Réalisé par: Jean Sutter, Sylvan Rossier , sous la direction de: Adam Alexander T. Smith
Pour le cours: "Solving Biological Problems that require Math 2016"

Investigating gene expression across the cell cycle using single-cell RNA-seq data

Réalisé par: Jean Sutter, Sylvan Rossier , sous la direction de: Adam Alexander T. Smith
Pour le cours: "Solving Biological Problems that require Math 2016"

Introduction

Contexte

- Cycle cellulaire:

  Le cycle cellulaire se divise en 4 phases, G1 est un stade de croissance, S celui de la duplication du génome , G2 prépare la mitose,et M est le stade de la division cellulaire.
  L'expression des gènes est différente selon les stades.

- Long ARN non-codant ("lncRNA")

  Les longs ARN non-codants comme leur nom l'indique sont des ARN ne codant pas et faisant plus de 200 nucléotides de long. Contrairement aux courts ARN non-codants qui font moins de 200 nucléotides. Les lncRNA sont très nombreux environ 3 fois plus que les mRNA (codant pour les protéines). Ils ont souvent des fonctions liées à la régulation de l'expression des gènes, ils ne sont encore que très peu étudiés.

- Méthode d'RNA-seq ("single cell", sur cellules isolées)

  Le single cell RNA-seq est une nouvelle technique qui permet le séquençage de l'ARN d'une cellule unique et non pas celui d'une population (environ 100'000 cellules). Ce procédé amène certains avantages comme le fait de ne plus travailler avec des moyennes d'expression, mais il y a néanmoins de nouveaux problèmes qui apparaissent (augmentation du bruit). Dans le single cell classique, on isole la cellule,on extrait l'ARN, on l'amplifie avec du cDNA par PCR. Pour finir, on séquence pour obtenir un jeu de données à analyser.

Objectifs

- Explorer les caractéristiques de nouvelles méthodes
- Parmi les transcriptions, déterminer s'il y a des «clusters» avec des motifs d'expression similaires
- Caractériser les différents motifs d'expression
- Analyse fonctionnelle d'enrichissement

Méthodologie

Préparation des données de RNA-seq sur cellules isolées

L'expérimentation porte sur des échantillons contenant chacune une seule cellule souche embryonnaire de souris isolée, dont le stade du cycle cellulaire est connu. Ces cellules sont obtenues par triage par cytométrie en flux sur base de marquage au Hoechst. Ce dernier marque l'ADN, dont la quantité varie selon le cycle cellulaire, permettant de différentier les cellules au stade G1, S et G2/M. Les cellules de chaque stade sont isolées grâce à une puce microfluidique Fludigm "C1" et y sont préparées pour le séquençage. Les détails de cette préparation sont disponibles dans les données supplémentaires de la publication originale.

Suite au séquençage de l'ARN extrait des échantillons, les données brutes obtenues sont des "lectures" (séquences de nucléotides). Pour chaque lecture, l'origine dans le génome de l'organisme étudié est retrouvé par une étape dite de "mapping". L'expression de chaque gène ou transcrit est alors quantifié à partir du nombre de lectures qui leur correspondent. Finalement, les gènes ou transcrits ayant des niveaux d'expression statistiquement différents entre groupes expérimentaux sont identifiés de par une analyse différentielle. Divers contrôles de qualité interviennent à chaque étape.

Dans cette étude, le mapping et la quantification ont été réalisées avec un algorithme récemment développé, Kallisto, avec une prise en compte des lncRNAs. L'analyse différentielle à été faite en utilisant un logiciel compagnon de Kallisto, Sleuth. Ce projet porte sur l'étude des résultats ainsi obtenus, qui sont:

- une matrice contenant pour chaque transcrit (en ligne) et pour chaque cellule (en colonne) un nombre de lectures.
- une matrice similaire effectuée sur des "bootstraps" (voir ci-dessous) des données de séquençage.
- pour chaque comparaison de stades (G1 vs S, S vs G2/M, G1 vs G2/M), une liste des transcrits trouvés comme étant différentiellement exprimés.

Normalisation

Les données de compte de lectures issues du RNA-seq ne suivent typiquement pas des distributions normales (Gaussiennes), ce qui rend leur analyses statistique difficile. Nous avons donc transformé nos données afin

de les rendre plus facile à analyser. Une transformation simple et couramment utilisée que nous avons appliquée ici est de prendre le “shifted log10 expression”: log10(expr+1).

ACP (Analyse en composantes principales)

L’ACP permettra une interprétation et visualisation faciles pour un grand nombre de variables quantitatives, de par une représentation en nuage de points que l’on peut observer dans un nombre réduit et choisi de dimensions. Le but d’une ACP est de simplifier et réduire les données pour donner une vue d’ensemble et il sera alors plus aisé par exemple de repérer les outliers, des regroupement d’échantillons, des gradients ou des trajectoires correspondant au plan expérimental.

Classification ascendante hiérarchique

La CAH (ou classification ascendante hiérarchique) est un moyen de regrouper des objets sur la base d’une similarité entre celles-ci. Ces regroupements sont ascendants (on regroupe graduellement de plus en plus d’objets) et hiérarchiques (les groupes sont imbriqués et peuvent donc être représentés sous forme d’arbre).

Les dendrogrammes (ou arbres) permettront alors de mettre en évidence le regroupement soit des transcrits, soit des cellules similaires. Il y a deux choix à faire pour faire une CAH: la distance/similarité entre objets, et l’algorithme pour regrouper les objets. Une fois l’arbre obtenu, un dernier choix repose sur comment “couper” celui-ci pour obtenir des groupes d’objets.

Ici, nous avons essayé comme distance transcrit-transcrit, la corrélation entre ceux-ci, ainsi que la valeur absolue ou le carré de cette corrélation. Nous avons testé les algorithmes comme Single, Ward.D2 ou encore Complet. Deux possibilités de découpe d’arbre s’offraient à nous. Nous pouvons effectuer des découpes à une hauteur h déterminée, ou alors une découpe en différents k groupes (clusters). Ces différentes découpes nous ont permis d’isoler des listes de gènes. Ce qui permettra alors d’effectuer des analyses d’enrichissement fonctionnel.

Enrichissement fonctionnel

Une annotation fonctionnelle associe à un gène la description d'une fonction biologique. Partant d'un ensemble de gènes, on peut alors regarder si ceux-ci contiennent davantage de certaines annotations fonctionnelles qu'attendues par hasard. Il existe différentes plateformes pour faire de l'enrichissement fonctionnel, par exemple ToppGene et DAVID.

ToppGene

ToppGene est une suite d'outils disponibles sur internet (disponible sur : https://toppgene.cchmc.org/) qui permettent d'accéder à une base de données. Dans ce travail, nous avons utilisé l'outil ToppFun qui permet une analyse d'enrichissement d'annotations fonctionnelles au sein de listes de gènes.

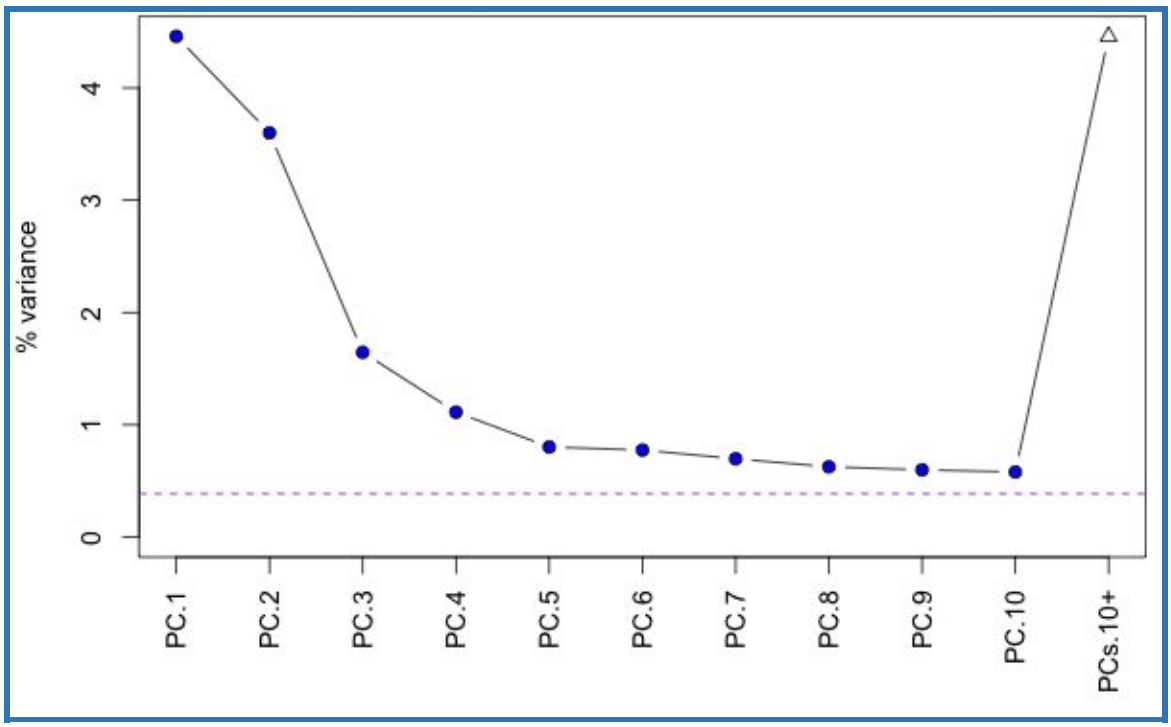
Résultats

Analyse en composante principale (ACP) de tous les transcrits

Cet ACP (centré réduit) est réalisé sur toutes les cellules et tous les transcrits.

Scree plot

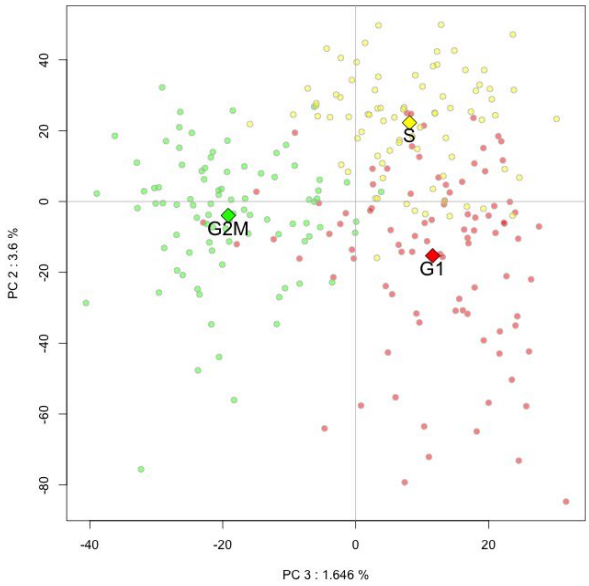
Le “scree plot” permet de visualiser le pourcentage de la variabilité totale présente dans les données qui est capturé par chaque composant principal.



=> il y a clairement de 2 à 4 composants majeurs. L’étude des graphes des observations permettra potentiellement d’interpréter les signaux à l’origine de ceux-ci.

Graphes des observations

Les jeux de données proviennent de cellules dans différents stades du cycle cellulaire. L’ACP nous permettra de vérifier si les cellules d’un même stade expriment les mêmes gènes/transcrits, ou de repérer s’il y a des cellules anormales, dites “outliers”.
Nous n’avons pas représenté l’ACP de la première dimension, car elle représente un artefact technique induit par le single cell RNA-seq.



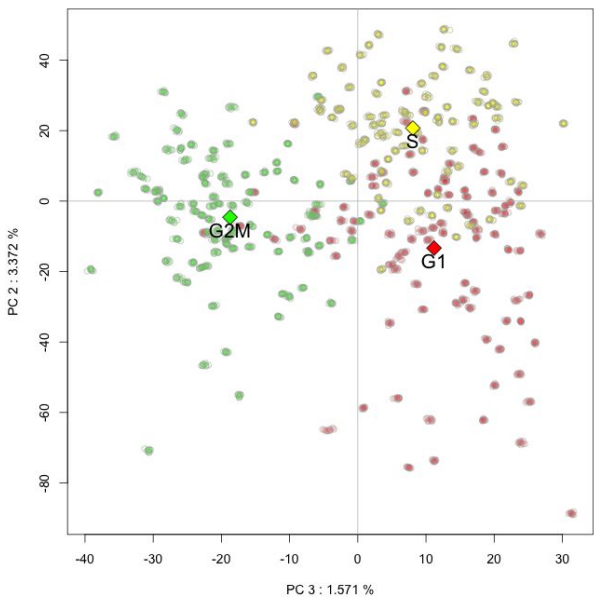
Code couleur des différents stades:

- G1 en rouge
- G2M en vert
- S en jaune

Sur ce graphe chaque cellule est représentée par un point, les observations montrent une bonne séparation des différents stades du cycle cellulaire.

ACP sur les Bootstraps

ACP de 10 bootstraps pour chaque cellule, sur l'ensemble des transcrits. On peut considérer que notre reéchantillonage est une forme de pseudo-réplicat technique.

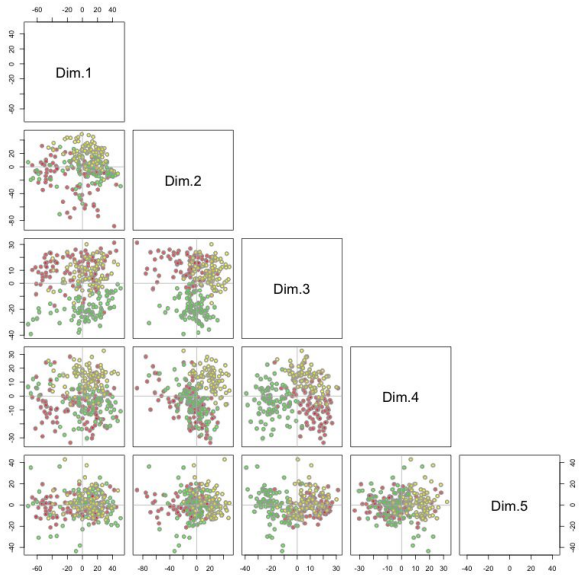


Code couleur des différents stades:

- G1 en rouge
- G2M en vert
- S en jaune

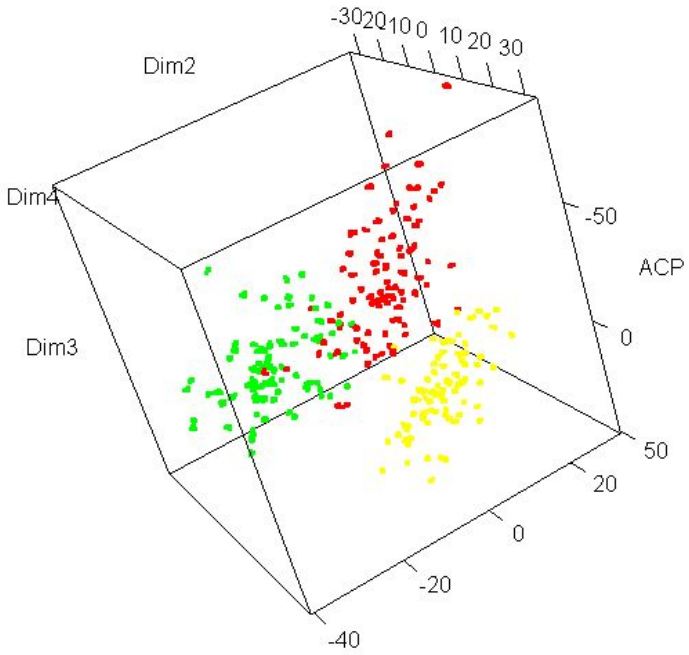
Chaque point correspond à un bootstrap des données de séquençage d'une des cellules originales. Les bootstraps d'une même cellule forment très clairement des groupes serrés sur nos graphiques. Ils n'apportent donc pas d'informations supplémentaires, mais cela nous permet d'augmenter la puissance statistique pour l'analyse différentielle.

Sur les 5 premières dimensions



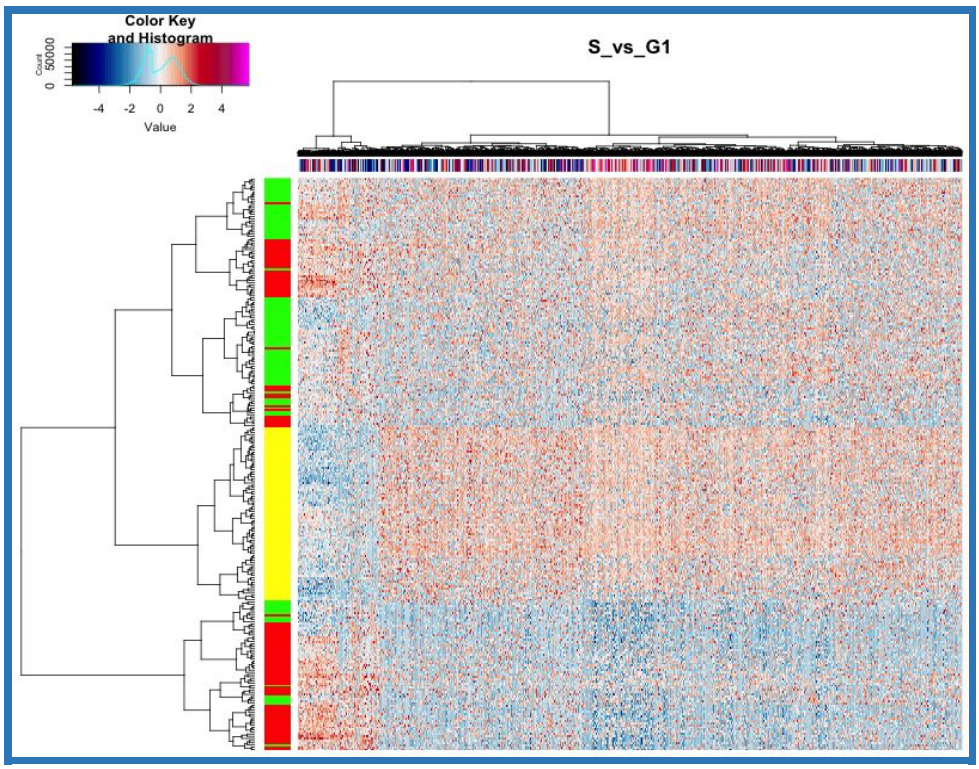
Le code couleur est le même que précédemment.

Notons qu'il est également possible de représenter les données avec un graphique interactif en trois dimensions. Nous avons choisi les dimensions 2, 3 et 4 car elles sont les plus intéressantes pour la séparation des différents stades du cycle cellulaire, comme nous pouvons le voir sur le graphique ci-dessous.



Le code couleur est le même que précédemment.

Heatmap des transcrits différentiellement exprimés



- Code couleur des différents stades (barre verticale à gauche)
  - G1 en rouge
  - G2M en vert
  - S en jaune
- Dans la Heatmap, les transcrits seront placés en colonne et les cellules en ligne
- Il y a aussi deux dendrogrammes à gauche et sur le haut de la heatmap

Le but de l'utilisation des heatmaps sera de déterminer s'il y a des groupes de transcrits qui ont des motifs d'expression semblables. Les transcrits sont représentés en colonne et les cellules en lignes. Il y a une barre en haut de la heatmap, celle-ci nous donnera une idée de la valeur d'expression des transcrits. On peut voir que les transcrits qui seront plus exprimés que la moyenne seront représentés en violet, et les transcrits moins exprimés que la moyenne seront représentés en bleu. Il y a également la clé avec le dégradé de couleurs qui nous informe sur l'expression des transcrits.

La heatmap que nous avons choisie de montrer ici concerne les transcrits différentiellement exprimés entre les stades S vs G1. Nous avions choisi ces deux stades car les différents blocs étaient particulièrement visibles. Ici, 5 blocs semblent se détacher.

Nous pouvons remarquer qu'il n'y a pas une séparation parfaite des différents stades, et qu'il y a une certaine hétérogénéité d'expression au sein des blocs. Diverses hypothèses peuvent expliquer ceci :

- Dans cette heatmap, nous nous intéressons aux transcrits différentiellement exprimés entre les stades cellulaires S et G1. La répartition des cellules du stade G2 ne sera donc pas forcément consistante.
- Si nous avions pris tous les transcrits qui ont un rapport avec le cycle cellulaire, les résultats auraient été plus consistants. Cependant, le jeu de données intégral ne pouvait être analysé en un temps acceptable sur nos machines, et le heatmap complet (avec environ 4000 transcrits) n'aurait pas été particulièrement lisible.
- De base, les données ne sont pas parfaites, et ceci pour plusieurs raisons. Premièrement, les données biologiques sont connues pour être hautement variables, et il y a de nombreux facteurs qui ne peuvent être parfaitement contrôlés qui créeront des "bruits" dans les données. Deuxièmement, ces données proviennent de séquençage sur des cellules isolées (single cell RNA-seq), contrairement aux méthodes traditionnelles de transcriptomique (puces à ADN, RNA-seq) qui s'effectuent sur des échantillons contenant des milliers voire des

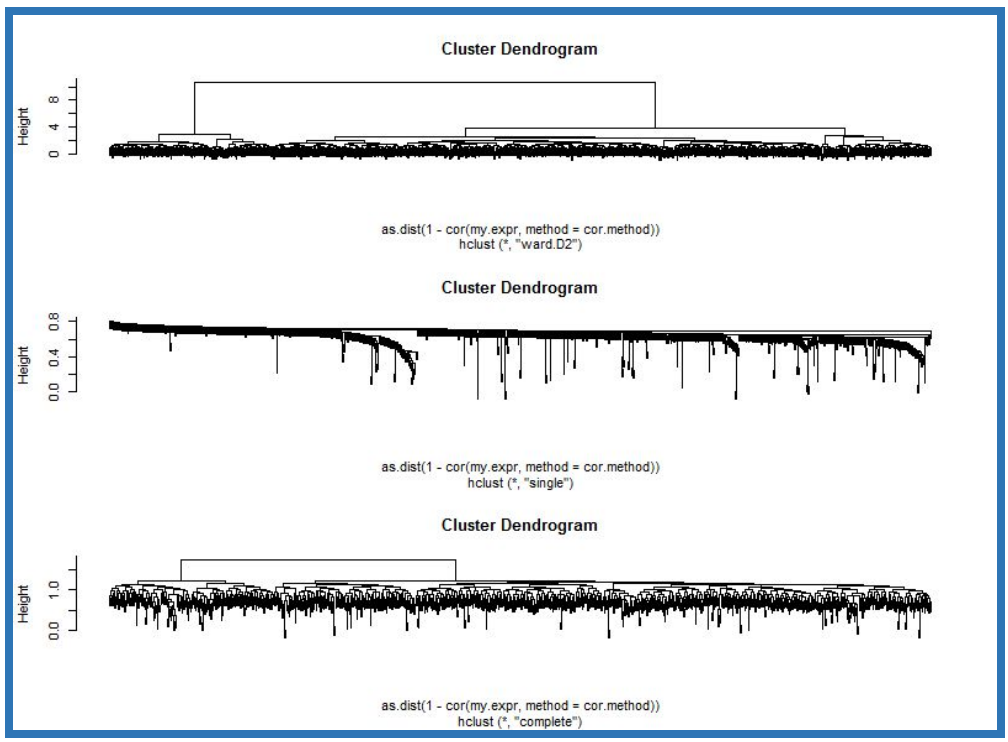
millions de cellules. Il y aura de fait plus de variabilité. Finalement, le tri (par cytométrie en flux) des cellules selon les différents stades n'est pas parfait non plus, et il est de surcroît toujours difficile de conserver les cellules dans le même stade du cycle qu'au moment du tri.

Les dendrogrammes

Le but maintenant sera d'isoler les différents transcrits qui auront des motifs d'expression similaires, et ainsi former des groupes (clusters). Nous pourront ainsi par la suite effectuer des analyse d'enrichissement sur ces clusters. La CAH va se baser sur des distances entre transcrits, en prenant soit les distances, soit la valeur absolue des distances, soit le carré des distances.

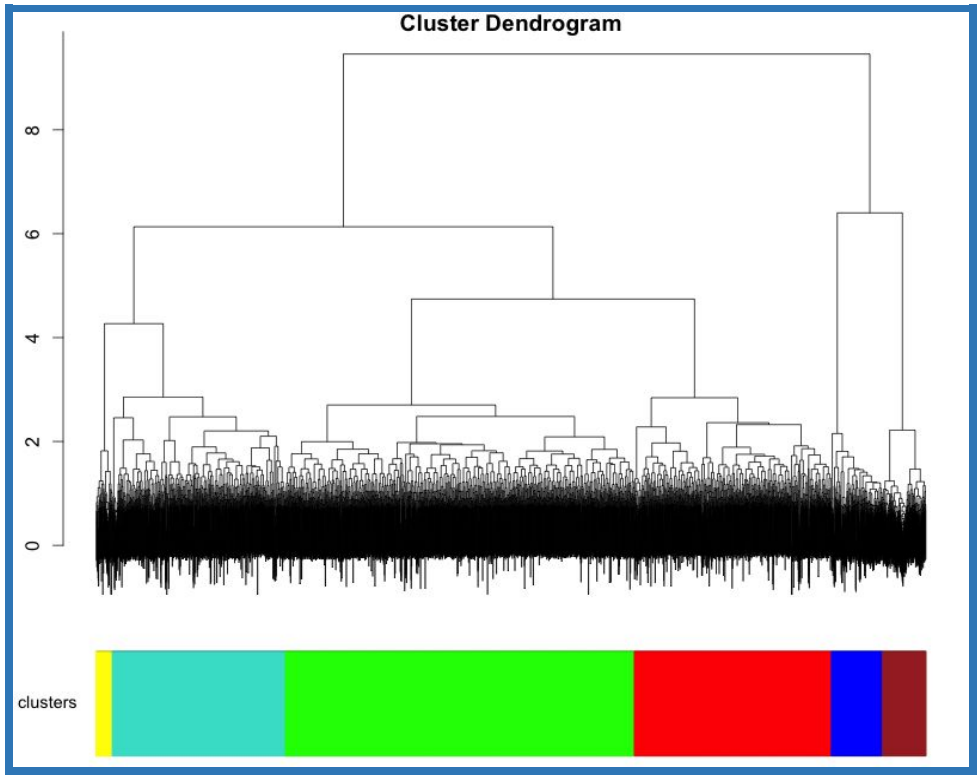
Il y a ensuite trois méthodes de création d'arbres :

- Single
- Ward.D2
- Complet



Et enfin, nous avons utilisé deux manières pour couper les arbres, soit selon une certaine hauteur h, ou selon un certain k nombre de clusters.

Voici un exemple pratique d'une coupe d'arbre avec 17'000 transcrits et en utilisant la méthode Ward.D2. Pour une séparations plus aisée des transcrits, nous avons utilisé une découpe à k = 6. Nous avons donc une séparation en six clusters que nous avons représentés avec six couleurs différentes pour une visualisation plus aisée.



Il est possible d'enregistrer les listes de gènes des différents clusters dans des fichiers Excel. L'étape suivante sera d'effectuer des analyses d'enrichissement sur les clusters.

Analyses d'enrichissement

On a réalisé une analyse d'enrichissement fonctionnelle sur les transcrits des clusters préalablement coupés. Premièrement, ils sont convertis pour être compatibles avec la base de données "Ensemble". Ensuite on choisit quelles analyses on veut faire, ici: "Biological process" (processus biologiques), "Molecular function" (fonctions moléculaires) et "Pathway" ("voies métaboliques").

On obtient comme attendu des résultats où les gènes différentiellement exprimés entre les stades cellulaires sont enrichis avec des annotations liées au cycle cellulaire. Nous avons généralement des p-values très satisfaisantes avoisinant les $x10^{-30}$ même après correction pour test multiple.

Certains de nos clusters présentaient un nombre important de lncRNA. Lors de l'analyse d'enrichissement fonctionnelle de ces derniers, nous obtenons des annotations telles que liaisons à l'ARN, liaisons à l'ADN ainsi que par ex: processus de régulation de mRNA. Nous avons donc confirmation que nos lncRNA identifiés en sont probablement.

Conclusions & perspectives

Nous avons effectué des ACP pour confirmer la ségrégation entre les stades du cycle cellulaire. Des heatmaps et des classifications ascendantes hiérarchiques pour repérer et isoler les transcrits qui ont des motifs d'expression semblables. Les clusters ont ensuite été analysés pour les caractériser. Les résultats des analyses d'enrichissement fonctionnels montre une appartenance aux fonctions du cycle cellulaire. Enfin, nous avons remarqué que les clusters avec beaucoup de long ARN non codant étaient enrichis en interactions soit avec l'ARN, soit avec l'ADN.

Nous avons prévu de continuer ce projet, avec d'autres jeux de données, de nouvelles méthodes statistiques, et des ordinateurs plus puissants pour pouvoir traiter les données en des temps raisonnables.