**AGROCAMPUS OUEST**

☐ **CFR Angers**

☑ **CFR Rennes**

Année universitaire : 2016-2017

Spécialité : Data Science

Spécialisation (et option éventuelle) :

……

**Mémoire de Fin d'Études**

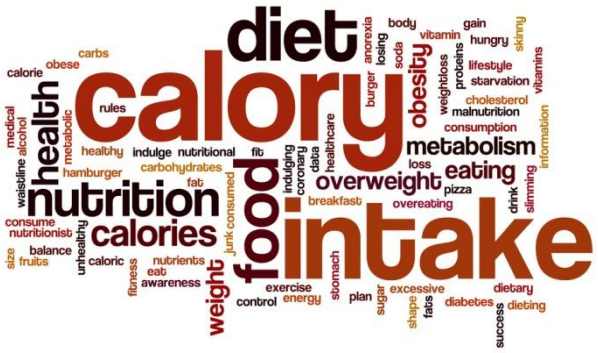
☐ d'Ingénieur de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☐ de Master de l'Institut Supérieur des Sciences agronomiques, agroalimentaires, horticoles et du paysage

☑ d'un autre établissement (étudiant arrivé en M2)

# Analyse statistique d'une protéine impliquée dans les problèmes d'obésité en relation avec l'expression des gènes

Par : Thibaut GUIGNARD




***Soutenu à Rennes le 06/09/2017***

***Devant le jury composé de :***

Président : Grégoire THOMAS

Maître de stage : Nathalie VILLA-VIALANEIX

Enseignant référent : Mathieu EMILY

Autres membres du jury (Nom, Qualité)

# Fiche de confidentialité et de diffusion du mémoire

## Confidentialité

☐ Non ☒ Oui si oui : ☐ 1 an ☒ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible (1).

Date et signature du maître de stage (2) : 16/06/2017

**A la fin de la période de confidentialité,** sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

## Droits d'auteur

L'auteur(3) **Nom Prénom** ----------

autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur :

## Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

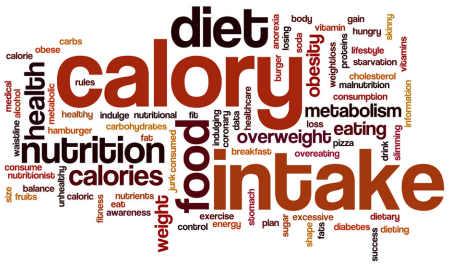
Date et signature de l'enseignant :

(1) L'administration, les enseignants et les différents services de documentation d'AGROCAMPUS OUEST s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3).Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Année scolaire 2016-2017

Stage 5e année

# Analyse statistique d'une protéine impliquée dans les problèmes d'obésité en relation avec l'expression des gènes





Tutrices

Nathalie Villa-Vialaneix
Alyssa Imbert
Nathalie Viguerie

Étudiant

Thibaut Guignard

# Table des matières

# Remerciements

Je tiens tout d'abord à remercier Nathalie Villa-Vialaneix (dit NV2), Chargée de Recherche - HDR au sein de l'unité MIAT à l'INRA de Castanet-Tolosan, et Nathalie Viguerie (dit NV1), Chargée de Recherche - HDR au sein de l'équipe I2MC à l'INSERM de Toulouse, mes maîtres de stage, de m'avoir accepté en tant que stagiaire, pour leurs conseils avisés et la confiance qu'elles m'ont accordé tout au long de mon stage.

Je remercie également Alyssa Imbert, thésarde au sein de l'unité MIAT, pour l'attention et l'aide qu'elle a pu m'apporter au quotidien pendant mon stage au sein de l'INRA.

Je souhaite également remercier Ronan Trépos et Eric Casellas, Ingénieurs d'études dans l'unité MIAT, pour leur accueil au sein de leur bureau MIAT46, pour leur bonne humeur, ainsi que pour leurs différents conseils.

Enfin, je tiens à remercier l'ensemble de l'équipe MIAT et également l'équipe de Flag rugby de l'INRA pour leur accueil pendant la durée de mon stage.

Toutes ces personnes ont contribué, par leur disponibilité et leur bonne humeur, à rendre mon stage enrichissant et motivant.

# Introduction

Ce rapport présente le travail que j'ai fourni lors de mon stage, qui s'est déroulé du 13 février 2017 au 11 août 2017, au sein de l'unité MIAT de l'INRA de Castanet-Tolosan.

Ce stage, d'une durée de 6 mois, a consisté à effectuer de la fouille de données et de l'apprentissage statistique. La variabilité des missions et des méthodes statistiques ont été des facteurs déterminants dans mon choix de postuler à ce stage. J'ai également choisi de postuler à ce stage car il répondait à mes attentes, tant sur le plan personnel que sur le plan professionnel. Le domaine de la biostatistique m'intéressait et ce stage fut une opportunité d'approfondir cet intérêt et de vérifier si j'ai les compétences nécessaires pour poursuivre dans ce domaine ou non.

L'objectif de ce stage n'est pas uniquement de me faire progresser sur le plan professionnel, d'apprendre nombre de choses et d'employer celles déjà vues dans mes années d'études, mais également de m'imprégner de l'ambiance qui règne dans une entreprise afin de me faire grandir humainement, pour être prêt à affronter différents problèmes dans les années à venir.

La mission réalisée s'est avérée très enrichissante pour mon expérience professionnelle. Grâce à ce stage, j'ai pu avoir un aperçu de ce qu'est la profession de Statisticien dans ce secteur d'activité.

Je vous expose dans ce rapport, dans un premier temps, une présentation des centres de recherche et unités avec lesquels j'ai collaboré pendant mon stage. Ensuite, je vous explique les différents aspects de mon travail (contexte, résultats...) durant ces 6 mois. Et enfin, en conclusion, je résume ce que ce stage m'a apporté.

Partie 1

# MIAT, L'Inserm, l'I2MC et l'équipe 4

J'ai passé les six mois de mon stage de fin d'année à l'Institut National de Recherche Agronomique (INRA) dans l'unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT, anciennement unité de Biométrie et Intelligence Artificielle). Ce laboratoire est situé sur la ville de Castanet Tolosan (31). J'ai également réalisé mon stage en étroite collaboration avec le laboratoire I2MC de l'INSERM de Toulouse.

## L'Institut National de la Recherche Agronomique (INRA)

L'INRA est un Etablissement Public à caractère Scientifique et Technologique (EPST, au même titre que le CNRS ou que l'INSERM), essentiellement financé par des fonds publics (il rend compte de son activité et de sa gestion à ses ministères de tutelle, le ministère de l'Enseignement supérieur et de la Recherche et le ministère de l'Alimentation, de l'Agriculture et de la Pêche). L'INRA compte environ 400 unités de recherche réparties dans 19 centres (localisations) et 14 départements (grandes thématiques) de recherche (une unité correspond à un centre et a un département de rattachement). Les unités expérimentales de l'INRA couvrent environ 12 000 hectares dont 3 000 hectares de forêts. Parmi le cheptel de l'INRA, on peut compter environ 6000 bovins, 16 000 ovins, 8 000 porcins, 300 équins, 34 000 volailles, une centaine de cervidés et une dizaine de lamas.
L'INRA renforce ses activités autour de trois champs :

1. le développement d'une agriculture durable,
2. l'alimentation et son rôle sur la santé humaine
3. l'environnement et les territoires

et a pour missions de produire et diffuser des connaissances scientifiques et des innovations, contribuer à la formation et, par la recherche, à la diffusion de la culture scientifique et au débat science/société, participer par son expertise à éclairer les décisions des acteurs publics et privés.
Les quatre priorités de recherche de l'INRA sont :

1. protéger les ressources naturelles,
2. manger sain et sûr,
3. passer des génomes aux populations végétales et animales,
4. travailler avec l'informatique et la biologie à haut débit.

### L'Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT)

L'Unité de Mathématiques et Informatique Appliquées de Toulouse est une unité propre du département de Mathématiques et Informatique Appliquées (MIA) de l'INRA. Comme toutes les unités de MIA, MIAT a pour mission scientifique de développer et mettre en œuvre des méthodes mathématiques et/ou informatiques pertinentes pour résoudre des problèmes identifiés avec les collaborateurs biologistes qui sont issus principalement d'autres départements de l'INRA.
L'unité comporte actuellement (depuis janvier 2011) deux équipes de recherche (MAD et SaAB) et trois équipes de service (Plateformes GENOTOUL, RECORD et SIGENAE) :

- MAD (Modélisation des Agro-écosystèmes et Décision)
- SaAB (Statistiques et Algorithmique pour la Biologie)
- GENOTOUL (Plateforme bioinformatique du GIS GENOTOUL - Génopole Toulouse Midi-Pyrénées)
- RECORD (Plateforme de modélisation et de simulation des agro-écosystèmes)
- SIGENAE (Plateforme Systèmes d'information des génomes des animaux d'élevage)

Relever le défi des sciences du vivant suppose de développer les relations entre calcul, modélisation et les diverses composantes des sciences du vivant (biologie, agronomie, écologie, sciences de l'environnement)
L'informatique, la statistique et, plus généralement, les mathématiques sont indispensables pour organiser les données sur ces objets, pour en permettre l'accès, pour représenter les modes d'organisation et pour simuler les interactions au sein de systèmes complexes qu'ils sont ou dans lesquels ils s'inscrivent. Cela requiert de concevoir des outils formels et calculatoires d'analyse de données, de prédiction et d'optimisation. Ces disciplines jouent un rôle fondamental dans la recherche de sens, dans la compréhension des fonctionnements des systèmes et dans le pilotage ou la gestion de ceux de nature anthropique. La mission de l'unité est de contribuer à apporter une réponse à ces besoins.
Ces recherches s'accompagnent d'une activité de production de logiciels, pour la valorisation des méthodes développées, et d'une activité de formation, pour leur diffusion. L'unité poursuit une double ambition de production disciplinaire sur ses domaines de compétences et de production finalisée en construisant des projets et en collaborant avec des biologistes et des agronomes.

### Statistiques et Algorithmique pour la Biologie (SaAB)

L'équipe a pour objectif de développer et de mettre à disposition des biologistes des méthodes mathématiques, statistiques et informatiques permettant de contribuer à la compréhension du vivant.
L'équipe s'intéresse à la localisation et l'identification d'éléments fonctionnels dans les génomes des bactéries, plantes et animaux, et de façon croissante aux interactions qui existent entre ces différents éléments :

- au niveau génétique
- au niveau molécule ADN/ARN
- au niveau molécule de protéine
- au niveau de l'expression de gènes

Pour traiter ces problèmes, l'équipe mobilise et développe des méthodes en mathématiques, statistiques, probabilités (modélisation, inférence, modèles de mélanges de lois, régression pénalisée, modèles graphiques stochastiques, processus) et en informatique (modélisation, optimisation combinatoire, réseaux de contraintes,

modèles graphiques déterministes, algorithmique) avec le but de valoriser les méthodes développées dans des outils logiciels directement utilisables par nos partenaires biologistes et rendant compte le mieux possible de la complexité et de la variété des données utilisables et en capitalisant les développements méthodologiques dans des logiciels génériques, éventuellement déclinés ensuite sur différentes applications.

## L'Institut national de la santé et de la recherche médicale (Inserm)

L'institut national de la santé et de la recherche médicale 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. L'Inserm s'est vu confier, en 2008, la responsabilité d'assurer la coordination stratégique, scientifique et opérationnelle de la recherche biomédicale. L'Inserm, et d'autres grands acteurs de la recherhe biomédicale française, composent l'Aviesan (Alliance nationale pour les sciences de la vie et de la santé) (huit grands établissements publics, CNRS, INRIA, Inserm, Institut Pasteur, auxquels s'associent les universités et les CHU). Cette alliance a été fondée en avril 2009 afin de coordonner les équipes et les programmes de la recherche biomédicale en France.
L'Inserm a 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, la physiopathologie, la thérapie génique, l'épidémiologie, l'imagerie médicale... De plus, l'Inserm joue un rôle important 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).
Il est structuré en unités de recherche de tailles variées, le plus souvent insérées au sein d'UFR de médecine, d'hôpitaux, et d'universités.

### 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 située sur le site du Centre universitaire hospitalier de Rangueil. Actuellement, l'Institut est constitué de 14 équipes de recherche, développées autour de trois thèmes :

- Intestins, Tissu Adipeux, Obésité et Diabète ;
- Thrombose, Athérosclérose et Vaisseaux ;
- Coeur et Rein.

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

### 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 voies inflammatoires dans le tissu adipeux.
Les résultats des projets peuvent contribuer :

- au développement d'une nouvelle classification des réponses aux régimes hypocaloriques et aux programmes d'activité physique ;
- à identifier des biomarqueurs pour caractériser les patients qui répondront le mieux à ces stratégies thérapeutiques (théragnostique) ;
- à développer des thérapies combinant une restriction calorique ou un exercice physique et un traitement pharmacologique avec une molécule antilipolytique ou des peptides natriurétiques.

Partie 2

# Quelques notions de biologie

Durant le stage, j'ai travaillé sur des données d'expression génique issues d'une expérience de « Polymerase Chain Reaction » (PCR). Dans ce premier chapitre, je présente quelques notions de biologie, qui permettent de comprendre à quoi servent ces données et comment elles sont obtenues, puis je présenterai le contexte et les données de l'étude sur laquelle j'ai travaillé.

## 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. Le génome est l'ensemble de l'ADN d'un organisme. Ce dernier est étudié par différentes techniques regroupées sous le terme de génomique. La structure standard de l'ADN est une double-hélice, composée de deux brins complémentaires (image 2.1). Chaque brin d'ADN est constitué d'un enchaînement de nucléotides qui sont formés d'un phosphate, d'un sucre (le désoxyribose)... On trouve quatre nucléotides différents dans l'ADN, notés A (Adénine), G (Guanine), C (Cytosine) et T (Thymine), 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.

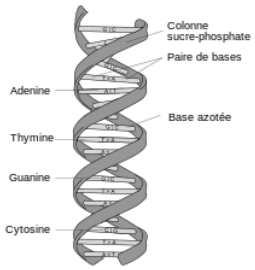



Image 2.1 – Structure d'une molécule d'ADN
Source : l'image provient de tchernobylfrance

L'ADN détermine 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. Cette copie est synthétisée dans le noyau et passe ensuite dans le cytoplasme où elle est traduite en la protéine correspondante par les ribosomes.

La synthèse des protéines se fait en deux étapes (image 2.2) :

1. la transcription, qui est le transfert de l'information génétique de l'ADN vers une autre molécule, l'ARNm ;
2. la traduction, qui est un transfert d'information depuis l'ARNm vers les protéines ;


Image 2.2 – 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.

## 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.
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. Par exemple, cela peut 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 RT-PCR. Les autres méthodes ne seront pas expliquées ici.

## La RT-PCR

La « Polymerase Chain Reaction » ou PCR, est une technique de réplication ciblée in vitro. Elle permet d'obtenir, à partir d'un échantillon complexe et peu abondant, d'importantes quantités d'un fragment d'ADN spécifique et de longueur définie. L'ordre de grandeur à retenir est celui du million de copies en quelques heures. C'est, généralement suffisant pour une utilisation ultérieure. Le principe et les conditions expérimentales qui en découlent sont très simples. Il s'agit de réaliser une succession de réactions de réplication d'une matrice double brin d'ADN. Chaque réaction met en oeuvre deux amorces oligonucléotidiques dont les extrémités 3' pointent l'une vers l'autre. Les amorces ou « primers » en anglais définissent alors, en la bornant, la séquence à amplifier (image 2.3). L'astuce consiste à utiliser les produits de chaque étape de synthèse comme matrices pour les étapes suivantes, au lieu de les séparer afin de ne réutiliser que la matrice originale. Au lieu d'être linéaire, l'amplification obtenue est exponentielle.

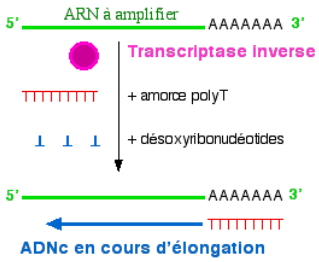



Image 2.3 – Schéma simplifié du principe de la réaction de transcription inverse en présence d'amorce polyT
Source : l'image provient de l'ens-Lyon

L'acronyme RT-PCR signifie Reverse Transcriptase PCR, soit une PCR après transcription inverse d'un acide ribonucléique (ARN) en ADN complémentaire (ADNc). En réalité, il s'agit d'une PCR « classique » réalisée sur un ADN complémentaire (ou ADNc), qui est une copie d'un ARN obtenue par une transcription inverse. L'une des difficultés de cette méthode est liée à la préparation des ARN qui peuvent être très facilement dégradés et contaminés par de l'ADN génomique.

La RT-PCR a été mise au point pour utiliser les ARN comme matrice d'amplification de la PCR. Elle est certainement la méthode la plus sensible pour détecter, les ARN messagers au niveau d'un organe, d'un tissu ou d'une cellule. Elle est également utilisée pour la construction de sondes d'ADN. La synthèse d'ADNc est catalysée par des transcriptases inverses (reverse transcriptase (RT) en anglais). Ces enzymes sont des ADN polymérases, ARN dépendantes, capables d'utiliser un brin d'ARN comme matrice pour catalyser la synthèse du brin d'ADN complémentaire. Cela correspond effectivement à l'« inverse » d'une réaction de transcription de l'ADN en ARN. Les transcriptases inverses sont issues de rétrovirus dont elles sont une des principales caractéristiques. Comme toutes les ADN polymérases, les transcriptases inverses ne peuvent pas initier seule la synthèse d'un brin d'ADN. Elles ont besoin d'une amorce possédant une extrémité 3'-OH libre (extrémité de la séquence polynucléotidique terminée par un ose (ribose ou désoxyribose), et notamment par un groupement hydroxyle OH porté par le carbone 3' du sucre).

Dans un premier temps, la Taq polymérase catalyse la synthèse du second brin d'ADNc en utilisant le premier brin comme matrice (La Taq polymérase est une variété d'ADN polymérases thermostables). Ensuite, la PCR permet d'amplifier le fragment d'ADNc lors d'une série de réactions en chaine jusqu'à 40 cycles de PCR. (image 2.4).

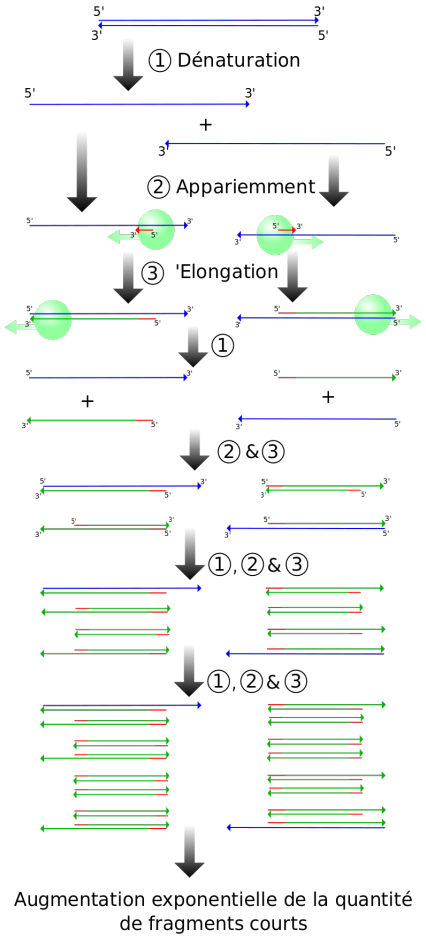



Image 2.4 – Diagramme des quatre premiers cycles de la PCR
Source : Wikipédia

Le produit final est un ADN dont l'un des brins est complémentaire de l'ARN d'intérêt et l'autre brin a la même séquence que cet ARN d'intérêt (à la substitution près de U par T).
Au cours des 40 cycles de temps, on va observer l'augmentation du niveau de fluorescence sur différents échantillons. (image 2.5) Les données vont être basées sur la différence de temps, à une même fluorescence, entre l'échantillon A et un échantillon de référence.

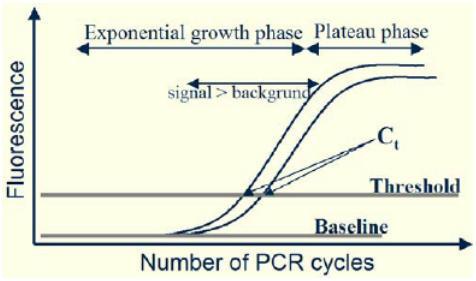



Image 2.5 – Modèle en temps réel d'une PCR quantitative
Source : l'image provient de reasearchgate

- la ligne de base (Baseline) est définie comme des cycles de PCR dans lesquels des signaux fluorescents sont accumulés mais qui restent en dessous des limites de détection de l'instrument ;
- La fluorescence est l'incrémentation des signaux fluorescent à chaque pas de temps. Les valeurs de fluorescence sont dessinées en fonction du nombre de cycles ;
- le seuil (Threshold) est un niveau arbitraire de fluorescence choisi par rapport à la variabilité de la ligne de base. Un signal détecté au dessus du seuil est considéré comme un signal pouvant définir le cycle seuil (CT) pour un échantillon. Le seuil peut être ajusté pour chaque expérience pour qu'il soit dans la région d'amplification exponentielle.
- CT est défini comme le nombre de cycles de PCR pour lesquels la fluorescence est plus grande que le seuil. Le CT est un principe de base de la PCR en temps réel et est un composant essentiel dans la production de données précises et reproductibles. En effet, les données sur lesquelles nous avons travaillé, sont des données d'expression génique sous la forme $2^{\Delta ct}$, avec $\Delta$CT qui correspond à la différence de deux pas de temps, entre un échantillon A et l'échantillon référence, à un niveau de fluorescence donné.

# Contexte du stage

## Contexte biologique

L'obésité est caractérisée par un excès du tissu adipeux (image 2.6) qui s'accompagne à long terme de complications métaboliques et cardio-vasculaires. Lors d'une obésité, la résistance à l'action de l'insuline est courante. L'insuline est une hormone sécrétée par des cellules du pancréas (glande située en arrière de l'estomac). Elle diminue le taux de glucose (sucre) dans le sang et favorise son utilisation par les tissus de l'organisme. On peut observer, chez les personnes obèses ou en excès de poids, une résistance à l'insuline, ce qui entraîne la destruction des cellules bêta du pancréas, celles qui sécrètent l'insuline. Ce phénomène a pour conséquence l'apparition d'un diabète de type II. Pour améliorer ce problème, la perte de poids et la pratique modérée d'exercice physique sont généralement prescrites [Capeau, 2003].

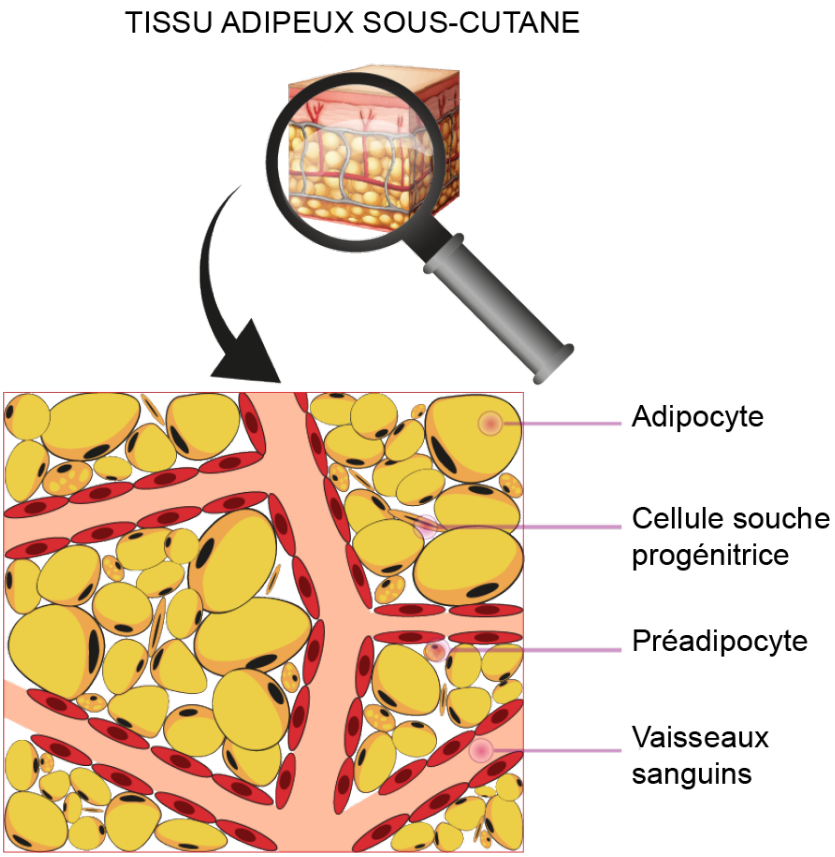



Image 2.6 – Schema simplifié du tissu adipeux sous cutané
Source : l'image provient de Stemcis.com

Le tissu adipeux se compose d'une multitude de cellules différentes : parmi elles, les adipocytes, qui servent à stocker l'énergie sous forme de lipides, ou les adipokines, qui englobent l'ensemble des facteurs protéiques sécrétés par le tissu adipeux, qui agissent soit au sein même du tissu, soit à distance sur d'autres tissus. La leptine est un exemple d'adipokine. La leptine a été la première hormone que l'on a reliée aux problèmes de poids : c'est elle qui informe le cerveau du niveau des réserves énergétiques et sa sécrétion augmente avec le poids. Généralement, plus le tissu adipeux augmente en volume, plus la quantité des adipokines augmente. Toutefois, une adipokine s'est révélée être moins présente chez les individus obèses : cette adipokine est l'adiponectine [Lacquemant et al., 2003].

Dans une étude préalable au stage et concernant le projet décrit dans la section 2.4.2, une autre adipokine, dont le profil d'expression est proche de celui de l'adiponectine, a été trouvée. Cette adipokine, produite par le tissu adipeux, est moins exprimée chez les personnes obèses et encore moins chez les personnes obèses présentant un syndrome de résistance à l'insuline : elle a une expression fortement liée à la sensibilité à l'insuline. Cette adipokine est l'APOM (apolipoprotéineM).
Le but de ce stage était d'étudier l'expression de l'APOM afin de comprendre ses relations avec les autres hormones et protéines du tissu adipeux et avec les variables permettant de caractériser la perte de poids et la sensibilité à l'insuline pendant un régime hypocalorique.

## Description des données

### 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 régime puis de stabilisation. 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.

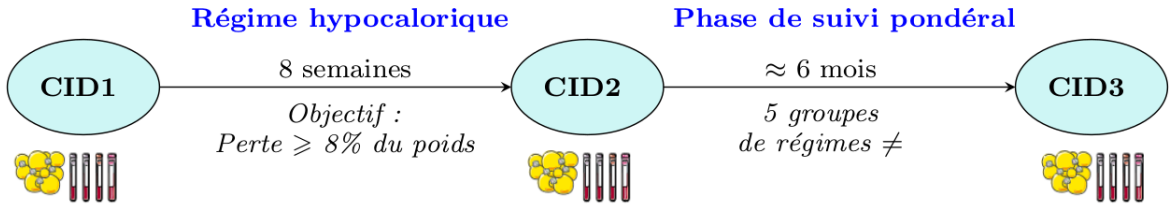



Image 2.7 – Protocole du projet DiOGenes

Une expérience a été réalisée dans 8 centres européens avec 632 patients obèses. Cette expérience se divise en 3 étapes (image 2.7) :

- la première consiste à faire une analyse clinique non exhaustive mais détaillée. Au début de l'étude, les patients ont été pesés, mesurés, des enquêtes diététiques et des questionnaires sur l'activité physique ont été réalisés. On réalise des prélèvements sanguins et urinaires, une biopsie du tissu adipeux sous cutané abdominal (pour quantifier l'expression des gènes à l'aide de RT-qPCR), on analyse la composition corporelle à l'aide de mesures de la matière grasse (« fatmass », « fat-free mass ») et on réalise un test de tolérance au glucose (« oral glucose tolerance test (OGTT) ») pour mesurer la sensibilité à l'insuline. Pour mesurer cette sensibilité, on va également calculer des indices de sensibilité à l'insuline tels que HOMA-IR, QUICKi et Matsuda (avec les données de OGTT). Cette étape sera notée CID1 (Clinical Intervention Day 1) dans la suite de ce rapport ;

- Dans une deuxième phase de l'étude, un régime hypocalorique fort est imposé aux participants. Ce régime dure 2 mois. L'objectif est de faire perdre aux sujets 8% de leur poids initial. Les sujets tiennent un carnet alimentaire où ils indiquent (nature et quantité) tout ce qu'ils mangent. La fin de la deuxième phase se conclut par une deuxième analyse clinique détaillée. Cette phase de restriction sera notée CID2 dans la suite de ce rapport ; Tous les individus ayant perdu au moins 8% de leur poids initial (10-11kg en moyenne) poursuivent l'étude.

- Lors de la troisième étape, des recommandations alimentaires sont proposées aux sujets dans une phase dite de « maintenance pondérale » où les patients ont reçu une éducation diététique et mangent autant qu'ils le veulent. De manière plus précise, 5 types de recommandations alimentaires ont été proposés (régime témoin dit « health diet », régime à faible teneur en protéines et faible teneur en glycémie, régime à faible teneur en protéines et forte teneur en glycémie, régime à forte teneur en protéines et faible teneur en glycémie, régime à forte teneur en protéines et forte teneur en glycémie). Le patient se voit proposer, de manière aléatoire, un de ces 5 types de régime. Le but de cette phase est de comprendre l'impact du type d'alimentation sur le contrôle pondéral. À la fin de cette dernière phrase d'une durée de 6 mois, une troisième analyse clinique détaillée est effectuée. Cette étape (qui marque la fin de la seconde phase) sera notée CID3 dans la suite de ce rapport.

Nous disposons d'une vaste gamme d'indicateurs biologiques (données cliniques ou d'expression des gènes dans le tissu adipeux) chez les individus en surpoids ou obèses avant et après régime hypocalorique.

**Les données étudiées**

À la suite de ces expériences, on dispose de trois jeux de données différents fournis par l'INSERM :

- le premier se compose de 1062 observations et 30 variables. Une observation correspond à un patient lors d'une des trois étapes (CID1, CID2 ou CID3) : le jeu de données contient donc les mesures de 354 individus distincts. 2 des 30 variables sont qualitatives et les autres quantitatives. Les 2 variables qualitatives sont l'ID et l'étape de l'expérience. Les variables quantitatives sont les expressions de différents gènes correspondant à des protéines identifiées et exprimées en ($\Delta$CT). En particulier, on y retrouve les expressions de l'adiponectine (ADIPOQ) et de l'APOM ;

- le second se compose de 978 observations et 85 variables. Une observation correspond à un patient lors d'une des trois étapes, donc on a 326 individus distincts dans le jeu de données. 2 des 30 variables sont qualitatives et les autres quantitatives. Les 2 variables qualitatives sont l'ID et l'étape de l'expérience. Les variables quantitatives sont également les expressions de gènes correspondant à différentes protéines mais ces protéines sont ici majoritairement des marqueurs adipocytes bruns qui dissipent l'énergie sous forme de chaleur. En particulier, on y retrouve les expressions de l'adiponectine (ADIPOQ) et la Leptine (LEP). On s'intéresse aux adipocytes bruns car la présence d'adipocytes bruns dans le tissu adipeux est associé à un profil métabollique sain comme une sensibilité à l'insuline qui s'améliore ;

- le troisième jeu de données correspond aux données cliniques récoltées lors des trois étapes.Il y a 632 observations (une observation correspond à un patient) et 127 variables. En particulier, le centre dans lequel le sujet a participé à l'expérience est indiqué.

Partie 3

# Analyse exploratoire des données

## Objectif de l'analyse exploratoire

L'objectif de cette partie est d'obtenir des données épurées, c'est-à-dire, des données dans lesquelles les biais expérimentaux ont été corrigés, les valeurs atypiques repérées et corrigées ou supprimées, les valeurs manquantes imputées, etc... À l'issue de cette phase, les données obtenues sont directement utilisables et interprétables pour les analyses statistiques réalisées pour répondre à la question biologique.
Dans cette phase, nous avons :

- uniformisé les unités de mesure des différents échantillons (transformation en log2 sur certains échantillons car les données avaient été fournies sous la forme « $\Delta$CT ») ;
- analysé et nettoyé les valeurs manquantes ;
- effectué une analyse exploratoire des données (ACP, analyse des distributions...) pour vérifier si il n'y a pas d'individus aberrants et identifier les biais expérimentaux ;
- normalisé les données pour supprimer les biais expérimentaux.

## Outils et méthodes

Dans cette partie, on notera $X_i$ le vecteur des observations de $p$ variables $(x_{ij})_{j=1,\dots,p}$ pour l'individu $i$, $i \in \{1, \dots, n\}$. On notera $X$ la matrice de l'ensemble des observations $(X_i)_{i=1,\dots,n}$. En outre, lorsque la variable $j$ contient des valeurs manquantes, on notera $X_{\text{obs},j}$ le vecteur des individus pour lesquels la variable est observée et $X_{\text{mis},j}$ le vecteur des individus pour lesquels la variable n'est pas observée.
Enfin, rappelons que les variables sont, dans notre cas d'étude, des expressions de gène ou bien des mesures cliniques.

### Imputation des données manquantes

Les jeux de données biologiques contiennent de nombreuses données manquantes. Supprimer tous les individus contenant des valeurs manquantes conduit donc à une perte d'information très importante et dommageable pour la qualité des analyses statistiques effectuées. Nous avons donc effectué une *imputation* des valeurs manquantes, c'est-à-dire, nous avons remplacé les valeurs manquantes par une valeur « probable ». De nombreuses méthodes d'imputation existent : nous décrivons et justifions dans cette section les choix effectués dans notre étude. Dans toute la suite, nous supposerons que les données sont MAR (« Missing At Random »), ce qui signifie que la probabilité d'avoir une valeur manquante ne dépend que des valeurs observées.

Les données manquantes dans les expressions de gènes correspondent à des données pour lesquelles le signal n'a pas dépassé le seuil de détection de la méthode. Pour ces données, on a donc choisi une méthode d'imputation simple, l'imputation par le minimum (toutes les valeurs manquantes ont été imputées à la valeur minimum observée dans les expressions de gènes).
Dans les données de mesures cliniques, les données manquantes ont des origines diverses (perte d'information, échec de la mesure, absence du sujet aux tests, etc). Nous avons choisi une méthode d'imputation plus sophistiquées, utilisant les forêts aléatoires [Breiman, 2001]. La méthode utilisée est celle du package **missForest** de R [Stekhoven and Bühlmann, 2012]. La méthode se déroule de cette manière :

1. La méthode est initialisée par une imputation « naïve » des valeurs manquantes (imputation par la moyenne des valeurs observées de la variable, par exemple) ;
2. Soit alors $k$ le vecteur des indices de colonnes de $X$ triées par quantité croissante de valeurs manquantes ;
3. Tant que $\Delta N$ diminue
   (a) $X^{\text{old}} \leftarrow X$ (matrice précédemment imputée)
   (b) Pour $s$ dans $k$ faire
      i. Estimer la régression de $x_{\text{obs},s}$ sur $X_{\text{obs},-s}$, l'ensemble des variables différentes de la variable $s$ pour les individus observés pour la variable $s$, par forêt aléatoire
      ii. Prédire $x_{\text{mis},s}$ avec $X_{\text{mis},-s}$, l'ensemble des variables différentes de la variables $s$ pour les individus manquants pour la variable $s$
      iii. Mettre à jour $X$ par imputation par les valeurs prédites
   (c) Mettre à jour le critère d'arrêt

$$\Delta N = \frac{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - x_{ij}^{\text{old}})^2}{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij})^2}.$$

### Analyse exploratoire des données

Pour les données d'expression de gènes, les vérifications de qualités ont consisté à vérifier que tous les individus possèdent 3 observations (si il y a bien une observation pour chaque étape de l'expérience), et si les distributions des variables sont symétriques (en utilisant des boîtes à moustaches). Lorsque des distributions de variables présentaient une forte asymétrie, on a supprimé ces variables pour éviter qu'elles influent sur une mauvaise interprétation des résultats.
Pour vérifier si il y a des individus aberrants, une ACP a été réalisée sur laquelle différentes informations additionnelles (liées au plan d'expérience) ont été représentées. L'ACP a été réalisée avec le package **FactoMineR**.

### Normalisation des données

Afin de normaliser les données et donc de supprimer les éventuels biais expérimentaux, on utilise la fonction ComBat du package sva de bioconductor ([Johnson et al., 2007]) On appelle biais expérimentaux tous les paramètres fluctuants non mesurés qui peuvent avoir un impact sur les mesures obtenues lors de l'expérience. De manière simplifiée, la méthode estime un modèle linéaire dans lequel les expressions des gènes, $X$, sont expliquées par des variables correspondant au plan d'expérience (par exemple, le centre de prélèvement des données) qui peuvent influencer leur niveau. L'estimation du modèle est basé sur une approche bayésienne permettant de travailler avec des échantillons de taille plus modérée. Après retrait des effets des variables du plan d'expérience, la mesure restante est alors utilisée comme données d'expression corrigées.

## Résultats

Comme je l'ai précisé précédemment, l'APOM est moins exprimée chez les individus obèses. Après la première phase, les individus ont perdu 8% de leur poids entre CID1 et CID2. Si on se réfère à cette information et si les données sont bien distribuées, on doit observer une augmentation des valeurs de l'APOM dans les données d'expressions génique entre CID1 et CID2. On observe bien une légère augmentation globale des valeurs de l'APOM entre CID1 et CID2 sur l'image 3.1.

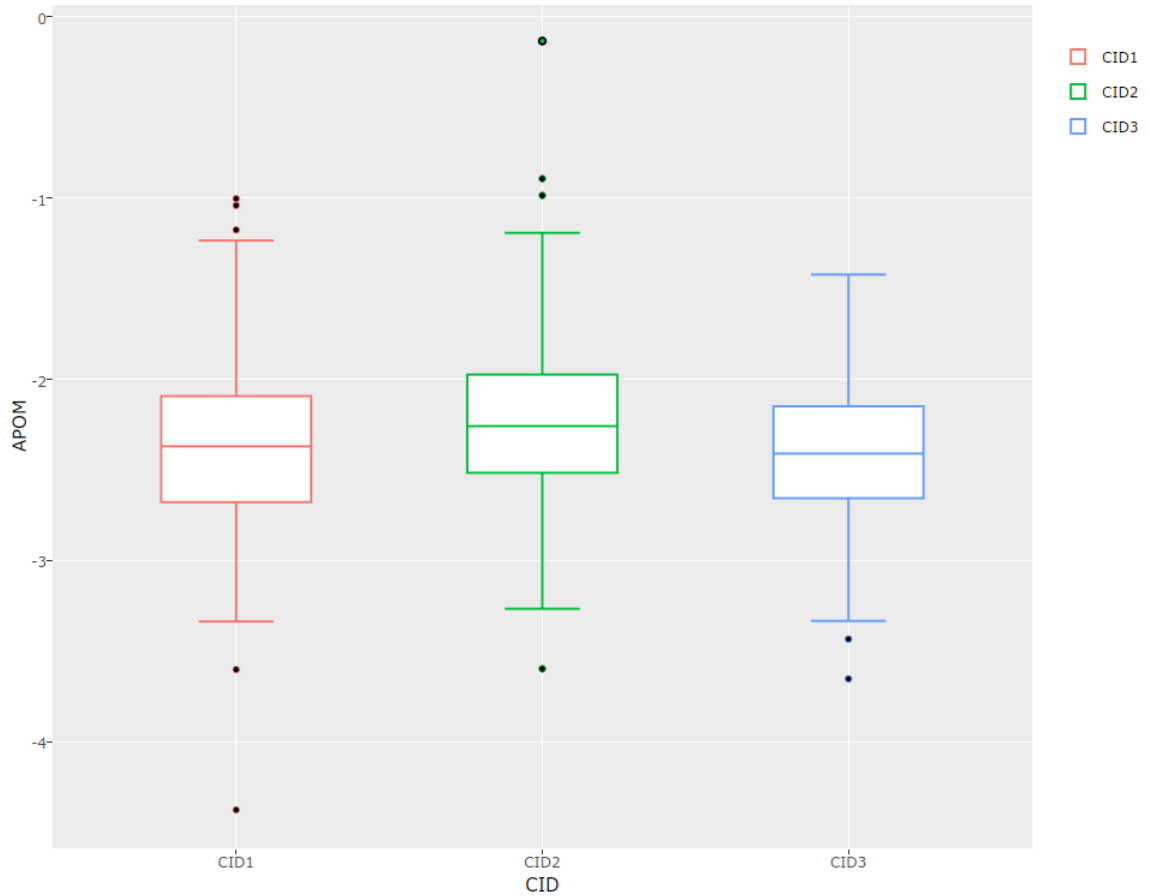



Image 3.1 – Distributions des données d'expression génique de l'APOM en fonction des différentes analyses cliniques

Je me suis ensuite intéressé aux individus qui pourraient être aberrants si il y en a. Pour cela j'ai réalisé différentes ACP sur les différents jeux de données, dont notamment une ACP sur les données d'expression génique en représentant les individus dans la projection avec une couleur correspondant à leur centre. (image 3.2)

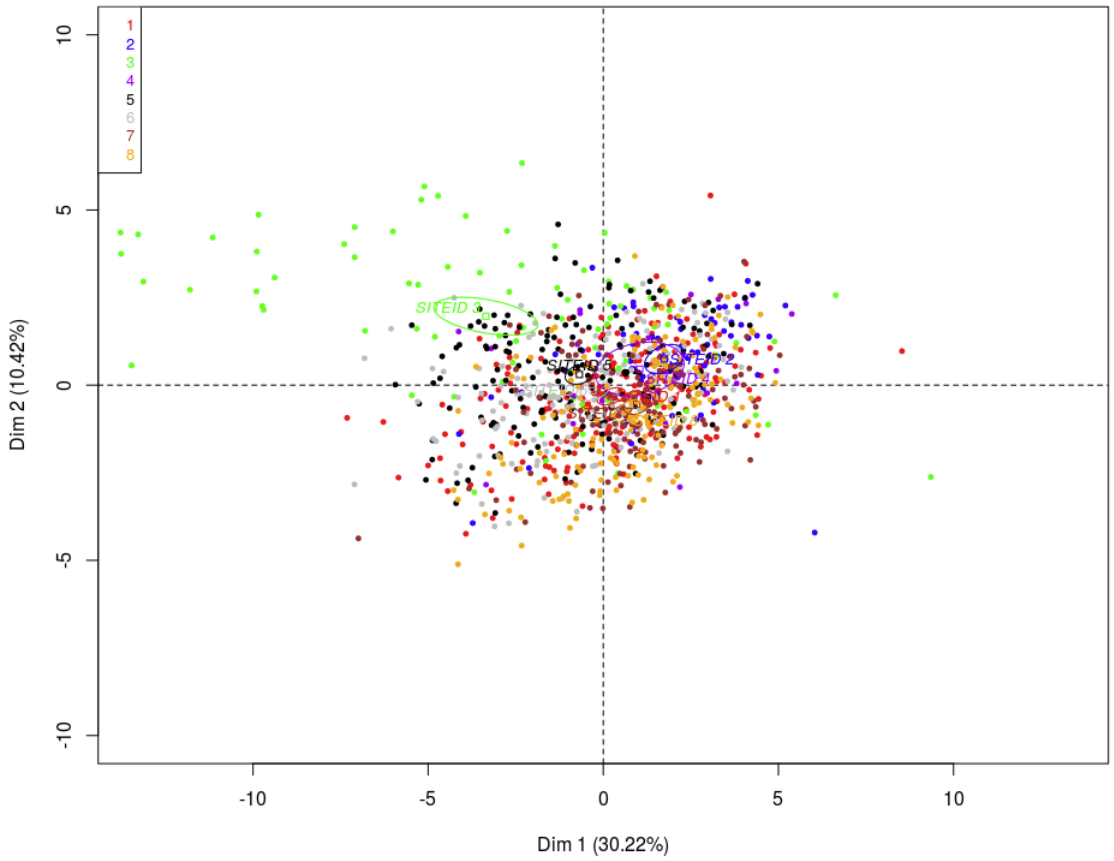



Image 3.2 – ACP du premier jeu de données (données d'expression génique). Les couleurs correspondent aux centres d'études dans lesquels les études ont été réalisées.

On constate par exemple sur l'image 3.2 que le nuage de points est très concentré au centre de gravité. On remarque que les points du troisième centre sont assez éloignés du centre de gravité global.

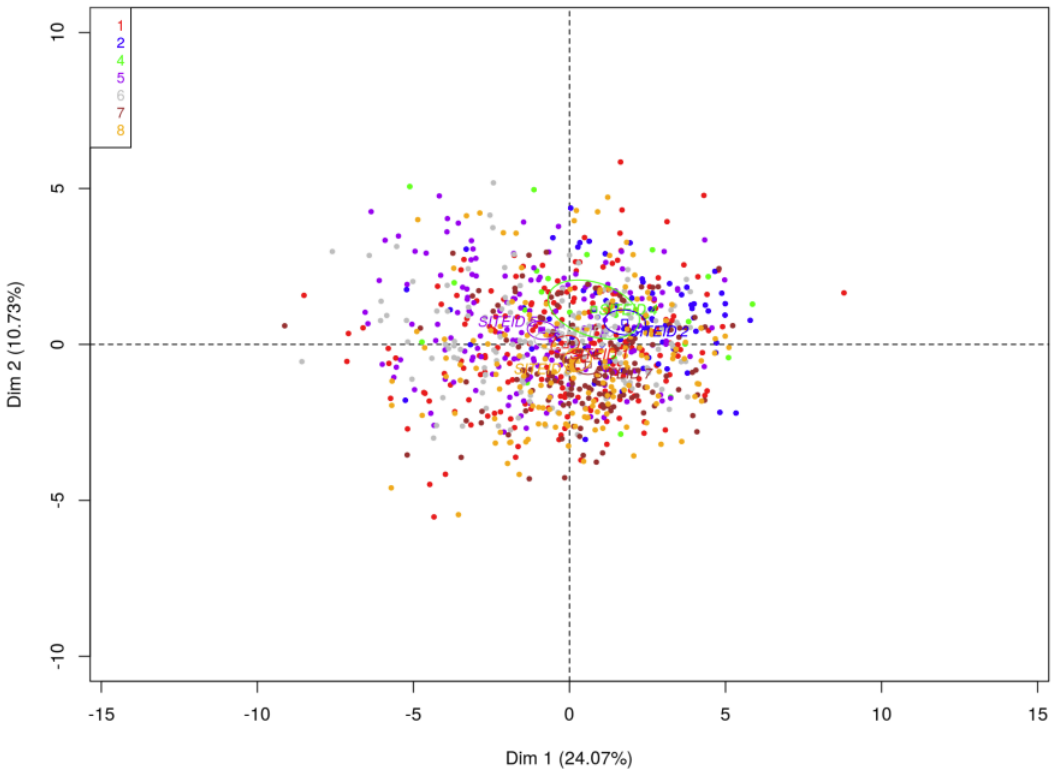



Image 3.3 – ACP du premier jeu de données (données d'expression génique) sans les données du troisième centre. Les couleurs correspondent aux centres d'étude.

Lorsque l'on supprime les données du troisième centre, la nouvelle ACP (image 3.3) montre que le nuage de points est plus éclaté (les distances entre les points sont plus importantes), ce qui montre que la suppression des individus du centre 3 a permis d'avoir des données plus homogènes et de retirer des individus qui pouvaient être considérés comme atypiques. Les centres de gravités de chaque centre sont proches les uns des autres et proches du centre de gravité global.

## Conclusion

L'objectif principal de cette partie était d'obtenir des données épurées et interprétables afin de réaliser les différentes analyses qui me permettrait de répondre à la problématique de ce stage. Après de nombreuses analyses et manipulations, notamment la suppression d'individus atypiques où l'imputation des données manquantes. Les données sont dites propres. On va pour voir commencer les analyses afin de déterminer quels gènes sont corrélés avec le gène d'intérêt qui est l'APOM.

Partie 4

# Corrélation de l'expression génique avec la protéine d'intérêt

## Objectif de cette partie

En probabilités et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques numériques, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. Le fait que deux variables soient « fortement corrélées » ne démontre pas qu'il y ait une relation de causalité entre l'une et l'autre, mais qu'il y a un lien fort entre elles.
Dans cette phase, nous avons :

- calculé les corrélations de Pearson et les valeurs du $\tau$ de Kendal entre chaque paire de variable ;
- effectué des tests de corrélations sur les différents jeux de données.
- réalisé des classifications de variables ;

## Outils et méthodes

Dans cette partie, on notera $X$ la matrice des $n$ observations de $p$ variables. Le vecteur des observations de la variable $j$ sera noté $X^j$ et le vecteur des observations de l'individu $i$ sera noté $X_i$. L'observation de la variable $j$ pour l'individu $i$ sera noté $x_{ij}$.

### Corrélations de Pearson et $\tau$ de Kendall

La corrélation de Pearson permet de détecter la présence ou l'absence d'une relation de corrélation linéaire entre deux caractères quantitatifs continus $X^j$ et $X^{j'}$.
Le coefficient de corrélation linéaire entre ces deux caractères est égal à :

$$r(X^j, X^{j'}) = \frac{COV(X^j, X^{j'})}{\sigma_{X^j} * \sigma_{X^{j'}}}$$

Le signe de $r$ indique le sens de la relation tandis que la valeur absolue de $r$ indique l'intensité de la relation.
Une alternative à la corrélation de Pearson est le $\tau$ de Kendall. Le $\tau$ de Kendall est une statistique qui mesure la corrélation de rangs entre deux variables.

On suppose, pour simplifier, que les valeurs $(x_{ij})_i$ sont distinctes pour tout $j$. Les paires d'observations $(x_{ij}, x_{ij'})$ et $(x_{i'j}, x_{i'j'})$ sont dites concordantes si $x_{ij} < x_{i'j}$ et $x_{ij'} < x_{i'j'}$ ou si $x_{ij} > x_{i'j}$ et $x_{ij'} > x_{i'j'}$. Elles sont dites discordantes si $x_{ij} < x_{i'j}$ et $x_{ij'} > x_{i'j'}$ ou si $x_{ij} > x_{i'j}$ et $x_{ij'} < x_{i'j'}$. Dans le cas où $x_{ij} = x_{i'j}$ ou $x_{ij'} = x_{i'j'}$, la paire n'est ni concordante ni discordante. Le $\tau$ de Kendall est alors défini comme :

$$\tau = \frac{\text{(nombre de paires concordantes)} - \text{(nombre de paires discordantes)}}{\frac{1}{2}n(n-1)}$$

### Classification des variables

On réalise une classification des variables à l'aide du package **ClustOfVar**, et plus précisément de la fonction `hclustvar` qui permet de construire la hiérarchie. Le critère d'agrégation est la baisse d'homogénéité pour les classes fusionnées où l'homogénéité d'une classe est définie comme la corrélation (au carré) totale entre les variables de la classe et une variable synthétique qui la représente. Cette variable synthétique est définie comme étant la première composante principale de l'ACP car c'est elle qui maximise la corrélation totale avec toutes les variables d'une classe donnée.
De manière plus précise, chaque classe $C_k$ est résumée par une variable synthétique :

$$c_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{j \in C_k} r^2(X^j, u)$$

où $c_k$ est aussi la première composante principale de l'ACP de $X$ restreinte aux variables de $C_k$.
La classification ascendante hiérarchique des variables est donc effectuée de la manière suivante :

1. la classification est initialisée à $p$ classes, une classe par variable. Les variables synthétiques de chaque classe sont déterminées (elles sont initialement égales aux variables elles-mêmes) ;

2. les deux classes de plus faible dissimilarité, $C_k$ et $C_{k'}$ sont agrégées où la dissimilarité est définie par

$$d(C_k, C_{k'}) = H(C_k) + H(C_{k'}) - H(C_k \cup C_{k'})$$

   avec $H(C_k)$ le critère d'homogénéité égal à $\sum_{j \in C_k} r^2(X^j, c_k)$ qui est aussi égal à la première valeur propre de l'ACP restreinte aux variables de $C_k$ ;

3. les variables synthétiques et les critères d'homogénéité sont recalculés ;

4. le processus s'arrête lorsque toutes les classes ont été fusionnées en une seule classe.

### Tests multiples

Il est courant en biostatistique de répéter un test donné sur un grand nombre de gènes (ici, on effectue des tests de significativité de la corrélation de Pearson entre le gène d'intérêt et tous les autres gènes). Pour un risque donné $\alpha$ et lorsque $p$ tests indépendants sont effectués, la probabilité d'avoir au moins un faux positif, sous l'hypothèse que tous les gènes satisfont l'hypothèse nulle, est égal à

$$1 - (1 - \alpha)^p.$$

Cette quantité converge vers 1 de manière rapide lorsque $p$ (le nombre de tests effectués) est grand.

Pour pallier ce problème, des procédures de correction des tests multiples ont été proposées. Elle contrôle, soit le FWER (Family Wise Error Rate, c'est-à-dire la probabilité d'avoir au moins un faux positifs parmi les $p$ tests), ou bien le FDR (False Discovery Rate, c'est-à-dire l'espérance de la proportion de faux positifs parmi les $p$ tests). Une des procédures les plus fréquemment employées pour contrôler le FDR est la méthode de Benjamini et Hochberg [Benjamini and Hochberg, 1995]. Dans cette approche, les p-valeurs, $u_1, \ldots, u_p$ sont rangées par ordre croissant, $u_{(1)}, \ldots, u_{(p)}$ et corrigées par :

$$\tilde{u}_{(j)}^{\mathrm{BH}} = \min\left\{u_{(j)} \times \frac{p}{j}, 1\right\}.$$

## Résultats

On réalise une matrice de corrélation de Pearson (matrice $C$ où à chaque croisement d'une ligne et d'une colonne on retrouve la valeur du coefficient de corrélation entre une variable $j$ et une autre variable $j'$) pour connaître le degré de linéarité entre les variables (les gènes) et notamment le lien APOM-ADIPOQ qui ont un profil d'expression similaire (image 4.1).

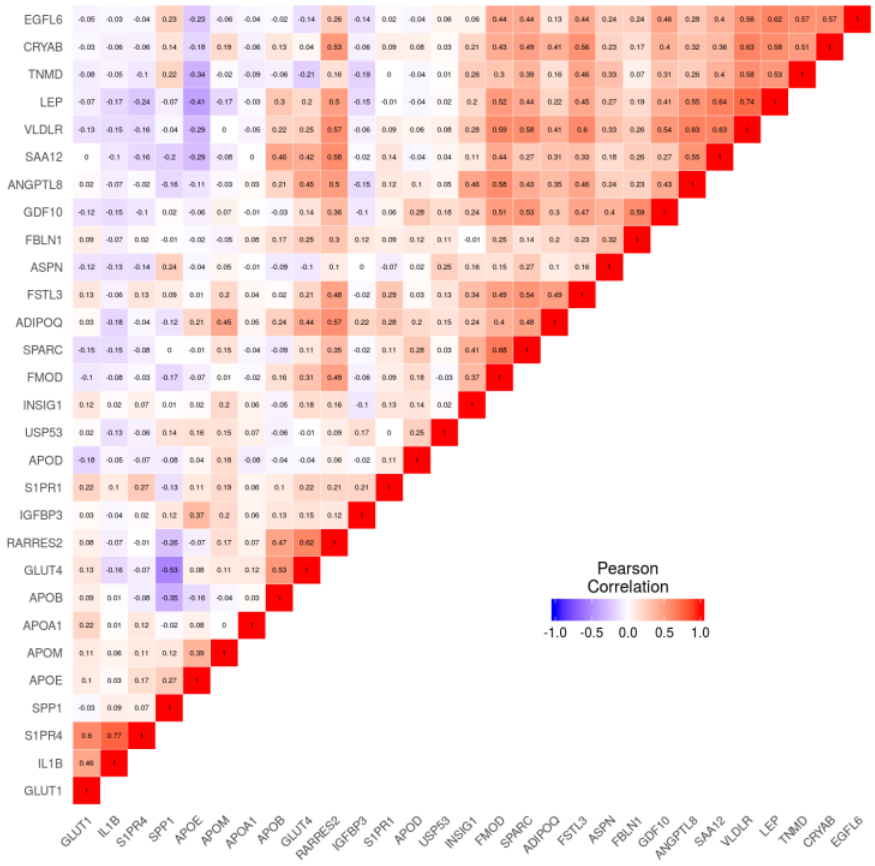



Image 4.1 – Matrice des corrélations d'expression génique

La plus forte corrélation est de 0,77 entre les gènes Interleukin 1 beta (IL1B) et Sphingosine-1-phosphate receptor 4 (S1PR4). Le taux de corrélation négatif le plus élevé est entre Glucose transporter type 4 (GLUT4) et secreted phosphoprotein 1 (SPP1) (-0,53). La variable APOM a une bonne corrélation avec ADIPOQ (0,45) et l'apolipoprotéineE (APOE) (0,39). La corrélation négative la plus forte de l'APOM est avec la leptine (LEP) (-0,17).

On obtient des liaisons équivalentes quand on utilise le $\tau$ de Kendall pour calculer ses liaisons. Les valeurs de $\tau$ sont cependant inférieures aux valeurs des corrélations de Pearson.
Afin d'approfondir l'analyse des liaisons, on réalise une classification ascendante hiérarchique sur les variables (image 4.2). Pour cela, on utilise le package **ClustOfVar**.

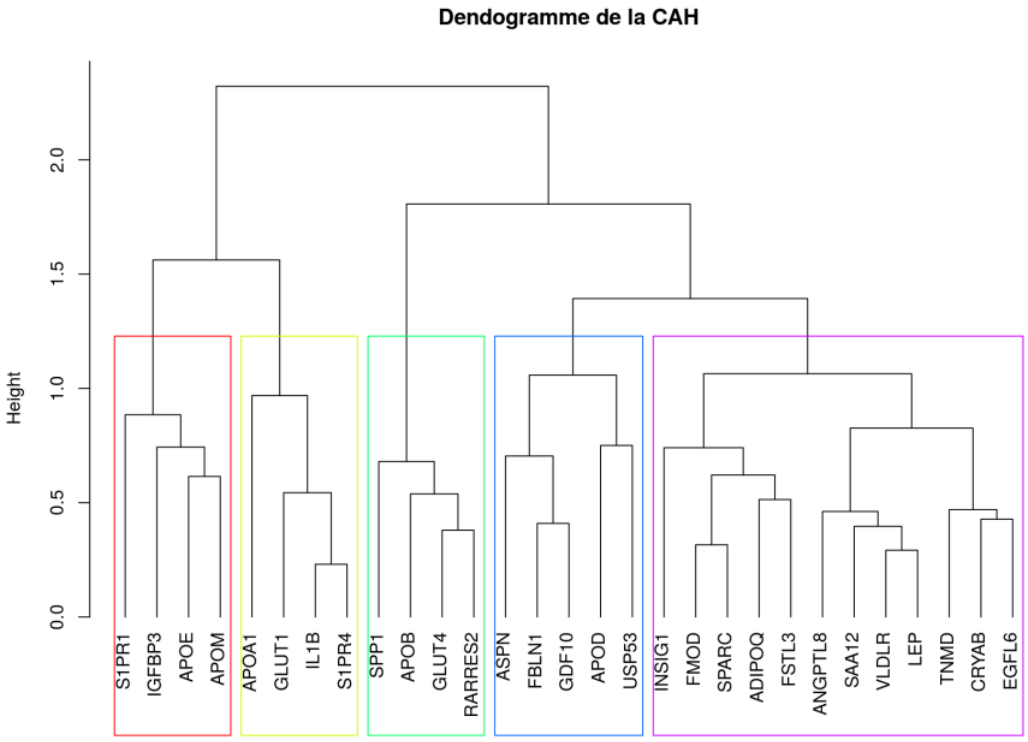



Image 4.2 – Classification ascendante hiérarchique sur les variables

Le gène le plus proche de APOM est APOE. ADIPOQ est plus éloigné car il a un coefficient de corrélation plus fort avec Follistatine-like 3 (FSTL3) (0.49) qu'avec APOM. IL1B et S1PR4 sont les variables les plus proches en termes de distance comme l'on devait s'y attendre. Il y a 12 variables dans la classe 1 dont ADIPOQ et LEP qui est proche de Very Low Density Lipoprotein Receptor (VLDLR), 4 dans la classe 2 dont APOM et APOE, 4 dans la classe 3 et 5, et 5 dans la classe 4.

Nous avons également étudié quelles variables ont une corrélation de Pearson significative (au seuil de 5%) avec l'APOM à chaque étape de l'expérience (jeu de données complet sans distinction de CID, valeurs de l'APOM uniquement à CID1, uniquement à CID2 et uniquement à CID3). Pour cela, nous avons calculé les coefficients de Pearson et ajusté les p-valeurs par la procédure de Benjaimi-Hochberg. Les relations entre gènes détectés significativement corrélés à l'APOM par les tests de significativité des corrélations de Pearson sont présentés dans le diagramme de venn (image 4.3).

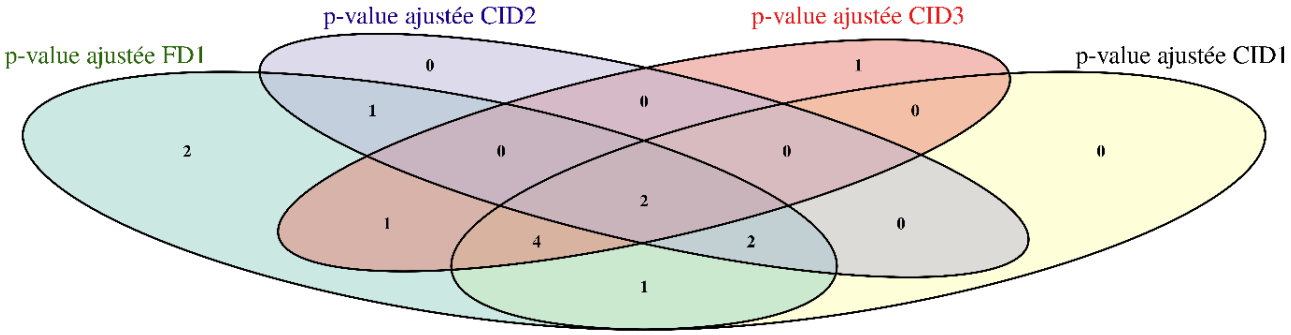



Image 4.3 – Diagramme de Venn présentant les tests de significativité des corrélations de Pearson aux différentes étapes

Sur la figure 4.3, on constate qu'au croisement de chaque cercle, représentant les tests de significativité des corrélations de Pearson à différentes étapes de l'expérience, on retrouve 2 variables. Cela signifie, que ces 2 variables ont leur coefficient de corrélation avec l'APOM significatif tout au long de l'expérience, ou du moins à chaque analyse clinique. Ces 2 variables sont ADIPOQ et APOE.
Nous avons, par la suite, étudié les relations entre le niveau de l'APOM et la variable clinique d'intérêt. Le quicki est un indice permettant de détecter l'insulino-résistance : nous avons effectué un test de comparaison de moyennes (Student) entre individus extrêmes pour les valeurs du quicki. Ici les deux échantillons sont constitués des valeurs de l'APOM pour les individus ayant un quicki inférieur à 1 (les 20 valeurs les plus faibles) et pour les individus ayant un quicki supérieur à 1 (les 20 valeurs les plus élevés) et ce à chaque étape de l'expérience.

Le graphique 4.4 montre la distribution des deux échantillons avec en rouge les individus insulinos résistant (quicki<1) et en bleu les individus non insulinos résistants (quicki>1) à l'étape 3 de l'expérience (CID3).

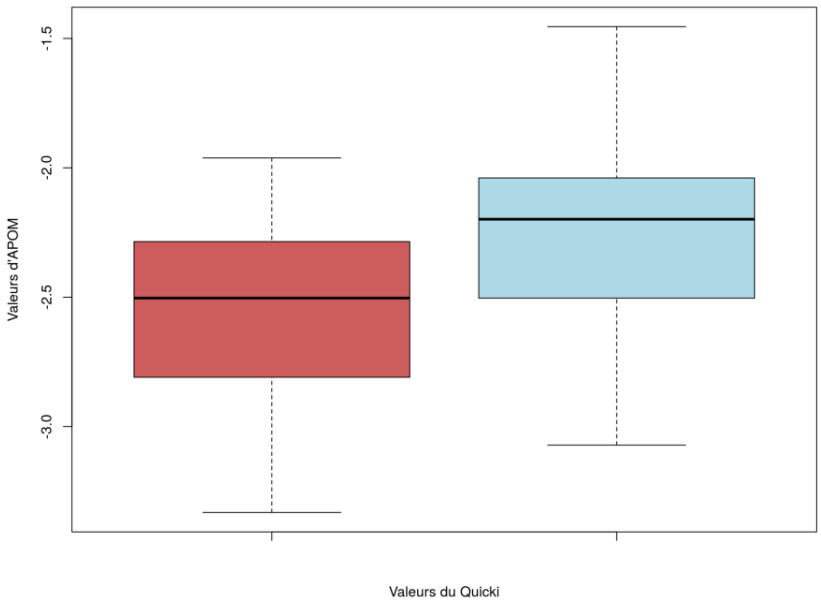



Image 4.4 – Boxplot des valeurs de l'APOM en fonction de l'insulino-résistance des individus à CID3 (échantillons des 20 valeurs extrmes)

La statistique $t$ de Student vaut -2,3984 et la p-valeur est de 0,02122. Donc à un seuil de 5% le test est significatif. Donc le quicki discrimine les valeurs de l'APOM à CID3. Le test de normalité de Shapiro-Wilks fournit une statistique égale à W = 0,95732 pour les individus insulinos-résistants (p-valeur = 0,4639) et à W = 0,9805 pour les individus non insulinos-résistants (p-valeur = 0,9323). Les p-valeurs étant supérieures à 0,05, on ne rejette pas l'hypothèse de normalité. Le test d'égalité des variances de Fisher-Snedecor est de F = 1,0033 et la p-valeur du test est de 0,9943. On remarque que la probabilité critique est supérieur à 0,05. Cela signifie que les variances entre les deux échanitllons ne sont pas significativement différentes.
En ce qui concerne les données à CID1, la statistique $t$ de Student vaut -0,18357 et la p-valeur est de 0,8553. Donc à un seuil de 5% le test n'est pas significatif. Ici W = 0,9374 pour les individus insulinos-résistants (p-valeur = 0,1935) et W = 0,96818 pour les individus non insulinos-résistants (p-valeur = 0,6925). Les p-valeurs étant supérieures à 0,05, on ne rejette pas l'hypothèse de normalité. Ici F = 1,532 et la p-valeur du test est de 0,348. On remarque que la probabilité critique est supérieur à 0,05. Cela signifie que les variances entre les deux échanitllons ne sont pas significativement différentes.
Enfin, à CID2, la statistique $t$ de Student vaut 0,013868 et la p-valeur est de 0,989. Donc à un seuil de 5% le test n'est pas significatif. Le test de normalité de Shapiro-Wilks est de W = 0,90092 pour les individus insulinos-résistants (p-valeur = 0,03646) et de W = 0,96832 pour les individus non insulinos-résistants (p-valeur = 0,6956). On rejette l'hypothèse de normalité pour les individus insulinos-résistants. Le test d'égalité des variances de Fisher-Snedecor est de F = 2,682 et la p-valeur du test est de 0,03257. On remarque que la probabilité critique est inférieur à 0,05. Cela signifie que les variances entre les deux échanitllons sont significativement différentes.

Les individus participant à ce programme ont une caractéristique commune, ils sont obèses. On pouvait donc s'attendre à ce que le quicki ne discrimine pas les valeurs de l'APOM à CID1. À CID2, les individus suivent tous le même régime drastique donc leur poids diminue et parallèllement l'expression de l'APOM augmente, ce qui explique pourquoi le quicki ne discrimine pas les valeurs de l'APOM à cette étape. En revanche, à CID3, les régimes diffèrent selon les individus et le suivi n'est pas le même, ce qui explique le fait que le quicki discrimine les valeurs de l'APOM à CID3. Nous avons vu, dans la partie contexte de stage, que l'APOM était moins exprimée chez les individus obèses et que ces derniers pouvaient être plus insulino résistants que des individus non obèses. Nous retrouvons cette référence sur l'image 4.4 où les valeurs de l'APOM pour les individus insulino résistants sont globalement plus faible que les valeurs de l'APOM pour les individus non insulino résistants.

## Conclusion

Dans cette partie, on a cherché à savoir quelles étaient les liaisons entre variables, et avec quelles variables l'APOM a de plus fortes liaisons. On a découvert que l'APOM avait des liaisons plus importantes avec ADIPOQ et APOE et que ces liaisons sont significatives au cours du temps. Cette référence était attendue car on savait déjà préalablement que les valeurs de ADIPOQ et que les valeurs d'APOM augmentaient avec la perte de poids. Nous nous sommes également demandé si la moyenne des valeurs de l'APOM pour les individus insulino résistants est significativement différente de celle pour les individus non insulino résistants. Nous observons que le quicki discrimine les valeurs de l'APOM à CID3 mais pas à CID1 et CID2, ce qui pourrait être expliqué par le fait qu'entre CID2 et CID3, tous les individus ne suivent pas le même programme et mangent selon leurs envies.

Partie 5

# Réalisation de différents modèles pour exprimer les relations du gène d'intérêt avec les gènes des autres protéines

## Objectif de cette partie

L'objectif de cette partie est de connaître la relation linéaire de l'APOM avec les autres variables. C'est-à-dire savoir si on peut construire un modèle prédictif de l'APOM et qu'on se donne comme objectif qu'il y ait peu de variables dans ce modèle. On va regarder quelles variables expliquent le mieux l'APOM dans le jeu de données d'expression génique dans un premier temps, puis dans le jeu de données des données cliniques ensuite.

## Outils et méthodes

Pour savoir quelles variables expliquent le mieux l'APOM, on va réaliser différentes régressions linéaires multiples avec sélection de variables de type pas à pas avec pénalisation BIC. BIC est un critère pour la sélection de modèles parmi un ensemble fini de modèles. Le modèle avec le BIC le plus bas est préféré. Il repose en partie sur la fonction de vraisemblance.
La fonction de vraisemblance, notée $L(X_1, \ldots, X_n|\theta_1, \ldots, \theta_k)$ est une fonction de probabilités conditionnelles qui décrit les valeurs $X_i$ d'une loi statistique en fonction des paramètres $\theta_j$ supposés connus. Si les $(X_i)_i$ sont indépendantes et identiquement distribuées, elle s'exprime à partir de la fonction de densité $f(X|\theta)$ par

$$L(X_1, \ldots, X_n|\theta) = \prod_{i=1}^{n} f(X_{ij}|\theta)$$

avec $f(X;\theta)$ la loi de X selon le vecteur de paramètres $\theta = \theta_1 \ldots \theta_k$
Cette formule n'est valable que si on suppose que les $x_i$ sont indépendants et identiquement distribués entre eux.

Le critère d'information bayésien BIC est défini par :

$$BIC = -2\ln(L) + k\ln(n)$$

avec $L$ la vraisemblance du modèle estimée, $n$ le nombre d'observations dans l'échantillon et $k$ le nombre de paramètres libres du modèle.
Un ensemble fini de modèles est construit par une approche pas à pas. À chaque étape de cette approche, une variable est ajoutée ou enlevée pour améliorer la qualité de prédiction.
Enfin on utilisera le coefficient de détermination $R^2$ pour connaître la qualité de la prédiction de la régression linéaire. Il est défini comme le rapport entre la variance expliquée et la variance totale :

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

avec $x_i$ les valeurs des mesures, $\hat{x_i}$ les valeurs prédites et $\bar{x}$ la moyenne des mesures.

# Résultats

## Données d'expression génique

On réalise une régression linéaire multiple sur les données d'expression génique dans un premier temps. La figure 5.1 montre le nuage de points des valeurs prédites de cette régression en fonction des valeurs réelles de l'APOM.

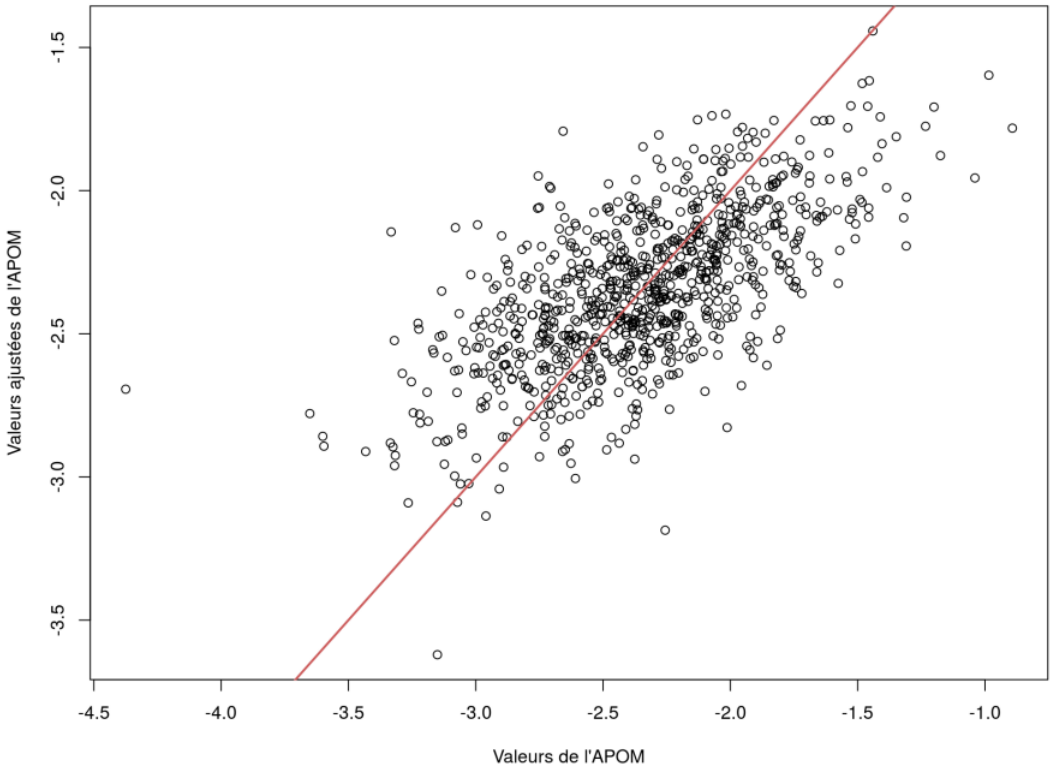



Image 5.1 – Nuage de points des valeurs ajustées de l'APOM par la régression linéaire en fonction des valeurs réelles de l'APOM sur les données d'expression génique

On réalise une sélection de variables de type stepwise avec pénalisation BIC sur ces données d'expression génique. Les variables selectionnées sont : ADIPOQ, ANGPTL8, APOD, APOE, CRYAB, FBLN1, FMOD, INSIG1, RARRES2 et LEP.
Ici le $R^2$ est de 0,422.

Pour voir l'influence de ces variables sur l'APOM, on va réaliser une nouvelle régression mais uniquement sur les variables selectionnées. Le nuage de points qui en découle (image 5.2) montre qu'il n'y a pas beaucoup de différences entre le nuage de points où l'on considère toutes les variables d'expression génique du jeu de données et le nuage de points où l'on considère uniquement les variables du modèle selectionné. Il y a une légère diminution des valeurs prédites.

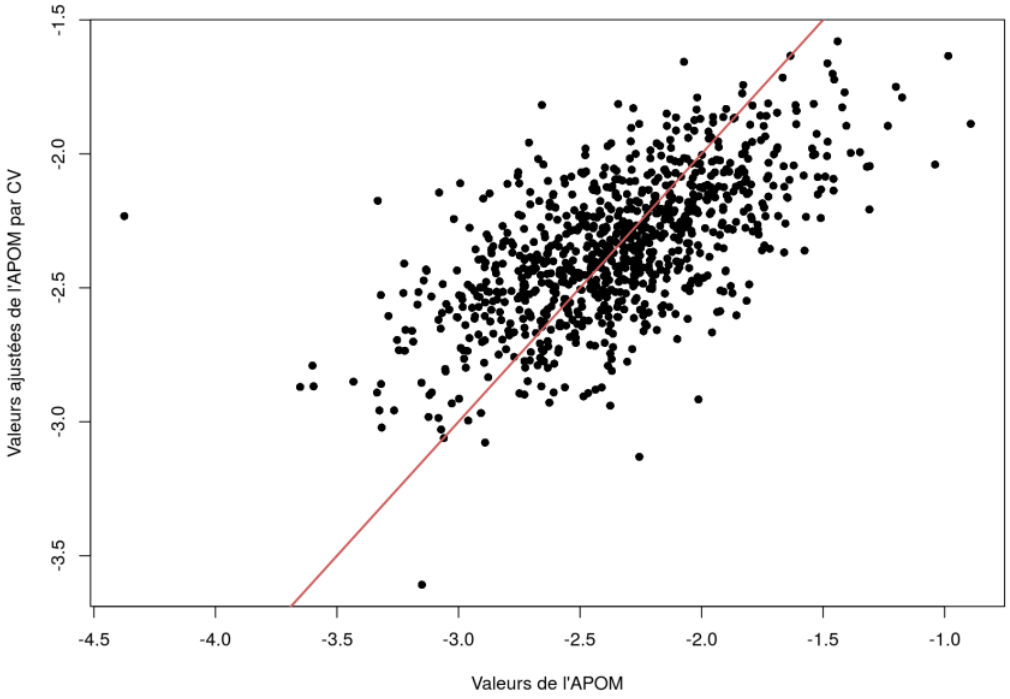



Image 5.2 – Nuage de points des valeurs ajustées de l'APOM par la régression linéaire du modèle selectionné en fonction des valeurs réelles de l'APOM sur les données d'expression génique

La valeur du nouveau $R^2$ est de 0,395.

### Données cliniques

On a également réalisé une régression linéaire multiple sur les données cliniques. À chaque étape, on réalise une regression afin de savoir quelles variables expliquent les valeurs de l'APOM que ce soit à CID1, à CID2 ou à CID3. On réalise de nouveau une sélection de variable de type stepwise avec pénalisation BIC sur les données clinique.
Par exemple, les variables selectionnées à CID2 sont Protéine C Réactive en mg/L, la valeur calorique journalière de l'alimentation en kJ/jour, le taux de glucose à CID1 et le poids total perdu à la 8ème semaine. Les 3 premières variables évoluent dans le sens inverse des variations de l'APOM, quand les valeurs de ces 3 variables diminuent, les valeurs de l'APOM augmentent, et le poids total perdu à la 8ème semaine évolue dans le même sens que l'APOM. La protéine C réactive est une protéine qui apparaît dans le sang lors d'une inflammation aiguë. Son taux augmente rapidement après le début de l'inflammation. On est soumis à des risques plus important d'inflammation en tant qu'obèse, ce qui peut expliquer la présence de cette variable comme variable explicative.

En ce qui concerne la variable « poids total perdu après la première phase de l'expérience », plus les valeurs de cette variable sont fortes, c'est-à-dire plus les individus vont perdre de poids après la première phase, plus l'APOM est exprimée. On observe donc bien que l'expression de l'APOM est expliquée par certaines variables caractérisant la perte de poids.
Ces regréssions linéaires avaient été réalisées pour savoir si le type de régime suivi entre CID2 et CID3 avait une influence sur les valeurs de l'APOM. Plus précisément, on voulait voir si l'expression de l'APOM augmentait selon le régime réalisé par le patient. Or la variable « Régime » n'est selectionnée dans aucun des 3 modèles. Lorsque l'on observe les distributions des taux d'évolution, entre CID2 et CID3, selon les différents régimes (figure 5.3), on observe très peu de différences entre ces distributions.

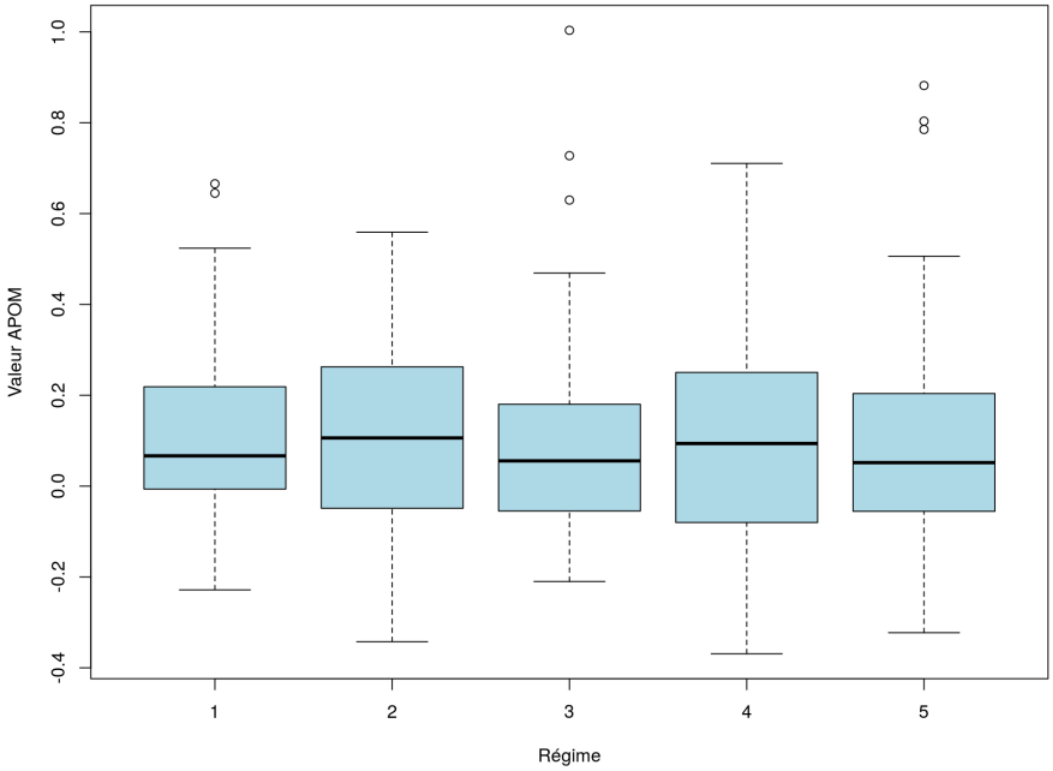



Image 5.3 – Distributions du taux d'évolution de l'APOM entre CID2 et CID3 en fonction des différents régimes

Pour confirmer on regarde le test de Shapiro-Wilks pour vérifier la normalité de la distribution. La p-valeur n'est pas significative au seuil $\alpha = 5\%$ On a réalisé un test de Kruskal Wallis qui est un test non paramétrique de comparaison de distributions. La valeur de la statistique de test est de 1,6792 et la p-valeur est de 0,7945. Il n'y a donc pas de différence significative de l'APOM selon le type de régime.

## Conclusion

En ce qui concerne les données d'expression génique, l'ADIPOQ, l'APOE et la LEP sont des variables explicatives des valeurs de l'APOM comme on pouvait s'y attendre au vu des résultats précédents. En effet, l'APOM avaient des coefficients de corrélation assez forts et significatifs avec ADIPOQ et APOE. En ce qui concerne les données cliniques, on avait donné l'hypothèse que la variable « Régime », soit une des variables explicatives des valeurs de l'APOM, ce qui n'est pas le cas, et ce dans aucune des 3 phases de l'expérience. Cependant les variables sélectionnées apportent de précieuses informations sur les valeurs de l'APOM notamment à CID2.

# Conclusion

Je ressors enrichi professionnellement et personnellement de ce stage car il m'a permis de découvrir dans le détail le secteur de la biologie, ses acteurs, contraintes... J'ai acquis beaucoup de choses, tant dans la dimension professionnelle que dans la dimension humaine de l'entreprise. Mon esprit critique, mon sens de la responsabilité et mon indépendance se sont développés à travers les missions qui m'ont été confiées. J'ai pu faire le rapprochement entre ce que j'avais appris en cours et ce qui se passe vraiment dans une entreprise. Il m'a permis de participer concrètement à ses enjeux au travers de mes missions variées comme celle qui m'a été confiée, que j'ai particulièrement appréciée.

Ce stage m'a aussi permis de comprendre que j'étais attentif aux conseils, aux consignes et aux remarques, mais que j'avais du mal à être sûr de moi, et à parfois à communiquer avec autrui, mais je pense tout de même avoir évolué quelque peu sur ces sujets. En effet, mon sens du dialogue s'est amélioré car je n'ai pas hésité à poser des questions quand je ne comprenais pas.

Fort de cette expérience, j'aimerais beaucoup par la suite me lancer dans le monde du travail, vers le secteur biologique qui m'a passionné.

# Bibliographie

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57 :289–300.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5–32.

[Capeau, 2003] Capeau, J. (2003). Voies de signalisation de l'insuline : mécanismes affectés dans l'insulinorésistance. *Médecine/Sciences*, 19(8-9) :834–839.

[Johnson et al., 2007] Johnson, W., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1) :118–127.

[Lacquemant et al., 2003] Lacquemant, C., Vasseur, F., Lepetre, F., and Froguel, P. (2003). Cytokines d'origine adipocytaire, obésité et développement du diabète. *Médecine/Sciences*, 19(8-9) :809–817.

[Stekhoven and Bühlmann, 2012] Stekhoven, D. and Bühlmann, P. (2012). Missforest-non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1) :112–118.

| | |
|---|---|
| | Diplôme : Master<br>Spécialité : Data Science<br>Spécialisation / option :<br>Enseignant référent : Mathieu EMILY |

| | |
|---|---|
| Auteur(s) : Thibaut Guignard<br><br>Date de naissance : 23/08/1991 | Organisme d'accueil : INRA de Castanet<br><br>Adresse :<br>24 Chemin de Borde Rouge, 31326 Castanet-Tolosan<br><br>Maître de stage : Nathalie VILLA-VIALANEIX |
| Nb pages : 41 Annexe(s) : 0 | |
| Année de soutenance : 2017 | |

Titre français :

Analyse statistique d'une protéine impliquée dans les problèmes d'obésité en relation avec l'expression des gènes

Titre anglais :

Statistic analysis of a protein involve with problem of obesity related to gene expression

Résumé (1600 caractères maximum) :

Ce rapport montre le travail que j'ai réalisé et les résultats que j'ai trouvé lors de mon stage dont la problématique s'intitule « Analyse statistique d'une protéine impliquée dans les problèmes d'obésité en relation avec l'expression des gènes ».

L'obésité est caractérisée par un excès de tissu adipeux qui s'accompagne à long terme de complications métaboliques et cardiovasculaires dont les mécanismes d'installation sont encore à clarifier. Le Laboratoire de Recherche sur les Obésités a mis en évidence une protéine produite par le tissu adipeux et qui n'a pas encore été étudiée dans le domaine de l'obésité et du diabète. L'objectif est ici de rechercher où se situe cette nouvelle protéine parmi les variables, paramètres plasmatiques ou expression de gènes, expliquant l'amélioration de la sensibilité à l'insuline pendant le régime hypocalorique et dans quelle mesure cet effet est indépendant de la perte de poids.

Abstract (1600 caractères maximum) :

This report show the work what i've done and the results what i've found during my internship which the problem is titled « Statistical analysis of a protein involved in obese problems related to genes expression ».

Obesity is characterised by an excess of adipose tissue which comes along in the long term with metabolics and cardiovasculars complications which the mechanisms of installation are far from clear. The Research Laboratory on Obesities has highligthed a protein produced by the adipose tissue which was not studied in the area of obesity and diabetes yet. The objective is to look for where is situated this protein among variables, serum parameters or genes expression, explaining enhancing insulin action during the hypocaloric diet and to what extent this effect is independant of the weight loss.

Mots-clés : Obésité, protéines, gènes, expérience, régime, données, apprentissage statistique, biostatistique

Key Words : Obesity, protein, gene, experience, diet, data, statistical learning, biostatistic