Projet Analyse de Donnée

CALLIS Guilhem et FERRERE HOAREAU Anthony

2025-01-17

Contents

1 Introduction

Ce rapport de projet présente les différents aspects de l'analyse que nous avons réalisée sur un jeu de données en rapport avec une expérimentation de différents traitements sur des plantes visant à observer l'évolution de l'expression des gènes chez celles-ci. Pour cela, nous avons décidé d'effectuer différentes analyses sur ce jeu de données de 542 gènes qui nous a été fourni :

-Une analyse unidimensionnelle sur chacune des variables du jeu de données pour mieux comprendre leur comportement. -Une analyse bidimensionnelle pour mettre en évidence les liens entre les variables des données. -Une analyse en composantes principales des variables quantitatives du jeu de données ainsi qu'une classification non supervisée de ces variables, pour ici analyser les traitements. -Une analyse en composantes principales ainsi qu'une classification non supervisée, mais ici en prenant les gènes comme variables. -Une classification non supervisée des gènes à partir ici des variables qualitatives "ExpT1", "ExpT2", et "ExpT3".

2 Description du jeu de données

Chargement du jeu de données: Décrivez l'ensemble du jeu de données en précisant la nature des variables.

En premier lieu, l'analyse du jeu de données révèle qu'il contient à la fois des variables qualitatives et quantitatives. Les variables quantitatives correspondent aux 36 premières colonnes, réparties en deux catégories : les colonnes 1 à 18 pour le réplicat 1 (R1) et les colonnes 19 à 36 pour le réplicat 2 (R2). Ces deux réplicats seront comparés au cours de l'analyse.

Chaque réplicat est subdivisé en trois traitements distincts (T1, T2, T3). Il est important de noter que T3 est une combinaison des traitements T1 et T2, une information essentielle pour les analyses ultérieures. De plus, chaque traitement est décomposé en six heures d'observation, allant de 1 heure à 6 heures. Ainsi, chaque colonne des variables quantitatives peut être désignée par la notation Tt_sH_Rr, où t représente le traitement, s l'heure, et r le réplicat.

Les variables qualitatives, quant à elles, sont contenues dans les colonnes 37 à 39. Ces colonnes (ExpT1, ExpT2 et ExpT3) indiquent les états "Sur", "Sous" ou "Non" de l'expression des gènes pour les différents traitements au bout de 6 heures.

```
## [1] "Table de la nature des différentes variables du jeu de données"
```

```
## 'data.frame':    542 obs. of  39 variables:
##  $ T1_1H_R1: num  -0.205 -0.62 0.309 0.192 0.108 ...
##  $ T1_2H_R1: num  -0.689 -0.856 0.817 0.148 0.288 ...
##  $ T1_3H_R1: num  -0.1811 -0.0211 -0.5615 0.2424 -0.1975 ...
##  $ T1_4H_R1: num  -0.06657 -0.14456 0.18148 0.56182 -0.00155 ...
##  $ T1_5H_R1: num  0.5217 0.4934 -0.337 0.0453 -0.2274 ...
##  $ T1_6H_R1: num  0.448 0.454 -0.373 -0.635 -0.571 ...
##  $ T2_1H_R1: num  -0.449 -0.572 -0.209 0.526 0.34 ...
##  $ T2_2H_R1: num  -1.5144 -1.4755 -1.29 1.4315 -0.0468 ...
##  $ T2_3H_R1: num  -3.815 -3.079 -2.633 1.842 -0.327 ...
##  $ T2_4H_R1: num  -2.5 -2.22 -2.4 1.83 -0.47 ...
##  $ T2_5H_R1: num  -2.9144 -2.2659 -2.4397 1.9242 0.0153 ...
##  $ T2_6H_R1: num  -3.57 -3.36 -2.03 2.19 2.17 ...
##  $ T3_1H_R1: num  -0.6645 -0.5427 -0.2709 0.4127 -0.0402 ...
##  $ T3_2H_R1: num  -2.522 -2.281 -1.176 1.688 0.179 ...
##  $ T3_3H_R1: num  -1.797 -1.597 -3.018 1.812 -0.192 ...
##  $ T3_4H_R1: num  -2.967 -2.635 -2.953 1.868 -0.553 ...
##  $ T3_5H_R1: num  -2.99182 -2.42474 -2.96356 2.14249 -0.00553 ...
##  $ T3_6H_R1: num  -2.84 -2.54 -2.49 2.1 2.09 ...
##  $ T1_1H_R2: num  -0.25 -0.527 0.303 -0.234 -0.33 ...
##  $ T1_2H_R2: num  -0.2376 -0.3474 0.5477 -0.2899 0.0044 ...
##  $ T1_3H_R2: num  -0.741 -0.64 0.589 0.725 0.171 ...
##  $ T1_4H_R2: num  0.504 0.274 -0.908 -0.488 -0.53 ...
##  $ T1_5H_R2: num  0.355 0.347 -1.428 -0.289 -0.481 ...
##  $ T1_6H_R2: num  0.698 0.663 -0.699 -0.649 -0.714 ...
##  $ T2_1H_R2: num  -0.671 -0.67 0.163 0.272 -0.373 ...
##  $ T2_2H_R2: num  -2.489 -2.416 -2.307 1.47 -0.109 ...
##  $ T2_3H_R2: num  -2.4 -2.21 -2.32 1.83 0.13 ...
##  $ T2_4H_R2: num  -2.552 -2.198 -3.294 1.618 -0.453 ...
##  $ T2_5H_R2: num  -2.474 -2.238 -2.967 2.191 0.888 ...
##  $ T2_6H_R2: num  -3.14 -2.47 -3.18 2.19 2.18 ...
##  $ T3_1H_R2: num  -0.62 -0.842 -0.195 0.17 -0.446 ...
##  $ T3_2H_R2: num  -2.7064 -2.4478 -2.1068 1.5124 0.0384 ...
##  $ T3_3H_R2: num  -2.828 -2.552 -2.624 2.051 -0.111 ...
```

```
##  $ T3_4H_R2: num  -2.849 -2.484 -3.024 1.559 -0.222 ...
##  $ T3_5H_R2: num  -2.94 -2.46 -2.81 2.32 1.19 ...
##  $ T3_6H_R2: num  -3.39 -2.97 -2.7 2.16 1.95 ...
##  $ ExpT1   : chr  "Non" "Non" "Non" "Non" ...
##  $ ExpT2   : chr  "Sous" "Sous" "Sous" "Sur" ...
##  $ ExpT3   : chr  "Sous" "Sous" "Sous" "Sur" ...
```

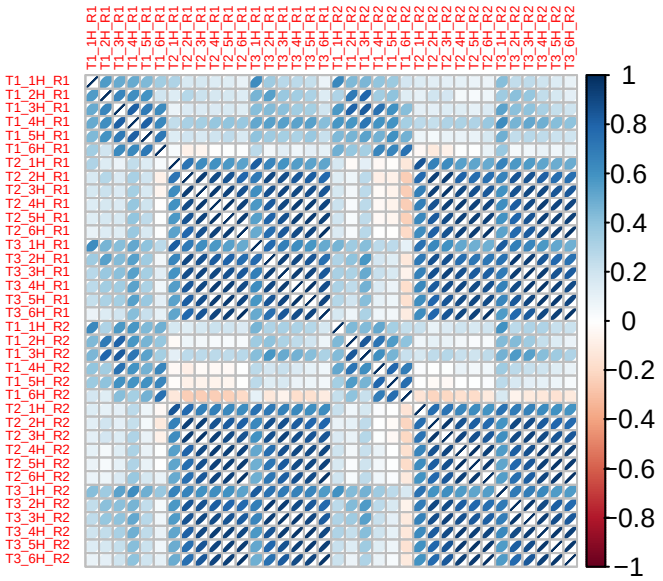

Figure 1: Matrice des corrélations entre les variables quantitatives (les Tt_sH_Rr)

On peut constater sur `la figure` 1 que, dans notre jeu de données, les traitements 2 et 3, quel que soit leur réplicat, sont fortement corrélés entre eux. En revanche, le traitement 1 montre un comportement différent, notamment à la 6ème heure pour le réplicat 2, où une correlation négative avec les autres variables est observée.Cela indique un comportement différent du traitement 1 par rapport aux deux autres. Cela nous permet également de poser l'hypothèse que le traitement 3 (rappelons qu'il s'agit d'une combinaison des deux autres traitements) serait majoritairement influencé par le comportement du traitement 2, entraînant des variations génétiques fortement similaires à celles observées dans le traitement 2.

3 Analyse dimentionnelle du jeu de données

3.1 Analyse uni-dimensionnelle

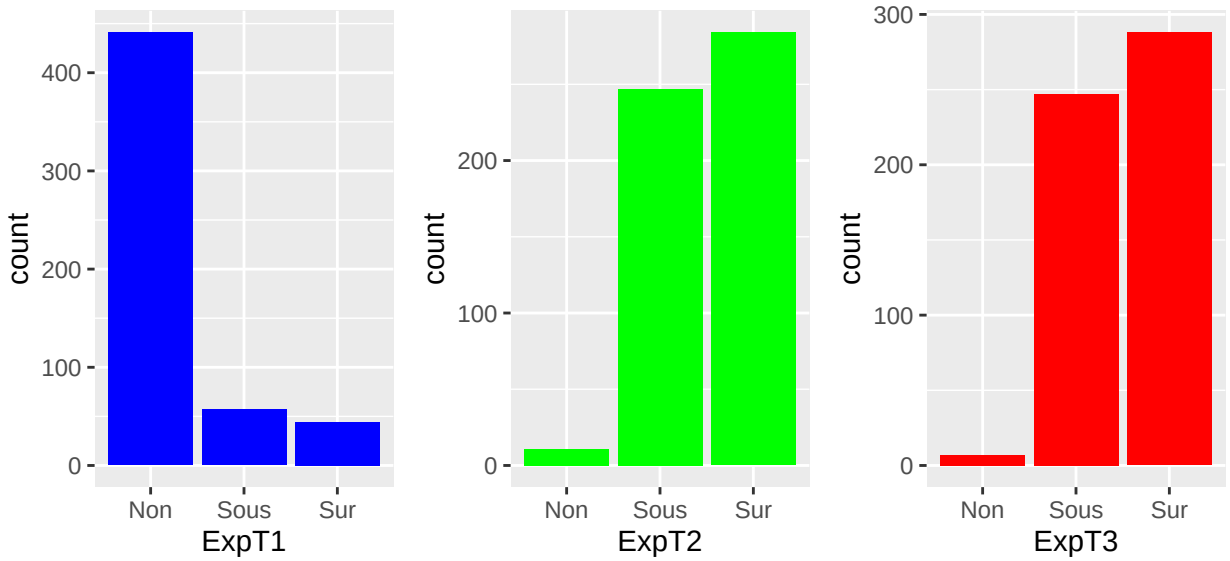

Figure 2: Représentation de la quantité respective de gènes Non, Sur et Sous-exprimés pour les traitements 1, 2 et 3 au bout de 6 heures

À la suite de cette analyse, plusieurs observations et conclusions peuvent être tirées sur les relations entre les variables : (on rappelle que T3 est une combinaison de T1 et T2, comme indiqué dans le sujet.) En ce qui concerne l'analyse Uni-dimensionnelle sur les variables qualitatives, on constate sur `la figure` 2 que les fréquences des gènes classés comme surexprimés, sous-exprimés et non exprimés sont globalement similaires pour ExpT2 et ExpT3. Cela peut s'expliquer par le fait que T3 est une combinaison de T1 et T2, ce qui entraîne des distributions proches. En revanche, ExpT1 se distingue clairement des deux autres, la majorité des gènes dans ExpT1 sont non exprimés, ce qui contraste avec les répartitions plus équilibrées observées pour ExpT2 et ExpT3. Cette observation suggère que le traitement T1 induit très peu de changements dans l'expression des gènes en réponse à un traitement.

Pour les variables quantitatives, on remarque plusieurs choses. En effet, sur `la figure` 3, on remarque que les distributions des valeurs d'expression pour R1 et R2 sont remarquablement similaires. Cela indique une bonne reproductibilité biologique entre les réplicats. La cohérence entre R1 et R2 valide la qualité des données et leur fiabilité pour les analyses ultérieures. Pour les traitements, T2 et T3 sont fortement liés, leurs médianes, leurs intervalles interquartiles sont très similaires. Cela renforce l'idée que T3, étant une combinaison de T1 et T2, hérite principalement des caractéristiques de T2. Cependant, les colonnes T2_1H_R1/R2 et T3_1H_R1/R2 présentent un grand nombre de valeurs aberrantes (outliers), ce qui peut indiquer des réponses génétiques atypiques à 1 heure pour ces traitements. Les intervalles interquartiles pour T1 sont beaucoup plus petits, suggérant que les données pour ce traitement sont plus concentrées autour de la médiane. Toutefois, T1 présente également de nombreux outliers, en plus grand nombre que pour T2 ou T3, ce qui peut indiquer des comportements génétiques spécifiques ou une variabilité accrue pour certains gènes sous ce traitement.

3.2 Analyse Bi-dimensionnelle

Afin de mieux comprendre les relations qu'il pourrait y avoir entre les variables, nous avons décidé de faire 3 analyses Bi-dimensionnelles.

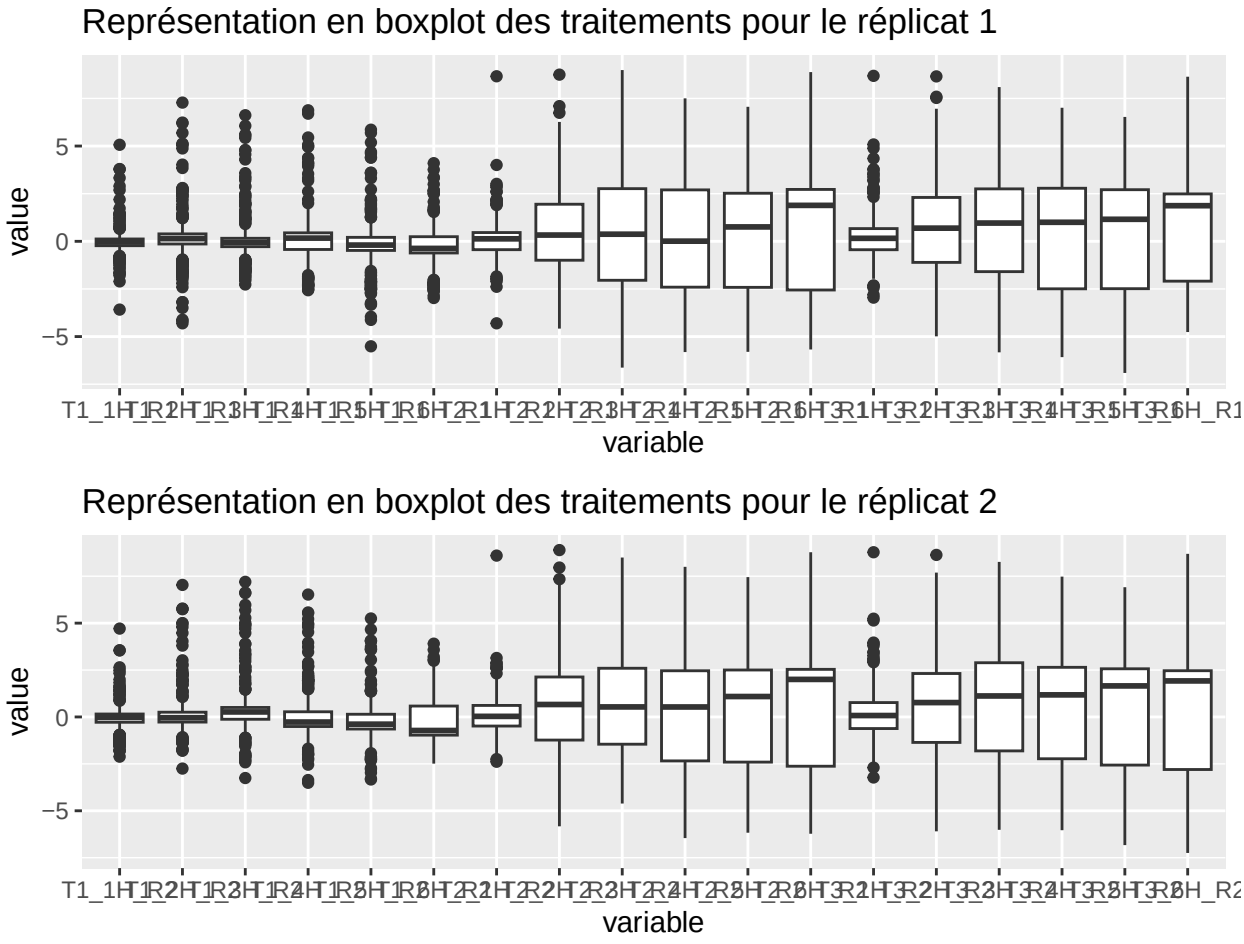

Figure 3: Représentation de la dispersion des valeurs pour chaque variable

3.2.1 Analyse sur les variables qualitatives

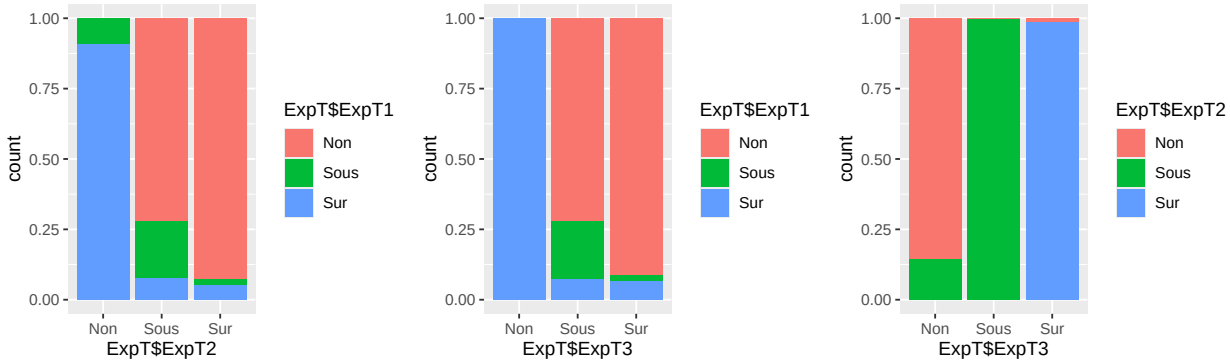

Figure 4: Comparaison entre l'expression des gènes des traitements deux à deux à 6 heures

On constate sur `la figure` 4 que les relations entre les expressions des gènes pour le 2ème et 3ème traitements sont très similaires, contrairement au premier traitement. En effet, la grande majorité, si ce n'est la totalité des gènes non exprimés des traitements 2 et 3 sont sur exprimés pour le traitement 1, et les gènes surexprimés et sousexprimés des traitements 2 et 3 sont majoritairement non exprimés pour le traitement 1. On peut donc en conclure que le traitement 1 est quasiment l'opposé du traitement 2 en termes d'expression des gènes. De plus, grâce au troisième graphique, on observe que le traitement 3, étant une combinaison des deux autres traitements, est quasiment identique au traitement 2 en termes d'expression des gènes surexprimés et sous-exprimés, bien qu'une différence minime soit notée pour les gènes non exprimés. Ce qui est cohérent avec les observations de l'analyse unidimensionnelle de l'expression des gènes.

3.2.2 Analyse sur les variables qualitatives et quantitatives

Cette analyse nous montre, sur la figure 5, différents graphiques représentant l'évolution du degré de liaison entre les traitements et l'expression des gènes au cours du temps. Ce degré de liaison reflète ici la proximité entre l'évolution de l'expression des gènes considérés au fil du temps lors du traitement et l'expression de ces mêmes gènes après 6 heures. On constate pour le premier traitement que, malgré une évolution croissante du degré de liaison, celui-ci varie fortement en fonction du temps, que ce soit pour le premier ou le second réplicat, avec des diminutions à 2h, 3h et 5h. De plus, ce degré atteint un maximum de 0.56 pour le premier réplicat et de 0.46 pour le deuxième réplicat, indiquant une forte évolution au cours du temps des gènes de la plante lors du traitement. Ce traitement prend plus de temps à agir que les deux autres, puisqu'il n'est semblable qu'à 50% au résultat final attendu pour ce traitement au bout de 6h. Les deux autres traitements, quant à eux, ayant une forte similarité, comme constaté précédemment, présentent une courbe d'évolution du degré de liaison très similaire. On peut donc en conclure que ces deux traitements agissent dans un même laps de temps et beaucoup plus rapidement que le premier traitement, avec beaucoup moins de variation de l'expression des gènes au cours du temps. Étant donné que les courbes sont logarithmiques, on constate de forts changements dans l'expression des gènes au départ, mais ces changements diminuent au fil du temps. On peut donc en déduire que, contrairement au premier traitement, qui reste fortement actif sur la plante tout au long de l'expérience malgré une forte présence de gènes non actifs et qui prend beaucoup de temps à agir, les deux autres traitements sont principalement actifs sur la plante au début de l'expérience. Leur influence et la variation de l'expression des gènes diminuent avec le temps. Ils sont donc, en comparaison avec le résultat mesuré à 6h, bien plus rapides et efficaces que le traitement 1.

3.2.3 Analyse sur les variables quantitatives

Pour cette analyse, nous avons décidé de représenter à nouveau la matrice de corrélations de la première partie, mais en réalisant un zoom sur les différents traitements.

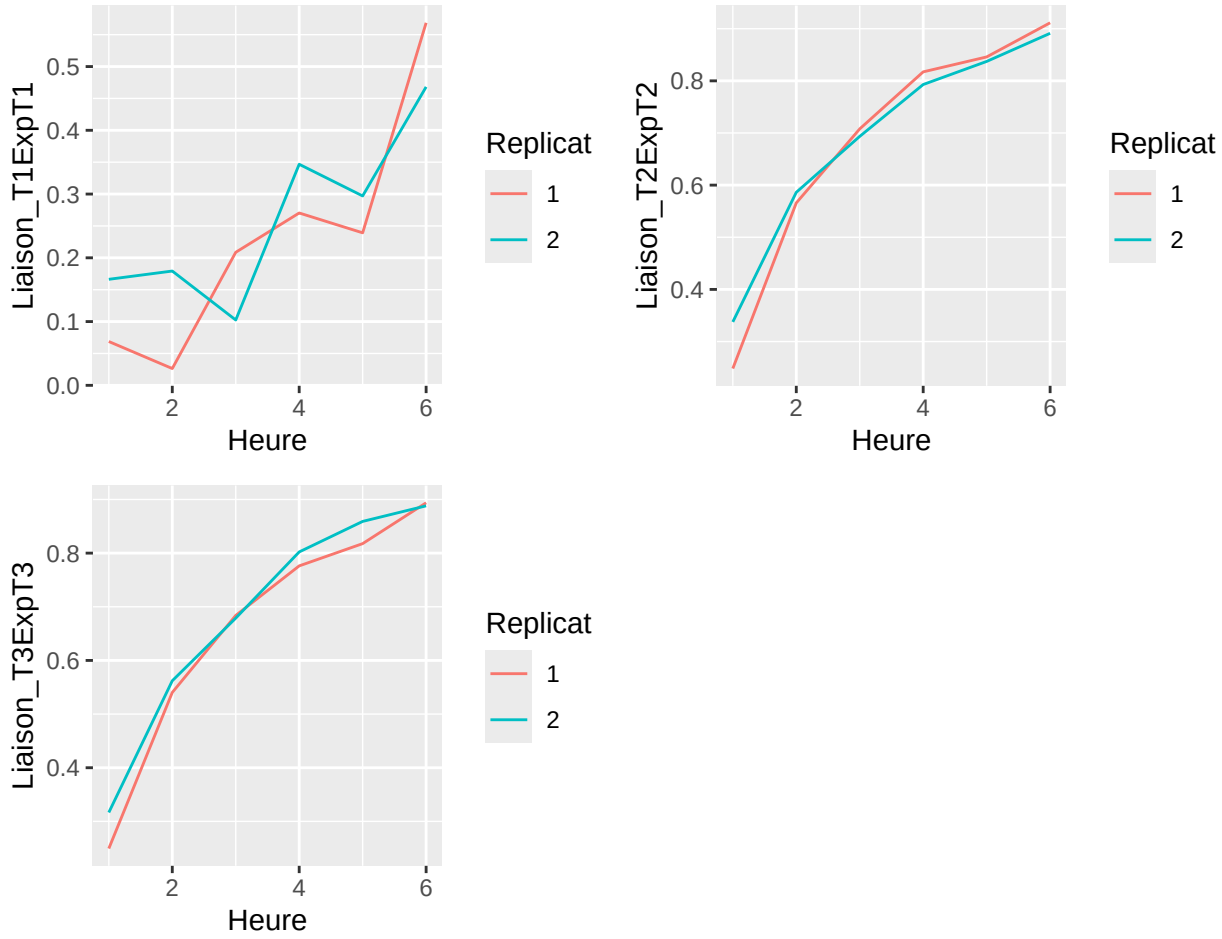

Figure 5: Degré de liaison entre l'expression des gènes obtenu dans le jeu de donnée à 6 heure et l'évolution de cette expression au court du temps lors des testes

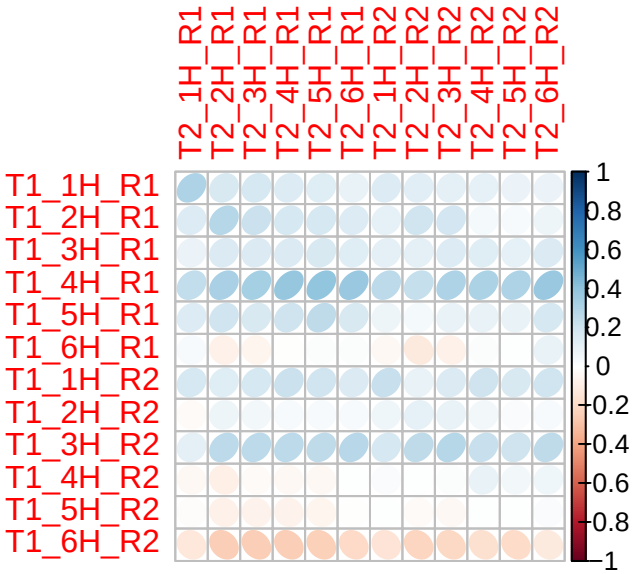

Figure 6: Correlations entre les variables du traitement 1 et du traitement 2 pour les deux réplicats

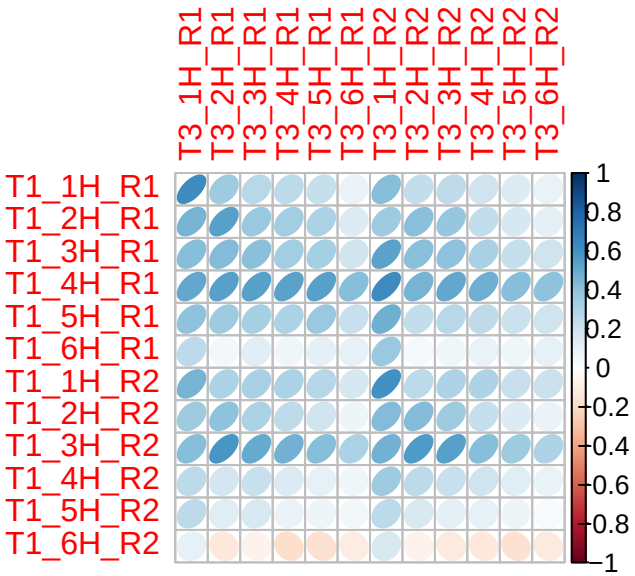

Figure 7: Correlations entre les variables du traitement 1 et du traitement 3 pour les deux réplicats

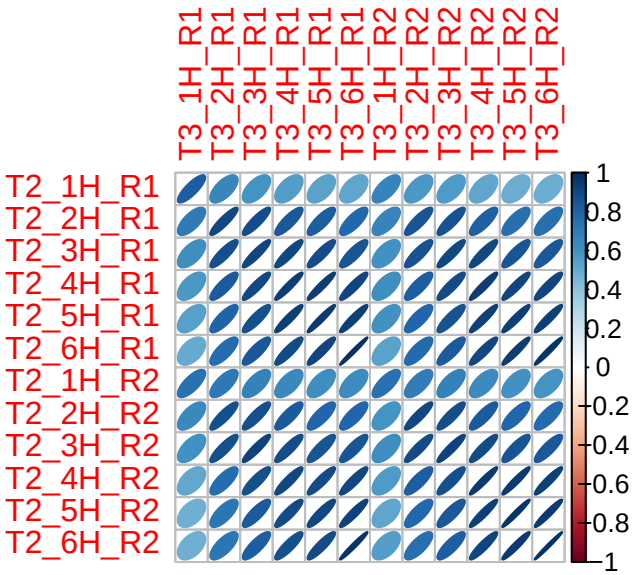

Figure 8: Correlations entre les variables du traitement 2 et du traitement 3 pour les deux réplicats

Ici en comparant l'expression des gènes au cours du temps des différents traitements, on remarque sur `la figure` 6 que le traitement 1 est très peu corrélé avec le traitement 2, voire même corrélé négativement pour la 6ème heure du réplicat 2. Un constat similaire peut être fait vis-à-vis du traitement 3, visible sur `la figure` 7 (bien que moins marqué, puisque le traitement 3 est une combinaison des traitements 1 et 2, et donc plus ou moins fortement corrélé avec le traitement 2 selon les heures), ce qui appuie l'idée que le traitement 1 se comporte différemment des deux autres. Les traitements T2 et T3, quant à eux, sont similaires dans leurs variations.

En effet, sur `la figure` 8, ces deux traitements sont très corrélés positivement, quelle que soit l'heure à laquelle les mesures ont été réalisées. Ils se comportent donc de la même manière et sont similaires en termes d'évolution de l'expression des gènes, confirmant que le traitement T3 hérite principalement des caractéristiques du traitement T2.

De manière générale, à l'exception du premier traitement, qui est peu corrélé positivement, voire négativement à certaines heures avec les deux autres, les variables sont fortement corrélées positivement. Cela implique qu'une forte réduction des dimensions est attendue pour l'ACP.

4 Analyse des Tt_sH_Rr

4.1 Analyse en composantes principales

Ici, les variables de notre jeu de donnée sont dèjà homogènes, elles ont des échelles similaires et il n'est pas nécessaire de centrer/réduire. Dans ce cas, les données sont déjà comparables sans manipulation supplémentaire. De plus, La centrage/réduction pourrait artificiellement altérer les relations naturelles entre les variables.

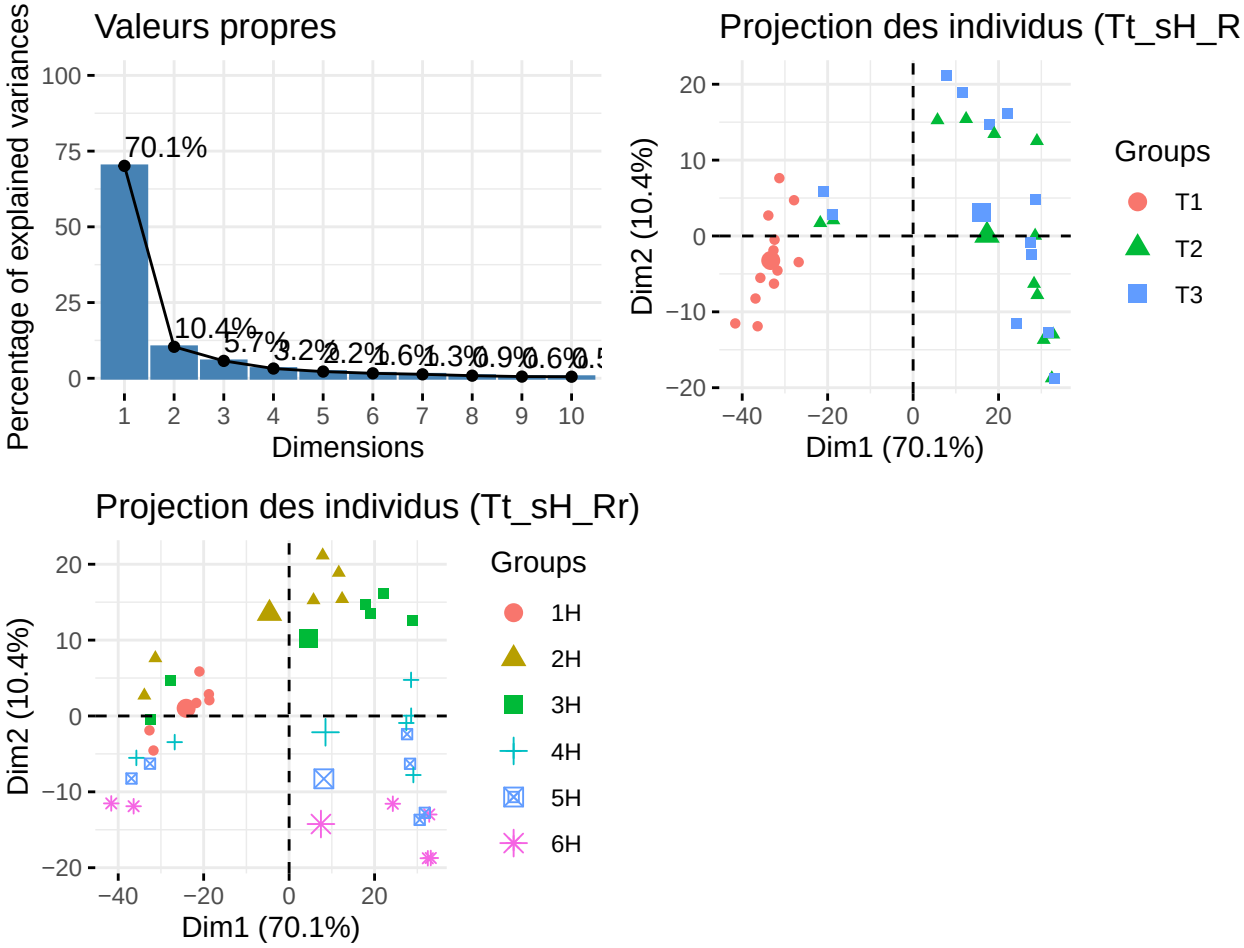

Figure 9: Représentation de l'ACP des Tt_sH_Rr regroupés selon les heures et selon les traitements

En premier lieu, lorsqu'on regarde notre graphe des valeurs propres sur `la figure` 9, on remarque bien que les 2 premières dimensions représentent environ 80% de la variance, ces deux dimensions sont alors suffisantes pour résumer les données. Il ne faut pas oublier pour la suite que la dimension 1 (70,1%) a beaucoup plus "d'informations" que la dimension 2 (10,5%). Chaque point sur notre ACP correspond à une colonne Tt_sH_Rr et donc aux variables.

Ensuite, lorsqu'on observe cette ACP, toujours sur `la figure` 9, en affectant des couleurs pour chaque traitement, on remarque quelque chose de très intéressant. Une fois de plus, on observe des comportements similaires pour les traitements 2 et 3. Cela signifie que les ensembles de données associés à ces traitements ont des caractéristiques similaires, ou que les données sont distribuées de manière proche dans l'espace des variables considérées. En revanche, le centre d'inertie du traitement 1 semble plus éloigné, et le traitement 1 semble être regroupé dans son coin. Lorsqu'on pousse cette analyse sur les heures, on remarque que les

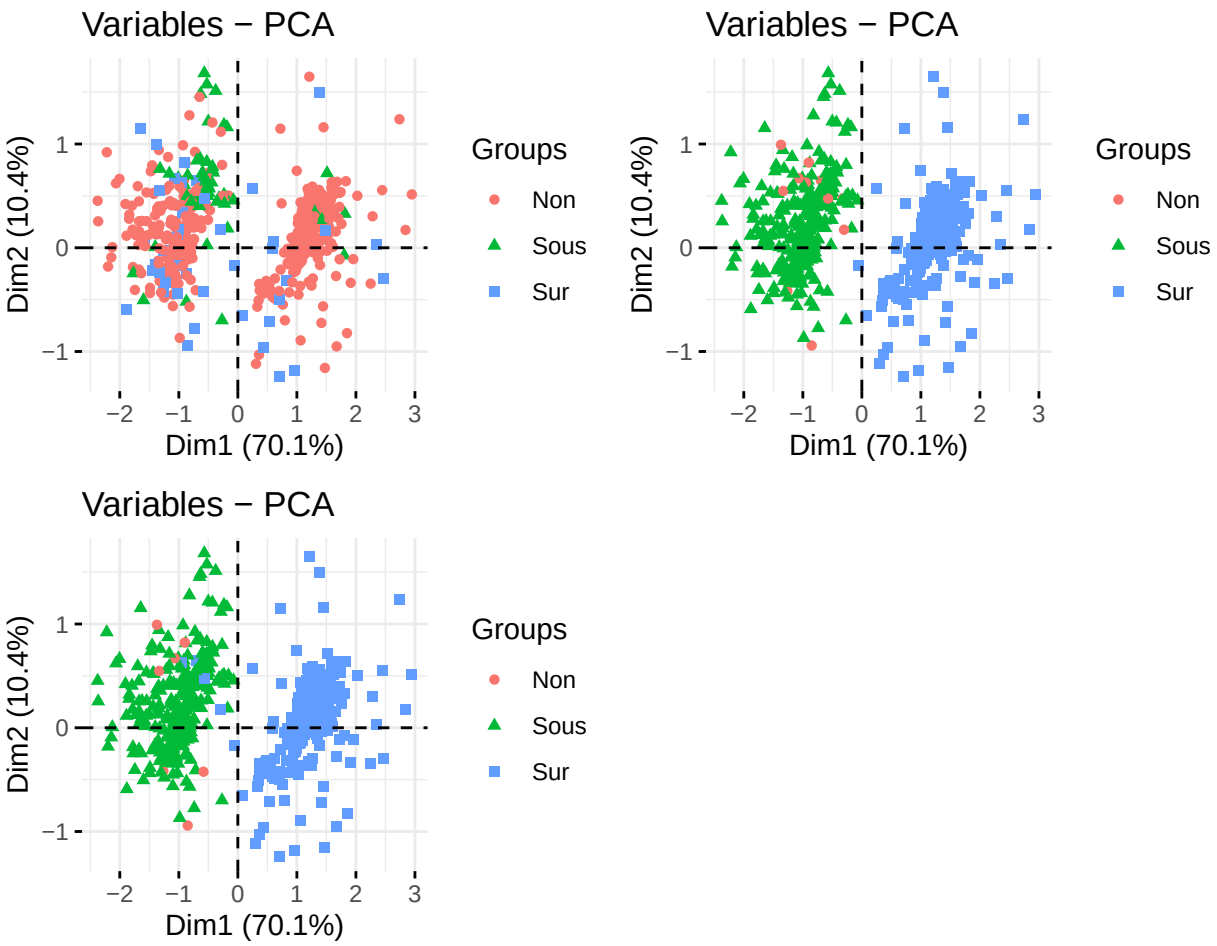

Figure 10: Représentation de l'ACP variable regroupés selon l'expression de leurs gènes pour les traitements 1, 2 et 3, respectivement

données pour l'heure 1 sont toutes très proches, peu importe le traitement auquel elles appartiennent, et qu'elles se comportent toutes de la même manière que le traitement 1 au départ (si l'on compare les deux graphiques). Par ailleurs, nous avons effectué une ACP en groupant les points par réplicats, et, encore une fois, on constate que les deux réplicats sont cohérents, car leurs centres d'inertie sont proches.

Lorsqu'on analyse le graphe des variables sur `la figure` 10, les conclusions sont les mêmes que dans la partie unidimensionnelle concernant la répartition des "Non", "Sous", "Sur" pour les traitements. On remarque que les gènes semblent plutôt bien répartis. On peut en déduire que centrer/réduire les variables n'était donc pas nécessaire.

Cependant, il reste possible d'étudier la contribution relative des individus. Les individus les plus "intéressants" sont ceux qui sont les plus éloignés de l'individu moyen, ainsi que ceux dont la contribution à la dispersion est la plus importante. Ici, on remarque que la variable la plus intéressante est T1_6H_R2. Elle est éloignée du centre, indiquant une réponse atypique ou spécifique par rapport aux autres combinaisons traitement-heure-réplicat. Sa position dans l'espace principal suggère qu'elle joue un rôle important dans la variation observée dans les données.

4.2 Classification non supervisée (Clustering)

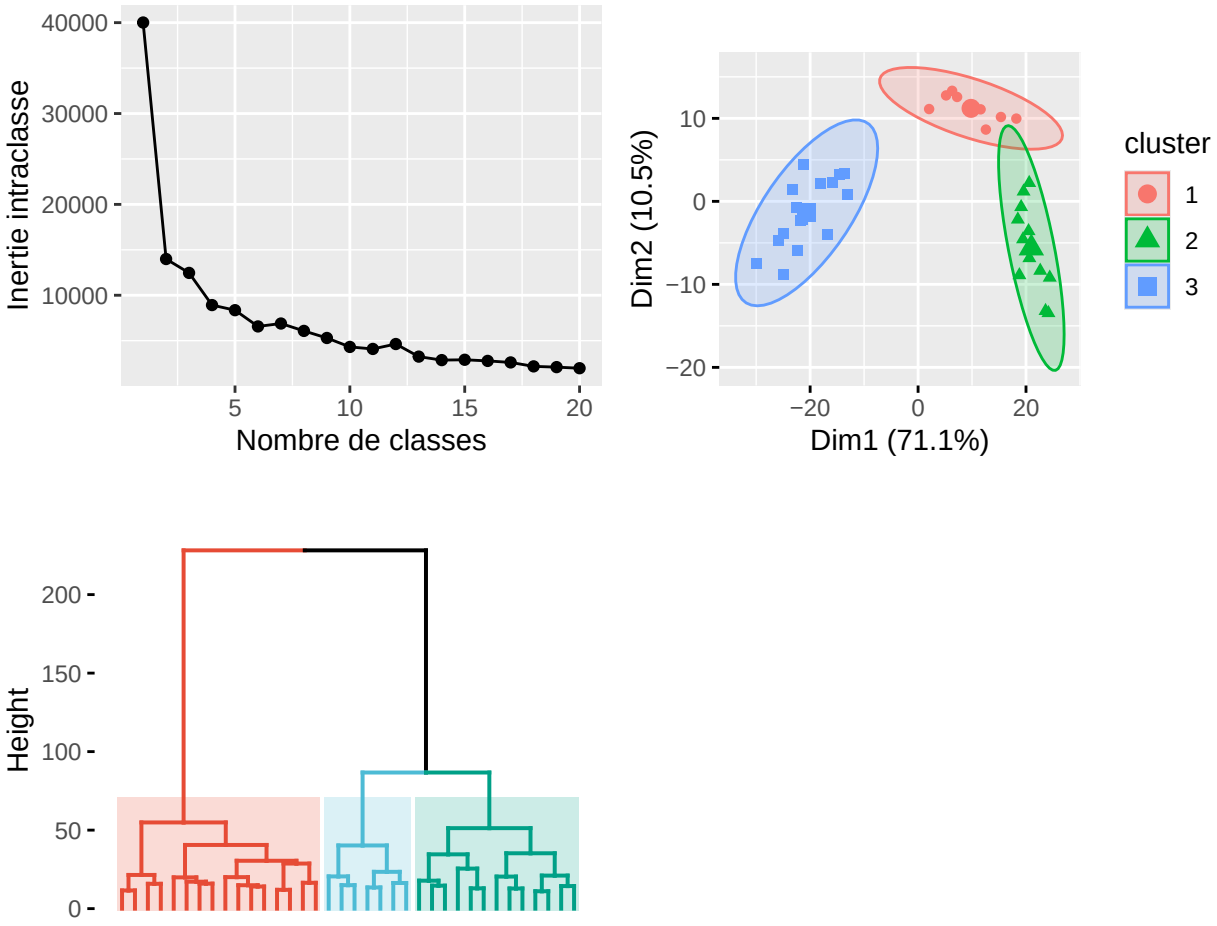

Figure 11: Représentation en cluster homogène de 3 classes sur les Tt_sH_Rr.

En premier lieu, concernant le graphe `La figure` 11 ,on remarque que nos clusters (avec les méthodes K-means et hiérarchique) nous donnent les mêmes résultats : on obtient les mêmes groupes avec le même

nombre d'individus pour les 3 clusters formés.

Dans le cadre du clustering avec la méthode des K-means, on observe que les clusters sont bien distincts, ce qui indique que les groupes d'individus présentent des différences marquées. Ce résultat de clustering n'est d'ailleurs pas surprenant : lorsqu'on le compare à l'ACP basée sur le classement par traitements, on constate qu'une des classes était prévisible. Je parle de la classe "bleue" (située à gauche), qui est composée presque exclusivement du traitement 1. Les seuls points supplémentaires dans cette classe sont au nombre de 4 et correspondent aux traitements 2 et 3 à l'heure 1 pour les deux réplicats.

En ce qui concerne les deux autres classes, elles apparaissent comme des "fusions" des traitements 2 et 3, suggérant une proximité ou une similarité entre ces deux traitements dans les dimensions considérées.

5 Analyse de DataExpMoy et de ExpT

5.1 Analyse en composantes principales de DataExpMoy

Pour la suite de l'analyse, nous allons construire une matrice correspondant à la moyenne des données des expressions des gènes pour chaque heure *DataExpMoy*.

```
##          eigenvalue percentage of variance cumulative percentage of variance
## comp 1  65.16145327            85.37262060                           85.37262
## comp 2   4.70336708             6.16221328                           91.53483
## comp 3   2.98279169             3.90796599                           95.44280
## comp 4   1.05225130             1.37862872                           96.82143
## comp 5   0.67947009             0.89022174                           97.71165
## comp 6   0.56145850             0.73560643                           98.44726
## comp 7   0.27872498             0.36517727                           98.81243
## comp 8   0.24923004             0.32653387                           99.13897
## comp 9   0.14114338             0.18492191                           99.32389
## comp 10  0.12794386             0.16762829                           99.49152
## comp 11  0.08410710             0.11019464                           99.60171
## comp 12  0.07066921             0.09258872                           99.69430
## comp 13  0.05832086             0.07641028                           99.77071
## comp 14  0.04954814             0.06491651                           99.83563
## comp 15  0.04416281             0.05786081                           99.89349
## comp 16  0.03142646             0.04117403                           99.93466
## comp 17  0.02629581             0.03445200                           99.96912
## comp 18  0.02357320             0.03088492                          100.00000
```

En premier lieu, lorsqu'on regarde notre graphe des valeurs propres sur `la figure` 12, on remarque bien que les 2 premières dimensions représentent environ 90-91% de la variance, ces deux dimensions sont alors suffisantes pour résumer les données.

Chaque point sur notre ACP correspond à un gène.

Dans ce cas, il est particulièrement pertinent d'analyser le graphe des variables (`figure` 13), car il semble plus lisible et offre des informations riches et significatives. Une observation très intéressante émerge notamment grâce à la deuxième méta-variable (Dim2). Selon le graphe, cet axe semble refléter l'appartenance des gènes à un traitement spécifique. On constate que :

- Les valeurs positives sur Dim2 correspondent principalement au traitement 1.
- Les valeurs négatives sont plutôt associées au traitement 2.
- En ce qui concerne les gènes du traitement 3, il semble que ses valeurs soient mieux réparties, certaines étant dans le positif et d'autres dans le négatif.

En ce qui concerne le premier axe (Dim1), qui explique 85,4 % de la variance, on observe qu'il est principalement influencé par les traitements 2 et 3 entre 3 h et 6 h. Ces points se situent à l'extrémité droite de la dimension 1. De plus, leur forte contribution est confirmée par leur couleur rouge sur le graphe, indiquant qu'ils jouent un rôle majeur dans l'explication de la variance sur cet axe.

- Les valeurs positives sur Dim1 correspondent principalement aux traitements 2 et 3. On peut donc en déduire que les valeurs positives de Dim1 correspondent au comportement des traitements 2 et 3, à l'opposé du traitement 1, quasiment nul sur cette dimension (confirmant les résultats précédents).

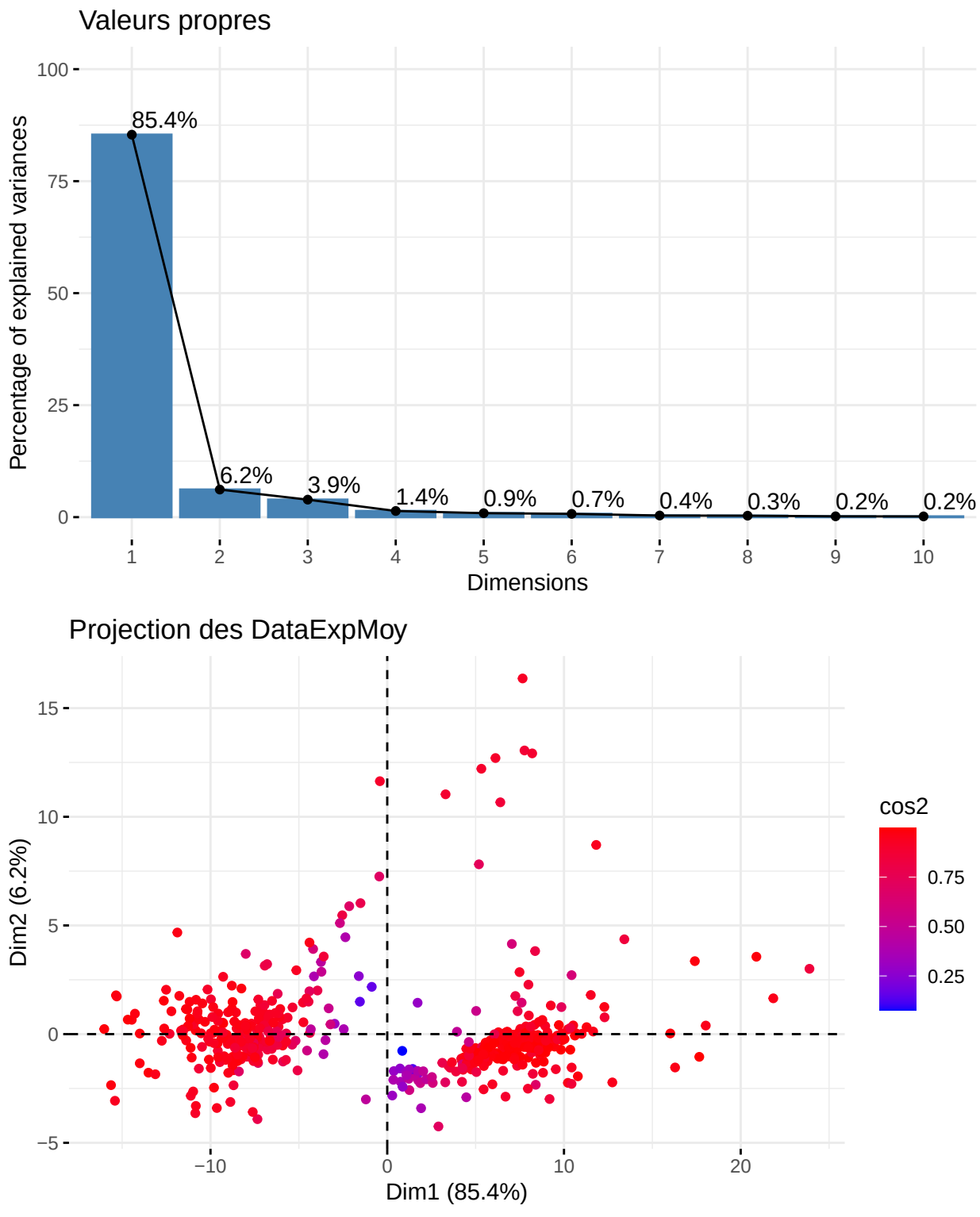

Figure 12: ACP de DataExpMoy

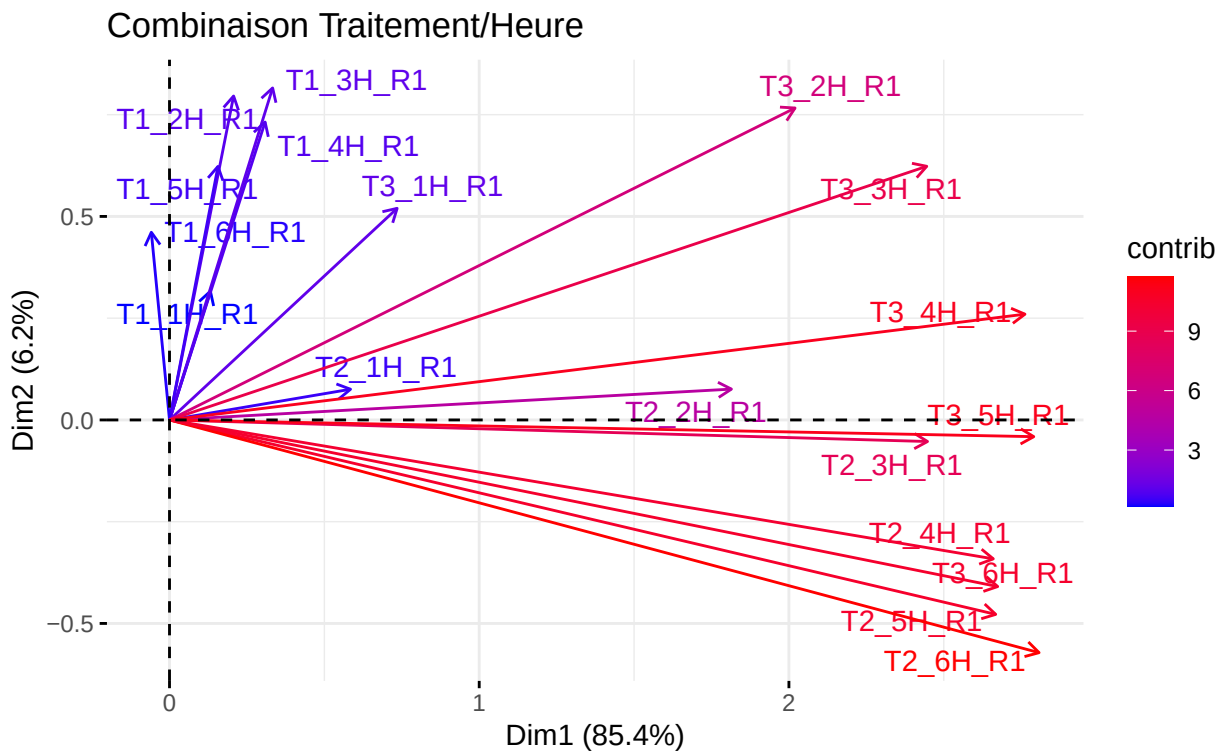

Figure 13: Projection des variables de l'ACP dans les dimensions 1 et 2

5.2 Classification non supervisée (Clustering) de DataExpMoy

Lorsque l'on utilise un diagramme alluvial entre nos 6 clusters (`figure` 14) et les variables qualitatives ExpT1,ExpT2 et ExpT3 (`figure` 15). On remarque clairement une connexion et un lien entre ces données. Pour ExpT2 et ExptT3, encore une fois très similaires, on remarque que les clusters ExpMoy3 et ExpMoy6 sont liés avec la valeur "Sur". Ce qui signifie que les gènes appartenant à ces clusters montrent une forte surexpression dans les traitements T2 et T3. Ensuite, ExpMoy4 et ExpMoy5 et ExpMoy2 sont liés à "Sous". Ce qui indique que ces gènes sont sous-exprimés dans les traitements T2 et T3 à 6 heures. Enfin, ExpMoy2 est lié à "Non", Cela signifie que les traitements 2 et 3 sont inefficaces face à ses gènes. Pour ExpT1, la majorité des clusters sont liés à la classe "Non", ce qui signifie que le traitement 1 semble plutôt inefficace. Cependant, on remarque que ExpMoy2 est lié avec "Sur", ce qui montre l'efficacité de ce traitement pour ces gènes.

Cette analyse montre que si l'on arrive à relier un gène à un certain cluster alors on sait quelle traitement utiliser pour les gènes soient surexprimés. Si le gène appartient aux clusters ExpMoy3 ou ExpMoy6 alors il est préférable d'utiliser les traitements 2 ou 3. Si le gène appartient au cluster ExpMoy2 alors il est préférable d'utiliser le traitement 1. Si le gène appartient à ExpMoy4, il est préférable d'utiliser le traitement 1 car même si ce cluster est faiblement lié à "Sur", il s'agit du seul traitement qui peut avoir un effet positif ou nul sur le gène. Si le gène appartient au cluster ExpMoy1 ou ExpMoy5, il semble qu'aucun traitement ne puisse surexprimé le gène.

5.3 Classification non supervisée (Clustering) de ExpT

Ce clustering en deux classes, visible sur `la figure` 16, nous permet de voir avec plus de précision le comportement des trois traitements, divisés en deux grandes classes de points relativement éloignés les uns des autres. On pourrait donc ici en déduire, à partir des résultats des analyses précédentes, que l'une regroupe

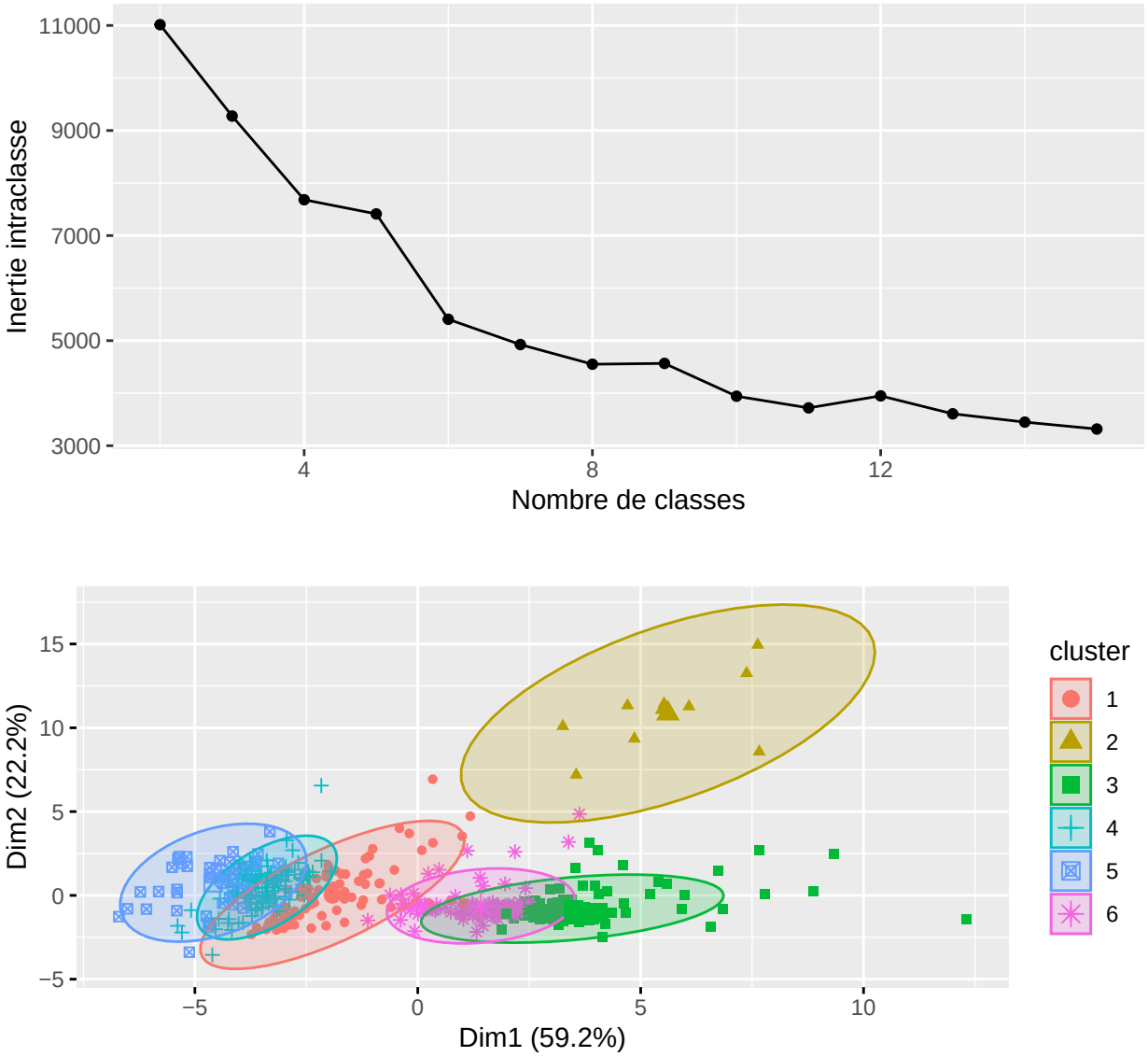

Figure 14: Clustering de DataExpMoy

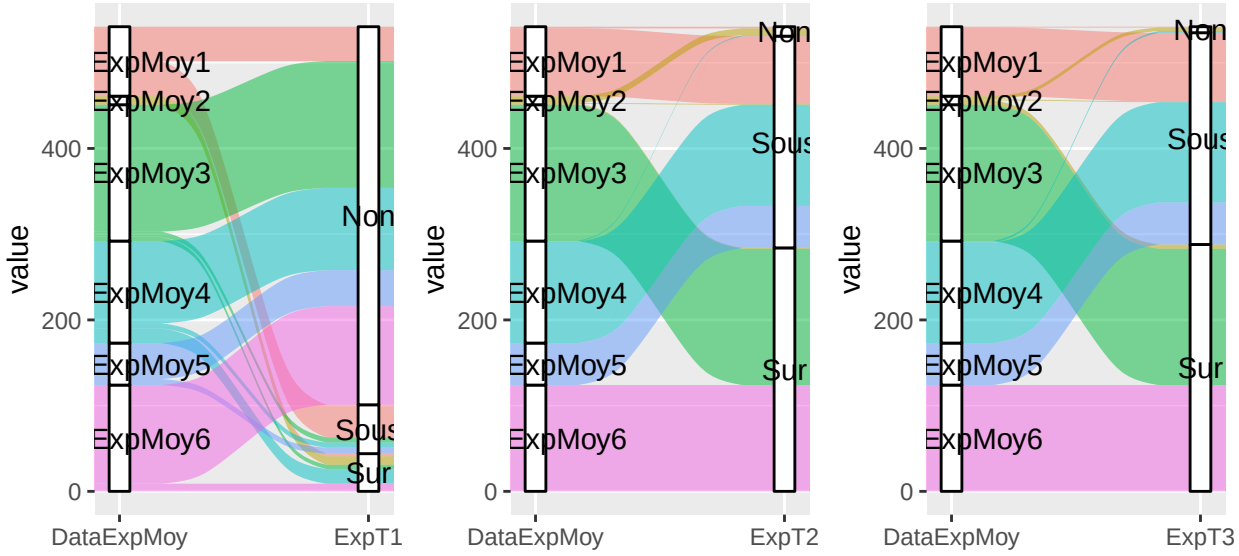

Figure 15: Alluvial diagram de la répartition des différents clusters en fonction des expressions des gènes

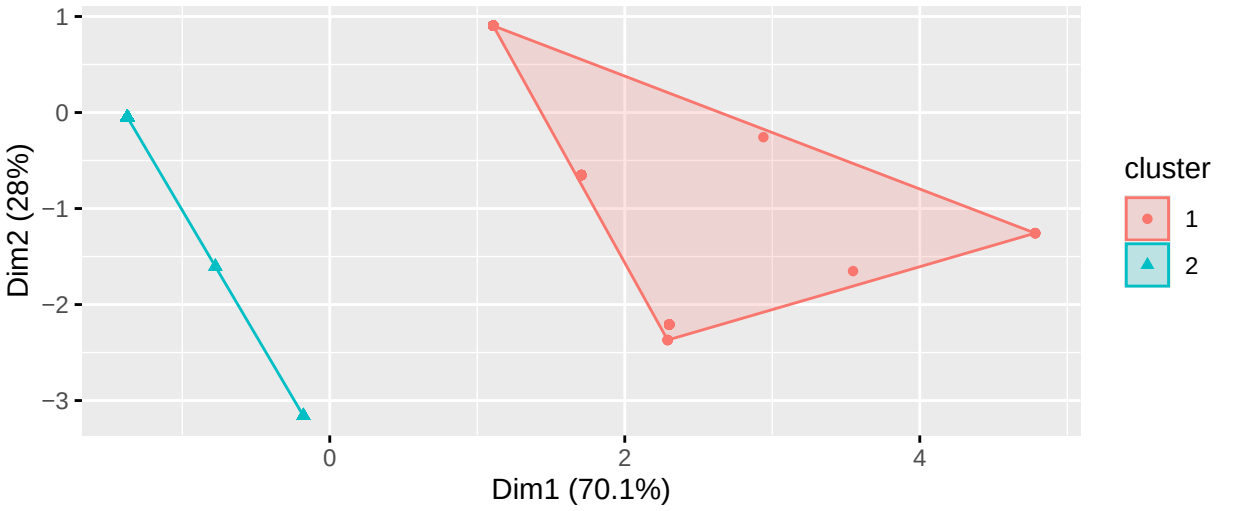

Figure 16: Cluster des gènes à partir des variables ExpT

les gènes dont l'expression au cours du temps suit le comportement des traitements 2 et 3, et l'autre regroupe les gènes dont l'expression suit le comportement du traitement 1.

Si nous comparons les résultats obtenus avec l'analyse précédente utilisant *DataExpMoy*, qui contient 6 clusters et deux dimensions principales pour l'ACP, ces résultats mettent en évidence les comportements concernant l'évolution de l'expression des gènes selon le traitement utilisé.

Ainsi, ce clustering, réalisé à partir des informations sur l'expression des gènes à 6 heures pour chaque traitement, met en évidence les mêmes résultats que ceux constatés sur le diagramme alluvial (`figure` 15) et plus généralement tout au long de l'analyse. Ces résultats soulignent une très forte similarité, pour un grand nombre de gènes, au niveau de leur expression chez les plantes ayant expérimenté le traitement 2 et chez celles ayant expérimenté le traitement 3, ainsi qu'une forte différence au niveau des gènes des plantes ayant expérimenté le traitement 1.

Le cluster de `la figure` 16 représenterait donc les gènes en fonction de si leur expression, 6 heures après le traitement, se rapproche ou non du comportement général issu du traitement 3 (ou du traitement 2) ou du traitement 1, moins efficace et présentant des divergences (comme l'expression des gènes à la 6ème heure pour le réplicat 2). Cela nous montrerait comment réagissent les gènes face aux traitements et nous permettrait de choisir le meilleur traitement (voire de mieux comprendre le fonctionnement de certains gènes afin de créer de nouveaux traitements adaptés à nos besoins) en fonction des résultats attendus de l'expression des gènes chez la plante.

6 Conclusion

Pour conclure, voici un bilan des principales hypothèses démontrées au cours de cette analyse de données. Tout d'abord, nous avons montré que les réplicats 1 et 2 étaient cohérents, grâce à l'ACP sur Tt_sH_Rr ainsi que l'analyse des variables qualitatives et quantitatives. Ensuite, nous avons mis en évidence les fortes similarités entre les traitements 2 et 3, qui présentent des comportements similaires. En revanche, nous avons observé que le traitement 1 agissait de manière différente, notamment à l'heure 6 dans le réplicat 2 (T1_6H_R2), une variable dont l'expression était atypique. Cette observation avait déjà été soulevée dans la matrice de corrélation, où nous avions remarqué que cette variable avait une corrélation négative avec les traitements 2 et 3, indiquant un comportement distinct. En outre, dans l'ACP, cette variable se situe loin du centre, jouant ainsi un rôle important dans la variation des données.

De plus, lors de l'analyse des expressions des gènes, nous avions déjà pu déduire, dans la partie bi-dimensionnelle sur les variables qualitatives, que le traitement 1 serait moins efficace que les autres, notamment en raison du grand nombre de gènes non exprimés au bout de six heures. Par la suite, grâce à l'ACP, au clustering basé sur les moyennes d'expression des réplicats de chaque gène et au diagramme alluvial, nous avons observé une corrélation entre le clustering et l'expression des gènes selon le traitement. Ce diagramme nous a permis de conclure que, bien que le traitement 1 ne soit pas aussi efficace que les traitements 2 et 3, il reste cependant pertinent dans certains cas, notamment pour les gènes appartenant aux clusters ExpMoy2 et ExpMoy4.