Ensai | LA GRANDE ÉCOLE DE LA STATISTIQUE ET DU TRAITEMENT DE L'INFORMATION

LEFORT Gaëlle

STAGE D'APPLICATION EN STATISTIQUE

Structure d'accueil : Inserm UMR 1048 – I2MC

Thème du stage : Analyse de données de puces à ADN

Tutrices de stage : Nathalie Viguerie et Nathalie Villa-Vialaneix

Lieu de stage : 1 avenue Jean Poulhès

Ville : Toulouse **Pays : France**

MODELLING DATA, CREATING KNOWLEDGE
MODÉLISER LES DONNÉES, CRÉER DU SAVOIR

Ecole Nationale de la Statistique et de l'Analyse de l'Information
Ensai

Ensai
LA GRANDE ÉCOLE DE LA STATISTIQUE
ET DU TRAITEMENT DE L'INFORMATION

Instituts thématiques — Inserm
Institut national de la santé et de la recherche médicale

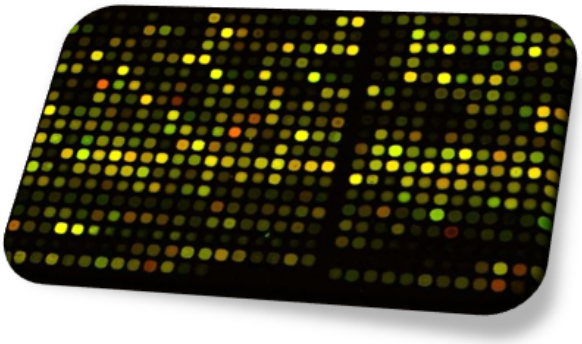
I2MC Institut des Maladies Métaboliques et Cardiovasculaires TOULOUSE

Année scolaire 2013-2014

Stage 2e année

Analyse et normalisation de données biopuces

Recherche de gènes différentiellement exprimés au cours d'un régime entre deux groupes d'individus obèses



Tutrices

Nathalie Viguerie
Nathalie Villa-Vialaneix

Élève

Gaëlle Lefort

Table des matières

Introduction

À l'échelle mondiale, le nombre de cas d'obésité a doublé depuis 1980 : en 2008, 5 % des adultes âgés de 20 ans et plus étaient en surpoids et 11 % étaient obèses [1]. Reconnue depuis 1997 comme une maladie par l'Organisation Mondiale de la Santé, l'obésité devient une problématique majeure de santé publique pour un grand nombre de pays. Cette maladie est caractérisée par un excès de tissu adipeux qui peut entraîner de graves problèmes de santé comme l'hypertension, le diabète, voire des cancers. Les causes peuvent être aussi bien environnementales que génétiques [2]. C'est pourquoi, plusieurs projets, comme le projet européen DiOGenes [1], sont mis en place pour trouver des solutions aidant les personnes en surpoids ou obèses.

À l'Institut des Maladies Métaboliques et Cardiovasulaires (I2MC), plusieurs équipes travaillent sur ce sujet. Durant un stage de deux mois, j'ai étudié la génomique du tissu adipeux, plus particulièrement l'expression des gènes d'obèses ayant suivi un régime. Dans le cadre du projet DiOGenes, plus de 600 obèses ont suivi un régime strict pendant huit semaines avant de le continuer plus librement pendant six mois. Au début et à la fin de ce protocole, elles ont subi une biopsie de tissu adipeux pour que l'on puisse mesurer l'expression de plusieurs milliers de leurs gènes (transcriptomique). Il est donc possible aux biologistes et aux statisticiens de rechercher des gènes différentiellement exprimés entre deux groupes d'individus n'ayant pas réagi de la même manière au régime pour permettre, par exemple, d'adapter les régimes à chaque individu. Dans ce contexte, j'ai analysé les expressions des gènes de plusieurs échantillons de tissu adipeux pour rechercher ceux différentiellement exprimés entre deux groupes d'obèses.

Dans ce rapport, je commencerai par décrire (chapitre 1) l'I2MC, le laboratoire dans lequel j'ai effectué mon stage. Ensuite, je décrirai (chapitre 2) les données utilisées dans cette étude ainsi que tout le travail d'apurement et de suppression des effets latents (chapitre 3). Enfin, le chapitre 4 sera consacré à la recherche de gènes différentiellement exprimés à l'aide de diverses méthodes statistiques telles que les tests paramétriques et non paramétriques de comparaison de deux groupes (ANOVA et Mann-Whitney) ou les forêts aléatoires.

1. http://www.diogenes-eu.org/

Partie 1 — L'Inserm, l'I2MC et l'équipe 4

J'ai passé plusieurs semaines comme stagiaire à l'Institut des Maladies Métaboliques et Cardiovasculaire (I2MC) dans le Laboratoire de Recherche sur les Obésités. Cet institut fait partie de l'Institut National de la Santé et de la Recherche Médicale (INSERM) et est situé sur le site hospitalier de Rangueil à Toulouse.

1.1 L'Institut National de la Santé et de la Recherche Médicale (INSERM)

L'INSERM est un établissement public français de recherche, créé en 1964. Il est entièrement dédié à la santé humaine et donc placé sous la double tutelle du ministère de la Santé et du ministère de la Recherche. Depuis avril 2009, il s'associe à l'Aviesan (Alliance nationale pour les sciences de la vie et de la santé) avec les grands acteurs de la recherche biomédicale en France (huit grands établissements publics comme le CNRS, l'INRIA ou Institut Pasteur, auxquels s'associent les universités et les CHU). Cette alliance a pour but de coordonner les équipes et les programmes de la recherche biomédicale en France.

L'INSERM a depuis lors pour mission l'étude de la santé humaine avec pour vocation d'investir le champ de la recherche biomédicale fondamentale et appliquée, dans les domaines de la biologie cellulaire, la biologie moléculaire, la génétique, la physiologie, l'épidémiologie... De plus, l'INSERM joue un rôle de première importance dans la construction de l'espace européen de la recherche et conforte sa position à l'international par d'étroites collaborations (équipes à l'étranger et laboratoires internationaux associés).

Enfin, l'INSERM est structuré en unités de recherche de tailles variées, le plus souvent insérées au sein d'UFR de médecine, d'hôpitaux, et d'universités.

1.2 L'Institut des Maladies Métaboliques et Cardiovasculaires (I2MC)

L'Institut des Maladies Métaboliques et Cardiovasculaires a été créé le 1er janvier 2011 par l'INSERM et l'Université Paul Sabatier à Toulouse. Il est situé sur le site hospitalo-universitaire de Rangueil.

Actuellement, l'Institut est constitué de 13 équipes de recherche, développées autour de trois thèmes :

- intestins, tissu adipeux, obésité et diabète ;
- thrombose, athérosclérose et vaisseaux ;
- cœur et rein.

Au total, plus de 280 personnes (chercheurs, médecins, ingénieurs, techniciens, étudiants, postdoctorants et administratifs) travaillent à l'I2MC.

1.3 Le Laboratoire de Recherche sur les Obésités (équipe 4)

Le Laboratoire de Recherche sur les Obésités fait partie de l'I2MC et s'inscrit dans l'axe de recherche sur les intestins, le tissu adipeux, l'obésité et le diabète. Il explore de nouveaux aspects du métabolisme des acides gras dans les cellules adipeuses et musculaires et étudie les relations entre voies métaboliques et résistance à l'action de l'insuline.

Les résultats des projets menés par cette équipe peuvent contribuer, entre autre, au développement d'une nouvelle classification des réponses aux régimes hypocaloriques et aux programmes d'activité physique ou aux recommandations individuelles pour les programmes de perte de poids avec une combinaison appropriée de restriction calorique et d'exercice physique.

Partie 2 — Les données de biopuces étudiées

Les données que j'ai eu à étudier durant ce stage ont été collectées grâce à la technologie biopuce (annexe A) dans le cadre du projet DiOGenes.

2.1 Le projet DiOGenes (Diet, Obesity and Genes)

Le projet DiOGenes est un projet de recherche clinique européen multicentrique dont un des buts est d'identifier et de caractériser des interactions gènes - nutriments associées aux variations pondérales, des marqueurs moléculaires de l'intervention diététique et des gènes prédicteurs des variations de poids. Il est basé sur une intervention diététique, contrôlée et randomisée, visant à analyser, entre autres, les effets de différents régimes sur le maintien du poids lors de la phase de stabilisation après un régime hypocalorique. Pour chacune des phases, différentes données ont pu être obtenues : des données dites cliniques, des données d'expression génique et des taux d'acides gras dans le tissu adipeux.

Au départ, 568 patients obèses ont initié un régime basse calorie durant huit semaines (phase de restriction). À la fin de cette première phase, les patients dont le poids a diminué d'au moins 8 % pouvaient continuer le protocole, et étaient randomisés dans une des cinq branches de régime « ad libitum » (*i.e.*, les patients pouvaient manger autant qu'ils le voulaient) pendant 6 mois (phase de stabilisation). Cette phase est une phase de suivi pondéral où les patients ont reçu une éducation diététique et suivent le régime chez eux [2].

Au début de l'étude et à la fin de chaque phase, les patients ont été pesés, mesurés, la masse grasse a été quantifiée, des prélèvements sanguins et urinaires ont été effectués ainsi que des biopsies de tissu adipeux sous cutané abdominal, des enquêtes diététiques et des questionnaires sur l'activité physique ont été réalisés. Les échantillons de tissu adipeux prélevé ont été utilisés pour quantifier l'expression des gènes à l'aide de la technologie biopuce.

2.2 Les données initiales

Les données étudiées dans la suite correspondent à la collecte d'expression de gènes au début de l'étude (CID1) et à la fin de la phase de stabilisation (CID3) pour :

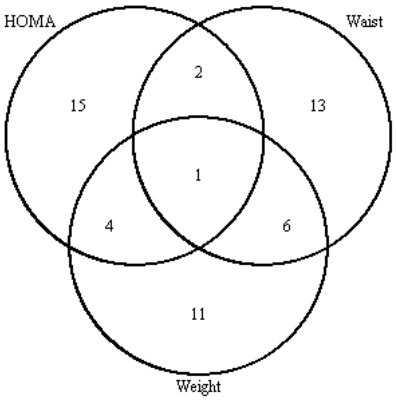
- les 11 individus de l'expérience qui ont la diminution la plus importante de poids (weight) entre CID1 et CID3 et les 11 individus qui reprennent le poids perdu durant la phase de restriction (stabilisation du poids entre CID1 et CID3) ;
- les 2×11 individus correspondant aux variations les plus importantes (stabilisation et diminution) de tour de taille (waist) ;
- les 2×11 individus correspondant aux variations les plus importantes de HOMA [1].

Certains individus sont présents pour plusieurs critères (graphique 2.1), il n'y a donc pas 66 patients mais seulement 52 patients (soit 104 biopuces, une pour CID1 et une pour CID3).

Les données sont composées de l'expression de plusieurs milliers de gènes (les variables) pour les 52 patients en CID1 et CID3. Ce qui correspond aux tailles typiques de données issues de biopuces. Elles sont fournies dans des fichiers texte où un fichier correspond à une biopuce. Dans ces fichiers, on trouve, entre autres, l'identifiant des sondes, leurs positions sur la puce, si elles sont ou non des sondes de contrôle, la mesure de l'intensité pour chaque sonde et le nom du transcrit correspondant. Chaque sonde de la biopuce correspond à un transcrit, une variante d'un gène, et non un gène. En effet, il peut exister plusieurs transcrits différents pour un seul et même gène. Dans la suite, le terme transcrit sera donc souvent utilisé. Pour les analyses statistiques, le ratio entre les intensité en vert et en rouge est utilisé pour mesurer l'expression des gènes.

Les données concernant les patients comme leur sexe ou le centre où ils ont réalisé l'étude clinique sont également disponibles. Ces données sont utilisées pour identifier ou bien interpréter d'éventuels effets latents dus

1. HOMA (HOmeostasis Model Assessment - Insulin Resistance) : mesure de la résistance à l'insuline, indicateur d'une prédisposition au diabète.



Graphique 2.1 – Diagramme de Venn des patients selon leur critère d'intégration à l'étude
Source : protocole de l'étude

au protocole expérimental. Enfin, on dispose de plusieurs variables indiquant le ou les critères d'inclusion dans l'étude (HOMA, poids et/ou tour de taille) ainsi que le groupe auquel appartient l'individu (diminution ou stabilisation du critère).

2.3 Apurement des données

Il est nécessaire de normaliser les données de biopuces pour corriger les différences systématiques entre les mesures qui peuvent être induites par des biais d'expérimentation dus à la technologie et aux différentes manipulations. L'objectif de cette normalisation est d'identifier des sources parasites de variation des expressions. Parmi ces sources, on cible plus précisément l'hétérogénéité du bruit de fond qui est corrigée par le logiciel Mapix : celui-ci quantifie l'expression des gènes et l'hétérogénéité du signal que l'on corrige grâce à des normalisations intra et inter lames. De plus, certaines sondes d'une même biopuce sont identiques, il faut donc les traiter pour supprimer les doublons avant d'effectuer d'autres analyses statistiques.

Pour manipuler ce type de données et réaliser ces traitements, on utilise le logiciel  et plus particulièrement le package `limma` de Bioconductor [2].

2.3.1 Normalisation intra-lame

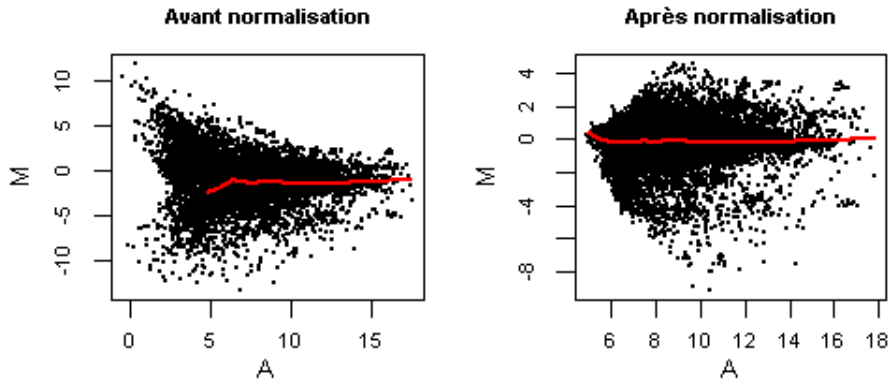
L'hétérogénéité des signaux mesurés sur une puce est souvent attribuée à des variations uniquement liées aux fluorochromes. Cette différence entre les signaux est décelable sur le nuage dont les points, représentant des sondes, ont pour abscisses les moyennes entre les logarithmes des signaux verts et rouges (dites valeurs A) et pour ordonnées les différences entre logarithmes des signaux verts et rouges (dites valeurs M). Ce graphique, dit graphe MA, est obtenu par la fonction `plotMA`.

$$M = \log \frac{R}{G} \quad et \quad A = \log \sqrt{RG} \tag{2.1}$$

La fonction `normalizeWithinArrays` corrige cette hétérogénéité en ajustant un modèle non-paramétrique par une méthode de régression locale (dite méthode loess).

La modélisation se fait par régression locale : on retient seulement les points proches du point où l'on cherche à faire une prédiction. La distance est prise en compte avec une fonction de poids cubique, comme cela avait initialement proposé par William Cleveland (fonction Lowess, LOcally WEighted Scatterplot Smoothing). Plus que la forme des poids, c'est la fenêtre définissant la distance du voisinage qui est importante. Autrement dit, plus la fenêtre est grande, plus on prend en compte de voisins, et donc plus la fonction sera lisse (mais biaisée). En revanche, avec une petite fenêtre, on suit les variations au plus près, la courbe est moins régulière (plus de variance), mais localement on peut penser avoir un biais plus faible. La fenêtre utilisée ici sera la fenêtre par défaut c'est-à-dire 0,3.
Dans notre étude, quelle que soit la biopuce, la normalisation intra-lame est faite sur un nuage de points ayant la même forme ce qui entraîne une même forme de nuage normalisé (graphique 2.2).

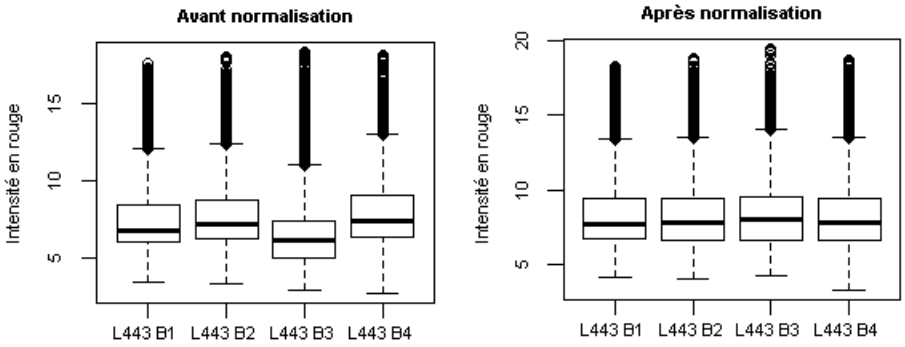
2. Projet *open source* qui fournit des outils (packages R) pour l'analyse et la compréhension de données génomiques



Graphique 2.2 – Effet d'une normalisation intra-lames (courbe de régression loess)
Source : biopuce 1 de la lame 443

2.3.2 Normalisation inter-lames

Après normalisation intra-lame des valeurs M, les distributions des intensités rouges et vertes devraient être essentiellement les mêmes pour toutes les puces (c'est une hypothèse biologique). Cependant, il peut être observé une grande variabilité. La fonction `normalizeBetweenArrays` permet de normaliser les distributions d'intensité en alignant les quantiles de A entre les différentes puces (graphique 2.3).



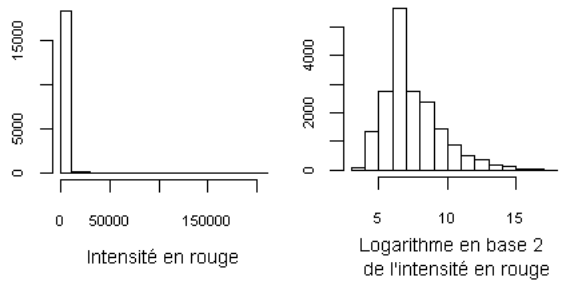
Graphique 2.3 – Effet d'une normalisation inter-lames pour l'intensité en rouge
Source : lame 443

2.3.3 Gestion des doublons et des valeurs manquantes

Lors de la lecture des données, il est possible d'appliquer un filtre aux sondes pour les pondérer par 0 ou 1. Ce filtre permet de pondérer par 0 les sondes qui sont des sondes de contrôle et celles dont l'intensité en vert ou rouge est insuffisante. Les sondes de contrôle sont supprimées et celles pour lesquelles l'intensité est insuffisante sont considérées comme des valeurs manquantes. Malgré cela, les données ne sont pas encore nettoyées : certaines sondes sont présentes plusieurs fois et peuvent représenter le même transcrit. Il est aussi possible pour certains transcrits d'avoir beaucoup de valeurs manquantes.

On ne travaille plus ici sur les données brutes de biopuces mais sur un tableau dont les lignes sont les transcrits, les colonnes les biopuces et les valeurs les logarithmes (en base 2) du ratio des intensités de rouge et de vert. Il n'est pas judicieux de travailler sur les données brutes car celles-ci ont une distribution très asymétrique. On utilise, généralement, les valeurs en logarithme en base 2 car la plupart des intensité en vert et en rouge sont très faibles. Une transformation en logarithme en base 2 va permettre d'obtenir une distribution plus proche d'une distribution normale, en la recentrant et en la rendant symétrique (graphique 2.4).

Doublons Les sondes identiques présentes plusieurs fois sur une même biopuce sont moyennées pour n'apparaître qu'une seule fois. On utilise comme identifiant celui des transcrits qui commence généralement par « NM_ ». Celui-ci est différent du nom des gènes aussi appelé symbole.



Graphique 2.4 – Effet d'une tranformation logarithmique en base 2 sur l'intensité rouge d'une biopuce
Source : biopuce 1 de la lame 443

Valeurs manquantes Seulement 45 % des transcrits ont moins de 20 % de valeurs manquantes. Pour beaucoup on dépasse les 80 % de valeurs manquantes. On ne va garder pour la suite que les transcrits ayant moins de 20 % de valeurs manquantes. Ce qui en fait environ 12 000 sur les 45 000 sondes présentes sur chaque biopuce.

Des données prêtes à être analysées

Il est maintenant possible d'utiliser les données pour différentes analyses statistiques.
La normalisation est importante et similaire quelles que soit les données de biopuces traitées. Pour permettre à des biologistes de refaire ces normalisations et les autres traitements ultérieurement, j'ai réalisé un script R détaillé expliquant les différents traitements sur les données de cette étude (annexe B).
Cette étape de nettoyage des données a été une partie importante du début de mon stage. Le but ayant été de réaliser un script permettant de reproduire ces traitements sur n'importe quel jeu de données mais aussi de réfléchir aux traitements à réaliser, par exemple, quelle méthode utiliser pour la normalisation intra-lame ou comment regrouper les transcrits (par l'identifiant de la sonde, l'identifiant du transcrit ou le nom du gène).

Partie 3 — Identification et suppression des effets latents

Avant de rechercher les gènes différentiellement exprimés entre les deux groupes (stabilisation ou diminution du critère d'inclusion dans l'étude), il faut regarder si d'autres facteurs expérimentaux influent. Pour cela, on peut réaliser une Analyse en Composantes Principales où les biopuces sont les individus et les transcrits les variables. Ensuite, il ne reste plus qu'à colorer les points du graphique des individus en fonction d'un critère comme le sexe de l'individu ou son centre d'appartenance, par exemple. Si on observe des points atypiques ou des groupes suivant les couleurs alors ce facteur sera considéré comme ayant un effet.

3.1 Des effets liées au centre d'appartenance et au critère de sélection des individus

Plusieurs ACP ont été réalisées : une générale sur toutes les biopuces et trois autres sur des sous-groupes correspondant aux différents types de variations (poids, tour de taille et HOMA). L'ACP sur toutes les biopuces est celle qui apporte le plus d'informations dans le sens où les résultats obtenus dans cette ACP sont retrouvés partiellement dans les ACP sur les sous échantillons. C'est celle-ci que l'on va détailler dans la suite.

3.1.1 Qu'est-ce que l'Analyse en Composantes Principales ?

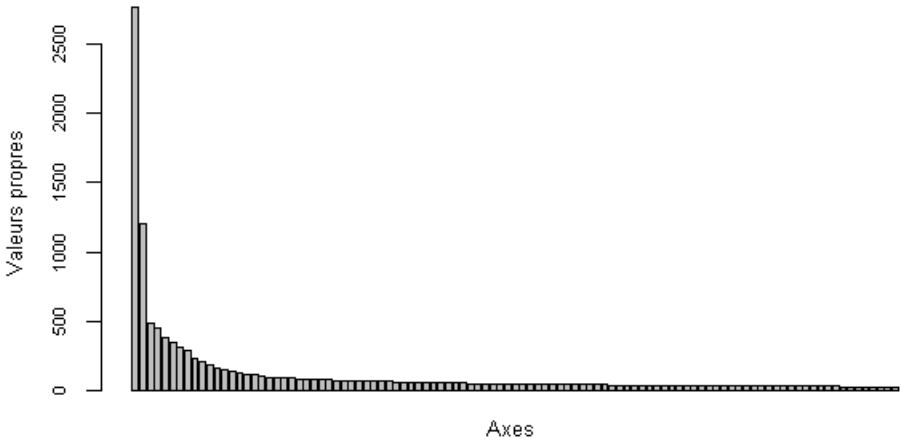
Objectif L'ACP a pour but de synthétiser de manière la plus pertinente possible les données initiales pour pouvoir ensuite les représenter graphiquement sur un nombre réduit de dimensions. En particulier, des représentations sur des sous-espaces pertinents de dimension 2 sont possibles.

Principe C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, car on analyse essentiellement la dispersion des données considérées. On utilise la matrice des corrélations à la place de celle des variances-covariances lorsque l'on effectue une ACP normée (*i.e.*, les variables ont été centrées et réduites au préalable). Cela permet d'homogénéiser les variables pour ne pas avoir de problèmes liés à des unités de mesure différentes. C'est la méthode qui sera utilisée pour cette étude.
D'un point de vue mathématique, l'ACP est un changement de base. On passe d'une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations. Cela va permettre de réaliser les graphiques désirés dans un espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (c'est à dire en choisissant les facteurs de manière à maximiser la dispersion du nuage de points projeté). C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées.

Mise en œuvre sur R Elle est très simple à mettre en œuvre sur grâce à la fonction `PCA` du package `FactoMineR`. Cependant, avant de réaliser une ACP, on doit imputer les valeurs manquantes. Pour ce faire, j'ai utilisé la méthode des k plus proches voisins (ici, k a été fixé à 3) grâce à la fonction `impute.knn` du package `impute` de Bioconductor. La méthode des k plus proches voisins va rechercher, pour un individu contenant des valeurs manquantes, les k invididus les plus proches du jeu de données sur les variables non manquantes. Les valeurs manquantes sont alors imputées de la valeur moyenne de la variable considérée pour les k voisins les plus proches.

3.1.2 Quatre dimensions nécessaires pour résumer l'information

Avant d'interpréter les résultats de l'ACP donné par le logiciel, il faut choisir le nombre de facteurs à retenir. Il existe plusieurs méthodes que l'on peut combiner pour trouver ce nombre. La plus couramment répandue est celle de la recherche d'un coude sur l'éboulis des valeurs propres (graphique 3.1).
Un coude est visible au niveau de la deuxième et le suivant, moins marqué, au niveau de la quatrième. Une formalisation de ce critère est le Scree-test de Catell : on calcule les différences secondes des valeurs propres



Graphique 3.1 – Éboulis des valeurs de l'ACP réalisée sur toutes les biopuces
Source : données nettoyées de toutes les biopuces

jusqu'à ce que le signe change, ce qui se produit ici entre la deuxième et la troisième valeur propre. Il est aussi possible de sélectionner les axes jusqu'à obtenir un pourcentage de l'inertie totale fixé a priori, généralement 80 %. Avec ce critère, on devrait interpréter 44 axes ce qui beaucoup trop. Seulement quatre axes seront donc retenus pour essayer d'identifier des effets.

3.1.3 Identification des effets

L'ACP réalisée permet de représenter les données dans un espace de faible dimension pour en observer la structure et détecter des effets expérimentaux non souhaités (points atypiques, groupe d'individus et individus aux mêmes caractéristiques répartis au même endroit du plan). Grâce à l'ajout de différentes couleurs selon les modalités de diverses caractéristiques des individus, plusieurs effets ont pu être identifiés (graphique 3.2).

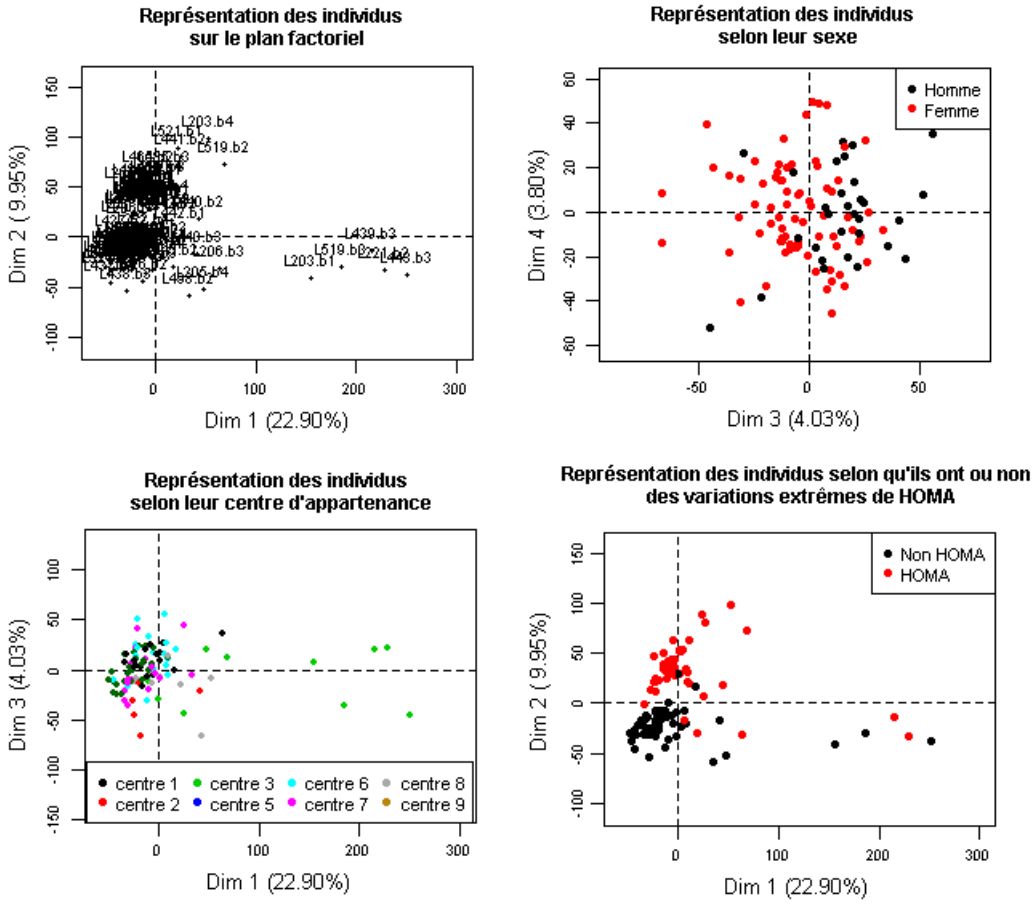
Sur la première dimension, on observe cinq individus atypiques qui appartiennent tous au centre 3. Après discussion avec les biologistes, nous avons décidé de supprimer tous les individus de ce centre, soit cinq individus, car cela ne déséquilibrait pas les trois critères HOMA, poids et tour de taille, ni les groupes « forte perte »ou « stabilisation ».

Sur la seconde dimension, les individus sont regroupés en deux groupes bien séparés. Cette séparation peut être expliquée par les trois critères d'inclusion des individus (HOMA, poids et tour de taille). En effet, quasiment tous les individus sélectionnés pour des fortes variations de HOMA sont du coté positif de l'axe au contraire de ceux sélectionnés pour des fortes variations de poids et/ou de tour de taille. Cet effet n'est pas problématique pour la suite de l'étude car toutes les analyses seront menées par critères d'inclusion.

Enfin, on observe un effet du sexe de l'individu sur la dimension 3 et un effet du centre d'appartenance sur le plan 1-3, que l'on va corriger après avoir refait la normalisation inter-lames sans les individus du centre 3.

3.2 Suppression des effets trouvés

Pour supprimer les effets Sexe et Centre, des normalisations par la médianes et par les quantiles vont être utilisées. L'effet du sexe de l'individu étant très léger, un alignement des médianes dans les deux groupes homme et femme est suffisant. On commence par calculer les médianes pour chaque sexe sur tous les gènes puis on la soustrait à l'expression des gènes des individus correspondants. L'alignement des quantiles pour supprimer l'effet Centre ajuste, quant à lui, tous les quantiles des divers groupes. Elle peut être réalisée avec le package `preprocessCore` et sa fonction `normalize.quantiles.in.blocks` [3].



Graphique 3.2 – Représentation des individus dans plusieurs plan selon plusieurs caractéristiques
Source : données nettoyées de toutes les biopuces

Des effets latents identifiés et supprimés

Après avoir supprimé les individus du centre 3 qui étaient atypiques au niveau des gènes, il a été possible de supprimer les effets Sexe et Centre identifiés.
Cependant, cela ne veut pas dire que tous les effets latents existants ont pu être identifiés. Pour approfondir cette partie, on utilisera, par la suite, la méthode FAMT (Factor Analysis for Multiple Testing) qui permet de supprimer des effets latents non observés pour ensuite rechercher des transcrits différentiellement exprimés entre deux groupes. Mais avant cela, on va réaliser des tests ANOVA (ANalysis Of VAriance) avec correction de tests multiples pour tenter d'identifier directement ces transcrits sur les données corrigées issues du travail décrit dans cette partie.

Partie 4 — Recherche de gènes différentiellement exprimés

On dit qu'un gène est *différentiellement exprimé* entre deux conditions lorsque son niveau d'expression moyen est significativement différent (au sens statistique du terme) entre les deux conditions. Pour rechercher les gènes différentiellement exprimés entre les deux groupes (stabilisation ou diminution du critère HOMA, poids ou tour de taille), on va utiliser plusieurs méthodes : des tests suivis de correction de tests multiples, la méthode FAMT (Factor Analysis for Multiple Testing). Alternativement, les gènes ayant un fort pouvoir prédicteur du groupe seront recherchés avec la méthode des fôrets aléatoires.
Dans ce chapitre, les données utilisées seront celles issues des sous-populations selon le critère d'inclusion dans l'étude (HOMA, poids ou tour de taille) et le temps de la mesure. Pour chaque critère, on va étudier séparément les données à CID1, celles à CID3 et les taux d'évolution entre CID1 et CID3. On va donc travailler sur 9 jeux de données différents en parallèle. Puis, pour chaque jeu de données, on mettra en commun les résultats obtenus par les différentes méthodes.

Dans la suite de ce rapport, seuls les résultats obtenus sur les individus sélectionnés pour leur variation de poids à CID3 seront présentés.

4.1 Des tests pour trouver des gènes différentiellement exprimés entre les groupes

Dans un premier temps, on cherche les transcrits dont l'expression à CID3 est significativement différente selon que l'individu diminue ou stabilise son poids. Plusieurs milliers d'ANOVA et de tests de Mann-Whitney sur tous les transcrits sont réalisés sur les données corrigées précédemment. Lorsque plusieurs tests sont réalisés en parallèle, il est nécessaire de corriger les p-valeurs obtenues pour éviter un nombre trop important de faux positifs (c'est-à-dire de gènes déclarés différentiellement exprimés alors qu'ils ne le sont pas).

4.1.1 ANOVA (ANalysis Of VAriance), tests de Mann-Whitney et correction des tests multiples

Objectif L'analyse de la variance ainsi que le test de Mann-Whitney permettent de mettre en évidence si une variable numérique a des valeurs significativement différentes selon plusieurs catégories.

ANOVA L'ANOVA est une généralisation d'un test paramétrique de Student d'égalité des moyennes lorsqu'il y a plus de deux catégories (annexe C.1) :

$$H_0 : \text{les moyennes sont égales}$$
$$\text{contre}$$
$$H_1 : \text{au moins l'une des moyennes est différente}$$

L'inconvénient de ce test de Fischer est qu'il repose sur deux hypothèses pour être valide. Il faut, d'une part, que la variable numérique suive une loi normale ce qui peut-être testé à l'aide d'un test de Shapiro-Wilk ou de Kolmogorov-Smirnov (annexes C.5 et C.6). Le test de Shapiro-Wilk peut être considéré comme plus puissant sur des échantillons plus petits [4]. Cependant les p-valeurs des deux tests étant très différentes pour les données étudiées, les deux sont ajoutées aux résultats produits. D'autre part, il faut que les variances des deux groupes soit égales. Le test de Bartlett est utilisé pour tester cette hypothèse (annexe C.4). Lorsque celle-ci n'est pas vérifiée, on peut pratiquer un test de Welch plutôt qu'une ANOVA classique (annexe C.2).

Test de Mann-Whitney Contrairement à l'ANOVA, ce test est non paramétrique et il ne repose sur aucune hypothèse (annexe C.3). Il permet de comparer les distributions de deux échantillons de petites tailles en utilisant l'ordre dans lequel apparaissent les observations des deux échantillons réunis :

H_0 : les échantillons sont identiquement positionnés (*i.e.* ils appartiennent à la même population)
contre
H_1 : les lois sous jacentes aux observations ne sont pas les mêmes

Correction de tests multiples Lorsque plusieurs tests statistiques sont réalisés simultanément, le risque global d'erreur de première espèce s'accroît. La répétition à chaque test du risque d'obtenir un résultat significatif augmente le risque global de conclure à tort à une différence significative entre les groupes. Le risque global de conclure à tort à l'efficacité à l'issu de cet essai n'est plus de 5 % (même si c'est le seuil retenu pour chaque test) mais il est bien supérieur. Pour corriger cela, il est possible d'utiliser plusieurs méthodes comme celle de Bonferroni ou de Benjamini & Hochberg. Dans la suite, la méthode de Benjamini & Hochberg sera utilisée [5]. Cette méthode contrôle le taux de faux positifs, ce qui est plus puissant que le contrôle de l'erreur de première espèce réalisé par la correction de Bonferroni.
Elle est réalisée en ordonnant les p-valeurs de chaque gène de la plus faible à la plus élevée. Puis, pour chaque p-valeur, la p-valeur corrigée est égale à la p-valeur initiale multipliée par le nombre total de gènes et divisée par son rang.

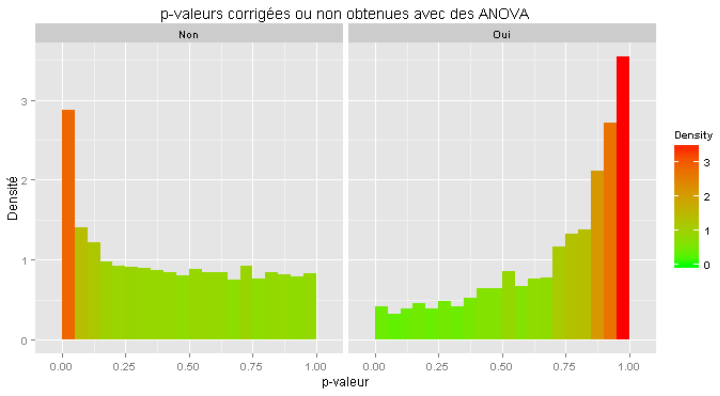
Mise en œuvre sur R La fonction `oneway.test` permet de réaliser une ANOVA et d'en récupérer la p-valeur. L'avantage de cette fonction par rapport aux autres implémentées sous R est qu'elle permet de gérer le cas où les variances ne sont pas égales en réalisant un test de Welch. Pour tester l'égalité des variances, on peut utiliser la fonction `bartlett.test`. Au contraire, la non normalité des données ne peut pas être gérée directement par ce test et donc la p-valeur associée au test ANOVA sera toujours accompagnée de celle associée à un test de normalité (fonctions `shapiro.test` et/ou `ks.test`). De plus, un test de Mann-Whitney est réalisé sur tous les transcrits grâce à la fonction `wilcox.test`. Enfin, pour corriger les p-valeurs, il suffit de les passer en paramètre de la fonction `p.adjust` et de spécifier la méthode de correction utilisée (`"BH"`).

4.1.2 Résultats des deux tests

A CID1, aucun des deux tests ne permet de trouver des gènes significativement différent entre les deux groupes. Ce n'est pas le cas à CID3.

Plus de 250 gènes différentiellement exprimés selon l'ANOVA

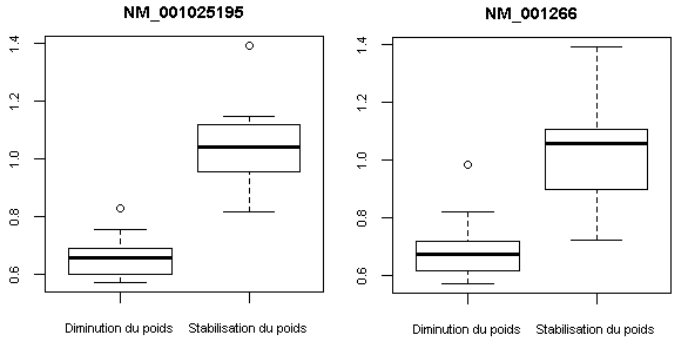
Les tests ANOVA réalisés sur tous les gènes donnent 256 p-valeurs corrigées inférieures au seuil de 5 % (graphique 4.1). Pour ces transcrits, l'expression des gènes des patients qui diminuent leur poids est donc significativement différente de ceux des patients qui le stabilisent entre CID1 et CID3.



Graphique 4.1 – Effet de la correction de la multiplicité des ANOVA sur les p-valeurs
Source : données corrigées des individus à fortes variations de poids à CID3

La manière la plus courante de représenter les expressions des gènes pour un ou plusieurs groupe est l'utilisation de boîtes à moustache. Elles sont généralement réalisées pour les p-valeurs les plus faibles : cela permet de visualiser la différence entre les niveaux d'expression en relation avec les variations d'expression des gènes extraits de l'ANOVA. Pour le transcrit NM_001025195, qui a l'une des plus faibles p-valeurs corrigées (0,002), ce graphique est fourni dans la figure 4.1 : on peut voir que pour ce transcrit, les deux distributions des expressions

ne se chevauchent pas pour les deux groupes. Il est l'un des transcrits du gène CES1. L'autre transcrit de ce gène (NM_001266) a une p-valeur de 0,018. Pour ces deux transcrits, les tests de Kolmogorov-Smirnov rejettent la normalité alors que ceux de Shapiro-Wilk l'acceptent avec un p-valeur environ égale à 0,15. Il faut donc être prudent lorsque l'on interprète les résultats de ces ANOVA.



Graphique 4.2 – Boîtes à moustache de l'expression des transcrits du gène CES1 selon le type de variations du poids
Source : données corrigées des individus à fortes variations de poids à CID3

Près de 150 gènes différentiellement exprimés selon le test de Mann-Whitney

Les tests de Mann-Whitney suivis de la correction des tests multiples permettent d'extraire 142 transcrits différentiellement exprimés entre les deux groupes d'individus extrêmes pour leurs variations de poids au seuil 5 %. Certains transcrits sont les mêmes que pour les ANOVA (section 4.4) comme les transcrits du gène CES1 qui ont des p-valeurs égales à 0,02 et 0,03.

4.2 FAMT (Factor Analysis for Multiple Testing)

Dans le chapitre précédent (chapitre 3), l'objectif était d'identifier et de supprimer des effets latents. Certains effets comme celui du centre ou du sexe ont pu être corrigés. Cependant, avec des ACP, les effets sont identifiés de manière visuelle, ce qui est subjectif et les effets identifiés ne le sont que selon des variables expérimentales déjà connues (observées). De plus, les effets trouvés sont corrigés par des méthodes de normalisations générales et non pas spécifiques aux données. La méthode FAMT permet de résoudre ces problèmes en recherchant de manière automatisée des effets latents, possiblement non observés. Cette méthode inclut une procédure de test pour rechercher les gènes différentiellement exprimés sur les données corrigées.

4.2.1 Description de la méthode

Objectif La méthode d'analyse en facteurs pour les tests multiples a pour but de sélectionner les gènes différentiellement exprimés entre deux groupes d'individus lorsque la structure de corrélation entre les expressions des gènes est forte [6].

Principe La méthode se décompose en trois étapes. Dans un premier temps, on estime le nombre de facteurs latents. Pour ce faire, on modélise la structure de covariance puis on cherche le nombre de facteurs k qui minimise l'inflation de variance du nombre de faux-positifs. Ensuite, on ajuste un modèle de régression (pour l'expression des gènes $m_k(x)$) combiné à un modèle d'analyse en facteurs (pour les facteurs latents Zb'_k et les résidus ϵ_k) grâce à l'algorithme EMFA (Expectation-Maximisation pour l'Analyse en Facteurs) :

$$Y_k = m_k(x) + Zb'_k + \epsilon_k$$

Enfin, des tests de Fischer sont réalisées sur les données ajustées sous les mêmes conditions que pour les ANOVA (section 4.1.1). Nous terminons ensuite l'analyse par une correction des tests multiples comme décrite dans la section précédente.

Mise en œuvre sur R Le package `FAMT` permet de réaliser les différentes étapes décrites ci-dessus [7]. Tout d'abord, il est possible d'obtenir un jeu de données utilisable grâce à la fonction `as.FAMTdata`. Cette fonction va permettre d'obtenir un jeu de données sans valeur manquante en imputant par la méthode des plus proches voisins. Ensuite, la fonction `modelFAMT` va estimer le nombre de facteurs optimal (par un critère BIC), ajuster le modèle et calculer les p-valeurs associées à chaque test de Fischer. Il est aussi possible d'interpréter les facteurs trouvés avec la fonction `defacto`. Celle-ci calcule les p-valeurs de tests de corrélation entre les facteurs et les différentes variables du plan d'expérience (sexe, centre...).

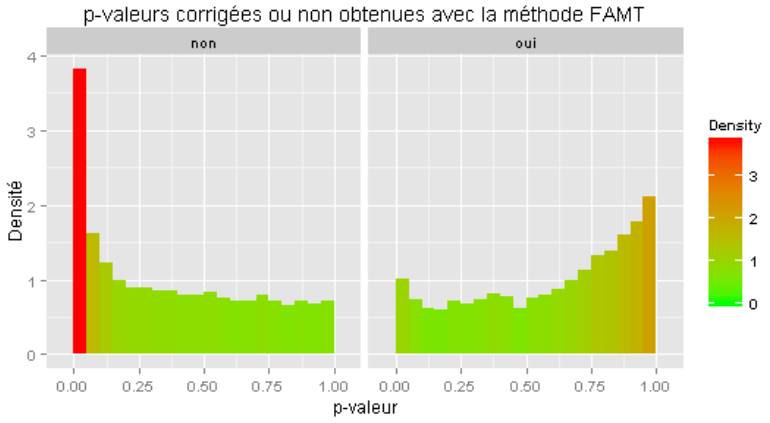
4.2.2 Deux facteurs latents identifiés et près de 700 transcrits différentiellement exprimés

La méthode FAMT fait ressortir deux facteurs pour les données de CID3 des individus à forte variation de poids. Ces deux facteurs sont corrélés avec le centre d'appartenance, le sexe de l'individu ainsi que la combinaison des critères d'inclusion dans l'étude (tableau 4.1). Ce sont les mêmes effets qui peuvent être observés avec des ACP. Le type de régime effectué pendant la phase de stabilisation n'a pas d'effet.

	Centre d'appartenance	Sexe	Type de régime	Combinaison des critères
Facteur 1	0,03	0,39	0,79	0,0005
Facteur 2	0,12	0,05	0,48	0,0002

Tableau 4.1 – p-valeurs des tests de corrélation entre les facteurs latents obtenus et les variables du plan d'expérience
Source : données nettoyées et imputées des individus à fortes variations de poids à CID3

Les ajustements des données réalisés par cette méthode permettent d'identifier 697 p-valeurs corrigées inférieures au seuil de 5 % (graphique 4.3), ce qui est très supérieur au nombre de transcrits différentiellement exprimés trouvés par les ANOVA.



Graphique 4.3 – Effet de la correction de la multiplicité des tests de Fischer réalisés sur les données ajustées par la méthode FAMT
Source : données nettoyées et imputées des individus à fortes variations de poids à CID3

4.3 Recherche de gènes fortement prédictifs

Dans cette section, nous utilisons une approche très différente des approches par test pour identifier des gènes importants pour expliquer les différences entre les deux groupes : nous utilisons une méthode de classification supervisée (ici, les forêts aléatoires) et recherchons des gènes dont le pouvoir prédictif est fort dans le modèle ajusté. Les forêts aléatoires sont généralement utilisées pour faire de la prédiction mais il est aussi possible d'en tirer un critère d'importance pour les variables explicatives.

4.3.1 Principe de la méthode

Objectif L'objectif des forêts aléatoires est de prédire une variable (régression ou classification supervisée), à partir de variables explicatives. Dans cette étude, on ne cherche pas réellement à prédire si, selon son profil d'expression génique, un individu perd du poids ou au contraire en gagne mais les transcrits qui sont important pour obtenir un taux d'erreur faible sont d'intérêt pour le biologiste. On va donc s'intéresser, plus particulièrement, à l'importance des variables.

Construction d'une forêt La construction d'une forêt aléatoire consiste à élaborer plusieurs centaines d'arbres de décisions puis à les aggréger [8]. Chaque arbre est réalisé à partir d'un échantillon bootstrap (avec remise) des patients et chaque scission de nœud est effectué à partir d'un sous-ensemble des transcrits initiaux, tirés aléatoirement. Il y a donc une double randomisation qui a pour effet de diminuer clairement la corrélation entre les arbres de la forêt et donc la variance du modèle agrégé. Le modèle final de prédiction utilise un principe de vote majoritaire : la forêt prédit la classe qui lui a été majoritairement attribuée par l'ensemble des arbres.

Critères d'analyse des résultats Classiquement, un certain nombre de critères sont associés à l'analyse des forêts : ils permettent d'une part d'estimer la qualité de prédiction de la forêt et d'autre part, de récupérer les variables explicatives les plus importantes du modèle.

- Taux d'erreur calculé sur les individus out-of-bag (OOB) : c'est le taux d'erreur calculé pour chaque arbre sur les individus dits « out-of-bag ». Ces individus sont ceux qui n'appartiennent pas à l'échantillon bootstrap ayant permis la construction de l'arbre. Pour le calculer sur la forêt, le vote de l'individu n'est pris en compte que sur les arbres qui ont été construits sans celui-ci.
- Importance des variables mesurée par la « mean decrease Gini » : elle représente la baisse de pureté des nœuds obtenue après randomisation des valeurs de la variable d'intérêt dans l'échantillon initial (le but est de voir si le fait de fournir des valeurs éronnées pour cette variable dégrade la qualité de la forêt). La pureté d'un nœud est mesurée par l'indice de Gini.

Choix des paramètres pour la construction de la forêt Les forêts aléatoires dépendent d'un certain nombre de paramètres qui sont à calibrer pour obtenir des résultats pertinents. Ces paramètres sont :

- nombre d'arbres dans la forêt : ce paramètre est choisi de manière à avoir un taux d'erreur OOB qui se stabilise. Nous avons choisi 2000 arbres ;
- nombre de variables pour la scission de chaque nœud : pour des arbres de classification, le choix de la partie entière de la racine carré du nombre total de prédicteurs est le paramètre par défaut que nous avons utilisé (soit ici 110 variables) ;
- effectif minimum de chaque feuille : ce paramètre correspond au nombre minimum d'individu inclu dans une feuille final d'un arbre. La valeur par défaut de 1 a été conservée.

Mise en œuvre sur R Pour réaliser des forêts aléatoires, il est possible d'utiliser le package `randomForest`. Avec la fonction du même nom, on obtient les taux d'erreur et les valeurs des critères d'importance des variables.

VSURF : un package R pour faire de la sélection de variables avec les forêts aléatoires En plus du package précédemment cité, il existe un package spécifique pour faire de la sélection de variables nommé VSURF (Variable Selection Using Random Forest) [9]. Il fournit deux sous-ensembles de variables associés à deux objectifs de sélection de variables. Le premier est un sous-ensemble de variables importantes pour l'interprétation. Le second est un sous-ensemble parcimonieux à l'aide duquel on peut faire de bonnes prédictions. La stratégie générale est basée sur un classement préliminaire des variables donné par l'indice d'importance des forêts aléatoires, puis utilise un algorithme d'introductions ascendantes de variables pas à pas. Une fonction du nom du package permet d'ajuster le modèle et de calculer les critères.

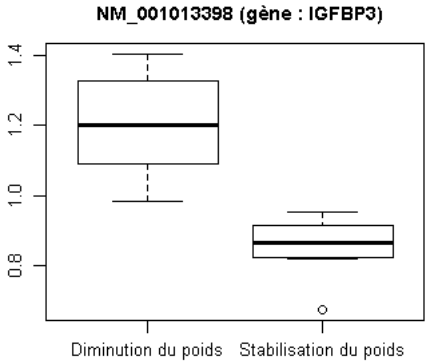
4.3.2 Une centaine de transcrits fortement prédicteurs

Comme son nom l'indique, la forêt aléatoire est aléatoire : lorsque plusieurs ajustements du modèle sont effectués, les résultats diffèrent. Pour les individus à forte variation de poids, je n'ai donc pas réalisé une seule forêt mais

cent différentes. Ensuite, pour trouver les transcrits fortement prédicteurs, j'ai sélectionné les 500 premiers de chaque forêt selon le critère « mean decrease Gini »puis je n'ai gardé que ceux présents dans plus de 80 % des forêts. Pour les autres jeu de données, les cent premiers transcrits seulement ont été sélectionnés car les résultats des autres méthodes ne donnent qu'une vingtaine de gènes différentiellement exprimés au maximum.

Avec cette approche, à CID3 et pour les individus à forte variation de poids, 169 transcrits sont fortement prédicteurs du groupe de l'indivdus. De plus, une forêt aléatoire basée sur ces seul transcrit a un taux d'erreur Out-Of-Bag de 0 % donc les individus sont toujours bien classés, ce qui n'est pas le cas sur les forêts avec tous les transcrits où le taux d'erreur OBB est d'environ 20 % et montre donc le bon caractère prédicteur des transcrits sélectionnés.

Avec la méthode VSURF, un seul transcrit ressort comme fortement prédicteur : le transcrit NM_001013398 du gène IGFBP3 (graphique 4.4). C'est le seul transcrit que fait ressortir cette méthode car le taux d'erreur de mauvais classement associé à un arbre ne contenant que cette variable est de 0 %. En conséquence, ce seul transcrit permet de classer les individus diminuant leur poids de ceux le stabilisant sans erreur. Qui plus est, ce transcrit est intéressant car c'est le seul du gène IGFBP3.



Graphique 4.4 – Boîtes à moustache de l'expression du transcrit du gène IGFBP3 selon le type de variations
Source : données corrigées des individus à fortes variations de poids à CID3

4.4 Un seul gène commun à toutes les méthodes

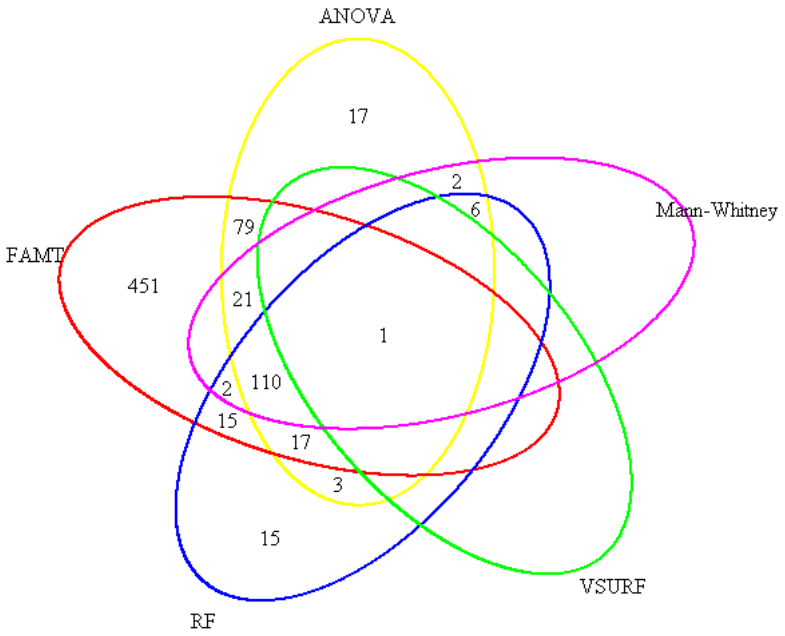
À CID1, certains gènes s'exprimeNT légèrement différemment entre les groupes à forte perte et à stabilisation de poids. Pour ces gènes, la différence n'est pas significative car ils sont mis en avant par les forêts aléatoires et non par les différents tests réalisés, au contraire de CID3 où la différence est beaucoup plus marquée.

Le gène IGFBP3

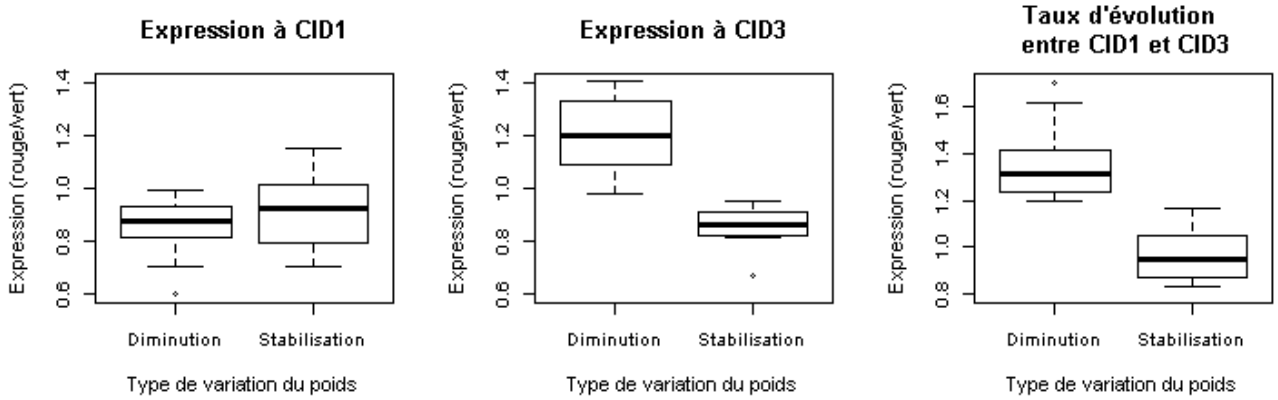
La figure 4.5 est le diagramme de Venn des gènes identifiés par les différentes méthodes. Le gène IGFBP3 est le seul qui ressort avec toutes les méthodes. De plus, 110 transcrits sont différents selon les groupes si l'on considère les résultats de toutes les méthodes exceptée celle de sélection de variables grâce aux forêts aléatoires (à l'aide du package `VSURF`). Le gène IGFBP3 est aussi le seul gène qui est déclaré avoir une évolution différente pour les deux groupes entre CID1 et CID3 par toutes les méthodes. Les boîtes à moustaches de l'expression de ce gène à CID1, CID3 et pour le taux d'évolution entre ces deux temps sont représentées dans le graphique 4.6.
Entre CID1 et CID3, le taux d'évolution de l'expression de ce gène pour les individus ayant diminué leur poids est, en moyenne, de 1,38 alors que celui des individus ayant stabilisé leur poids est, en moyenne, de 0,98. L'expression de IGFBP3 a donc augmenté pour le premier groupe et très légèrement diminué pour le second. À CID3, les boîtes à moustaches ne se chevauchent plus. Pour les individus du groupe ayant diminué leur poids, l'expression de IGFBP3 est plus élevée (moyenne de 1,21 ± 0,15 d'écart-type). Pour ceux de l'autre groupe, les expressions ce sont rapprochées de la moyenne (0,86 ± 0,08 d'écart-type).

Deux autres gènes potentiellement intéressants : VGLL3 et CES1

L'interprétation des résultats ne s'arrête pas au seul gène IGFBP3 qui ressort avec toutes les méthodes. D'autres peuvent être aussi intéressants d'un point de vue statistique ou biologique. J'ai choisi d'étudier l'expression des



Graphique 4.5 – Diagramme de Venn représentant les résultats communs aux différentes méthodes
Les zéro n'apparaissent pas pour plus de lisibilité.
Source : transcrits dont l'expression est différente à CID3 pour les individus à fortes variations de poids obtenus par cinq méthodes



Graphique 4.6 – Boîtes à moustache de l'expression du transcrit du gène IGFBP3 selon le type de variations à CID1, à CID3 et pour le taux d'évolution entre CID1 et CID3
Source : données corrigées des individus à fortes variations de poids

gènes VGLL3 et CES1 pour montrer que des différences existent aussi entre les deux groupes d'individus pour d'autres gènes. Le choix de ces gènes est subjectif mais s'explique par leurs caractéristiques misent en avant dans la suite du rapport.

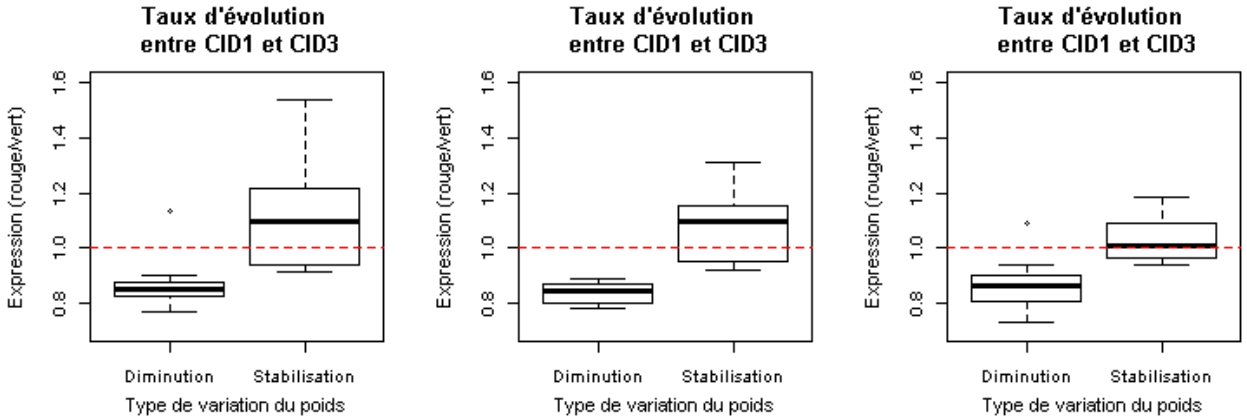
Le gène VGLL3 a un seul transcrit qui fait partie des 110 précédemement cités. De plus, il fait parti des quatre dont les boîtes à moustaches par groupe ne se chevauchent pas et son taux d'évolution ressort avec les forêts aléatoires. À CID3, l'expression de VGLL3 est presque 1,5 fois plus élevée en moyenne dans le groupe 2 des patients qui stabilisent leur poids que dans le groupe 2 de ceux qui le diminuent (tableau 4.2).

Le gène CES1 a, lui, deux transcrits (NM_001266 et NM_001025195). Ces deux transcrits ressortent aussi bien à CID3 que pour les taux d'évolution. Seul le premier transcrit cité n'évolue pas différentiellement mais il apparait dans les résultats des forêts aléatoires. L'expression de ces transcrits est plus faible pour les individus ayant diminué leur poids que pour les autres à CID3 (tableau 4.2). De plus, le groupe des individus pour lesquels le régime n'a pas marché est moins homogène que celui pour lequel il a marché.
L'évolution de ces trois transcrits est assez semblable (graphique 4.7). Pour les trois, l'expression a tendance à plus diminuer dans le groupe des individus ayant diminué leur poids alors que pour ceux l'ayant stabilisé, la

Gène	VGLL3	CES1	
Transcrit	NM_016206	NM_001266	NM_001025195
Moyenne dans le groupe 1	0,72	0,70	0,66
Moyenne dans le groupe 2	1,07	1,02	1,04
Ecart-type dans le groupe 1	0,09	0,12	0,08
Ecart-type dans le groupe 2	0,18	0,19	0,16
Erreur standard dans le groupe 1	0,01	0,004	0,01
Erreur standard dans le groupe 2	0,02	0,02	0,02

Tableau 4.2 – Statistiques descritptives sur les transcrits des gènes VGLL3 et CES1 à CID3
Source : données nettoyées et imputées des individus à fortes variations de poids à CID3

tendance est plutôt à la hausse.



Graphique 4.7 – Boîtes à moustache du taux d'évolution entre CID1 et CID3 de l'expression des transcrits des gènes VGLL3 et CES1
Source : données corrigées des individus à fortes variations de poids à CID3

D'autres gènes apparaissent dans les résultats des quatre méthodes mais ils ne peuvent pas tous être présentés ici. C'est aux biologistes d'en regarder les fonctions et les intéractions qu'ils peuvent avoir avec d'autres gènes pour en tirer les conclusions biologiques adéquates et ce sera l'objet pour eux du travail futur.

Quelques gènes trouvés par différentes méthodes statistiques

Pour les individus à fortes variations de poids, les différentes méthodes statistiques utilisées (tests et forêts alétoires) ont donné des résultats probants, surtout à CID3. Les gènes IGFBP3, CES1 et VGLL3 sont différentiellement exprimés entre le groupe d'individus à forte perte de poids et celui le stabilisant. De plus, leur expression est différente selon les groupes d'après quasiment toutes les méthodes utilisées.

Néanmoins, tous les résultats ne sont pas présentés ici mais ils sont à la disposition des biologistes sous forme de fichier csv pour qu'ils puissent les analyser. Il sera, ensuite, possible de combiner ces résultats avec ceux obtenus par d'autres méthodes statistiques. De plus, avec les scripts réalisés pendant mon stage, les méthodes peuvent être appliquées sur d'autres échantillons de données plus conséquents pour obtenir des résultats plus robustes.

Conclusion

Ce stage m'a permis de travailler pour la première fois en entreprise dans le monde de la recherche et de la biologie. Cela n'a fait que confirmer mon intérêt pour la biostatistique. De plus, j'ai pu travailler en collaboration avec des personnes de différents horizons : des biologistes, des statisticiens ainsi que des cliniciens. Cette collaboration est très enrichissante d'un point de vue personnel mais elle permet aussi d'aborder les problèmes et les résultats obtenus avec un sens différent : ce ne sont pas seulement des chiffres.

Au cours de ce stage, j'ai pu apprendre de nouvelles méthodes statistiques puis les appliquer aux différents jeux de données à ma disposition. Celles-ci étaient aussi bien générales comme les forêts aléatoires ou les ANOVA que plus spécifiques à la biostatistique comme la normalisation des biopuces. Ensuite, j'ai réalisé plusieurs scripts en RMarkdown pour transmettre l'application de ces méthodes ainsi que leurs résultats à mes tutrices. Enfin, la mise en commun des résultats sur chaque jeux de données a permis de faire ressortir quelques gènes différentiellement exprimés entre les deux groupes d'individus aux variations extrêmes de poids, de tour de taille ou de HOMA. Par exemple, le gène IGFBP3 s'exprime différemment à CID3 entre les individus à forte augmentation de poids et ceux qui le diminuent.

Durant ce stage, je n'ai pas pu appliquer la totalité des méthodes prévues au départ comme la PLS-DA. Il serait intéressant de les réaliser puis de comparer ces résultats à ceux préalablement obtenus. De plus, je n'ai travaillé que sur un nombre restreint d'individus. Il faudrait essayer de travailler sur des échantillons plus conséquents pour obtenir des résultats plus fiables.

Annexes

Annexe A — Quelques notions de biologie

Pour cette étude, des données issues de puces à ADN (biopuces) ont été analysées. Pour bien comprendre leur fonctionnement, il est nécessaire de s'approprier quelques notions de biologie.

A.1 L'ADN et l'ARNm

L'acide désoxyribonucléique (ADN) est une molécule, présente dans toutes les cellules vivantes, qui renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme. Il porte l'information génétique (génotype) et constitue le génome des êtres vivants.

La structure standard de l'ADN est une double-hélice, composée de deux brins complémentaires (image A.1). Chaque brin d'ADN est constitué d'un enchaînement de nucléotides. On trouve quatre nucléotides différents dans l'ADN, notés A, G, C et T, du nom des bases correspondantes. Ces nucléotides se regroupent par paires spéciales : A avec T, T avec A, C avec G et G avec C. Aucune autre paire n'est possible (sauf dans le cas de mutations génétiques).

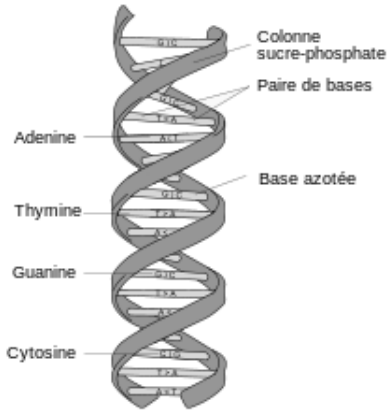


Image A.1 – Structure d'une molécule d'ADN
Source : l'image provient de Wikimedia Commons et est attribuable à MesserWoland

L'ADN est à l'origine de la synthèse des protéines, par l'intermédiaire de l'acide ribonucléique messager (ARNm) qui est une copie transitoire d'une portion de l'ADN correspondant à un ou plusieurs gènes. La synthèse des protéines se fait en plusieurs étapes (image A.2) :

1. La transcription, qui est le transfert de l'information génétique de l'ADN vers une autre molécule, l'ARN.
2. La traduction, qui est un transfert d'information depuis l'ARN vers les protéines.
3. L'activité des protéines.

Image A.2 – Étapes de la synthèse des protéines
Source : l'image provient de Wikimedia Commons et est attribuable à Toony

L'activité des protéines détermine l'activité des cellules, qui vont ensuite déterminer le fonctionnement des organes et de l'organisme. Pour cette étude, on ne va s'intéresser qu'à la transcription.

A.2 L'expression des gènes

L'expression des gènes désigne l'ensemble des processus biochimiques par lesquels l'information héréditaire stockée dans un gène est lue pour aboutir à la fabrication de molécules qui auront un rôle actif dans le fonctionnement cellulaire, comme les protéines ou les ARN.

Même si toutes les cellules d'un organisme partagent le même génome, certains gènes ne sont exprimés que dans certaines cellules, à certaines périodes de la vie de l'organisme ou sous certaines conditions. La régulation de l'expression génique est donc le mécanisme fondamental permettant la différenciation cellulaire, la morphogenèse et l'adaptabilité d'un organisme vivant à son environnement. Par exemple, les multiples couleurs d'un chat (image A.3) sont le résultat de différents niveau d'expression des gènes responsable de la pigmentation à plusieurs endroits de sa peau.

Image A.3 – Exemple d'expression génique pour la couleur d'un chat
Source : l'image provient de Wikimedia Commons et est attribuable à Sannse

Mesurer l'expression des gènes est une part importante de beaucoup de recherches en sciences de la vie. Pouvoir quantifier le niveau d'expression d'un gène particulier dans une cellule, un tissu ou un organisme peut apporter beaucoup d'informations sur le fonctionnement de la cellule. Cela peut, par exemple, permettre de déterminer le susceptibilité d'un individu à un cancer ou de trouver si une bactérie est résistante à un médicament.

Pour mesurer l'expression des gènes, on va quantifier le niveau d'ARNm. Pour cela il existe plusieurs méthodes : *Northern blot*, RT-PCR, RNAseq et les puces à ADN (biopuces). Dans cette étude, les données utilisées sont issues de biopuces. Les autres méthodes ne seront donc pas expliquées ici.

A.3 La technologie biopuce

Une biopuce, ou puce à ADN, est un ensemble de molécules d'ADN fixées en rangées ordonnées sur une petite surface qui peut être du verre, du silicium ou du plastique (image A.4) et repose sur une lame. Dans le cadre de cette étude, on compte quatre puces par lames. Cette biotechnologie récente permet d'analyser le niveau d'expression des gènes (transcrits) dans une cellule, un tissu, un organe, ou un organisme, à un moment donné et dans un état donné par rapport à un échantillon de référence.

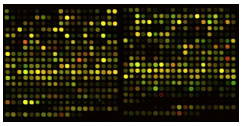


Image A.4 – Fragment de biopuce
Source : l'image provient de Wikimedia Commons et est attribuable à Mangapoco

Le principe de la puce à ADN repose sur la propriété que possède l'ADN dénaturé de reformer spontanément sa double hélice lorsqu'il est porté face à un brin complémentaire (réaction d'hybridation). Les quatre bases azotées de l'ADN (A, G, C, T) ont en effet la particularité de s'unir deux à deux. Par exemple, si un patient est porteur d'une maladie, les brins extraits de l'ARN d'un patient (et rétrotranscrits en ADN), vont s'hybrider avec les brins d'ADN synthétiques représentatifs de la maladie.

Pour mesurer l'expression des gènes, on commence par créer une plaque contenant de l'ADN connu, avec un gène sur chaque « sonde » (image A.5). Les cibles qui vont s'hybridées aux sondes sont complexes : l'ARNm de la cellule à étudier sont marquées par un fluorochrome émettant dans la rouge ou dans le vert. Ensuite,

on met en contact les cibles (marquées en vert ou rouge) et la plaque. Si les brins d'ADN sont identiques, ils s'associeront, et le sonde sera plus ou moins fluorescente selon la quantité de brins présents. Enfin, on récupère comme information, pour chaque gène, la mesure de son expression à travers la quantité de fluorescence émise.

On utilise toujours un échantillon de référence marqué en vert contre un échantillon de l'individu dont on souhaite étudier l'expression des gènes marqué en rouge. On a donc une sur-expression du gène chez le patient si le sonde est rouge et une sous-expression si elle est verte.

Pour les puces utilisées dans cette étude, deux sondes peuvent correspondre à deux variantes de transcrits d'un seul et même gène. C'est pour cela que dans la suite le terme transcrit sera utilisé à la place de gène sauf lorsque le gène n'a qu'un seul transcrit.

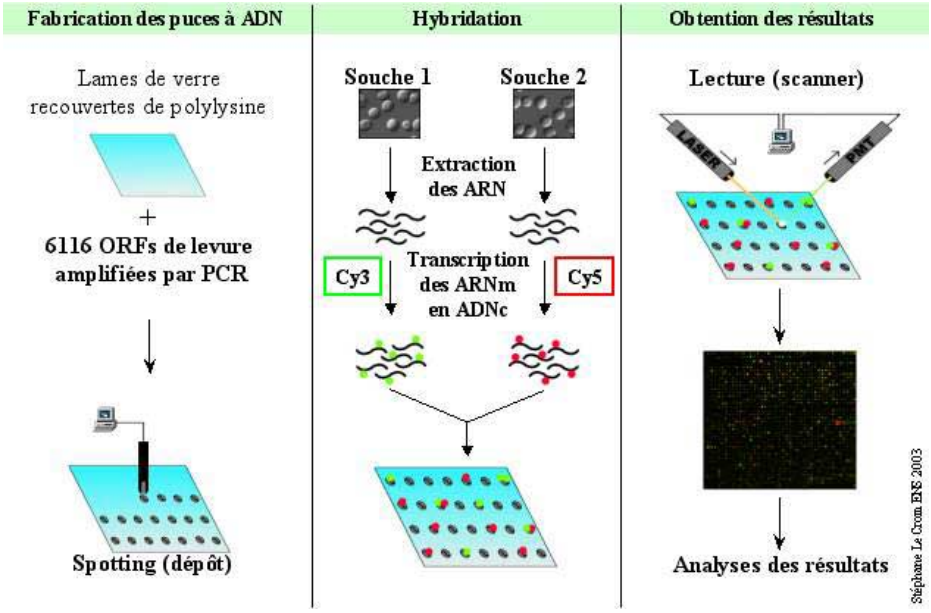


Image A.5 – Principe de la puce à ADN
Source : l'image provient de la page de Stéphane Le Crom et Philippe Marc sur le site de la plate-forme transcriptome de l'ENS

Annexe B — Scripts en RMarkdown réalisés

Pour permettre à d'autres personnes, statisticiens ou non, d'utiliser mon travail pour des études similaires, j'ai produit des fichiers en RMarkdown avec le package `knitr` [1] de  et RStudio [2]. On peut grâce à cela produire, assez facilement, des documents html contenant du texte simple mais aussi du code R et des graphiques. Lors de la compilation le code R sera exécuté pour donner des résultats numériques ou des graphiques qui seront directement intégrés au fichier html. L'avantage des fichiers réalisés en RMarkdown est que l'on va obtenir le code R, les résultats ou graphiques et les commentaires sur un document structuré simple à réalisé, le langage Markdown  [3] étant un language à la syntaxe très simplifiée.

J'ai, dans un premier temps, réalisé un fichier pour l'apurement des données et un autre pour la correction des effets. Le premier cité est donné comme exemple ci-dessous. Ensuite, j'ai réalisé un fichier par type de variation (HOMA, poids ou tour de taille) et par type de méthode utilisée. Tous ces fichiers ainsi que des fichiers Excel contenant les résultats ont été remis à mes tutrices à la fin du stage pour qu'elles puissent voir les étapes des démarches statistiques et interpréter les résultats de manière biologique.

1. http://cran.r-project.org/web/packages/knitr/index.html
2. http://www.rstudio.com/
3. http://daringfireball.net/projects/markdown/

Normalisation des données biopuces

Gaëlle Lefort

Première version : 10 juin 2014
Dernière mise à jour : 17 juillet 2014

Ce script explique les différentes étapes de la normalisation et du nettoyage de données de biopuces.

Il peut être executé avec **R** version 3.1.0 en utilisant le package suivant :

- limma version 3.20.4 pour manipuler et normaliser les données de biopuce (Bioconductor).

Importation du package :

```
# Décommenter pour installer limma...
# source("http://bioconductor.org/biocLite.R")
# biocLite("limma")
library(limma)
```

1. Données

Protocole expérimental : dans le cadre du projet DiOGenes, des individus obèses ont effectué un régime. Pendant les 8 premières semaines (phase de restriction), le régime était un régime basse calorie. Ensuite, pendant 6 mois (phase de stabilisation), les patients dont le poids a diminué d'au moins 8% pouvaient continuer le protocole, et étaient randomisés dans une des cinq branches de régime « ad libitum ».

Biopuces réalisées : des biopuces ont été réalisées pour collecter l'expression des gènes :

- des 11 individus de l'expérience qui ont l'augmentation la plus importante de poids (weight) entre CID1 (début de l'étude) et CID3 (fin de la phase de stabilisation) et les 11 individus qui ont la diminution la plus importante de poids entre ces deux mêmes moments ;
- les 2x11 individus correspondant aux variations les plus importantes (augmentation et diminution) de tour de taille (waist) ;
- les 2x11 individus correspondant aux variations les plus importantes de HOMA.

Certains individus sont présents pour plusieurs critères, il y a 52 patients (soit 104 biopuces, une pour CID1 et une pour CID3).

Convention de nommage des fichiers : DIOGENES-[Date]-[heure]-[Numéro de lame]-5µm-b[Numéro de bloc].txt

2. Importation des données

Les données sont au format texte (.txt) avec 1 fichier par individu. Pour faciliter l'importation, ils sont tous dans un même dossier.

```
cheminImport<-"../../1. rawdata/Toutes les lames/"
# cheminImport <- "../0-rawData"
cheminExport<-"../../2. normdata/"
# cheminExport <- "../1-normData/"
cheminCorresp <- "../../1. rawdata/"
# cheminCorresp <- "../0-rawData/"
```

2.1. Configuration de l'importation

Pour faciliter la lecture dans les fichiers qui seront créés ensuite, on peut donner des noms aux différentes biopuces. Si on ne souhaite pas le faire, les noms seront ceux des fichiers texte. Ici j'ai choisi de prendre les 3 derniers chiffres du numéro de lame et le numéro du bloc sur la lame pour obtenir un nom du type "L[n° de lame] B[n° du bloc]". Ces données sont extraites à partir d'une expression régulière basée sur le nom du fichier pour limiter les risques d'erreur.

```
files <- dir(path=cheminImport, pattern="/*.txt")

allB <-  unlist(lapply(regmatches(files,regexec("_(b[1-4])",files)), function(u)
  u[2]))
allL <- paste0("L",
               substr(unlist(lapply(regmatches(files,
                                               regexec("_([0-9]*)_",files)),
                                    function(u) u[2])), 10, 12))
nomLame <- paste0(allL,".",allB)
```

Ensuite, il faut créer la fonction de filtre nécessaire à l'importation. Un seuil de 1.2 a été choisi par rapport à des expériences similaires menées à l'INSERM.

```
# Fonction de filtrage
myfun <- function(x,threshold=1.2) {
    okred<-x[,"F635 Median"]/x[,"B635 Median"] > threshold
    okgreen<-x[,"F532 Median"]/x[,"B532 Median"] > threshold
    okFlag<-(x[,"Flags"]>=0 & x[,"ControlType"]=="false")
    as.numeric((okgreen | okred) & okFlag)
}
```

Fait après avoir interprété les résultats du script ACPResumeAvecC3 :
On a pu voir grâce à l'analyse exploratoire que les individus atypiques venaient exclusivement du centre 3. On va donc tous les supprimer (5 individus).

```
# Importation des correspondance lame-individu
proto <- read.csv(paste0(cheminCorresp,"corresp.csv"), header=TRUE, sep=",",
                  dec=".")

# Obtention du tableau (dans l'ordre des fichiers dans le dossier) contenant les centres de chaque individus
correspondance <- merge(data.frame(lame=nomLame), proto, by.x="lame",
                        by.y="Lame_Bloc", sort=FALSE)

# Suppression des fichiers du centre 3
centre3 <- correspondance$Centre=="3"
files <- files[!centre3]
nomLame <- nomLame[!centre3]
```

2.2. Importation

Importation des données en spécifiant le nom des fichiers à importer, le chemin où ils se trouvent, leur source, la fonction de filtrage et le nom à attribuer à chaque biopuce :

```
raw.data.AvecControls <- read.maimages(files, path=cheminImport,
                                       source="genepix", wt.fun=myfun,
                                       names=nomLame)
```

On obtient un objet contenant 94 biopuces de 45220 spots.

2.3. Suppression des spots de contrôle

Pour les analyses ultérieures, on supprime les spots de contrôle des données. Pour les détecter, on change le filtre d'importation des données.

```
# Filtre pour obtenir les spots de contrôle
control <- function(x) {
    okred <- TRUE
    okgreen <- TRUE
    okFlag <- (x[,"ControlType"]=="false")
    as.numeric((okgreen | okred) & okFlag)
}

# Importation d'un fichier pour connaître l'emplacement des spots de contrôle
raw.data.Controls <- read.maimages(files[1], path=cheminImport,
                                   source="genepix", wt.fun=control)

# Emplacement des spots de contrôle
control <- raw.data.Controls$weights[,1]==0

# Données sans les spots de contrôle
raw.data <- raw.data.AvecControls[!control,]
```

Pour le traitement des données filtrées, on choisit ici de les remplacer par des données manquantes.

```
raw.data$R[raw.data$weights==0] <- NA
```

```
raw.data$G[raw.data$weights==0] <- NA
```

On obtient un objet contenant 94 biopuces de 43118 spots.

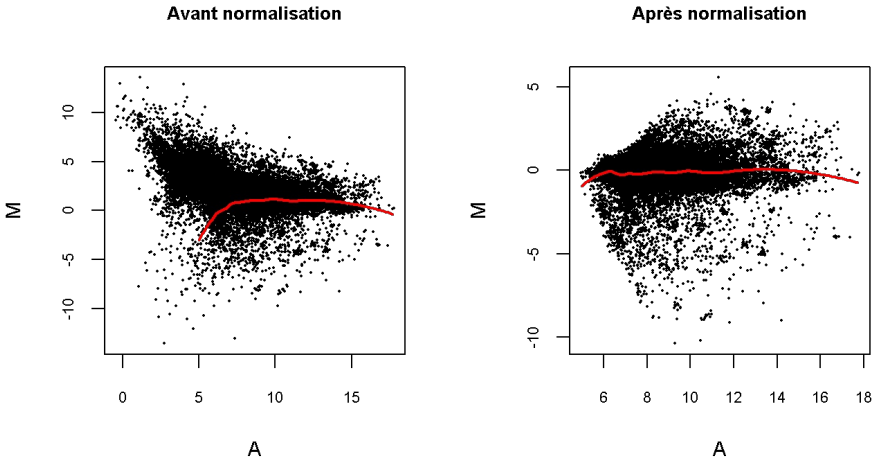
3. Normalisation intra et inter lames

3.1. Normalisation intra-lame

On utilise ici la méthode LOESS avec comme paramètre de lissage 0,3 (paramètre par défaut) sans soustraction du bruit de fond (celle-ci a déjà été faite au préalable ; dans le cas contraire, il aurait fallu remplacer bc.method="none" par "subtract").

```
MA <- normalizeWithinArrays(raw.data,method="loess",bc.method="none")
```

Visualisation d'une normalisation intra-lames avec régression loess sur un bloc :

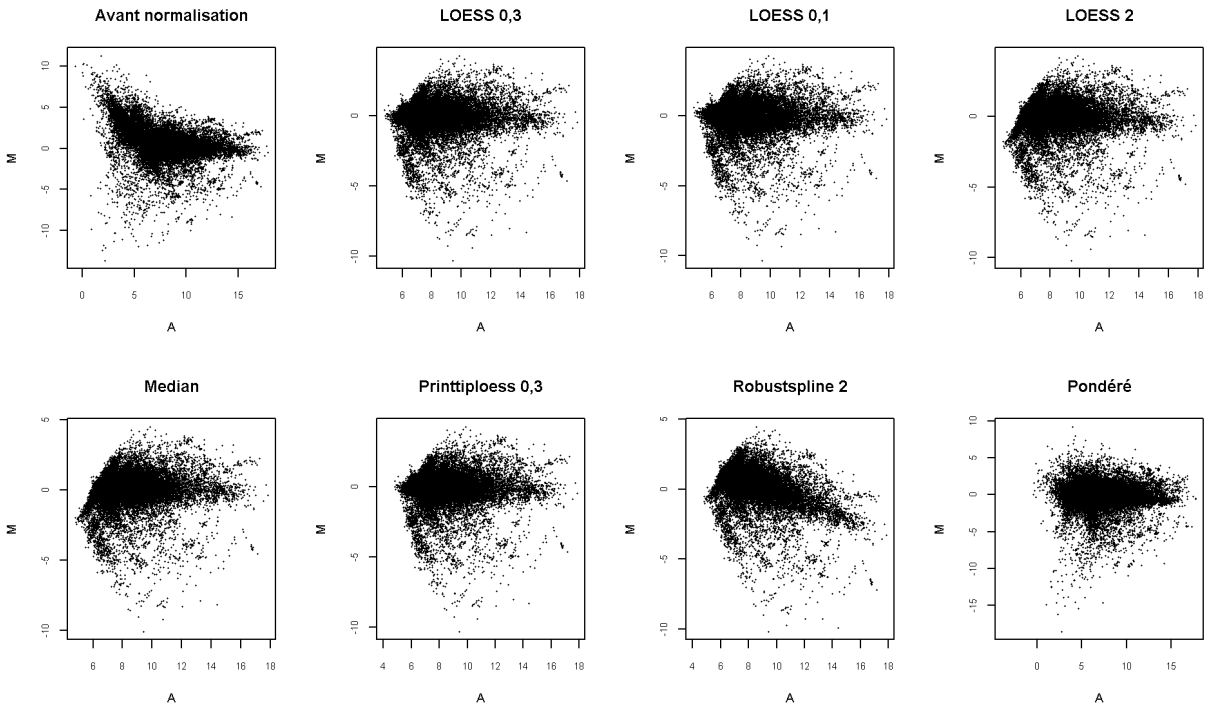


Essai des autres méthodes

Ici, d'autres méthodes de normalisation (ou d'autres paramètres de lissage) sont testés pour voir si il existe une différence avec la méthode de base.

```
essai1 <- normalizeWithinArrays(raw.data[,34], method="loess", bc.method="none",
                                span=0.3)
essai2 <- normalizeWithinArrays(raw.data[,34], method="loess", bc.method="none",
                                span=0.1)
essai3 <- normalizeWithinArrays(raw.data[,34], method="loess", bc.method="none",
                                span=2)
essai4 <- normalizeWithinArrays(raw.data[,34], method="median",
                                bc.method="none")
essai5 <- normalizeWithinArrays(raw.data.AvecControls[,34],
                                method="printtiploess", bc.method="none")
essai9 <- normalizeWithinArrays(raw.data.AvecControls,
                                weights=raw.data.AvecControls$weights)
essai8 <- normalizeWithinArrays(raw.data.AvecControls[,34],
                                method="robustspline", bc.method="none",
                                df=2)
```

```
## Loading required package: splines
```



Les méthodes alternatives ne semblent pas donner de meilleurs résultats sur la lame considérée. La méthode initiale est donc conservée.

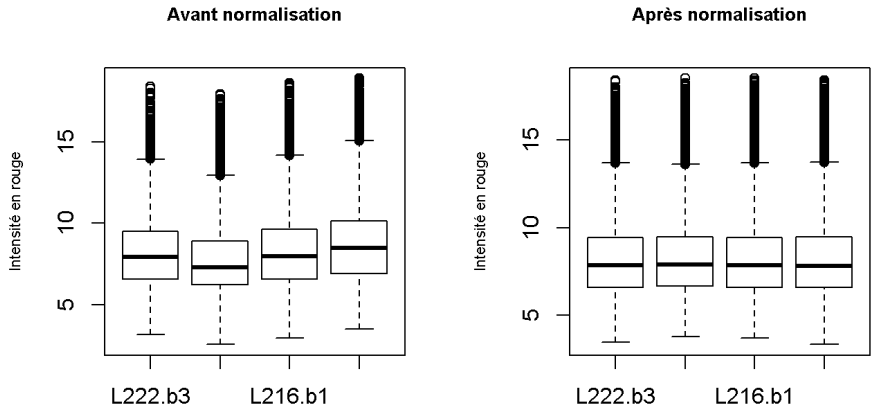
3.2. Normalisation inter-lames

On utilise pour cette normalisation la méthode “Aquantile” pour aligner les quantiles de A entre les lames.

```
MA <- normalizeBetweenArrays(MA, method="Aquantile")

# Données normalisées
norm.data <- RG.MA(MA)
```

Visualisation d’une normalisation inter-lames sur une lame pour le rouge :



4. Exportation du tableau de données

Avant l’exportation, on supprime les gènes qui n’ont que des valeurs manquantes :

```
diff <- log2(norm.data$R)-log2(norm.data$G)
normdata <- data.frame(ID=norm.data$genes$ID, Name=norm.data$genes$Name, diff)
```

```
# Part de gènes avec des valeurs manquantes seulement
sum(apply(normdata, 1, function(x) sum(is.na(x))) == ncol(normdata)-2)/nrow(normdata)
```

```
## [1] 0.01143
```

```
# Suppression de ces gènes
toKeep <- which(apply(normdata,1,function(x) sum(is.na(x))) != ncol(normdata)-2)
normdata <- normdata[toKeep,]
```

Exportation des données sous la forme d'un tableau où les lignes sont les spots, les colonnes les blocs et les valeurs les différences des logarithmes (en base 2) des intensités de rouge et de vert.

```
write.table(normdata, paste0(cheminExport, "normdata.csv"), sep=",", dec=".")
```

5. Spots en double

Certains spots ont le même nom de gène (NM_...), on les moyenne pour que n'apparaisse qu'une seule fois chaque nom.

```
# Vecteur contenant les noms de manière unique
nom <- as.vector(unique(normdata$Name))

# Fonction permettant de moyenner les colonnes pour un nom donné
moy <- function (x, data) {
    colMeans(data[(1:nrow(data))[na.omit(data$Name)==nom[x]], 3:ncol(data)],
             na.rm=TRUE)
}
```

```
# Obtention des moyennes pour chaque nom
uniqueExp <- t(sapply(1:length(nom), moy, normdata))
```

```
# Tableau de données sans doublons
normdata.sansD <- data.frame(uniqueExp)
row.names(normdata.sansD) <- nom
```

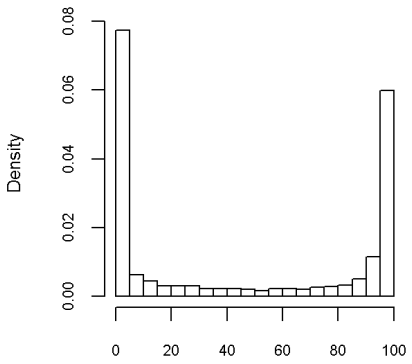
On obtient un objet contenant 94 biopuces de 26838 spots.

6. Valeurs manquantes

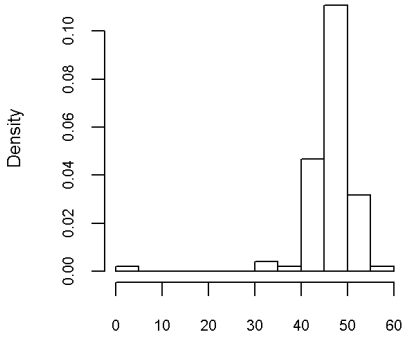
Certains gènes ont beaucoup de valeurs manquantes ce qui peut poser problème dans les analyses statistiques ultérieures.

```
par(mfrow=c(1,2))
hist(apply(normdata.sansD, 1, function(x) sum(is.na(x))/(ncol(normdata)-2)*100),
     freq=FALSE, xlab="Poucentage de valeurs manquantes par gène", main="",
     cex.lab=0.8, cex.axis=0.7)

hist(colSums(is.na(normdata.sansD))/nrow(normdata.sansD)*100, freq=FALSE,
     xlab="Pourcentage de valeurs manquantes par individu", main="", cex.lab=0.8,
     cex.axis=0.7)
```



Poucentage de valeurs manquantes par gène



Pourcentage de valeurs manquantes par individu

Seulement 46% des gènes ont moins de 20% de valeurs manquantes. Pour la suite, on ne va garder que les gènes ayant moins de 20% de valeurs manquantes.

```
normdata.sansM <- normdata.sansD[!apply(normdata.sansD, 1, function(x)
  sum(is.na(x))) > 20/100*ncol(normdata.sansD),]
```

On obtient un objet contenant 94 biopuces de 12236 spots.

Enfin, les expressions normalisées correspondant aux gènes uniques et ayant moins de 20% de valeurs manquantes sont exportées :

```
write.table(normdata.sansM, file=paste0(cheminExport,"cleandata.csv"), sep=",", dec=".")
```

Annexe C — Statistiques des tests utilisés

C.1 Test de Fischer

Soit n le nombre total d'individus, p le nombre de groupes d'individus) et $y_1, ..., y_n$ l'expression d'un gène. La statistique du test de Fischer est

$$F = \frac{\frac{SCE_{facteur}}{p-1}}{\frac{SCE_{residu}}{n-p}} \sim Fischer(p-1, n-p) \tag{C.1}$$

où

$$SCE_{facteur} = \sum_{i=1}^{p} n_i(\bar{y_i} - \bar{y})^2 \sim \chi^2(DDL_{facteur}) \quad avec \;\; DDL_{facteur} = p-1$$

et

$$SCE_{residu} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(y_i^j - \bar{y_i})^2 \sim \chi^2(DDL_{residu}) \quad avec \;\; DDL_{residu} = n-p$$

C.2 Test de Welch

Soit μ_i, σ_i^2 et n_i la moyenne, la variance et la taille du groupe i. La statistique du test de Welch est alors :

$$T = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim Student(\lfloor \frac{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^2}{\frac{\frac{\sigma_1^2}{n_1}}{n_1-1} + \frac{\frac{\sigma_2^2}{n_2}}{n_2-1}} \rfloor) \quad avec \;\; \lfloor c \rfloor \; la \; partie \; entire \; de \; c \tag{C.2}$$

C.3 Test de Mann-Whitney

Pour obtenir la statistique U_{n_1,n_2} du test de Mann-Whitney, on procède à des calculs successifs :

1. on classe par ordre croissant l'ensemble des observations des deux groupes ;
2. on affecte le rang correspondant ;
3. on effectue les sommes des rangs pour chacun des deux groupes, notées R_{n_1} et R_{n_2} ;
4. on en déduit les quantités U_{n_1} et U_{n_2} qui se calculent ainsi :

$$U_{n_1} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_{n_1} \quad et \quad U_{n_2} = n_1 n_2 - U_{n_1} \tag{C.3}$$

Par conséquent, la statistique U_{n_1,n_2} se définit comme étant la plus petite des deux valeurs U_{n_1} et U_{n_2}. Lorsque les tailles des n_1 et n_2 des deux groupes sont inférieures ou égales à 20, comme dans ce cas, les tables de Mann-Whitney fournissent une valeur critique c. Si la valeur de la statistique de test est inférieure ou égale à c alors on rejette H_0.

C.4 Test de Bartlett

Soit k échantillons de taille n_i et de variances empiriques S_i^2, alors le test de Bartlett est tel que

$$X^2 = \frac{(N-k)\ln(S_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(S_i^2)}{1 + \frac{1}{3(k-1)}(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k})} \sim \chi_{k-1}^2 \tag{C.4}$$

où

$$N = \sum_{i=1}^{k} n_i \quad et \quad S_p^2 = \frac{1}{N-k} \sum_i (n_i - 1) S_i^2 \tag{C.5}$$

C.5 Test de Shapiro-Wilk

Soit $x_{(i)}$ la i^e statistique d'ordre, $\bar{x}$ la moyenne de l'échantillon et a_i la constante donnée par $(a_i, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$ où m est la transposée du vecteur des espérances et V est la matrice de variance-covariance. La statistique de test est :

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{C.6}$$

Cette statistique suit une loi de Shapiro-Wilk.

C.6 Test de Kolmogorov-Smirnov

Soit F_n et F, respectivement les fonctions de distribution observée et théorique de X, la statistique D du test de Kolmogorov-Smirnov est alors

$$D = \max_x |F_n(x) - F(x)| \tag{C.7}$$

Cette statistique suit une loi de Kolmogorov-Smirnov à n degré de libertés.

Bibliographie

[1] Organisation Mondiale de la Santé (OMS), Aide-mémoire N° 311, Mai 2014.

[2] Viguerie N., Montastier E., Maoret J., Roussel B., Combes M., Valle C., Villa-Vialaneix N., Iacovoni J., Martinez J., Holst C., Astrup A., Vidal H., Clément K., Hager J., Saris W., et Langin D., *Determinants of human adipose tissue gene expression : impact of diet, sex, metabolic status and cis genetic regulation*, PLoS Genetics, 2012.

[3] Bolstad, B. M., Irizarry R. A., Astrand, M, and Speed, T. P., *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. Bioinformatics 19(2), pp 185-193, 2003.

[4] Razali N.M, Wah Y. B., *Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lillefors and Anderson-Darling tests*, Journal of Statistical Modeling and Analytics Vol.2 No.1, 21-33, 2011.

[5] Benjamini Y., Hochberg Y., *Controlling the false discovery rate : a practical and powerful approach to multiple testing* J. Royal Stat. Soc. B, 85, 289-300, 1995.

[6] Friguet C., Kloareg M., and Causeur D., *A Factor Model Approach to Multiple Testing Under Dependence* Journal of the American Statistical Association, 104 :488, p.1406-1415, 2009.

[7] Causeur D., Friguet C., Houee-Bigot M., Kloareg M., *Factor Analysis for Multiple Testing (FAMT) : An R package for large-scale significance testing under dependence*, Journal of Statistical Software, May 2011.

[8] Breiman L., *Random Forests*, Machine Learning, 45, 5–32, 2001.

[9] Genuer R., Poggi J.-M., Tuleau-Malot C., *Variable selection using Random Forests* Pattern Recognition Letters 31 :2225-2236, 2010.