



Master 2 Machine Learning for Data Science

TP1

Reconstruction et analyse d'un réseau de gènes

Réalisé par:

Nadia RADOUANI – 21911973 FI

Amira KOUIDER - 21904040 FI

INTRODUCTION:

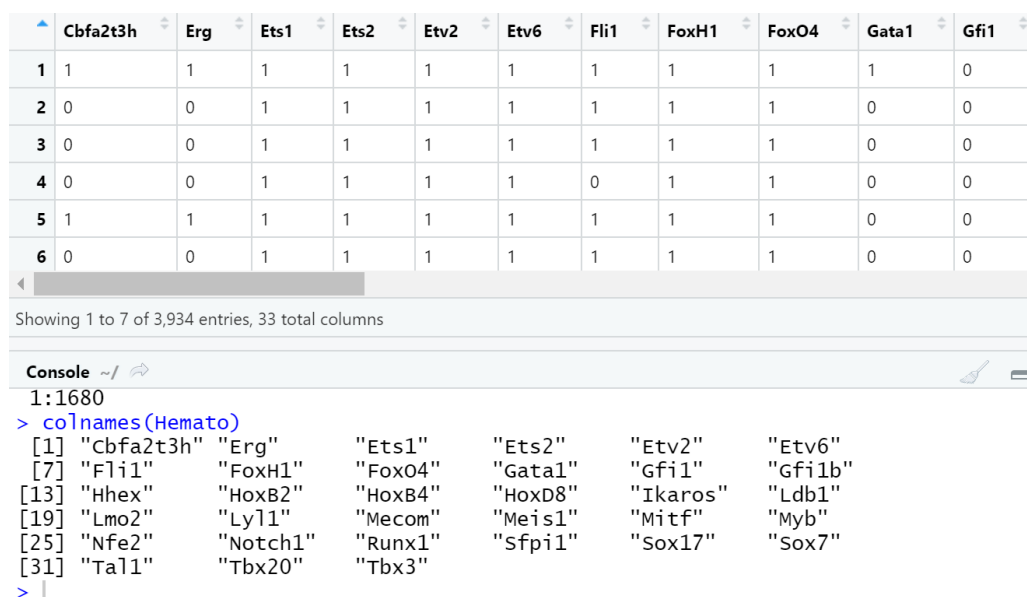
Afin de mieux comprendre le fonctionnement des gènes, les différents mécanismes parfois complexes des maladies et d'importants processus biologiques des organismes vivants, l'outil de reconstruction des réseaux de gènes est fortement utilisé. Or cela constitue une tâche très difficile car il est important d'identifier des relations de régulation entre les facteurs de transcription et les gènes cibles. Cette action s'acquies suite à une perturbation des gènes ce qui n'est pas très éthique.

Notre travail consiste donc à créer un réseau de régulation à partir d'un jeu de données *HematoData* : des données d'expression binarisées de 33 facteurs de transcription impliqués dans la différenciation précoce des cellules érythroïdes et endothéliales primitives (3934 cellules)¹ extrait du package *miic*.

TF CORRELATION NETWORK:

1. Exploration de la base:

La base utilisée est une base binaire qui contient 33 facteurs de transcription en colonnes et 3934 cellules en lignes.



	Cbfa2t3h	Erg	Ets1	Ets2	Etv2	Etv6	Fli1	FoxH1	FoxO4	Gata1	Gfi1
1	1	1	1	1	1	1	1	1	1	1	0
2	0	0	1	1	1	1	1	1	1	0	0
3	0	0	1	1	1	1	1	1	1	0	0
4	0	0	1	1	1	1	0	1	1	0	0
5	1	1	1	1	1	1	1	1	1	0	0
6	0	0	1	1	1	1	1	1	1	0	0

Showing 1 to 7 of 3,934 entries, 33 total columns

```
Console ~/ |
1:1680
> colnames(Hemato)
[1] "Cbfa2t3h" "Erg"      "Ets1"     "Ets2"     "Etv2"     "Etv6"
[7] "Fli1"      "FoxH1"    "FoxO4"    "Gata1"    "Gfi1"     "Gfilb"
[13] "Hhex"      "HoxB2"    "HoxB4"    "HoxD8"    "Ikaros"    "Ldb1"
[19] "Lmo2"      "Ly11"     "Mecom"    "Meis1"    "Mitf"      "Myb"
[25] "Nfe2"      "Notch1"   "Runx1"    "Sfp1"     "Sox17"     "Sox7"
[31] "Tal1"      "Tbx20"    "Tbx3"
```

Figure 1 – Aperçu de la base

¹ <https://rdrr.io/cran/miic/man/hematoData.html>

2. Calcul de corrélation de toutes les paires:

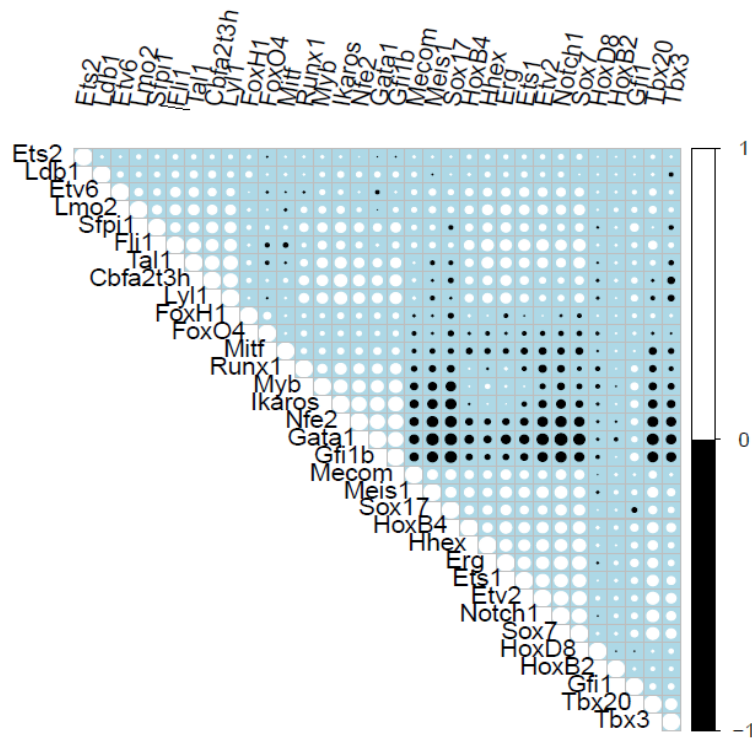


Figure 2 – Corrélation

On remarque qu'il y a des faibles corrélations entre les variables et on peut distinguer des corrélations négatives et d'autres positives. On remarque aussi que les deux variables *HoxB2* et *HoxD8* sont très faiblement corrélées avec toutes les autres variables.

3. Plot le réseau de corrélation:

Afin de s'assurer de garder des liens entre les variables, le seuil ne doit pas être trop grand par rapport à la moyenne des coefficients de corrélation des variables ($\text{mean}(r) = 0.13$), car, par exemple, à partir du seuil 0.1, on commence à avoir des variables qui n'ont aucun lien avec les autres (Cf Figure 7). D'où le choix de nos seuils.

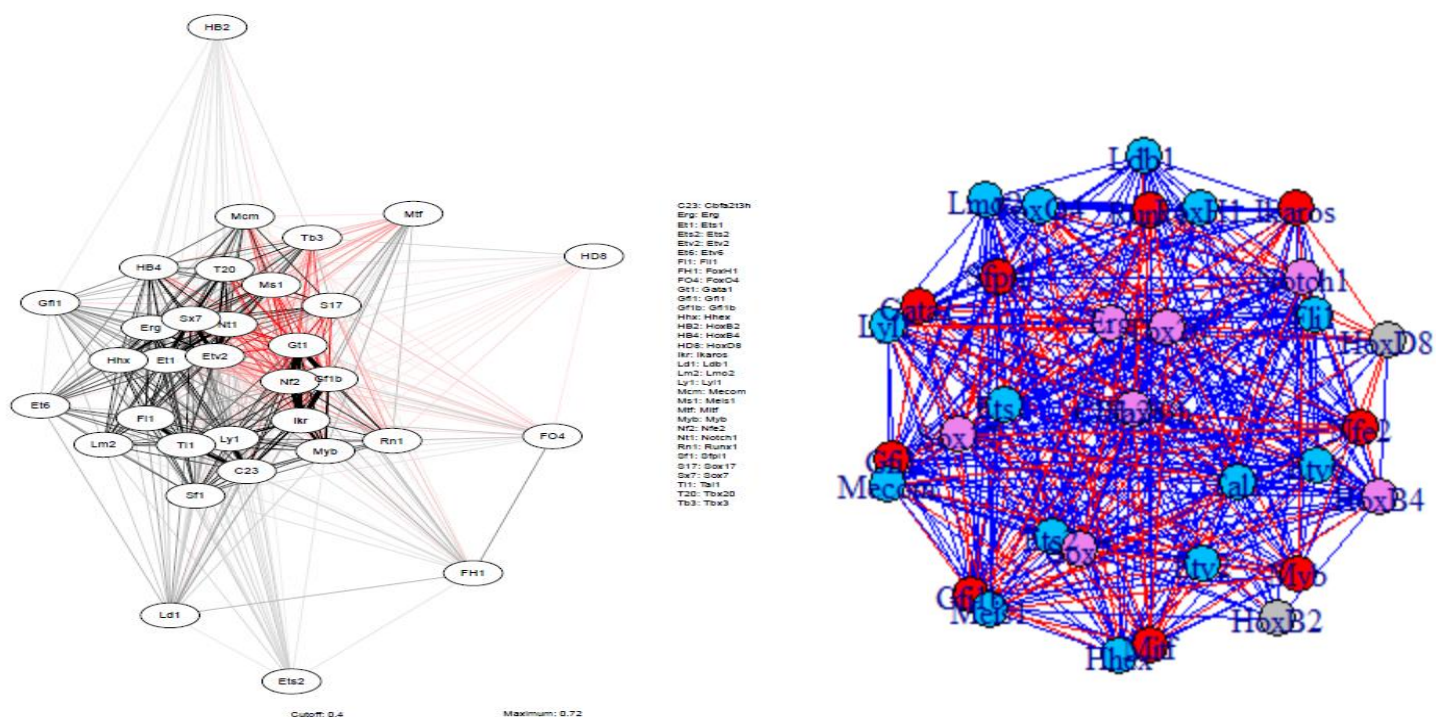


Figure 5 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.02)

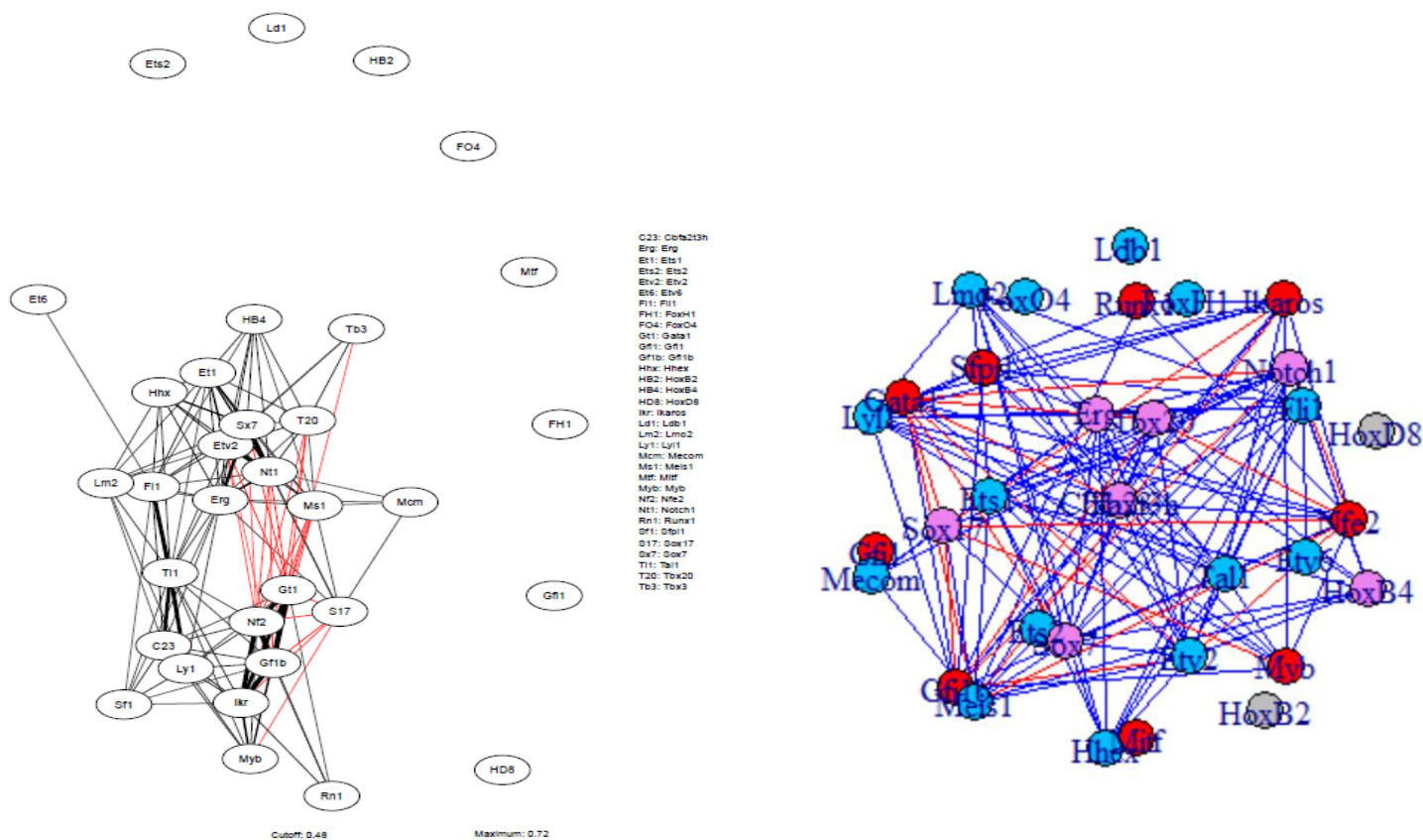


Figure 6 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.3)

4. CONCLUSION:

Après avoir calculé et tracé la corrélation totale en utilisant plusieurs seuils, on peut distinguer des corrélations fortes positives (en noir foncé) et d'autres fortes négatives (en rouge foncé), par ailleurs on remarque d'autres variables avec une légère corrélation positive/négative. On peut distinguer aussi la variable HoxD8 qui n'est corrélée avec aucune autre variable.

Le graphe ressemblant le plus à celui de Verny et al est celui obtenu avec un seuil = 0.05.

TF PARTIAL CORRELATION NETWORK:

1. Calcul de la covariance et la matrice inverse:

Le déterminant de la matrice de covariance tend fortement vers 0 ($1.078972e-37$). Afin de calculer l'inverse de la matrice de covariance, on fait une régularisation de la matrice en ajoutant une petite valeur λ à sa diagonale.

2. Plot les réseaux de la corrélation partielle:

a) $\lambda = 10^{-5}$

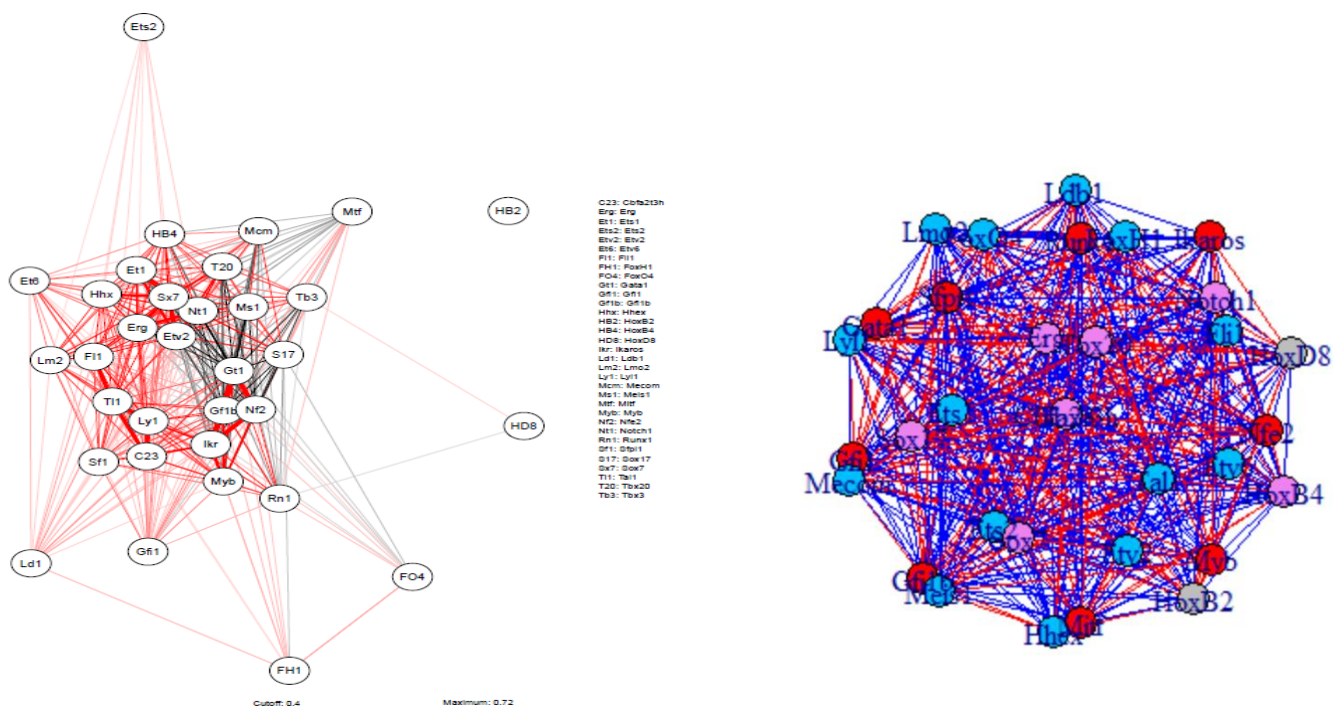


Figure 7 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.05)

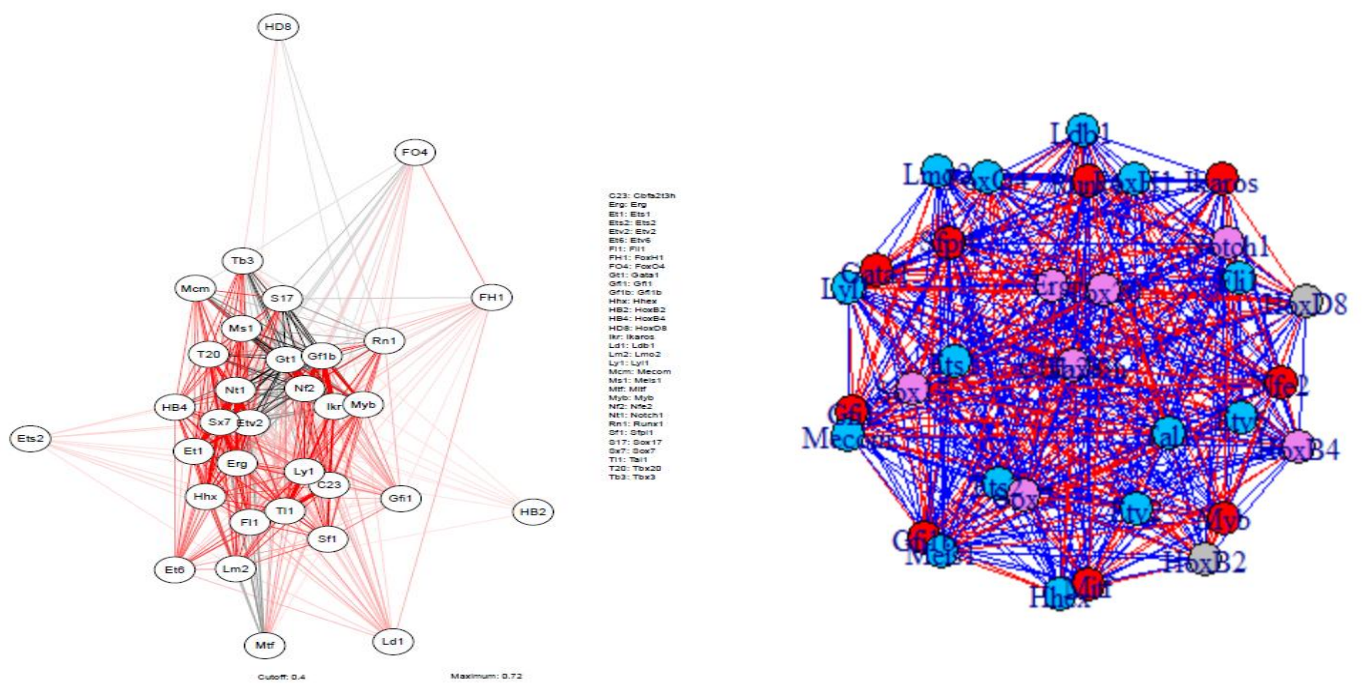


Figure 8 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.04)

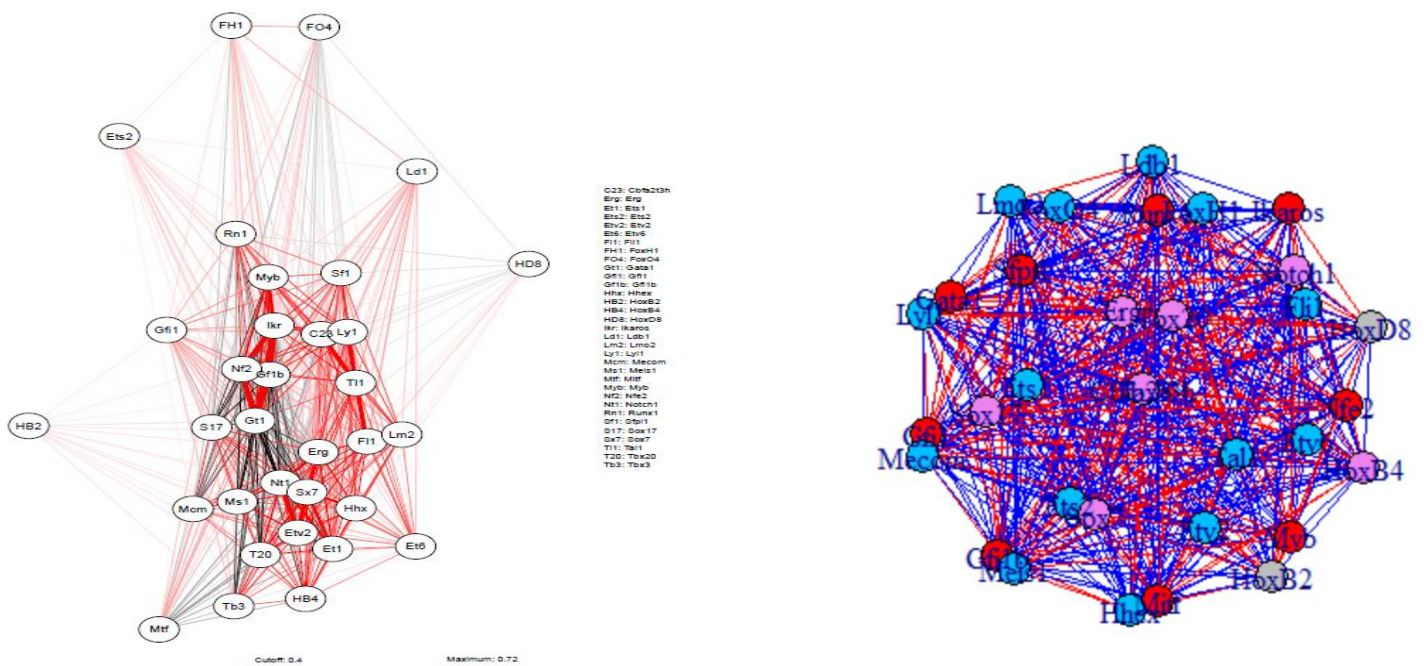


Figure 9 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.02)

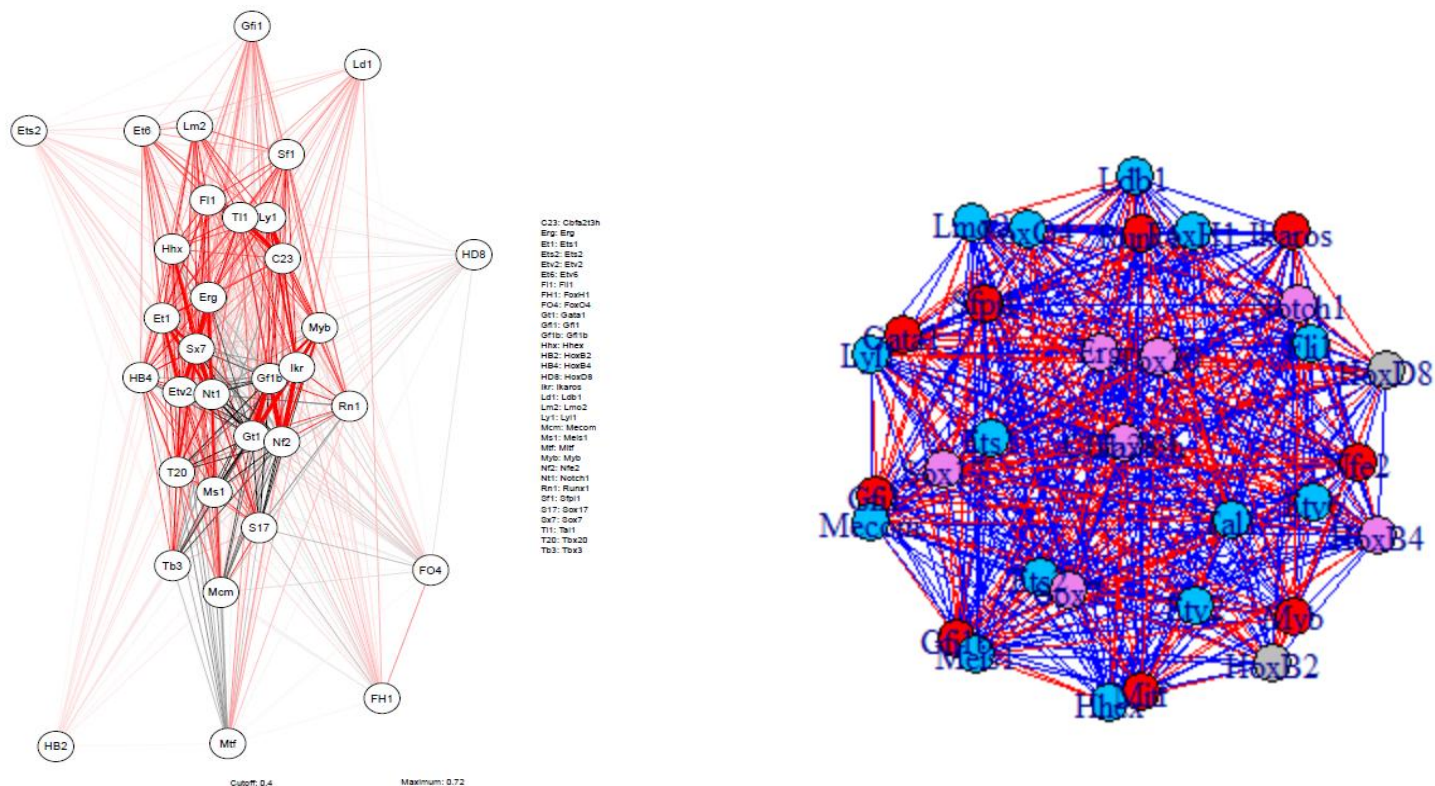


Figure 10 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.01)

b) $\lambda = 10^{-10}$

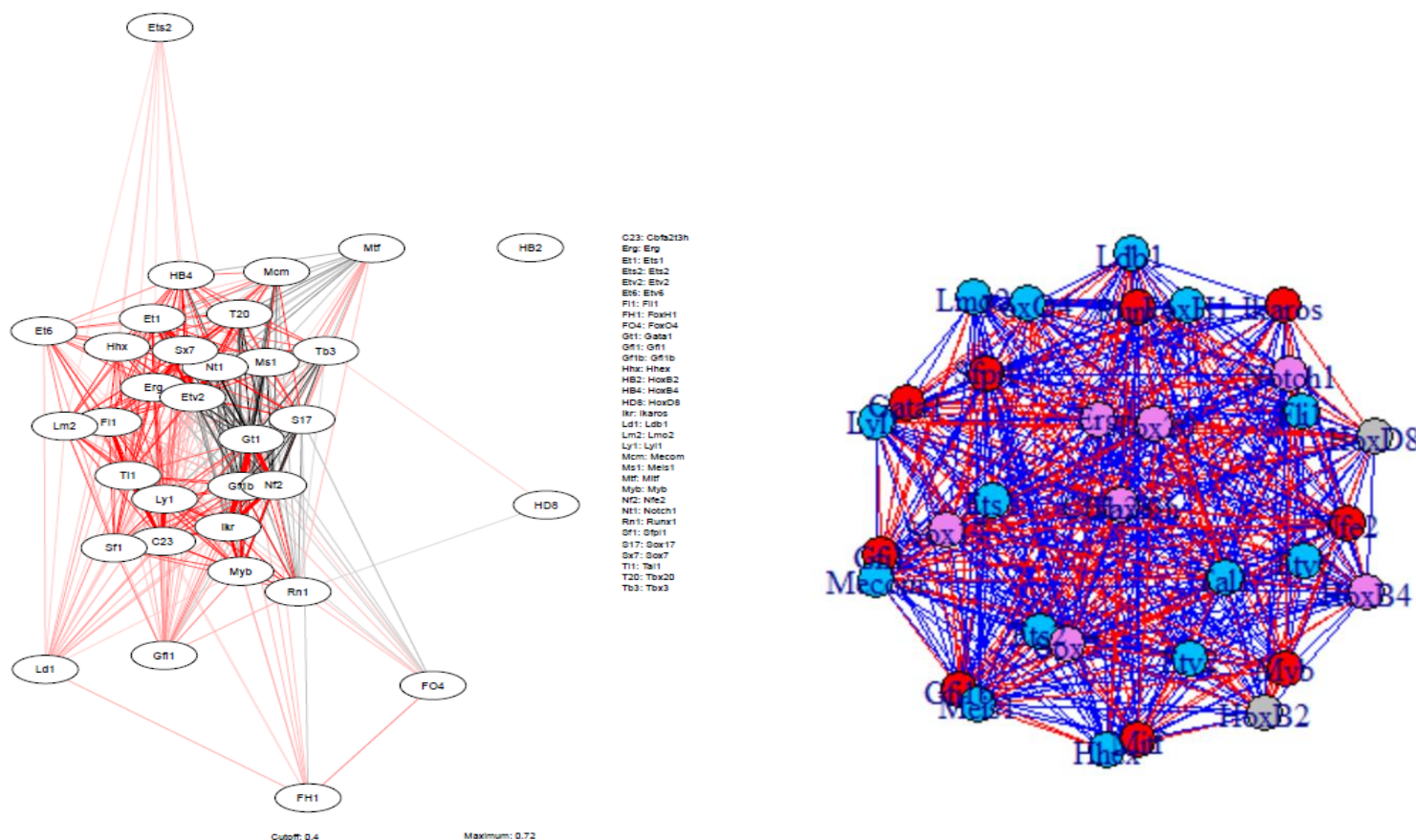


Figure 11 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.05)

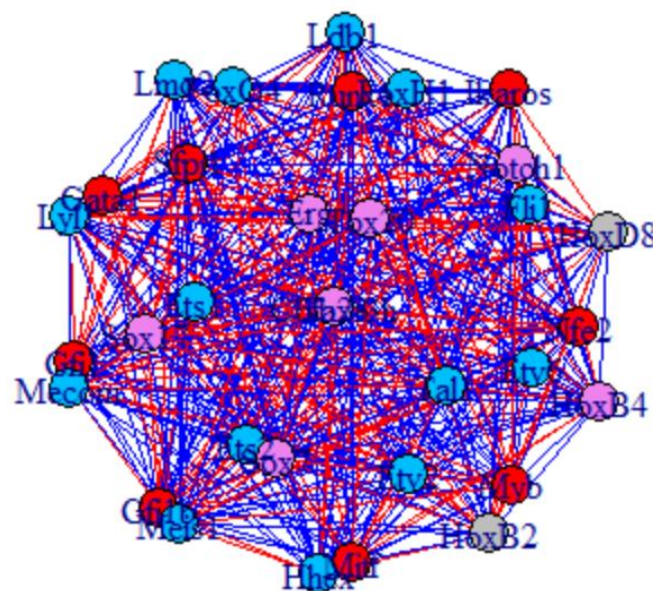


Figure 13 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.02)

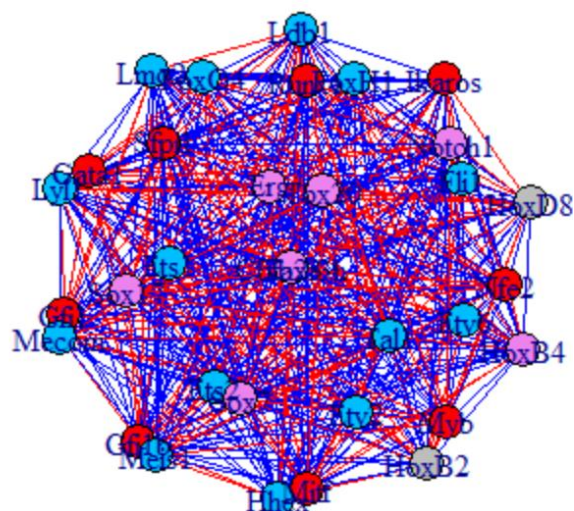


Figure 13 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.02)

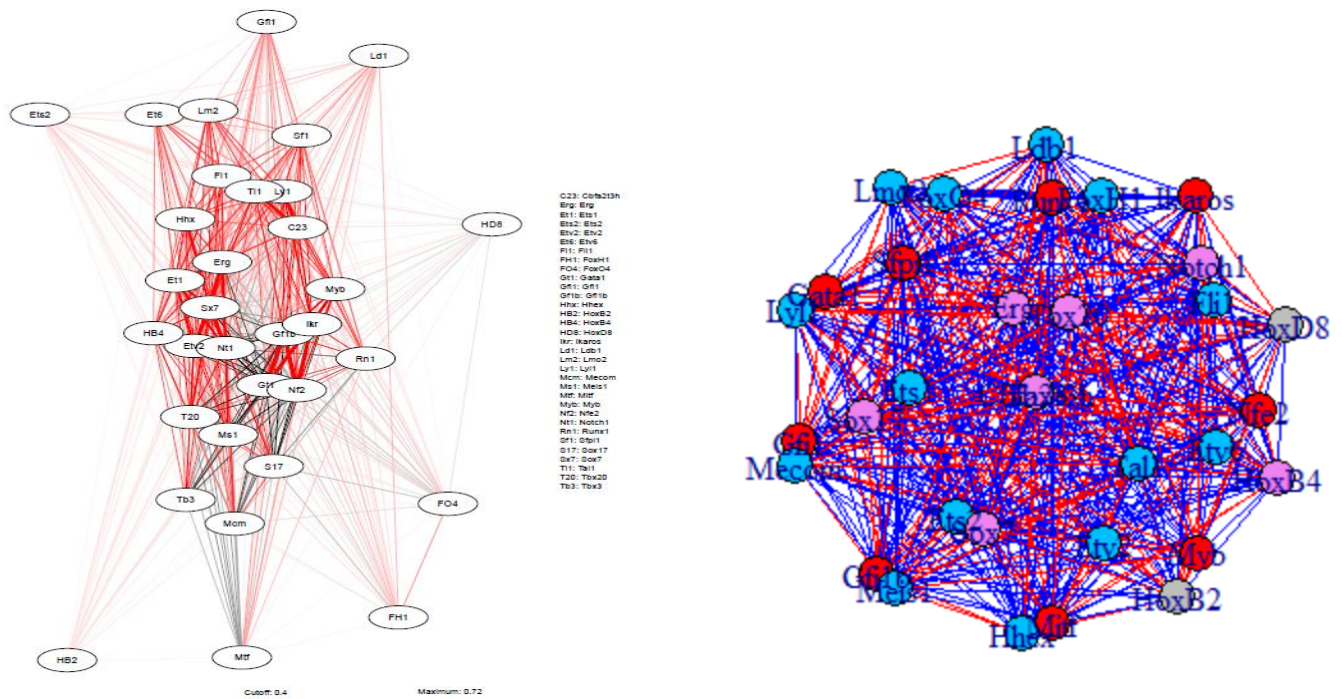


Figure 14 – Corrélation avec package *qgraph* à gauche et *igraph* à droite (seuil 0.01)

3. CONCLUSION:

Pour la corrélation partielle, on a utilisé les valeurs du seuil utilisées auparavant avec différentes valeurs de lambda. On remarque qu'au-delà du seuil 0.04 pour les deux valeurs de lambda, on commence à avoir des variables qui ne sont corrélées avec aucune des autres variables.

On remarque aussi que les graphes des corrélations partielles pour les deux valeurs de lambda sont dominés par des liens négatifs.