par la société Apis-Gene (www.apis-gene.com) qui regroupe les professionnels des filières d'élevage des espèces ruminantes (bovin, ovin, caprin). L'objectif du projet est de mimer les environnements naturels de l'ovocyte et l'embryon bovin afin d'améliorer la production des embryons in vitro. Il vise in fine le développement de milieux de culture optimisés, entièrement synthétiques et dépourvus d'additifs d'origine animale. Ce projet a généré des données transcriptomiques et métabolomiques. Mon travail a consisté à analyser ces données en vue de définir le phénotype moléculaire des embryons issus de différents micro-environnement +/- riches en PGE2 et à les corréler aux meilleurs taux de réussite de gestation.

Ce stage a pour objectif d'évaluer plusieurs outils méthodologiques permettant l'analyse statistique des données issues des technologies de séquençage du transcriptome (RNA-seq). D'après plusieurs auteurs, les difficultés de modélisation de données de comptage RNA-seq sont liées à leur caractère discret et au faible nombre d'échantillons disponibles en comparaison du nombre important de variables décrites. Le faible nombre d'échantillons de la plupart de ces études est limité par le coût financier important du séquençage.

Dans un premier temps, je me suis familiarisé avec les notions de biologie de la reproduction liées au projet afin de comprendre la problématique et cerner la démarche de mes analyses ultérieures. Ensuite, il faut fouiller et préparer les données RNA-seq pour permettre l'analyse de ces données au moyen des procédures statistiques dédiées. Les données issues de la technologie de RNA-seq sont des données bruitées issus de l'identification des nucléotides (A,C,T,G) des séquences nucléotidiques. Elles doivent donc être « débruitée » afin de rendre des tableaux de comptages.

Une seconde partie des travaux de ce stage porte sur l'analyse différentielle de données d'expression de gènes à l'aide des package DESeq2 et edgeR. L'objectif de l'analyse différentielle est la détection des gènes dont les expressions changent significativement entre plusieurs conditions expérimentales. Nous allons aussi comparer la performance de ces packages du logiciel R. Dans la dernière partie des travaux de ce stage, nous allons nous focaliser sur l'inférence de réseaux qui a pour but la détection des relations de dépendance entre les gènes basée sur leurs niveaux respectifs d'expression.

# Chapitre 1

# Contexte biologique

## 1.1 Notions de biologie

### ADN

L'acide désoxyribonucléique (ADN) est une molécule présente dans toutes les cellules vivantes, c'est le support de l'hérédité ou de l'information génétique car il constitue le génome des êtres vivants et se transmet en totalité ou en partie lors des processus de reproduction.

L'ADN forme une double-hélice, composée de deux brins complémentaires enroulés l'un autour de l'autre. Chaque brin d'ADN est constitué d'un enchaînement de nucléotides. On trouve quatre différents nucléotides dans l'ADN, notés Adénine(A), Cytosine(C), Guanine(G), Thymine(T) des bases azotées qui les composent. Les nucléotides trouvés dans un brin possèdent des nucléotides complémentaires situés en vis-à-vis sur l'autre brin avec lesquels ils peuvent interagir. Le génotype résulte de l'ordre dans lequel s'enchaînent les quatre nucléotides.

### ARN

L'acide ribonucléique est une copie d'une région d'un des deux brins de l'ADN. Les ARNs sont des molécules constituées par l'assemblage de ribonucléotides reliés entre eux par des liaisons nucléotidiques. L'ordre de cet enchaînement (structure primaire) est dicté par la séquence des désoxyribonucléotides portés par l'ADN à partir duquel il est copié. En effet, l'ARN provient de la transcription de l'ADN par une enzyme (l'ARN polymérase) qui recopie, en quelque sorte, la séquence.

Il y a plusieurs classes d'ARNs. Certains ARNs sont dits codants, d'autres non-codants. Il y a de nombreuses catégories d'ARNs non codant qui ont un rôle de régulateur de la transcription. Nous ne nous attarderons pas avantage sur ces catégories d'ARNs qui ne sont jamais traduits en protéines. À l'inverse les ARN codant sont traduits en protéines.

Il existe de nombreuses familles d'ARN (ARNr, ARNm, ARNt,...) dont chacune possède une structure ou une fonction particulière :

- Les ARN ribosomique (ARNr) entrent dans la composition des ribosomes, avec les protéines ribosomiques.
- Les ARN messagers (ARNm) serviront de matrice pour la synthèse des protéines.

- Les ARN de transfert (ARNt) portent des acides aminés et permettre leur incorporation dans les protéines.

**Gène**

Un gène est un élément génétique correspondant à un segment d'ADN ou d'ARN (virus), situé à un endroit bien précis (locus) sur un chromosome. Chaque région de l'ADN qui produit une molécule d'ARN fonctionnelle est un gène, soit une unité d'hérédité contrôlant un caractère particulier.

L'information génétique portée par l'ADN constitue le génotype d'un organisme qui s'exprime pour donner naissance à un phénotype, cette expression du génome se fait en plusieurs étapes dont nous détaillons ci-dessous :

- **La transcription**, qui consiste à copier des régions dites codantes de l'ADN en molécules d'ARN.
- **La translation**, qui est un transfert d'informations depuis l'ARN vers les protéines.
- **L'activité des protéines**, qui détermine l'activité des cellules et vont ensuite déterminer le fonctionnement des organes et de l'organisme.

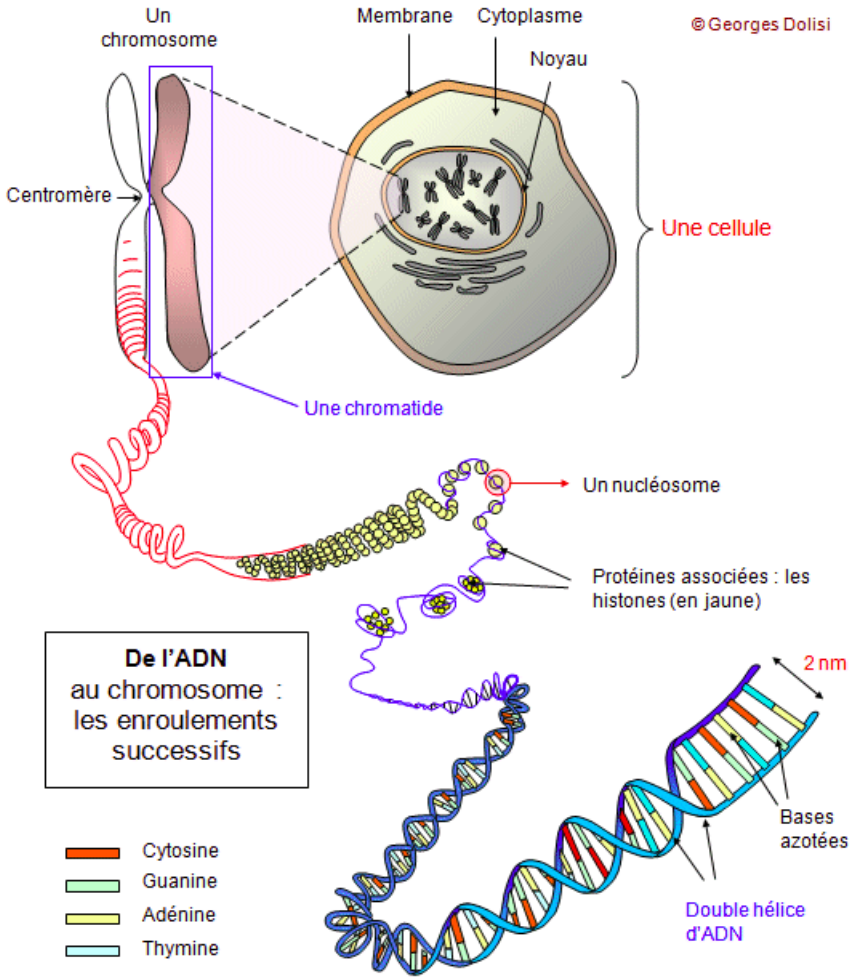



FIGURE 1.1 – L'information génétique portée par l'ADN

## 1.2 La technologie RNA-seq

La technologie RNA-seq est une technologie récente apparue vers 2005. Il s'agit d'une technologie de séquençage à haut débit (séquençage de seconde génération) qui mesure l'abondance des séquences d'ARN dans différentes cellules pour les milliers de gènes simultanément. Autrement dit, cette technologie est faite pour mesurer le transcriptome, c'est-à-dire d'évaluer l'ensemble de toutes les molécules d'ARN produites dans une population de cellules. Le transcriptome peut varier avec les conditions environnementales externes et reflète les gènes qui ont été exprimés de façon active à un temps donné.

Le séquençage de l'ARN consiste à déterminer l'ordre d'enchaînement des nucléotides pour un fragment d'ARN donné. Rappelons que, pour des raisons de la plus grande stabilité des molécules d'ADN comparent à celle d'ARN, l'ARN qui est extrait de la cellule et retranscrit en un brin d'ADN complémentaire (ADNc).

Le principe du séquençage seconde génération consiste à découper l'ADNc en petits morceaux, en découpant X paires de bases (nombre de nucléotides) selon la capacité de la machine utilisée pour le séquençage. Ensuite, des copies des fragments d'AND sont synthétisées en grand nombre par une étape nommée l'amplification qui permet d'avoir plus de matériel pour réaliser le séquençage. Enfin, le séquenceur identifie chaque base nucléotidique en fonction d'une réaction de fluorescence spécifique à chacune d'elles. La machine retourne des bouts de séquence génomique, appelée des *reads*, sous forme d'un fichier au format texte avec des nucléotides (A,T,C,G).

Dans ce stage, nous nous intéressons principalement au traitement des données issues de la technologie RNA-seq.

## 1.3 Modèle expérimental

Des études antérieures ont montré que l'aptitude d'un ovocyte à donner un embryon de bonne qualité est positivement corrélée à la concentration en prostaglandine E2 (PGE2) présente dans l'environnement de l'ovocyte durant la période péri-ovulatoire. La PGE2 est synthétisée à partir de l'acide arachidonique par l'activité de la prostaglandine G/H synthase 2 (PTGS2). Les prostaglandines sont des médiateurs lipidiques de la signalisation cellulaire intervenant à plusieurs étapes de la reproduction des mammifères. Dans notre étude, nous avons utilisé un modèle in vitro pour produire des embryons bovins issus de différents environnements en PGE2 pendant la maturation ovocytaire et la fécondation. Le complexe ovocyte-cumulus (COC) est une structure anatomique constituée d'un ovocyte entouré de cellules somatiques dites du cumulus. Les COCs sont collectés à partir d'ovaires d'abattoir par aspiration des follicules antraux de 3 à 6 mm de diamètre. Ils sont mis en culture pour qu'ils réalisent la maturation puis ils sont soumis à la Fécondation in Vitro (FIV : consiste à réaliser la fécondation en dehors du corps humain ou mammifère). Le NS-398 (8µM), un inhibiteur sélectif de PTGS2 ou 10µM de PGE2 exogène sont ajoutés au milieu de maturation et de fécondation.

Ensuite, la fécondation donne naissance à une cellule diploïde et totipotente, appelée zygote qui est à l'origine des tissus de l'embryon. Le jour de fécondation est considéré comme le jour 0. L'activation génomique embryonnaire (EGA) s'effectue au stade 8-16 cellules chez le bovin, l'organisme qu'on étudie durant notre stage. Elle commence par la reprogrammation épigénétique du génome embryonnaire pour permettre son activation transcriptionnelle. Cette activation a

pour conséquence le passage progressif d'un contrôle maternel à contrôle embryonnaire du développement. L'embryon quitte l'oviducte et entre dans l'utérus au stade 16-32 cellules, 5 à 6 jours après la fécondation.

Au $7^{ème}$ jour de développement embryonnaire survient la formation du blastocyste chez le bovin. Le blastocyste comporte une cavité, appelée blastocèle, qui est rempli d'un liquide cellulaire et est entouré par le trophoblaste (ou trophectoderme) qui donnera le placenta et les autres annexes extra-embryonnaires. La masse cellulaire interne (ICM) formera le futur disque embryonnaire, donnera le fœtus et ses annexes embryonnaires.

Au $15^{ème}$ jour de développement, le conceptus est constitué du disque embryonnaire entouré du tissu extra-embryonnaire (trophoblaste). Les données brutes de RNA-seq utilisées par la suite sont issues de ces deux tissus embryonnaires, nommés "Trophoblaste" et "Disque". Les conceptus J15 dérivés des différentes conditions de culture MIV/FIV seront appelés "inhibé", "stimulé" et "contrôle" pour désigner les traitements respectifs par le NS-398, la PGE2 ou sans traitement. L'objectif principal est d'identifier un impact tardif de la prostaglandine PGE2 périconceptionnelle sur le transcriptome des deux compartiments du conceptus bovin après 15 jours de développement.

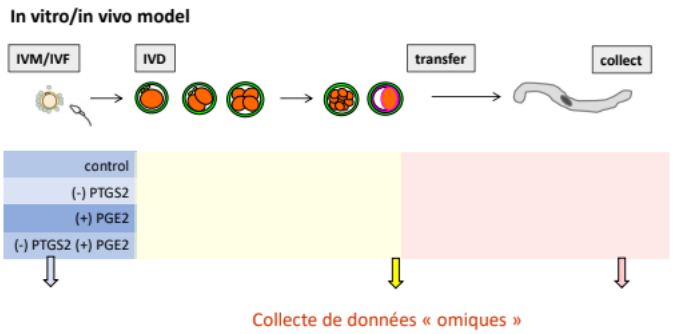



FIGURE 1.2 – Schéma du modèle expérimental

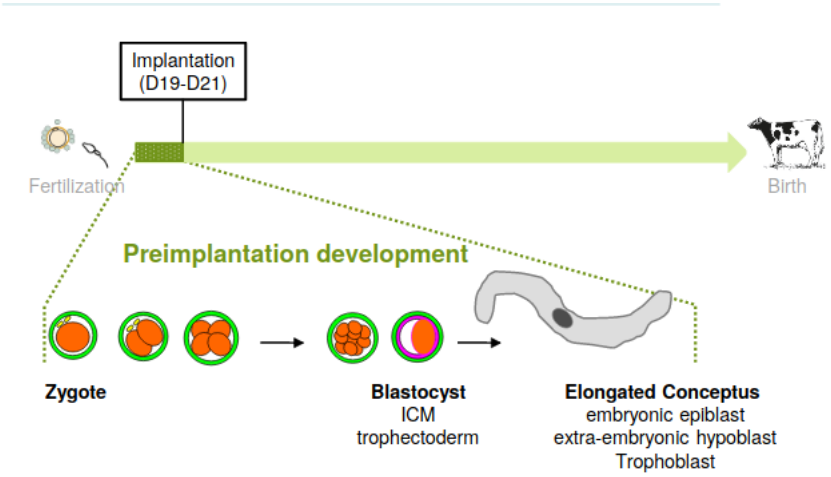



FIGURE 1.3 – Étapes clé du développement pré-implantatoire

# Chapitre 2

# Traitement de données bio-informatique

## 2.1 Description de données brutes

La technologie de séquençage à haut débit nous fournit au final des données dites brutes en format FASTQ, qui peut être ouvert avec un simple éditeur de texte. Ce fichier contient des séquences nucléotidiques ainsi que des valeurs de qualité de séquençage. Cependant ce fichier fastq n'apporte aucune information relative à un génome (ou transcriptome), ces données se présentent sous la forme suivante.

```
@K00201:110:HJ35VBBXX:3:1101:1468:1050 1:N:0:NAGGCATG+TATTACGC
NCTGAGGTGTCCTCTATTCTTGAAGAGCGTATTCTTGGAGCTGATACCTC
+
#AAFFJJFJJFJJJJJJJJJJJJJJJJJJJAJA7J<AJJJJJ7FJJJJJ7J
@K00201:110:HJ35VBBXX:3:1101:1651:1050 1:N:0:NAGGCATG+TCTTACGC
NCACCACAGAAGGTCAGCTTGAACCAAAACAAGCTGAAATAACCTCAAAA
+
#AAFFJJFFJJJFAFJJJJJJJJJJJFJJJJJJJJFJJJJJFJJJJJJJJ
@K00201:110:HJ35VBBXX:3:1101:2849:1050 1:N:0:NAGGCATG+TCTTACGC
NCTGCTGAGTGGCCTGGGACTTGTTTCAAACCTTCTTGTCCCAGCCCTAC
+
#AAFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
```

FIGURE 2.1 – Un fichier fastq

Chaque bloc de quatre lignes, commençant par un "@" correspond un *read* d'un échantillon donné, chacune de ces lignes décrit une structure à laquelle nous détaillons ci-dessous. La première ligne nous donne l'identifiant de l'échantillon :

K00201 :110 :HJ35VBBXX :3 :1101 :1468 :1050 1 :N :0 : NAGGCATG

- K00201 : nom du séquenceur
- 110 : identifiant du run
- HJ35VBBXX : identifiant de la flowcell
- 3 : numéro de ligne
- 1101 : numéro du tile
- 1468 : coordonnée X
- 1050 : coordonnée Y

- 1 : numéro de la paire (1 ou 2)
- N : booléen indiquant le passage du filtre qualité
  - Y : la séquence est de mauvaise qualité
  - N : la séquence a passé le filtre de qualité
- 0 : 0 lorsque aucun des bits contrôlés n'est activé, sinon c'est un nombre.

La deuxième ligne représente la séquence de nucléotides et la dernière est un code composé de lettres et de symboles qui correspond à la fiabilité de lecture de chaque nucléotide du read. La qualité de séquençage sera vérifiée grâce au logiciel FastQC qui nous indiquera si la séquence est de mauvaise ou bonne qualité.

## 2.2 Fiabilité de données

Avant d'aligner les reads, nous devons vérifier la qualité des données brutes des séquences avec le logiciel FastQC pour améliorer la fiabilité de la lecture. Autrement dit, il faut s'assurer d'abord que le séquenceur de la technologie de RNAseq a bien fait son travail. Il s'agit, pour chaque échantillon récupéré d'un fichier avec la liste des reads obtenus.

Le logiciel FastQC nous fournit à partir de chaque fichier de .fastqc un ensemble d'informations et des graphes qui permettent d'évaluer la qualité de séquençage de l'échantillon. La qualité est évaluée sur une échelle de 1 à 40 divisée en trois zones telles que mauvaise, moyenne et bonne. Pour un échantillon donné, on considère l'exemple suivant dont nous présentons le graphe qui détermine la qualité de la lecture des nucléotides en fonction de leur position sur le read grâce à la construction de boîtes à moustaches. La courbe blue est la qualité moyenne en fonction de la position de reads et nous décrit la fiabilité de séquences. Si cette courbe reste globalement dans la zone verte sans atteindre la zone rouge, nous pouvons conclure que la qualité de séquençage de l'échantillon est suffisamment bonne pour passer à l'alignement.

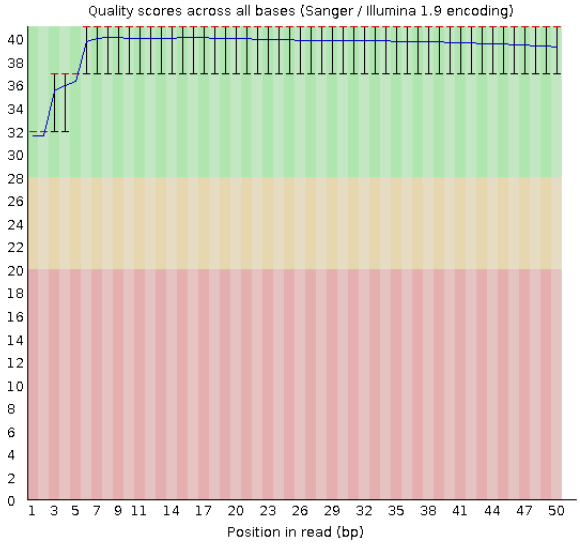



FIGURE 2.2 – Fiabilité de séquençage

## 2.3 Alignement

Une fois que l'on aura vérifié la qualité de séquençage, on pourra aligner les reads de chaque échantillon sur un génome de référence de l'organisme étudié. Rappelant où l'organisme étudié dans le cadre de notre projet est la vache donc on récupère le transcrit de référence auprès de site Ensembl (www.ensembl.org). L'alignement peut se faire sans un génome de référence ni un transcriptome (en utilisant le logiciel Cufflincks et TopHat). Nous nous intéressons dans le cas où le transcriptome de référence est disponible. L'objectif de l'alignement est de transformer les données brutes décrit précédemment en tableau de comptage pour pouvoir le comparer les niveaux d'expression des gènes selon les conditions expérimentales.

En particulier, pour aligner des lectures sur un transcriptome, on assigne une position transcriptomique aux lectures en les alignant directement sur celui-ci. L'identification de transcrits et leurs quantifications, mais aussi la détection de variants dépendent fortement de la qualité de l'alignement sur le transcriptome de référence. Il existe plusieurs méthodes d'alignement. Nous avons choisi les méthodes dite exon-first qui cherchent d'abord à aligner les lectures en un seul bloc sur le transcrit. Cette étape permet de définir les exons (région codante de l'ARNm inclus dans le transcrit). Elles utilisent ensuite les lectures non alignées pour trouver les jonctions entre les exons.

On considère l'exemple suivant dont on prend une ligne de génome de référence et lectures de brin d'ADNc qu'on aimerait aligner :

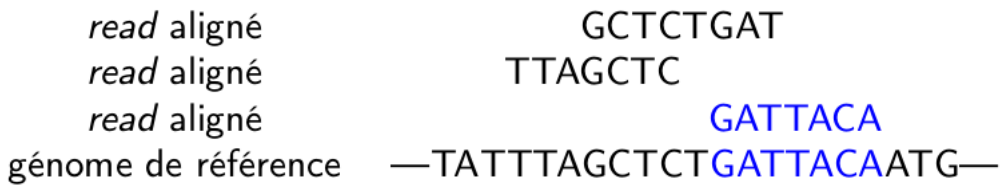
1. Lecture des brins d'ADNc, appelés reads :

   **GATTACA, GTTTTTAGCTG, TAATTAG**

2. Génome de référence :

   **TATTTAGCTCTGATTACAATG**

3. Alignement des reads :




En pratique, nous devons commencer par l'indexation de transcriptome de référence. Le transcriptome dispose plusieurs millions de paires de bases et le nombre de reads à aligner sur ce transcriptome étant de plusieurs millions aussi, donc il faut rendre le processus d'alignement le plus rapide possible. Pour l'indexation, nous utilisons la fonction **bowtie2-bluid** du logiciel Bowtie2. Toutes les fonctions utilisées pour l'indexation et l'alignement seront exécutées en ligne de commande depuis le terminal de l'environnement Ubuntu. Une fois l'index créé, nous pouvons aligner les reads grâce à la fonction **bowtie2** dont nous présenterons dans l'annexe , d'ailleurs tout le code effectués durant ce stage. Finalement, on comptabilise le nombre de reads pour chaque région d'intérêt et après on obtient un tableau de comptage dont on présentera ci-dessous.

**Description de comptage obtenu :**

Pour chaque échantillon, on obtient un fichier contenant des comptages, ce qu'il veut dire le nombre de reads alignés sur chaque gène. Comme on a déjà parlé dans la sous-section 1.3 que le stade de Conceptus J15 dispose deux tissus, trophoblaste et disque, chacun a ses propres échantillons mais ils partagent les mêmes gènes. Le trophoblaste incite 18 échantillons et 21 000 gènes, idem pour le disque 17 échantillons et 21 000 gènes. Nous illustrons un exemple à quoi ressemblent les données que nous allons traiter par la suite. On a les échantillons en colonne et les gènes (identités) en lignes , chaque échantillon compose un numéro d'identifiant (par exemple 1121,1138) plus une lettre T ou D (T1,T2,D1,D2) qui nous indique si cet échantillon est extrait du trophoblaste ou de disque. On utilise le terme d'échantillon pour désigner les réplicats biologiques et l'on précise les différentes conditions expérimentales si besoin. Les réplicats biologiques désignent les échantillons collectés sur des individus différents. Au contraire, il y a les réplicats techniques qui désignent les échantillons collectés sur un même individu pour une même condition expérimentale. Les jeux de données analysées dans ce stage contiennent des réplicats biologie.

| | 1121T3 | 1138T1 | 1121D1 | 1138D6 |
|---|---|---|---|---|
| ENSBTAG00000000522 | 4489 | 2505 | 40 | 3047 |
| ENSBTAG00000017531 | 5265 | 4460 | 543 | 6510 |
| ENSBTAG00000020512 | 126 | 41 | 2 | 57 |
| ENSBTAG00000022759 | 714930 | 538261 | 26865 | 349 |

Table 2.1 – Exemple des comptages

**Description des conditions expérimentales :**
On dispose aussi un fichier .csv pour chaque tissu qui décrit le traitement donné ou pris par chacun d'échantillon, c'est-à-dire les conditions expérimentales des échantillons. Ces fichiers sont des tableaux à 18 échantillons pour le trophoblaste et 17 échantillons pour le disque et une seule colonne qui décrit si la molécule NS398 ou la prostaglandine PGE a été ajouté au milieu de maturation et de fécondation. On note :

- **ns1pg1** : contrôle, l'absence de PGE2 et l'absence du molécule NS398,
- **ns1pg3** : inhibé, la présence de PGE2 mais l'absence du molécule NS398,
- **ns3pg1** : stimulé, l'absence de PGE2 par contre la présence du molécule NS398.

L'objectif général est de savoir l'impact causé par la prostaglandine (PGE2) dans le développement embryonnaire.

# Chapitre 3

# L'analyse différentielle de données RNA-seq

Dans cette partie nous allons étudier l'analyse différentielle qui est le premier bu de mon stage, l'analyse différentielle est généralement décomposé en trois étapes : normalisation de comptage, modélisation d'expression des gènes, tests statistiques.

**Notations**

On considère les échantillons comme des variables (en colonne) et les gènes comme des individus (en ligne). Les jeux de données analysées se présentent sous la forme d'un tableau de comptage de taille $p \times n$, où chaque élément $(i, j)$ $y_{ij}$ correspond au nombre de copie due transcrit du gène $i$, $i \in \{1, ..., p\}$, pour l'échantillon $j$, $j \in \{1, ..., n\}$. On note également $Y_{ij}$ la matrice d'expression.

## 3.1 Statistique descriptive

Avant toute autre chose, il convient de décrire les données afin de se familiariser avec leur distribution. La statistique descriptive permet de résumer et présenter les données observées de la manière la plus pertinente possible. Nous avons fait cette analyse sur l'ensemble de données par échantillon. Elle consiste à calculer des paramètres qui vont résumer nos tableaux de comptages. Ces paramètres appartiennent à deux grandes catégories :

- les paramètres de position (la moyenne, la médiane) qui témoignent d'un niveau.
- les paramètres de dispersion (l'écart type, la variance) qui renseignent sur la répartition des données autour de la moyenne.

Nous devons également vérifier si nos tableaux de comptages contiennent des données manquantes, c'est une information très importante pour réaliser une analyse. Il est éventuel d'éliminer la variable, respectivement l'individu qui a des valeurs manquantes ou remplacer par la moyenne, respectivement la médiane des valeurs observées pour cette variable. Heureusement, il n'y a pas des données manquantes dans nos jeux des données.

On considère l'exemple suivant une simple description (la fonction summary de R dont nous avons ajouté une colonne qui donne la variance et l'autre le nombre de zero) sur les données des embryons du stade J15 de disque.

| | min | Q1 | median | mean | Q3 | max | variance | zero |
|---|---|---|---|---|---|---|---|---|
| 1121D1 | 0 | 3 | 125 | 1777.481 | 1148.00 | 437982 | 65798400 | 4058 |
| 1134D1 | 0 | 2 | 89 | 1337.104 | 832.00 | 302616 | 33050255 | 4383 |
| 1138D5 | 0 | 2 | 101 | 1444.025 | 827.75 | 441624 | 51843515 | 4273 |
| 1138D6 | 0 | 2 | 92 | 1482.625 | 860.00 | 332401 | 47169240 | 4512 |
| 1138D7 | 0 | 3 | 98 | 1550.138 | 962.75 | 342097 | 48191531 | 3851 |
| 1140D4 | 0 | 2 | 91 | 1548.286 | 977.75 | 311967 | 42806024 | 4183 |

FIGURE 3.1 – Exemple d'un Summary

**Remarque :**

Dans tous les échantillons, des milliers de gènes sont complètement non détectés (zéros). On trouve aussi qu'il y a une énorme différence entre la tendance centrale dont la médiane est bien inférieure à la moyenne, ce qui indique une asymétrie à droite et le maximum reflète la présence de valeurs aberrantes, ce qu'il veut dire quelques gènes ont des centaines de milliers de comptages. Le nombre moyen par gène montre des fluctuations importantes, indiquant la nécessité d'une normalisation. Finalement, on constate que la variance est généralement plus élevée que la moyenne.

Comme nos données sont des données discrètes, on y attend que nos comptages suivent une distribution discrète, soit la loi de Poisson, la loi de Poisson sur-dispersés ou loi binomiale négative...etc. Rappelant la définition d'une loi de Poisson, $y \sim P(u)$ de paramètre u, si la moyenne et la variance sont égaux, ce qui n'est pas le cas. Lorsque dans un jeu de données la variance excède la moyenne, cette distribution est dite sur-dispersée par rapport au Poisson par contre si on a le contraire, on se trouve dans le cas sous-dispersée.

On en conclut que nos données suivent une loi binomiale négative (possède deux paramètres r et p) car c'est une alternative intéressante à la loi de Poisson et elle est particulièrement utile pour les données discrètes dont la variance empirique excède la moyenne empirique.

## 3.2 Filtrage

Dans une étude d'analyse différentielle des données RNA-seq, le filtrage est une étape préliminaire et motive la nécessité de réduire la taille du jeu de données afin d'améliorer la puissance de détection des tests d'analyse différentielle. Il s'agit de retirer les gènes dont le profil d'expression ne paraît pas informatif, autrement dit nous allons supprimer les gènes qui n'apportent aucun intérêt à notre analyse statistique. Il existe plusieurs méthodes de filtrage utilisées régulièrement pour les données RNA-seq, par exemple des méthodes qui sont basées sur le choix d'un seuil fixé arbitrairement. Cependant nous avons utilisé une méthode de filtrage inventé par Andrea Rau (inra.fr) et ses collègues en 2013 qui s'appuie sur un seuil calibré à l'aide d'un indice de Jaccard qui mesure la similarité entre les différents réplicats d'une même condition. Nous considérons deux réplicats biologiques $j$ et $j^{'}$ appartenant à une même condition expérimentale, un seuil fixé s. Pour un seuil donné, on calcule un indice de jaccard, coefficient de similarité entre les deux réplicats $j$ et $j^{'}$ à partir d'un tableau de contingence.

| | $\forall i$ tels que $y_{ij} > s$ | $\forall i$ tels que $y_{ij} \leq s$ |
|---|---|---|
| $\forall i$ tels que $y_{ij'} > s$ | $M_1$ | $M_2$ |
| $\forall i$ tels que $y_{ij'} \leq s$ | $M_3$ | $M_4$ |

TABLE 3.1 – Tableau de contingence du nombre de gènes

$$J_s(y_j, y_{j'}) = \frac{M_1}{M_1+M_2+M_3}$$

Notons $C$ le cardinal de l'ensemble des couples d'échantillons possibles appartenant à une même condition. On calcule la moyenne de ces coefficients sur l'ensemble des couples d'échantillons appartenant à une même condition possible :

$$J_s^*(y) = \frac{1}{C}\Sigma_{j<j'} J_s(y_j, y_{j'}) \tag{3.1}$$

Le seuil calibré est celui qui maximise la quantité $J_s^*(y)$ :

$$s^* = \text{argmax}_s \, J_s^*(y)$$

Le choix du seuil selon cette méthode garantit que la similarité entre les profils d'expression des échantillons d'une même condition expérimentale soit la plus forte possible. Cette méthode est implémentée R dans le package **HTSFilter** disponible sur Bioconductor.

Nous prenons un exemple pour savoir la raison dont laquelle le filtrage est si important avant d'aller plus loin sur les analyses statistiques. Le gène 2 comporte un alignement de 24 reads pour l'échantillon $1120T1$ de la condition ns1pg1, et uniquement pour cet échantillon. On est conscient que ce gène ne semble pas apporter d'information et pourra difficilement être considéré comme différentiellement exprimé entre ns1pg1 et ns1pg3. Donc il serait retiré de tableau de comptage avant d'effectuer l'analyse différentielle.

| • | gène 1 | gène 2 | condition |
|---|---|---|---|
| 1120T1 | 155 | 24 | ns1pg1 |
| 1140T2 | 2403 | 0 | ns1pg3 |
| 1138D3 | 0 | 0 | ns1pg3 |
| 1121T3 | 25 | 0 | ns1pg1 |
| 1141T5 | 893 | 0 | ns1pg1 |
| 1151T3a | 5 | 0 | ns1pg3 |

TABLE 3.2 – Extrait d'un tableau de comptages

## 3.3 Normalisation

La normalisation permet de corriger les biais techniques systématiques et rendre les comptages comparables entre les différents niveaux d'expression mesurés. Un grand nombre de méthodes de normalisation ont été développées et utilisées, plusieurs statisticiens en ont fait des revues et ont conclu qu'il y a pas une méthode universelle mais plutôt des méthodes adaptées à chaque cas. Nous allons présenter deux méthodes de normalisation qui sont souvent utilisées pour les données RNA-seq et que nous allons appliquer pour nos données des embryons.

### 3.3.1 RLE

La méthode Relative Log Expression est développée par Andres et Hubers dans le package Bioconductor **DESeq2**, elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimés. Il s'agit d'utiliser, pour chaque échantillon $j$, un facteur RLE qui sera utilisé en coefficient multiplicateur pour corriger le nombre de comptage dans l'échantillon. Ce facteur est obtenu en calculant pour chaque gène $i$ la médiane des ratios de ses comptages par rapport à sa moyenne géométrique entre les différents échantillons.

- Soit $M_i$ la moyenne géométrique des comptages $Y_{ij}$ de chaque gène $i$ au travers de tous les échantillons, notant $n$ le nombre d'échantillons :

$$M_i = (\Pi_{j=1}^{n} Y_{ij})^{1/n} \tag{3.2}$$

On obtient ainsi un pseudo échantillon de référence $j^{'}$ composé de ces comptages moyens.

- Le ratio pour chaque comptage sera :

$$\frac{K_{ij}}{Y_{ij^{'}}}$$

- Enfin, on calcule, pour chaque échantillon la médian $S_j$ du ratio d'échantillons :

$$S_j = median(\frac{Y_{ij}}{M_i}) \tag{3.3}$$

Cette médiane sera notre facteur RLE par lequel tous les comptages seront multipliés pour rendre les échantillons comparables et corriger le biais de profondeur de séquençage. Cette méthode de normalisation utilise la moyenne géométrique qui est plus robuste que la moyenne arithmétique car moins sensible aux valeurs extrêmes. Elle donne donc une meilleure estimation de la tendance centrale des données.

### 3.3.2 TMM

La méthode Trimmed Mean of M-values est développé par Robinson et Oshlack(2010) dans le package bioconductor **edgeR**. le facteur d'échelle $s_j^{TMM}$ sera utilisé en coefficient multiplicateur pour corriger les biais techniques de comptage, il est calculé, pour chaque échantillon, l'un d'eux étant considéré comme l'échantillon de référence et les autres comme des échantillons teste. Pour chaque échantillon teste le facteur d'échelle est la moyenne pondérée des logs-ratios entre ce test et la référence.

Soient $p$ gènes et $n$ échantillons, $y_{ij}$ et $y_{ir}$ les comptages de gène $i$ et d'échantillon $j$, respectivement $r$ (référence). $s_j^{TMM}$ est le facteur de TMM et se définit comme suit :

$$log_2(s_j^{TMM}) = \frac{\Sigma_{i\in p} w_{ij} M_{ij}}{\Sigma_{i\in p} w_{ij}} \tag{3.4}$$

où

$$M_{ij} = log_2(\tfrac{y_{ij}/N_j}{y_{ir}/N_r}),$$

$$w_{ij} = \tfrac{N_j - y_{ij}}{N_j y_{ij}} + \tfrac{N_r - y_{ir}}{N_r y_{ir}},\ y_{ij}, y_{ir} > 0$$

$N_j, N_r$ sont les nombres totaux de reads pour l'échantillon test et référence.

Remarque :
- les facteurs de normalisation fournis par TMM s'appliquent aux nombres totaux de reads, pas au comptage.
- S'il y a peu de gènes différentiellement exprimés, le facteur de l'échelle $s_j^{TMM}$ doit être proche de 1 donc $log_2(s_j^{TMM}) = 0$

On obtient les facteurs de normalisation suivante pour RLE et TMM.

| • | 1121T3 | 1134T4 | 1138T1 | 1151T2 | 1140T1a |
|---|---|---|---|---|---|
| RLE | 0.5948730 | 0.8276890 | 0.5264549 | 1.2203183 | 1.3904211 |
| TMM | 0.6008421 | 0.7689937 | 0.6240440 | 0.9729026 | 1.2477488 |

TABLE 3.3 – Facteurs de normalisation

## 3.4 Détection des gènes différentiellement exprimés

La détection des gènes différentiellement exprimés à pour objectif notamment d'identifier les gènes dont les expressions changent significativement entre deux conditions ou plus. Le but peut être la recherche de l'identité des gènes, dits marqueurs, spécifiques d'un tissu, d'un organe ou spécifiques d'un dysfonctionnement (maladie). Nous allons présenter ci-dessous la modélisation, la correction de tests statistiques et les packages utilisés pour détecter le nombre de gènes différentiellement exprimés.

### 3.4.1 Modélisation

Après avoir normalisé les données, il convient de modéliser l'expression des gènes dans chaque condition par une variable aléatoire. Pour un échantillon donné, le nombre de *reads* alignés dépend fortement la profondeur de séquençage qui est une contrainte spécifique de la technologie RNA-seq. Sachant $N_j = \Sigma_{i=1}^{p} y_{ij}$, le vecteur de l'échantillon $Y_j$ modélisant la distribution des *reads* sur les $n$ régions transcriptomiques dus transcrivent pour l'échantillon $j$ peuvent être modélisées par une loi multinomiale des paramètres $N_j$ et $(\pi_{1j}.....\pi_{pj})$.

$$p(Y_{1j} = y_{1j}, ..., Y_{pj} = y_{pj}) = \frac{N_j}{y_{1j}!....y_{pj}!}\pi_{1j}.....\pi_{pj} \tag{3.5}$$

Chaque composante $Y_j$ suite une loi binomiale de paramètre $N_j$ et $\pi_i$. On remarque que le nombre de *reads* alignés est élevé et $\pi_i$, la proportion de *reads* alignés sur le gène $i$ est petite. On peut approcher une loi binomiale à une loi de Poisson :

$$Y_{ij} \sim P(\lambda_{ij}),$$

$$p(Y_{ij} = y_{ij}) = \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}} e^{-\lambda_{ij}},$$

$$E(Y_i j) = V(Y_i j) = \lambda_i j.$$

Rappelant dans la section 3.1, nous avons prouvé que la variance excède la moyenne cela implique que la loi de Poisson ne peut pas modéliser nos données. En particulier, le modèle de Poisson est utile pour l'analyse de réplicats technique d'une même condition d'après Marioni et al. Par contre, il modélise mal la grande variabilité inter-échantillon observée dans les données RNA-seq lorsqu'on dispose de réplicats biologie. Afin de prendre en compte cette grande variabilité, nous proposons les modèles de loi de Poisson sur-dispersée ou de loi binomiale négative comme solution alternative au modèle de Poisson :

$$Y_{ij} \sim NB(\lambda_{ij}, \sigma_{ij}^2), \tag{3.6}$$

$$p(Y_{ij} = y_{ij}) = (\tfrac{\varphi_i}{\varphi_i + \lambda_i j})^{\varphi_i} \tfrac{\Gamma(\varphi_i + y_{ij})}{y_{ij}!\Gamma(\varphi_i)} (\tfrac{\lambda_{ij}}{\varphi_i + \lambda_{ij}})_i^y j,$$

$$E(Y_{ij}) = \lambda_{ij},$$

$$\sigma_{ij}^2 = \lambda_{ij} + \lambda_{ij}^2 \varphi_i \ .$$

$\varphi_i$ est le paramètre de dispersion du gène $i$. Il convient de modéliser le paramètre $\lambda_{ij}$ de manière appropriée, par exemple en supposant que $\lambda_{ij} = g_i s_j$ où $s_j$ est un paramètre spécifique à l'échantillon $j$ et $g_i$ est un paramètre spécifique au gène $i$.

### 3.4.2 Test statistique

Pour chaque gène $i$, l'analyse différentielle détermine si une différence d'expression est observée entre deux ou plusieurs conditions expérimentales à l'aide de tests d'hypothèse. Afin de mieux comprendre les notations, nous considérons la comparaison de deux conditions expérimentales uniquement condition *ns1pg1* et condition *ns1pg3*. Le test d'hypothèse détermine s'il existe une différence entre la moyenne $\lambda_i^{ns1pg1}$ d'expression du gène $i$ dans la condition *ns1pg1* et la moyenne $\lambda_i^{ns1pg3}$ d'expression du gène $i$ dans la condition *ns1pg3* :

$$H_0 : \lambda_i^{ns1pg1} = \lambda_i^{ns1pg3} \; vs \; H_1 : \lambda_i^{ns1pg1} \neq \lambda_i^{ns1pg3} \tag{3.7}$$

pour tout gènes $i = 1, ...., p$

De plus, il faut prendre en compte le fait qu'un grand nombre de tests statistiques sont effectués, ce qui augmente considérablement la probabilité de faire une erreur d'identification des gènes différentiellement exprimés. Par exemple, on choisit un taux d'erreur de 5%, nous testions 20 000 gènes et que tous soient non différentielle exprimés, ce qui est inacceptable. Cependant la solution sera d'utiliser une p-value ajustée qui est adaptée aux tests multiples, autrement dit on utilise une méthode qui contrôle cette probabilité d'erreur.

Les méthodes les plus utilisées sont :
- Benjamini-Hochberg (BH) ont proposé une mesure appelée **FDR** pour False Discovery Rate, qui permet de contrôler le taux d'erreurs attendu. Cette mesure est basée sur l'idée que l'on peut tolérer plus d'erreur lorsque le nombre de tests est grand. Soit $pval_i$ la p-value associée au gène $i$, les gènes étant classés par p-value croissante. Les gènes déclarés différentiellement exprimés au taux $\beta$ fixé sont tous ceux indices par $i \leq i_0$ ou $i_0$ est le plus grand indice $i^{'}$ tel que :

$$pval_{i'} \leq \frac{\beta}{p} i^{'}$$

- La procédure de Bonferroni qui consiste à contrôler le Family Wise Error Test (FWER), c'est-à-dire la probabilité d'avoir au moins un faux positif. Pour garantir $FWER \leq \beta$, on réalise chacun des tests à un taux de $\frac{\beta}{G}$ avec G le nombre de gènes testés.

Pour ajustée la p-value, nous avons choisi d'utiliser la méthode de Benjamini-Hochberg (BH).

Pour le modèle de loi binomiale négative, nous avons utilisé les deux méthodes proposées par Love et al (2014) et Robinson et Oshlack qui sont disponibles respectivement dans les packages R de bioconductor **DESeq2** et **edgeR**. L'intérêt d'appliquée ces packages à nos données est d'estimer le paramètre de dispersion de la loi binomiale négative, nous expliquons l'utilisation de ces deux méthodes ci-dessous, dont on a déjà présenté ces méthodes de normalisation dans la section 3.3.

### 3.4.3 DESeq2

C'est une évolution de package **DESeq** dans laquelle le test exact est remplacé par un test utilisant le modèle linéaire généralisé, il modélise la tendance moyenne-variance afin d'estimer ce paramètre de dispersion. L'objectif de ce package est de chercher les gènes différentiellement exprimés en se basant sur un modèle de distribution binomiale négative. Nous modélisons $\lambda_{ij}$ de (3.6) en supposant :

$$\lambda_{ij} = w_{ij} u_j,$$

$$log(w_{ij}) = \beta_i D_j.$$

La moyenne $\lambda_{ij}$ est composée d'un paramètre spécifique à la taille de la librairie $u_j$ et d'un paramètre $w_{ij}$ proportionnel à l'expression du gène $i$ dans l'échantillon $j$. $\beta_i$ est un vecteur qui modélise les variations de l'expression du gène $i$ en fonction des conditions expérimentales de chaque échantillon résumé dans la matrice de design D.

L'utilisation de ce package est décomposé en trois étapes :

- La construction de l'objet DESeqDataSet contenant une table de comptage et également un tableau décrivant le plan d'expérience nomme *coldata* (conditions expérimentales).
- L'application de la fonction DESeq() qui réalise la normalisation et l'analyse différentielle en une seule étape.
- Pour les résultats de l'analyse, nous utilisons la fonction results(..,contrast) dont l'argument contrast permet d'indiquer les deux conditions comparées lorsqu'on dispose plusieurs conditions expérimentales, c'est le cas de notre plan d'expérience. Pour chaque gène nous obtiendrons un tableau qui contient la p-value, la p-value ajustée par la méthode de Benjamini-Hochberg, le logarithme de Fold Change (log2FoldChange) où Fold Change est le ratio entre deux niveaux d'expérience.

Finalement, nous pouvons extraire une liste des gènes différentiellement exprimés avec la fonction subset, en fixant le pourcentage de déclarations comme par exemple 5% et les sélectionnant par rapport à leur p-value ajustée (padj).

**Simulation**

Nous réalisons l'analyse différentielle sur le jeu de données RNAseq des embryons du stade J15 dont on dispose deux tissus le trophoblaste et le disque. Le trophoblaste, respectivement le disque possède 17 et 18 replicats biologie (échantillons) et 21 000 gènes, et nous effectuons également l'analyse de deux tissus séparément. L'intérêt de faire l'analyse différentielle est de détecter les gènes qui expriment différemment dans les différentes conditions. Les tableaux suivants nous indiquent le nombre de gènes déclarés différentiellement exprimés à 5% (autrement dit on sélectionne les gènes dont leur *p_value* ajustée est inférieur à 0.05). Nous avons comparé deux à deux les conditions expérimentales, ce qui veut dire, on distingue *ns1pg1* vs *ns1pg3* (groupe 1) ,*ns1pg1* vs *ns3pg1* (groupe 2) et *ns1pg3* vs *ns3pg1* (groupe 3).

Nous trouvons ci-dessous deux tableaux contenant le nombre de gènes différentiellement exprimés pour le trophoblaste, respectivement le disque.

| • | ns1pg3 | ns3pg1 |
|---|---|---|
| ns1pg1 | 42 | 976 |
| ns3pg1 | 199 | • |

TABLE 3.4 – GDE de Tropho

| • | ns1pg3 | ns3pg1 |
|---|---|---|
| ns1pg1 | 32 | 273 |
| ns3pg1 | 53 | • |

TABLE 3.5 – GDE de Disque

**Remarque :**

Tout d'abord, on constate que le trophoblaste possède plus de gènes différentiellement exprimés que le tissu de disque embryonnaire. Ensuite, en termes biologie lorsqu'on ajoute la prostaglandine (PGE2), on parle le groupe 1 et 3. Il nous semble qu'il y a moins de gènes DE (exemple pour le tropho : DE groupe 1=42 et DE groupe 3=199) par contre le nombre de gènes DE du groupe 2 (l'absence de PGE2) est plus élevé que les deux autres groupes quel que soit le tissu.

Pour en savoir un peu plus sur l'impact de PGE2 dans le développement embryonnaire, nous représentons un diagramme de Venn pour montre, est-ce que ces sont les mêmes gènes qui déclarent différentiellement exprimer entre les différents groupes. le Diagramme de Venn représente un schéma des intersections entre différentes listes de gènes différentiellement exprimés en prenant les identités de ces gènes. On voit que 976 gènes déclarés DE à l'absence de la prostaglandine, 104 de ces gènes s'expriment différemment dans le groupe 3, ce qu'il veut dire $9,6\%$ de gènes communs entre le groupe 2 et 3. On peut en tirer que la prostaglandine a affecté ces 104 gènes $(9,6\%)$. On aperçoit que le nombre de gènes différentiellement exprimés pour le groupe 1 et 3 qu'il y a la présence de la prostaglandine sont trop affaiblis par rapport au groupe 2 donc la PGE2 impacte bien les gènes déclarant différentiellement exprimés, cela implique que la prostaglandine influence le développement embryonnaire.

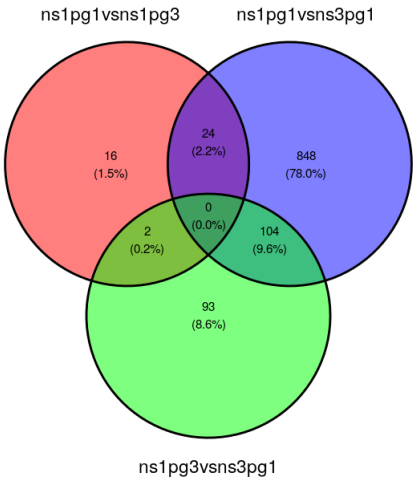



FIGURE 3.2 – Diagramme de Venn pour le trophoblaste

Après avoir identifié les gènes différentiellement exprimés et préparer des listes de gènes, les biologistes s'intéressent à donner sens à ses listes de gènes exprimés, ils utilisent PANTHER (http ://www.pantherdb.org/geneListAnalysis.do) dont on peut trouver la famille et le nom de chaque gène. J'ai personnellement essayé de travail sur ce site pour en savoir un peu plus des identités de gènes exprimés.

### 3.4.4 edgeR

edgeR est spécifié pour détecter les gènes déclarés différentiellement exprimés, se base sur une méthode Bayésienne empirique et implémente un test exact ou un modèle linéaire généralisé fondés sur la loi binomiale négative. En particulier, la procédure de Bayésienne empirique est utilisée pour modérer le degré de sur-dispersion entre les gènes en empruntant des informations entre eux. Le test permet d'évaluer l'expression différentielle de chaque gène. Il utilise un compromis entre une dispersion commune à tous les gènes et une dispersion spécifique à chaque gène.

La modélisation de $\lambda_{ij}$ est similaire à celle du package **DESeq2** sauf qu'on a ajouté le logarithme de $N_i$, sachant $N_j = \Sigma_{i=1}^{p} y_{ij}$ donc on a :

$$log(\lambda_{ij}) = \beta_i D_j + log(N_j) \tag{3.8}$$

Nous détaillons ci-dessous les fonctions R de ce package qui nous a permis d'identifier les gènes différentiellement exprimés :

- Pour commencer, on doit créer un objet *DGEList* qui contient les données de comptage et la variable correspondant les conditions expérimentales.

- La fonction calcNormFactors réalise la normalisation en choisissant la méthode **TMM** décrit dans la section 3.3.2, estimateDisp effectue l'estimation de dispersion.

- Le modèle linéaire généralisé(GLM) est une extension de modèle linéaire classique aux données qui ne suivent pas une loi normale, afin d'appliquer ce modèle on utilise la fonction glmFit et glmLRT effectue des tests de rapport de vraisemblance pour un ou plusieurs coefficients dans le modèle linéaire généralisé.

- Finalement les résultats de l'analyse sont accessibles via la fonction topTags, elle nous fournit le logarithme de Fold Change (logFC), log de comptage par millions, la p-value et la p-value ajustée par la mesure de FDR (FDR). Nous adoptons les gènes différentiellement exprimés par la fonction subset pour sélectionnant par rapport à leur **FDR**.

**Résultats :**

Nous effectuons l'analyse différentielle la même manière que le package **DESeq2**, en suivant la méthode du package **edgeR** décrit ci-dessus. les tableaux suivants nous indiquent le nombre de gènes différentiellement exprimés à 5%. On a juste changé la méthode mais les données d'expressions et les conditions expérimentales restent les mêmes. Dans le tableau du trophoblaste, on a la stabilité, c'est-à-dire que le groupe 2 est toujours plus élevé que les deux autres groupes (1 et 3) dont on sent la présence de la prostaglandine. Par contre le tableau de résultat de disque a changé, nous parlerons plus tard dans la sous-section suivante pour comparer ces deux packages.

| • | ns1pg3 | ns3pg1 |
|---|---|---|
| ns1pg1 | 33 | 859 |
| ns3pg1 | 48 | • |

TABLE 3.6 – GDE du package edgeR Tropho

| • | ns1pg3 | ns3pg1 |
|---|---|---|
| ns1pg1 | 2 | 38 |
| ns3pg1 | 98 | • |

TABLE 3.7 – GDE du package edgeR, Disque

### 3.4.5 Comparaison de deux packages

Tout d'abord, nous citons les similitudes et les différences entre les deux packages (**edger** et **DESeq2**). Ils utilisent la même loi qui est la loi binomiale négative, appliquent aussi la méthode linéaire généralisée. Les points qui diffèrent ces packages sont : l'estimation de la dispersion, la gestion des comptages outliers (aberrants) et le filtrage des faibles comptages. Il y a plusieurs approches qui nous permettent de choisir laquelle de ces deux méthodes semble meilleure pour déterminer ou identifier le nombre de gènes déclarant différentielle exprimés. La première méthode dans laquelle nous avons utilisé, c'est la courbe ROC, elle revient à dessiner un graphique avec ordonnée True Positive Rate (sensitivity) et en abscisse False Positive Rate (1-specificity où specificity= True Negative Rate). Le sensitivity est égal à S/R, c'est-à-dire sur le nombre d'hypothèses rejetées alors que H1 est vrai. Le False Positive rate, c'est le risque de première espèce, c'est-à-dire le nombre de tests qui rejette l'hypothèse H0 alors que celle-ci est vraie.

Nous prenons par exemple la comparaison du groupe 1 (ns1pg1 vs ns1pg3) des données RNA-seq du stade Conceptus J15 du trophoblaste. Pour DESeq2, on a obtenu 42 gènes déclarés DE. Pour edgeR, on a détecté 33 gènes déclarés DE. On a 15 gènes différentiellement exprimés communs aux deux méthodes. Ensuite, on choisit 200 gènes aléatoires, on construit un tableau contenant 3 variables : les deux premières variables sont vecteurs de p-values ajustée de deux packages et la 3è*me* variable décrit le statut des 200 gènes sélectionnés (DE ou non DE), elle est binaire, on donne 1 si les gènes DE, sinon 0.

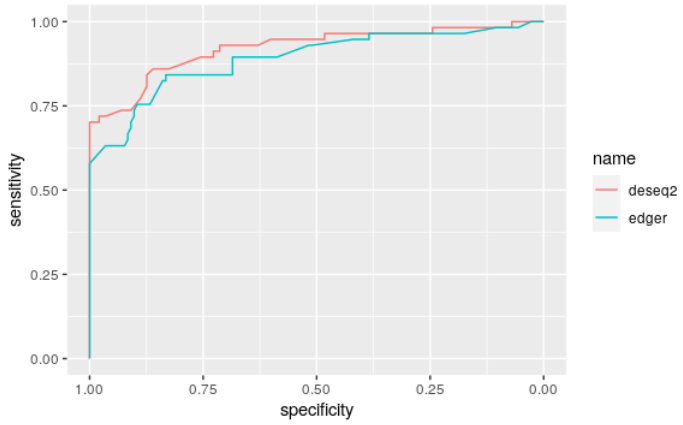



FIGURE 3.3 – La courbe ROC

On voit que la courbe ROC de DESeq2 est au-dessus de celle edger, l'AUC (Area under the curve) qui signifie l'aire sous la courbe ROC. Par définition, l'AUC mesure l'intégralité de l'aire à deux dimensions situées sous l'ensemble de la courbe ROC (par calcul d'intégrales) de 0 à 1. Pour DESeq2, on obtient 0.9217 auc. Pour edgeR on a 0.8903 auc. On constate qu'il y a une différence de 0.0314 entre les deux packages. On peut en conclure pour le groupe 1, DESeq2 donne une meilleure approche que celle de edgeR. On ne s'autorise pas en général à dire définitivement laquelle des deux méthodes donne les meilleurs résultats pour nos données car si on applique avec les données de tropho du groupe 2 (ns1pg1 vs ns3pg1), on trouve qu'edgeR est mieux que DESeq2. Donc la courbe ROC nous précise seulement la performance d'un package pour un groupe donné mais elle ne généralise pas l'efficacité des méthodes pour toutes nos données ou groupes de comparaison.

Nous allons éclaircir que les deux méthodes donnent souvent des résultats proches par la régression linéaire simple en expliquant la liaison entre les vecteurs p-values ajustées des deux packages pour l'ensemble des gènes donc on aura fait affaire un tableau de 12966 gènes et 2 variables.

Le $R^2$ ajusté entre les deux vecteurs de p-valeurs ajustées EdgeR et DEseq2 est de 0.946, cela nous montre que les résultats de ces deux packages sont très significatifs et proches. On a aussi présenté un graphique pour comprendre la liaison dont nous suggérons.

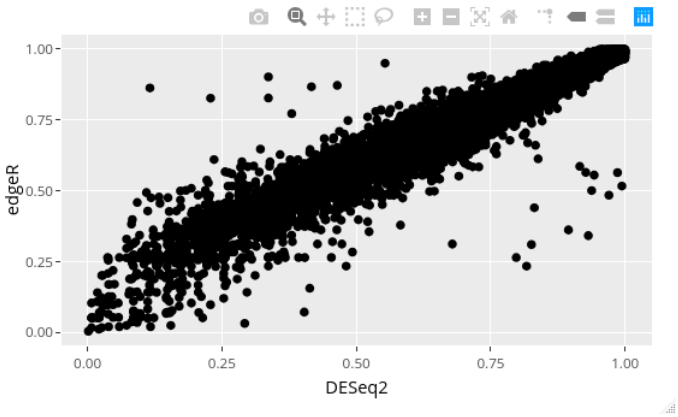



FIGURE 3.4 – DESeq2 en fonction d'edgeR

# Chapitre 4

# Classification Ascendante Hiérarchique

La classification ascendante hiérarchique a pour but de construire une hiérarchie sur les individus et se présente sous la forme d'un dendrogramme. Il s'agit de regrouper les individus les plus proches afin de former des classes qui elles-même seront regroupées selon leur distance pour former de nouvelles classe et ainsi de suite jusqu'à l'obtention d'un dendrogramme. Ici, nous utilisons une distance en matière de corrélation. Ainsi, plus les classes sont éloignées moins elles sont corrélées. Donc l'objectif est de vérifier que les individus que l'on suppose à priori être proches le sont effectivement.

En classification des données RNA-seq, certains statisticiens classent les réplicats biologiques et non les gènes, par contre nous nous intéressons la classification des gènes exprimés donc on considère les gènes comme des individus et les échantillons comme des variables. Autrement dit, il s'agit de grouper les gènes ayant des profils d'expression similaires à travers les différentes conditions biologie, on nomme cette analyse, l'analyse de co-expression. Particulièrement, elle est intéressante dans le cas où plus de deux conditions expérimentales sont étudiées, qui est notre cas, car nous disposons trois conditions biologie (*ns1pg1*,*ns1pg3*,*ns3pg1*).

L'analyse de gènes co-exprimés permet d'aller plus loin que l'analyse différentielle dans l'étude du transcriptome. La raison est simple, l'analyse différentielle essaie de comparer les conditions en groupe, ce qui veut dire elle compare deux à deux les conditions (par exemple comme on a déjà bien étudié dans la section précédant la comparaison de groupe *ns1pg1 vs ns1pg3*, *ns1pg3 vs ns1pg3*) et elle nous fournit simplement le nombre des gènes différentiellement exprimés. Par contre, l'analyse de gènes co-exprimés nous indique la répartition de ces gènes en fonction de leurs conditions expérimentales. Le but précis en terme biologie est de regrouper les gènes impliqués dans un même processus biologique.

Nous illustrons un exemple de classification ascendante hiérarchique étudié sur le jeu de données des embryons du stade J15 des échantillons de disque, on dispose 21000 gènes et 17 échantillons mais nous ne pourrons pas directement classer tous ces gènes. Nous avons extrait 20 gènes qui ont une variance très grande. On a utilisé le package R **pheatmap**.

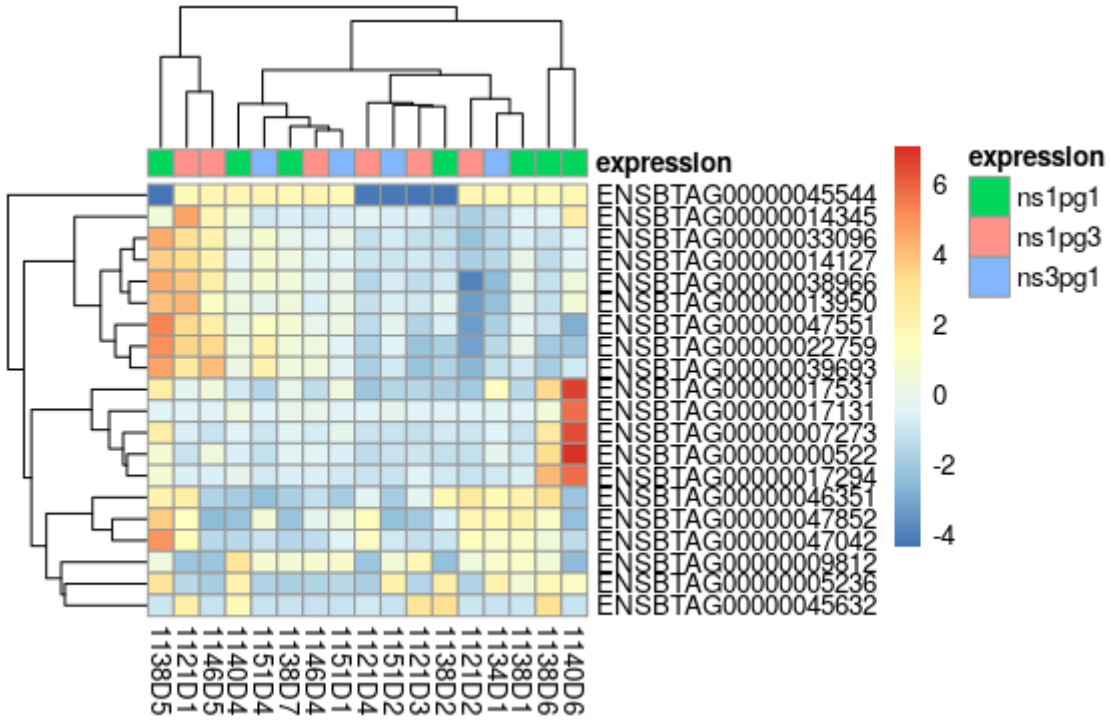



FIGURE 4.1 – CAH de Disque pour les 20 gènes top

# Chapitre 5

# Analyse en Composantes Principales

Tout d'abord, l'objectif d'une Analyse en Composantes Principales (ACP) est de résumer et visualiser un tableau de données croisées individus×variables. Dans le cadre de notre travail, nous considérons l'ACP du tableau échantillons×variables où les échantillons sont considérés comme les individus et les gènes comme les variables. L'ACP permet d'étudier les ressemblances entre les échantillons du point de vue de l'ensemble des gènes et éclaircit les profils des échantillons. Elle permet également de réaliser un bilan des liaisons linéaires entre variables (gènes) à partir des coefficients de corrélation.

La question biologique est d'identifier les principales sources de variation dans les données et de déterminer si ces sources correspondent au modèle expérimental ou si elles sont dues aux seules variations individuelles. L'objectif est ensuite d'identifier les variables (gènes) clés qui contribuent à expliquer le modèle expérimental et qui sont à l'origine delà plupart de la séparation des individus.
Nous nous intéressons à l'utilisation de l'ACP en tant qu'outil de visualisation des données RNA-seq normalisée par les deux méthodes de normalisation étudiées précédemment ou plus simple la transformation logarithmique (**rlog**).

Par exemple, on veut faire l'ACP d'un jeu de données des embryons du stade de conceptus J15 du tissu trophoblaste dont on dispose 21000 gènes (variables) et 18 échantillons (individus).

Dans un premier temps, il est important de choisir les gènes actifs car ce sont ces gènes qui contribuent majoritairement à la construction des axes de l'ACP. Ce sont ces variables qui seront utilisées pour calculer les distances entre les échantillons. Nous avons choisi les deux cents gènes présentant la plus grande variance. Nous avons sélectionné les échantillons correspondant les conditions expérimentales *ns1pg1*,*ns3pg1*, *ns1pg3*. Nous somme en présence d'un tableau réduit à 18 échantillons et 200 gènes.

**Choix des axes de l'acp :**

Dans une analyse en composantes principales le choix des axes joue un rôle très important. Il y a deux manières pour déterminer le nombre d'axes à prendre en compte :

- Critère de Kaiser : on ne retient que les axe dont les valeurs propres sont supérieures à 1 (1=I/P c'est l'inertie moyenne).
- Critère du coude : on conserve les valeurs propres qui dominent les autres, en se référant au graphique en barres des valeurs propres (screeplot).

Nous avons utilisé le critère du coude, en représentant le diagramme en barres de valeurs propres associées à chaque axe de l'acp. Nous cherchons alors une décroissance ou une cassure apparente sur ce diagramme.

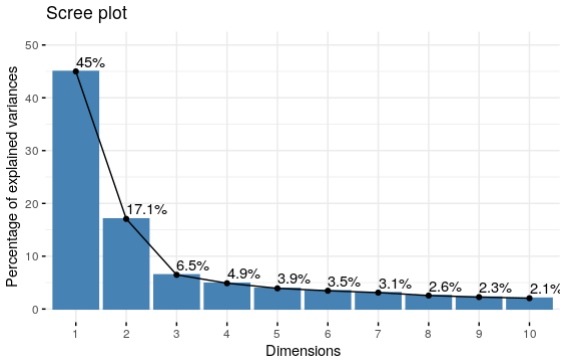



FIGURE 5.1 – Valeurs propres associée à chaque dimension de l'ACP.

D'après la figure 4, on constate que 45% de la variation sont expliqués par la première valeur propre et 17.1% par la deuxième valeur propre. Ainsi, 62.1% de la variance totale est expliquée par les deux premières valeurs propres. Nous choisirons les deux premières dimensions pour représenter les axes principaux à conserver pour la représentation de l'ACP. Nous pourrions également considérer que 62.1% des informations contenues dans les données sont conservées par les deux premières composantes principales.

Nous visualisons ci-dessous le graphe des individus de l'ACP. L'ACP est réalisé sur les 200 gènes dont leurs variances sont les plus grandes.

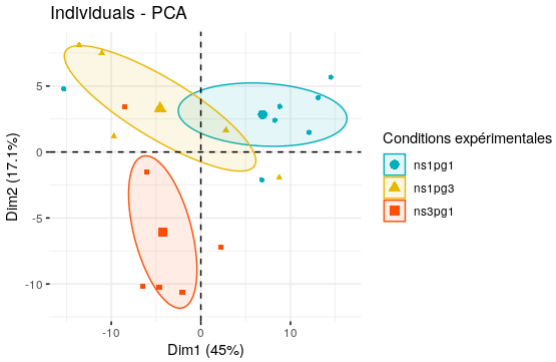



FIGURE 5.2 – Graphe des échantillons de l'ACP

**Interprétation**

L'unique interprétation qui nous intéresse, consiste à vérifier que les échantillons se séparent en fonction des conditions expérimentales sur les premiers axes de l'ACP. Donc le graphe des individus de l'ACP présente ci-dessus nous montre que 62% de la variabilité totale exprimée par le tissu du trophoblaste des différents traitements en PGE2 et N3-398 (condition expérimentales) appliquée en période péri-conceptionnelle est expliquée par les deux premières dimensions. Le premier axe de composante principale (45% de la variabilité) sépare surtout les échantillons correspondant à la condition *ns3pg1* des autres échantillons ayant une condition *ns1pg1* ou *ns1pg3*, tandis que les échantillons de *ns1pg1* et *ns1pg3* se regroupent et s'opposent aux échantillons de *ns3pg1* sur la deuxième composante principale disposante 17% de la variabilité. Cette analyse met en évidence un effet du traitement durant la période péri-conceptionnelle sur l'expression

des gènes du trophoblaste des conceptus J15 de gestation.

La sélection des gènes clés a été demandé aussi, nous avons identifié les 20 gènes qui contribuaient à l'explication de la plupart des écarts dans l'ensemble de nos échantillons. Nous avons listé ces gènes pour en donner du sens, leurs identités sont : FRS2, MEX3C, TBC1D13, MAB21L3, RASSF6, PRTG, ENSBTAG00000047908, PIK3R1, ATF7IP2, CSGALNACT2, ZNF652, FCER2, ENSBTAG00000048185, ZFP36L2, ATP13A3, ENSBTAG00000045630, TSPAN1, LYZ3, ADGRE5, TKDP2.

Nous représentons le graphe des échantillons et celui des variables(gènes) de l'acp pour le tableau réduit aux 20 plus importants gènes. Le graphe des gènes présente pour chaque gène un vecteur dont les coordonnées traduisent sa contribution dans la répartition des échantillons du trophoblaste dans les dimensions. L'exploration par ACP des 20 gènes du trophoblaste montre que 92.5% de l'inertie totale sont expliqués par les deux premières dimensions (Figure 5.4). Nous pouvons constater que les échantillons du groupe *ns1pg3* sont dispersés dans les deux axes. Ainsi les échantillons [ 1134T1, 1140T4, 1121T1, 1121T2 ] sont mal représentés car ils sont très proches de l'origine et assez loin des axes. Rappelons que la corrélation de chaque point sur un axe exprime la qualité de représentation du point sur l'axe, elle prend des valeurs entre 0 et 1. Si cette valeur est proche de 1, alors le point est bien représenté sur l'axe, sinon le contraire. On s'intéresse donc essentiellement aux points bien représentés (i.e. situés loin du centre).

La dimension 1 ($68,4\%$ de la variabilité) oppose des échantillons tels que 1138T4a, 1146T5,1140T5a (à droite du graphe des échantillons, caractérisées par une coordonnée fortement positive sur l'axe) à des échantillons comme 1121T3,1138T5a.
Le groupe auquel les échantillons 1138T4a,1146T5,1140T5a appartiennent partage :

- de fortes valeurs pour les gènes qui se trouvent à droite du graphe des gènes (par exemple : ZNF652,MAB21L3).
- de faibles valeurs pour les gènes ENSBTAG00000048185 et ENSBTAG00000047908.

Le groupe auquel les échantillons 1121T3, 1138T1 appartiennent (caractérisées par une coordonnée négative sur l'axe) partage :

- de fortes valeurs pour les gènes ENSBTAG00000048185 et ENSBTAG00000047908.
- de faibles valeurs pour les gènes qui sont à droite du graphe des variables en classant de la moins extrême à la plus extrême.

Les échantillons du groupe *ns1pg1* (sauf 1140T4) sont corrélés à cette dimension (corrélation 0.54 et 0.96) et la plupart d'entre eux se regroupent à droite du graphe des échantillons. En s'appuyant sur les coefficients de corrélations, dans cet axe les gènes FRS2,MEX3C,TBC1D13,..ect (les 13 gènes qui se trouvent à droite du graphe des gènes) sont le plus corrélés à la dimension, autrement dit sont très proches de l'axe 1, cela nous indique que ces gènes sont bien représentés dans cet axe et également contribuent le plus à la répartition des échantillons.

En ce qui concerne l'axe 2 ($24,1\%$ de la variabilité), nous pouvons remarquer que les échantillons du groupe *ns3pg1* (sauf 1134T4) se regroupent, sont très corrélés entre eux et proches de cet axe. les échantillons des conditions *ns1pg1* et *ns1pg3* sont loin de la dimension 2 donc on peut en tirer qu'ils sont peu représentés par cet axe. Enfin, les gènes ENSBTAG00000045630, LYZ3, TSPAN1, TKDP2, ADGRE5 sont très bien corrélés avec l'axe et participent donc le plus à la séparation des conditions.
Les coefficients de corrélation pour chacun des 20 gènes en fonction de chacune des deux dimensions sont fournis dans le tableau ci-dessous.

| Gènes | Dimension 1 | p-values |
|---|---|---|
| MEX3C | 0.9790555 | 1.743702e-12 |
| MAB21L3 | 0.9746837 | 7.837213e-12 |
| FRS2 | 0.9741628 | 9.208934e-12 |
| TBC1D13 | 0.9719869 | 1.746469e-11 |
| RASSF6 | 0.9663397 | 7.455590e-11 |
| PRTG | 0.9600106 | 2.900157e-10 |
| CSGALNACT2 | 0.9519763 | 1.223055e-09 |
| PIK3R1 | 0.9516493 | 1.289967e-09 |
| ATF7IP2 | 0.9491498 | 1.915349e-09 |
| FCER2 | 0.9394159 | 7.539073e-09 |
| ZFP36L2 | 0.9280394 | 2.880074e-08 |
| ZNF652 | 0.9274908 | 3.055102e-08 |
| ATP13A3 | 0.9232305 | 4.758317e-08 |
| ENSBTAG00000048185 | -0.9365355 | 1.083183e-08 |
| ENSBTAG00000047908 | -0.9601212 | 2.837578e-10 |
| | Dimension 2 | p-values |
| ENSBTAG00000045630 | 0.9848359 | 1.340676e-13 |
| LYZ3 | 0.9728811 | 1.351025e-11 |
| TSPAN1 | 0.9651458 | 9.816347e-11 |
| TKDP2 | 0.9509683 | 1.439563e-09 |
| ADGRE5 | 0.9480759 | 2.256198e-09 |

TABLE 5.1 – les coefficients de corrélations

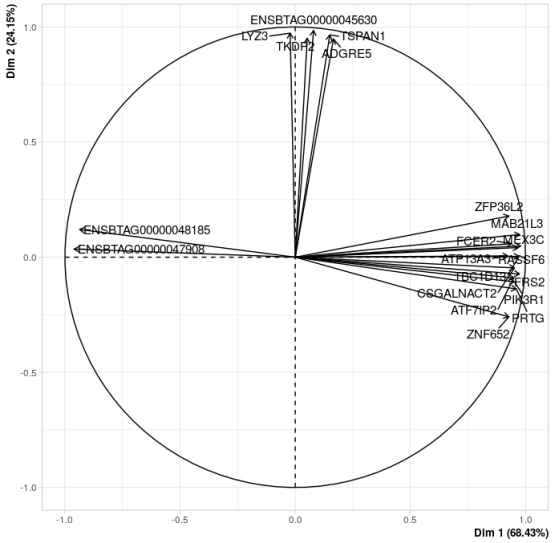



FIGURE 5.3 – Graphe des gènes de l'ACP

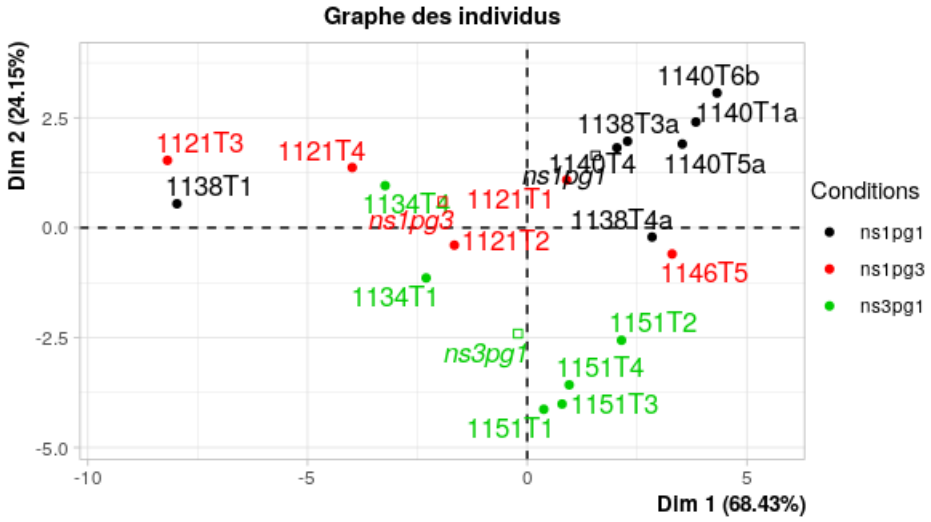



FIGURE 5.4 – Graphe des échantillons

# Chapitre 6

# Intégration des données omiques

L'objectif de l'intégration des données omiques est d'identifier une signature multi-omique hautement corrélée discriminante des groupes d'échantillons connus, en intégrant par exemple les données de deux tissus de l'embryon du stade J15 (Trophoblaste et Disque) ou des données transcriptome et celle de lipidique issus du stade J7. Dans le cadre de notre projet, on a analysé les données collectées de cellule blastocyste (stade J7) et le conceptus J15 dont nous avons bien précisé leurs descriptions dans la première partie du rapport.

## 6.1 PLS

La régression PLS (Partial Least Square) permet de modéliser la liaison entre un bloc de variable Y et un bloc de variables X observées sur les mêmes individus (échantillons). cette méthode consiste à rechercher dans un premier temps des composantes orthogonales, combinaisons linéaires des variables X, expliquant au mieux à la fois les variables X et Y. Ensuit les équations de régression PLS sont obtenues en régressant chaque variable Y sur les composantes, puis en exprimant ces régressions en fonction des variables X d'origine. En général, lorsque le nombre $q$ de variables explicatives est très grand, voire plus grand que le nombre $n$ d'individus à analyser, c'est le cas de nos données RNA-seq, il est difficile ou voire impossible, d'utiliser la méthode des moindres carrés. On peut dire que cette méthode est adaptée à ce cas de données en grande dimension comme les données transcriptome (e.g génétique ).

Il existe différentes versions de régression PLS et elles ont été proposées en fonction de l'objectif poursuivi :

- PLS1 : Une variable cible Y quantitative est à expliquer, modéliser, prévoir par $p$ variables explicatives quantitatives $X^j$.
- PLS2 : Il y a deux versions de PLS2 :
  - la version canonique qui permet de mettre en relation un ensemble de $q$ variables quantitatives Y et un ensemble de $p$ variables quantitatives X.
  - la version régression qui cherche à expliquer, modéliser un ensemble de $q$ variables Y par un ensemble de $p$ variables explicatives quantitatives.
- PLS-DA : Version discriminante. Cas particulier du cas précédent. La variable Y qualitative à q classes est remplacée par q variables indicatrices de ces classes.

Dans cette partie on s'intéresse la PLS2 de la version canonique pour mettre en relation les données transcriptomiques et lipidiques prisées par les mêmes échantillons.

Nous voulons intégrer les niveaux d'expression transcrite (transcriptome) avec les concentrations d'acides gras (lipidique). Il s'agit que X représente les données transcriptomique collectées du stade J7 de l'embryon, plus précisément dans la cellule blastocyste, cet ensemble dispose 16 échantillons et 30 gènes sélectionnés (top gènes), Y est l'ensemble des données de lipidomique des embryons au stade J7, incite 16 échantillons et 15 molécules qui décrivent ces individus et les variables de l'ensemble X et Y sont quantitatives. Nous avons utilisé la méthode sparse PLS pour que nous choisissons le nombre de variables qui participe la répartition des échantillons pour les deux premières composantes principales. On a découvert le package mixOmic du logiciel R qui peut faire **sPLS** en une seule commande en utilisant la fonction *spls* qui prend en argument les ensembles X et Y toujours quantitatives, $keepX = (a, b)$ où a et b sont les nombres des variables (gènes) qu'on veut sélectionner sur les deux premières composantes de PLS de l'ensemble X, idem pour $keepY = (d, f)$, on peut quand même choisir arbitrairement ces vecteurs. Pour notre analyse, on prend a et b égal à 10 pour X, d et f égal à 5, ce qui veut dire on sélectionne les 15 variables dont leurs variances sont les plus grandes pour chaque axe.
Nous avons mis en parallèle le graphe des échantillons des données transcriptomiques et celui de lipidomiques, on a aussi représenté le graphe des variables pour identifier les variables qui contribuent la répartition.

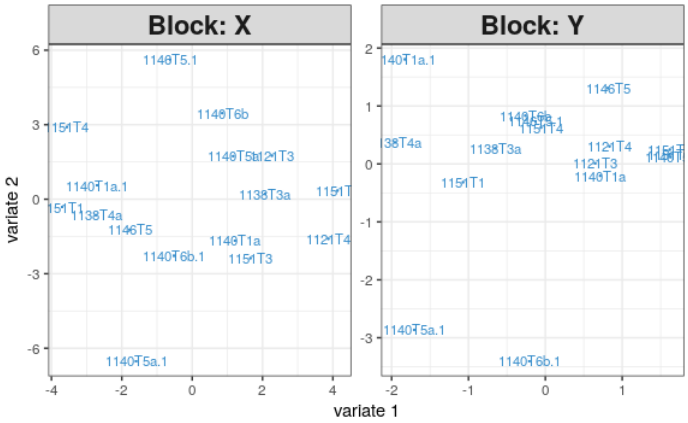
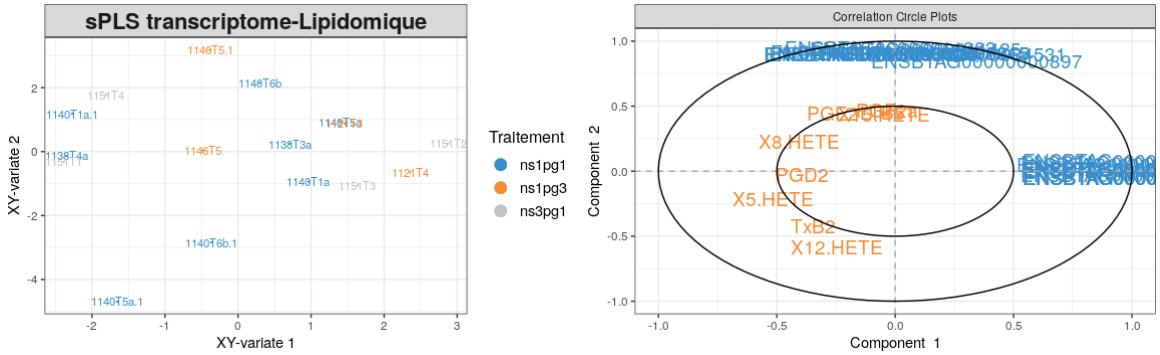



FIGURE 6.1 – les graphes des échantillons pour les données transcriptome et lipidique




(a) Graphe des individus pour X et Y

(b) les graphes des variables

Dans la figure 1.1, on remarque que la plupart des échantillons sont proches de l'origine dans le bloc Y, les échantillons ne se trouvent pas le même endroit entre les deux blocs X et Y. Pour mieux distinguer la corrélation entre les données transcriptome et lipidique, on a utilisé la méthode dite **DIABLO** qui consiste à identifier la corrélation multi-omiques. Le résultat associé est représenté ci-dessous.

Dans le graphe des variables, nous constatons que les gènes qui décrivent les échantillons trans-

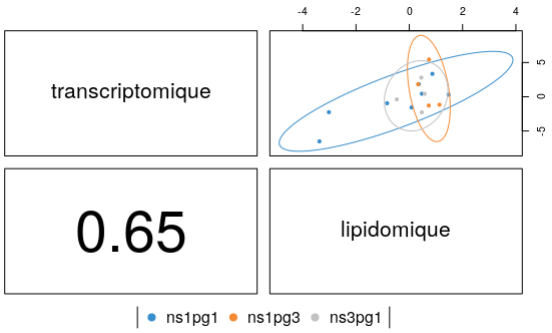
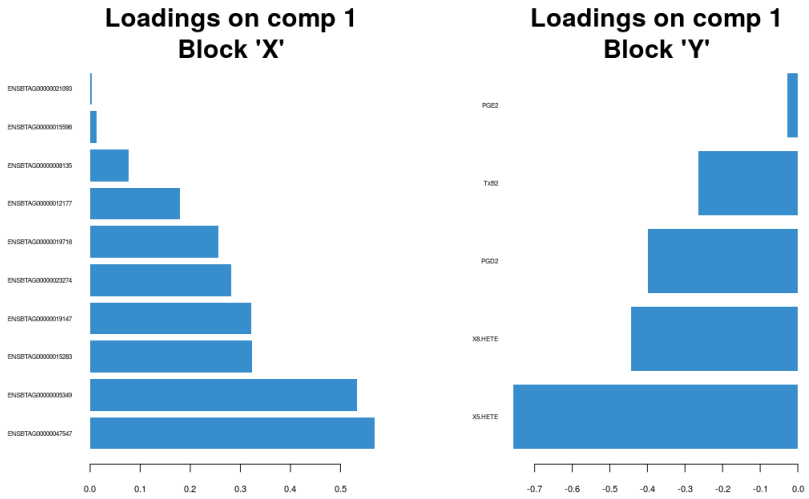
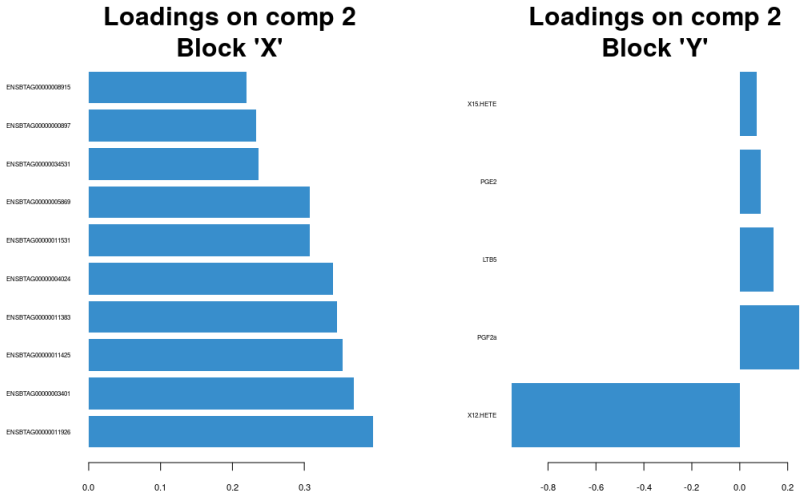



FIGURE 6.2 – Degré de corrélation

criptomiques sont corrélés aux axes, mais les variables lipidiques sont assez loin des axes et proches de l'origine (la plupart). Cela nous montre que les gènes contribuent plus pour la séparation des échantillons entre X et Y. On identifie les degrés de contribution pour les deux premiers axes :




(a) contribution axe 1




(b) Axe 2

## 6.2 PLS-DA

L'objectif de la régression PLS-DA est d'analyser un seul ensemble de données (par exemple les données transcriptomique) et classer les échantillons en grouper connus et prédire la classe de nouveaux échantillons. On est également intéressé à identifier les gènes (variables) clés qui conduisent à la discrimination. La PLS-DA est une technique de classification qui nécessite des groupes d'appartenance de type qualitatives (classes) des différents objets qui composent le jeu de données. Elle prend en entrée des échantillons d'une variable qualitative (classe) à prédire, notée Y et codant l'appartenance des échantillons à une région (tempérée ou tropicale) ainsi que des échantillons de variables quantitatives $X = (X_1, ...., X_p)$. Les classes sont au nombre de trois, *ns1pg1*, *ns1pg3*, *ns3pg1*.

Nous supposons que X et Y peuvent être projetées sur un espace de dimensions réduites. Ces matrices sont alors décomposées en matrice de scores et de loadings selon les équations :

$$X = SL_X^t + R_X \tag{6.1}$$

$$Y = L_Y Q^t + R_Y \tag{6.2}$$

où S et $L_X$ représentent les scores et les loadings de X, $L_Y$ et Q sont les scores et les poids de chaque classe de Y, $R_X$ et $R_Y$ contiennent les résidus.
Pour calculer S, on va utiliser le poids $p_j$ de chaque variable mesuré par leur covariance et contenant dans la matrice P. On définit la matrice P selon la formule suivante :

$$S = XP \tag{6.3}$$

Les scores de X étant de bons prédicteurs de Y, on a également :

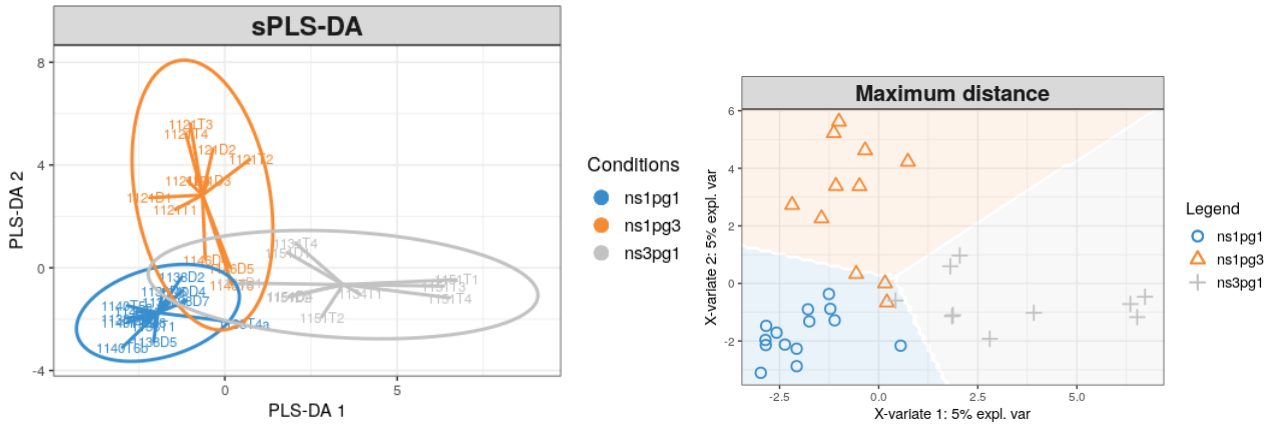
$$Y = SQ^t + G \tag{6.4}$$

Les équations précédent peuvent être combinées la manière suivante :

$$Y = XB + R_Y \qquad o\grave{u} \quad B = PQ^t \tag{6.5}$$

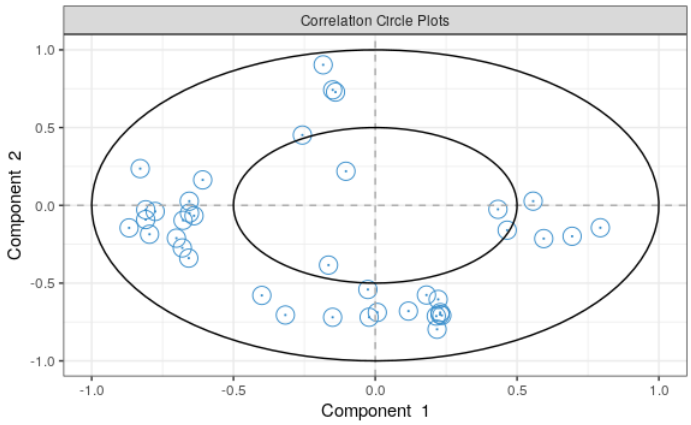
Les coefficients de cette matrice permettent de prédire la valeur de Y par des nouveaux échantillons n'ayant pas servi. Dans notre analyse X représente un tableau de 35 échantillons considérés des observations et 21 000 gènes considérés des variables, il s'agit la concaténation des données des embryons du stade de conceptus J15 du trophoblaste et disque (17 échantillons de disque et 18 échantillons du trophoblaste) et Y est les conditions expérimentales d'où on a trois modalités, bien précisé précédemment. On se contente de comparer les échantillons de deux tissus, soi-disant, on vérifie s'ils sont corrélés ou non. Nous avons donc utilisé la PLS-DA dans un contexte où les variables explicatives sont nombreuses ce qui peut rendre l'interprétation des données difficile.

Dans le but d'améliorer l'interprétation des données ainsi que la séparation des classes, on applique nos données la sPLS-DA (sparse Partial Least Square Discriminant Analysis) qui est une extension de la PLS-DA. Cette méthode nous facilite de choisir un nombre de gènes sélectionnés selon les composantes principales, ce qui veut dire qu'on doit réduire la dimension de l'espace pour mieux visualiser et d'en tirer des informations pertinentes.




(a) Projection des échantillons

(b) Maximum de distance des échantillons




(c) Projection des gènes

Dans la pratique, nous avons sélectionné 20 gènes dans chacune de deux première composantes. Les éléments contenus dans la matrice score et loading pour chacune des variables sont représentés ci-dessous. Autrement dit, nous observons la répartition des échantillons ainsi que les gènes sur les nouveaux axes créés.

Dans le graphe de projection des échantillons, on constate que les échantillons de toutes les trois conditions se regroupent ensemble. Les échantillons 1146D4,D5,T5 du groupe *ns1pg3* sont très proches du groupe *ns3pg1* par rapport à leur propre groupe, il y a aussi certains échantillons du groupe *ns1pg1* qui se trouvent dans l'intersection de trois cercles de classement. Chaque vache (exemple d'identité : 1121) de tissu trophoblaste et disque, leurs valeurs sont corrélées entre eux. Donc peu import la condition, les échantillons issus d'une même vache sont toujours proches. Dans le graphe des variables (gènes), on observe qu'il y a 4 gènes qui sont proches de l'origine et assez loin des axes donc ils sont mal représentés et on ne s'intéresse que les gènes qui sont loin de l'origine (ne se trouvent pas dans le petit cercle).

Nous listons les gènes qui contribuent la répartition des échantillons dans chaque composante principale, en décrivant le pourcentage de leur contribution par le graphe suivant. On considère un exemple sur le premier axe de notre sPLS-DA et voici les vingt gènes de l'axe 1. On remarque que les gènes du groupe *ns1pg1* participent plus que les deux autres groupes à la séparation des échantillons. ENSBTAG00000016836 est le seul et unique gène qui agit pour l'attribution des échantillons du groupe *ns1pg3*, pour *ns3pg4* on a repéré que 5 gènes et les restes pour la condition contrôle. L'information pertinente qu'on peut en tirer, est que la présence de la prostaglandine dans un échantillon rend faible les gènes décrivant cet échantillon donc on sent l'impact de la

prostaglandine.

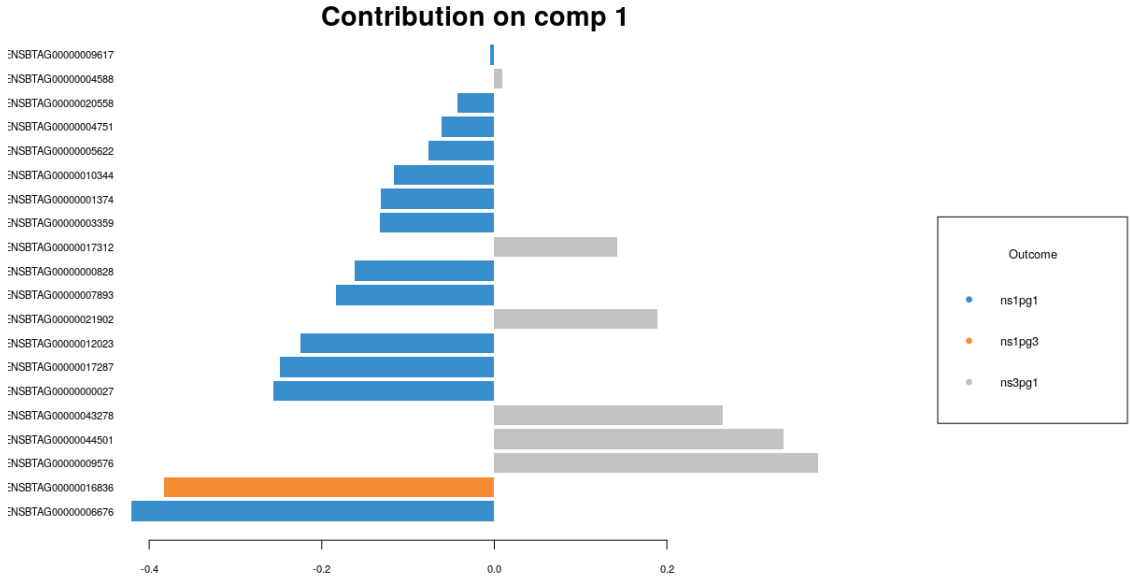



FIGURE 6.3 – Contribution des gènes pour axe 1

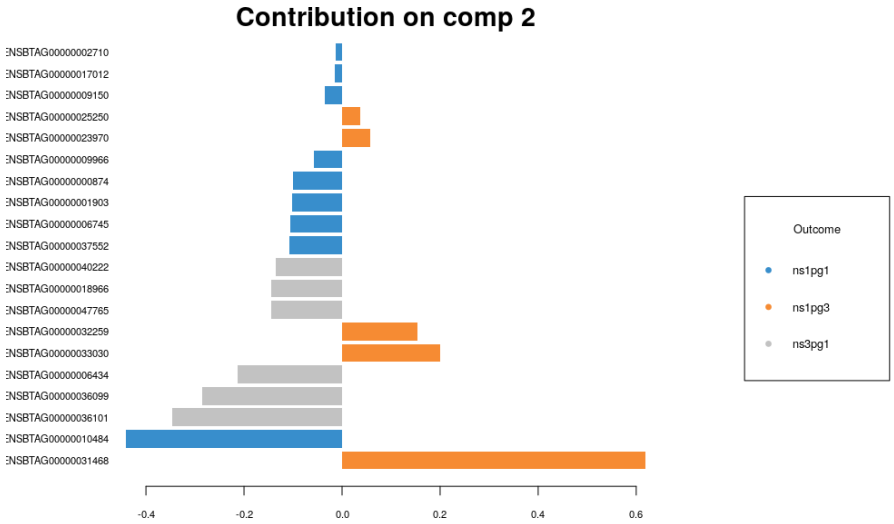



FIGURE 6.4 – Contribution des gènes pour axe 2

# Chapitre 7

# Inférence de réseaux

L'analyse différentielle des données d'expression ou la classification de ces données ne peut pas modéliser la dépendance entre les gènes. Par contre l'inférence de réseaux nous aidera à déterminer les liaisons entre les gènes.

## 7.1 Principe

L'inférence de réseau permet de relier les gènes qui interagissent entre-eux d'un point de vue statistique autrement dit il modélise la dépendance entre les gènes. La biologie des systèmes s'intéresse aux interactions entre les différents acteurs biologiques donc dans un contexte biologie, on recherche les liaisons entre un ensemble de gènes. Il existe plusieurs types de réseaux comme les réseaux métaboliques, réseaux d'interaction de protéines et réseaux de régulation génétique. Dans le cadre de notre stage, on se focalise sur les réseaux de régulation des gènes dont nous présentons la relation de dépendance entre les gènes par un graphe $G = (V, E)$ où :

- $V = \{1, ..., p\}$ ensemble des nœuds représentant les gènes,
- $E$ ensemble des arêtes représentant les interactions entre les gènes.

Nous rappelons ci-dessous la définition d'un graphe.

**Définition 1.** *On appelle graphe (ou réseau) $G = (V, E)$ (non pondéré) ou parfois $G = (V, E, W)$ (pondéré) un ensemble d'entités, $V$ , appelées nœuds (ou vertex en anglais) qui peuvent (ou pas) être reliées, deux à deux, par une relation donnée (appelée arêtes ou edge en anglais). L'ensemble des paires de sommets liés par une relation est noté $E \subset V * V$.*
*Dans le cas pondéré, les relations sont chacune munie de poids (notés $W$ ) qui sont des réels positifs.*
*Les arêtes peuvent être orientées ou non et la matrice des poids, $W$ est alors symétrique ou non.*

**Remarque 1.** *Un graphe $G = (V, E)$ non orienté, avec $p$ nœuds est équivalent à une matrice triangulaire $A$ dite adjacence de dimension $p * p$, définissant pour tout $j$, $j^{'}$ $(j \neq j^{'})$ la présence éventuelle d'une arête entre le nœud $j$ et le nœud $j^{'}$ de la manière suivante :*

1. *si $\alpha_{j,j'} = 1$, alors il existe une arête entre le nœud $j$ et le nœud $j^{'}$,*
2. *si $\alpha_{j,j'} = 0$, alors il n'existe pas une arête entre le nœud $j$ et le nœud $j^{'}$.*

*Et $\forall j$, $\alpha_{j,j} = 0$*

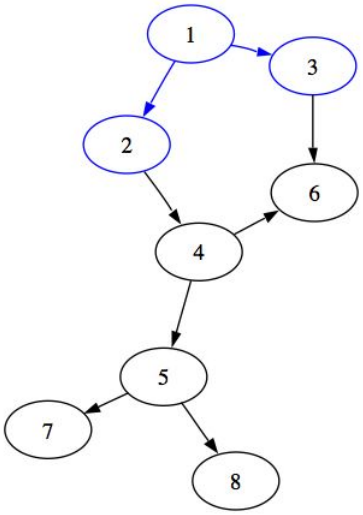



FIGURE 7.1 – Exemple d'un simple graphe

De plus, il est possible d 'effectuer un certain nombre d'opérations sur les graphes qui vont nous donner des informations intéressantes d'un point de vue biologique. À partir de ces graphes, il se peut qu'on effectue un regroupement de gènes, qui permet de visualiser les gènes qui appartiennent au même processus biologique, ou bien qui sont régulés par les mêmes mécanismes.

Nombreuses méthodes ont été proposées par différents auteurs pour reconstituer un graphe comme les réseaux bayésiens qui donnent des graphes orientés ainsi les graphes non-orientés qui sont les modèles graphiques gaussiens, log-linéaire de Poisson.

## 7.2 Modèle graphique gaussien

L'hypothèse sous-jacente du modèle graphique gaussien est que les données suivent une loi normale ce qui n'est pas le cas pour nos données qui sont des données RNA-seq discrète. Dans un premier temps, il est préférable de transformer les données en utilisant la méthode de transformation la plus simple log-transformée, qui peut nous conduire à des données approximativement normales. Par la suite, on considère que nos données suivent une loi normale de dimension $n \times p$, de moyenne nulle et de variance $\Sigma$, matrice définie positive de taille $p \times p$. Les comptages normalisés $(Y_1, ...., Y_n)$ sont indépendantes et identiquement distribuées :

$$Y_i \sim N(0, \Sigma) \text{ pour tout i} \in \{1, ..., n\}.$$

Les dépendances conditionnelles entre deux gènes conditionnellement aux autres gènes se reposent sur les coefficients de la matrice de corrélations partielles entre les expressions de gènes (par définition la forme normalisée de la matrice de covariance est bien la matrice de corrélation). Ses coefficients sont définis de la manière suivante :

$$\forall j \neq j^{'} (j < p, j^{'} < p), \quad \Gamma_{jj'} = Cor(Y_j, Y_k | Y_k, k \neq j, j^{'}) \tag{7.1}$$

Deux gènes seront liés par une arête dans le graphe si et seulement si leur corrélation partielle est significativement non nulle. En effet, le coefficient de corrélation partielle $\Gamma_{jj'}$ entre les gènes $j$ et $j^{'}$ vérifie la relation suivante :

$$\Gamma_{jj'} = \frac{\rho_{jj'}}{\sqrt{\rho_{jj}\rho_{jj'}}}$$

où $\rho_{jj'}$ sont les coefficients de la matrice $\Theta = \Sigma^{-1}$, cette matrice est appelée la matrice de concentration, est également l'inverse de la matrice des variances-covariances $p \times p$. On doit se rendre compte que l'inverse d'une matrice dans le cas des hautes dimensions ($p >> n$) se révèle réellement un problème mathématique mal posé. Nous avons estimé $\Theta$ à l'aide du maximum de vraisemblance, il existe deux approches dans cette méthode, celle de [ Freidman et al] inventé en 2008 et celle de [ Meinshausen et Bühlmann] créée en 2006. L'hypothèse Gaussienne peut nous permettre d'associer à ce contexte un modèle linéaire :

$$Y_j = \sum_{j' \neq j} \beta_{jj'} Y_{j'} + \epsilon \tag{7.2}$$

dans lequel l'expression du gène $j$ s'écrit en fonction de l'expression de tous les autres gènes $Y_{j'}$, $(\forall j \neq j')$.

L'inférence de réseau se réduit à estimer les coefficients $\beta_{jj'}$ à partir des données donc nous avons choisi d'appliquer la méthode par maximum de vraisemblance dans l'approche [ Freidman et al] qui cherchera à minimiser en fonction des coefficients $(\beta_{jj'})$ pour chaque gène $j$ :

$$\min_{(\beta_{jj'})_{jj',j' \neq j}} \sum_j logML_j + \gamma \sum_{j' \neq j} |\beta_{jj'}| \tag{7.3}$$

où $logML_j \sim -\sum_{i=1}^{n}(Y_{ij} - \sum_{j' \neq j} \beta_{jj'} Y_{ij'})$, il s'agit le logarithme du maximum de vraisemblance pour le modèle linéaire. Le second membre du critère est la régularisation qui dépend de la norme $L^1$ de tous les $\beta_{jj'}$. Le principe de la pénalisation de LASSO est qu'une grande partie des coefficients et fait varier le paramètre de régularisation noté $\gamma$, pour obtenir plus ou moins des coefficients non nuls dans la matrice d'adjacence, ce qu'il veut dire d'avoir plus ou moins d'arêtes dans le graphe. Notant $B = (\beta_{jj'})_{jj'=1,\ldots,p}$ la matrice d'adjacence obtenue, elle est symétrique, c'est-à-dire la corrélation du $j^{ème}$ gène avec le $j'^{ème}$ $(j' \neq j)$ est la même que celle entre le $j'^{ème}$ et le $j^{ème}$.
Après avoir construit la matrice d'adjacence, il est facile de repérer les gènes qui sont dépendants ou indépendants et de tracer les graphes. En disant, un coefficient nul $\beta_{jj'} = 0$ indique l'indépendance des gènes $j$ et $j'$ conditionnellement à tous les autres gènes des tableaux de comptage normalisés. En pratique nous pourrons se servir le package **Simone**, en choisissant la méthode *graphical lasso* qui permet d'estimer la matrice de covariance inversée dans l'approche de Friedman et al.

Le tableau utilisé pour l'inférence de réseau correspond 18 échantillons de trophoblaste et 30 gènes(variables) dont nous avons cherché leurs noms dans le site l'Ensembl (www.ensembl.org). Les résultats associés sont représentés sous la forme des figures 7.2 et 7.3. ces figures mettent en évidence la dépendance de 30 gènes sélectionnés parmi les gènes différentiellement exprimés du trophoblaste.

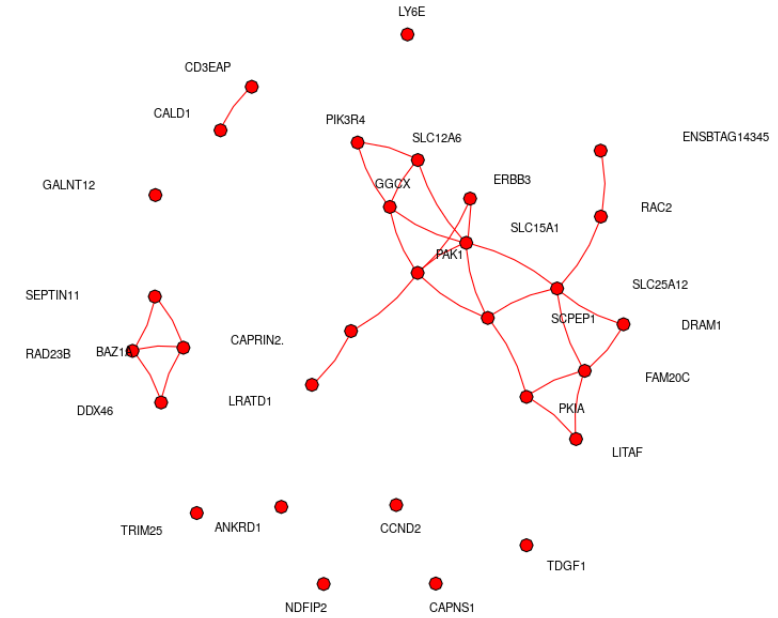



FIGURE 7.2 – le graphe de dépendance

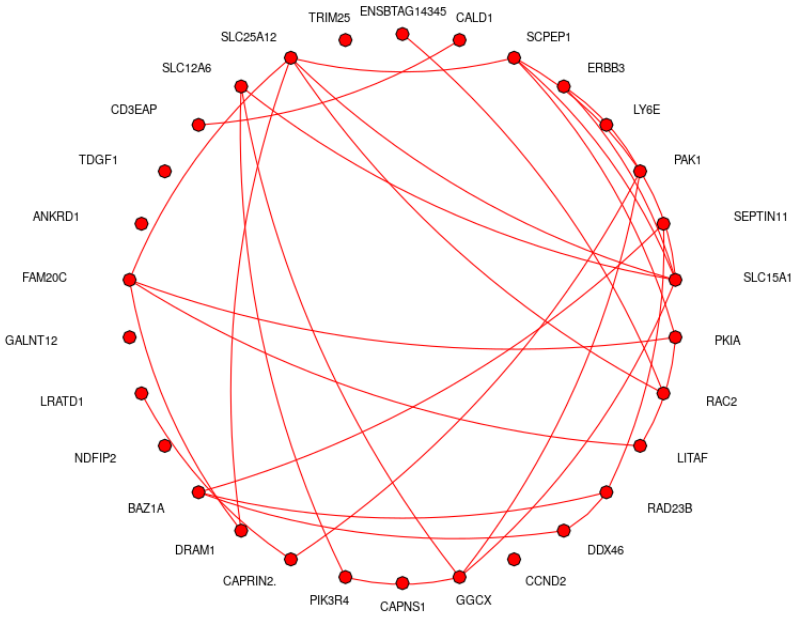



FIGURE 7.3 – Cercle

# Discussion

Dans cette partie, nous discutons la comparaison entre les différentes méthodes d'analyse traitées auparavant, plus particulièrement l'analyse différentielle, l'analyse en composantes principales et la PLS-DA. Chacune de ces méthodes a effectué une mission différente des autres, l'analyse différentielle est faite pour nous fournir des listes des gènes différentiellement exprimés, la raison qui nous a poussé d'effectuer une ACP, était de décrire les données issues de chacun des tissus embryonnaires et visualiser la répartition des échantillons, on a aussi listé les gènes qui contribuaient cette répartition dont nous avons cité quelques-uns précédemment. La PLS-DA visait à expliquer les trois réponses (*ns1pg1*,*ns1pg3*,*ns3pg1*) à partir de l'ensemble des données analytiques, nous avons également identifié les gènes qui participaient plus la plupart d'explication. Pour donner sens à ces listes de gènes, les biologistes se posent " est-ce que ces sont les mêmes gènes qui ont été identifiés pour toutes ces analyses ?" on a représenté un diagramme de venn pour voir l'intersection des gènes sélectionnés des trois méthodes (voire la figure 7.4). On constate qu'il y a 15 gènes communs entre la PLS-DA et l'analyse différentielle, 5 gènes communs entre l'ACP et l'analyse différentielle, mais aucun gène en commun entre la PLS-DA et l'ACP.

On a constaté que ces ne sont pas forcément les mêmes gènes qui réagissent à chaque fois dans les différentes méthodes.

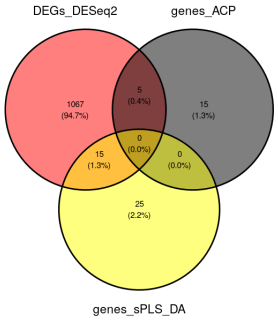



FIGURE 7.4 – Les gènes en communs

# Conclusion

Pour conclure, les notions de base en biologie à laquelle j'étais complètement étranger, m'ont permis de comprendre la problématique de ce stage. Ce défi m'a aussi ouvert les yeux et je me suis rendu compte qu'un statisticien doit avoir une grande capacité d'adaptation face au large éventail de disciplines dans lesquelles il peut être amené à travailler. La partie bioinformatique consiste à transformer les données brutes en tableau de comptage cela m'a appris qu'un statisticien doit pouvoir manipuler les données brutes qu'il doit traiter, quelle que soit leur nature. La préparation des données est la première chose qu'on s'occupe avant d'aller plus loin sur les analyses et joue un rôle très important dans une analyse de données afin de garantir la véracité des données.

L'objectif de ce stage était de mener à bien une étude d'analyse différentielle et exploratoire sur des données transcriptomiques afin de mettre en évidence l'impact des prostaglandines péri-conceptionnelles présentes dans le micro-environnement ovocytaire sur le développement embryonnaire chez la vache. À l'aide de l'analyse différentielle, la présence de la prostaglandine PGE2 affecte les gènes différentielle exprimés en rendant faible la quantité de gènes exprimés, quel que soit le tissu embryonnaire (trophoblaste et disque), il nous semble que l'inhibiteur sélectif de PTGS2 (NS398) contrôle la PGE2 car l'assistance du molécule NS398 rend un peu plus fort que lorsqu'il y a la PGE2 seulement (comparaison de *ns3pg1* contre *ns1pg3*).

L'analyse en composantes principales permet de découvrir que les échantillons dans lesquelles la prostaglandine a été ajoutée dans leurs milieux de maturation et de fécondation des COCs, sont dispersés entre eux et la plupart de ces échantillons ne sont pas bien représentés ce qui veut dire qu'ils sont proches de l'origine du graphe des échantillons de l'acp. On s'est contenté d'étudier la relation entre les échantillons de deux tissus dont on a utilisé la méthode PLS-DA. On a pu rendre compte que les gènes impacté par la prostaglandine, contribuent moins peu à la répartition des échantillons. Nos résultats indiquent que la concentration en PGE2 péri-conceptionnelle affecte la cinétique de développement des conceptus et influence également l'expression de gènes des tissus embryonnaire.

Enfin, ce stage m'a permis de compléter mes connaissances en statistique, de progresser dans l'utilisation du logiciel R et j'ai discerné plusieurs packages que je n'ai jamais rencontrés dans mon cursus universitaire. Cette expérience professionnelle m'a confortée dans mon choix de chercher un travail dans le domaine biostatistique.

# Bibliographie


1. Elie Maza (2016). In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. *FRONT GENET*, 7 :164.
2. S. Anders and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(R106) :1–28.
3. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014
4. Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(10)
5. Andrea Rau, Mélina Gallopin, Gilles Celeux, and Florence Jaffrézic. Data-based filtering for replicated high-throughput transcriptome sequencing. experiments *Bioinformatics*,09(13) :2146-52.
6. Jiang, D., Tang, C., and Zhang, A. (2004). Cluster analysis for gene expression data : a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11) :1370–1386.
7. Kim-Anh Lê Cao, Simon Boitard et Philippe Besse. Sparse PLS discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 253(2011).
8. Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié and Philippe Besse. A Sparse PLS for Variable Selection when Integrating Omics Data. *BMC Bioinformatics*, 34(2009).
9. Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso.*Biostatistics*, 9(3) :432–441.
10. Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9 :485–516.

# Annexes

```
##### Installation des librairies #####
library(DESeq2)
library(HTSFilter)
library(gplots)
library(ggvenn)

##### DESeq2

# Importation des données
Tropho<-read.csv("Images/Travaux/Tropho.csv",header = TRUE,row.names = 1)
conditionT<-read.csv("Images/Travaux/expreTropho.csv",
                     header = TRUE, row.names = 1)
colnames(Tropho)=rownames(conditionT)

# Construction de l'objet de DESeq2
dds<-DESeqDataSetFromMatrix(Tropho,conditionT,design = ~condition)
class(dds)
colData(dds)
design(dds)

# Normalisation
dds<-DESeq(dds)
dds<-HTSFilter(dds)$filteredData # Filtrage de Jaccord
data_normT<-counts(dds) # Comptage normaliser

# Manipulation des comptages normalisees
dim(data_normT)
head(data_normT)
summary(data_normT)

# ns1pg1 vs ns1pg3
res_N1<-results(dds,contrast = c("condition","ns1pg1","ns1pg3"))
gene_N1<-subset(res_N1,padj<0.05)
nrow(gene_N1)
DGE_N1<-rownames(gene_N1) # identites des genes DE

# ns1pg1 vs ns3pg1
res_N2<-results(dds,contrast = c("condition","ns1pg1","ns3pg1"))
gene_N2<-subset(res_N2,padj<0.05)
nrow(gene_N2)
DGE_N2<-rownames(gene_N2)

# ns1pg3 vs ns3pg1
res_N3<-results(dds,contrast = c("condition","ns1pg3","ns3pg1"))
gene_N3<-subset(res_N3,padj<0.05)
nrow(gene_N3)
DGE_N3<-rownames(gene_N3)

# Diagramme de venn
ven<-list(ns1pg1vsns1pg3=DGE_N1,
```

```
            ns1pg1vsns3pg1=DGE_N2,
            ns1pg3vsns3pg1=DGE_N3)
ggvenn(ven,columns = NULL,
       fill_color = c("red","blue", "green"))
```

Listing 1 – DESeq2

```
library(edgeR)
# on crée une liste contenant les trois conditions qui nous intéressent
parametre<-list(c1="ns1pg1",c2="ns1pg3",c3="ns3pg1")
#cherchant les positions de chacune condition dans la data condition (condition expérimentale)
ns1pg1<-which(conditionT$condition==parametre$c1) # ns1pg1, les positions qu'elle s'occupe
#dans le fichier condition
ns1pg3<-which(conditionT$condition==parametre$c2)
ns3pg1<-which(conditionT$condition==parametre$c3)

# Groupe de comparaison
P1<-c(ns1pg1,ns1pg3) # ns1pg1 vs ns1pg3
P2<-c(ns1pg1,ns3pg1) # ns1pg1 vs ns3pg1
P3<-c(ns1pg3,ns3pg1) # ns1pg3 vs ns3pg1

## ns1pg1 vs ns1pg3
dataT1<-Tropho[,P1] # comptage correspondant les échantillons qui ont pris le traitement ns1pg3
#et sans (controle)
condiT1<-conditionT[P1,] # data condition pour ns1pg1 et ns1pg3

# Matrice de design
con<-relevel(factor(condiT1$condition),ref = "ns1pg1")
designMT<-model.matrix(~con)

# L'objet de edgeR
edgeT1<-DGEList(dataT1,group=con)
edgeT1<-calcNormFactors(edgeT1,method = "TMM") # normalisation

#HTSFilter
edgeT1<-HTSFilter(edgeT1,plot=FALSE)$filteredData # Filtrage

# Estimer la dispersion pour tous les comptages de lecture dans tous les échantillons
edgeT1<-estimateDisp(edgeT1,designMT)
# Adapter au modèle binomial négatif
edgeT1_fit<-glmFit(edgeT1,design = designMT)
# Effectuer le test pour chaque gène en utilisant le modèle binomial négative
edgeT1_LRT<-glmLRT(edgeT1_fit)

# Résultat
res_edgeT1<-topTags(edgeT1_LRT,n=nrow(edgeT1_LRT$table),
                    adjust.method = "BH",sort.by = "PValue")
DGE_edgeT1<-subset(res_edgeT1$table,FDR<0.05)
nrow(DGE_edgeT1) # nombre de gènes DE=33
name_DGE_edgeT1<-rownames(DGE_edgeT1) # les gènes DE (identités)

# Nous ferons le même travail pour les autres comparaison
```

Listing 2 – edgeR

```
library(mixOmics)
# On concaténe le comptage de tropho et disque de J15 pour identifier la corrélation
# entre les échantillons de deux tissus
```

```r
countTable<-read.csv("Images/P1/data_sansns3pg3.csv",header = TRUE, row.names = 1)
condTable<-read.csv("Images/P1/condsans_ns3ps3.csv",header = TRUE,row.names = 1)
colnames(countTable)=rownames(condTable)

# considérons les gènes comme des variables et les échantillons comme des individus
log_dataTD<-t(log2(countTable+1))

expression<-as.factor(condTable$condition) # factor de condition
result.splsda<-splsda(log_dataTD, expression ,keepX = c(20,20)) # pls_da pour seulement 40 gènes top

# graphe des échantillons
plotIndiv(result.splsda, ind.names=TRUE, group=expression,legend=TRUE,
          ellipse= TRUE, star=TRUE, title='sPLS-DA',
          X.label='PLS-DA 1',Y.label='PLS-DA 2',
          legend.title = 'Conditions')

plotVar(result.splsda,var.names = FALSE) # visualisation des corrélations de 40 gènes
selectVar(result.splsda,comp=1)$name # les identités de gènes cléf dans le 1er composant
plotLoadings(result.splsda, contrib = 'max', method = 'mean') # contribution des gènes

background<-background.predict(result.splsda, comp.predicted = 2,
                               dist = "max.dist") # prédiction pour utiliser plot des individus
plotIndiv(result.splsda, comp = 1:2, group = expression,
          ind.names = FALSE, title = "Maximum distance",
          legend = TRUE,  background = background)
```

Listing 3 – PLS-DA

```r
## Importation de données
lipido<-read.csv("Images/DonneesJ7/lipidomique.csv",header = TRUE,row.names = 1)
J8ARN<-read.csv("Images/DonneesJ7/J7ARN.csv",header = TRUE,row.names = 1)
condtrolip<-read.csv("Images/DonneesJ7/condtro_lip.csv",header =TRUE,row.names = 1)

## Relation des données transcriptomiques et lipidomiques
pls_tro_lip<-spls(J8ARN,lipido,keepX = c(10,10),keepY = c(5,5)) # sparse PLS
plotIndiv(pls_tro_lip, group = condtrolip$condition,
          rep.space = "XY-variate", legend = TRUE,
          legend.title = 'Traitement',
          ind.names = rownames(condtrolip),
          title = 'sPLS transcriptome-Lipidomique')
plotIndiv(pls_tro_lip) # graphe des échantillons en séparant le bloc X et Y
selectVar(pls_tro_lip,comp = 2) # identités des variables qui contribuent la PLS
plotVar(pls_tro_lip) # graphe des variables

plotLoadings(pls_tro_lip, comp =1 , size.name = rel(0.5)) # contributions des variables

##############################################################
###L'utilisation de DIABLO pour voir le pourcentage de corrélation entre les deux ensembles
X<-list(transcriptomique=J8ARN,
        lipidomique=lipido)
Y<-condtrolip$condition
list_keepX<-list(transcriptomique=c(10,10),lipidomique=c(5,15))

diablo<-block.splsda(X,Y,keepX=list_keepX)
```

```
27 plotDiablo(diablo,ncomp=1)
```

Listing 4 – PLS

```
1  network_data<-read.csv("Images/Network method/networkdata.csv",header = TRUE,
2                         row.names = 1)
3  simofrei<-simone(network_data,
4  control = setOptions(edges.steady = "graphical.lasso",
5                                                    clusters.crit=30))
6  get_net<-getNetwork(simofrei,30)
7  plot(get_net)
8  plot(get_net, type = "circles")
```

Listing 5 – Réseau