# Méthodes statistiques pour le traitement des données de protéomique quantitative longitudinale

Hélène Borges

# THÈSE

Pour obtenir le grade de

## DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

Spécialité : **Modèles, méthodes et algorithmes en biologie, santé et environnement**

Arrêté ministériel : 25 mai 2016

Présentée par


## Hélène BORGES


Thèse dirigée par **Thomas BURGER, Directeur de recherche, CNRS**, co-dirigée par **Virginie BRUN, Directrice de recherche, CEA** et co-encadrée **par Yohann COUTÉ, Ingénieur de recherche, INSERM**

préparée au sein du **Laboratoire Exploring the Dynamics of Proteomes** dans **l'École Doctorale d'Ingénierie pour la Santé, la Cognition et l'Environnement**

# Méthodes statistiques pour le traitement des données de protéomique quantitative longitudinale

Thèse soutenue publiquement le **19 octobre 2021**,
devant le jury composé de :

**Mme. Sophie ROUSSEAUX**
Directrice de recherche, Université Grenoble Alpes, Présidente
**Mme. Christine CARAPITO**
Chargée de recherche, Université de Strasbourg, Rapportrice
**M. Jacques COLINGE**
Professeur, Université de Montpellier, Rapporteur
**M. Quentin GIAI GIANETTO**
Ingénieur de Recherche, Institut Pasteur, Examinateur
**M. Laurent GATTO**
Professeur associé, Université catholique de Louvain, Examinateur
**M. Thomas BURGER**
Directeur de Recherche, Université Grenoble Alpes, Directeur de thèse

*A la patience, la confiance et le soutien infinitésimaux*
*de mes parents, de ma sœur, de ma famille, de ma moitié et de mes amis.*

# Remerciements

Je souhaite en tout premier lieu remercier Mme Christine Carapito et M. Jacques Colinge d'avoir accepté d'être les rapporteurs de mes travaux de thèse, ainsi que Mme Sophie Rousseaux, et Messieurs Quentin Giai Gianetto et Laurent Gatto d'avoir accepté d'évaluer ce travail.

Mes sincères remerciements vont aussi à mon comité de suivi de thèse pour leurs conseils avisés : Valérie Siroux et Olivier François.

Je remercie Thomas Burger, Virginie Brun et Yohann Couté de m'avoir confié ce projet et de m'avoir guidé et conseillé pendant tout ce temps. Merci Thomas d'avoir été un directeur de thèse très présent, tu as contribué de manière décisive à la réalisation de cette thèse, en te souciant de sa viabilité et de sa cohérence. Je partais de loin, et j'ai vraiment beaucoup appris pendant cette thèse, aussi bien sur le plan humain, scientifique et rédactionnel. Merci pour tout cela !

Je remercie tout le labo, EDyP, aussi bien ceux qui y sont encore, que ceux qui en sont partis et les nouveaux de 2020/2021 que je n'ai malheureusement pu croiser qu'en de rares occasions. Tous vous m'avez accueilli malgré mon mauvais caractère et mon franc-parler dérangeant (que vous m'avez appris à dompter un peu cela dit) ! Merci pour toutes ces conversations scientifiques, culturelles et ces débats interminables mais très intéressants. Un grand merci à Thomas Fortin, pour m'avoir fait découvrir l'escalade ! Grâce à toi je pouvais grimper du 6b et puis j'ai un jour compris avec Julia qu'en fait tu m'aidais beaucoup et que je suis à peine capable de faire du 4c haha ! Mais je vais m'entraîner promis ! To Szu-Hsueh, thank you sooooo much for your listening, your advice, your encouragement, your knowledge on so many things and our long discussions about the different languages and cultures that enrich humanity. David, Lucas et Vaitson, vous allez y arriver ! La thèse c'est très facile ! On comprend rien au début mais on fini par y voir clair !

Durant cette thèse, il y aussi eu la naissance de nombreuses amitiés : Olga, Ohlala Olga, je te dois tellement ! Sans toi la thèse aurait été tellement différente, que je ne veux même pas imaginer une autre branche spacio-temporelle ! On finissait seules chaque soir au labo quand il faisait nuit en hiver et qu'il n'y avait plus personne, c'était bien calme, parfait pour coder ! Lucid, merci pour tous ces bons moments au café, en soirée, et surtout à m'avoir appris à mieux comprendre l'ironie et la diplomatie ! Sara, un merci infini pour

tes conseils, ton écoute, les petis moments passés (et à passer) ensemble à parler de tout, de rien, de l'avenir, de la philosophie de la vie, et surtout de se dire qu'il faut arrêter de se prendre la tête pour tout. Julia, merci pour tout, tout simplement. Pour les balades, l'escalade, les tricothés, les soirées jeux, les bars, m'avoir fait découvrir ton beau Pays Basque (il faut qu'on y retourne pour manger du jambon), pour tout ce temps que tu as passé à m'écouter parler et râler de ma thèse, et à me dire de tenir et de réussir !

A mes amis de promo bioinfo, Floflo, Juju, Chacha, Jayjay, Nanis, Nico et les autres qui se reconnaîtrons : merci pour votre écoute, vos conseils et votre amitié.

Et le meilleur pour la fin, ma famille, que j'aime par-dessus tout, qu'elle soit en France ou en Espagne, vous êtes toujours avec moi, même si c'est à des centaines et des milliers de kilomètres. Mi Mamá, tu as toujours été là pour me gronder, me pousser, me dire que je suis capable et que je ne dois jamais rien lâcher, car à partir d'un certain niveau c'est le baccalauréat tous les jours ! et 634 km c'est vraiment loin tu as raison... Mi Papá, tu m'avais dit que les études c'est très important, et j'ai finalement tellement bien intégré ce conseil que je suis allée jusqu'au doctorat ! Ma ch'tite Anne adorée, ma petite soeur qui a toujours et sera toujours là pour moi (la réciproque est vraie), et qui est toujours venue jusqu'à Grenoble pour me voir (la réciproque n'est pas vraie, et je te demande pardon). Ma mamie, merci d'être encore là, et de m'avoir toujours encouragée à faire des études, et de continuer l'école jusqu'à 28 ans ! Mi Tio Pruden, siempre attento, con quien he compartido mis dudas y mis penas. Mis tias, Mari y Pili, también siempre conmigo, no hos he dado noticias bastantes veces, perdonadme por eso. La tesis se acabó, pronto volveré a veros ! A toute ma famille, qui n'a jamais vraiment tout compris de ce que je racontais mais qui savait que je faisais de mon mieux. A mon Thibaut, je pense que tu es celui qui a le plus souffert de ces 3 ans et demi de thèse, et à l'avoir vécu tous les jours : ta patience est effarante et ton affection inébranlable (cette réciproque est également vraie).

Merci à tous du fond du cœur.

# Résumé

L'objectif de la protéomique est l'identification et la quantification des protéines présentes dans des échantillons biologiques. Une de ses applications est la recherche de biomarqueurs, c'est-à-dire des entités mesurables décrivant précisément un état biologique spécifique. Ces biomarqueurs peuvent ensuite être utilisés dans un contexte clinique, pour le diagnostic ou le suivi de patients atteints de pathologies, notamment chroniques et ainsi assister les cliniciens dans leur prise en charge et leur traitement. La découverte de biomarqueurs passe par l'analyse différentielle des protéines, autrement dit, la mise en évidence d'une altération de l'expression des protéines entre les différents échantillons par une analyse statistique. Cependant, l'analyse de grandes cohortes cliniques nécessite des chaînes instrumentales spécifiques produisant des données complexes en raison de biais techniques et d'une variabilité inter-patient pouvant altérer les résultats. Afin de prendre cela en compte, tout en maintenant un haut niveau d'automatisation (indispensables au bon fonctionnement d'une plateforme d'analyse gérant de manière concomitante de multiples projets), des développements méthodologiques ainsi que leur implémentation logicielle sont nécessaires. Ce travail de thèse cherche à répondre à ce besoin, sous la forme de trois contributions principales. La première est la création de *Well Plate Maker*, un logiciel facilitant la conception de protocoles expérimentaux plus robustes. Le logiciel génère automatiquement un placement des échantillons sur des plaques à puits qui minimise les biais potentiels dans l'expérience et permet d'obtenir en aval des résultats statistiques plus reproductibles. La seconde est l'adaptation fiable et reproductible de l'Analyse de la Variance (une approche classique en statistique) afin de répondre aux spécificités des données de protéomique. Cette adaptation est combinée à des méthodes de représentation et de visualisation des profils d'expression des protéines, tout en préservant une utilisation facile pour les protéomiciens dans un contexte applicatif de plateforme d'analyse. La troisième contribution est la mise en application concrète de la méthodologie ainsi proposée sur une cohorte clinique de patients atteints de stéatose hépatique non-alcoolique. Nous avons identifié des protéines présentant des profils d'expression décrivant la progression de la pathologie, qui pourront être intéressantes à approfondir dans des études cliniques ultérieures. Au-delà du cas des stéatoses hépatiques non-alcooliques, ce travail illustre l'intérêt de la protéomique comme un outil complémentaire fiable dans le contexte clinique du suivi et de la prise en charge de patients.

# Abstract

The goal of proteomics is the identification and quantification of proteins present in biological samples. One of its applications is the search for new biomarkers, *i.e.*, measurable entities that precisely describe a specific biological state. These biomarkers can then be used in a clinical context, for the diagnosis or the medical monitoring of patients suffering from pathologies, in particular chronic ones hereby assisting the clinicians in their care and treatment. Biomarker discovery involves the differential analysis of proteins, in other words, the demonstration of an alteration in the expression of proteins between the different samples by a statistical analysis. However, the analysis of large clinical cohorts requires specific numerous instruments producing complex data. These data are difficult to process, due to technical biases and inter-patient variability. Eventually, inadequate processing of these data can lead to erroneous results. To tackle this challenge, while maintaining a high level of automation (essential for the daily work of an analytic platform managing multiple concomitant projects), methodological developments as well as their software implementation are necessary. This work seeks to meet this need thanks to three main contributions. The first one is the development of the *Well Plate Maker* software to assist in the design of more robust experimental protocols. The software automatically generates a well plate filling strategy that minimizes potential biases in the experiment and consequently allow more reproducible downstream statistical results. The second one is the reliable and reproducible adaptation of Analysis of Variance (a classic statistical approach) to account for the specificities of proteomics data. This adaptation is combined with methods of representation and visualization of protein expression profiles, while preserving ease of use for proteomicians in an application context of an analysis platform. The third contribution is the concrete application of the above methodology on a clinical cohort of patients with non-alcoholic fatty liver disease. We have identified proteins with expression profiles describing the progression of the disease, which may be of interest to explore in further clinical studies. Beyond the case of non-alcoholic fatty liver disease, this work illustrates the interest of proteomics as a reliable complementary tool in the clinical context of patient monitoring and care.

# Avant-propos

Bien que le titre ne le laisse pas vraiment transparaître, ce travail de thèse comporte deux volets : d'une part, une recherche de développement méthodologique pour l'analyse statistique en protéomique, et d'autre part, une mise en application dans un projet de recherche biomédicale dans l'espoir de mieux comprendre une pathologie.

Bioinformaticienne de formation, j'ai toujours été attirée par les sujets appliqués à la médecine humaine. C'est donc tout naturellement que ma curiosité a été piquée par ce sujet de thèse. Le travail de recherche que j'ai décidé d'entreprendre associe une réflexion méthodologique à une mise en pratique concrète afin de proposer des outils à la communauté scientifique, en particulier celle de protéomique. Répondre aux différentes problématiques auxquelles j'ai été confrontée s'est traduit par la rencontre de plusieurs domaines de recherche : la protéomique, la clinique, les statistiques et la bioinformatique. Bien que plus ou moins éloignés, ils se complètent à merveille, et permettent de répondre à des questionnements (biologiques) toujours plus complexes.

Mon travail de thèse a d'abord consisté à compléter mes quelques connaissances en statistiques et à découvrir une pathologie dont j'ignorais l'existence et qui est pourtant bien réelle dans nos sociétés modernes : la stéatose hépatique non-alcoolique (« *Non-Alcoholic Fatty Liver Disease* » ou NAFLD). En parallèle, j'ai découvert la protéomique, une discipline complexe mais pleine de promesses dont les besoins en méthodologie statistique avec des outils adaptés sont encore très présents.

Mes trois années de thèse m'ont donc permis de découvrir de nouvelles disciplines, de compléter des outils statistiques pré-existants et d'explorer le protéome plasmatique de patients NAFLD à la recherche de potentiels biomarqueurs candidats. Dans son ensemble, la thèse s'est révélée être pour moi un parcours semé d'imprévus et d'inconnues auxquels j'ai dû faire face en apportant des idées et des approches différentes. Ces trois années ont été l'occasion d'assimiler une avalanche de nouvelles connaissances, d'acquérir une méthode de travail bien particulière qu'est celle du chercheur, mais aussi, et surtout, de faire de belles rencontres humaines.

# Liste des abréviations

| | |
|---|---|
| **AA** | Acide Aminé |
| **AC** | Assignation Croisée |
| **ANOVA** | *ANalysis Of VAriance* - Analyse de la variance |
| **AP** | Affinity Propagation clustering |
| **BH** | Benjamini & Hochberg |
| **BP** | *Biological Process* |
| **CID** | *Collision Induced Dissociation* |
| **CME** | Carrés Moyens Expliqués |
| **CMR** | Carrés Moyens Résiduels |
| **CMT** | Carrés Moyens Totaux |
| **DA** | Différentiellement Abondant |
| **DDA** | *Data Dependent Acquisition* - Acquisition indépendante des données |
| **DIA** | *Data Independent Acquisition* - Acquisition dépendante des données |
| **ESI** | *ElectroSpray Ionization* |
| **ETD** | *Electron Transfert Dissociation* |
| **FDP** | *False Discovery Proportion* - Proportion de fausses découvertes |
| **FDR** | *False Discovery Rate* - Taux de fausses découvertes |
| **FLIP** | *Fatty Liver Inhibition of Progression* |
| **GO** | *Gene Ontology* |
| **GUI** | *Graphic User Interface* - Interface Graphique Utilisateur |
| **HCD** | *High Collision Dissociation* |
| **HPLC** | *High Performance Liquid Chromatography* |
| **IMC** | Indice de Masse Corporelle |
| **LC** | *Liquid Chromatography* - Chromatographie Liquide |
| **LC-MS** | Chromatographie Liquide combinée à la Spectrométrie de Masse |
| **LOESS** | *LOcally weighted regrESSion* |
| **MALDI** | *Matrix-Assisted Laser Desorption/Ionization* |
| **MEC** | *Missing in Entire Condition* |
| **MF** | *Molecular Fonction* |
| **MS** | *Mass Spectrometry* - Spectrométrie de Masse |
| **MS/MS** | Spectrométrie de Masse en tandem |
| **MTC** | *Mutiple Testing Correction* - correction de tests multiples |
| **NAFLD** | *Non Alcoholic Fatty Liver Disease* - Stéatose hépatique non-alcoolique |
| **NAS** | *NAFLD Activity Score* |
| **NASH** | *Non-Alcoholic Steato Hepatitis* - Stéatohépatite non-alcoolique |
| **POV** | *Partially Observed Values* |
| **PSM** | *Peptide-to-Spectrum Match* |
| **QC** | Contrôle Qualité |
| **SAF** | Score de Stéatose, Activité et Fibrose |
| **SCE** | Somme des Carrés Expliqués |
| **SCR** | Somme des Carrés Résiduels |
| **SCT** | Somme des Carrés Totaux |
| **T2D** | Diabète de Type 2 |
| **vsn** | *variance stabilization normalisation* |
| **wpm** | Well Plate Maker |
| **XIC** | *eXtracted Ion Chromatogram* |

# Table des figures

# Liste des tableaux

# Table des matières

# Chapitre 1

# Introduction

## 1.1 Contexte

### 1.1.1 Résumé des objectifs de thèse

Physique, chimie, mathématiques et informatique ont permis de développer des technologies dites « omiques » et d'ouvrir de nouvelle perspectives pour la recherche en biologie. Il s'agit d'une ingénierie d'analyse automatisée et systématique du vivant à l'échelle moléculaire, qui a pour ambition de fournir une vision aussi exhaustive que possible. Parmi ces « omiques » nous pouvons citer la génomique, consacrée à l'étude de l'ensemble des gènes, et particulièrement connue du grand public (notamment en raison du retentissement médiatique qu'a eu le premier séquençage complet du « génome » humain à l'orée des années 2000 [1]). Une autre omique plus récente, est la protéomique, dont les principaux objectifs sont l'identification et la quantification de l'ensemble des protéines présentes dans un échantillon biologique (que l'on nomme protéome). De par son sujet d'étude, la protéomique ambitionne de caractériser précisément un état biologique à un instant donné, afin de mieux comprendre les mécanismes cellulaires [2]. Les technologies développées et utilisées en protéomique sont à la fois récentes et en constant perfectionnement (une description détaillée de celles employées dans notre laboratoire est donnée dans le chapitre 2). Ce perfectionnement induit des chaînes instrumentales de plus en plus complexes et produisant des données plus volumineuses et plus difficile à analyser, induisant de nouveaux défis bioinformatiques et biostatistiques.

Notamment, il est à l'heure actuelle compliqué pour les plateformes de protéomique de

réaliser une analyse statistique fiable, reproductible, tenant compte des potentiels biais expérimentaux, et adaptée aux spécificités des données produites. En effet, cela nécessite à la fois des développements méthodologiques originaux, mais aussi l'amélioration des outils logiciels disponibles, afin d'atteindre les niveaux d'automatisation nécessaires au bon fonctionnement d'une plateforme gérant de manière concomitante de multiples projets. Cela est encore plus vrai lorsque l'on souhaite utiliser la protéomique à grande échelle, car la complexité du vivant implique une diversité des conditions biologiques observables (par exemple, différents états de progression d'une maladie).

Dans bien des cas, l'investigation biologique ou clinique est menée de manière différentielle : L'étude cherche à comprendre quels phénomènes pourraient expliquer la différence entre deux états précisément définis : l'état basal (également appelé « contrôle », c'est l'état où *a priori* il n'y a pas de perturbation) et l'état altéré (ou plus précisément, l'état pathologique dans le cas d'une étude clinique). Le plan d'expérience (ou plan expérimental) permet de comparer les données biologiques disponibles entre ces deux états et d'identifier les mécanismes qui pourraient expliquer les différences de phénotypes (notamment, une pathologie). Plus précisément, par plan d'expérience, nous entendons ici l'ensemble ordonné d'essais expérimentaux, qui permettent d'acquérir de nouvelles connaissances en contrôlant un ou plusieurs paramètres (appelés *facteurs*) d'intérêt afin de répondre à une question biologique précise et définie. Cette question biologique se traduit classiquement en statistiques par la formulation d'une hypothèse à tester. Les outils statistiques permettant de traiter un plan expérimental en protéomique sont maintenant assez courants. Le Chapitre 3 en présente les principes généraux.

Cependant, pour des raisons principalement liées à la méthodologie analytique sous-jacente, la majorité des études protéomiques ont jusqu'à présent été menées selon des plans d'expérience « binaires », c'est-à-dire n'impliquant la comparaison que de deux conditions. Néanmoins, cela est en train de changer [3, 4], notamment grâce à *i)* l'augmentation du débit analytique des technologies employées (elles permettent de traiter un nombre croissant d'échantillons à temps et coûts fixés) et à *ii)* l'influence de la vision « *big data* » de la recherche moderne, d'où émerge le besoin de fournir une vision plus complète des phénomènes biologiques, supportée par toujours plus de données permettant des statistiques à la fois plus robustes et puissantes. En effet, il ne s'agit plus de traiter quelques dizaines d'échantillons mais désormais plusieurs centaines, voire des milliers [5], ce qui permet de s'intéresser à des phénomènes beaucoup plus complexes, ou n'induisant que des signaux biologiques faibles. Cependant, un affinement des méthodologies statistiques

est encore nécessaire pour tenir compte de ces plans d'expérience tout en conservant un traitement fin des spécificités induites par la chaîne instrumentale utilisée en protéomique (cela sera décrit plus en détails aux chapitres 2 et 3). Par ailleurs, la théorie statistique comme le bon usage des outils qui en découle ne sont que peu maîtrisés par les chimistes analyticiens ou les biologistes travaillant en protéomique, de sorte qu'un réel effort interdisciplinaire de mise à disposition est nécessaire. Notamment, nous avons identifié un réel besoin en termes de logiciels donnant accès aux protéomiciens à des outils statistiques à la fois fiables et spécifiques à leurs données. Les chapitre 5 et 6 présentent les solutions que je propose pour pallier une partie de ces besoins, tant méthodologiques qu'applicatifs.

### 1.1.2 Objectif du projet support : étude de la cohorte NAFLD

Mon sujet de thèse ayant pour objet l'analyse de données de protéomique quantitative, c'est tout naturellement qu'une partie de mon travail a été influencée par le projet universitaire finançant mon contrat ainsi que la recherche clinique devant directement bénéficier de la méthodologie et des outils que j'allais développer pendant ces trois années : le projet LIFE. Parmi les différents objectifs de ce projet, celui qui nous intéresse porte sur la recherche de biomarqueurs [1] chez les patients atteints d'une pathologie chronique qu'est la stéatose hépatique non-alcoolique (« *Non Alcoholic Fatty Liver Disease* », ou NAFLD). Il s'agit d'une dégénérescence hépatique progressive causée par une accumulation chronique de lipides dans le foie. Cette recherche a pour objectif de remplacer la biopsie hépatique (actuelle méthode de référence pour le diagnostic de la maladie) par une simple prise de sang, car cette dernière faciliterait grandement le suivi des patients et permettrait de s'affranchir des nombreuses limites de la biopsie (pour plus de détails sur la pathologie et les méthodes de diagnostic, voir la Section 7.1). Les médecins du CHU Grenoble-Alpes impliqués dans le projet LIFE, nous ont confié les échantillons d'une cohorte de 160 patients NAFLD avec l'objectif d'identifier de potentiels biomarqueurs circulant dans le plasma sanguin, et permettant la stratification de la pathologie en différents stades de gravité. Le plasma contenant beaucoup de protéines sécrétées par le foie, l'objectif est légitimement d'identifier les protéines pouvant avoir ce rôle de biomarqueur. Malgré l'expertise reconnue de notre laboratoire en termes de production et de traitement de données de protéomique, ce projet amène trois défis. Le premier est la complexité de la NAFLD. Le second est l'ambiguïté du plan d'expérience (cf. Chapitres 4 et 7). Le troisième est la taille de la cohorte qui est beaucoup plus importante que celles habituellement analysées

1. Entité mesurable décrivant précisément un état biologique spécifique.

au laboratoire. Ma thèse a pour objectif de relever le second aspect de ce défi, qui sera explicité dans le Chapitre 4.

De manière plus précise, cet objectif peut être décliné en deux volets : Le premier volet consiste à mettre en place une méthodologie statistique fiable et reproductible, qui soit *i)* adaptée aux données de protéomique quantitative ; *ii)* Adaptée aux plans d'expériences cliniques, plus complexes que les comparaisons binaires que l'on trouve majoritairement en biologie fondamentale ; et *iii)* qui soit intégrée dans un outil logiciel la rendant facile d'utilisation pour les protéomiciens dans un contexte applicatif de plateforme d'analyse. Le second volet concerne la validation de cette méthodologie sur des données réelles pour l'identification de potentiels biomarqueurs de la NAFLD. Au-delà de ma thèse proprement dite, la mise en application de la méthodologie et des outils d'analyse ainsi développés a l'ambition de permettre à son projet cadre de constituer une preuve de concept ; et de montrer que la protéomique peut devenir un outil fiable dans le contexte clinique du suivi et de la prise en charge de patients souffrant de maladies chroniques (obésité, diabète, troubles cardiovasculaires, ...).

## 1.2 Depuis les protéines jusqu'à la protéomique

### 1.2.1 Les protéines comme briques du vivant

Avant de décrire les technologies issues de la chimie analytique qui sont employées en protéomique quantitative, nous allons dans un premier temps revenir sur certains concepts fondamentaux de biologie. En effet, comme décrit précédemment, la protéomique s'intéresse à l'étude du protéome. Un protéome se définit comme l'ensemble des protéines constituant un organisme, un tissu, une cellule ou encore un fluide biologique dans des conditions spécifiques à un instant donné [6].

Les protéines sont des macromolécules biologiques. Elles constituent un enchaînement de longueur très variable d'acides aminés reliés entre eux par des liaisons peptidiques. Un acide aminé (AA) est composé d'un groupement carboxylique (COOH), d'un groupement amine ($NH_2$) et d'une chaîne latérale qui peut être plus ou moins ramifiée (voir Figure 1.1). La liaison peptidique est une liaison covalente qui s'établit entre la fonction carboxyle (COOH) portée par le carbone $\alpha$ d'un acide aminé et la fonction amine ($NH_2$) portée par le carbone $\alpha$ de l'acide aminé suivant avec libération d'une molécule d'eau. Les acides aminés peuvent s'assembler ainsi en séquences peptidiques dont la longueur peut

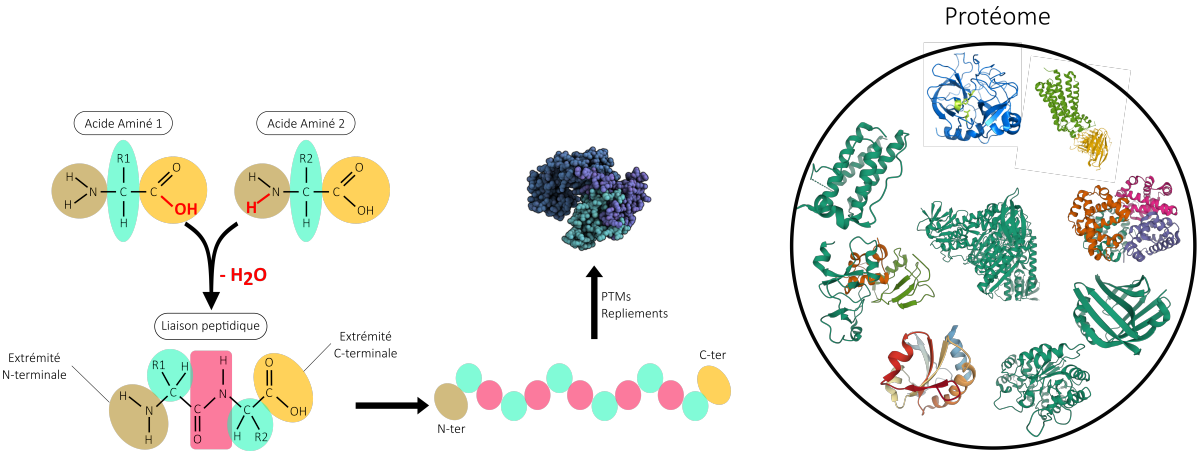



FIGURE 1.1 – L'association de deux acides aminés forme une liaison peptidique avec une molécule d'eau. La partie constituée du groupement amine ($NH_2$) est nommée N-terminale, tandis que l'extrémité composée de l'acide carboxylique (COOH) est nommée C-terminale. L'enchaînement de plusieurs acides aminés donnent une séquence peptidique qui va subir des modifications post-traductionnelles (*Post-Translational Modification* ou PTMs) et des repliements pour former une protéine. L'ensemble des protéines considéré à un instant $t$ constitue le protéome.

varier de quelques dizaines à plusieurs milliers pour former des protéines. En protéomique, le terme « peptide », qui peut habituellement désigner toute chaîne d'acides aminés, est généralement réservé à celles obtenues par le clivage des protéines étudiées lors du processus analytique. L'extrémité de la séquence peptidique qui porte le groupe $NH_2$ du premier AA est appelée extrémité N-terminale (N-ter). L'autre extrémité qui comporte le groupement COOH du dernier AA est appelée C-terminale (C-ter). Le « sens de lecture » de la séquence peptidique se fait usuellement de N-ter vers C-ter.

Bien qu'il existe plusieurs centaines d'acide aminés chez les eucaryotes, seulement 21 acides aminés permettent la synthèse des protéines chez l'homme [7], appelés acides aminés protéinogènes. Ce nombre d'acides aminés « briques », de prime abord « faible », permet pourtant de générer un nombre d'enchaînements possibles (les protéines) considérable : par exemple le nombre de peptides distincts formés de 10 acides aminés est de l'ordre de $10^{13}$. Les 21 acides aminés ont chacun un nom propre associé à une lettre (voir Figure 1.2). La séquence peptidique peut alors être lue comme un « mot ». Cette nomenclature est à la base de l'identification des peptides et des protéines (cf Section 2.3.1).

La biosynthèse d'une protéine s'effectue résidu après résidu, depuis l'extrémité N-terminale jusqu'à l'extrémité C-terminale. Cet enchaînement d'AA possède des propriétés physico-chimiques qui lui sont propres, et qui détermine la structure tridimensionnelle de la protéine ainsi que sa fonction biologique (structurelle, de catalyse enzymatique, de trans-

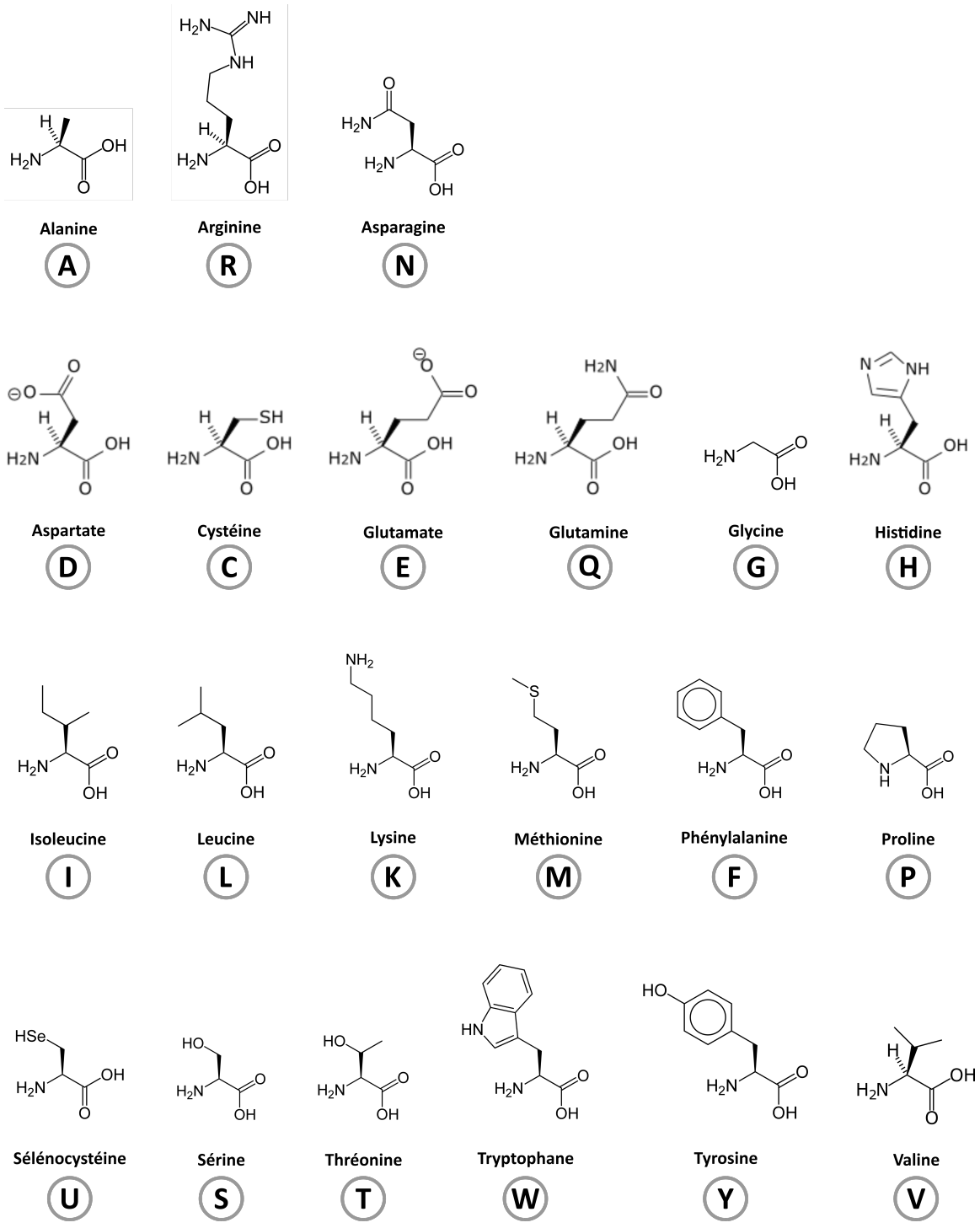



FIGURE 1.2 – Structure des 21 acides aminés protéinogènes chez l'homme avec leur nom et leur abréviation à une lettre (encerclé). Les acides aminés protéinogènes sont les unités de base des protéines. Les séquences peptidiques peuvent être lues comme des « mots » : par exemple, un peptide lu « MENHIR » sera composé d'une méthionine, d'un glutamate, d'une asparagine, d'une histidine, d'une isoleucine et d'une arginine.

port, *etc*) [8]. Par ailleurs, les protéines peuvent interagir entre elles afin de former des complexes [9] (par exemple les canaux membranaires) ; avoir une activité enzymatique, c'est-à-dire qu'elles vont catalyser (accélérer) les réactions chimiques nécessaires au bon fonctionnement cellulaire. Certaines ont un rôle dans le métabolisme (par exemple, la leptine est une hormone impliquée dans le phénomène de satiété [10] ; l'ATP synthase est impliquée dans le métabolisme énergétique), d'autres permettent la signalisation cellulaire (par exemple, les récepteurs membranaires, les hormones) [11], d'autres encore dans la structuration de la charpente cellulaire (par exemple, la tubuline est impliquée dans la formation des microtubules constitutifs du cytosquelette cellulaire) ou bien sont impliquées dans la régulation de l'expression génique (tels les histones [12] ou les facteurs de transcription). C'est notamment pour cela que les protéines peuvent être qualifiées de véritables « briques du vivant » : elles assurent le fonctionnement des cellules, et plus globalement des tissus chez les organismes pluricellulaires.

### 1.2.2 Histoire de la protéomique

Le terme « protéomique » est relativement récent puisqu'il est apparu pour la première fois vers 1997 dans le contexte de l'électrophorèse bi-dimensionnelle sur gel [6]. Évidemment, ce terme est directement issu du mot *protéome*, lui-même apparu pour la première fois en 1995 [13]. L'étude du protéome a connu un intérêt croissant avec l'amélioration des technologies usuellement utilisées, telles que l'électrophorèse bi-dimensionnelle [14] et la spectrométrie de masse (« *Mass Spectrometry* » ou MS) [15]. En effet, bien qu'existantes depuis plusieurs dizaines d'années, ces technologies nécessitaient des améliorations afin de permettre l'analyse non seulement de protéines individuelles, mais aussi de plusieurs milliers de protéines simultanément, afin de caractériser un échantillon biologique complexe. Plus précisément, la MS fut développée au début du XX$^{\text{ème}}$ siècle et est depuis une technologie extrêmement utilisée en chimie analytique. Cependant, son application en biologie pour l'identification des peptides constitutifs des protéines n'a réellement pris son essor que dans les années 1980 [16] avec l'avènement de techniques d'ionisation capables de générer des ions stables en phase gazeuse à partir de molécules thermiquement instables, à savoir désorption/ionisation laser assistée par matrice (« *Matrix-Assisted Laser Desorption/Ionization* » ou MALDI) et l'ionisation par électrospray (« *ElectroSpray Ionization* » ou ESI) [17]. Ce sont concrètement ces améliorations technologiques qui ont permis d'initier le fulgurant développement de la pratique de l'analyse protéomique pour les recherches en biologie [15, 18].

Par ailleurs, l'usage de la protéomique en biologie a aussi évolué. Elle a d'abord constitué une approche qualitative d'analyse des échantillons biologiques. Puis, notamment avec l'avènement de la MS, elle est devenue peu à peu quantitative [15, 19]. Il existe désormais une multitude de technologies en protéomique, depuis les différentes méthodes de séparation des protéines sur gel jusqu'aux nombreuses approches d'identification et quantification des peptides constituant les protéines. Une méthode d'analyse très utilisée et performante est la combinaison de la chromatographie en phase liquide (« *Liquid Chromatography* » ou LC) couplée à la spectrométrie de masse (MS), abrégée en LC-MS. Elle permet l'identification et la quantification très précise des analytes et c'est désormais la technologie la plus employée en protéomique. D'autre part la protéomique par LC-MS s'est divisée en deux grandes catégories : la protéomique « top-down » et la protéomique « *bottom-up* » [20]. La première consiste à analyser directement les protéines intactes. La seconde consiste à d'abord découper les protéines en peptides, puis à séparer ces peptides grâce à la LC, puis d'identifier ces peptides comme substituts des protéines grâce à la MS. Dans la suite de ce document, nous nous focaliserons exclusivement sur la protéomique « *bottom-up* » utilisant un mode particulier de la spectrométrie de masse, dit « en tandem ». La raison est double : non seulement, elle est celle qui est majoritairement utilisée au sein de notre laboratoire, mais il s'agit également de celle dont le traitement de données nécessite le plus de compétences en science des données. L'approche est décrite plus en détails dans le Chapitre 2.

### 1.2.3 Intérêts et applications de la protéomique

Comme dit précédemment, l'objectif principal de la protéomique est « simplement » d'analyser la composition protéique d'un échantillon biologique, ce qui revient à déterminer quelles protéines sont présentes et en quelles quantités. Ces deux étapes sont classiquement suivies par une comparaison, qui permet de déterminer les possibles similarités et différences existantes entre les différents échantillons analysés. Compte-tenu de ces objectifs, la lourdeur expérimentale et instrumentale peut sembler excessive, de sorte qu'il est légitime de questionner son efficacité. On pourrait par exemple se contenter d'exploiter au maximum les autres « omiques » (génomique et transcriptomique [2] notamment) moins onéreuses, moins complexes, ou plus exhaustives. L'intérêt de la protéomique est cependant réel et il peut être résumé en trois principaux points :

2. L'étude de l'ensemble des ARN messagers. La transcriptomique est souvent mise en parallèle avec la protéomique car elle analyse les produits issus du génome (les ARNm) responsables de la régulation de l'expression des protéines.

Tout d'abord, la diversité des protéines et de leurs rôles est telle, et si dépendante des conditions physiques et environnementales, qu'il n'existe ainsi pas *un* protéome, mais plutôt *des* protéomes. Un protéome d'un même individu ou d'une même cellule peut être différent d'un instant à un autre. Par exemple, chez l'homme l'abondance d'insuline circulante dans le sang évolue durant la journée. Le protéome est donc un phénomène dynamique et changeant, contrairement au génome. Un autre exemple intéressant pour comprendre l'importance d'étudier directement les protéines est celui de la chenille et du papillon. Les deux forment un même individu (à différents instants de vie) avec le même génome, mais le protéome (ainsi que le phénotype) de chacun est complètement différent.

Ensuite, un même gène peut encoder plusieurs protéines aux structures et fonctions bien distinctes. Le niveau d'expression des protéines n'est donc pas un simple reflet de l'expression des gènes [21]. Ainsi, pour un même ensemble de gènes dont l'expression des ARN messagers est plus ou moins homogène au sein d'un tissu biologique, le niveau d'expression des protéines effectivement observées peut être variable.

Enfin, les protéines peuvent subir des modifications chimiques (réalisées le plus souvent par une enzyme), leur permettant de modifier leur structure et leur fonction [22]. Il s'agit de modifications post-traductionnelles (PTMs). Cela peut consister en un clivage, une glycosylation, une phosphorylation, une méthylation, une acétylation, *etc.* (pour plus d'informations sur les PTMs voir [23]). Ces phénomènes physico-chimiques sont des mécanismes essentiels pour la régulation et la grande diversité du protéome. Or les PTMs ne sont généralement pas identifiables directement sur la séquence génique codant pour ces protéines, par conséquent étudier directement le protéome permet de mieux comprendre la physiologie cellulaire.

Ainsi, pour comprendre au mieux un organisme, l'étude du protéome semble essentielle car elle permet de mieux appréhender le phénotype, c'est-à-dire l'ensemble des traits biologiques observables. Au-delà de l'aspect dynamique, dans un contexte clinique, la caractérisation des protéines est plus spécifique que les études génomique ou transcriptomique. En effet, le protéome d'un patient est aussi dépendant de son état physiologique et de son contexte environnemental (pollution, alimentation, traitement médicamenteux, *etc.*), de sorte que l'analyse protéomique permet d'obtenir un aperçu global de la complexité du système biologique étudié, tant au niveau traductionnel que post-traductionnel.

## 1.3 La recherche des biomarqueurs plasmatiques, ses enjeux et ses défis

### 1.3.1 Définition et classification des biomarqueurs

Un biomarqueur est une entité biologique que l'on peut mesurer et quantifier de manière à la fois précise et reproductible, et qui reflète les signes cliniques indicateurs de santé ou de maladie [24]. Cette entité peut être une protéine, un métabolite, un gène ou encore un ensemble de mesures physiques. Dans notre contexte de protéomique clinique, un biomarqueur est une protéine, une PTM d'une protéine ou encore un ensemble de protéines (aussi appelée signature biologique ou signature protéique) qui va permettre de décrire précisément l'état des patients, et d'accélérer le développement de traitements adaptés. Plus précisément, il est possible de classer les biomarqueurs en fonction de leur rôle [25, 26, 27], tel que résumé dans la Table 1.1. Un exemple de biomarqueur de suivi est l'hémoglobine glyquée dont on mesure la concentration pour surveiller la glycémie chez les personnes atteintes de diabète de type 2.

| **Type** | **Définition - Objectif** |
|---|---|
| **Suivi** | Mesuré plusieurs fois dans le temps afin de suivre l'état de la pathologie/condition physique d'un patient donné. |
| **Diagnostic** | Détecter ou confirmer la présence d'une pathologie (ou une condition d'intérêt) et d'identifier les individus présentant un sous-type de la maladie en question. |
| **Pronostic** | Identifier la probabilité d'un événement clinique, d'une récidive ou d'une progression de la maladie chez les patients atteints de la maladie ou de l'état médical d'intérêt. |
| **Prédiction** | Identifier les individus les plus à même de connaître un effet positif/négatif lors de l'exposition à un médicament ou agent environnemental. |
| **Réponse** | à la prise d'un médicament, pour évaluer l'efficacité de ce dernier. |

Tableau 1.1 – Définition des principaux types de biomarqueurs

Classiquement, on espère d'un biomarqueur qu'il caractérise systématiquement correctement un état et qu'il soit facilement accessible. Un « bon » biomarqueur va donc contribuer à la prise de décisions médicales qui permettront d'améliorer les conditions physiologiques en identifiant de manière fiable l'état actuel/futur du patient. Ainsi, un biomarqueur est-il

évalué en fonction de sa sensibilité (c'est-à-dire la capacité du biomarqueur à identifier les individus malades comme effectivement malades, aussi appelés vrais positifs) et de sa spécificité (c'est-à-dire la capacité à identifier les individus sains comme effectivement sains, aussi appelés vrais négatifs). Le biomarqueur « parfait » est celui qui atteint 100% de sensibilité et 100% de spécificité. Mais dans les faits, les biomarqueurs sont évidemment imparfaits. La difficulté réside donc dans le choix de l'équilibre entre sensibilité/spécificité que l'on souhaite obtenir pour une pathologie donnée. Cela nécessite de définir quels cas sont les plus souhaitables et ceux qui le sont le moins. Imaginons l'exemple d'une pathologie devenant mortelle en cas de stade avancé mais pour laquelle il existe des traitements efficaces qui permettent d'empêcher la progression vers le stade fatidique. Dans un tel cas, il faut éviter que le biomarqueur de diagnostic utilisé n'identifie pas correctement des patients comme malades alors qu'ils le sont (mauvaise sensibilité). Cela signifierait que ces patients « faux négatifs » ne seraient pas pris en charge à temps.

Une étude dont la recherche porte sur les biomarqueurs peut se résumer ainsi : *i)* Définir un plan d'expérience pertinent et collecter des échantillons en conséquence ; *ii)* analyser les échantillons par des techniques à haut-débit ; *iii)* utiliser des méthodes informatiques et statistiques afin d'identifier un ensemble de biomarqueurs potentiels. Classiquement en protéomique clinique, la recherche de nouveaux biomarqueurs suit une logique très précise et structurée en plusieurs phases [5, 28] :

**Phase exploratoire** (ou de découverte) : Identification de nouveaux biomarqueurs candidats (plusieurs centaines) avec de petites cohortes d'échantillons (quelques dizaines). L'objectif principal de cette phase est d'identifier les protéines les plus à même de réussir les phases suivantes. Des analyses statistiques permettent de déterminer les candidats qui expliquent le mieux le phénomène étudié. Il ne s'agit pas de se contenter d'une simple corrélation entre l'expression d'un candidat et la pathologie, mais d'établir une inférence statistique fiable et reproductible des résultats.

**Phase de vérification** : les biomarqueurs candidats identifiés et retenus (plusieurs dizaines) lors de la phase exploratoire sont évalués dans des études ayant des cohortes plus grandes (environ une centaine d'échantillons). L'objectif de cette étape est de déterminer si la différence d'expression des candidats entre les différentes conditions étudiées montre de potentielles applications cliniques.

**Phase de validation** : Elle comporte une cohorte d'échantillon beaucoup plus grande, de l'ordre de plusieurs centaines. Le nombre de biomarqueurs candidats se résume

le plus souvent à moins d'une dizaine de candidats. L'objectif ici est d'évaluer les performances des biomarqueurs retenus, et de trouver une signature protéique (c'est-à-dire un ensemble de protéines) qui va précisément décrire la pathologie d'intérêt.

**Phase d'évaluation clinique** : Le nombre de biomarqueurs candidats devient très faible, pour des milliers d'échantillons. L'objectif est de déterminer la précision avec laquelle les biomarqueurs sélectionnés permettent de catégoriser les différents états physiologiques d'intérêt.

Mes travaux se focalisent sur la phase la plus préliminaire qu'est la phase exploratoire.

### 1.3.2 Les atouts et les contraintes de la stratégie adoptée

La stratégie adoptée pour le projet LIFE se confronte à des contraintes inhérentes à la recherche de biomarqueurs et à la discipline de la protéomique. Mais elle dispose aussi de leurs atouts. Même si les choix stratégiques ne font pas partie de mon travail de thèse, il me semble important de mentionner les avantages et inconvénients, car ils impactent indirectement l'interprétation des résultats statistiques et les conclusions biologiques.

La recherche de biomarqueurs plasmatiques présente plusieurs avantages : Tout d'abord, le plasma possède un rôle central dans le sang et plus globalement dans la physiologie humaine. Les protéines du plasma peuvent être réparties en plusieurs catégories [5] : *i)* Les protéines ayant un rôle fonctionnel du sang ; *ii)* celles résultant de fuites tissulaires et *iii)* celles jouant un rôle de signalisation. Elles peuvent donc constituer un reflet relativement complet de l'état physiologique d'un patient (ou plus largement de son phénotype). Ensuite, les protéines plasmatiques étant sécrétées par le foie, elles constituent un reflet relativement complet de l'état physiologique pour les pathologies hépatiques. Elles constituent donc des cibles particulièrement intéressantes pour comprendre la NAFLD. Enfin, la prise de sang constitue à l'heure actuelle le moyen le plus facile (rapide et à moindre coût) et le plus utilisé en médecine pour accéder à des protéines de l'organisme. Par conséquent, rechercher des biomarqueurs parmi les protéines circulant dans le plasma est une stratégie particulièrement répandue, aussi bien pour le diagnostic que le pronostic de nombreuses pathologies [29]. Néanmoins, il est possible que les protéines les plus intéressantes soient les plus faiblement exprimées et donc les plus difficilement détectables. De plus, il est tout à fait envisageable que pour une pathologie donnée, les protéines caractéristiques ne soient jamais sécrétées dans le sang et que donc les protéines plasmatiques ne soient pas

les plus pertinentes pour comprendre la maladie.

L'utilisation de la protéomique pour la recherche de biomarqueurs présente elle aussi plusieurs avantages : Tout d'abord, la MS permet de mesurer très précisément la masse des peptides issus des protéines de l'échantillon étudié. C'est donc une méthode très spécifique et qui permet d'être extrêmement précis dans l'identification des protéines présentes dans un échantillon. Ensuite, la MS est une technologie à « haut-débit » car elle permet d'analyser une grande quantité d'échantillons en relativement peu de temps par rapport à d'autres techniques, et cela nous permet d'espérer une recherche préliminaire à « large spectre ». Cependant, le choix d'utiliser la protéomique présente plusieurs contraintes. La MS possède une certaine capacité à détecter des protéines plus ou moins abondantes. On parle de profondeur et de couverture du protéome. Or le plasma est extrêmement complexe, et son analyse précise et complète reste un défi [30]. En effet il possède une plage extrêmement large de concentrations des protéines [5, 31, 32], certaines protéines sont très abondantes, comme l'albumine et les immunoglobulines, et d'autres protéines sont présentes en très faibles quantités ou bien exprimées uniquement en cas de besoin, telles que les cytokines et les hormones. Un des moyens de réduire cette complexité est de réaliser une déplétion des protéines les plus abondantes, c'est-à-dire que l'on va les retirer de l'échantillon à analyser. La concentration de quelques protéines connues du plasma humain est représentée dans la Figure 1.3 avec les protéines ordonnées de la plus abondante (à gauche) à la moins abondante (à droite) et est inspirée du travail de Geyer et al. [5]. Plus un appareil est capable de détecter des protéines à très faible concentration, plus l'analyse est profonde et la couverture, complètes. A l'heure actuelle, la couverture et la profondeur d'analyse peuvent être limitantes et les améliorer reste un objectif des développeurs d'instruments. De plus, les protéines très abondantes ont tendance à « masquer » les protéines faiblement abondantes (qui pourraient être intéressantes, mais qui ne seront alors malheureusement pas détectées). Enfin, la capacité à trouver des protéines qui seront effectivement cliniquement valides en aval de toute la recherche exploratoire constitue une limite à part entière de l'approche protéomique.

Finalement, ce projet (consistant à rechercher grâce à la protéomique des biomarqueurs de la NAFLD dans le plasma sanguin) porte une part de risque, comme de réelles opportunités, dont la gestion ne relève pas de mon rôle de doctorante. Néanmoins, il m'a semblé important de les identifier car ce projet de recherche bénéficie directement de tout ce que j'ai développé durant ma thèse. Notamment, même si j'ai conscience que l'approche méthodologique que j'ai développée ne pourra pas entièrement compenser les risques liés

aux choix stratégiques ou aux limites instrumentales, ils ont vocation à les limiter et ainsi avoir au final une influence sur le projet global (concrètement, la qualité des biomarqueurs qui seront finalement proposés, une fois le projet complètement terminé).

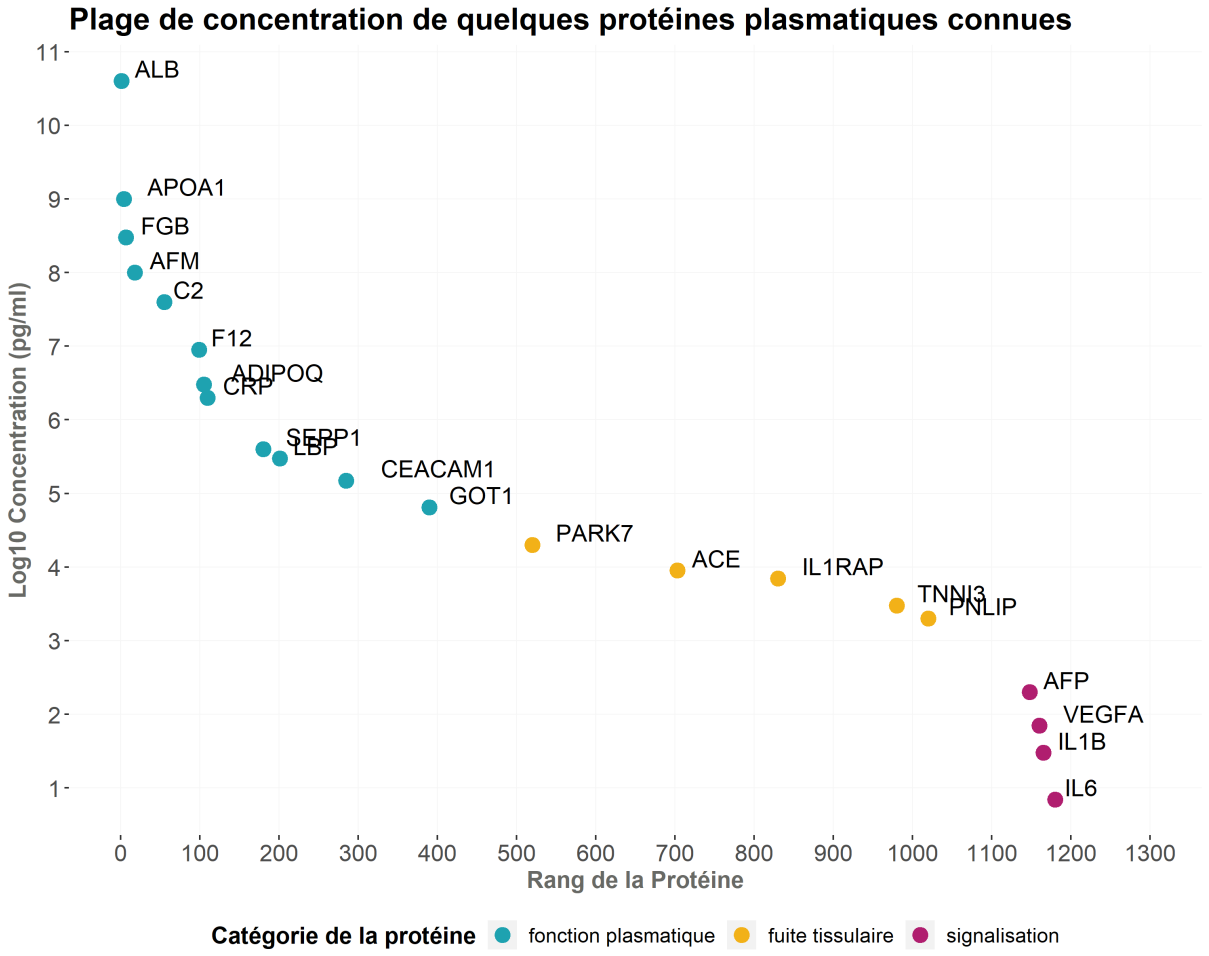



FIGURE 1.3 – Plage de concentration de quelques protéines plasmatiques connues. Adapté de Geyer *et al.* (2017) [5]. Les concentrations correspondent à des mesures réalisées par diverses méthodes à partir du sérum ou du plasma dans une situation physiologique normale. Les protéines sont représentées par leur nom de gène. L'ordre de grandeur varie de quelques pg/ml pour l'Interleukine 6 (IL6) à plus de 50 mg/ml pour l'albumine (ALB). Le rang correspond au classement de chaque protéine au sein de toutes les protéines plasmatiques connues en fonction de leur concentration. Ainsi, IL6 est une protéine ayant une des plus faibles concentrations plasmatiques connues et ALB possède la plus grande concentration (rang 1). Les protéines ayant une fonction plasmatique (en bleu) sont logiquement très abondantes dans le plasma. Celles ayant un rôle dans les fuites tissulaires (en jaune) ou la signalisation (en rouge) et qui peuvent potentiellement jouer un rôle dans une maladie, sont généralement beaucoup moins abondantes.

# Chapitre 2

# Production des données de protéomique quantitative

L'objectif premier de ce chapitre est de présenter l'approche protéomique qui est utilisée au sein de notre équipe, pour permettre l'identification et la quantification des protéines présentes dans un échantillon biologique. Il s'agit d'une approche basée sur l'utilisation couplée d'une chromatographie liquide (LC), d'un spectromètre de masse en mode tandem (MS/MS) et d'un ensemble d'outils bio-informatiques. Le second objectif est d'apporter une meilleure compréhension des contraintes techniques de la protéomique présentés dans le chapitre précédent. Cela permettra notamment au lecteur de mieux appréhender le type de données pour lequel nous avons réalisé les développements méthodologiques décrits dans la suite du manuscrit.

## 2.1 Présentation de l'approche *bottom-up*

### 2.1.1 Principe général

La protéomique basée sur l'utilisation de la MS peut être subdivisée en deux stratégies bien distinctes : la stratégie *top-down* et la stratégie *bottom-up* [20]. La stratégie *top-down* consiste à analyser les protéines entières, sans digestion préalable, directement dans le spectromètre de masse. Cette méthode est intéressante lorsque l'on souhaite par exemple étudier les modifications post-traductionnelles de certaines protéines, car les traitements préalables sont minimisés, contrairement à la seconde stratégie. Cependant, elle devient

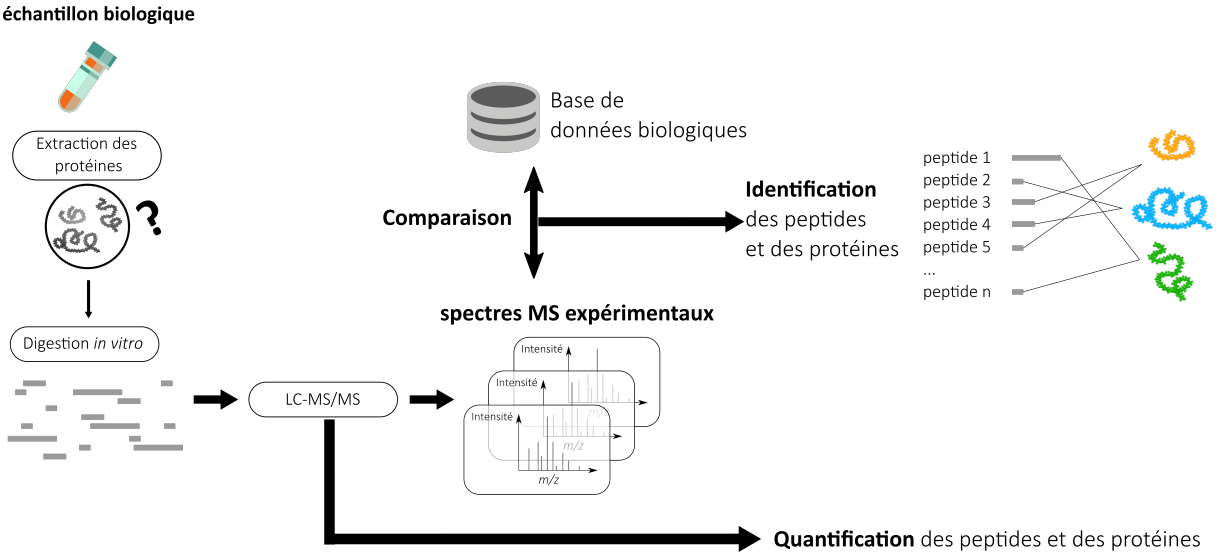



FIGURE 2.1 – Représentation schématique de la stratégie *bottom-up* en protéomique quantitative. Les protéines sont d'abord extraites de l'échantillon biologique, dont nous ignorons la composition exacte. Ces protéines inconnues sont digérées grâce à une protéase (la trypsine), donnant des peptides. Ces peptides sont ensuite analysés par LC-MS/MS pour générer des spectres de masse qui sont ensuite comparé à des bases de données afin d'identifier les peptides, et par inférence, les protéines.

extrêmement difficile à mettre en œuvre lorsqu'il s'agit d'explorer le protéome d'échantillons biologiques complexes [33]. La stratégie *bottom-up* consiste à identifier et quantifier les protéines à partir des peptides qui les constituent. Pour ce faire, les protéines sont digérées *in vitro*. Les peptides ainsi obtenus sont ensuite séparés par chromatographie liquide (LC), puis analysés par spectrométrie de masse en tandem (MS/MS), comme cela est illustré sur la Figure 2.1, puis détaillé dans la Section 2.2. Enfin, les peptides sont identifiés et quantifiés par des outils bio-informatiques (cf. Section 2.3), puis les protéines sont inférées à partir de ces peptides caractérisés. L'approche *bottom-up* fournit donc une caractérisation indirecte des protéines. Elle est particulièrement adaptée lorsque l'on souhaite explorer le plus largement possible les protéines présentes dans l'échantillon étudié (protéomique de découverte), car elle ne nécessite pas de connaître la composition en protéines de l'échantillon. Il s'agit également d'une approche plus économique, avec des protocoles de préparation plus simples à mettre en place que la stratégie *top-down* [20]. L'approche *bottom-up* est principalement utilisée au laboratoire, et c'est celle mise en uvre pour l'analyse de la cohorte NAFLD, de sorte que seule celle-ci est détaillée dans les sections suivantes.

### 2.1.2 Préparation des échantillons

La première étape d'une analyse protéomique *bottom-up* est la préparation des échantillons, afin qu'ils soient analysables par spectrométrie de masse (Figure 2.1). Cela consiste d'abord en une extraction des protéines contenues dans l'échantillon biologique étudié, puis en une séparation des protéines et enfin en la digestion de celles-ci afin d'obtenir des peptides. La digestion peut être subdivisée en plusieurs sous-étapes : *i)* la dénaturation, qui consiste à éliminer les repliements des protéines afin de faciliter l'accès à la séquence en acides aminés ; *ii)* la réduction qui permet la rupture des ponts disulfures ; *iii)* l'étape d'alkylation empêche la reformation des ponts disulfures grâce à des agents alkylants qui se fixent sur les résidus souffrés ; et enfin *iv)* la digestion qui est généralement réalisée avec de la trypsine, en raison de son fonctionnement bien particulier : elle coupe les séquences d'AA pratiquement uniquement après deux AA particuliers (la lysine, symbolisée K, ou l'arginine, symbolisée R). L'avantage de ce découpage spécifique est lié à la fréquence d'apparition et la distribution de ces deux AA. D'une part, la trypsine sectionne la séquence d'une protéine relativement souvent et de manière assez régulière. D'autre part, elle génère des peptides ayant une longueur idéale pour être « observables » avec la MS, car ni trop courts ni trop longs. Le second avantage de la digestion par la trypsine est que K et R sont les seuls AA chargés positivement, ce qui facilite l'étape d'ionisation (voir la section 2.2.2) car il suffit d'avoir une solution à pH acide.

## 2.2 L'étape de chimie analytique

### 2.2.1 La Chromatographie en phase Liquide

La LC permet de séparer les peptides en fonction de leurs propriétés physico-chimiques, notamment leur hydrophobicité. Il existe de nombreuses façons de réaliser une chromatographie en phase liquide, mais nous nous concentrons uniquement sur la méthodologie employée dans notre équipe.

Le principe général de la LC consiste à faire passer une phase mobile à l'état liquide au travers d'une colonne chromatographique contenant une phase stationnaire, solide et poreuse. La phase mobile est constituée de deux solutions : une première solution A (mélange d'eau et d'acide formique) et une seconde nommée B (mélange d'acétonitrile ($CH_3CN$) et d'acide formique). La phase stationnaire correspond à la colonne et est généralement constituée de chaînes carbonées qui vont influer sur l'interaction hydrophobe avec les pep-

tides. Plus les chaînes carbonées sont longues, plus l'interaction sera forte [34]. On injecte d'abord de la solution A avec l'échantillon contenant les peptides grâce à l'injecteur afin que les peptides interagissent avec la colonne et se fixent dessus. Puis une fois les peptides fixés dans la colonne, la solution B est progressivement mélangée à la solution A, augmentant ainsi la proportion de solvant organique dans la phase mobile (Figure 2.2). Ce changement progressif de solvant est appelé *gradient*. Ainsi au cours du temps d'élution, plus la solution B est présente, et plus les peptides hydrophobes vont pouvoir se détacher de la colonne et en sortir pour être analysés par la MS en aval. Ce sont donc les peptides les plus hydrophiles qui sont élués en premier, jusqu'aux plus hydrophobes qui seront élués en dernier.

Originellement, la phase mobile passait au travers de la phase solide par gravité, mais un système à haute pression, nommé « *High Performance Liquid Chromatography* » (HPLC) fut développé afin de gagner en débit et en précision de séparation. La technologie *Nano-LC-MS* utilise un capillaire d'environ 50-100 micromètres pour l'injection dans le spectromètre de masse, ce qui permet d'utiliser une quantité moindre d'échantillon pour obtenir la même précision de mesure.

### 2.2.2 La spectrométrie de masse

#### 2.2.2.1 Principe

La spectrométrie de masse consiste à identifier des molécules en mesurant précisément leur masse et en décrivant leur structure chimique. Plus formellement, le principe consiste en une séparation en phase gazeuse de molécules chargées (les ions) en fonction du rapport masse/charge ($m/z$). On obtient alors des spectres présentant des pics d'intensité des différents fragments détectés en fonction de leur rapport $m/z$, qui sont ensuite analysés pour identifier les peptides qui ont été mesurés.

#### 2.2.2.2 Structure d'un spectromètre de masse

Un spectromètre de masse peut être schématiquement décomposé en trois parties : une source d'ionisation, un analyseur, et un détecteur (Figure 2.3). L'étape d'ionisation des peptides est nécessaire car les spectromètres utilisent des champs électromagnétiques pour séparer les fragments et mesurer précisément leur masse. Il existe une multitude de technologies pour chaque partie du spectromètre [35], dont les plus utilisées aujourd'hui sont résumées dans la Figure 2.3. Mais nous ne détaillons que celles employées par notre équipe.

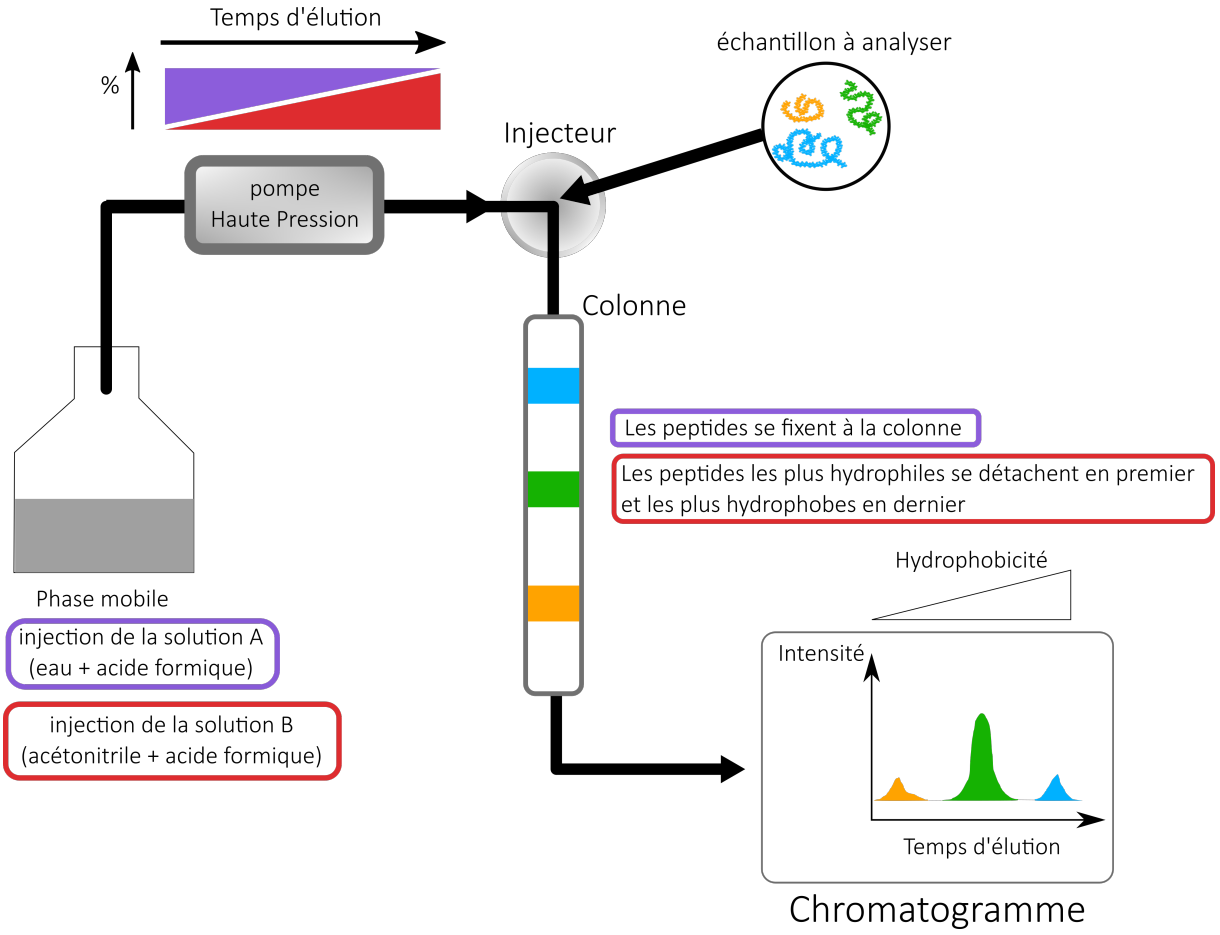



FIGURE 2.2 – Principe de l'HPLC. Dans un premier temps l'échantillon est injecté dans un solvant hydrophile (majoritairement A) afin que les peptides se fixent à la colonne plus hydrophobe. Puis le solvant devient graduellement plus hydrophobe, en augmentant la part de solvant B, entraînant alors les peptides en fonction de leur hydrophobicité. Dans cet exemple, le peptide jaune est le plus hydrophile et le bleu le plus hydrophobe.

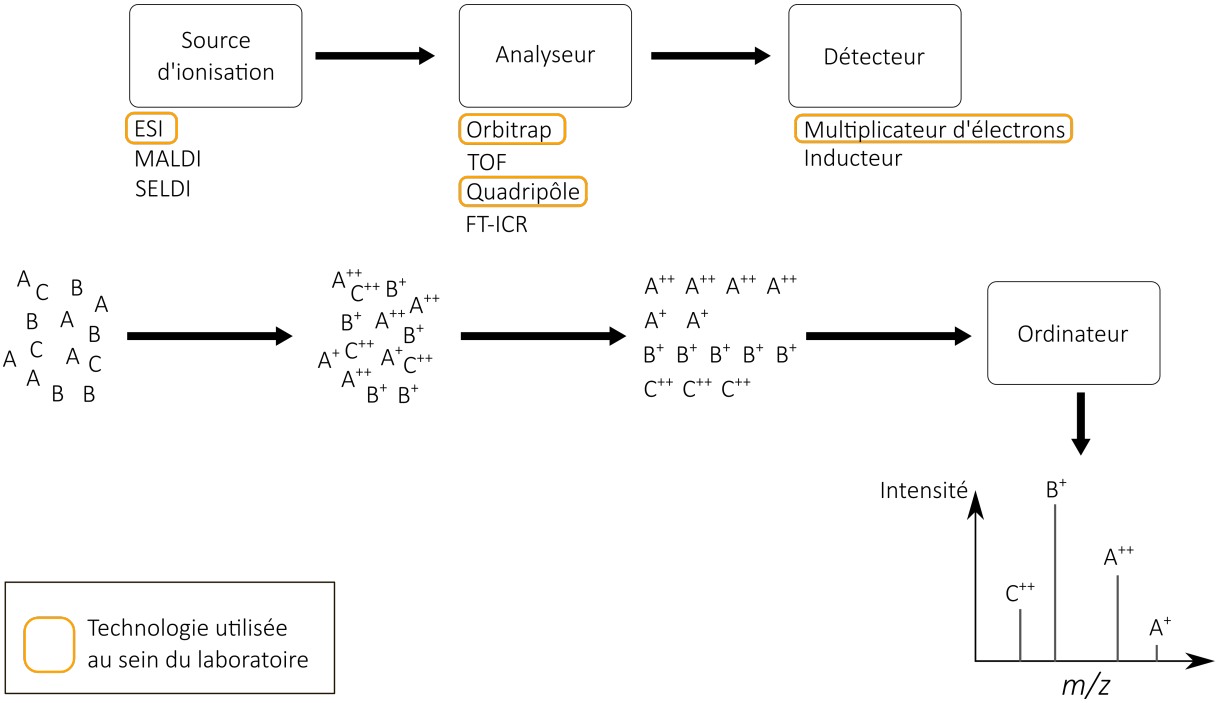



FIGURE 2.3 – Vue d'ensemble schématique d'un spectromètre de masse. Cet appareil peut être divisé en trois principaux segments : la source d'ionisation, l'analyseur et le détecteur. Pour chaque élément constitutif du spectromètre sont listés les technologies les plus répandues. Les technologies utilisées au laboratoire sont entourées en jaune.

L'étape d'ionisation consiste en l'ajout de protons ($H^+$) aux peptides à mesurer. Il existe plusieurs technologies permettant de réaliser cette étape, notamment l'ESI [17] et le MALDI [36]. Celle qui est utilisée dans notre laboratoire est l'ESI. Avec cette méthode, les peptides (ou plus généralement les analytes) sont amenés dans la source d'ionisation par l'intermédiaire d'un fluide à pression atmosphérique. Ce liquide est ensuite pulvérisé à partir d'un capillaire métallique dans un champ électromagnétique puissant, créant ainsi un nuage de gouttelettes chargées qui se retrouve orienté vers l'entrée de l'analyseur du spectromètre (Figure 2.4).

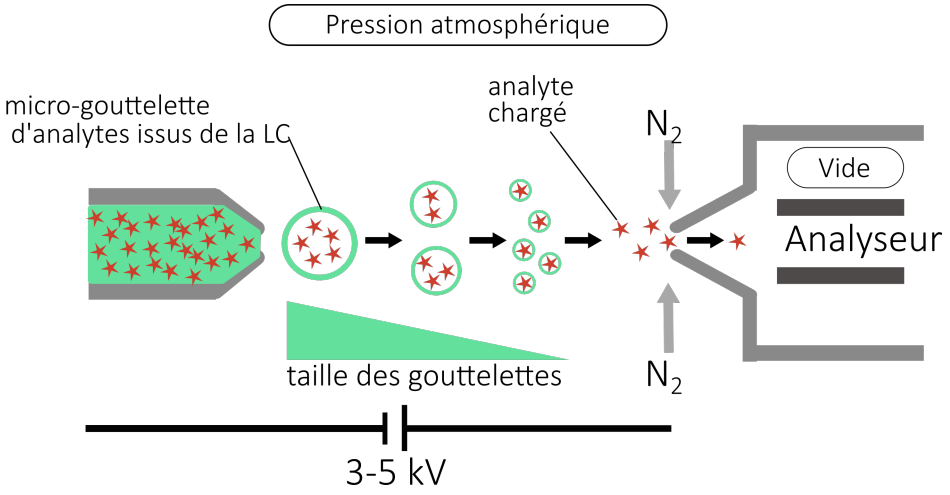



FIGURE 2.4 – Principe général de l'ESI. Un capillaire métallique étroit est maintenu sous haute tension (3-5 kilovolt) et une différence de potentiel est produite entre le capillaire et l'orifice du spectromètre de masse. Une solution contenant des ions d'analytes est pulvérisée à partir du capillaire. Un flux de gaz inerte ($N_2$) est également utilisé pour nébuliser le flux de liquide en microgouttelettes. Ce gaz oriente la dispersion des microgouttelettes vers l'analyseur du spectromètre de masse.

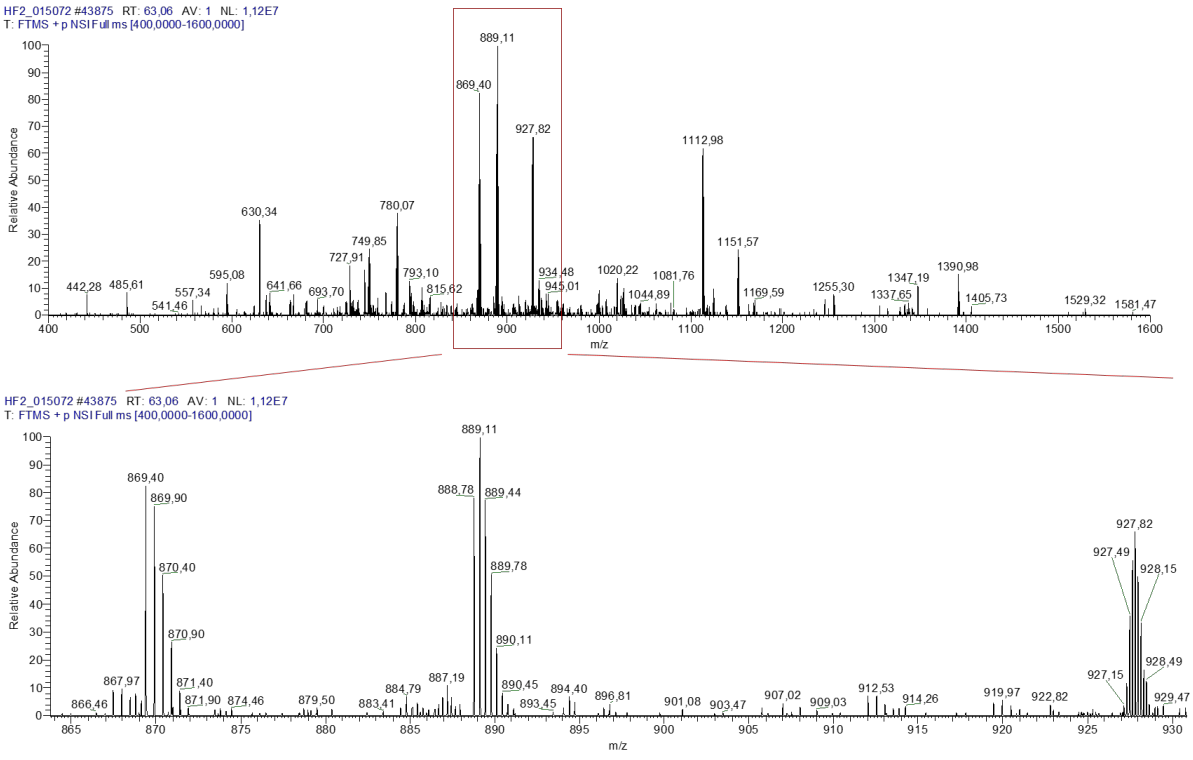



FIGURE 2.5 – Exemple d'un spectre MS1 (en haut). Un zoom sur l'intervalle $m/z$ [865-930] permet de voir la décomposition en différentes charges : des pics possédant une différence de $m/z$ de 1, correspond à une différence de charge de 1.

Une fois les fragments ionisés (chargés), ils sont transférés dans l'analyseur. Celui-ci va alors séparer les composés en fonction de leur rapport $m/z$. Notre laboratoire utilise un Orbitrap [37]. L'analyseur est l'élément clé d'un spectromètre de masse, puisqu'il mesure le rapport $m/z$. Il détermine les performances en termes de précision, de résolution et de sensibilité de l'instrument. L'Orbitrap s'avère être un analyseur particulièrement performant [38]. Il est composé d'une électrode externe et d'une électrode centrale. Lorsque les ions entrent dans l'Orbitrap, ils sont capturés par le champ électrostatique et vont osciller le long de l'axe de l'électrode centrale (ils sont alors en orbite autour de l'électrode, d'où le nom d'Orbitrap) avec une fréquence qui leur est propre car elle dépend du $m/z$ de chaque ion. Le détecteur va enregistrer ces oscillations et les transmettre à un calculateur qui va générer un spectre de masse $m/z$ grâce à une transformation de Fourier. Pour plus d'informations sur cette technologie, se référer aux travaux de Hu et al. (2005) [37]. Le spectre de masse obtenu est un diagramme composé du rapport $m/z$ sur l'axe des abscisses, et l'intensité du signal sur l'axe des ordonnées (voir l'exemple sur la Figure 2.5).

### 2.2.3 La LC combinée à la spectrométrie de masse en tandem

#### 2.2.3.1 Principe général

L'approche dans son ensemble consiste à combiner HPLC et spectrométrie de masse. Le couplage consiste donc à séparer dans un premier temps les peptides obtenus par digestion enzymatique des protéines puis à les ioniser et les injecter progressivement dans le spectromètre. Cependant, le spectromètre est utilisé en mode tandem. Cela signifie que les peptides de l'échantillon sont analysés une première fois pour générer ce que l'on appelle un spectre MS1. Ensuite, les peptides détectés à l'issue de l'étape de MS1 sont fragmentés en sous-unités ionisées. Enfin, ces ions sont eux aussi analysés pour produire ce que l'on appelle des spectres MS2 (ou des spectres de fragmentation, ou des spectres MS/MS) qui permettent l'identification des peptides. La fragmentation du peptide ionisé consiste à briser les liaisons amides entre les acides aminés ionisés. Les ions produits ne peuvent être détectés que s'ils possèdent au moins une charge. Si la charge est portée au niveau de la partie N-terminale de l'AA, l'ion est classé en catégorie $a$, $b$ ou $c$. Si la charge est portée sur la partie C-terminale de l'AA, l'ion est classé en $x$, $y$ ou $z$. Chaque fragmentation donne naissance à deux espèces, une neutre et une autre chargée dans le cas d'un précurseur mono-chargé, et seule la forme chargée est détectée par le spectromètre de masse. La plupart des modes de fragmentation utilisés en protéomique produisent essentiellement des ions $b$ et $y$, telles que les méthodes de dissociation induite par collision

(« *Collision Induced Dissociation* » ou CID) et de dissociation par collision à haute énergie (« *High Collision Dissociation* » ou HCD). D'autres méthodes comme la dissociation par transfert d'électrons (« *Electron Transfert Dissociation* » ou ETD), produisent des ions $c$ et $z$. La Figure 2.6 présente un peptide constitué de 4 AA (adapté des travaux de Zhang et al. (2010) [38]) où le peptide est fragmenté en deux ions produits $b_2$ et $y_2$.

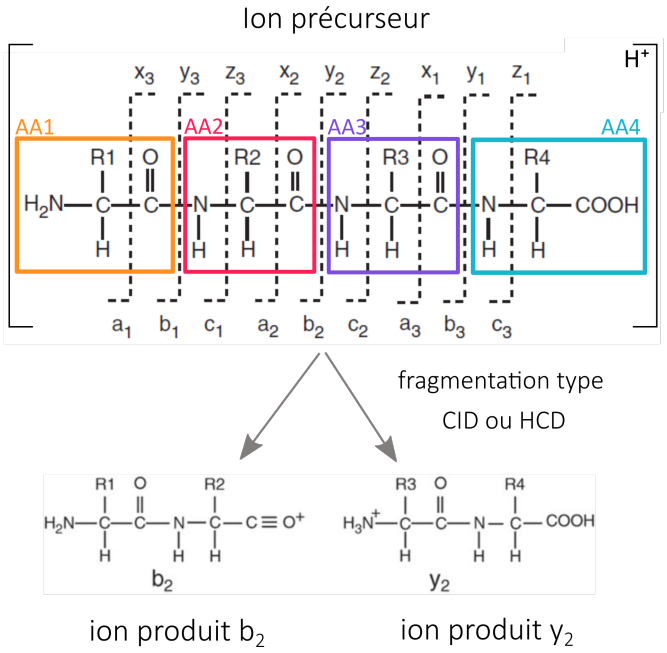



FIGURE 2.6 – Fragmentation CID ou HCD d'un ion précurseur qui donnera des ions de type $b$ et $z$. Seulement deux ions produits ($b_2$ ou $y_2$) sont présentés ici après fragmentation. Adapté des travaux de Zhang et al. (2010) [38].

Pour donner un ordre de grandeur de production de données, un spectromètre de masse moderne peut facilement produire 35 000 spectres par heure de fonctionnement et est souvent utilisé 24 heures sur 24. Le couplage LC-MS/MS permet d'analyser plusieurs milliers de peptides et de protéines avec des gradients de seulement quelques heures [39].

#### 2.2.3.2 Production de spectres : les différents modes d'acquisition

Il existe différents modes d'acquisition des données pour produire des spectres MS2 à partir des peptides détectés lors de l'analyse MS1. Nous décrivons ici les deux approches utilisées couramment au laboratoire et qui ont été employées pour produire les données de la cohorte NAFLD. La principale différence entre les deux est la manière dont sont sélectionnés les peptides qui seront fragmentés pour obtenir l'analyse MS2.

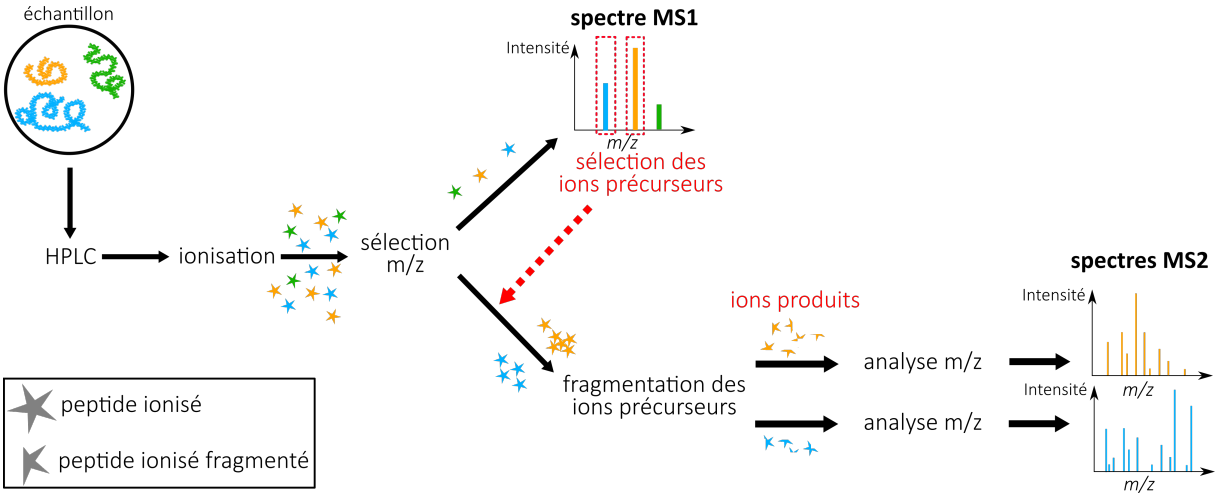



FIGURE 2.7 – Principe général de l'approche par DDA. Dans cet exemple, seuls les pics les plus intenses sont sélectionnés (le jaune et le bleu). Ainsi, seuls les peptides jaunes et bleus détectés par la suite seront fragmentés et analysés sur une période donnée.

La première approche est l'acquisition dépendante des données (« *Data Dependent Acquisition* » ou DDA). La première étape permet de générer les spectres de masse (appelés MS1) des peptides dans une gamme de masses définie par l'analyticien (généralement, entre 400 et 1600 $m/z$). Sur ces spectres, il est possible d'identifier automatiquement les valeurs de $m/z$ pour lesquelles les intensités (la hauteur des pics) sont les plus élevées, et d'isoler les peptides ionisés correspondant. Par exemple, ne seront sélectionnés que les 10 à 20 peptides les plus intenses d'un spectre MS1. Ces peptides sélectionnés sont appelés ions précurseurs (voir Figure 2.7). Ensuite, ces ions sont fragmentés (en ions produits) pour générer des spectres MS2. Ce spectre MS2 permet de reconstituer l'ordre de la séquence du peptide analysé (Figure 2.8).

Un exemple très simplifié du principe DDA est présenté dans la Figure 2.9 : Une analyse MS détecte la présence de 6 peptides (A, B, C, D, E et F). L'appareil sélectionne les pics les plus intenses pour la poursuite de l'analyse. Admettons que seulement les trois peptides les plus intenses sont sélectionnés (A, B et C). Il y aura une analyse MS/MS exclusivement pour les ions issus de la fragmentation du peptide A, puis une autre analyse MS/MS pour B et une analyse MS/MS pour C. Une fois ces trois analyses réalisées, une nouvelle analyse MS1 est faite. L'intervalle de temps nécessaire pour chaque analyse se situe généralement entre 0,1 et 8 secondes. A titre d'exemple, les appareils utilisés par notre équipe ont une vitesse d'acquisition de 18 Hz, c'est-à-dire que 18 spectres MS2 sont générés par seconde, d'où le fait que seulement les 10 à 20 pics les plus intenses sont sélectionnés. On comprend alors que la principale limite du mode DDA est le sous-

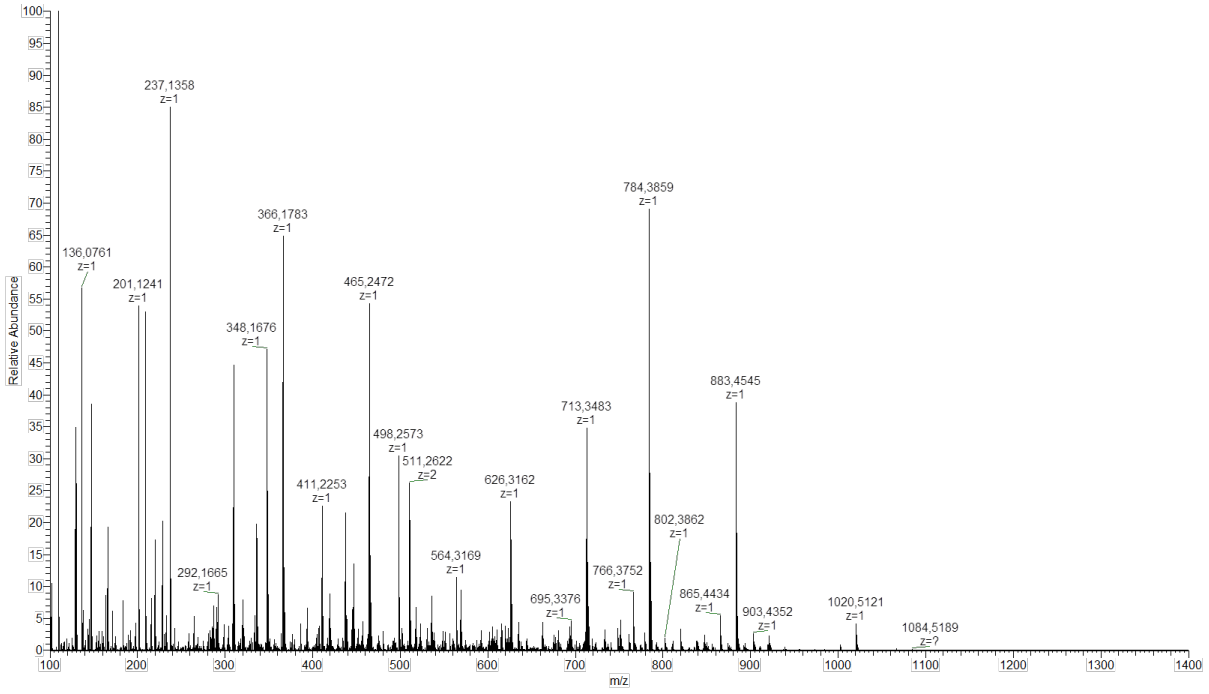



FIGURE 2.8 – Exemple de spectre MS2 issu de la fragmentation d'un peptide ayant un rapport $m/z$ de 449.

échantillonnage, car tous les peptides ne seront pas analysés. Seul une partie des peptides les plus intenses sera sélectionnée pour l'analyse MS2.

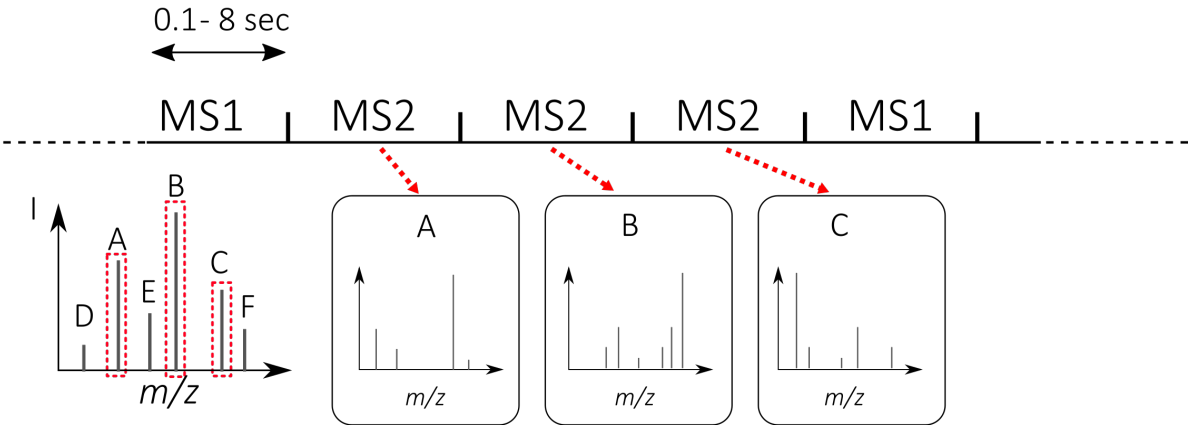



FIGURE 2.9 – Schéma décrivant l'enchaînement des analyses MS1 et MS2 dans le temps pour l'approche DDA. Une analyse MS1 est réalisée en 0,1 à 8 secondes. Puis il y a des analyses MS2 successives pour chaque ion précurseur sélectionné. Le désavantage est que l'on n'analyse qu'un sous-ensemble des peptides détectés lors de la MS1. L'avantage de cette approche est de produire des spectres MS2 simples à interpréter, car un spectre MS2 correspond à la décomposition d'un peptide donné.

La deuxième approche est l'acquisition indépendante des données (« *Data Independent Acquisition* » ou DIA). Elle s'est développée au début des années 2000 [40]. Il existe désormais une multitude d'implémentations [41] mais le principe général est d'échantillonner de manière exhaustive et répétée chaque peptide, produisant un ensemble complexe de spectres de masse (Figure 2.10).

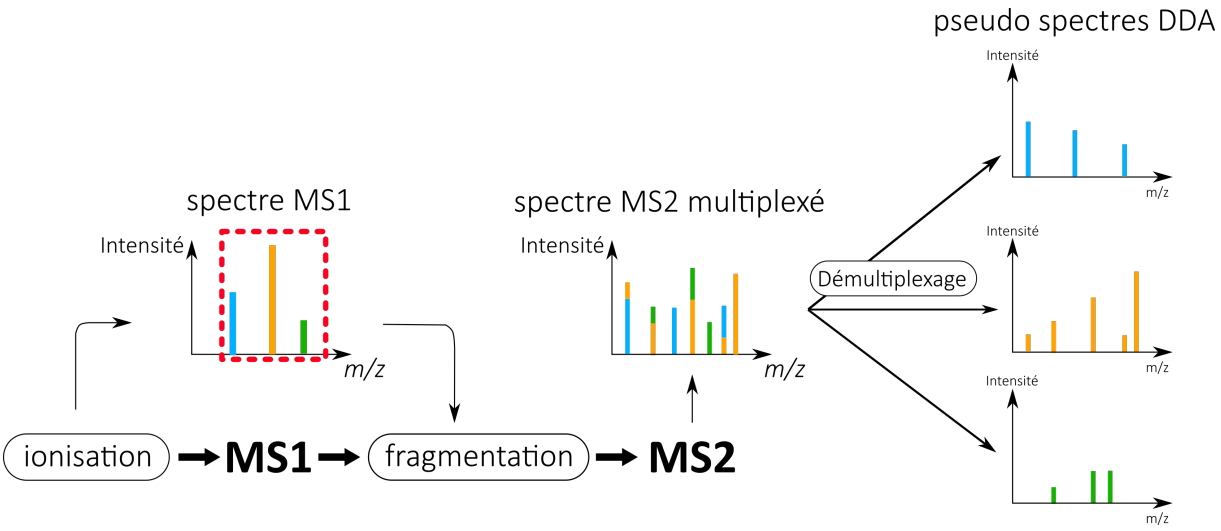



FIGURE 2.10 – Schéma du principe général de la DIA. L'avantage est que l'on analyse l'ensemble des peptides détecté lors de la MS1. Il s'agit donc d'une analyse plus exhaustive de l'échantillon biologique que par l'approche DDA. Ce processus a le désavantage de produire des spectres MS2 multiplexés difficile à interpréter, et qui demande une étape délicate de démultiplexage afin d'obtenir des spectres spécifiques pour chaque peptide.

Contrairement à la DDA qui produit des spectres MS1 et MS2 en tandem via l'isolement et la fragmentation de précurseurs de peptides spécifiques, les méthodes DIA parcourent des fenêtres de masse dans lesquelles tous les peptides d'une fenêtre sont fragmentés. Cette approche a l'avantage de l'exhaustivité (le biais dû à la sélection des ions précurseurs est supprimé, surmontant ainsi les limites de la non-reproductibilité et du sous-échantillonnage inhérentes à la DDA). Malheureusement cette approche conduit à la production de spectres MS2 dits multiplexés, c'est-à-dire qu'ils ne contiennent pas l'information de séquence de seulement un peptide donné, mais de plusieurs peptides. La principale difficulté en DIA est alors l'interprétation de ces spectres multiplexés. Plusieurs méthodes de démultiplexage ont vu le jour depuis les années 2010 et ont permis à la DIA de se démocratiser au sein de la communauté protéomique [42, 43, 44, 45, 46]. La plupart des méthodes utilisés en routine pour réaliser ce démultiplexage repose sur la mise en correspondance des spectres avec une bibliothèque de spectres DDA précédemment acquise.

Cependant, de plus en plus de méthodes sont basées sur une approche purement formelle, comme par exemples [47], qui ont notamment été développées au laboratoire.

## 2.3 L'analyse bioinformatique

### 2.3.1 L'identification des peptides

Les séquences en acides aminés des peptides détectés expérimentalement sont identifiées grâce aux spectres MS/MS (MS2) par différentes stratégies [38, 48]. Nous nous focalisons sur l'approche prédominante à l'heure actuelle, qui consiste à comparer les spectres expérimentaux à des spectres dit « théoriques » qui ont été générés à partir d'une base de données de séquences protéiques [49] par simulation numérique du processus physico-chimique d'analyse. L'idée est qu'un peptide possède une séquence (un enchaînement d'AA) qui lui est propre avec des fragments bien spécifiques, et que ces fragments sont caractérisés par des spectres MS/MS spécifiques. Les spectres peuvent donc être assimilés à des « empreintes » qui caractérisent le peptide, de la même façon que les empreintes digitales permettent de distinguer les individus.

Concrètement, il s'agit d'une procédure impliquant trois choix : *i)* celui de bases de données protéiques qui va servir de référence à l'identification des peptides ; *ii)* celui des filtres permettant de contrôler le volume des peptides théoriques à générer ; *iii)* celui du score quantifiant la similarité entre les spectres expérimentaux et les spectres théoriques.

Les bases de données actuelles étant très grandes (par exemple la *Universal Protein Knowledge Base* [50], abrégée en UniProtKB, `http://www.uniprot.org`), le nombre de spectres théoriques possibles est extrêmement grand. Un premier filtre consiste donc généralement à sélectionner les peptides de la base de données qui ont une masse théorique compatible avec celle de l'ion précurseur à identifier. Un second filtre classique consiste à choisir les PTMs que l'on souhaite observer.

Parmi les algorithmes les plus couramment utilisés pour évaluer le niveau de similarité entre un spectre et un peptide théorique, nous pouvons citer ceux ayant donnés naissance à l'outil du même nom : MASCOT [51], SEQUEST [52], X! Tandem [53] et OMSSA [54]. Plus le spectre expérimental ressemble au spectre théorique, plus le score de similarité est élevé. On comprend alors que l'efficacité de cette approche dépend de la qualité des spectres expérimentaux obtenus. Pour une revue plus détaillée des principaux algorithmes d'identification et notamment les calculs de score, se reporter à [55]. Les peptides candidats

associés aux spectres théoriques sont classés et filtrés pour créer des correspondances de spectre peptidique, que l'on appelle aussi PSM (« *Peptide-to-Spectrum Match* »). Les meilleures correspondances, c'est-à-dire les PSM ayant les scores de similarité supérieurs à une valeur seuil choisie, définissent alors les peptides identifiés. Ce seuil est généralement défini en fonction de critère statistiques [56] dont la description nécessite pour prérequis un certain nombre de notions décrites dans le Chapitre 3.

### 2.3.2 La quantification *label-free* des peptides

En protéomique, la quantification des peptides (et ensuite des protéines) peut être subdivisée en deux grandes catégories [57] : la quantification absolue [23, 58] et la quantification relative. C'est la seconde qui est la plus utilisée à l'heure actuelle au laboratoire, et c'est celle qui a permis d'obtenir les données quantitatives pour la cohorte NAFLD.

La quantification absolue repose sur l'utilisation d'étalons dont les abondances sont prédéfinies et qui sont introduits dans les échantillons [23], et qui permettent de compenser la non-connaissance de la relation liant l'intensité du signal mesuré à la concentration chimique (relation qui est propre à chaque ion, et qui dépend de la composition de l'échantillon analysé). L'intensité du signal d'un étalon correspond alors à une abondance connue qui sert de référence pour les peptides à quantifier. En revanche, la quantification relative « *label-free* » n'utilise pas d'étalon interne et par conséquent il n'y a pas d'abondance connue pouvant servir de référence. La seule mesure directe qui peut être réalisée est la différence d'intensité d'un même peptide entre deux échantillons. La quantification est donc relative à l'échantillon [57]. Les techniques de quantification absolue imposent un coût supplémentaire conséquent pour garantir la pureté et la qualité des étalons. Le nombre limité d'étalons ne permet également pas une couverture exhaustive des peptides et donc des protéines contenues dans l'échantillon, contrairement à la quantification relative.

Parmi toutes les méthodes de quantification relative disponibles [19, 57, 59, 60], notre laboratoire utilise principalement la quantification « *label-free* » basée sur le calcul des courants ioniques extraits (« *eXtracted Ion Chromatogram* » ou XIC) qui permet d'obtenir une bonne couverture des protéines présentes dans un échantillon ainsi qu'une quantification relative assez fine, tout en étant applicable à de larges cohortes d'échantillons [19]. Elle consiste à extraire les pics chromatographiques de chaque peptide d'une analyse LC-MS/MS, puis à calculer l'aire sous la courbe correspondant au pic d'intensité sur un temps d'élution donné pour en déduire l'abondance du peptide. Bien que cette méthode soit actuellement la plus performante, elle n'est pas exempte de limitations :

La première limitation réside dans l'obtention de mesures reproductibles lors de l'analyse LC-MS [61]. En effet, bien que la stabilité des instruments LC et MS se soit considérablement améliorée ces dernières années, celle-ci n'est pas parfaite. A titre d'exemple, les acquisitions sont échantillonnées tout au long d'une élution continue, ce qui induit nécessairement des erreurs de mesures. Dès lors, un contrôle de qualité minutieux reste nécessaire pour garantir la comparaison des intensités des peptides entre les différents échantillons. Ce contrôle qualité se fait notamment par l'alignement des temps de rétention entre les analyses des différents échantillons, puis par différentes normalisations des intensités. La seconde limitation de la quantification est directement liée au pouvoir de détection des instruments. En effet, lorsqu'un peptide est présent en très faibles quantités, son signal d'intensité se confond avec la limite de détection inférieure de l'instrument. Par conséquent, le signal du peptide n'est pas interprétable et le peptide est considéré comme absent. Ce phénomène explique en partie la génération de valeurs manquantes (notés « N.A. ») pour des peptides dans certaines conditions. Plusieurs logiciels ont été développés pour la quantification des peptides [62], le plus connu étant MaxQuant [63, 64]. Cependant, notre équipe développe et utilise son propre outil, Proline [65] pour réaliser l'étape de quantification.

Nous comprendrons que le nombre important d'étapes réalisées lors de l'identification et la quantification des peptides ainsi que leur complexité, tend à produire une proportion non négligeable de valeurs manquantes, ce qui se traduit par une perte d'information biologique non souhaitable. Une stratégie populaire pour palier à ces valeurs manquantes est de réaliser une étape d'assignation croisée (AC). L'AC consiste à aligner l'ensemble des chromatogrammes analysés grâce au temps de rétention, puis d'assigner les identifications de peptides obtenues dans certaines analyses MS à des analyses pour lesquelles les peptides n'ont pas été identifiés mais dont les pics chromatographiques correspondent à des ratios masse sur charge, des temps de rétention et des états de charge similaires [65]. Cette AC permet donc d'augmenter le nombre de peptides pouvant être quantifiés dans chaque échantillon car il permet la quantification de peptides non identifiés par MS/MS dans certains échantillons. Ce procédé permet de diminuer le nombre de valeurs manquantes dans le tableau de données final. En effet, en l'absence d'AC, seuls les peptides identifiés par MS/MS seraient éligibles à la quantification dans chaque échantillon. Cependant l'AC n'est pas exempte de limitations. En effet, elle ne permet de compléter qu'une partie seulement des valeurs manquantes et elle peut s'avérer être conservative.

### 2.3.3 Passage des peptides aux protéines

Une fois un ensemble de peptides dont les identités ont été validées, il est possible de remonter aux protéines parentes, grâce aux mêmes bases de données que celles utilisées pour l'identification peptidique. Cependant cette tâche est en pratique complexe pour deux principales raisons :
Premièrement, les protéines caractérisées par des peptides spécifiques (c'est-à-dire propre à une unique protéine), ne sont pas systématiques : D'une part parce que la couverture complète d'une séquence ne peut quasiment jamais être obtenue, de sorte que l'identification d'une protéine repose sur une partie de sa séquence seulement (sans garantie que celle-ci soit spécifique). D'autre part, parce qu'il existe de nombreuses familles de protéines, au sein desquelles les membres partagent des séquences peptidiques similaires. Ainsi, après digestion enzymatique, elles produisent parfois les mêmes peptides, alors appelés peptides partagés [66]. Par conséquent, il est impossible d'être certain de l'origine des peptides partagés lors de l'agrégation des peptides en protéines, et plusieurs protéines identifiées sur la base du même ensemble de peptides sont signalées comme un groupe plutôt que des protéines individuelles.
Deuxièmement, l'utilisation de certaines bases de données, notamment comme TrEMBL [67], pose des problèmes de redondances : une même protéine peut être référencée avec plusieurs noms différents. Ainsi, un peptide correspondant à cette protéine va renvoyer un résultat comprenant plusieurs entrées (qui en réalité ne correspondent qu'à une seule protéine). Une possibilité pour éviter les redondances est d'utiliser des bases de données vérifiées et annotées telle que Swiss-Prot [67]. Malheureusement de telles bases de données contiennent en général beaucoup moins de protéines, et leur usage induit une moindre couverture du protéome analysé.

Concernant la quantification des protéines, celle-ci est le plus souvent réalisée en moyennant ou en sommant les intensités des peptides qui leur sont associées [68]. Bien que des méthodes plus élaborées aient été publiées [64, 69, 70], obtenir une quantification précise reste difficile pour deux raison.
La première est que le réel poids de chaque peptide attribué dans la part de l'intensité protéique est inconnu, ceci étant une cause directe de la nature relative de la quantification.
Une seconde difficulté surgit en présence de peptides partagés. En effet, comment attribuer l'abondance d'un peptide à une protéine, sachant qu'elle partage ce peptide avec d'autres protéines ? Il serait inapproprié de compter plusieurs fois les peptides partagés

| **Protéines** | **Condition** 1 | **Condition** 2 | **...** | **Condition** $k$ |
|---|---|---|---|---|
| **Prot** 1 | abondance | abondance | ... | abondance |
| **...** | ... | ... | ... | ... |
| **Prot** $N$ | abondance | abondance | ... | abondance |

TABLEAU 2.1 – Tableau représentant la structure des données après identification et quantification des protéines pour une étude où l'on souhaite comparer $k$ conditions biologiques (en colonne) et où l'on a identifié $N$ protéines (en ligne).

en les attribuant à plusieurs protéines distinctes, car cela consisterait à exacerber l'abondance des protéines en question. Une des approches consiste à attribuer à chaque protéine une quantité d'un peptide partagé qui soit proportionnelle aux peptides spécifiques de chaque protéine du groupe [64]. Autrement dit, une protéine qui possède une grande abondance d'expression de peptides spécifiques, se verra attribuer une plus grande quantité des peptides partagés par rapport aux autres protéines moins exprimées. Une autre manière plus simple de procéder consiste à ne pas utiliser les peptides partagés pour la quantification [68], avec tous les risques que cela comporte (notamment la perte de signal de quantification pour les protéines n'ayant que des peptides partagés).
Finalement, les étapes d'identification et de quantification des protéines permettent de générer des tableaux résumant les données comme présenté dans la table 2.1. Ce sont ces tableaux de quantification que nous explorons avec l'analyse statistique dans les chapitres suivants.

# Chapitre 3

# L'analyse statistique en protéomique quantitative

Le premier objectif de ce chapitre est de présenter un état de l'art sur l'utilisation et l'application des statistiques lors de l'analyse de données de protéomique quantitative. Le second objectif est de présenter les principaux concepts statistiques nécessaires à la compréhension des chapitres suivants.

## 3.1 Analyse de données de protéomique

### 3.1.1 L'absence de consensus en méthodologie statistique

Malgré une amélioration constante et progressive, l'analyse statistique reste parmi les sections les moins décrites et les moins justifiées dans les publications de protéomique (et de biologie en général). Bien souvent, seuls sont spécifiés le nom du test statistique utilisé et son seuil de significativité associés (le choix d'un seuil souvent égal à 5% ou 1%, dont une définition plus formelle est donnée en Section 3.3), ce qui rend difficile la reproduction des résultats, dans un contexte où il n'y a pas de consensus autour de la méthodologie d'analyse statistique. Des publications parues dans un même journal peuvent avoir des choix de seuil de significativité et de taux de fausses découvertes (voir Section 3.4.3) différents [71, 72], alors que le test statistique réalisé au départ est le même. Par ailleurs, 5%, 1% ou même 1‰peuvent tous être des seuils de significativité ou de taux de fausses découvertes valides, mais chacune de ces valeurs doit néanmoins être justifiée. Enfin, il existe une multitude d'outils dédiés au traitement des données de protéomique

quantitative, pour la plupart développés sur la base des bibliothèques statistiques de langages scientifiques comme R [73, 74, 75] ou Python [76]. Néanmoins, les plus répandus sont Perseus [77] et Excel, ce qui s'explique à la fois par des causes historiques et par l'importance des interfaces graphiques, car celles-ci permettent une utilisation conviviale et facile en l'absence de compétences en programmation scientifique.

### 3.1.2 Transposition de méthodes développées pour d'autres technologies à haut-débit

Actuellement en protéomique, même les outils statistiques dédiés sont largement inspirés de ceux issus d'autres domaines d'application proches, notamment la transcriptomique [78, 79]. Cela s'entend par la relative similarité entre ces deux types de données [23, 80, 81, 82, 83, 84]. En effet, les deux technologies mesurent l'abondance d'entités biologiques. D'un point de vue pratique, les données résultantes de ces deux approches sont représentées de la même façon, c'est-à-dire sous forme d'un tableau de valeurs numériques où les lignes représentent les entités biologiques mesurées et les colonnes représentent les différents échantillons analysés. Les quantifications au sein de chaque échantillon ont la plupart du temps une distribution log-normale [85]. Cela signifie qu'il y a un grand nombre d'observations de faible abondance, avec moins d'observations d'abondance élevée ; et que l'on retrouve une distribution gaussienne après transformation logarithmique des quantités mesurées. Dans les deux cas, le nombre de variables mesurées est beaucoup plus grand que le nombre d'échantillons, limitant de ce fait drastiquement la palette d'outils statistiques applicables. Pour donner un ordre de grandeur, il est classique de considérer plusieurs centaines, voire milliers de protéines (les variables) pour moins d'une dizaine d'échantillons biologiques. Par ailleurs, la transcriptomique est une discipline plus ancienne (les puces à ADN ont été développées dans les années 1990 [86, 87]) pour laquelle plus d'outils sont déjà disponibles. Il paraît donc naturel de s'en inspirer, voire d'en adapter certains aux spécificités des données protéomiques.

### 3.1.3 Les défis du traitement des données avant l'analyse statistique

La production des données de protéomique quantitative implique de nombreuses étapes de manipulation biochimique, analytique et bioinformatique. Celles-ci ont de nombreuses conséquences sur les données finales, dont notamment : *i)* des biais, c'est-à-dire des va-

riations systématiques entre les groupes d'échantillons qui ne sont pas dû au phénomène que l'on étudie ; et *ii)* la présence d'une proportion importante de valeurs manquantes.

Les biais peuvent avoir de multiples causes, aussi bien matérielles qu'humaines : ils peuvent être dus aux traitements biologiques, expérimentaux et techniques [76, 88]. Ces sources techniques de variabilité sont aussi connues sous le nom d'effet de lot (ou « *batch effect* » en anglais) [89, 90, 91]. Ces processus techniques non-biologiques peuvent interférer avec la réelle variabilité biologique que l'on souhaite étudier. Par conséquent, les conclusions de l'étude ne seront potentiellement pas reproductibles, c'est-à-dire que des études indépendantes ultérieures ne retrouveront pas forcément les mêmes résultats. Dans le contexte clinique, cela consisterait à échouer dans l'identification de biomarqueurs réellement efficaces. Afin d'éviter cela, différentes méthodes de normalisation ont été proposées dans la littérature, avec pour objectif de pouvoir limiter certains effets de lots (avec un succès dépendant à la fois de l'amplitude de l'effet et de son origine) [92].

Les valeurs manquantes correspondent à des observations pour lesquelles la mesure n'est pas disponible. Malheureusement, en raison de la chaîne instrumentale utilisée, les données de protéomique quantitative comportent souvent une part relativement importante de valeurs manquantes, pouvant aller jusqu'à 50% [93]. Les origines de ces valeurs manquantes sont multiples [76, 94, 95]. La procédure d'identification et de quantification résulte d'un enchaînement de plusieurs étapes au cours duquel des erreurs peuvent se propager. Ainsi, à titre d'exemple, si un peptide est mal identifié dans un échantillon, la quantification ne sera pas associée à la bonne séquence, induisant une valeur manquante pour celui-ci. Des valeurs manquantes peuvent aussi être simplement dues à l'absence de certains peptides, ou à des peptides dont les abondances sont extrêmement proches des limites de détection du spectromètre, de sorte que ces valeurs manquantes sont finalement proches de mesures d'intensité nulle. Dans le cas de la DDA (voir Section 2.2.3.2), un peptide produira des valeurs manquantes parce qu'il est sélectionné pour la fragmentation dans un échantillon mais pas dans un autre, ou parce qu'il peut entraîner un spectre MS2 de qualité inférieure. Or, ces valeurs manquantes sont extrêmement problématiques. En plus de la perte pure et simple d'information, elles peuvent entraîner un biais dans les estimations des paramètres et des erreurs lors de l'analyse statistique [95]. Les tests statistiques perdent en puissance (voir définition à la Section 3.2.3) plus le nombre de valeurs manquantes est important. Cela conduit potentiellement les protéomiciens à écarter des protéines pourtant biologiquement intéressantes. C'est pourquoi, à l'heure actuelle, le consensus en protéomique quantitative est de réaliser une étape d'AC (vois Section 2.3.2) et/ou

une étape d'imputation des valeurs manquantes (c'est-à-dire de les remplacer par des valeurs estimées), afin de ne pas pénaliser l'analyse statistique en aval [76, 93, 94, 96]. Le développement de méthodes et algorithmes d'imputation est devenu un champ actif de la protéomique computationnelle [93, 97].

Même si les nombreuses méthodes de normalisation [92, 98] et d'imputation permettent déjà d'améliorer sensiblement la puissance des résultats [88], l'absence de consensus dans la manière de les utiliser comme la jeunesse de l'expression de ce besoin font qu'il y a toujours un réel manque d'outils logiciels facile d'utilisation permettant de proposer un cadre méthodologique défini, justifié et fiable pour mener à bien l'analyse statistique des données de protéomique.

### 3.1.4 La stratégie d'analyse adoptée à EDyP

Par souci de reproductibilité sur la plateforme du laboratoire, EDyP a fait le choix de standardiser autant que possible le traitement des données de quantification préalablement à l'analyse différentielle. Pour rendre ces étapes facilement réalisables, l'équipe développe activement une suite logicielle qui est aujourd'hui le support de mes travaux : les paquets R DAPAR et Prostar [73]. DAPAR est le paquet contenant toutes les fonctionnalités qui permettent de réaliser le prétraitement et l'analyse différentielle. Prostar est l'interface graphique permettant une utilisation facilitée de DAPAR, dédiée aux utilisateurs n'ayant pas de compétences de programmation R. Le prétraitement consiste en trois étapes : le filtrage de certaines protéines, la normalisation des abondances et l'imputation des valeurs manquantes, et qui sont résumées dans la figure 3.1.

Le filtrage a pour objectif de supprimer les protéines non pertinentes (voir figure 3.1) : celles correspondant à des erreurs, des contaminants, *etc.* ; ou celles pour lesquelles le nombre de valeurs manquantes est supérieur au seuil maximal permettant une bonne exploitation des données. En pratique, c'est le protéomicien qui, en fonction de sa connaissance des données, décide de ce seuil maximal, soit au sein de chaque condition biologique, soit sur l'ensemble de l'expérience. Formellement, la normalisation consiste à ajuster les valeurs mesurées pour chaque échantillon sur une gamme de valeurs commune [99, 82] (voir figure 3.1). En protéomique *bottom-up* il est possible de normaliser les données à une multitude d'étapes de traitement différentes [100]. Ici, nous nous concentrons sur la normalisation à l'étape de prétraitement des données de quantification, c'est-à-dire une fois la LC-MS/MS réalisée et les données acquises, identifiées et quantifiées. Au sein de la grande diversité et complexité des méthodes de normalisation couramment utilisées

en protéomique et qui sont proposées dans DAPAR/Prostar, nous pouvons citer les plus simples (sans prétendre à l'exhaustivité compte-tenu du nombre considérable de méthodes qu'il faudrait décrire ici) :

**Le centrage sur la médiane** : Cela consiste à calculer la médiane sur l'ensemble des protéines de chaque échantillon, puis à remplacer ces médianes par la médiane globale de tous les échantillons. Cela suppose que le nombre de protéines respectivement sur-exprimées et sous-exprimées entre les conditions comparées sont relativement similaires.

**L'Alignement des quantiles** : Cela consiste à ordonner les abondances des protéines dans chaque échantillon, puis à attribuer les mêmes abondances aux protéines ayant le même rang. Cette normalisation suppose que la majorité des protéines ont des quantités inchangées entre les conditions comparées.

**L'usage d'abondances relatives intra-échantillon** : Pour chaque échantillon, chaque valeur d'abondance est divisée par la somme totale des abondances de l'échantillon. Cela suppose que la quantité de matériel biologique est constante entre les échantillons.

**Le centrage-réduction** : les abondances sont centrées et réduites autour de la moyenne globale des échantillons.

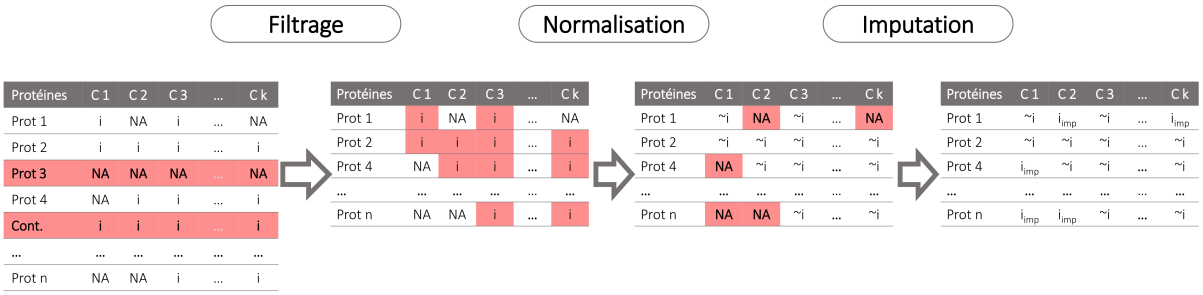



FIGURE 3.1 – Schéma des différentes étapes de prétraitement des données de protéomique avant de réaliser l'analyse statistique. Un tableau correspond à des données quantifiées pour n protéines et k conditions. Les cellules surlignées en rouge correspondent aux valeurs concernées par l'étape à réaliser. La première étape consiste à filtrer les protéines correspondant à des contaminants ou pour lesquelles le nombre acceptable de valeurs manquantes est atteint ou dépassé, réduisant ainsi le nombre de protéines à analyser par la suite. La seconde étape consiste à normaliser les échantillons, afin que les valeurs d'abondances soient comparables entre elles. La troisième et dernière étape consiste en l'imputation des valeurs manquantes restantes.

Dans DAPAR/Prostar, les valeurs manquantes sont considérées selon deux catégories :

l'ensemble des valeurs est absent pour une condition donnée (« *Missing in Entire Condition* »ou MEC) ou bien seulement une partie des valeurs sont manquantes pour une condition donnée (« *Partially Observed Values* » ou POV). Cette catégorisation permet de réaliser une imputation plus fine car plus adaptée à la nature des valeurs manquantes concernées. Il est important de souligner qu'en cas d'absence d'AC (voir Section 2.3.2), les étapes de pré-traitement telles que décrites dans la Section 3.1.4 sont modifiées : Soit il est nécessaire d'être plus stringent lors de l'étape de filtrage des valeurs manquantes, soit il faut imputer beaucoup plus de valeurs manquantes. Dans le cas du filtrage, se présente le risque de perdre beaucoup de protéines candidates. Dans le cas de l'imputation, une proportion non négligeable de valeurs n'ayant pas de signification biologique réelle va être introduite, ce qui risque d'engendrer une augmentation des faux positifs. L'éventail des possibilités d'imputation en fonction du statut MEC/POV et de la diversité des algorithmes est considérable. Comme ces questions sont assez éloignées de mon travail de thèse, elles n'ont pas vocation à être discutées ici, mais le lecteur intéressé peut se référer à [88, 94, 97, 101].

## 3.2 Hypothèses et test statistique

### 3.2.1 Définitions et notation des hypothèses

Lorsque l'on réalise une étude pour répondre à une question biologique, on souhaite que la réponse s'applique à l'ensemble des individus concernés par cette question, autrement dit la population. Malheureusement dans la pratique, nous ne pouvons pas réaliser des observations sur l'ensemble de la population, mais seulement sur une fraction de celle-ci : l'échantillon statistique. Ce dernier permet, sous condition de représentativité de la population, d'inférer des conclusions sur cette dernière. La question biologique du chercheur est formulée par une hypothèse à tester. L'objectif est de prédire si cette hypothèse est fondée pour l'ensemble de la population, sur la seule base des observations disponibles (c'est-à-dire les échantillons étudiés). Formellement, cette hypothèse est appelée l'hypothèse nulle et est notée $H_0$. Pour une analyse différentielle en protéomique, elle est formulée pour chaque protéine (à la fois identifiée et quantifiée) et représente généralement l'absence de différence entre les conditions biologiques étudiées [102, 103]. De manière contre-intuitive, c'est son rejet qui est intéressant pour le chercheur l'ayant formulé. L'hypothèse alternative, notée $H_1$, est implicitement définie par le rejet de l'hypothèse nulle. Par conséquent, elle correspond à l'existence d'une différence entre les conditions étudiées, qu'elle soit po-

sitive ou négative. Cette nature implicite de $H_1$ fait qu'il existe une certaine asymétrie entre $H_0$ et $H_1$ : on peut apporter des preuves uniquement contre $H_0$, mais on ne peut pas l'accepter ni porter des conclusions sur $H_1$. Le test statistique permet de quantifier cette preuve contre $H_0$ (pour une définition plus théorique, se référer à la Section 3.2.4) et donc de prendre une décision binaire en fonction de ce niveau de preuve, c'est-à-dire de conclure au rejet ou au non rejet de l'hypothèse nulle. A partir de ces définitions, nous pouvons distinguer quatre types de résultat suite à un test statistique. Ceux-ci sont résumés dans le Tableau 3.1.

| **Type de résultat** | **Conclusion par rapport à $H_0$** |
|---|---|
| Vrai Positif (VP) | Rejet de $H_0$ lorsqu'elle est fausse |
| Vrai Négatif (VN) | Non rejet de $H_0$ lorsqu'elle est vraie |
| Faux Positif (FP) | Rejet de $H_0$ lorsqu'elle est vraie |
| Faux Négatif (FN) | Non rejet $H_0$ lorsqu'elle est fausse |

TABLEAU 3.1 – Correspondance entre les différents types de résultat reportés dans les études qui sont définis selon les conclusions à l'issue du test statistique.

Concrètement, dans notre contexte protéomique, si nous souhaitons comparer deux conditions biologiques A et B, nous pouvons formuler :

$H_0$ : « il n'existe pas de différence d'abondance entre A et B »

$H_1$ : « il existe une différence d'abondance entre A ou B » qu'il s'agisse d'une sur-abondance ou d'une sous-abondance dans A vis-à-vis de B.

Ainsi à titre d'exemple, une protéine pour laquelle nous avons conclu qu'elle présentait une différence d'abondance entre les conditions alors qu'en réalité il n'y en a pas, est appelée un faux positif (FP).

### 3.2.2 Les risques et probabilités d'erreur de décision

Suite aux correspondances présentées dans la tableau 1, nous pouvons définir différentes probabilités correspondant aux erreurs de conclusion ou bien aux conclusions correctes respectivement à $H_0$ [104], et qui sont résumées dans le tableau 3.2.

On appelle risque d'erreur de première espèce (ou risque d'erreur de Type I), la probabilité

de rejeter $H_0$ alors que $H_0$ est vraie, autrement dit $H_0$ est rejetée à tort (c'est-à-dire que l'on soit face à un FP). Ce risque est noté $\alpha$ :

$$\alpha = P(rejet\ de\ H_0 \mid H_0\ vraie) \tag{3.1}$$

On appelle risque d'erreur de seconde espèce (ou erreur de Type II), la probabilité de ne pas rejeter $H_0$ alors que $H_0$ est fausse(c'est-à-dire que l'on soit face à un FN). Ce risque est noté $\beta$ :

$$\beta = P(non\ rejet\ de\ H_0 \mid H_0\ fausse) \tag{3.2}$$

La probabilité de ne pas rejeter $H_0$ lorsqu'elle est vraie et donc de conclure correctement est notée $1 - \alpha$ et est définie par :

$$1 - \alpha = P(non\ rejet\ de\ H_0 \mid H_0\ vraie) \tag{3.3}$$

Enfin, on appelle la puissance du test, la probabilité de rejeter l'hypothèse nulle lorsqu'elle est fausse et est notée $1 - \beta$ :

$$1 - \beta = P(rejet\ de\ H_0 \mid H_0\ fausse) \tag{3.4}$$

| | **Non rejet de $H_0$** | **Rejet de $H_0$** |
|---|---|---|
| $H_0$ vraie | $1 - \alpha$ | $\alpha$ (erreur de Type I) |
| $H_0$ fausse | $\beta$ (erreur de Type II) | $1 - \beta$ (puissance) |

TABLEAU 3.2 – Table de contingence des différentes probabilités correspondant aux différentes conclusions que l'on peut mener respectivement à $H_0$.

En protéomique, et en particulier dans la recherche de biomarqueurs, le plus intéressant est d'identifier les protéines pour lesquelles il y a effectivement une réelle différence d'abondance, et qui apportent une réelle plus-value clinique [105]. Obtenir un nombre conséquent de faux positifs aura des conséquences désastreuses sur les études ultérieures de validation. Par conséquent, les erreurs de type I (les faux positifs) sont beaucoup plus grave que celles de type II (les faux négatifs), même si augmenter la puissance statistique

est toujours désirable : cette asymétrie étant intrinsèquement présente dans le formalisme du test d'hypothèse, son utilisation en protéomique est particulièrement naturelle.

### 3.2.3 La puissance statistique

Admettons que $H_0$ soit fausse. Dès lors, il est nécessaire d'avoir rassemblé un certain nombre d'observations pour qu'un test statistique devienne capable de détecter la différence que l'on veut mettre en évidence. Mais ce nombre d'observations nécessaires est différent selon le test. En effet, moins le test a besoin d'échantillons pour mettre en avant une différence donnée, plus le test sera considéré comme puissant. De nombreux facteurs influencent la puissance d'un test statistique et les anticiper permet d'optimiser la collecte de données expérimentales (souvent laborieuse et coûteuse). Les principaux facteurs sur lesquels il est possible d'agir sont *i)* le nombre d'échantillons (plus la taille d'échantillon statistique est grande, plus la puissance est élevée [106]). *ii)* La taille des effets de l'expérience. Elle correspond à la plus petite différence entre les moyennes ou les proportions que l'on estime intéressante à détecter. La puissance statistique sera d'autant plus élevée, que l'effet biologique étudié sera important. Il s'agit du paramètre le plus difficile à déterminer, car il repose sur les connaissances biologiques du chercheur. *iii)* L'équilibre du plan d'expérience : Si le nombre d'échantillons est très différent entre les conditions, alors la puissance du test sera affectée [107]. *iv)* Le niveau d'erreur dans les mesures expérimentales. L'erreur de mesure correspond à la variance qui masque les vrais effets de l'expérience. Améliorer la précision et la cohérence des mesures diminue cette variance non biologique et maintien la puissance statistique [106]. Pour un approfondissement de l'analyse de la puissance statistique se référer à [108, 109].

### 3.2.4 La p-valeur

Pour réaliser un test statistique, on définit $X$ la variable aléatoire associée à l'aléa de l'expérience, et on note $x$ la réalisation de cette variable aléatoire (il s'agit en somme de la notation générique d'une observation). Dans notre cas, $X$ représente donc l'abondance aléatoire d'une protéine dans un ensemble d'échantillons, et $x$ l'abondance précisément mesurée pour une protéine pour un échantillon biologique donné. On définit ensuite la statistique du test, notée $S$. Il s'agit aussi d'une variable aléatoire et elle caractérise un élément invariant (aux fluctuations du hasard près) des données quand celles-ci sont sous l'hypothèse nulle. Par exemple, pour les tests où l'on réalise des comparaisons de moyennes, $S$ sera liée à la distribution supposée des différences de moyennes empiriques (c'est-à-dire

observées en pratique) dans la population. La distribution de $S$ pour les données sous $H_0$ est notée $S_{H_0}$. Pour pouvoir réaliser un test statistique, il faut que la loi de $S_{H_0}$ soit connue (pour des tests classiques, cette loi peut-être par exemple, la loi de Student, la loi du $\chi^2$ ou la loi de Fisher-Snedecor) [110]. En revanche, la distribution de $S$ sous $H_1$ est le plus souvent inconnue. Pour chaque observation à tester, on peut calculer la même quantité que celle qui sert à définir $S$ (dans notre exemple, la quantité liée à la différence de moyenne). Notons cette valeur $s_{cal}$. Une fois en sa possession, il suffit de comparer celle-ci à la distribution de $S_{H_0}$ pour estimer un risque d'erreur et donc prendre une décision. Pour concrètement réaliser cela, on définit simplement le niveau de significativité du test, noté $s_{seuil}$, qui est la probabilité de rejeter à tort l'hypothèse nulle. Le $s_{seuil}$ correspond à la valeur que prend $S_{H_0}$ pour laquelle la probabilité d'erreur de Type I, $\alpha$, est fixée. Ensuite, la conclusion concernant le rejet/non rejet de l'hypothèse nulle est obtenue en comparant $s_{cal}$ et $s_{seuil}$. Si $s_{cal} \geq s_{seuil}$ alors nous rejetterons $H_0$, tandis que si $s_{cal} < s_{seuil}$, nous ne rejetterons pas $H_0$.

Le test statistique permet de calculer la probabilité qu'une valeur de $S_{H_0}$ soit au moins aussi éloignée de son espérance que la valeur observée $s_{cal}$. Donc on souhaite quantifier à quel point il est probable que la valeur $s_{cal}$ obtenue avec les données observées provienne de la distribution de $S_{H_0}$. Cette probabilité associée à $s_{cal}$ est la $p_{valeur}$ (ou probabilité critique). Il s'agit de la probabilité d'observer une valeur sous $S_{H_0}$ au moins aussi extrême que $s_{cal}$ :

$$p_{valeur} = P(S_{H_0} \geq s_{cal}) = P(S \geq s_{cal}|H_0) \tag{3.5}$$

Autrement dit, la $p_{valeur}$ est calculée à partir de l'aire sous la courbe de distribution supposée de $S_{H_0}$ et permet donc de quantifier à quel point les données sont incompatibles avec la distribution de $S_{H_0}$ [103, 110, 111, 112, 113, 114]. Ainsi, une $p_{valeur}$ très faible signifie qu'il est improbable que l'on observe une valeur sous $S_{H_0}$ aussi extrême que $s_{cal}$. Lorsque $p_{valeur} < \alpha$, le résultat est considéré comme statistiquement significatif au risque $\alpha$. En revanche, la $p_{valeur}$ ne reflète pas l'importance de l'effet (la différence d'abondance dans notre cas) [110]. Un effet peu important (une faible différence d'abondance entre conditions) peut être associé à une $p_{valeur}$ faible (résultat significatif). A l'inverse, un effet important (grande différence d'abondance entre conditions) peut avoir une $p_{valeur}$ très élevée donc non significative. Autrement dit, le sens biologique du résultat doit être bien distingué de la significativité du test statistique.

Pour finir, donnons un exemple concret de la manière d'interpréter la $p_{valeur}$, avec la com-

paraison de l'abondance d'une protéine entre un état sain et la présence d'une pathologie. On établit l'hypothèse nulle telle que : « il n'existe pas de différence d'abondance de la protéine entre l'état sain et l'état pathologique ». Suite au test statistique, on obtient une $p_{valeur} = 0,0025$. Cette valeur signifie que si l'on avait réalisé 400 comparaisons entre deux états identiques, on pourrait espérer que l'une d'entre elles induise une différence empirique plus importante que celle obtenue dans la pratique. Comme on peut le voir ci-dessus, l'interprétation correcte de la $p_{valeur}$ ne permet pas de statuer directement sur la probabilité d'être face à un FP ou un FN, ce qui est contre-intuitif, vue le lien entre la $p_{valeur}$ et $\alpha$. Néanmoins cela s'explique par le fait que le risque de prendre une mauvaise décision pour le chercheur se calcule par rapport à l'ensemble des décisions prises (ou à prendre), et non sur l'ensemble des observations sous $H_0$, comme cela apparaît dans l'équation 3.5.

## 3.3 Deux principaux tests statistiques utilisés en protéomique

Deux tests statistiques couramment appliqués dans les études de protéomique, sont le t-test de Student et l'Analyse de la Variance. Le t-test [115] et ses variantes constituent le principal outil proposé pour l'analyse différentielle dans la suite logicielle DAPAR/Prostar [73]. L'objectif de ces tests statistiques est de déterminer si $k$ groupes (les conditions biologiques étudiées) proviennent de la même population en comparant les moyennes, c'est-à-dire, que la distribution de la variable d'intérêt est la même dans les $k$ groupes. Les deux sous-sections suivantes sont consacrées à la description de ces tests afin d'alerter les utilisateurs sur les limites de tels tests. En effet, ce sont des tests paramétriques, autrement dit il est nécessaire de respecter certaines conditions d'application pour pouvoir les utiliser sans prendre le risque de perdre en puissance statistique ou d'avoir une inflation du nombre de faux positifs.

### 3.3.1 Le t-test : étude d'un seul facteur [1] à 2 conditions

Le t-test permet de traiter le cas où l'on étudie l'influence d'un seul facteur à deux conditions sur l'expression des protéines. Formellement, il s'agit d'expliquer les observations d'une variable quantitative (l'abondance d'une protéine) en fonction d'une variable qualitative à deux classes (les conditions biologiques étudiées). Soit deux conditions A et B

1. Comme définit dans la section 1.1.1

ayant chacune un nombre d'observations de taille $n_A$ et $n_B$. On définit la moyenne observée de chaque condition, notées $m_A$ et $m_B$. L'hypothèse nulle suppose que les moyennes observées proviennent de la même population relativement à la variable étudiée. Autrement dit, les hypothèses nulle et alternative sont définie par :

$$\begin{cases} H_0 : m_A = m_B \\ H_1 : m_A \neq m_B \end{cases} \tag{3.6}$$

Ce test statistique suppose que *i)* les observations sont indépendantes et identiquement distribuées selon une loi normale au sein de chaque condition ; et *ii)* les données sont homoscédastiques, c'est-à-dire que la variance empirique des deux conditions est la même (et la valeur correspondante est notée $s^2_{emp}$). La statistique $S$ sous $H_0$ suit une distribution $t$ de Student et la statistique empirique (notée $t_{cal}$) est calculée telle que :

$$t_{cal} = \frac{m_A - m_B}{\sqrt{\frac{s^2_{emp}}{n_A} + \frac{s^2_{emp}}{n_B}}} \tag{3.7}$$

On détermine un $t_{critique}$ obtenu dans la table de Student pour le seuil $\alpha$ fixé par l'utilisateur, que l'on compare ensuite à $t_{cal}$. Si $t_{cal} \geq t_{critique}$ alors on rejette $H_0$ pour le risque $\alpha$ fixé. Des tests alternatifs au t-test classique existent [116, 117, 118], notamment le t-test de Welch [119] qui ne présuppose pas l'homoscédasticité des observations, ou encore la version proposée dans Limma, qui propose un estimateur de la variance mutualisé sur l'ensemble des protéines [78].

### 3.3.2 L'Analyse de la Variance : étude d'un seul facteur à $k \geq 2$ conditions

L'Analyse de la Variance à un facteur (« *1-way ANOVA* ») est une généralisation du t-test [120, 121]. Elle est donc utilisée dans le cas où l'on s'intéresse à une seule variable qualitative présentant plus de deux conditions afin de réaliser une comparaison de moyennes entre conditions [122]. Le test statistique est basé sur la formulation d'une hypothèse nulle $H_0$ qui correspond à l'hypothèse où toutes les moyennes sont égales. L'hypothèse alternative $H_1$ correspond au cas où au moins une des moyennes s'écarte des autres moyennes. Soit $k$ le nombre de conditions étudiées, $n$ l'effectif total du jeu de données et $m$ la moyenne

pour une condition donnée. $H_0$ et $H_1$ peuvent alors être formulées ainsi :

$$\begin{cases} H_0 : m_1 = m_2 = ... = m_k \\ H_1 : \exists(i,j) \; tel \; que \; m_i \neq m_j \end{cases} \tag{3.8}$$

Un exemple d'application peut consister à étudier l'effet de différents traitements sur la concentration plasmatique d'une protéine. On définit $H_0$ « il n'y a pas de différence de concentration moyenne entre les différents traitements », et $H_1$ « il y a une différence de moyenne entre au moins deux traitements, mais on ne sait pas entre lesquels ». L'ANOVA repose sur l'analyse de la dispersion des données autour de la moyenne, autrement dit la variance des données observées [123]. Cette variance peut avoir deux origines : *i)* l'effet du facteur étudié, aussi appelée variabilité factorielle ou inter-classe, qui représente la dispersion des données imputable aux différentes conditions étudiées ; et *ii)* la variabilité intrinsèque à chaque condition étudiée, aussi appelée variabilité résiduelle ou intra-classe (pour une condition donnée, les observations ne sont pas strictement identiques, il existe une certaine variabilité dans les observations qui n'est pas expliquée par le facteur étudié). Ce test cherche à quantifier dans quelle mesure l'ensemble de la variabilité observée autour de la moyenne sur l'ensemble des données (appelée variabilité totale) est expliquée par la variabilité inter-classe (entre les conditions) et la variabilité intra-classe (au sein d'une même condition). Cette variabilité totale est alors définie comme la somme des variabilités inter-classe et intra-classe. En faisant les sommations appropriées sur cette variabilité totale des observations, on obtient l'équation de l'analyse de la variance, c'est-à-dire la somme des carrés totaux (notée $SCT$). La $SCT$ correspond à la somme de la somme des carrés expliqués (notée $SCE$ c'est à dire la variabilité expliquée par le facteur étudié), et de la somme des carrés résiduels (notée $SCR$, c'est-à-dire la variabilité résiduelle non expliquée par le facteur étudié) : $SCT = SCE + SCR$. A partir de là, on définit les Carrés Moyens Totaux ($CMT$) tels que :

$$CMT = \frac{SCT}{n-1} \tag{3.9}$$

puis les Carrés Moyens Expliqués ($CME$) et les Carrés Moyens Résiduels ($CMR$) tels que :

$$CMT = \frac{SCE}{k-1} \tag{3.10}$$

$$CMR = \frac{SCR}{n-k} \tag{3.11}$$

Pour l'ANOVA, la statistique $S$ sous $H_0$ suit une distribution $F$ de Fisher. Et la statistique seuil F théorique, notée $F_{théorique}$ est définie telle que :

$$F_{théorique} = \frac{CME}{CMR} = F_{n-k}^{k-1} \tag{3.12}$$

La variance inter-classe des données observées est notée $s^2_{inter}$ et correspond au $CME$ obtenu pour la distribution de $S_{H_0}$. La variance intra-classe des observations est notée $s^2_{intra}$ et correspond au $CMR$ obtenu pour la distribution de $S_{H_0}$. On peut alors définir la statistique F observée, notée $F_{cal}$ telle que :

$$F_{cal} = \frac{s^2_{inter}}{s^2_{intra}} \tag{3.13}$$

Pour plus de détails sur la construction du modèle mathématique et la décomposition de la variance, voir [121]. L'hypothèse nulle est rejetée si l'on obtient pour un risque d'erreur $\alpha$ fixé :

$$|F_{cal}| \geq F_{théorique,\alpha} \tag{3.14}$$

Dans notre contexte protéomique, le rejet de l'hypothèse nulle avec l'ANOVA signifie que la protéine est significativement différentiellement exprimée au risque $\alpha$ entre les conditions étudiées, mais on ignore précisément quelles conditions diffèrent des autres. L'ANOVA nécessite le respect de trois conditions d'applications pour assurer un contrôle du risque d'erreur $\alpha$ et une puissance statistique suffisante [124] : *i)* la distribution des données de chaque condition étudiée suit une loi normale ; *ii)* l'homogénéité de la variance des données entre les conditions, c'est-à-dire qu'elles doivent être égales ; *iii)* l'indépendance des erreurs résiduelles. Il est communément admis que L'ANOVA n'est pas un test robuste au non-respect d'une de ces hypothèses, notamment celle de l'homoscédasticité [124, 125, 126]. Par conséquent, en cas d'hétéroscédasticité, la puissance statistique est moindre, avec une plus grande propension aux erreurs de première espèce [126, 127]. C'est pour pallier ces défauts que des alternatives plus robustes ont été développées [128, 129].

## 3.4 Les tests multiples en analyse à haut débit

### 3.4.1 Définitions

La protéomique quantitative donne accès à un grand nombre de protéines. Or, comme cela a été montré au Chapitre 2, la quantification des protéines est relative, autrement dit les abondances ne sont pas directement comparables entre elles. Par conséquent, pour chaque protéine, on réalise un test à propos de son abondance différentielle indépendamment des autres protéines. Or, si l'on suppose qu'un nombre $k$ d'hypothèses nulles testées sont vraies, il existe une certaine probabilité qu'au moins un test ressorte significatif uniquement par hasard, autrement dit on obtient un faux positif (comme cela a été illustré en clôture de la Section 3.2). Sachant que pour un test donné, la probabilité de réaliser une erreur de type I est égale à $\alpha$, et que l'on réalise $k$ tests indépendants [2], alors la probabilité globale d'obtenir au moins un test significatif uniquement par hasard sur l'ensemble des tests réalisés est de :

$$Pr(rejet\ d'au\ moins\ une\ H_0) = 1 - (1 - \alpha)^k \tag{3.15}$$

Autrement dit, plus le nombre de tests augmente, plus la probabilité de rejeter $H_0$ à tort au moins une fois augmente. Plus concrètement, si l'on réalise 100 tests pour lesquels chaque $H_0$ est vraie (les 100 protéines ne sont pas différentiellement exprimées entre les conditions), et que l'on fixe $\alpha = 0.05$ (c'est-à-dire que l'on accepte de se tromper à hauteur de 5% de manière globale), alors on obtient en réalité une probabilité de 99,4% d'obtenir un résultat significatif par hasard (et non pas une probabilité de 5% comme on le voudrait). Il apparaît donc nécessaire de contrôler ce phénomène indésirable dû à la multiplicité des tests réalisés [130, 131, 132]. C'est pourquoi des procédures de correction de tests multiples (« *multiple testing correction* » ou MTC) ont été mises en place. L'objectif des MTC est de ramener le risque de type I à un niveau global (*i.e.* sur l'ensemble des tests réalisés), autrement dit que le risque $\alpha$ soit fixé pour les $k$ tests réalisés et non pas seulement au niveau d'un seul test. Les MTC dérivent le plus souvent de l'équation 15 ci-dessus, telle que la correction de bonferroni [133] (pour laquelle $\alpha$ est recalculé en fonction du nombre $k$ de tests réalisés), et sont notamment employées dans les tests statistiques dit *post hoc* présentés dans la section suivante. Une approche différente de MTC très populaire à l'heure actuelle est celle développée par Benjamini & Hochberg (BH) [134], avec laquelle on tolère une certaine proportion contrôlée de FP parmi les hypothèses nulles rejetées.

2. Les tests sont considérés indépendants, c'est-à-dire que l'on ne tient pas compte des éventuelles corrélations d'expression qui pourraient exister entre certaines protéines.

### 3.4.2 Les tests *post hoc*

De manière générale, après une ANOVA à un facteur, le chercheur souhaite approfondir sa question, afin de savoir entre quelles conditions est différemment exprimée la protéine en question. Il devient alors nécessaire de tester de manière exploratoire toutes les possibilités de comparaisons entre les conditions (induisant un besoin de MTC) d'observations, sans *a priori* [135] : c'est l'objectif des tests statistiques *post hoc*. Il existe de multiples tests post hoc différents, chacun dédié à un type de question formulée *a posteriori* [130, 135] de la réalisation d'une ANOVA. Les principales différences entre ces tests sont *i)* le type de comparaisons qu'ils réalisent (et donc le nombre de comparaisons) ; et *ii)* la façon dont est faite la MTC, c'est-à-dire que les différents tests *post hoc* ne contrôlent pas l'erreur de Type I de la même manière. Ils seront donc plus ou moins conservateurs avec une puissance statistique variable. Parmi les tests les plus connus nous pouvons citer celui de Scheffé [136], le « *Honestly Significant Difference* » (HSD) de Tukey [137, 138], le test de Dunnett [139] et le « *Least Significant Difference* » (LSD) de Fisher [130, 140, 141]. Une présentation générale de ces tests est faite dans le tableau 3.3, du test le plus conservateur (faible puissance statistique) au plus laxiste (pas de contrôle de l'erreur de type I). Pour plus de précisions concernant ces différentes méthodes, se référer à [120, 135].

Le choix du test *post hoc* à réaliser peut s'avérer difficile à cause de la grande diversité de tests disponibles. La littérature suggère qu'une manière de choisir est de décider [120, 135, 140] si la priorité est de contrôler l'erreur de Type I de manière plus ou moins stringente, et quel est l'équilibre souhaité entre erreur de Type I et puissance statistique. Par ailleurs, des situations problématiques peuvent être rencontrées par l'expérimentateur [121] : *i)* un résultat significatif est obtenu à l'ANOVA mais le test *post hoc* utilisé en aval n'exhibe pas de résultat significatif ; ou *ii)* il n'y a pas de résultat significatif à l'ANOVA mais il y en a un à l'issue du test *post hoc*. Dans la pratique, les tests *post hoc* ne sont réalisés que pour les protéines qui ont un résultat significatif à l'ANOVA, et non pas pour l'ensemble des protéines. Cette manière de procéder est initialement motivée par des raisons historiques. En effet, quand la théorie avait été formulée, les capacités calculatoires étaient bien moindres qu'actuellement, de sorte que cette manière de procéder permettait de limiter le nombre d'opérations réalisés. Cependant avec cette approche, la situation problématique *ii)* n'est pas contrôlée, pouvant mener à des problèmes de puissance statistique [121], et certains auteurs recommandent de réaliser les tests *post hoc* quelle que soit la conclusion de l'ANOVA [135].

| **Test** | **Comparaisons réalisées** | **Distribution de la statistique** | **Statistique des données observées** |
|---|---|---|---|
| Scheffé | toutes les combinaisons possibles | $F$ de Fisher | $F = \frac{(\bar{X}_i - \bar{X}_j)^2}{CMR\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ |
| Tukey | chaque paire possible de tous les groupes : $\frac{k(k-1)}{2}$ | étendues studentisées | $q = \frac{(\bar{X}_i - \bar{X}_j)}{\sqrt{\frac{CMR}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$ |
| Dunnett | comparaison d'un groupe de référence à chacun des autres groupes : $(k-1)$ | $t$ de Dunnett | $d = \frac{(\bar{X}_{contrôle} - \bar{X}_i)}{\sqrt{CMR\left(\frac{1}{n_{contrôle}}+\frac{1}{n_i}\right)}}$ |
| Fisher | chaque paire possible de tous les groupes : $\frac{k(k-1)}{2}$ | $t$ de Student | $t = \frac{(\bar{X}_i - \bar{X}_j)}{\sqrt{CMR\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}$ |

TABLEAU 3.3 – Comparaison des principaux tests *post hoc*.
Principaux tests *post hoc* classés du plus conservateur (faible puissance statistique) au plus laxiste (forte propension aux faux positifs). Avec $n$ le nombre d'observation pour un groupe ; le couple $(i, j)$ deux groupes parmi les $k$ groupes à comparer ; « *contrôle* » est le groupe de référence pour le test de Dunnett ; $CMR$ est le Carré Moyen Résiduel calculé lors de l'ANOVA (voir section 3.3.2).

### 3.4.3 Le contrôle du taux de fausses découvertes

Considérons $m$ le nombre d'hypothèses nulles testées, avec un nombre $m_0$ de $H_0$ vraies. Soit $R$ le nombre d'hypothèses nulles rejetées (quelles soient vraies ou non), que l'on appelle aussi découverte (ou « *discovery* »). Soit $V$ le nombre d'hypothèses nulles vraies rejetées à tort, et $S$ est le nombre d'hypothèses nulles fausses rejetées correctement. On obtient alors la table de contingence détaillée dans le tableau 3.4. $R$ est une variable aléatoire observable, tandis que $U$,$V$,$S$ et $T$ sont des variables aléatoires non observables ; puisqu'on ne connaît pas la vérité sur les $H_0$ testées.

| | **Déclaré non significatif** | **Déclaré significatif** | **Total** |
|---|---|---|---|
| $H_0$ vraies | U | V | $m_0$ |
| $H_0$ fausses | T | S | $m - m_0$ |
| Total | $m - R$ | $R$ | $m$ |

TABLEAU 3.4 – Table de contingence présentant les différents nombres associés aux conclusions possibles respectivement à $H_0$.

On définit la Proportion de Fausses Découvertes (« *False Discovery Proportion* » ou FDP) comme la proportion d'hypothèses $H_0$ qui ont été rejetées de manière erronée sur l'ensemble des hypothèses testées, telle que :

$$FDP = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases} \tag{3.16}$$

Mais la FDP n'est pas directement calculable car V n'est pas accessible. En revanche, il est possible d'estimer une valeur proche, appelée le taux de fausses découvertes (« *False Discovery Rate* » ou FDR) [134] et qui peut s'écrire :

$$FDR = E(FDP) = E\left(\frac{V}{R}\right) \tag{3.17}$$

Notre objectif de contrôler l'erreur de type I dans un contexte de test multiple peut ensuite être ramené à l'objectif de contrôler le FDR. En d'autres termes, il s'agit de maintenir le risque d'erreur de Type I à un niveau acceptable « en moyenne ». De nombreuses méthodes ont été développées pour contrôler le FDR et c'est un sujet toujours très étudié à l'heure actuelle [142, 143, 144]. La méthode la plus utilisée est celle développée par BH en 1995 [134]. Le contrôle du FDR tel que défini par BH consiste à ordonner les $p_{valeurs}$ $P_1, \cdots, P_m$ obtenues à partir du test des $m$ hypothèses, de la plus faible à la plus élevée (ordre croissant). On définit $j$ le rang d'une $p_{valeur}$ considérée et $\alpha$ la valeur critique que l'on souhaite contrôler. BH ont montré que la liste induite par les $j$ premières $p_{valeurs}$, avec $j$ tel que :

$$P_j \leq \frac{j}{m}\alpha \tag{3.18}$$

Permettait de prendre une décision sur l'ensemble des tests avec un FDR contrôlé à un niveau $\alpha$ [3]. Ainsi, seuls les tests ayant obtenus une $p_{valeur}$ de rang inférieur à $j$ seront considérés comme significatifs. La méthode de BH suppose que l'ensemble des tests réalisés sont indépendants [134], mais d'autres approches ont été développées pour s'affranchir de cette hypothèse [145, 146].

Le contrôle du FDR est une solution populaire [147] pour maîtriser l'erreur statistique sur l'ensemble des tests réalisés, que ce soit grâce à l'approche initiale de BH [134] ou grâce à d'autres plus récentes [145, 144, 147]. Comme montré en début de cette section, le FDR est uniquement calculé avec les $p_{valeurs}$ des protéines. Les protéines sont toutes supposées correctement identifiées et quantifiées, indépendamment de toute erreur antérieure qui se serait propagée (erreurs d'identification et/ou de quantification des peptides, erreurs lors de l'agrégation des peptides en protéines, erreurs lors l'étape de filtrage, méthode de normalisation inappropriée, imputation inexacte des valeurs manquantes, *etc.*). Or, il est très improbable que ce soit effectivement le cas en situation réelle. Par conséquent, ce FDR obtenu présente une certaine limite pour l'analyse statistique globale de l'étude. C'est d'ailleurs pour cela que le contrôle du FDR est aussi essentiel à d'autres étapes du traitement des données de protéomique. Notamment, à l'étape d'identification des peptides, une certaine proportion de PSM (voir Chapitre 2) est incorrecte et certains spectres sont donc assignés à tort à des peptides. Ces PSM sont donc des faux positifs, dont on souhaite estimer la proportion afin d'assurer un contrôle essentiel de la qualité des résultats. Cette estimation se fait généralement grâce à une approche empirique qui diffère de l'approche BH : la stratégie dite du « *Target-Decoy* » [56, 147, 148, 149, 150]. Elle consiste à rechercher dans deux bases de données dont la première, dite « *Target* », contient les séquences protéiques authentiques, et la seconde, dite « *Decoy* », contient des séquences artificielles. Un spectre expérimental peut se voir attribuer différents peptides candidats, aussi bien des « *targets* » que des « *decoys* ». En supposant que les non-correspondances de « *targets* » et les correspondances de « *decoys* » ont la même probabilité de survenir, et en considérant les peptides « *decoys* » comme une mesure du nombre de PSM aléatoires et incorrectes, le nombre de correspondances de « *decoys* » peut être utilisé pour estimer le

3. Dans cette formule, $\frac{j}{m}\alpha$ est la valeur attendue pour $P_j$ si les p valeurs suivent une distribution uniforme entre 0 et $\alpha$ (cette uniformité est obtenue par construction si l'on ne considère que des tests sous $H_0$).

nombre de non-correspondances de « *targets* ». Cette procédure intuitive d'estimation peut en pratique être un peu adaptée pour aboutir à une procédure de contrôle du FDR [151].

# Chapitre 4

# Problématiques et positionnement de mon travail

## 4.1 Intégration des spécificités du projet support

Comme détaillé au Chapitre 1 (cf. Section 1.1.2), les principales attentes des cliniciens pour le projet NAFLD concerne le suivi de patients tout au long de leur maladie chronique (ce qui est appelé « suivi longitudinal »). Cependant, les échantillons du projet ne correspondent pas à une étude longitudinale. En effet, nous disposons d'un seul échantillon de plasma par patient, et non de plusieurs échantillons prélevés sur chaque patient à différentes dates. Par ailleurs, la NAFLD est une pathologie chronique avec une évolution assez linéaire, de sorte que l'on peut aisément confondre le stade de la maladie et la durée de celle-ci ; et cette ambiguïté a probablement augmenté la confusion entre les objectifs des médecins et la mise en place du plan d'expérience (antérieure au commencement de ma thèse). Les données disponibles ne nous permettent donc que d'étudier l'expression des protéines selon un plan d'expérience stratifié : La progression de la maladie est discrétisée en un ensemble de stades dans lesquels se répartissent les échantillons des patients, en fonction de la gravité de leur maladie. Une difficulté a donc été de déterminer cette stratification sur la base du grand nombre de variables cliniques et paracliniques dont nous disposions. Finalement, le choix s'est porté sur le score SAF [152], qui discrimine le stade de progression de la NAFLD selon trois critères histologiques (la Stéatose, l'Activité et la Fibrose). Une description plus approfondie de ce score est présentée dans la Section 7.2.1. Nous avons choisi de prioriser l'analyse de la Fibrose plutôt que les autres car il s'agit du

facteur avec le plus gros enjeu clinique, aussi bien sur le plan médicamenteux que sur le pronostic de survie des patients. Cependant, la stratification induite par le score de Fibrose induit un plan expérimental déséquilibré : les effectifs d'échantillons diffèrent entre les différents stades, ce qui réduit potentiellement la puissance de l'étude (cf Section 3.2.3 du Chapitre 3).

Par ailleurs, sur un tel projet, le traitement statistique des données de protéomique quantitative ne peut être réalisé de manière semi-automatisée (par le protéomicien en charge de l'analyse) via le logiciel Prostar (le logiciel développé dans ce but au laboratoire, cf. Chapitre 2). En effet, l'étude portant sur des facteurs (d'abord F, puis A et S) induisant plus de deux conditions à comparer, le plan d'expérience ne peut être pris en compte de manière optimale par Prostar. De plus, les cohortes cliniques humaines présentent des difficultés techniques et méthodologiques supplémentaires par rapport aux modèles animaux, plus classiquement utilisés sur des expériences de biologie fondamentale, et pour lesquelles notre laboratoire a une meilleure expertise : Il s'agit de plus grandes cohortes d'échantillons (celle du projet NAFLD étant la plus grosse analysée jusqu'à présent au laboratoire), et le type d'échantillon analysé est différent (plasma sanguin humain, nécessitant une préparation biochimique spécifique). Enfin, une cohorte de patients humains est beaucoup moins contrôlée, de sorte qu'il faut prendre en compte des variabilités biologique et non-biologique nouvelles (avoir un contrôle plus fin sur les variabilités intergroupe et intra-groupe ; ainsi que minimiser l'influence de nombreux facteurs confondants et effets de lot par des approches statistiques jusque-là peu exploitées en protéomique).

Les méthodes d'analyses nécessaires ainsi que les outils logiciels implémentant lesdites méthodes de manière spécifique aux données de protéomique, et permettant un usage quotidien par une plateforme d'analyse (qui requiert à la fois simplicité d'utilisation et garanties de reproductibilité) sont autant de besoins qui dépassent de loin les objectifs du projet NAFLD. L'objectif de ma thèse a donc autant été de participer à l'analyse des données de ce projet, que de m'appuyer sur l'expertise acquise sur ces données pour proposer des outils génériques, qui permettront au laboratoire EDyP de réaliser par la suite et de manière autonome d'autres projets cliniques. C'est ce double objectif clinique/méthodologique qui a implicitement défini les 3 axes de recherche que j'ai poursuivis.

## 4.2 Résumé des contributions

### 4.2.1 Réduire l'influence de la variabilité biologique en pré-analytique et en post-analytique

Afin de réduire l'impact potentiellement fort des effets de lot sur des projets cliniques, je suis intervenue aussi bien en amont de l'analyse protéomique qu'en aval de celle-ci. En amont de l'analyse protéomique, je propose une méthode permettant de réaliser le placement des échantillons sur les plaques à puits, que j'ai développé sous la forme dun paquet R [153] (afin de le rendre accessible à une large part de la communauté scientifique). En aval, la variabilité biologique importante entre les échantillons et l'influence de facteurs non-biologiques non maîtrisés peuvent être partiellement réduits par l'application de méthodes bien choisies de normalisation des données. Je propose des méthodes adaptées de la trancriptomique comme alternative aux méthodes déjà proposées dans Prostar. Ces deux contributions sont décrites dans le chapitre suivant (Chapitre 5).

### 4.2.2 Améliorer l'analyse différentielle

Dans la pratique, l'approche statistique sur des données stratifiées consiste à réaliser une ANOVA à 1 facteur selon différentes approches classiquement décrites dans la littérature [127, 78]. Cependant, comme déjà mentionné dans la Section 3.4.2 du Chapitre 3, l'adéquation de ces approches à la question biologique ou clinique n'est souvent, au mieux, qu'une hypothèse implicite ; de sorte que nous avons ressenti le besoin de clarifier leur usage dans un contexte protéomique, et de fournir les outils logiciels correspondants (cf Chapitre 6).

### 4.2.3 Application de l'approche à la cohorte NAFLD et identifier des biomarqueurs potentiels

Le dernier axe de travail est d'appliquer l'approche développée dans les chapitres 5 et 6 à la cohorte NAFLD, afin : *i)* d'identifier les protéines statistiquement différentiellement abondantes entre les différentes conditions ; *ii)* d'identifier des groupes de profils d'expression intéressants pour décrire les différents stades de la pathologie et *iii)* de fournir une liste de protéines potentiellement intéressantes à étudier dans des études ultérieures pour approfondir la question initiale.

# Chapitre 5

# Première contribution : Solutions pour la gestion des effets de lot dans les grandes cohortes

## 5.1 Introduction aux effets de lots

### 5.1.1 Origine et description du problème

Les études cliniques impliquent le plus souvent l'existence de facteurs confondants, c'est-à-dire des facteurs distincts du facteur étudié, et qui pourtant, peuvent potentiellement interagir avec ce dernier. Par conséquent, s'ils ne sont pas considérés lors de l'expérimentation et de l'analyse statistique, l'effet de causalité entre le facteur étudié et les données observées sera biaisé. A ces facteurs confondants s'ajoute une grande variabilité biologique au sein des données prélevées dans un contexte clinique. Celle-ci peut être minimisée par l'analyse d'un plus grand nombre d'échantillons, mais le volume de données induit nécessite alors de gérer les analyses par lots, ce qui, nous allons le voir, constitue intrinsèquement une source de facteurs confondants.

Cette gestion « par lots » [90, 91] consiste dans la pratique à répartir des échantillons sur plusieurs plaques à puits, analyser des échantillons sur plusieurs jours, sur plusieurs instruments, faire intervenir plusieurs préparateurs pour les étapes biochimiques, *etc.* Ainsi différents lots sont-ils formés, au sein desquels règne une certaine forme d'homogénéité induite par la construction des lots elle-même. Comme nous l'avons décrit précédemment

à la Section 3.1.3, ces lots peuvent être à l'origine de variations non désirées, les effets de lots, qui peuvent interférer avec les facteurs étudiés. La variabilité observée dans les données ne dépend alors plus uniquement des facteurs d'intérêt, ce qui engendre un biais statistique si les variables qui prévalent au découpage des lots ne sont pas prises en compte dans l'analyse. Un exemple particulièrement classique est l'effet de plaque. Il est introduit lorsque le nombre d'échantillons est supérieur au nombre de puits disponibles sur une plaque, de sorte que l'utilisation de plusieurs plaques devient nécessaire. Le risque est alors que tous les échantillons d'un même groupe soient placés sur une seule plaque. Dans ce cas, si une plaque s'avère inexploitable, il en sera de même pour le groupe d'échantillons en question. Ainsi, ces lots augmentent les conséquences des aléas de la recherche : en cas de problèmes techniques, de perte de matériel, si une condition biologique est majoritairement impactée, cela peut compromettre l'analyse des résultats en aval de l'expérimentation. De la même façon, bien que les échantillons soient placés aléatoirement sur la plaque à puits, il se peut que du point de vue d'un facteur confondant (par exemple « être diabétique ou non ») tous les échantillons appartenant à un même groupe soient placés sur une même partie de la plaque. En outre, ces phénomènes ont pour conséquence directe un risque accru de non-reproductibilité des résultats expérimentaux et statistiques. En effet, dans le cadre plus précis de la recherche clinique, la découverte d'un biomarqueur de diagnostic dans une étude peut potentiellement échouer à classifier correctement d'autres cohortes de patients provenant d'autres études. En effet, il pourrait avoir été choisi sous l'influence d'un effet de lot absent dans les autres études, et sera donc incapable de généraliser ses conclusions à la population d'intérêt. Enfin, si des effets de lots ont un impact sur la variabilité observée, alors il sera plus délicat de considérer comme différentiel les entités qui ont un réel phénomène biologique. De même, des entités qui normalement n'ont pas d'effet, se retrouvent différentiels à cause de l'effet de lot, et le nombre de faux positifs augmente. Dès lors, tenir compte des effets de lots lors de l'expérimentation permet *i)* d'assurer la validité de l'expérience, *ii)* de réduire les erreurs d'interprétation lors de l'analyse statistique, et donc *iii)* de limiter la perte de puissance statistique [89, 154].

### 5.1.2 État de l'art

Trois grandes stratégies sont classiquement utilisées pour limiter l'influence des effets de lots :

La première est de répartir aléatoirement les échantillons (le néologisme « randomisation » est souvent utilisé pour désigner cela) dans des groupes définis préalablement. Concrète-

ment, lors de la préparation de l'étape analytique, la randomisation consiste à allouer aléatoirement les échantillons à des positions sur plusieurs plaques à puits tout en respectant une homogénéité des facteurs confondants (par exemple, l'âge, le sexe, présence/absence d'une pathologie particulière, *etc.*) et des variables d'intérêt entre les différentes plaques. Bien sûr, certains facteurs confondants peuvent s'avérer inaccessibles ou inconnus, de sorte qu'ils seront par définition difficilement anticipables et maîtrisables. Cependant, la randomisation reste malgré tout une des approches les plus efficaces pour en limiter les conséquences [107, 155]. En effet, la répartition des échantillons au sein des groupes formés étant aléatoire, les risques de regrouper ensemble des échantillons présentant les mêmes caractéristiques du facteur inconnu seront plus faibles. Cela permet, *in fine* et dans une certaine mesure, de limiter l'influence des effets inconnus dans les conclusions statistiques. Cependant, une simple randomisation est insuffisante car elle peut induire un nouveau biais : la possibilité d'obtenir des groupes avec des effectifs différents pour une modalité particulière de la variable d'intérêt. Il y a donc un risque non négligeable d'obtenir un plan d'expérience déséquilibré [107] (cela est d'autant plus vrai que le nombre d'échantillons est petit, ou que le nombre de facteurs confondants est élevé). Une manière classique d'éviter ce type de biais est de s'appuyer sur une randomisation un peu plus élaborée, dite « par bloc ». Cette méthode de randomisation consiste à former des blocs, chacun formé par les groupes à analyser, eux-mêmes représentés par un échantillon pris aléatoirement. Bien que cette approche soit décrite comme une solution adaptée à l'élaboration de plans d'expérience valides et robustes dès 1965 [156], elle semble peu démocratisée en protéomique [154, 157, 158].

La seconde approche possible est de normaliser les données obtenues une fois l'expérimentation réalisée, afin de corriger *a posteriori* le ou les effets de lot. La normalisation a pour objectif d'estimer quantitativement une variation non-biologique présente dans les données observées et de la retirer de ces dernières. Ceci permet de réduire la perte de puissance statistique [92, 154]. Par exemple, une différence de quantité de matériel biologique analysé peut être normalisée entre les échantillons par ajustement des différentes concentrations en divisant par la quantité totale. De nombreux travaux portent sur la manière de normaliser les données en tentant de détecter des variations non biologiques et de les prendre en compte une fois l'analyse à haut-débit réalisée [82, 88, 89, 92, 99, 159].

Enfin, une troisième approche pour limiter les conséquences des effets de lots est de réaliser une analyse en variables latentes les prenant en compte. L'analyse en variables latentes, ou analyse factorielle, fut initialement décrite par Pearson [160] au début du XX$^{\text{ème}}$ siècle

puis développée par Thurstone [161] et Everitt [162]. C'est une procédure de statistique multivariée dont l'objectif est de mettre en évidence les relations existantes entre les différentes variables observées et variables « latentes ». Une variable latente est une variable dont on suppose l'existence, elle n'est mesurable ni directement, ni à l'aide de variables observées. C'est une variable qui interfère avec les variables observées et dont les effets se répercutent sur les observations [162]. Les facteurs confondants non anticipés ou les effets de lots peuvent être assimilés à des variables latentes. Ainsi, un modèle prenant en compte l'ensemble des variables est construit. Il est ensuite estimé sur les données observées afin d'identifier la variabilité expliquée par les variables latentes et ainsi de mieux comprendre l'impact des effets de lots sur les données observées.

### 5.1.3 Motivations pour le projet NAFLD et solutions adoptées

En protéomique le placement des échantillons sur les plaques se fait classiquement à la main, principalement en raison de leur faible nombre. Cependant, cette tâche devient rapidement fastidieuse et source d'erreur lorsqu'il s'agit de plusieurs centaines d'échantillons (comme c'est le cas dans notre projet NAFLD), ou milliers, comme l'on peut supposer que cela arrivera à l'avenir sur d'autres projets de protéomique clinique. Ce placement est d'autant plus délicat lorsque l'on souhaite tenir compte de plusieurs variables d'intérêt ou confondantes, ainsi que des effets de lots. De plus, la spectrométrie de masse comporte un effet de lot spécifique, appelé le phénomène de dérive (ou « *drift* » en anglais) [157]. Ce phénomène est dû à une perte de sensibilité de détection des ions au cours du temps d'analyse. Ainsi, si ce phénomène n'est pas considéré et que l'on analyse d'abord tous les échantillons d'un groupe donné, puis tous les échantillons d'un autre groupe, les intensités mesurées pour le premier groupe seront plus élevées que celles pour le second. On pensera à tort qu'il y a une plus forte expression de protéines dans le premier groupe par rapport au second, alors qu'il s'agit en réalité d'un manque de reproductibilité de la mesure.

Nous souhaitions prendre en compte plusieurs contraintes pratiques concernant notre projet clinique. La première contrainte considérée est « l'effet de plaque », pour les raisons évoquées dans la Section 5.1.1. La seconde contrainte est le chauffage des plaques lors de l'expérimentation, qui maintien les échantillons à une température permettant la digestion trypsique (cf. Section 2.1.2) et qui n'est pas nécessairement uniforme sur une plaque à puits. De même, nous voulions éviter que des échantillons appartenant à un même groupe ne se retrouvent côte-à-côte sur la plaque à puits (que nous entendons par la suite par contrainte de voisinage), pour éviter la perte d'un même groupe en cas de défaillance

matérielle ou bien de contamination. Enfin, nous voulions placer précisément des types d'échantillons particuliers tels que les échantillons de contrôle qualité (QC) ou encore les solutions tampons afin de maîtriser les potentielles contaminations inter-échantillons. Ces règles de positionnement, à savoir la règle de voisinage ainsi que le placement précis des échantillons QC et des tampons, constituent ce que l'on appelle dans la suite de ce manuscrit les contraintes spatiales et seront détaillées dans la Section 5.2.2.

Dans cette thèse, nous nous sommes tournés vers les deux premières stratégies (randomisation et normalisation) présentées au 5.1.2, plutôt que la troisième (analyse factorielle). Notre intérêt pour la randomisation vient du fait qu'un nombre de plus en plus important de travaux décrivent la nécessité de prendre en compte les effets de lots avant l'analyse expérimentale [89, 157]. Or, aucun logiciel ne permettait jusqu'à présent de réaliser des plans de plaques automatiquement en respectant les contraintes de placement que nous nous étions fixés. Certains logiciels permettent simplement la visualisation des placements d'échantillons sur les plaques [163, 164, 165, 166, 167], tandis que d'autres sont dédiés à la correction des biais *a posteriori* [89] de l'étape analytique. Enfin, l'application la plus proche de nos besoins [168], ne permettait qu'un placement purement aléatoire des échantillons et sans pouvoir spécifier des contraintes spatiales particulières. C'est ce qui m'a conduit à réaliser la librairie R *Well Plate Maker*. Concernant la normalisation, il s'agit d'une approche facile à mettre en place, disposant d'une riche littérature proposant des méthodes élaborées et performantes [82, 88, 92], et déjà bien ancrée dans les chaînes de traitement bioinformatiques utilisées en protéomique. Dès lors, il était intéressant de fournir aux utilisateurs de nos outils, des méthodes supplémentaires adaptées à des données cliniques car elles sont généralement moins formattées que les données issues de projet de biologie fondamentale. Enfin, notre (relative) réticence à s'appuyer sur des modèles à variables latentes peut s'expliquer : il faudrait dans notre cas tenir compte d'un nombre important de variables cliniques, qui parfois ne sont pas suffisamment (ou approximativement) renseignées, au risque de rendre les résultats statistiques difficilement interprétables. De plus, nous avons peu d'échantillons par rapport au nombre de variables dont il faudrait potentiellement tenir compte, ce qui risque de poser un problème de dimensionalité qui affecterait l'estimation du modèle, et donc le rapport robustesse/puissance des statistiques finales. Enfin, la nature de ces variables et la manière de les prendre en compte dans un modèle factoriel peuvent grandement varier d'un projet à un autre, réduisant d'autant la généricité des outils que nous proposons. Dans ce contexte, et sachant qu'une thèse se réalise sur un temps limité, nous avons fait le choix de nous concentrer sur les deux premières approches plutôt que la troisième.

## 5.2 Solution en amont de l'analyse à haut-débit : création de Well Plate Maker

### 5.2.1 Une solution algorithmique : le retour sur trace

Placer les échantillons tout en considérant les contraintes spatiales telles que décrites dans la Section 5.1.3, peut être ramené à un problème de satisfaction de contraintes. Un algorithme connu pour être adapté à la résolution de ce type de problème est celui de retour sur trace (« *backtracking* » en anglais) [169] Le *backtracking* est une alternative à la recherche exhaustive de solutions, qui permet néanmoins de trouver une solution à un problème où l'on a défini des règles de décision (également appelées contraintes). Le principe général consiste à tester des possibilités afin de trouver la solution finale qui respecte l'ensemble de ces contraintes. On définit l'espace de recherche comme l'ensemble des configurations possibles et celui-ci peut être à la fois conceptualisé et représenté sous la forme d'un arbre des possibilités, que l'on va explorer par un parcours en profondeur (voir la Figure 5.1).

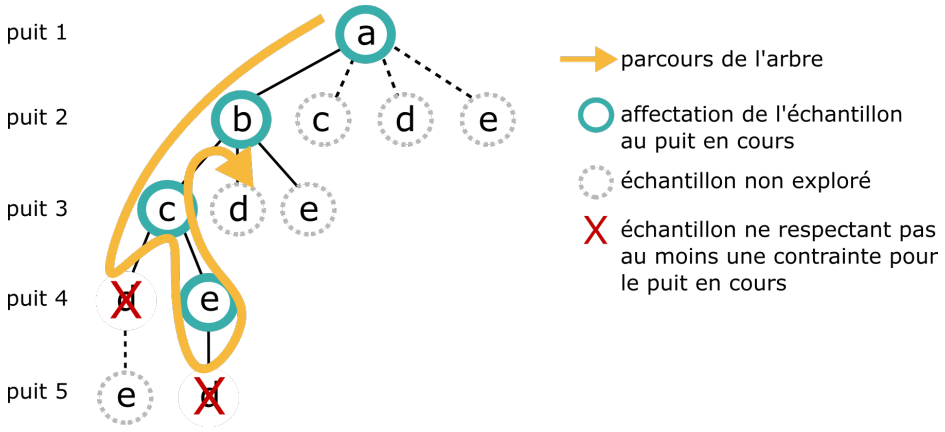



FIGURE 5.1 – Parcours en profondeur réalisé par l'algorithme de *backtracking*. Dans cet exemple, l'objectif est de placer 5 échantillons (a, b, c, d et e) dans 5 puits distincts en respectant des contraintes. L'arbre représente toutes les combinaisons possibles. Le sommet de l'arbre correspond au premier puits parcouru (initialisation) rempli avec l'échantillon a. Puis pour le puits n°2, un premier échantillon (le b) est testé parmi les 4 possibles et affecté au puits n°2, ainsi de suite jusqu'au puits n°4 où l'échantillon testé (le d) ne permet pas de compléter la solution partielle. L'algorithme retourne en arrière (il revient au puits 3 avec l'échantillon c), c'est-à-dire qu'il désaffecte l'échantillon d pour le puits n°4 et explore l'échantillon e qui mène aussi à un échec. L'algorithme retourne au puits n°2 et part explorer la branche suivante.

Le *backtracking* procède par incrémentation, autrement dit à chaque itération l'algorithme construit progressivement une solution partielle et l'étend jusqu'à obtenir une solution complète, en abandonnant les affectations qui ne peuvent pas être complétées car ne

respectant pas au moins une des contraintes définies. En d'autres termes, l'algorithme va éliminer l'examen de certaines configurations par élagage de l'espace de recherche. De cette manière, cette approche converge beaucoup plus rapidement vers une solution qu'une simple recherche par force brute (ou recherche exhaustive). Concrètement pour le remplissage d'une plaque à puits, l'initialisation consiste à choisir un puits et à lui affecter un échantillon, puis le puits voisin est rempli avec un autre échantillon qui respecte les contraintes choisies, et ainsi de suite jusqu'à ce que tous les échantillons soient correctement placés. Classiquement, l'algorithme parcourt la plaque de manière séquentielle en ligne ou en colonne, mais mon approche consiste à combiner le *backtracking* et la randomisation. L'initialisation de l'algorithme et le parcours de la plaque sont aléatoires. Cette astuce permet de résoudre en partie le problème d'effet de lot en cas de plan déséquilibré car elle permet d'éviter des motifs en forme de « damier » (Figure 5.2). En effet, si le parcours de la plaque est linéaire et que certains groupes sont surreprésentés, alors la partie de la plaque parcourue en dernier ne sera remplie qu'avec les échantillons de ces groupes. Ce phénomène est illustré dans la Figure 5.2, où l'on a représenté une plaque de 96 puits remplie avec des échantillons appartenant à des groupes différents (symbolisés par des couleurs différentes).

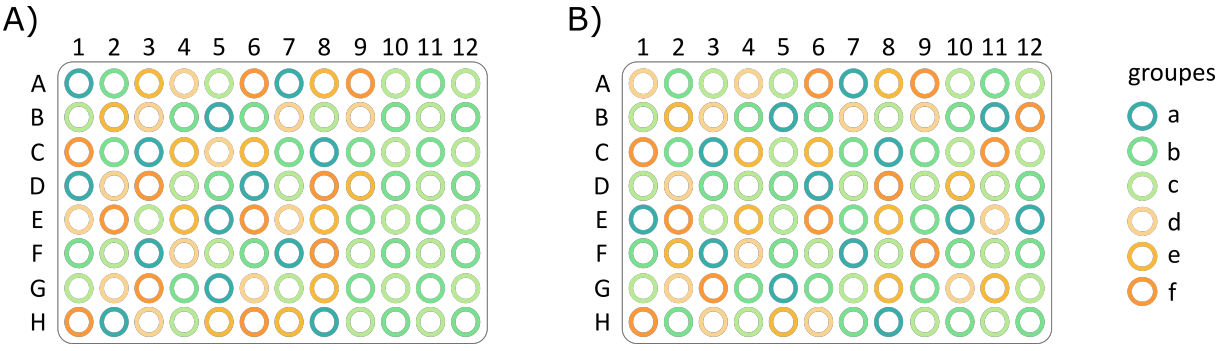



FIGURE 5.2 – Deux plaques à 96 puits contenant des échantillons appartenant à 6 groupes différents avec comme effectifs respectifs (plan déséquilibré) 13 échantillons dans le groupe a, 24 dans le groupe b, 25 dans le groupe c, 12 dans le groupe d, 11 dans le groupe e et 11 dans le groupe f. **A)** La plaque a été parcourue de façon linéaire (de haut en bas et de gauche à droite) et les échantillons tirés aléatoirement (avec la même probabilité de tirage pour chaque groupe), ce qui a conduit à une fin de remplissage uniquement composée d'échantillons vert clair et vert foncé générant un effet de lot. **B)** Le parcours de la plaque est aléatoire (en plus du choix aléatoire de l'échantillon à placer) générant une plaque plus homogène et minimisant ainsi l'effet de lot.

L'algorithme proposé consiste en un parcours en profondeur non récursif de l'arbre, associé à un ensemble de contraintes définies que sont les contraintes spatiales (Section 5.1.3). Il ne s'agit donc pas d'un *backtracking* à proprement parler, et lors d'un échec d'affec-

tation, l'algorithme réinitialise l'arbre à parcourir. C'est un inconvénient notoire, car ce fonctionnement ne nous assure pas d'explorer l'ensemble des solutions possibles, de sorte que la solution peut ne pas être trouvée. Cependant, cette exploration partielle permet de réduire les temps de calculs ; par ailleurs puisque l'initialisation et le parcours d'arbre sont aléatoires, relancer l'algorithme après un échec peut permettre de parcourir des arbres non parcourus dans un premier temps et de trouver une solution. C'est pourquoi Well Plate Maker (*wpm*) propose par défaut de relancer l'algorithme 20 fois. Ce paramètre est modifiable par l'utilisateur en fonction de la complexité des plans à réaliser, en sachant que le nombre de tentatives nécessaires augmente avec la complexité. La procédure générale de l'algorithme est représentée par le pseudo-code suivant (voir Algorithme 1) dans le cas du remplissage d'une seule plaque. Le code développé est utilisable dans le cadre d'une expérience nécessitant la mise en place de plans de plaques relativement complexes. Ce travail a pu donner lieu à une publication dans *Bioinformatics* [153] et un déploiement de la librairie sous le nom de *wpm*, sur le dépôt Bioconductor (`http://www.bioconductor.org/packages/release/bioc/html/wpm.html` ).

**Algorithme 1** Procédure générale simplifiée de *wpm* pour réaliser le remplissage d'une plaque à puits. Les paramètres en entrée : $d$ contient les données de l'utilisateur concernant les échantillons à placer ; *max_iter* est le nombre maximum d'itération que l'algorithme est autorisé à réaliser ; *contrainte_voisinage* est la contrainte spatiale choisie par l'utilisateur.

1 : **procedure** REMPLIR UNE PLAQUE($d$,*nombre_lignes*, *nombre_colonnes*, *max_iter*, *contrainte_voisinage*)
2 : *nombre_iteration* $\leftarrow 0$
3 : **while** *résultat* $= 1$ **ET** *nombre_itération* $\leq$ *max_iter* **do**
4 : Initialiser *matrice* avec des puits vides, aux dimensions (*nombre_lignes*, *nombre_colonnes*)
5 : Initialiser *puits_à_visiter* (puits vides de *matrice*)
6 : **if** *puits_interdits* **then**
7 : *puits_à_visiter* $\leftarrow$ *puits_à_visiter* $-$ *puits_interdit*
8 : **end if**
9 : *puits_visités* $\leftarrow$ *NULL*
10 : *échantillons_à_placer* $\leftarrow$ nombre d'échantillons de $d$
11 : *échantillons_placés* $\leftarrow$ *NULL*
12 : **while** nombre d'éléments dans *puits_visités* ! = nombre d'éléments dans *échantillons_à_placer* **do**
13 : *puit* choisi aléatoirement parmi les *puits_à_visiter*
14 : Ajouter *puits* à *puits_visités*
15 : $i \leftarrow$ numéro de ligne de *puits*
16 : $j \leftarrow$ numéro de colonne de *puits*
17 : Identifier les échantillons voisins de *puits* en fonction de *contrainte_voisinage*
18 : Identifier le groupe d'appartenance de chaque échantillon voisin de *puits*
19 : *groupes_possibles* $\leftarrow$ les groupes auxquels les échantillons voisins n'appartiennent pas
20 : **if** *groupes_possibles* $\leftarrow$ *NULL* **then**
21 : **return** *résultat* $= 1$
22 : **else**
23 : **if** pas d'échantillon possible parmi *groupes_possibles* **then**
24 : **return** *résultat* $= 1$
25 : **else**
26 : Prendre aléatoirement un *échantillon_choisi* appartenant à *groupes_possibles* et qui n'est pas dans *échantillons_placés*
27 : Ajouter *échantillon_choisi* dans *matrice* selon $(i, j)$
28 : Ajouter *échantillon_choisi* à *échantillons_placés*
29 : **end if**
30 : **end if**
31 : *Ligne* de $d \leftarrow i$ ; *Colonne* de $d \leftarrow j$
32 : Retirer *puits* de *puits_à_visiter*
33 : *résultat* $\leftarrow d$
34 : **end while**
35 : **if** *résultat* !=1 **then**
36 : **return** *résultat*
37 : **end if**
38 : *nombre_itération* $\leftarrow$ *nombre_itération* $+ 1$
39 : **end while**
40 : **end procedure**

### 5.2.2 Description du programme développé

Le fonctionnement de *wpm* repose fondamentalement sur la définition de quatre types de puits que l'on rencontre très souvent en expérimentation : *i)* Un puits ***interdit*** ne sera rempli d'aucun type d'échantillon, soit parce que l'utilisateur ne le souhaite pas (par exemple, les extrémités de la plaque en cas de distribution de chaleur non uniforme), soit en raison de contraintes matérielles (par exemple, puits sales, pipettes cassées). ; *ii)* Un puits ***tampon*** correspond à un puits rempli de solution mais sans matériel biologique. Cela permet par exemple d'éviter ou de vérifier la contamination croisée. ; *iii)* les puits qui contiendront les ***échantillons fixes*** correspondant à un échantillon ou un standard de contrôle qualité. La localisation précise de ces échantillons doit être contrôlée par le chercheur ; en dernier lieu *iv)* les puits qui contiendront les ***échantillons biologiques à randomiser***.

*Wpm* gère ces types de puits par ordre de priorité décroissante, autrement dit, si un puits (par exemple le puits A1 situé en haut à gauche d'une plaque) est spécifié par l'utilisateur aussi bien comme un puits *interdit* qu'un puits *tampon*, alors *wpm* considérera ce puits comme *interdit*. En effet, on estime la catégorie *interdit* comme la plus prioritaire des quatre car il s'agit de puits dans lesquels l'expérimentateur ne veut pas ou ne peut pas faire d'expérimentation. De même, un puits défini comme *interdit* ou *tampon* ou *fixe*, ne pourra pas contenir d'échantillon biologique et ainsi de suite. Concernant les solutions tampons, il est possible de choisir le type de placement souhaité entre un positionnement par ligne, par colonne, en damier ou bien manuel, qui ont été représentés dans la Figure 5.3. Le placement choisi sera identique pour toutes les plaques s'il y en a plusieurs à remplir.

On appelle contrainte de *voisinage* une règle contrôlant le positionnement d'un échantillon en fonction de son groupe d'appartenance. Cela permet de maîtriser l'effet de lot dû aux facteurs confondants pris en compte par la définition des différents groupes. Le principe consiste à spécifier si des échantillons appartenant au même groupe peuvent être placés côte-à-côte ou non. Actuellement *wpm* propose 4 types de contraintes de *voisinage* qui sont résumées dans la Figure 5.4.

On spécifie le nombre de plaques nécessaires ainsi que le facteur confondant que l'on souhaite contrôler lors du placement des échantillons. C'est ce facteur qui détermine les groupes pour les échantillons. Le nombre d'échantillons à placer sur chaque plaque est alors calculé afin de les répartir équitablement en fonction des groupes et ainsi obtenir les plans de plaques les plus équilibrés possibles. Par exemple, s'il y a 2 plaques à remplir, alors

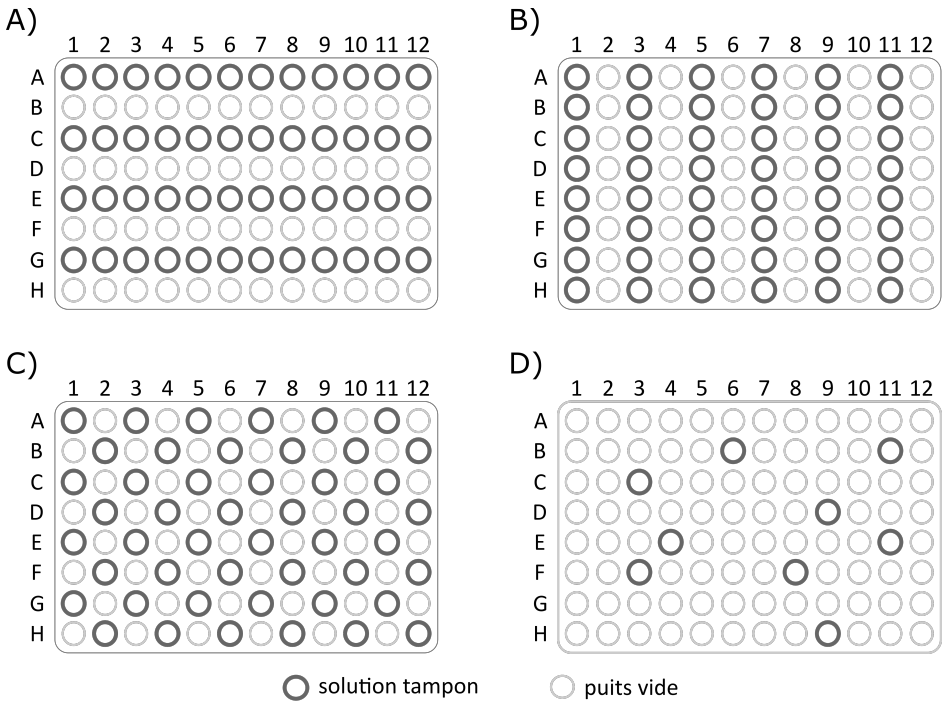



FIGURE 5.3 – Schéma des différentes configurations possibles de solutions tampons pour une plaque de 96 puits. **A)** répartition en lignes. **B)** répartition en colonnes. **C)** répartition en damier. **D)** répartition manuelle. Pour les cas **A)** et **B)**, il est possible de choisir une répartition sur les rangs pair ou impair.

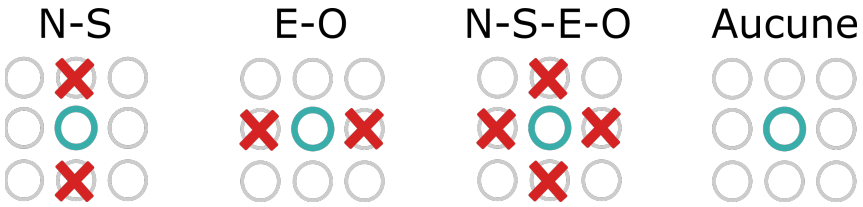



FIGURE 5.4 – Les quatre possibilités de *voisinage* pour le positionnement des échantillons selon leur appartenance à un groupe. En bleu l'échantillon actuellement considéré. La croix rouge symbolise une interdiction. La règle NS (Nord-Sud) correspond à l'interdiction de placer des échantillons du même groupe que celui considéré sur l'axe de la colonne. EO (Est-Ouest) interdit de placer sur l'axe de la ligne. NSEO (Nord-Sud-Est-Ouest) interdit de placer sur les deux axes ligne et colonne. La quatrième possibilité (aucune interdiction) consiste à autoriser toutes les positions.

50% des échantillons de chaque groupe se retrouveront sur chaque plaque. Si le nombre d'échantillons et le nombre de plaque impliquent des effectifs différents pour les plaques, alors *wpm* se rapprochera le plus possible des proportions idéales (par exemple pour deux plaques et 175 échantillons, il y aura respectivement 87 et 88 échantillons placés sur chaque plaque). En cas d'effectifs d'échantillons différents au sein des groupes, *wpm* fera également en sorte que le nombre d'échantillons soient les plus équitables possibles. Cela permet de compenser partiellement le plan d'expérience déséquilibré car on obtient une certaine homogénéité d'échantillonnage entre les plaques. Cependant cette correction reste

limitée, et ne permet pas toujours de compenser un plan d'expérience trop déséquilibré, que l'utilisateur doit malgré tout veiller à éviter. A titre d'exemple, la Figure 5.5 montre les plans obtenus avec *wpm* pour 192 échantillons de patients fictifs placés sur deux plaques de 96 puits. Les groupes ont été définis en fonction du facteur « diabète » (présence [1] / absence [0]) et du facteur « âge » (5 tranches d'âge). Les différents effectifs par sous-catégorie de diabète et d'âge sont détaillés dans le tableau 5.1 ci-après. Le plan expérimental est volontairement déséquilibré pour vérifier que *wpm* est capable de trouver une solution pour ce type de plan complexe.

| | | **Catégorie d'âge** | | | | | |
|---|---|---|---|---|---|---|---|
| | | 20-30 | 30-40 | 40-50 | 50-60 | 60-70 | **Total** |
| **Diabète** | Non [0] | 21 | 13 | 15 | 20 | 21 | 90 |
| | Oui [1] | 25 | 25 | 17 | 17 | 18 | 102 |
| | **Total** | 46 | 38 | 32 | 37 | 39 | 192 |

TABLEAU 5.1 – Table de contingence du jeu de données fictives pour tester *wpm*. On peut donc définir 10 groupes distincts grâce aux deux facteurs « diabète » et « âge ». Lecture : un premier groupe rassemble les 21 patients non-diabétiques (Non [0]) de 20-30 ans

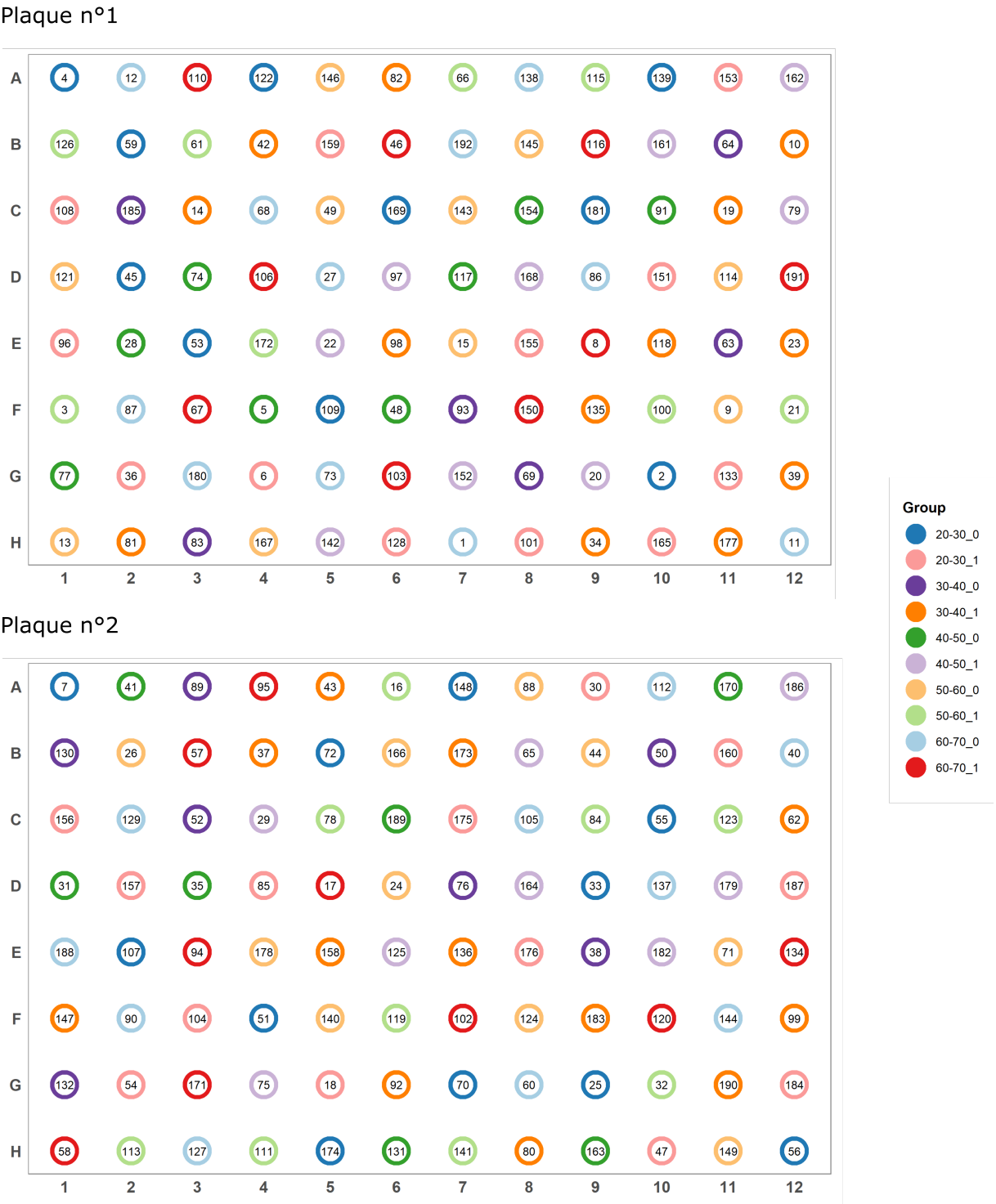



FIGURE 5.5 – Plans obtenus avec *wpm* pour 2 plaques de 96 puits où l'on a placé aléatoirement 192 patients appartenant à 10 groupes distincts. Dans cet exemple fictif, les groupes ont été défini selon une variable appelée « Group » qui rassemble les facteurs confondants âge (5 tranches d'âge) et diabète (présence/absence). Par exemple le groupe « 20-30_0 » correspond à un patient non diabétique classé dans la catégorie 20-30 ans. Les échantillons ont été placés selon la contrainte spatiale NSEO.

Certaines améliorations de *wpm* seraient souhaitables, notamment la modification de l'algorithme actuel en une version récursive plus proche du véritable *backtracking* et ainsi fournir une exploration plus complète et plus optimisée de l'espace des solutions possibles. Il serait également intéressant de proposer, pour les projets de protéomique, une option de sérialisation des échantillons pour l'injection dans un spectromètre de masse, dans le même esprit que celle proposée dans [157]. Cette option est relativement simple à implémenter et permettrait de faciliter cette étape fastidieuse pour les expérimentateurs. Dans le cas de l'étude NAFLD, nous avons alterné des échantillons de la première plaque et de la seconde plaque en effectuant une lecture séquentielle de ces dernières (de gauche à droite pour la première et de droite à gauche pour la seconde), et cette approche pourrait facilement être généralisée via une fonctionnalité supplémentaire dans *wpm*. Enfin, prendre en compte automatiquement plusieurs facteurs confondants permettrait de faciliter la saisie des paramètres, car actuellement ce sont les utilisateurs qui doivent faire ce travail de regroupement de facteurs. A titre d'exemple, si un utilisateur souhaite tenir compte des facteurs *diabète* et *âge*, qui existent en tant que deux colonnes distinctes dans son fichier, il faut créer une colonne supplémentaire dans le fichier qui contiendra la jointure des deux colonnes précédentes, comme cela est présenté dans la Figure 5.5. Automatiser cela permettrait donc à l'utilisateur de gagner un peu de temps.

### 5.2.3 Comparaison avec des logiciels similaires

Il existe deux logiciels, décrits rapidement ci-après, aboutis et qui ont le même objectif que *wpm* de contrôler des effets de lots lors de la réalisation des plans de plaques à puits.

Le premier, *PlateDesigner*, existe depuis 2019 [168] et est accessible via une page web (https://clinicalresearch-apps.shinyapps.io/platedesigner/). La principale différence avec *wpm* est la stratégie de placement des échantillons. En effet, le remplissage des plaques est purement aléatoire dans *PlateDesigner*, sans possibilité de spécifier des contraintes spatiales. Or, pour les raisons évoquées dans la section 5.1.2, cette approche simpliste présente le risque de proposer un contrôle limité de certains effets de lot. En revanche, il est possible dans *PlateDesigner* de regrouper tous les réplicats techniques d'un échantillon biologique sur la même plaque, ce que ne propose pas *wpm*. Par ailleurs, contrairement à *wpm*, il n'est pas possible de placer des échantillons de contrôle qualité ou des tampons, ni d'interdire des puits pour le placement, ce qui limite le contrôle du chercheur sur de potentiels problèmes techniques et/ou pratiques.

Le second logiciel est *Omixer*, également disponible sur le Bioconductor. Cependant, il

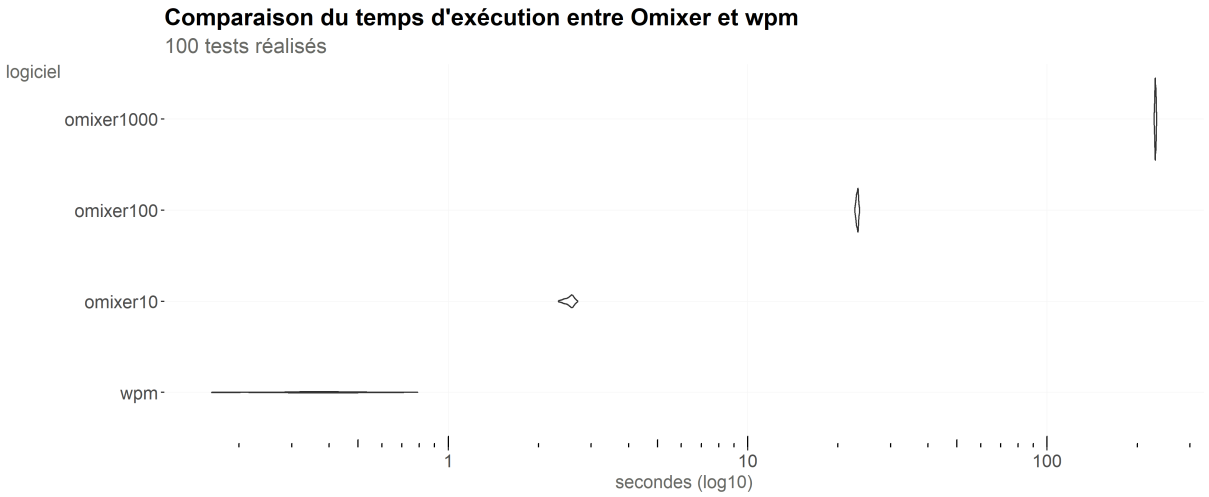
fut publié peu de temps après *wpm* [170], de sorte que ses auteurs, probablement ignorant l'existence de *wpm*, n'ont pas comparé les performances de leur outil avec *wpm*, comme il est normalement d'usage. Le principal inconvénient d'*Omixer* par rapport à *wpm* est son absence d'interface graphique pour l'utilisateur (GUI). Il apparaît également qu'*Omixer* ne permet pas de réaliser des plans de plaque lorsque le nombre de puits est supérieur au nombre d'échantillons à placer, tandis que *wpm* gère ce cas de figure. *Omixer* ne permet de spécifier que des puits interdits, mais ne prend pas en compte d'autres types tels que les solutions tampons et les échantillons QC. Enfin, contrairement à *wpm*, il ne veille pas à ce que les groupes soient répartis équitablement entre les plaques. Autrement dit le nombre d'échantillons pour un groupe donné sera déséquilibré entre les plaques, ce qui a été souligné comme un risque de biais d'expérimentation dans la Section 5.1.2. *Omixer* réalise donc un placement purement aléatoire des échantillons comme *PlateDesigner*, avec la seule particularité de compenser cette approche par la sélection du plan pour lequel le test des corrélations de Pearson présente les meilleurs résultats pour l'ensemble des facteurs confondants testés. Autrement dit, *Omixer* effectue pour chaque plan généré des tests de corrélation entre les variables techniques (par exemple le facteur « plaque ») et les facteurs confondants sélectionnés (par exemple « diabète » ou « âge »). Un plan parmi tous ceux qui ont été générés est choisi en fonction de deux critères. Le premier critère est le résultat statistique des tests de corrélation. Si l'ensemble des tests obtiennent une $p_{valeur}$ supérieure à 0,05, alors le plan est considéré comme valide car aucun test n'a suggéré une corrélation entre les variables. Le second critère permet de départager les plans qui remplissent le premier critère. Ainsi, le plan considéré comme le meilleur, sera celui pour lequel la somme des valeurs absolues des coefficients de corrélation est la plus faible.

Étant donné qu'*Omixer* réalise ces tests de corrélation pour chaque plan généré, il a donc besoin de réaliser un grand nombre de plaques pour espérer obtenir les meilleurs résultats statistiques (par défaut il génère 1000 plans) et sélectionner ainsi un plan final considéré de bonne qualité. Outre le nombre plus restreint d'options sur les contraintes de placement et l'interface graphique (qui sont en elles-mêmes importantes pour les utilisateurs), la principale différence conceptuelle entre *Omixer* et *wpm* est donc la méthode de sélection du plan optimal. *Omixer* adopte une approche statistique tandis que *wpm* utilise une approche algorithmique (respect de l'ensemble des contraintes spatiales). Le Tableau 5.2 ci-dessous approfondi la comparaison des trois logiciels.

| Logiciel | PlateDesigner | Omixer | wpm |
|---|---|---|---|
| GUI | oui | non | oui |
| Algorithme pour générer une plaque | Remplissage aléatoire / réplicats placés ensemble | Remplissage Aléatoire | Backtracking + Aléatoire |
| Gère plus d'une plaque | Oui | Oui | Oui |
| « effet plaque » maîtrisé | Oui/Non selon l'option de randomisation choisie | Non | Oui |
| Gestion des facteurs confondants | Oui | Oui | Oui |
| Méthode pour sélection du plan optimal | Aucune | Statistique | Algorithmique |
| Cases vides possibles | Oui | Non | Oui |
| Puits interdits possibles | Non | Oui | Oui |
| Choix de positionnement des puits interdits | Non | Oui | Oui |
| Prise en compte d'échantillons particuliers (contrôle, tampon) | Oui | Non | Oui |
| Choix de positionnement des éch. particuliers | Non | Non | Oui |

TABLEAU 5.2 – Tableau comparatif des trois logiciels (*PlateDesigner*, *Omixer* et *wpm*). *PlateDesigner* apparaît plus limité que les deux autres approches pour la réalisation des plans de plaque. *wpm* se démarque principalement par son approche algorithmique pour générer les plans, contrairement aux deux autres logiciels qui réalisent un placement purement aléatoire.

Nous avons comparé les performances de temps d'exécution d'*Omixer* et de *wpm* avec la librairie *microbenchmark* [171] sur un ordinateur Windows10 i-7 avec 32 Go de RAM. Le jeu de données qui a servi de test est le même que celui qui a été décrit dans la Section 5.2.2 pour générer la Figure 5.5. Un test correspond à l'exécution du code R d'une librairie afin d'obtenir les deux plans de plaque adaptés au jeu de données. Nous avons donc mesuré le temps nécessaire à la réalisation d'un test, et les résultats pour 100 tests sont reportés dans le tableau de la Figure 5.6. Le terme itération est employé différemment par *Omixer* et *wpm*. Pour *Omixer*, une itération (le paramètre `iterNum`) correspond à la génération d'un plan pour les deux plaques. Pour *wpm*, une itération correspond à une tentative pour générer chaque plan. Concrètement, pour le test où il faut générer deux plans de plaque, exécuter Omixer avec `iterNum` = 10 signifie donc qu'il va générer 10 paires de plans, et ne garder que la paire de plans obtenant les meilleurs résultats statistiques. En revanche, exécuter *wpm* avec `max_iter` = 10 signifie qu'il aura la possibilité de relancer l'algorithme au maximum 10 fois uniquement en cas d'échec d'une tentative précédente. Par exemple, exécuter *wpm* avec `max_iter`=10 ou `max_iter`=20 peut donner le même temps d'exécution si jamais il trouve les plans dès l'itération n°1 dans les deux cas. Étant donné cette différence de définition, nous avons choisi de conserver le paramètre par défaut de *wpm* (`max_iter`=20) et nous avons testé le temps d'exécution pour trois valeurs d'`iterNum` pour *Omixer* : 10, 100 et 1000 (valeur par défaut du paramètre). La librairie *microbenchmark* mesurant les temps d'exécution en nanosecondes, les différents résultats présentés dans la Figure 5.6 ont été converti en secondes pour plus de lisibilité.




| logiciel | nombre de test | temps min (sec.) | temps moyen (sec.) | temps médian (sec.) | temps max (sec.) |
|---|---|---|---|---|---|
| wpm | 100 | 0.1621557 | 0.3828389 | 0.3810751 | 0.790824 |
| omixer10 | 100 | 2.3340433 | 2.5402299 | 2.5590801 | 2.703793 |
| omixer100 | 100 | 22.8032156 | 23.2392711 | 23.2646217 | 23.616866 |
| omixer1000 | 100 | 227.7207232 | 230.0429346 | 229.9619226 | 232.022072 |

FIGURE 5.6 – Analyse comparative des performances d'exécution d'*Omixer* et *wpm* pour 100 tests réalisés avec la librairie *microbenchmark*. Le temps d'exécution des tests est présenté en secondes. Exemple de lecture : Omixer10 : *Omixer* exécuté avec le paramètre `iterNum=10` ; Omixer100 : `iterNum=100` ; Omixer1000 : `iterNum=1000`. *wpm* a été exécuté avec son paramétrage par défaut (`max_iter=20`).

Nous pouvons constater que *wpm* est plus rapide qu'*Omixer* comme le présente la Figure 5.6. Cela est principalement dû à la manière de procéder d'*Omixer*. En effet, il réalise d'abord l'ensemble des plans de plaques puis l'ensemble des tests de corrélations de Pearson, tandis que *wpm* se contente de retourner un résultat dès que deux plans de plaques compatibles ont été trouvés. L'intervalle de temps important entre le temps d'exécution minimal et maximal de *wpm* s'explique par le nombre de tentatives qu'il a réalisé avant de trouver les plans. Le temps minimal de 0,16 secondes correspond à une situation où *wpm* a trouvé un plan dès la première itération, et le temps maximal de 0,79 secondes correspond à des plans trouvés après 10 tentatives. Nous avons ensuite comparé les statistiques de corrélation obtenues pour les plans retenus par *Omixer* et *wpm*. Pour cela, nous nous sommes basés sur les critères d'*Omixer*, et avons utilisé la fonction `OmixerCorr()` d'*Omixer* afin de calculer les corrélations et les $p_{valeurs}$ associées aux tests de Pearson entre le facteur « plaque » et les facteurs confondants « âge » et « diabète ». Les résultats pour 10 tests sont présentés en Figure 5.7. Les résultats Omixer10, Omixer100 et Omixer1000 correspondent aux mêmes tests que ceux réalisés pour évaluer les performances d'exé-

cution des deux librairies (cf. Figure 5.6). Le résultat nommé wpm1 sur la Figure 5.7 correspond à un test où l'on exécute *wpm* qu'une seule fois, il n'y a donc qu'un seul plan généré à chaque test. Pour compléter la comparaison entre *Omixer* et *wpm*, nous avons réalisé wpm100, qui correspond aux résultats obtenus pour 10 tests pour lesquels nous avons sélectionné le meilleur plan parmi 100 plans générés avec *wpm*. Il s'agit donc d'un test équivalent à Omixer100.

Nous pouvons observer sur la Figure 5.7 que les coefficients de corrélation (Figure 5.7A) obtenus avec Omixer1000 sont tous très proches de 0, tandis que ceux obtenus avec Omixer10 se répartissent sur un intervalle plus large. Par conséquent, et conformément à ce à quoi nous pouvions nous attendre, plus *Omixer* génère un nombre important de plans, meilleurs en résultent les corrélations retenues. Concernant les coefficients de corrélation obtenus par wpm1 (Figure 5.7A), celui-ci fournit de meilleurs résultats qu'Omixer10, et des résultats équivalents à ceux obtenus avec Omixer100. Mais les coefficients obtenus avec Omixer1000 sont nettement plus faibles, bien que dans l'ensemble, les résultats de wpm1 restent très corrects, avec des coefficients de corrélation compris entre -0,021 et 0,022 (Figure 5.7A) et des de test de corrélation compris entre 0,726 et 0,959 (Figure 5.7B). L'ensemble des résultats suggèrent d'utiliser *Omixer* avec au minimum 1000 itérations afin d'être certain d'obtenir des plans de plaque de qualité. Concernant les résultats de wpm100 (Figure 5.7A et 5.7B), ils sont nettement meilleurs que ceux de wpm1 et d'Omixer1000, aussi bien concernant les coefficients de corrélation que les $p_{valeurs}$. Nous noterons que les coefficients obtenus sont compris entre -0,02 et 0,05, quelle que soit l'approche utilisée. Ces valeurs étant très faibles, la différence de qualité des plans obtenus avec *Omixer* et *wpm* est donc relativement négligeable. Il s'avère que wpm1 est 600 fois plus rapide que Omixer1000, et que ce dernier ne permet d'améliorer que faiblement les résultats de corrélation (de l'ordre de $10^{-2}$ pour l'exemple présenté). Par conséquent, l'ensemble des résultats montrent que les plans obtenus avec *wpm* sont de bonne qualité et cela dans un temps d'exécution minime, et que la stratégie algorithmique est adaptée pour maîtriser les effets de lots en aval de l'expérimentation. Par ailleurs, l'approche algorithmique retenue dans *wpm* permet de prendre en compte un plus grand nombre de contraintes, ce qui avec son interface graphique, devrait permettre une adoption plus facile par la communauté des utilisateurs Néanmoins, la qualité des plans de *wpm* pourrait éventuellement être améliorée en intégrant à notre librairie un système de sélection inspiré de celui d'*Omixer*, basé sur la génération de plusieurs plans, parmi lesquels un seul est retenu selon des critères statistiques.

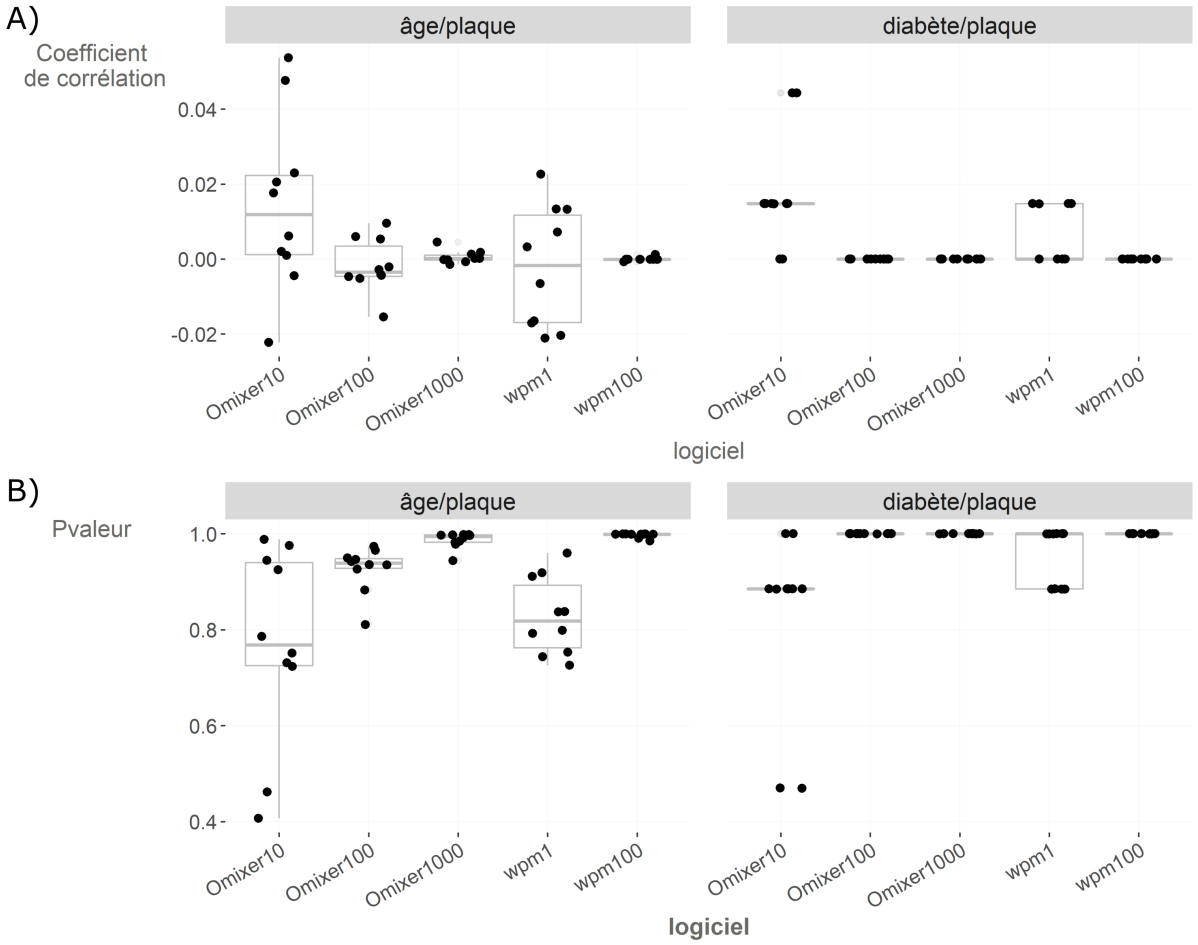



FIGURE 5.7 – Comparaison des résultats des test de corrélations entre les facteurs confondants (« âge », « diabète ») et les plaques pour *Omixer* et *wpm* pour le jeu de données fictif présenté en section 5.2.2. Chaque test est représenté par un point sur les boîtes à moustache (10 tests pour chaque logiciel). Pour Omixer10, Omixer100, Omixer1000 et wpm100 les 10 points représentent les 10 meilleurs plans obtenus. **A)** coefficients de corrélation des 10 tests réalisés avec *Omixer* et *wpm*. **B)** $p_{valeurs}$ des tests de corrélation pour chaque logiciel.

## 5.3 Solution en aval de l'analyse protéomique

Dans le cadre du projet NASH, j'ai proposé l'utilisation (puis l'implémentation dans le logiciel Prostar) de deux méthodes de normalisation, déjà connues en transcriptomique, mais permettant des corrections plus subtiles que celles utilisées classiquement au laboratoire. Bien que ces dernières soient tout à fait adaptées à des projets « fondamentaux », elles sont insuffisantes pour des données cliniques acquises dans des conditions expérimentales moins bien contrôlées. C'est deux méthodes sont « *LOcally weighted regrESSion* » (LOESS) [172, 173] et « *variance stabilization normalization* » (vsn) [79, 174, 175], et leur

choix a été motivé par trois principales raisons. Tout d'abord, ces approches sont robustes face aux valeurs manquantes, c'est-à-dire qu'elles assurent une normalisation appropriée des données malgré l'absence de certaines valeurs. Cette robustesse est une propriété indispensable étant donné la grande proportion de valeurs manquantes en protéomique *label-free* (cf Section 3.1.3) ; sachant que dans le pipeline de *Prostar*, l'imputation est réalisée après la normalisation. Ensuite, ces deux méthodes permettent de préserver suffisamment la variabilité biologique entre les différentes modalités étudiées tout en réduisant une variabilité non-biologique qui serait non négligeable [92]. Enfin, ces deux méthodes ont été testées sur des données de protéomique *label-free* [92] avec des performances satisfaisantes.

L'objectif de cette section est d'étudier l'applicabilité à la protéomique clinique de ces deux méthodes et non pas de réaliser uniquement un état de l'art sur ces dernières.

### 5.3.1 Le choix LOESS

La régression locale, ou LOESS (« LOcally weighted regrESSion ») [172], est une méthode de régression non paramétrique [176] qui repose sur la méthode des $k$ plus proches voisins. La LOESS combine la régression linéaire par les moindres carrés avec la régression non linéaire, en effectuant une régression sur des sous-ensembles locaux de données. La LOESS peut être une alternative possible aux méthodes habituelles de régression, telle que par exemple la régression par les moindres carrés. Le principe de la LOESS a été initialement décrit par Cleveland [177], puis développé par Cleveland [178] et Cleveland et Devlin [172]. Il consiste en la construction d'un modèle de régression ajusté à l'ensemble de points, à partir d'un ensemble contigu de polynômes locaux calculés sur des sous-ensembles de points. Soit un ensemble de points $(x, y)$ représentant les données, avec $X$ et $Y$ deux variables continues. L'objectif est de modéliser la relation existante entre ces deux variables. Les points $(x, \hat{y})$ représentent la fonction polynomiale de la régression LOESS associée à ces données. La LOESS est non paramétrique dans le sens où elle ne nécessite pas que la fonction décrivant la relation entre les variables soit définie à l'avance.

Pour calculer la valeur du polynôme local au point d'abscisse $x_i$ , noté $\hat{y}_i$, LOESS utilise trois paramètres : la fenêtre de voisinage notée $m$ ; le degré du polynôme de la LOESS, noté $d$ ; et la fonction de pondération, notée $W$. La fenêtre $m$, est une valeur comprise entre 0 et 1, et représente le nombre d'observations de $X$ considérés comme les plus proches voisins de chaque observation $x_i$. Par exemple, si $m$=0,5 alors la régression sera modélisée avec 50% des points les plus proches d'un point donné $(x_i, y_i)$. Le degré $d$ du polynôme de régression détermine l'estimation $\hat{y}_i$ au point $x_i$. En effet, si $d$=1 alors l'estimation se fera

à l'aide d'une régression linéaire, tandis que si $d$=2 , la régression utilisée pour calculer $\hat{y}_i$ sera quadratique.

L'ajustement de la LOESS nécessite une pondération des points impliqués dans la régression, c'est-à-dire que les points les plus proches de $x_i$ auront davantage de poids dans l'ajustement que les points les plus éloignés. Ainsi pour chaque $x_i$, les poids $w_k(x_i)$ sont définis pour tous les $x_k$ avec $k = 1, \ldots, n$ et sont calculés grâce à la fonction de pondération $W$ qui est définie par une fonction tricube telle que :

$$W(x) = \begin{cases} (1 - x^3)^3 & \forall |x| < 1 \\ 0 & \forall |x| \geq 1 \end{cases} \tag{5.1}$$

La fonction tricube a été choisie par les auteurs car elle permet une estimation robuste de la régression, c'est-à-dire que la présence de valeurs extrêmes dans les données influencera peu la LOESS [177]. Pour plus de détails sur la procédure, notamment concernant la définition des poids dits robustes $\delta_k w_k(x_i)$, se référer à la démonstration présentée dans l'article de Cleveland [177]. La Figure 5.8 illustre l'influence du paramètre $m$ dans le calcul de la LOESS sur un nuage de points $(x, y)$ obtenus où $Y$ est définie comme $Y = sin(2\pi X) + z$ où et $z \sim U[-1, 1]$, $U[a, b]$ étant la distribution uniforme définie sur l'intervalle $[a, b]$. On constate que plus la valeur de $m$ augmente, plus la courbe est lissée.

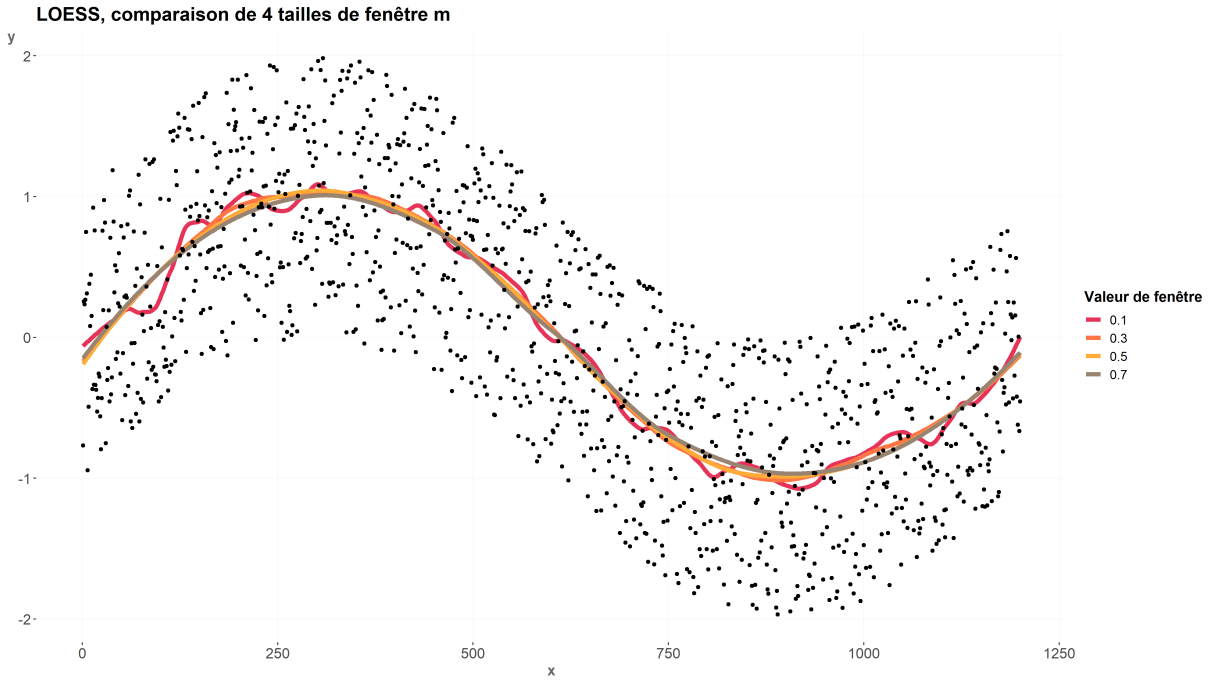



FIGURE 5.8 – Comparaison de différentes valeurs de largeur de fenêtre pour une régression LOESS sur un nuage de points définis par deux variables continues $X$ et $Y$.

La relation existante entre deux échantillons donnés peut être facilement visualisée à l'aide d'un MA-plot [179]. Ce graphique permet de visualiser les différences entre les mesures prises dans deux échantillons, en transformant les données en échelles $M$ (ratio entre les données log-transformées) et $A$ (la moyenne des données log-transformées). Soient deux échantillons $i$ et $j$, $M_{ij}$ et $A_{ij}$ sont calculées pour les abondances mesurées $x$ :

$$M_{ij} = \log_2 \left( \frac{x_i}{x_j} \right) \tag{5.2}$$

$$A_{ij} = \log_2(x_i x_j) \tag{5.3}$$

La normalisation par l'approche LOESS interprète comme un biais tout écart systématique dans le nuage de points par rapport à la droite en zéro sur l'axe $M$. La normalisation LOESS procède en deux étapes. Dans un premier temps, elle modélise la relation entre $A_{ij}$ et $M_{ij}$ en effectuant une LOESS sur le nuage de points. Ensuite, les points sont réajustés en soustrayant les valeurs ajustées $\hat{y}_i$ aux valeurs $M_{ij}$, centrant ainsi la dispersion sur zéro sur l'axe $M$. A titre d'exemple, la Figure 5.9 présente le MA-plot des données

de protéomique quantitatives pour deux conditions biologiques étudiées pour le projet NAFLD, en l'occurrence les stades 0 et 1 du score de Fibrose. Les données présentent une relation non-linéaire mise en évidence par la régression LOESS (courbe bleue), elles ne sont pas centrées sur zéro sur l'axe $M$. Les données sont ensuite normalisées grâce à la LOESS (la courbe verte correspond à la courbe bleue centrée sur zéro) et sont centrées sur zéro sur l'axe $M$.

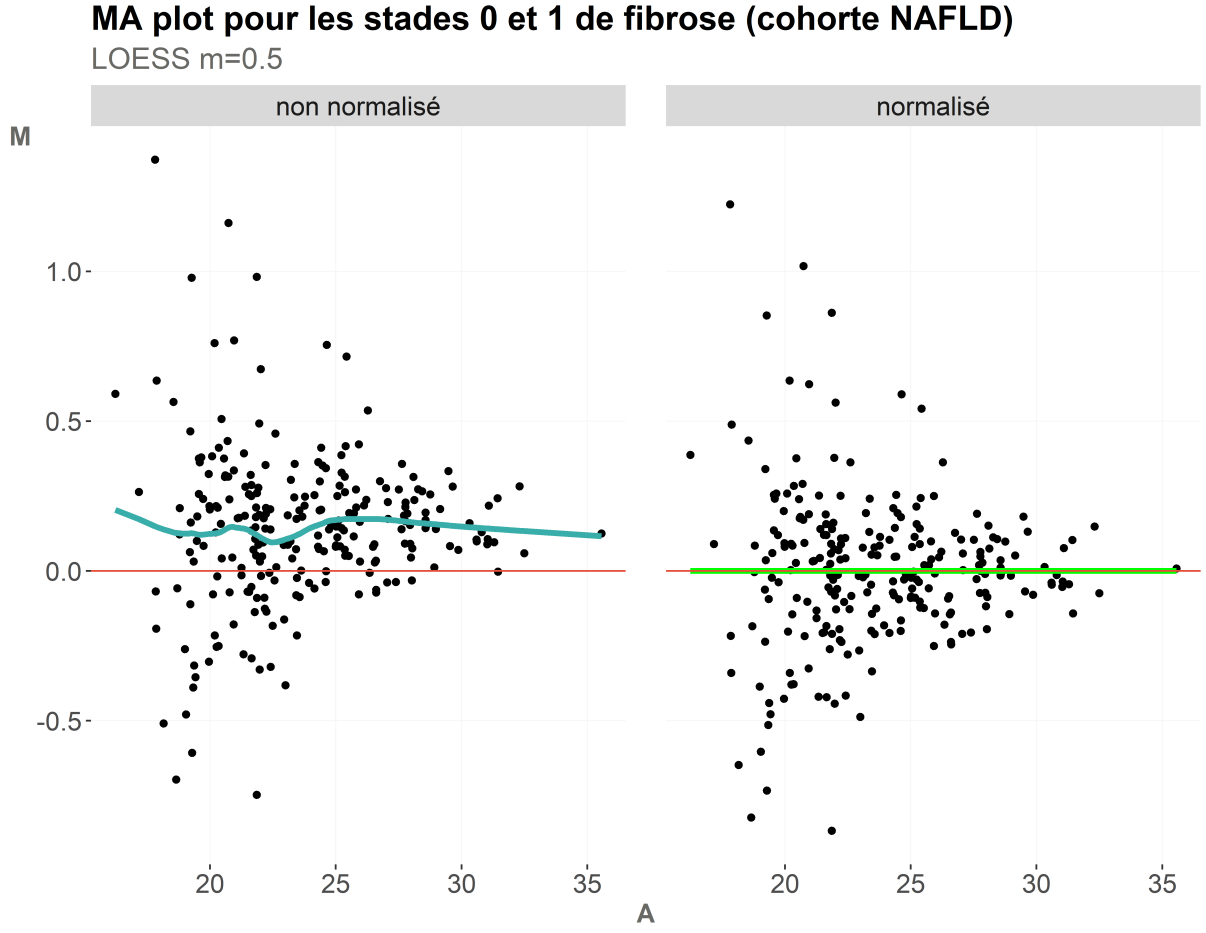



FIGURE 5.9 – MA-plot des deux conditions 0 et 1 du score de Fibrose pour la cohorte NAFLD. Un point représente une protéine, l'abondance de chaque protéine a été moyennée pour chaque condition étant donné que les effectifs sont différents (pour chaque protéine, nous avons 24 abondances pour la condition F0 et 30 abondances pour la condition F1). Sur le MA-plot « non normalisé » la courbe bleue correspond à la régression LOESS ($m = 0.5$ et $d$=1). Sur le MA-plot « normalisé », les données ont été recentrées sur 0 grâce à la régression LOESS. La courbe verte correspond à la courbe bleue qui a été centrée sur 0.

En protéomique, il est rare de ne considérer qu'une seule paire d'échantillons, et une méthode de normalisation adaptée doit pouvoir considérer un nombre quelconque (disons $k$) d'échantillons. Cela nécessite une généralisation de l'approche binaire du MA-plot. La

comparaison binaire pourrait être réalisée à partir d'un échantillon considéré comme la référence avec laquelle chacun des échantillons devrait être comparé et normalisé. Cependant, le choix de l'échantillon de référence n'est pas trivial, et en un choisir un au hasard risquerait de rendre la méthode non reproductible d'une étude à une autre. J'ai donc choisi d'utiliser l'approche de LOESS cyclique développée par Ballman *et al.* [173]. L'intérêt de cette méthode est la manière dont est choisi l'échantillon de référence. En effet, cet échantillon correspond à la moyenne de l'ensemble des échantillons à normaliser. Ainsi, chaque échantillon est normalisé par rapport à cet échantillon de référence. Cette méthode est relativement rapide car le temps de calcul est linéaire par rapport au nombre d'échantillons à normaliser. Cette linéarité est d'autant plus avantageuse pour des cohortes comportant de nombreux échantillons comme c'est le cas pour la cohorte NAFLD. Le second paramètre important à définir est la fenêtre de voisinage $m$. Il est intéressant de constater que la valeur par défaut de la fonction est $m = 0,7$, tandis que Cleveland recommande dans ses travaux [177] de débuter avec une valeur $m = 0,5$ lorsque l'utilisateur n'a aucune idée du voisinage nécessaire. Cette différence n'est pas clairement justifiée par Ballman *et al.*, mais cette valeur reste toutefois modifiable par l'utilisateur, si bien qu'il peut par la suite modifier $m$ s'il estime cela nécessaire à l'obtention d'une normalisation plus adaptée à ses données.

L'un des principaux avantages de la normalisation LOESS [173] est qu'il suffit de spécifier le paramètre $m$, lui conférant ainsi une certaine flexibilité pour estimer la régression sur un ensemble de points. Cependant la normalisation LOESS présente quelques limites. Notamment, elle hérite des propriétés de la LOESS qui nécessite des jeux de données relativement conséquents et denses pour pouvoir réaliser des modèles performants. LOESS est relativement intensive en calculs, d'autant plus lorsqu'il s'agit de modéliser des polynômes de second degré [176], mais l'implémentation optimisée de Ballman *et al.* pallie ce problème. Enfin, il s'agit d'une méthode relativement sensible aux valeurs extrêmes. Bien que les auteurs aient développé une pondération robuste face aux points extrêmes [172], il est toujours possible pour l'utilisateur de les supprimer avant de procéder à la normalisation, ou bien de leur attribuer un poids inférieur aux autres points ou nul (ce qui équivaut à leur suppression), s'il estime cela judicieux.

### 5.3.2 Le choix vsn

La normalisation par stabilisation de la variance (« *variance stabilization normalisation* » ou vsn) [79] consiste à appliquer une transformation sur les intensités mesurées, afin de

corriger la dépendance existante entre la variance et la moyenne des intensités [99]. Pour cela, les auteurs de cette méthode ont proposé une transformation non-linéaire paramétrique [79, 174] qui permet de conserver une variance constante sur l'ensemble des intensités mesurées, et donc d'obtenir une variance indépendante de la moyenne. Le modèle est une généralisation de celui de Chen *et al.* [180]. En effet, ces derniers avaient développé leur approche dans un contexte d'analyse transcriptomique avec uniquement deux échantillons (des puces à ADN marqué par deux types de fluorochromes). Le modèle proposé par Huber *et al.* généralise la normalisation à un nombre d'échantillons supérieur ou égal à 2, il peut donc être appliqué au cas de la protéomique. La vsn réalise une transformation, notée $h$, sur les données quantifiées $y$, telle que :

$$h(y) = \gamma \times arsinh(a + by) \tag{5.4}$$

Où $\gamma$, $a$ et $b$ sont des paramètres estimés par maximum de vraisemblance dont la démonstration détaillée est présentée dans [174]. Les auteurs ont montré que cette transformation $h$ permet d'obtenir des données normalisées, ayant une variance homogène indépendante de la moyenne des intensités. Le graphique en Figure 5.10 permet d'illustrer concrètement la transformation  et la transformation logarithmique par rapport à un jeu de données de puces à ADN. Ce graphique est directement issu de l'article de Huber *et al.* [79].

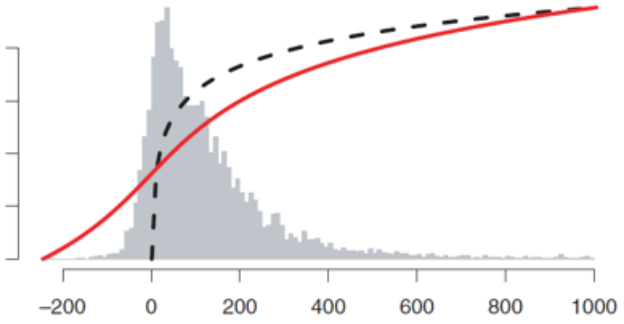



FIGURE 5.10 – Représentation de la transformation $h$ par vsn (ligne rouge), et la fonction logarithme (ligne en pointillés). L'histogramme correspond à la distribution des intensités pour un jeu de données d'une seule puce à ADN. Les paramètres de la transformation h ont été obtenus à partir de la comparaison avec les intensités obtenues sur une autre puce. Graphique issu de la publication de Huber *et al.* [79]

Il est important de noter que cette méthode a été développée dans le contexte de la transcriptomique, où une partie des intensités mesurées peuvent être négatives, comme cela est visible sur la Figure 5.10. La transformation *arsinh* permet donc d'obtenir des

valeurs réelles nulles ou négatives, contrairement à une simple transformation logarithmique. Cette propriété n'est pas nécessaire dans le cas de la protéomique, car il n'y a tout simplement pas de valeurs négatives.

Le principal avantage de vsn est qu'elle ne nécessite pas de paramétrisation de la part de l'utilisateur. Par conséquent son implémentation dans la librairie *DAPAR* utilise simplement la fonction fournie par la librairie *vsn*, en prenant soin de réaliser la normalisation sur les données non log2-transformées. Un autre avantage est qu'elle stabilise correctement la variance quelle que soit la taille des données [79]. De plus, il est possible de calculer les estimateurs du modèle soit sur un jeu de données de référence pour les réutiliser sur un autre jeu de données similaire, soit de les calculer directement à partir du jeu de données étudié. Une limitation de cette méthode est qu'elle suppose que l'ensemble des protéines analysées ont une abondance constante entre les différentes conditions. En ce qui concerne le plasma sanguin, son protéome est suffisamment stable pour que l'on puisse considérer l'hypothèse comme vérifiée. Par ailleurs, les auteurs supposent pour leur modèle que la relation des mesures entre les échantillons est estimée par une transformation linéaire [79]. Si cela n'est pas le cas dans les données étudiées, alors il est plutôt recommandé de se tourner vers des modèles ne se basant pas sur cette hypothèse, tel que LOESS, le centrage sur la médiane ou le centrage-réduction (cf. Section 3.1.4).

### 5.3.3 Discussion

Il est légitime de se demander si la normalisation des données est une étape indispensable pour obtenir de bonnes performances lors de l'analyse différentielle. Il n'y a pas de réponse définitive à cette question, dans le sens où elle se pose systématiquement (ou du moins devrait se poser) lors de chaque étude, car cela dépend des facteurs influençant l'étude et des données obtenues. Malheureusement, il peut s'avérer que les données présentent toutefois des biais de distribution, la simple log-transformation des intensités (cf. chap 3 section 3.1.2) n'ayant pas permis de résoudre le problème. Dans ce cas, il apparaît nécessaire d'effectuer une normalisation des données.

Si la normalisation est jugée pertinente pour parfaire l'analyse, alors le choix de la méthode à appliquer s'impose. De manière générale, il faut s'assurer que la méthode de normalisation choisie n'introduise pas de biais supplémentaire dans l'ensemble de données car cela affecterait les analyses statistiques en aval [181]. Cependant, ce choix est loin d'être trivial, et nécessite dans l'idéal d'évaluer les performances de différentes méthodes potentiellement intéressantes pour les données étudiées. Il est d'autant plus difficile de choisir

une méthode lorsque les performances de plusieurs d'entre elles s'avèrent similaires. En ce qui concerne cette thèse, notre choix s'est porté sur la visualisation de la calibration des $p_{valeurs}$ plutôt que sur des approches usuelles telles que *i)* la mesure de la réduction de la variance globale [82, 98] ; *ii)* la capacité à identifier des protéines comme différentiellement abondantes (DA) lorsqu'elles le sont en vérité ; ou *iii)* la comparaison de LogFC théoriques avec ceux obtenus à partir des données normalisées [92]. L'approche *i)* pose question : à trop réduire la variance des données, il est possible de créer artificiellement des faux positifs. L'approche *ii)* suppose que la méthode est testée sur des données simulées où la vérité est connue et maîtrisée, ou bien sur des données biologiques parfaitement connues, ce qui n'est pas le cas des données NAFLD. De même, l'approche *iii)* n'a d'utilité que si l'on souhaite réaliser des comparaisons binaires et qu'il y a une connaissance expérimentale des données étudiées, or nous souhaitons évaluer les performances pour la comparaison statistique globale que réalise l'ANOVA et nous travaillons sur une pathologie dont la biologie n'est pas encore entièrement comprise. Pour le cas particulier de LOESS, ses auteurs précisent [172] qu'une manière de s'assurer que la régression est adaptée aux données, est de comparer graphiquement les résidus $\epsilon_i = y_i - \hat{y}_i$ aux valeurs ajustées $\hat{y}_i$. Cependant cette comparaison serait difficile d'interprétation car cela consisterait à visualiser autant de graphiques qu'il y aurait de comparaison. Finalement, l'évaluation de la calibration des $p_{valeurs}$ constitue une approche originale vis-à-vis de la littérature en protéomique, et elle est plus adaptée à l'analyse de notre cohorte.

La calibration des $p_{valeurs}$ peut être visualisée sous la forme d'un calibration-plot. Ce dernier fut conceptualisé par Schweder *et al.* en 1982 [182] sous le nom de « *Pvalue-plot* ». Il s'agit d'une méthode graphique permettant d'évaluer les résultats d'un ensemble de tests statistiques en utilisant les $p_{valeurs}$ obtenues à cet ensemble de tests. Concrètement le graphique représente la distribution cumulée de $(1 - p_{valeur})$ en fonction de $(1 - p_{valeur})$. Cette méthode suppose qu'il existe une certaine proportion inconnue $\pi_0$ de protéines non différentiellement abondantes (DA) dont la distribution des $p_{valeurs}$ doit théoriquement être uniforme sur l'intervalle [0,1] sous $H_0$, si les données respectent les conditions d'application du test statistique choisi. À l'inverse, les $p_{valeurs}$ sous $H_1$, qui correspondent aux protéines DA, sont concentrées dans un intervalle de valeurs très proche de zéro. Par conséquent, le calibration-plot présente une droite croissante de pente $\pi_0$ correspondant aux $p_{valeurs}$ sous $H_0$, puis cette droite augmente fortement lorsque $(1 - p_{valeur})$ approche la valeur de 1 car elle correspond aux $p_{valeurs}$ sous $H_1$, qui sont très faibles.

Lorsque les conditions d'applications du test sont vérifiées, les $p_{valeurs}$ sont dites calibrées.

Or, pour que les conditions d'application d'un test statistique soient vérifiées, les données sont le plus souvent transformées, notamment par les étapes de pré-traitement décrites à la Section 3.1, et dont fait partie la normalisation. Par conséquent, si l'on observe que les $p_{valeurs}$ sont mal calibrées, alors cela laisse supposer que la transformation des données n'était pas adaptée (soit parce qu'insuffisante, soit parce qu'ayant introduit de nouveaux biais) pour qu'elles respectent suffisamment les conditions d'application du test statistique. Par extension, choisir la normalisation qui propose la meilleure calibration semble judicieux.

D'un point de vue pratique, j'ai utilisé la librairie *cp4p* [183] pour construire facilement les calibration-plot, puis vérifier les calibrations. Pour évaluer la pertinence des normalisations LOESS et vsn, les données de la cohorte NAFLD ont été filtrées et imputées de la même façon, et la même procédure statistique (ANOVA à 1 facteur) a été appliquée. La Figure 5.11 montre les calibration-plots des $p_{valeurs}$ de l'ANOVA obtenus lorsque les données n'ont pas été normalisées (Figure 5.11A), et lorsqu'elles ont été normalisées avec vsn (Figure 11B), LOESS (Figure 5.11C) et le centrage sur la médiane (voir section 3.1.4) qui est une méthode habituellement utilisée au laboratoire (Figure 5.11D). Il s'avère que les $p_{valeurs}$ obtenues à l'ANOVA réalisée sur des données normalisées par le centrage sur la médiane (Figure 5.11D) sont moins bien calibrées que celles obtenues après une normalisation LOESS (Figure 5.11C) et vsn (Figure 5.11B) par rapport à une absence de normalisation (Figure 5.11A). Par conséquent, non seulement le centrage sur la médiane ne semble pas une méthode pertinente pour notre cohorte NAFLD, mais LOESS semble plus adaptée que vsn.

Nous avons apporté quelques éléments de réponse pour LOESS et vsn en présentant les calibrations-plots comme critère d'évaluation et de comparaison. Par ailleurs, plusieurs études ont comparé des méthodes de normalisation en utilisant d'autres critères de performance [82, 92, 98], avec des conclusions finalement assez différentes. En effet, ils n'utilisent ni les mêmes jeux de données pour les tests, ni les mêmes critères d'évaluation. La seule conclusion commune que l'on peut identifier est qu'il est toujours préférable de tester plusieurs méthodes différentes et de comparer leurs résultats pour chaque jeu de données étudié, et cela en utilisant différents critères, notamment ceux brièvement abordés dans cette section.

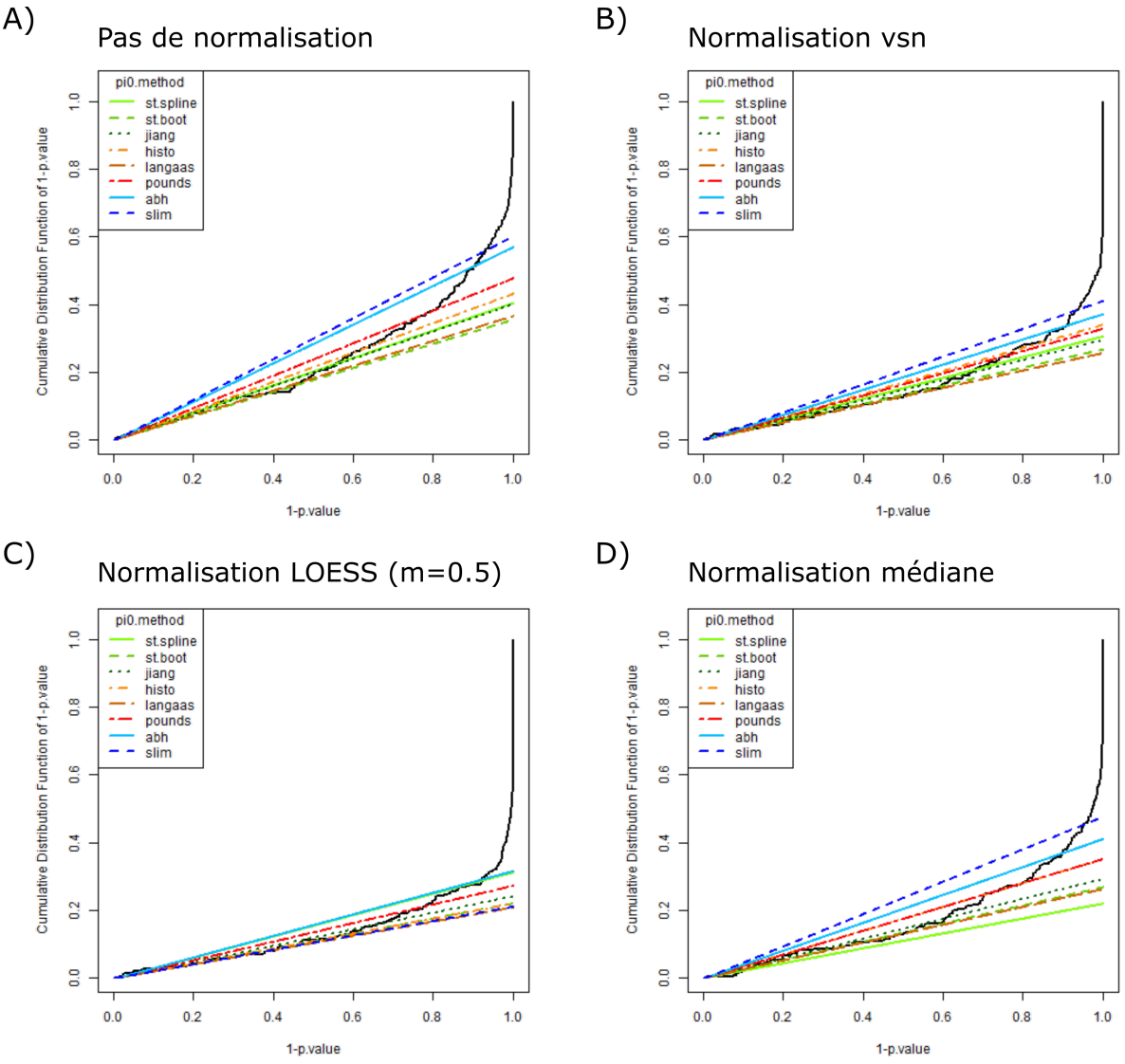



FIGURE 5.11 – Calibration plots obtenus pour les $p_{valeurs}$ issues de l'ANOVA à 1 facteur pour les données de la cohorte NAFLD. L'estimation du nombre $\pi_0$ de protéines non différentiellement exprimées a été réalisée avec l'ensemble des méthodes disponibles dans la librairie R *cp4p*. **A)** pas de normalisation réalisée, les pvaleurs sont mal calibrées. **B)** normalisation avec vsn, la calibration est clairement améliorée mais perfectible. **C)** normalisation avec LOESS (m=0,5), la calibration est meilleure que celle obtenue avec vsn. **D)** normalisation par centrage sur la médiane, la calibration est à peine améliorée par rapport à celle des données non normalisées, cette méthode ne semble donc pas pertinente pour notre jeu de données.

# Chapitre 6

# Deuxième contribution : Contrôle du FDR en cas de comparaisons multiples

## 6.1 Problématique

Comme cela a été expliqué dans le chapitre 3, il est nécessaire de réaliser autant de tests statistiques qu'il y a de protéines identifiées, de sorte que la correction de la multiplicité des tests (MTC) entre protéines devient indispensable. Parmi toutes les méthodes permettant cela, la communauté protéomique a principalement recourt au contrôle du FDR. Par ailleurs, dans le cas d'un projet nécessitant de comparer de nombreuses conditions biologiques entre elles, le risque d'inflation de l'erreur de type I doit aussi légitimement appeler à une seconde correction. Dans la pratique, la communauté de protéomique s'appuie principalement sur le cadre méthodologique de l'ANOVA pour réaliser de telles comparaisons multiples (pour les raisons évoquées précédemment, cf. Section 3.2), ainsi que sur les tests *post hoc*, permettant, entre autres, cette seconde correction de la multiplicité des tests (cf. Section 3.4.2).

Finalement, si l'on considère les données quantitatives sous la forme d'un tableau, avec en ligne les protéines identifiées, et en colonnes les conditions à comparer, alors on constate que la MTC des lignes (le contrôle du FDR) et celle des colonnes (celle intégrée par les tests *post hoc*) doivent cohabiter. Or, la question de leurs interactions n'a pas été particulièrement discutée ou théorisée dans la littérature des biostatistiques appliquées à la

protéomique, même si quelques auteurs la pointent du doigt comme possiblement problématique[184]. Théoriser et proposer une nouvelle méthodologie pour résoudre cette difficulté sort clairement du cadre de mon sujet de thèse, comme de celui de mes compétences théoriques. Dès lors, l'objectif de ce chapitre est plus modeste : il s'agit de s'interroger sur l'importance relative des deux types de MTC en fonction de la question biologique, afin de les prioriser.

## 6.2 Analyse des différents rôles des MTC

### 6.2.1 Les rôles du contrôle du FDR

Le contrôle du FDR a pour rôle premier évident celui de la rigueur statistique. Cependant, dans la pratique, on s'aperçoit qu'il a tendance à jouer deux autres rôles. Le premier est un rôle normatif. En effet, il s'agit d'une procédure bien décrite [131, 185, 186, 187], et dont l'application à des données de protéomique suit un protocole très standardisé. Son utilisation permet de réaliser une comparaison facile et interprétable entre différentes études, car la définition d'une protéine significativement différentiellement abondante sera la même pour les différentes publications. Le second rôle (que l'on appellera par la suite stringence expérimentale) concerne la pertinence des résultats d'un point de vue biologique. En effet, l'ajustement du seuil de risque d'erreur de type I permet, en étant plus ou moins stringent, d'augmenter ou de réduire le nombre de protéines statistiquement différentiellement abondantes. Il est bien connu qu'en pratique, leur nombre n'est pas régi que par des considérations statistiques, et que l'expérimentateur peut avoir une connaissance *a priori* dont il souhaite tenir compte. Il peut également y avoir des contraintes expérimentales, par exemple, sur la possibilité de ne valider qu'un nombre limité de candidats dans des expériences post-protéomiques. Dès lors, l'ajustement du seuil de contrôle n'a pas qu'une vocation statistique, mais aussi, implicitement, des vocations respectivement normative et pratique.

### 6.2.2 Le rôle des tests *post hoc*

Les tests *post hoc* ont deux rôles principaux. Le premier est de passer d'un test d'hypothèse global à ceux correspondant à un ensemble de comparaisons binaires. Le second est de contrôler l'inflation de l'erreur de type I due aux multiples tests réalisés. En ce qui concerne le premier rôle, il est possible de réaliser une ANOVA avant les tests *post hoc*, ou bien de répondre directement à cette question en s'appuyant sur la notion de « contraste ».

Un contraste $c$ permet de formaliser la comparaison spécifique d'un ensemble de $k$ moyennes parmi les différents groupes étudiées. C'est un ensemble de coefficients qui définissent la comparaison spécifique de l'ensemble des $k$ moyennes à comparer, et dont la somme est nulle $\sum_{i=1}^{k} = 0$. Ainsi, un contraste donné teste un ensemble spécifique d'hypothèses. Soit un contraste qui teste $k$ moyennes, notées $\mu_i, i \in [1, k]$, alors l'hypothèse nulle testée peut être formulée telle que :

$$H_0 : \sum_{i=1}^{k} c_i \mu_i = 0 \tag{6.1}$$

Et l'hypothèse alternative telle que :

$$H_1 : \sum_{i=1}^{k} c_i \mu_i \neq 0 \tag{6.2}$$

Des contrastes sont orthogonaux lorsque la somme des produits de leurs coefficients est nulle, autrement dit, pour deux contrastes notés $c^1$ et $c^2$, nous avons :

$$\sum_{i=1}^{k} c_i^1 c_i^2 = 0 \tag{6.3}$$

A titre d'exemple, deux contrastes spécifiés pour trois classes, notés $c^1(1, 0, -1)$ et $c^2(\frac{1}{2}, -1, \frac{1}{2})$ sont orthogonaux :

$$\sum_{i=1}^{k} c_i^1 c_i^2 = \left(1 \times \frac{1}{2}\right) + (0 \times -1) + \left(-1 \times \frac{1}{2}\right) = 0 \tag{6.4}$$

L'ANOVA à un facteur à $k$ groupes, également appelée test omnibus, correspond à un test avec $k-1$ contrastes orthogonaux non-redondants. Chaque contraste possède un degré de liberté et est indépendant des autres, par conséquent l'ensemble des $k-1$ contrastes orthogonaux permet de partitionner la somme des carrés expliqués ($SCE$) (cf. Section 3.3.2). En reprenant notre exemple précédent de deux contrastes orthogonaux ($a$ et $b$), nous avons :

$$SCE = SC_a + SC_b \tag{6.5}$$

La statistique $F$ est obtenue grâce aux $k-1$ contrastes orthogonaux, telle que :

$$\begin{aligned} F &= \frac{\frac{SCE}{CMR}}{k-1} \\ &\Leftrightarrow F = \frac{CME}{CMR} \\ &\Leftrightarrow F = F_{omnibus} \end{aligned} \tag{6.6}$$

Ainsi, réaliser des comparaisons binaires par contrastes orthogonaux est équivalent à réaliser un test omnibus, autrement dit une ANOVA à un facteur. C'est notamment pour cela que certaines librairies (dont notamment *limma* [78]) implémentent l'ANOVA (ainsi que plusieurs autres tests) en passant par ce formalisme générique des contrastes.

Néanmoins, réaliser un ensemble de contrastes pour un grand nombre de protéines peut s'avérer couteux d'un point de vue calculatoire. C'est notamment ce qui explique que la manière traditionnelle de réaliser une ANOVA ne suive pas cette logique algorithmique. En effet, comme cela a été souligné dans la Section 3.4.2, jusqu'à récemment dans l'histoire des statistiques, limiter les calculs inutiles et optimiser les calculs nécessaires était un impératif. Nous avions évoqué à titre d'exemple le fait que l'ANOVA permettait alors de limiter le nombre d'opérations calculées en aval (en ne réalisant les tests *post hoc* qu'après le rejet de l'hypothèse nulle omnibus). Or, cette optimisation n'a plus vraiment d'importance à l'heure actuelle car malgré l'augmentation du débit analytique (et donc de protéines à tester) les puissances de calculs ont progressé encore plus rapidement, de sorte que le calcul de plusieurs milliers d'ANOVA simultanément n'est en aucun cas problématique (même si elles sont implémentées *via* des contrastes). Par ailleurs, la démarche historique ne doit pas être forcément prise comme référence : Dans la mesure où les conclusions de l'ANOVA et du test *post hoc* peuvent même s'avérer conflictuelles [121], plusieurs auteurs suggèrent aux expérimentateurs de réaliser l'ensemble des tests *post hoc* [121, 135] quelle que soit l'issue de l'ANOVA (cf. Section 3.4.2). Dès lors, nous pouvons nous demander quel est l'intérêt à l'heure actuelle d'appliquer des ANOVA systématiquement suivie d'autant de tests *post hoc*, plutôt que d'utiliser directement les contrastes permettant les comparaisons que l'expérimentateur trouve intéressantes.

Le second rôle des tests *post hoc* peut avoir une importance différente selon l'objectif du projet. Prenons deux exemples : dans le cas d'une étude de comparaison multiple entre conditions indépendantes (telle que la comparaison entre un placebo et différents traitements distincts), l'ensemble des comparaisons conduit nécessairement à l'inflation de

l'erreur de type I et le second rôle des tests *post hoc* (pour rappel, la MTC) est primordial. A l'inverse, dans le cas d'une étude de type chronologique (« *time-course* » en anglais), où les différentes conditions correspondent à différents temps, mais ne correspondent pas à des mesures répétées sur les mêmes individus, sa nécessité peut être questionnée en fonction de ce que l'expérience cherche à mettre en évidence. Notamment, si l'on s'intéresse au comportement global d'une protéine au cours du temps, les fluctuations stochastiques locales (ainsi que les possibles erreurs de type I associées) n'auront que peu d'importance.

Finalement, aussi bien pour les tests *post hoc* comme pour le contrôle du FDR, nous observerons que leur rôle va au-delà de la MTC, mais que cette MTC peut avoir une importance relative en fonction du projet. C'est pourquoi, dans la suite de ce chapitre, nous présentons différents scénarios, et discutons de la manière de procéder respectivement aux deux types de MTC, de l'utilité de l'ANOVA et des rôles joués par les tests *post hoc*. Les trois premiers reposent sur l'ANOVA à proprement parler, alors que les trois autres sont construits à partir de contrastes.

## 6.3 Les différents scénarios possibles

Pour la suite de cette section nous considérons que les conditions d'application de l'ANOVA sont respectées (cf. Section 3.3.2). En effet, bien que l'ANOVA soit un test facile à mettre en place et à interpréter, il est essentiel de s'assurer en amont que les conditions d'application sont appliquées dans tous les cas, afin d'éviter des résultats biaisés.

### 6.3.1 Scénarios impliquant des ANOVA

#### 6.3.1.1 ANOVA, contrôle du FDR et tests *post hoc*

La méthodologie apparemment la plus répandue est celle qui consiste à réaliser une ANOVA à un facteur, puis un contrôle du FDR et enfin un test *post hoc* uniquement sur les protéines qui ont obtenu une $p_{valeur}$ passant le seuil induit par le contrôle du FDR. Dans ce scénario, le test *post hoc* permet surtout de réaliser les comparaisons binaires qui ne sont pas faites avec l'ANOVA, tout en introduisant une correction de tests multiples entre celles-ci, comme cela est décrit dans la Figure 6.1. Par exemple, le test *post hoc* de Tukey (cf. Section 3.4.2) réalise une correction inspirée de celle de Bonferroni. L'ensemble des étapes permet donc de déterminer quelles sont les protéines statistiquement différentiellement abondantes ; et plus précisément entre quelles conditions elles sont diffé-

rentielles. Nous pouvons voir sur la Figure 6.1 qu'il résulte de ce scénario autant de listes de protéines différentes qu'il y a de comparaisons faites grâce aux tests *post hoc*.

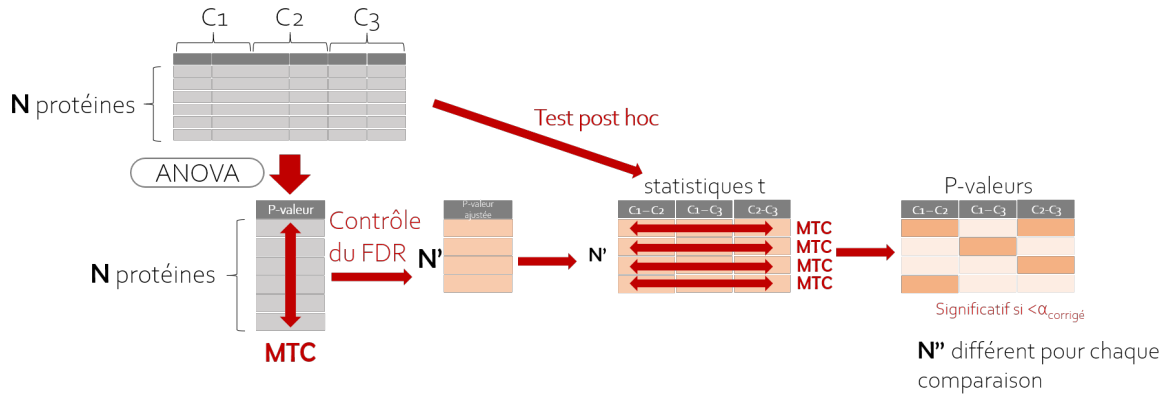



FIGURE 6.1 – Schéma récapitulatif de la procédure classiquement utilisée en protéomique lorsque l'on souhaite comparer plus de deux conditions expérimentales. Une ANOVA à un facteur est d'abord réalisée sur le tableau des données quantitatives. Le contrôle du FDR est réalisé sur les résultats de l'ANOVA, ce qui constitue une première correction de tests multiples (MTC) entre les lignes. Les tests *post hoc* ne sont réalisés que pour les protéines ayant obtenu un résultat significatif à l'issue du contrôle du FDR et réalisent une MTC entre les colonnes. Finalement ce sont autant de listes de protéines significatives que de comparaisons qui sont obtenues.

Cependant, avec ce scénario il est difficile de clairement établir le rôle joué par le contrôle du FDR, et ce pour plusieurs raisons. Tout d'abord, les multiples listes obtenues grâce au test *post hoc* n'ont pas de contrôle de FDR associé, alors que plusieurs tests ont été réalisés. Par conséquent le rôle de contrôle statistique (qui reste le rôle principal) que pourrait fournir le contrôle du FDR est partiellement perdu (bien que contrôlé après l'ANOVA, le FDR n'est pas contrôlé sur chacune des listes). Concernant l'aspect normatif du FDR, celui-ci est préservé dans le sens où la procédure est systématique (et donc facilement transposable) : les tests *post hoc* sont réalisés uniquement pour les protéines qui ont obtenu une $p_{valeur}$ inférieure au seuil fixé avec le contrôle du FDR. Ensuite, d'un point de vue biologique, le contrôle du FDR est principalement utilisé comme un filtre aux tests *post hoc*. Autrement dit, on pourrait penser qu'il permet dans une certaine mesure de prendre en compte les attentes de l'expérimentateurs, puisque le seuil de contrôle du FDR succédant à l'ANOVA peut être modifié en fonction. Cependant, une fois les tests *post hoc* réalisés, nous nous retrouvons avec autant de listes de protéines différentiellement abondantes qu'il y a de comparaisons binaires, et ces listes peuvent être de tailles très différentes. Ces résultats ne correspondent plus à la liste de protéines obtenue avec le contrôle du FDR de l'ANOVA, si bien que le rôle de stringence expérimentale de ce dernier

n'est que partiel. Néanmoins, ce scénario peut avoir un intérêt si le nombre de conditions à comparer est important, et donc la correction des tests multiples est absolument nécessaire. Il peut aussi présenter un intérêt lorsque le nombre de protéines obtenant une $p_{valeur}$ ajustée (la $q_{valeur}$) significative au seuil fixé est très faible, et que les processus biologiques associés sont indépendants les uns des autres. La correction des tests multiples entre les quelques protéines DA à la suite des tests *post hoc* n'est alors plus aussi importante ; comme cela peut être le cas lorsque l'on compare l'efficacité entre différents traitements en ne testant qu'un nombre très restreint de variables observées et indépendantes. Du moins, dans de tels cas, les MTC à invoquer peuvent sortir du cadre méthodologique du contrôle du FDR. En revanche, il s'agit de cas spécifiques, dont le contexte ne correspond pas à la plupart des expérimentations de protéomique. Finalement, il apparait que ce scénario pourtant très répandu, n'a que peu d'intérêt en dehors de sa fonction normative évidente.

#### 6.3.1.2 ANOVA, tests *post hoc* et multiples contrôles du FDR

Au regard des limitations évidentes du premier scénario, nous pouvons penser qu'il suffit d'intervertir l'étape de contrôle du FDR et celle des tests *post hoc* (cf. Figure 6.2), pour récupérer le rôle premier et principale du FDR, à savoir maintenir la rigueur statistique. Cela signifie que les tests *post hoc* doivent être réalisés pour l'ensemble des protéines quelle que soit l'issue de l'ANOVA, puis qu'un contrôle du FDR soit réalisé pour chaque comparaison binaire.

Ce second scénario présente plusieurs difficultés concernant les différents rôles du FDR. En effet, on génère autant de listes de protéines qu'il y a de comparaisons, avec pour chacune une MTC spécifique. Par conséquent, même si le rôle de contrôle statistique est présent, les deux autres rôles sont difficiles à appréhender du fait des multiples listes de protéines obtenues. Si le contrôle du FDR est réalisé après le test *post hoc*, alors c'est que l'on s'intéresse vraisemblablement plus aux résultats des comparaisons binaires, qu'à ceux de l'ANOVA. Dès lors, il serait tout aussi pertinent de réaliser une comparaison des contrastes d'intérêts, et éventuellement d'appliquer une MTC équivalente à celle incluse dans le test *post hoc*, sans passer par l'ANOVA. En fait, il serait même plus pertinent de procéder ainsi, dans la mesure où la MTC apparaîtrait explicitement, de sorte qu'il deviendrait possible d'en discuter l'intérêt (comme nous l'avons fait précédemment avec les deux exemples de la comparaison de traitements et de l'analyse d'un time-course en fin de Section 6.2.2).

Finalement, ce scénario permet une plus grande rigueur statistique que le précédent, mais son cheminement est dépourvu de logique (puisqu'il ne devrait pas impliquer une ANOVA, mais des contrastes) et l'interprétabilité des résultats qu'il fournit est tout aussi difficile.

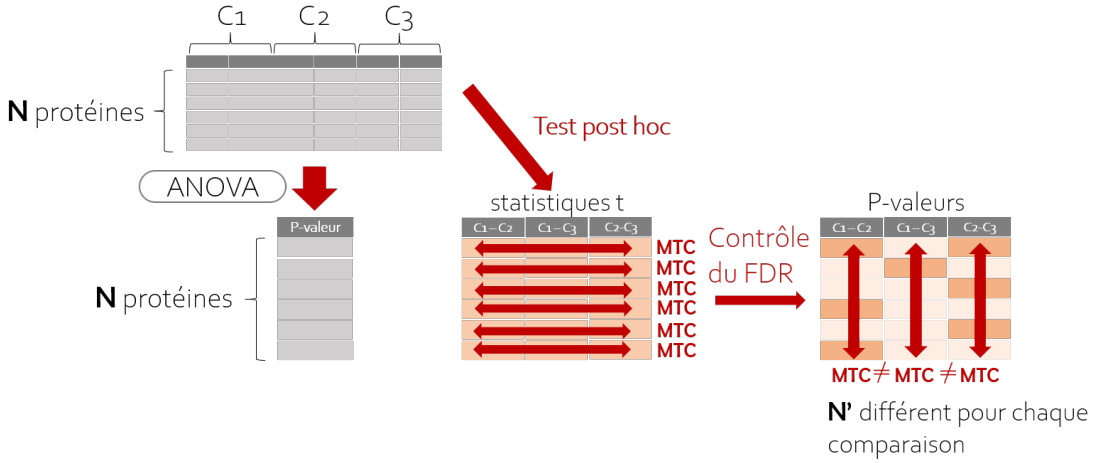



FIGURE 6.2 – Schéma du scénario où le contrôle du FDR est réalisé en aval des tests *post hoc*. MTC : correction de tests multiples. Avec ce scénario, on obtient autant de contrôles du FDR qu'il y a de comparaisons binaires.

#### 6.3.1.3 ANOVA, contrôle du FDR et clustering de profils d'abondance

Nous avons vu que l'implication des tests *post hoc* impacte les rôles du contrôle du FDR et que leur utilité est finalement questionnable lorsque l'on s'intéresse à l'ensemble des comparaisons pour décrire un unique phénomène et avoir ainsi une vision globale. C'est pourquoi nous proposons dans ce scénario-ci de nous passer des tests *post hoc* ; et de réaliser une ANOVA puis un contrôle du FDR, puis de réaliser un clustering de profils d'abondance pour contextualiser les résultats statistiques à l'aide des données quantitatives. Autrement dit, cela consiste à combiner visuellement les $p_{valeurs}$ inférieures au seuil obtenu avec le contrôle du FDR lors de l'ANOVA avec les résultats du clustering, tout en les associant aux profils de chaque protéine, comme cela est résumé dans la Figure 6.3. Evidemment, ce scénario ne s'applique pas à toutes les situations, car il n'a de sens que pour les études où ce qui prévaut est l'interprétation globale des conditions comparées. ; autrement dit, dans le cas particulier où les différentes conditions étudiées peuvent être naturellement ordonnées et où les profils d'abondance ont un sens biologique fort, telles que des données stratifiées, ou les études time-course. Nous définissons un profil d'abondance comme la représentation d'une protéine par ses abondances au cours des différentes conditions étudiées qui correspondent à un ordre biologique précis. Par exemple, lorsque

chaque condition correspond à une étape de progression, l'évolution de l'abondance d'une protéine peut être visualisée en fonction de ces étapes ordonnées, donnant alors un profil d'abondance pour cette protéine. Le partitionnement (ou « *clustering* » en anglais) de profils d'abondance est une méthode d'apprentissage non supervisé qui est déjà très utilisée de manière complémentaire à l'analyse statistique afin de visualiser les données dans la phase exploratoire [188] sans qu'il s'agisse d'une intégration explicite des résultats statistiques au sein de ceux du clustering.

Néanmoins, certaines études utilisent le clustering pour l'analyse des profils d'expression mais uniquement en complément d'une analyse statistique réalisant une ANOVA puis un contrôle du FDR suivi de tests *post hoc* [189, 190, 191], et cela sans tenir compte des problèmes que nous avons soulevés pour l'interaction entre les deux types de MTC.

Les deux premiers rôles du contrôle du FDR sont préservés : l'absence d'interaction avec une autre correction de tests multiples garantit la rigueur statistique du résultat ; quant à la dimension normative, elle résulte de la simplicité de la procédure, même si elle doit être tempérée par son manque de généricité. En effet, elle ne permet pas la comparaison binaire de plusieurs conditions indépendantes (contrairement aux deux premiers scénarios qui permettent d'y répondre grâce aux tests *post hoc*). Quant au troisième rôle du FDR (un filtrage plus ou moins stringent en fonction des attentes de l'expérimentateur), il reste possible dans une certaine mesure, et peut être compléter en jouant sur le nombre et la taille des clusters.

Finalement, cette procédure ne permet pas de répondre à toutes les situations, mais semble particulièrement intéressante lorsque l'on souhaite préserver une vision globale des résultats, tout en conciliant les contraintes statistiques, expérimentales et d'interprétation, comme c'est le cas sur le projet clinique qui nous intéresse.

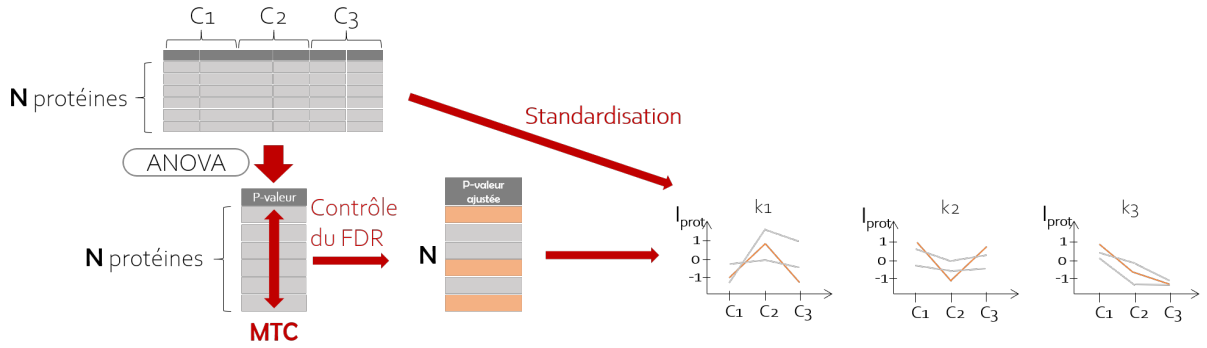



FIGURE 6.3 – Schéma du scénario combinant ANOVA, contrôle du FDR et clustering de profils d'abondance des protéines. Une standardisation des abondances est réalisée afin de rendre comparables les profils d'abondances des protéines et ainsi réaliser le clustering. Chaque profil de chaque protéine est coloré en fonction de sa $q_{valeur}$.

### 6.3.2 Scénarios reposant sur des contrastes

Lorsque plusieurs contrastes sont testés simultanément pour un très grand nombre de protéines, la problématique de l'interaction entre les différentes MTC se pose de la même manière qu'avec l'usage de tests *post hoc* suivant une ANOVA. Smyth *et al.* [184] ont signalé ce problème et ont proposé trois scénarios dans la librairie *limma* afin de le contourner. Ce sont ces scénarios que nous proposons de discuter dans cette section.

#### 6.3.2.1 Méthode Globale

La première méthode proposée dans [184], dite globale, consiste à réaliser un ensemble de contrastes qui fournissent pour chaque protéine une $p_{valeur}$ associée à la statistique $F$. Puis l'ensemble des $p_{valeurs}$ est ajusté sans tenir compte ni de la protéine associée, ni du contraste auquel elles correspondent, comme cela est montré dans la Figure 6.4. Par conséquent cela suppose que les contrastes étudiés sont extrêmement liés. L'avantage ici est que le seuil de $p_{valeur}$ est cohérent entre tous les contrastes. Malgré un rôle normatif évident, le principal inconvénient, signalé par l'auteur lui-même, est qu'aucun théorème ne prouve que le FDR sera effectivement contrôlé entre les contrastes. Par conséquent le rôle de contrôle statistique n'est ici pas assuré.

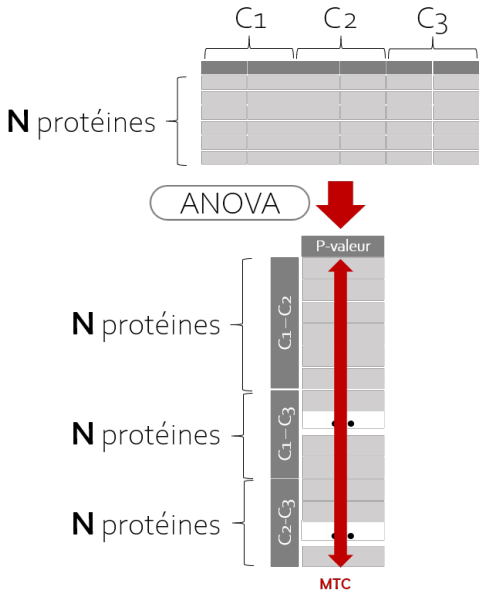



FIGURE 6.4 – Schéma résumant la méthode globale de *limma* pour réaliser l'analyse différentielle des protéines en fonction des contrastes spécifiés.

#### 6.3.2.2 Méthode hiérarchique

Conceptuellement, cette méthode est équivalente au tout premier scénario (cf. Section 6.3.1.1) mais reformalisé en utilisant des contrastes, puisque ceux-ci sont à la base de l'implémentation de la librairie *limma*. En effet, comme nous pouvons le voir dans la Figure 6.5, les $p_{valeurs}$ sont obtenues à la suite du test des contrastes, puis un contrôle du FDR est appliqué ; enfin des tests statistiques sont réalisées pour chaque contraste uniquement pour les protéines ayant obtenues une $q_{valeur}$ significative. Une correction de tests multiples est réalisée lors de ces tests comme cela est fait avec les tests *post hoc*. Cette procédure et celle de la Section 6.3.1.1 étant très similaires, nous avons naturellement ici les mêmes limitations concernant les rôles joués par le contrôle du FDR. Notamment, l'auteur confirme nos craintes, puisqu'il indique que le FDR est potentiellement mal maîtrisé [184].

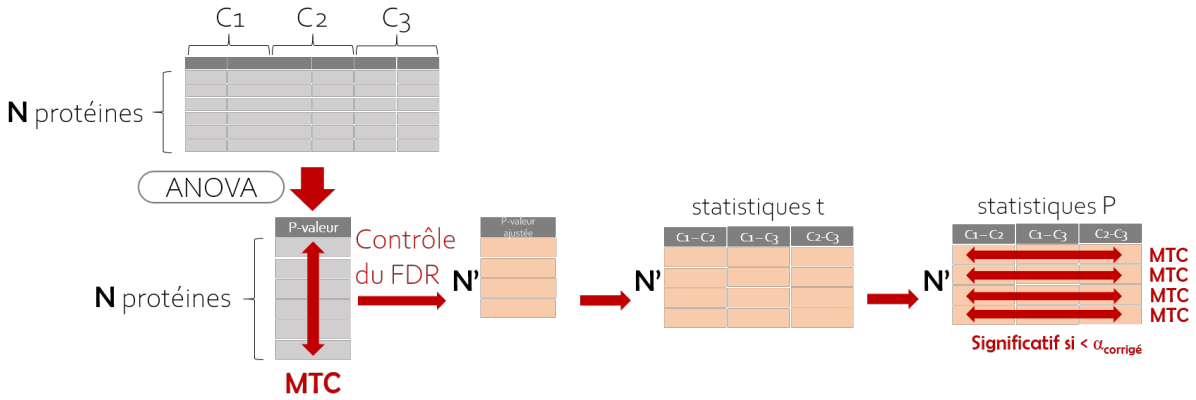



FIGURE 6.5 – Schéma résumant la méthode hiérarchique de *limma* pour réaliser l'analyse différentielle des protéines en fonction des contrastes spécifiés.

#### 6.3.2.3 Méthode imbriquée

Cette troisième méthode diffère des deux autres, notamment par la question à laquelle elle répond. En effet, le principe est de réaliser le test statistique omnibus, puis de contrôler le FDR, puis de réaliser des t-tests pour chaque contraste spécifié (voir Figure 6.6). Enfin, ne sont considérées comme différentiellement abondantes, que les protéines obtenant une $q_{valeur}$ significative au seuil FDR fixé et qu'elles obtiennent au moins deux $p_{valeurs}$ significatives pour les t-tests. Par conséquent elle répond à la question : « *quelles sont les protéines différentiellement abondantes dans plus de deux contrastes et desquels s'agit-il ?* ». L'objectif ici est donc de donner plus de poids aux protéines qui sont différentiellement abondantes pour plus de deux contrastes. Avec cette méthode, les rôles du contrôle du FDR sont encore une fois remis en question, car l'auteur affirme qu'il faut la considérer comme expérimentale et que le FDR n'est potentiellement pas correctement contrôlé. Par ailleurs, à la fin de la procédure le seuil fixé lors du contrôle du FDR pour les $p_{valeurs}$ n'est plus cohérent entre les contrastes. De plus, bien que l'on apporte une réponse à une question biologique intéressante, on ne répond plus à la question la plus classiquement posée lorsqu'on utilise des contrastes ou un test omnibus.

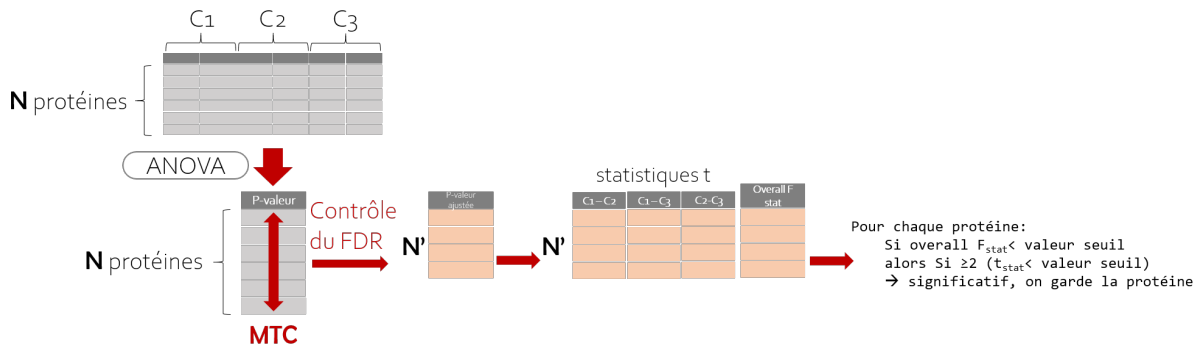



FIGURE 6.6 – Schéma résumant la méthode imbriquée de *limma* pour réaliser l'analyse différentielle des protéines en fonction des contrastes spécifiés.

Finalement, parmi les trois solutions proposées par Smyth *et al.* [184], aucune ne garantit le contrôle du FDR. Une telle absence de garantie pour la librairie probablement utilisée en biostatistique montre bien la difficulté sous-jacente à la problématique discutée dans ce chapitre. A défaut de proposer une nouvelle méthode, notre présentation et notre discussion de ces différents scénarios permettent à l'expérimentateur de mieux cerner les différentes possibilités qui s'offrent à lui, ainsi que dans quelle mesure elles sont impactées par l'interaction des différentes MTC.

## 6.4 Implémentation

Les différents scénarios décrits plus haut permettent chacun de répondre à des cas d'application différents. J'ai donc implémenté de manière aussi unifiée que possible dans DAPAR l'ensemble des briques qui permettent de les construire, afin qu'ils puissent être proposés ultérieurement aux utilisateurs de Prostar, soit à des fin d'usage pratique, soit à des fins de comparaison, dans une logique similaire à celle suivi dans ce chapitre. La Figure 6.7 résume ces différentes briques implémentées, à l'exception de la brique concernant le volcano plot et de celle du contrôle du FDR qui étaient déjà disponibles dans DAPAR.

Ainsi, à titre d'exemple, le scénario de la Section 6.3.1.2 (ANOVA à un facteur, puis tests *post hoc* suivi du contrôle du FDR) peut être construit en utilisant les briques représentées sur la gauche de la Figure 6.7, ainsi que l'outil de contrôle du FDR déjà présent dans DAPAR. Par ailleurs, je donne la possibilité de choisir entre différents tests *post hoc*, en fonction de la présence d'une condition de référence (utilisation du test de Dunnett, cf. Section 3.4.2) ou non (utilisation du test de Tukey, cf. Section 3.4.2). La visualisation

ultérieure de ces résultats, puis le contrôle du FDR se font par l'intermédiaire d'un volcano plot pour chaque comparaison binaire.

D'autre part, j'ai implémenté la procédure de la Section 6.3.1.3 (partie droite de la Figure 6.7), en laissant la possibilité de réaliser l'ANOVA soit avec une régression linéaire classique, soit avec la méthode de *limma*, puis de la compléter grâce au clustering. L'ensemble de ces implémentations ne sont que brièvement évoquées car elles ne présentent pas d'enjeu particulier. Néanmoins, l'étape de clustering et de visualisation des profils est plus détaillée dans le paragraphe suivant, car sa mise en place a suscité plusieurs interrogations et autant de choix de réalisation.

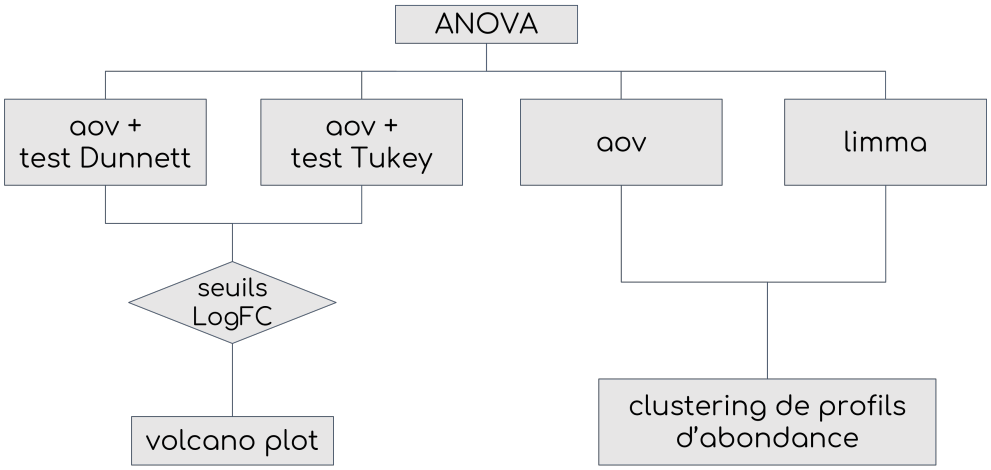



FIGURE 6.7 – Description des procédures statistiques possibles à partir d'une ANOVA à un facteur. Aov : fonction R pour réaliser l'ANOVA à un facteur. Limma : ANOVA à un facteur sous la forme de contrastes. Les différentes briques (en gris) ont été implémentées à l'exception de celle du volcano plot qui est un outil déjà disponible dans DAPAR.

Réaliser le clustering des profils d'abondance ainsi que leur visualisation impliquait de procéder à deux étapes de pré-traitement. La première étape consiste à moyenner les abondances des échantillons pour chaque condition en une unique valeur d'abondance (premier encadré sur la Figure 6.8). Ainsi, pour un exemple où l'on compare cinq conditions, le profil d'une protéine sera représenté par seulement cinq points reliés entre eux pour former une courbe, quel que soit le nombre de réplicats par condition, comme cela est visible sur la Figure 6.8. Ensuite, la seconde étape consiste à standardiser les abondances moyennes, c'est-à-dire les centrer et les réduire afin de les ramener sur la même distribution de moyenne 0 et d'écart-type égal à 1. Ce prétraitement permet ainsi de rendre comparables les protéines qui initialement ne le sont pas, comme cela a déjà été souligné dans la Section 3.4.1. Par ailleurs, cela réduit la difficulté de la tâche de clustering.

Pour intégrer les résultats statistiques obtenus à l'ANOVA au clustering de profils d'expression, les profils d'abondance moyenne des protéines sont représentés avec les autres profils de leur clusters d'appartenance, tout en étant colorés selon la $q_{valeur}$ obtenue lors de l'ANOVA après contrôle du FDR : plus la $q_{valeur}$ associée à une protéine est faible (donc statistiquement significative), plus le profil d'abondance est mis en valeur sur le graphique représentant les différents clusters (voir Figure 6.8). Nous avons choisi de visualiser les profils sous la forme d'une courbe d'abondance entre les différentes conditions, comme cela est présenté dans l'étude de Chen *et al.* [192], plutôt que sous la forme d'une « *heatmap* ». En effet, bien que cette dernière soit l'outil de visualisation le plus souvent associé au clustering, elle n'est pas nécessairement très intuitive, ni pour appréhender le profil d'expression d'une protéine dans son entièreté, ni pour saisir le profil global caractérisant chaque cluster. De plus, l'augmentation (ou diminution) d'abondance entre deux classes étant représentée par la couleur sur une heatmap, il semblait difficile d'associer graphiquement de manière visible l'information fournie par la $q_{valeur}$.

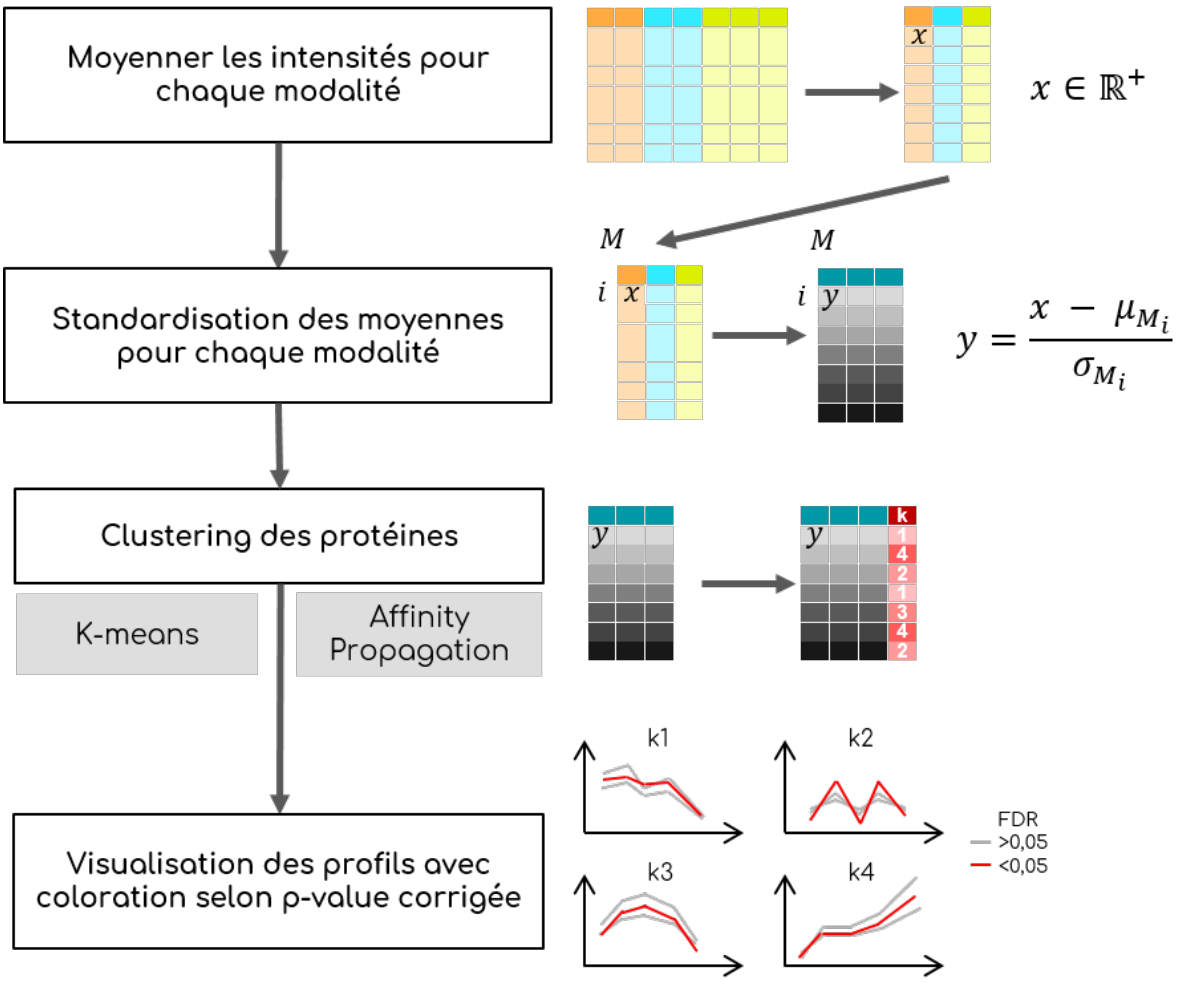



FIGURE 6.8 – Les différentes étapes de préparation des données quantitatives afin de réaliser le clustering et la visualisation des profils d'abondance. $M$ est la matrice des abondances. $\mu_{M_i}$ correspond à la moyenne des abondances pour une protéine $i$ donnée et $\sigma_{M_i}$ est l'écart-type.

La méthode de clustering la plus largement répandue, dans les publications de protéomique du moins, est la classification hiérarchique (« *hierarchical clustering* » en anglais ou HC). J'ai cependant choisi d'implémenter deux autres méthodes que sont le partitionnement des k-moyennes (ou « *k-means* » en anglais) [193], et la propagation d'affinité [194] (ou « *affinity propagation* » en anglais, ou AP). D'une part car les *k-means* sont très populaires dans la communauté de l'apprentissage automatique et sont facile d'utilisation et d'interprétation ; d'autre part, car l'AP permet de répondre à certains désavantages des *k-means* (point que nous développerons par la suite). Le principe des *k-means* est dans un premier temps d'assigner aléatoirement chaque point à un des clusters dont le nombre a été initialement fixé par l'expérimentateur, et de calculer le barycentre de chaque cluster (appelé centroïde) en utilisant une distance spécifiquement choisie (la plupart du temps,

la distance euclidienne). Ensuite, l'algorithme itère en deux étapes : *i)* Réassigner chaque point au cluster dont le centroïde est le plus proche ; *ii)* Calculer un nouveau centroïde pour chaque cluster. Ces 2 étapes sont répétées jusqu'à ce que la variation intra-cluster ne puisse plus être réduite. Cette dernière est calculée par la somme des distances entre les points et leurs centroïdes respectifs. En ce qui concerne l'AP, il s'agit d'une méthode moins connue, mais qui a l'immense avantage de ne pas nécessiter de connaître *a priori* le nombre de clusters, car les points représentatifs des clusters, appelés « exemplaires », sont identifiés automatiquement, et *a fortiori* le nombre $k$ de clusters aussi. Le principe repose sur l'échange de « messages » entre les points. Lors de l'initialisation, chaque point possède un score d'affinité, noté $s$, qui à chaque itération va permettre de quantifier la similarité entre deux points donnés. Si l'on considère trois point $a$, $b$ et $c$, et si le point $a$ est plus similaire à $b$ qu'à $c$, alors les scores d'affinité seront définis tels que : $s(a, b) > s(a, c)$. L'algorithme procède en alternant entre deux étapes de passage de message, qui mettent à jour deux matrices : La matrice de « responsabilité », notée $R$, ayant des valeurs $R(a, b)$ qui quantifient à quel point $b$ est bien adapté pour servir d'exemplaire pour le point $a$, par rapport à d'autres exemplaires candidats pour le point $a$. La matrice de « disponibilité », notée $A$, contient des valeurs $A(a, b)$ qui représentent à quel point il serait « approprié » pour le point $a$ de choisir le point $b$ comme exemplaire, en tenant compte de la préférence des autres points pour ce point $b$ comme exemplaire. Ces deux matrices $R$ et $A$ sont initialisées à zéro, puis l'algorithme réalise deux étapes par itération : *i)* la mise à jour de la matrice $R$ ; *ii)* la mise à jour de la matrice $A$. L'algorithme itère jusqu'à ce que les limites des clusters restent inchangées après qu'un certain nombre d'itération ait été réalisé ou qu'un nombre prédéterminé d'itérations soit atteint. Les points extraits des matrices finales dont la somme « responsabilité + disponibilité » est positive sont considérés comme les exemplaires des clusters. Pour une description plus détaillée de la méthode, notamment les calculs de mise à jour des matrices $R$ et $A$, se référer aux travaux de Frey et Dueck [194].

Les *k-means* présentent quelques désavantages. En effet, ils sont très sensibles aux valeurs extrêmes, et si les données sont légèrement modifiées, les clusters finaux qui en résultent s'en retrouvent fortement modifiés. Cette instabilité est renforcée par l'initialisation de l'algorithme, qui est aléatoire. Par conséquent, plusieurs répétitions des *k-means* ne permettra pas de retrouver les mêmes clusters. Enfin il est nécessaire de spécifier le nombre $k$ de clusters en amont. Or, dans notre cas de protéomique de découverte, on ne connaît pas *a priori* le nombre de clusters, et déterminer le $k$ optimal est un problème d'optimisation difficile [195] et qui est toujours très étudié [196]. Dans une situation où l'expérimentateur

ne saurait pas choisir le $k$ approprié pour partitionner ses données, nous proposons dans notre implémentation de tout d'abord vérifier qu'il est possible de partitionner les données, comme cela est présenté dans les travaux de Adolfsson *et al.* [197]. Pour cela, le *dip test* [198] est utilisé. Ensuite, Le $k$ optimal est déterminé grâce à la statistique d'écart (ou « *gap statistic* » en anglais). Cependant, d'un point de vue pratique, nous conseillons à l'utilisateur de spécifier $k$ pour les *k-means*, car la statistique d'écart donne le $k$ le plus petit possible, tandis qu'en biologie un petit nombre de clusters peut s'avérer non informatif. En revanche, comme indiqué plus haut, l'AP présente l'avantage de ne pas nécessiter la spécification du nombre $k$. Par ailleurs, la nature même de l'algorithme fait que les clusters obtenus pour un ensemble donné de points seront toujours les mêmes, ce qui constitue un second avantage par rapport aux *k-means*. Enfin, l'AP est beaucoup moins sensible aux valeurs extrêmes. Cependant, elle a tendance à générer un grand nombre de clusters, ce qui constitue l'autre extrême à ne pas atteindre, car au-delà d'une dizaine de clusters il devient également difficile pour l'expérimentateur de tirer des conclusions biologiques interprétables. Pour remédier à cela, nous incitons l'utilisateur à s'appuyer sur une paramétrisation de l'algorithme, que j'identifie comme « *affinity propagation reduced* » dans mon implémentation, car elle permet de minimiser le nombre de clusters.

## 6.5 Conclusion

Nous avons vu que dans notre contexte d'analyse à haut-débit, les deux types de MTC (contrôle du FDR et celle intégrée aux tests *post hoc*) doivent cohabiter, or cette interaction n'a non seulement pas été théorisée, mais elle a aussi été désignée comme problématique dans la littérature. Face à cela, nous nous sommes demandé de quelle manière cette cohabitation pouvait avoir lieu, et dans quelle mesure l'une ou l'autre de ces deux MTC pouvait être priorisée en fonction de la question biologique d'intérêt. Nous avons défini les différents rôles que peuvent remplir le contrôle du FDR et les tests *post hoc*, puis nous avons présenté et discuté de la pertinence de six différents scénarios afin d'aider un expérimentateur à faire son choix.

Il s'avère que l'approche la plus répandue est peu pertinente dans bien des cas. Pire, elle peut même poser des difficultés, tant au niveau du contrôle du FDR qu'au niveau de l'interprétation biologique finale. De plus, Nous avons vu qu'il n'est pas toujours indispensable de réaliser l'ANOVA suivie d'un test *post hoc*. Le scénario s'appuyant sur une contextualisation du contrôle du FDR au moyen d'un clustering permet de garantir les

trois rôles du contrôle du FDR, sans nécessiter des compétences statistiques avancées. Cependant, il ne permet pas de répondre à toutes les questions biologiques, notamment celle où l'on souhaite savoir s'il y a une différence entre deux conditions en particulier parmi l'ensemble des conditions considérées. Dans l'ensemble, aucune procédure n'apporte de solution satisfaisante pour préserver les trois rôles (statistique, normatif et biologique) du contrôle le FDR, comme cela est résumé dans le Tableau 6.1.

| **Scénario** | **Contrôle statistique** | **Fonction normative** | **stringence expérimentale** |
|---|---|---|---|
| ANOVA + FDR + PH | N | O | N |
| ANOVA + PH + FDR | O | N | N |
| ANOVA + FDR + clustering | O | O | $\sim$ |
| limma global | N | O | N |
| limma hiérachique | N | O | N |
| limma imbriqué | N | N | $\sim$ |

TABLEAU 6.1 – Tableau résumant les différents rôles du contrôle du FDR qui sont effectivement respectés dans les six scénarios présentés dans ce chapitre. PH : tests *post hoc* ; N : le rôle nest pas maintenu ; O : le rôle est préservé ; $\sim$ : le rôle est partiellement préservé.

Enfin, l'ensemble de la discussion sur les différents rôles du FDR et les différents scénarios abordés montre *i)* qu'il est primordial de définir clairement la question biologique à laquelle l'on souhaite répondre ; *ii)* qu'il est important de déterminer quel(s) sont les rôle(s) du FDR qu'il est nécessaire de préserver ; et *iii)* que plusieurs procédures sont possibles mais elles présentent chacune leurs limitations.

En ce qui concerne mon implémentation du clustering, j'ai choisi de proposer un clustering dit « dur », c'est-à-dire qui affecte une protéine à un seul cluster. Certains auteurs [188] proposent d'appliquer des algorithmes plus « doux » pour les études en biologie, telles que le « *fuzzy c-means* », qui consiste à affecter chaque protéine à plusieurs clusters avec une certaine probabilité associée. Ce type de partitionnement est souvent décrit comme plus intéressant car les protéines ne sont pas nécessairement impliquées dans un seul processus biologique. Cependant, associer une protéine à plusieurs clusters sous le seul argument qu'elle est associée à plusieurs processus biologiques revient finalement à utiliser le clustering uniquement comme un outil de visualisation soucieuse de correspondre à

ce qui est déjà connu, sans nécessairement chercher à mettre en évidence de nouvelles structures ou relations entre protéines. Par ailleurs, cela ne permet pas forcément de tirer des conclusions nécessairement claires et facilement interprétables, car le paramètre de « fuzzification » est difficile à fixer. Dès lors, augmenter le nombre de paramètre à régler par l'expérimentateur pour augmenter l'interprétabilité semble paradoxal.

# Chapitre 7

# Troisième contribution : Application à la cohorte NAFLD

## 7.1 État de l'art de la pathologie

### 7.1.1 Une maladie évolutive, complexe et mondiale

La stéatose hépatique non-alcoolique (« *Non Alcoholic Fatty Liver Disease* », ou **NAFLD**) est aussi dénommée la « maladie du foie gras » ou encore la « maladie du soda » en raison de son étiologie. En effet, il s'agit d'une pathologie hépatique où le foie accumule des graisses, sans que cela soit associé à une consommation d'alcool ou à une hépatite virale [152]. Plus particulièrement, elle est associée à une mauvaise hygiène de vie avec une alimentation trop grasse et trop sucrée, qui caractérise nos sociétés industrialisées. Plus précisement, la stéatose se définit par une accumulation de lipides dans le cytoplasme des hépatocytes (les cellules hépatiques). La maladie est déclarée lorsqu'au moins 5% des hépatocytes présentent une stéatose [152]. Il s'agit d'une pathologie chronique et asymptomatique dans ses stades précoces, avec une évolution progressive de l'atteinte hépatique (illustration en Figure 7.1). La NAFLD débute par une simple stéatose, qui s'accompagne progressivement d'une inflammation des tissus hépatiques, la NAFLD est alors simplement dénommée NAFL. A ce stade, la situation est encore réversible et une prise en charge adaptée du patient conduit à la régression de la NAFLD [199]. Cependant, si la NALF progresse, elle peut évoluer vers une forme symptomatique plus grave de la NAFLD, qu'est la stéatohépatite non-alcoolique (« *Non-Alcoholic Steato Hepatitis* » ou NASH) avec une aggravation de l'inflammation, des lésions hépatocytaires (gonflement et

mort des cellules) et le développement d'une fibrose. La fibrose résulte directement des phénomènes de cicatrisation, conséquence de l'inflammation, du gonflement et de la mort des hépatocytes. A ce stade, l'évolution est encore réversible mais si la NASH progresse, le patient développe une cirrhose ou un cancer hépatique, dont les seules issues sont soit la transplantation hépatique, soit le décès du patient.

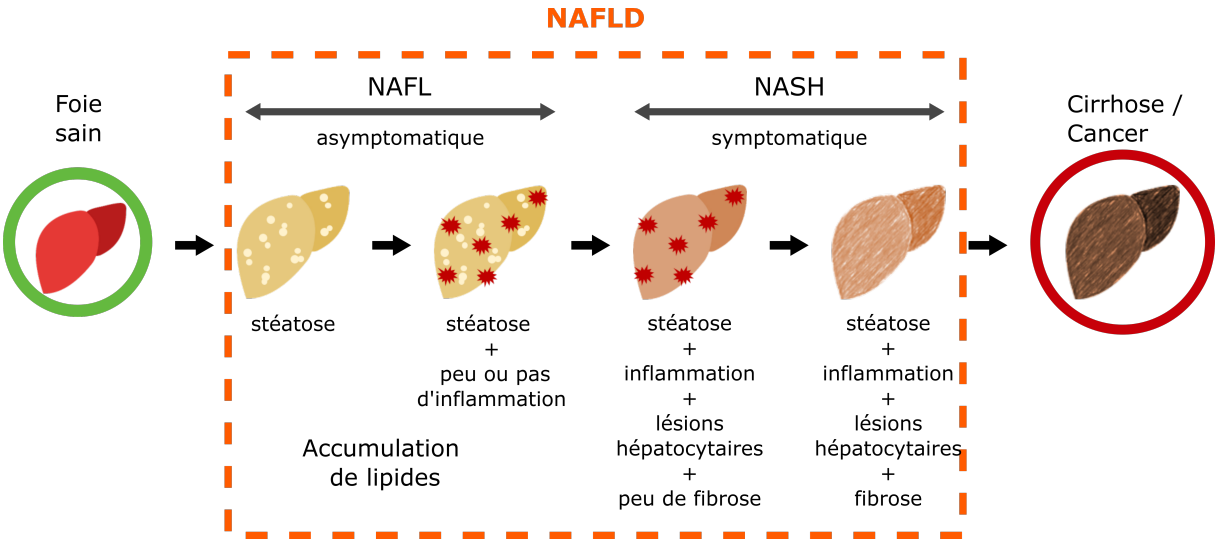



FIGURE 7.1 – Progression de la NAFLD : Les patients développent tout d'abord la NAFL qui est la forme asymptomatique de la maladie, puis ils progressent vers une forme symptomatique plus sévère qu'est la NASH.

Une étude de 2015 [200] a montré qu'un patient NAFL met en moyenne 14 ans pour atteindre le stade de fibrose, tandis qu'un patient NASH met en moyenne 8 ans pour évoluer vers un stade de fibrose plus sévère. Par conséquent, il s'agit d'une pathologie particulièrement lente mais qui s'accélère au fur et à mesure que la situation s'aggrave. Elle est également évolutive, c'est-à-dire que la NAFLD peut régresser, comme elle peut stagner, ou bien se dégrader vers des formes plus graves, et les mécanismes physiopathologiques sous-jacents sont complexes et partiellement élucidés [199]. Il s'avère que la proportion de patients atteints de NAFL qui progressent vers la NASH est relativement faible [199, 201, 202] et que la prévalence de la NASH dans le monde était estimée à seulement 2-3% en 2014 [203] (pour rappel, la prévalence mondiale de la NAFLD est de 25%).

La NAFLD est une pathologie complexe, qui est influencée aussi bien par des facteurs comportementaux qu'environnementaux, notamment l'alimentation et la sédentarité [199]. Des prédispositions héréditaires ont également été identifiées [199, 201]. La NAFLD est

associée à de nombreux facteurs de risques, notamment : l'obésité, le diabète de Type 2 (T2D), le syndrome métabolique, l'hypertension, la résistance à l'insuline et l'hyperlipidémie [201]. L'obésité et le T2D sont les facteurs de risque les plus importants dans le développement d'une NAFLD (notamment, plus de 60% des patients atteints de NAFLD sont diabétiques [201]). Par ailleurs, la prévalence de la NAFLD augmente avec l'âge [201], mais ce phénomène s'explique principalement par la progression plutôt lente de la maladie. Enfin en ce qui concerne la prévalence de la NAFLD en fonction du sexe, les résultats dépendent des pays et des études [201], ce qui en fait un facteur de risque moins décisif. On estime que 25% des adultes de la population mondiale présente une NAFLD, ce qui place cette pathologie comme la première maladie hépatique chronique à l'échelle mondiale, et la NASH est la première cause de transplantation hépatique [199, 202]. En revanche, la prévalence de la NASH est estimée seulement entre 3% et 5% selon les études [201]. Par ailleurs, la prévalence de la NAFLD n'est pas équivalente selon les continents. A titre d'exemple, on l'estime à 24.13% en Amérique du Nord, à 27.37% en Asie, à 13.48% en Afrique et à 23.71% en Europe [201]. L'augmentation la plus importante de la prévalence est en Asie car c'est une région du monde où la population a connu un changement rapide de mode de vie. Il a été montré que la prévalence de la NAFLD est proportionnelle à l'augmentation de l'Indice de Masse Corporelle (IMC) [201] et que les patients atteints de NAFLD ont deux fois plus de risques de développer des maladies cardiovasculaires que la population saine [201]. En France, la cohorte CONSTANCES, pilotée par l'INSERM depuis 2009, recense les données épidémiologiques de plus de 200 000 personnes chaque année. Une étude de 2020 s'est basée sur cette cohorte pour montrer que la prévalence de la NAFLD en France métropolitaine est de 18,2% [204], avec une prévalence augmentée chez les personnes obèses (79,1%), chez les personnes diabétiques (62,4%) et plus généralement, chez les personnes présentant ces deux facteurs de risques (91,5%). De plus, la prévalence de fibrose avancée était estimée à plus de 200 000 personnes en France métropolitaine.

Actuellement le seul traitement réellement efficace est une modification des habitudes alimentaires pour une perte de poids et une augmentation de l'activité physique [201] ; mais cela nécessite une prise en charge précoce et impose aux patients un changement radical de style de vie dans la durée, ce qui est souvent difficile à respecter. En effet, il n'y a actuellement aucun médicament autorisé sur le marché pour traiter la NASH [205] (bien qu'un traitement à base de vitamine E soit employé pour certains patients [206]), et seulement quatre médicaments candidats sont entrés en phase III de développement [206]. Il faut noter que des études montrent que les patients NAFLD/NASH ont une forte réponse au placebo, ce qui complexifie la capacité des études à déceler la réelle efficacité

des médicaments pour le traitement [201, 199] de la NAFLD.

### 7.1.2 Les différentes méthodes de diagnostic

La méthode universelle de diagnostic de la NAFLD (le « gold standard ») est la biopsie hépatique associée à un examen histopathologique [206, 207]. Cependant cette méthode présente des limitations importantes [206, 208, 209]. En effet, il s'agit d'une intervention à la fois lourde et coûteuse avec des risques de complications (hémorragie), voire de décès, pour le patient. De plus, il existe un biais d'échantillonnage car la biopsie représente une très petite fraction du foie. Ce biais d'échantillonnage se double par ailleurs d'un biais d'interprétation car ce dernier est praticien-dépendant et ne fait pas consensus. Enfin, la biopsie étant une méthode très invasive, elle n'est généralement pratiquée que sur les patients dont la NAFLD est déjà très avancée et les patients NASH. Elle n'est donc adaptée ni à un diagnostic précoce et massif de la NAFLD, ni à un suivi médical de longue durée.

En ce qui concerne les méthodes non-invasives, nous pouvons distinguer deux familles d'approches : physiques et biologiques. Les approches physiques rassemblent l'ultrasonographie, l'élastographie et l'imagerie par résonance magnétique (IRM). Bien que l'ultrasonographie soit facile d'accès et peu couteuse (en comparaison de l'IRM), elle a pour principaux désavantages une variabilité inter-praticiens et une sensibilité diminuée en présence de fibrose (notamment pour les patients atteints de NASH, donc). Plus globalement, ces trois approches échouent à différencier les patients NAFL des patients NASH [206], ce qui motive la recherche clinique à se tourner vers les approches biologiques. Ces dernières permettent d'élaborer des tests de diagnostic fournissant des scores de sévérité, en utilisant des mesures cliniques [206]. Les scores de référence pour évaluer le stade de progression de la NASH chez un patient sont ceux basés exclusivement sur la biopsie hépatique : le score « *NAFLD Activity Score* » (NAS) [210] et le score SAF [152]. Cependant, la biopsie présentant de fortes limitations, la recherche clinique a pour objectif de la remplacer par d'autres types de prélèvements, notamment le prélèvement de plasma sanguin. En effet, des biomarqueurs plasmatiques peuvent aider à caractériser les différents stades de progression et la prise de sang est facile à réaliser et peu coûteuse, ce qui présente un clair avantage par rapport à la biopsie pour réaliser des dépistages massifs. Etant donné la grande diversité des phénomènes métaboliques impliqués dans la NAFLD et la NASH, les biomarqueurs recherchés sont principalement ceux impliqués dans l'inflammation, la fibrogenèse, le stress oxydatif, l'apoptose et le métabolisme des lipides et

du glucose [206, 209]. Concernant plus spécifiquement le diagnostic de la NASH, il se base sur les phénomènes fibrotiques, inflammatoires et de mort cellulaire (l'apoptose) [206], tandis que l'on s'intéresse plus à la stéatose pour diagnostiquer la NAFL. L'Alanine Aminotransférase (ALAT1) [1] est le biomarqueur sérique le plus utilisé pour le diagnostic de la NAFLD/NASH, mais il a été montré que les niveaux d'ALAT1 ne permettent pas de prédire efficacement la NASH [206], ce qui encourage la recherche de nouveaux biomarqueurs plus spécifiques. Les biomarqueurs candidats principaux de l'inflammation sont la « *C-Reactive Protein* » (CRP), la « *Tumor Necrosis Factor* » (TNFA) et les InterLeukines 6 (IL6) et 8 (IL8). Concernant la fibrogenèse, les biomarqueurs candidats les plus prometteurs sont les protéines impliquées dans la synthèse du collagène, telles que la « *N-terminal propeptide of type III procollagen* » (PIIINP), la Pro-C3 (CO3) et la « *Tissue Inhibitor of Metalloproteinase 1* » (TIMP1) [208, 211]. Les biomarqueurs utilisés pour le stress oxydatif ne sont pas détaillés ici, car il s'agit principalement de molécules lipidiques qui ne sont pas analysées dans notre étude. Concernant l'apoptose, les biomarqueurs utilisés sont la Cytokératine 18 (K1C18) et l'antigène de surface médiateur de l'apoptose (TNR6). Enfin, le métabolisme des lipides est évalué à l'aide d'hormones telles que l'Adiponectine (ADIPO), la Leptine (LEP), « *Retinol-Binding Protein 4* » (RET4), « *Fatty Acid-Binding Protein 4* » (FABP4) et « *Fibroblast Growth Factor 21* » (FGF21).

Il existe des ensembles de mesures biologiques, aussi appelés panels, qui sont alternatifs aux scores NAS et SAF. Les plus connus pour permettre de diagnostiquer la NAFLD et d'évaluer la stéatose hépatique, sont le FLI (« *Fatty Liver Index* ») [212], le HSI (« *Hepatic Steatosis Index* ») [213], et le SteatoTest [214]. Ces panels diffèrent par la combinaison de marqueurs cliniques et de marqueurs plasmatiques, avec l'IMC qui est systématiquement pris en compte [206]. Les panels pour le diagnostic de la NASH sont l'ActiTest, le FibroTest [215], le Fibrosis-4 (FIB-4) et le « *Enhanced Liver Fibrosis* » (ELF) qui sont pour certains disponibles sous forme de tests commerciaux [206]. Le FibroTest mesure la concentration en GGT, alpha-2-macroglubuline (A2MG), l'apolipoprotéine A1 (APOA1) et l'haptoglobine (HPT) ; il apparaît moins efficace dans le diagnostic des stades précoces de fibrose. FIB-4 est un panel qui utilise l'aspartate aminotransférase (AATC) et ALAT1, qui s'avère modérément efficace dans le diagnostic de la fibrose. Le ELF est performant pour le diagnostic des stades avancés de fibrose mais pas pour les stades précoces [206]. Tous ces panels et biomarqueurs se heurtent au problème de la validation clinique. En effet, pour qu'un biomarqueur soit considéré comme un critère valide de substitution à

1. Dans ce chapitre, les protéines sont abrégées en utilisant leur nom d'entrée UniProtKB

la biopsie, il faut qu'il y ait suffisamment de preuves cliniques montrant une efficacité établie pour le diagnostic de la pathologie [208]. Par « efficacité », nous entendons une excellente sensibilité et spécificité, telles que définies au Chapitre 1. Or, parmi tous ces biomarqueurs candidats, seul K1C18 est cliniquement validé, les autres étant controversés selon les études cliniques ou bien nécessitant des études complémentaires pour leur validation [206, 199], comme c'est le cas notamment de PIIINP et de Pro-C3 qui ne sont validés qu'au niveau expérimental [208]. TIMP1 a montré une excellente capacité à différencier des patients NASH de patients sains [206]. Par ailleurs, ces biomarqueurs candidats ne sont pas spécifiques du foie, et peuvent être indicateur d'autres troubles inflammatoires ou métaboliques ; de sorte qu'ils ne peuvent pas être utilisés seuls. La description détaillée des différents biomarqueurs ainsi que des scores présentés succinctement ci-dessus peut être trouvée dans les références suivantes [199, 206, 208].

## 7.2 Plan d'expérience et analyse protéomique

### 7.2.1 Description de la cohorte étudiée

La cohorte NAFLD étudiée dans ce projet était accompagnée d'une base de données cliniques comportant 127 variables décrivant l'état clinique et physiologique des 160 patients qui ont été prélevés. Certaines variables étaient peu renseignées, notamment celle de la prise de médicaments (qui ne comportait l'information que pour un unique patient) et n'ont donc pas été explorées pour la caractérisation de la cohorte. De même, l'information sur la variable « obésité » était parfois manquante et a été complétée en se basant sur la variable « IMC ». Finalement, et après échanges avec l'équipe clinique, ce sont seulement 5 variables qui ont été plus approfondies pour décrire la cohorte : l'obésité, le diabète, l'élastographie hépatique, l'algorithme FLIP (« *Fatty Liver Inhibition of Progression* ») et le SAF. D'une part parce qu'elles ne comportaient pas de valeurs manquantes, et d'autre part parce qu'elles fournissaient une information sur la fibrose, facteur que nous voulions étudier en priorité (cf. Chapitre 4). En outre, le diabète et l'obésité sont les facteurs de risques les plus importants, et le score SAF et l'algorithme FLIP ont été développés conjointement pour fournir un diagnostic [207]. Il semblait donc intéressant d'étudier ces derniers pour l'analyse descriptive de la cohorte. Le facteur retenu pour réaliser l'analyse statistique des données quantifiées pour les protéines est le score SAF, principalement du fait de sa construction. En effet, le score SAF est en fait constitué d'un ensemble de trois scores : S pour la Stéatose, A pour l'Activité et F pour la fibrose [207]. Il permet donc de

catégoriser facilement les patients selon ces trois critères histologiques bien distincts. Le score S correspond au pourcentage de gouttelettes lipidiques présentes dans les hépatocytes. Le score A correspond à la somme de deux scores : le gonflement des hépatocytes (aussi dénommé « *ballooning* » par l'anglicisme) et l'inflammation du tissu hépatique. Le score de F de fibrose utilise la même notation que celle employée par le NAS [199, 207]. Une description plus détaillée du score SAF et de l'algorithme FLIP est présentée dans la Figure 7.2. L'algorithme FLIP se base sur ce score pour établir si un patient est NAFLD, NASH ou s'il n'est pas NAFLD. Ainsi, un patient n'est pas considéré comme ayant une NAFLD (et donc une NASH a fortiori) à partir du moment où il n'a pas de stéatose (S0). En revanche, un patient est considéré comme NAFLD dès lors qu'il présente une stéatose (S$\geq$1) et qu'il ne présente pas d'inflammation (Figure 7.2B). Un patient est considéré comme NASH à partir du moment où il présente un score S1-A2, quel que soit son score de F [207] (Figure 7.2B).

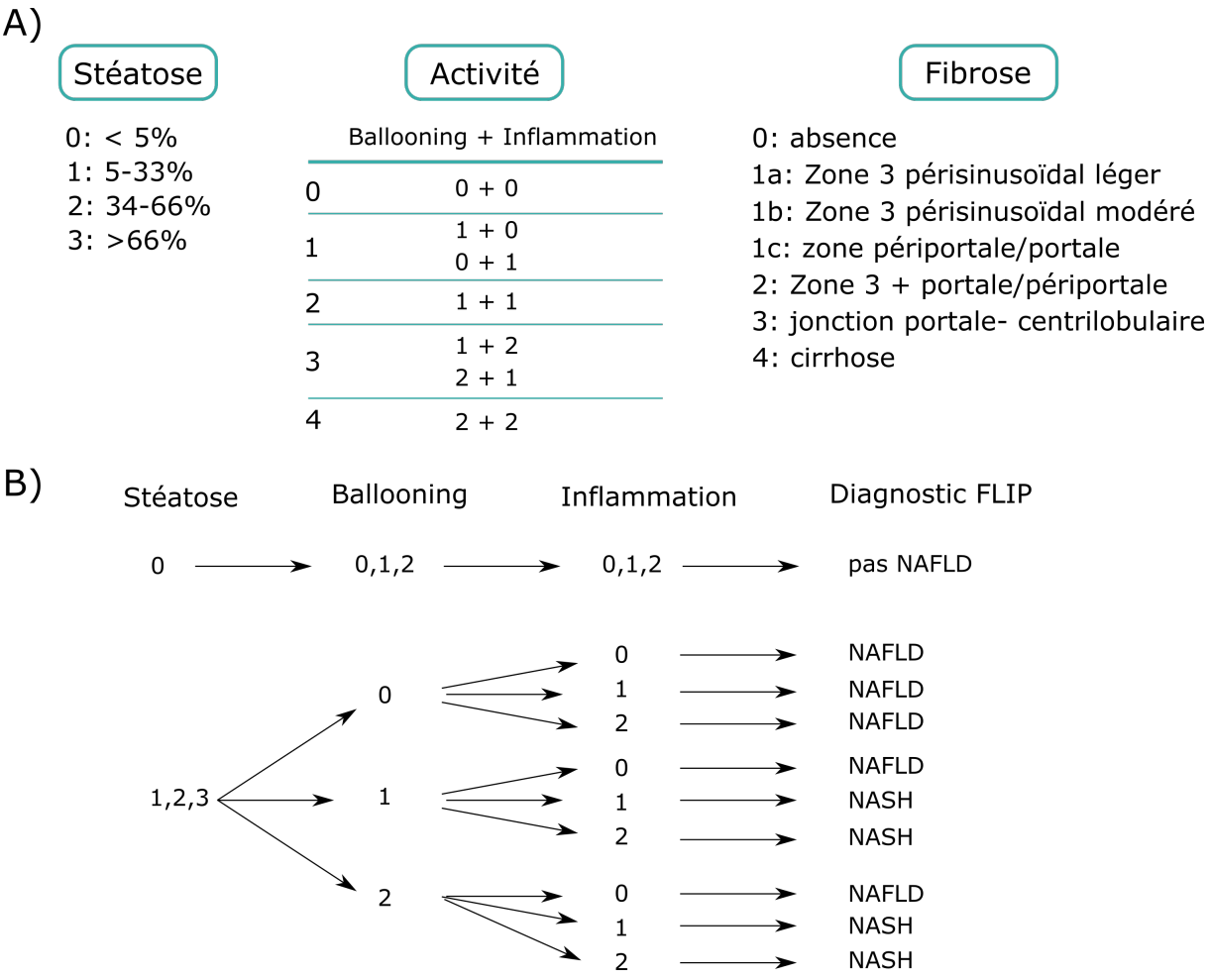



FIGURE 7.2 – Description de la construction **A)** du score SAF, **B)** de l'algorithme FLIP (repris de Bedossa *et al.* [207]).

### 7.2.2 Étude statistique descriptive et exploratoire

La cohorte de patients comporte 64 femmes (39,5%) ; 54,9% de patients sont obèses ; 40,7% sont diabétiques et 28,39% sont obèses et diabétiques. Nous avons réalisé une étude univariée du score SAF, c'est-à-dire que nous avons regardé la répartition des patients en fonction de chaque sous-score S, A et F dont les résultats sont résumés dans le Tableau 7.1. On constate de manière intéressante que 11 patients ne sont pas NAFLD selon les critères de l'algorithme FLIP. Parmi ces 11 patients, 6 étaient S0A0F0 dont un seul obèse et un seul diabétique, ce qui suggère que ce sont vraisemblablement des patients n'ayant pas d'atteinte hépatique. Cependant, ces patients ont tout de même été retenu pour l'analyse protéomique, car cela aurait réduit considérablement le nombre d'échantillons pour le stade F0.

| Score | S | A | F |
|---|---|---|---|
| 0 | 11 | 20 | 25 |
| 1 | 32 | 31 | 30 |
| 2 | 54 | 41 | 46 |
| 3 | 63 | 54 | 33 |
| 4 | | 14 | 26 |

TABLEAU 7.1 – Répartition des patients selon les sous-scores du SAF. Lecture : 11 patients ont un score de Stéatose égal à 0.

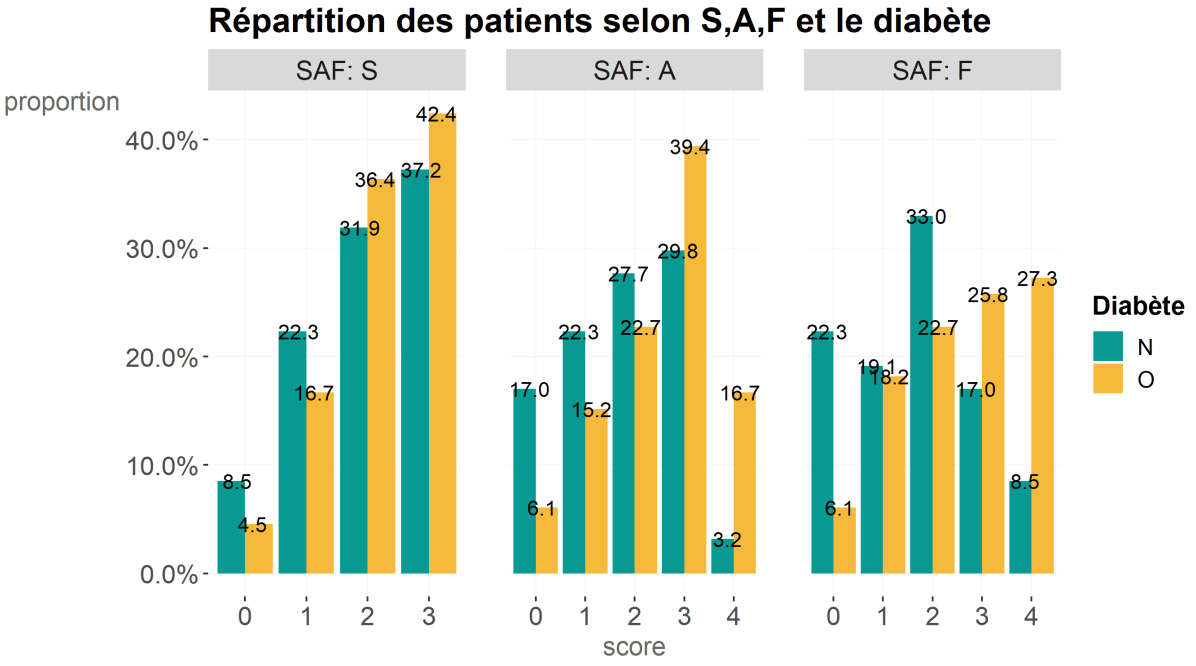



FIGURE 7.3 – Répartition des patients en fonction du score SAF et du diabète (présence : O / absence : N). Lecture : 4,5% des patients diabétiques ont un score de stéatose S0, 6,1% des patients diabétiques ont un score d'activité A0, 6,1% des patients diabétiques ont un score de fibrose F0. On remarque que les patients diabétiques (O, en jaune) ont tendance à avoir une activité (A) généralement plus avancée et une fibrose (F) plus sévère que les patients non diabétiques (N, en bleu) avec seulement 8,5% des non diabétiques ayant une cirrhose contre 27,3% des diabétiques.

La Figure 7.3 représente la proportion relative des patients diabétiques ou non en fonction du score SAF. Cette répartition suggère que les patients diabétiques ont tendance à avoir un score de stéatose (S) plus élevé que les patients non diabétiques (Figure 7.3, SAF : S). Concernant le score d'activité (A), la proportion de patients diabétiques augmente avec la sévérité du score, avec un maximum de 39,4% des patients diabétiques ayant un score A3 (Figure 7.3, SAF : A). En outre, les patients diabétiques présentent une fibrose (F) globalement plus sévère que pour les patients non diabétiques (Figure 7.3, SAF : F). La Figure 7.4 représente la proportion de patients obèses ou non en fonction du

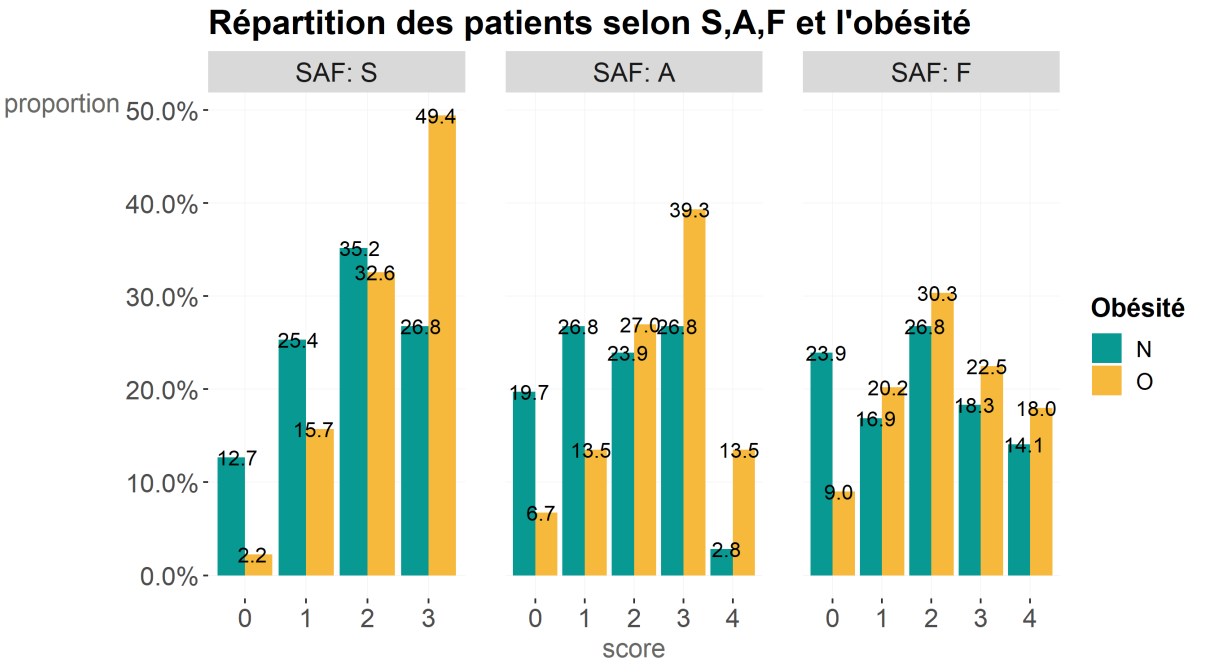



FIGURE 7.4 – Répartition des patients en fonction du score SAF et de l'obésité (présence : O / absence : N). Lecture : 2,2% des patients obèses ont une stéatose S0. On remarque que les patients obèses (O, en jaune) présentent plus fréquemment une stéatose plus sévère ainsi qu'une activité beaucoup plus développée par rapport aux patients non obèses (N, en bleu).

score SAF. Cette répartition montre très clairement que la grande majorité des patients obèses présentent une stéatose plus sévère que les patients non obèses (Figure 7.4, SAF : S), avec 49,4% des patients obèses ayant un score S3. De même, une grande majorité des patients obèses présentent un score d'Activité plus sévère que pour les patients non obèses, avec des scores A3 (39,3%) et A4 (13,5%), alors que seulement 2,8% des patients non obèses ont un score A4. Enfin en ce qui concerne le score F, les patients obèses représentent majoritairement les stades F1 (20,2%), F2 (30,3%), F3 (22,5%) et F4 (18%). Par conséquent, ces résultats confirment ce qui est relevé dans la littérature, à savoir que les patients obèses et/ou diabétiques, ont tendance à avoir une NAFLD/ NASH plutôt avancée au moment du diagnostic étant donné que ce sont des facteurs aggravants de progression de la NAFLD/NASH.

Le score SAF, étant donné sa construction, permet d'obtenir 100 classes possibles de scores. L'histogramme présenté en Figure 7.5, illustre la disparité des effectifs selon les 57 classes représentées dans la cohorte.

On comprend avec la Figure 7.5, que réaliser une analyse statistique sur le score SAF pris dans son ensemble, autrement dit réaliser une analyse statistique multivariée sur les

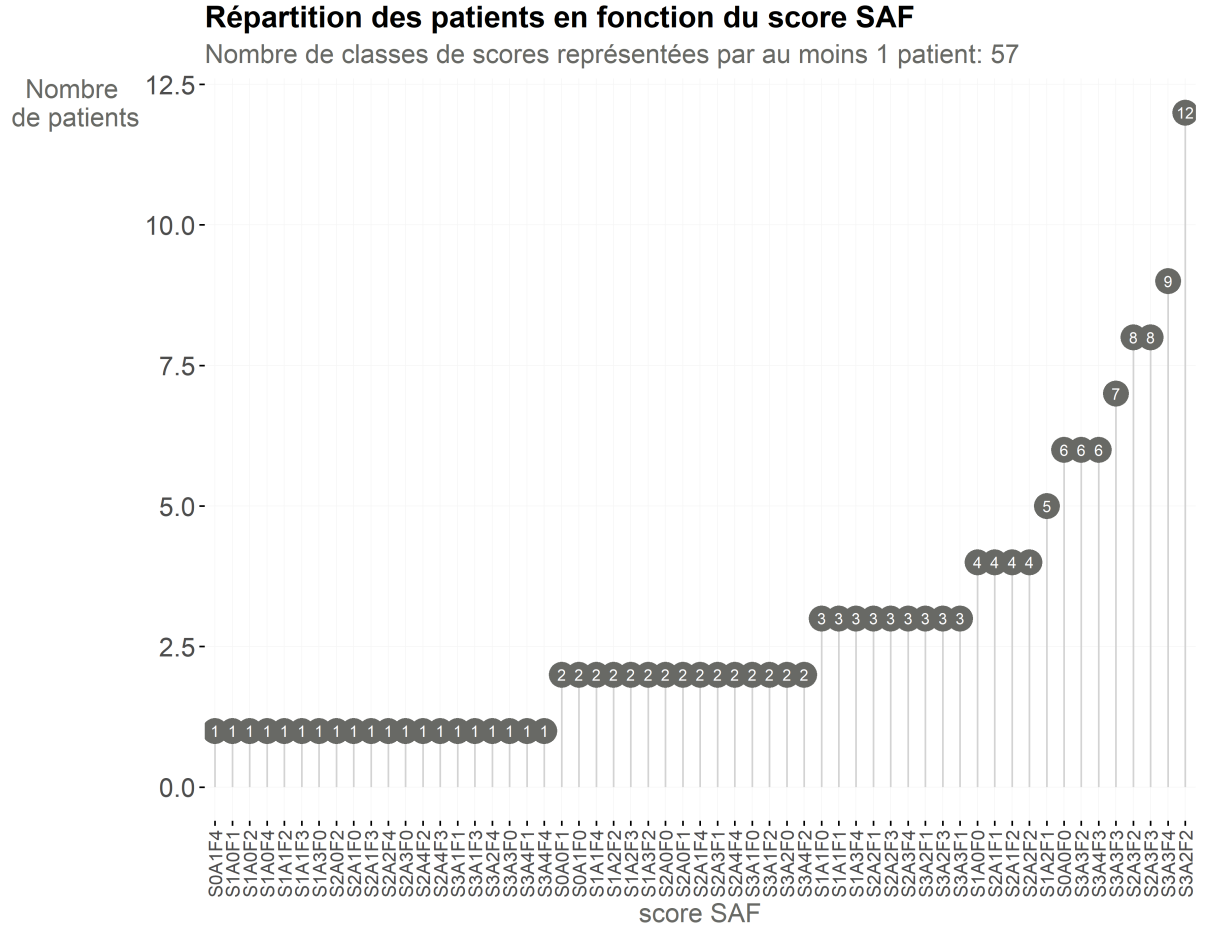



FIGURE 7.5 – Répartition des patients selon le score SAF. Seulement 57 parmi les 100 scores possibles du SAF sont représentés dans la cohorte. Lecture : 12 patients ont un score SAF S3A2F2.

trois scores, ne serait pas raisonnable compte-tenu de la taille de la cohorte, mais aussi compte-tenu des déséquilibres entre classes (certaines ne comportant qu'un seul patient). Plus précisément, deux patients présentant un score F1 n'auront pas forcément le même score de A ni de S, et donc chaque score implique des groupes composés d'individus différents. De plus, si l'on réalisait une étude multivariée, par exemple entre A et F, les groupes formés seraient encore plus déséquilibrés et petits que si l'on considère chaque score pris séparément (voir Tableau 7.2). Par conséquent l'analyse multivariée risquerait d'être peu puissante, ce qui nous a conduit à écarter ce type d'analyse pour l'identification de protéines différentiellement abondantes. Nous avons choisi de réaliser une analyse statistique uniquement sur un seul des sous-scores, afin de maximiser le nombre d'échantillon par groupe du facteur étudié. Nous nous sommes focalisés sur l'analyse du score F, représenté par 5 groupes (F0, F1, F2, F3 et F4), comme cela a été justifié à la Section 4.1.

| | | F | | | | |
|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 |
| **A** | 0 | 12 | 5 | 2 | 0 | 1 |
| | 1 | 8 | 8 | 7 | 3 | 5 |
| | 2 | 2 | 11 | 18 | 8 | 2 |
| | 3 | 3 | 5 | 16 | 15 | 15 |
| | 4 | 0 | 1 | 3 | 7 | 3 |

TABLEAU 7.2 – Répartition des patients en fonction des scores A et F du score SAF. On remarque que les effectifs sont très déséquilibrés avec des scores sur-représentés tels que A2F2 et A3F2 et des scores n'ayant pas de patients (A0F3 et A4F0) ou un seul patient (A4F1 et A0F4).

### 7.2.3 Stratégie de préparation des échantillons pour l'analyse MS

Afin de limiter les effets de lots, notamment l'effet « plaque » décrit au Chapitre 5, nous avons utilisé wpm (cf. Section 5.2) pour répartir les 160 échantillons sur des plaques 96 puits. L'objectif est de tenir compte des trois scores S, A et F tout en évitant d'avoir 57 groupes aux effectifs très déséquilibrés comme cela a été montré dans la Figure 7.5. Par ailleurs, réduire le nombre de groupe permet de trouver plus facilement un positionnement valide avec wpm. Nous avons donc partitionné les échantillons de plasma en 12 groupes représentatifs de la sévérité du score SAF, comme cela est détaillé dans la Figure 7.6. Il est important de noter que ces groupes ne correspondent pas aux conditions comparées lors de l'analyse statistique (Section 7.3.2), mais ont été choisis uniquement pour réaliser le placement des échantillons sur les plaques à puits afin de limiter de potentiels effets de lots. Le regroupement présenté en Figure 7.6 nous permet d'obtenir 10 groupes non-vides (ainsi que 2 groupes vides, c'est-à-dire pour lesquels nous n'avons pas de patient) décrivant la sévérité de la maladie.

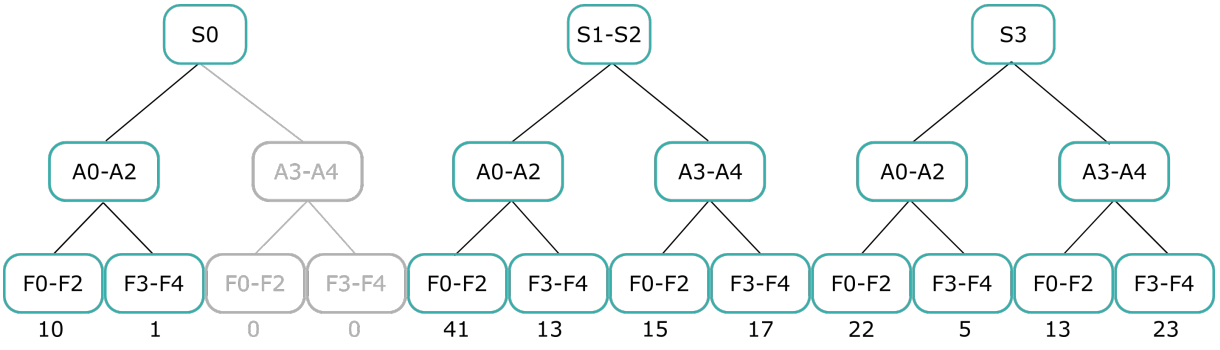



FIGURE 7.6 – Répartition des échantillons en fonction des différents scores S, A et F. les scores sont regroupés par sévérité : (S1-S2) ; (A0-A2) ; (A3-A4) ; (F0-F2) ; (F3-F4). Les groupes pour lesquels il n'y a aucun échantillon sont colorés en gris.

Afin de maximiser l'utilisation de l'espace sur les plaques de 96 puits, nous avons réparti les 160 échantillons sur deux plaques en affectant 80 échantillons à chacune et en veillant à les répartir le plus uniformément possible selon les groupes. Les plans de plaques sont présentés en Figure 7.7. Par mesure de précaution, nous avons considéré les 16 puits de chaque extrémité de plaque comme des puits interdits (voir Figure 7.7) car si le chauffage des plaques nécessaire à la solubilisation des protéines, à la réduction des ponts disulfures et à l'alkylation des cystéines (chauffage à 95°C dans un appareil à PCR) n'était pas homogène, les échantillons placés sur les bords de plaque risqueraient de subir des biais de préparation en comparaison de ceux du centre de la plaque. Nous avons utilisé la contrainte de voisinage NSEO (Nord-Sud-Est-Ouest) qui interdit de placer deux échantillons de même groupe côte-à-côte sur les deux axes horizontal et vertical. Pour la sérialisation des échantillons pour l'analyse MS, nous avons alterné des échantillons provenant de chaque plaque et réalisé la séquence en partant du coin supérieur gauche pour la première plaque, et inférieur droit pour la seconde : 113, 194, 106, 23, 108, 64, ..., 208, 16, 132, 164.

Les analyses DDA et DIA ont été réalisées par les ingénieurs de la plateforme analytique du laboratoire. L'analyse DIA a été réalisée en plus de la DDA afin d'évaluer les différences potentielles de résultats générés par ces deux méthodes (profondeur d'analyse, taux de valeurs manquantes de quantification, reproductibilité de la quantification, *etc.*). A noter que pour mon projet, la priorité fut donnée à l'analyse des données DDA. Les données DIA ne furent pas analysées car cela aurait nécessité une adaptation de la méthodologie développée, aussi bien au niveau des étapes de pré-traitement qu'au niveau de l'analyse statistique, ce qui nécessitait un recul et une expertise différents pour être en mesure d'évaluer la pertinence de l'approche que je propose dans ce travail de thèse.

## 7.3 Analyse statistique des données DDA

### 7.3.1 Pré-traitement des données de quantification

L'analyse MS en DDA de la cohorte NAFLD a permis d'identifier et de quantifier 352 protéines pour 159 échantillons sur les 160 échantillons initiaux. En effet, malgré les précautions décrites plus haut sur le placement des échantillons, l'un d'eux fut mal digéré et fut par conséquent écarté de l'analyse. Une étape d'AC des peptides a été réalisée lors du traitement bioinformatique (cf. Section 2.3.2).

Lors de l'étape de filtrage, trois contaminants ont été écartés car ces protéines ne pro-

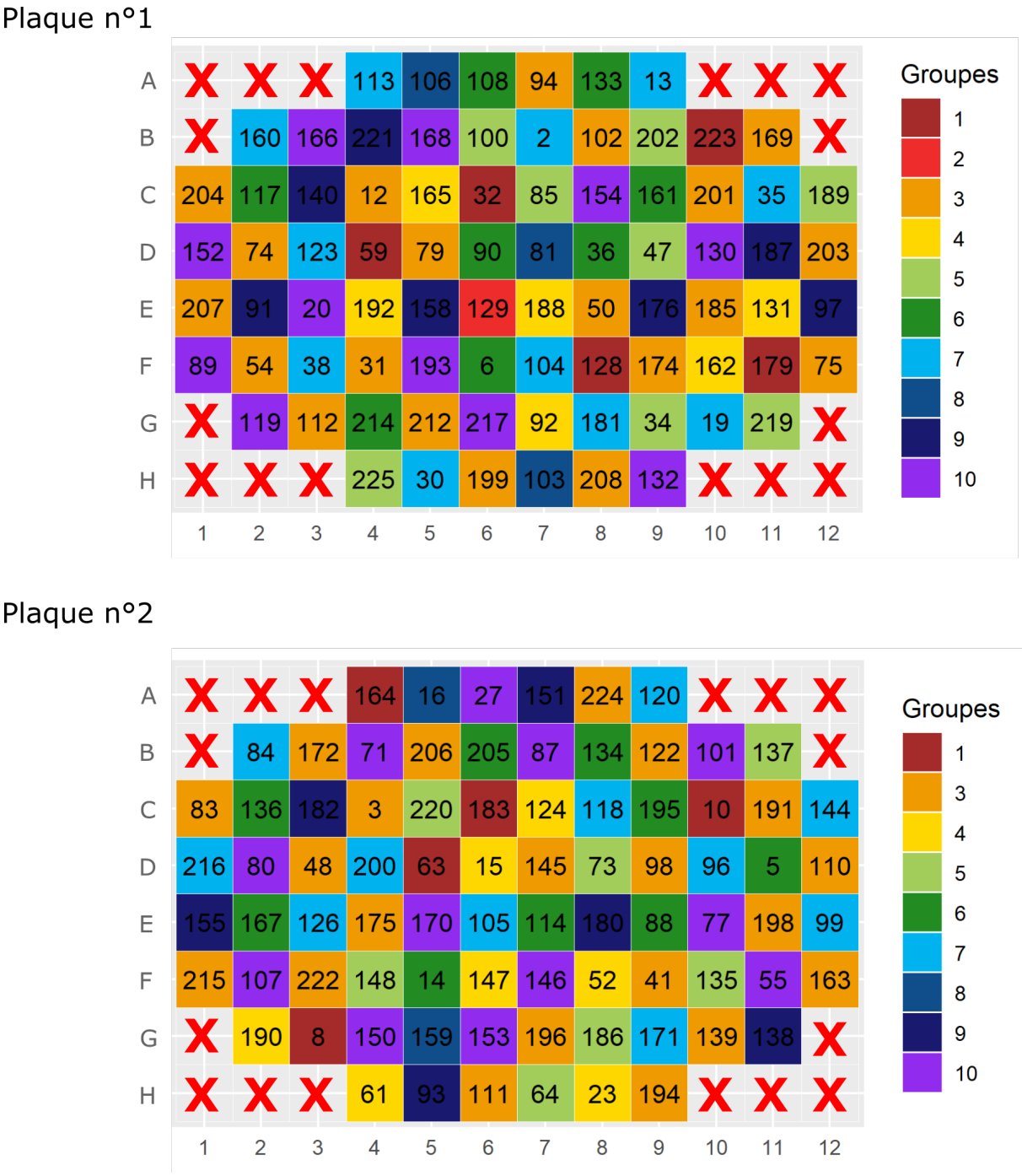



FIGURE 7.7 – Répartition des échantillons de la cohorte NAFLD à l'aide de WPM. Les groupes correspondent à ceux définis dans la Figure 7.6. Les puits interdits sont symbolisés par une croix rouge.

viennent pas des échantillons plasmatiques mais ont été rajoutés lors de leur préparation. Il s'agissait de la trypsine porcine (code UniProt P00761), du cytochrome C bovin (code UniProt P62894) et des standards iRT de Biognosys. Nous avons réalisé un filtrage en

conservant les protéines pour lesquelles il y avait au moins 75% de valeurs observées dans au moins une condition. Ce seuil permettait de ne pas perdre trop de protéines dès cette première étape de pré-traitement, tout en s'assurant de ne pas avoir à imputer largement le jeu de données. Ce filtrage nous a permis de conserver 244 protéines, soit 69% du nombre initial de protéines.

Concernant l'étape de normalisation, nous n'avons pas testé des méthodes de normalisation telles que l'alignement des quantiles, l'usage d'abondances relatives intra-échantillon ou encore le centrage-réduction (voir définitions dans la Section 3.1.4). En effet, les variations d'abondance des protéines dans le plasma sanguin sont très fines, or ces méthodes réduisent fortement la variabilité entre les échantillons, par conséquent les utiliser sur nos données aurait vraisemblablement supprimé une grande part de la variabilité biologique. Nous avons donc testé trois autres méthodes : LOESS (cf. Section 5.3.1), vsn (cf. Section 5.3.2) et le centrage sur la médiane (cf. Section 3.1.4). La méthode de normalisation retenue a été la méthode LOESS avec la valeur m de fenêtre fixée à 0,5. En effet, cette méthode présentait la meilleure calibration des $p_{valeurs}$ respectivement à vsn et le centrage sur la médiane (cf. Chapitre 5). De plus, les calibrations étaient meilleures lorsque la normalisation était réalisée pour chaque condition plutôt que sur l'ensemble des conditions, ce qui peut s'expliquer par le fait que chaque condition biologique possède suffisamment d'échantillons et que ceux-ci se trouvent dans des gammes d'abondance similaires. Les données ainsi filtrées et normalisées présentaient des valeurs manquantes de façon partielle pour chaque condition (POV *partially Observed Values*) mais aucune valeur manquante pour une condition entière (MEC *Missing on the Entire Condition*). Par conséquent, la méthode d'imputation slsa [183] a été appliquée car il s'agit de celle préconisée dans *Prostar* [73] dans ce type de cas, ainsi que celle généralement employée au laboratoire. La Figure 7.8 présente les boîtes à moustaches des données brutes et des données après les étapes de pré-traitement (filtrage, normalisation et imputation). On constate que la dispersion des données brutes est très similaire entre les différentes conditions comparées (les cinq stades de F), et que la différence avec les données prétraitées n'est pas graphiquement visible. L'apport de la normalisation a été principalement mis en évidence grâce aux calibration-plots (cf. Section 5.3.3).

### 7.3.2 Analyse statistique du facteur de fibrose des données DDA

Comme l'objectif de l'analyse différentielle est ici de trouver les protéines pour lesquelles une différence d'abondance moyenne entre les différents stades du score F (F0, F1, F2,

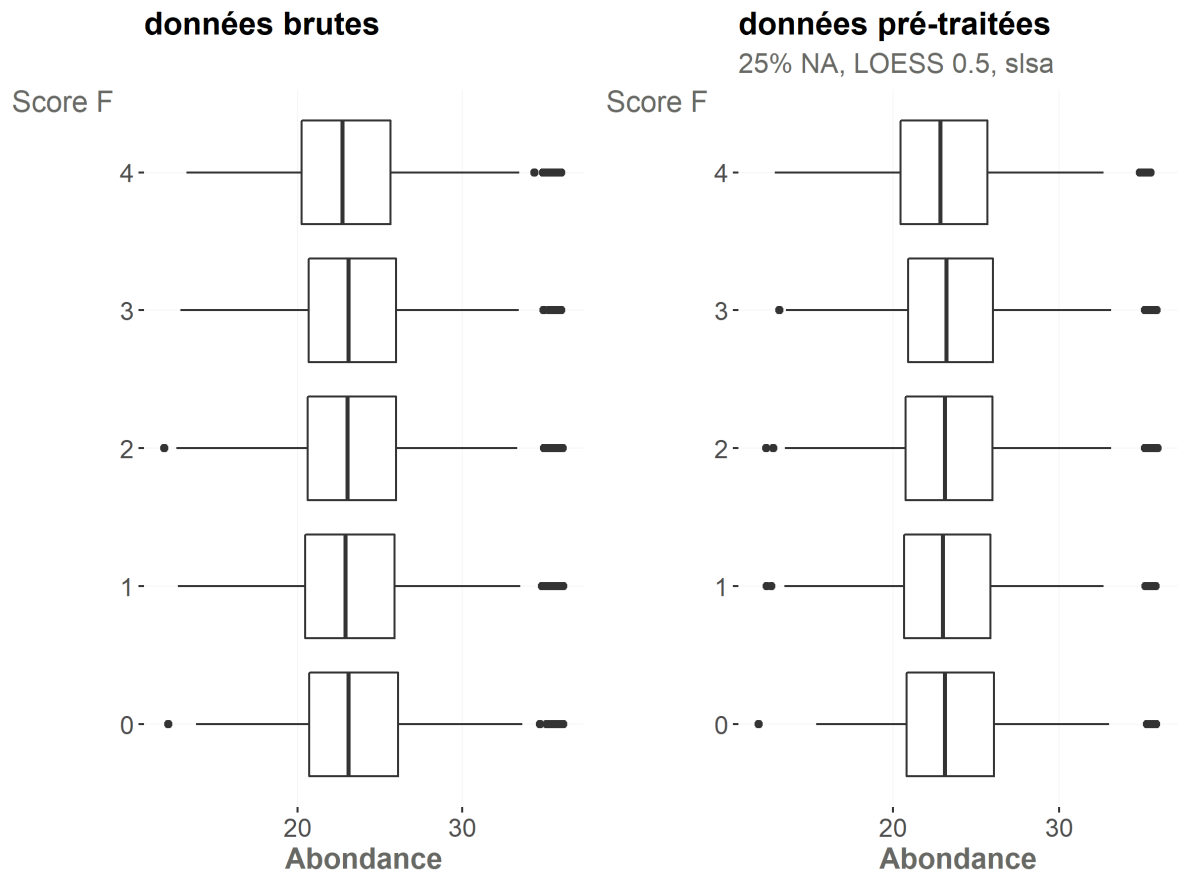



FIGURE 7.8 – Comparaison entre les données brutes issues de l'analyse MS (à gauche), et les données pré-traitées (à droite). Le pré-traitement des données a consisté en un filtrage des protéines présentant au moins 75% de valeurs observées (25% de valeurs manquantes (NA)), une normalisation LOESS m=0,5 et une imputation par la méthode slsa.

F3, F4) est statistiquement significative, les hypothèses nulle et alternatives choisies sont celles de l'ANOVA :

$H_0$ **:** « il n'existe pas de différence d'abondance entre les différents stades du score F. »

$H_1$ **:** « il existe un stade pour lequel une différence d'abondance est observée par rapport aux autres stades pour le score F », qu'il s'agisse d'une surabondance ou d'une sous-abondance entre les différents stades.

Nous avons vérifié les conditions d'application de l'ANOVA. Pour rappel, ces conditions d'application sont décrites à la Section 3.3.2. En ce qui concerne la normalité des distributions, il existe de nombreux tests statistiques [216], tels que le test de Shapiro-Wilk et le test de Kolmogorov-Smirnov. Bien que le test de Shapiro-Wilk présente une puissance

statistique élevée [217, 218, 219], les méthodes graphiques sont une bonne alternative à ces tests [220]. Nous avons donc utilisé la méthode graphique du diagramme Quantile-Quantile, aussi appelé *Q-Q plot*, qui consiste à comparer la distribution des quantiles observés en fonction des quantiles théoriques d'une distribution gaussienne. Si la relation est linéaire, autrement dit si les points sont alignés sur la première bissectrice ($x = y$), alors la distribution théorique proposée (ici une distribution normale) est une modélisation adaptée des valeurs observées. Cependant cela suppose de générer un Q-Q plot pour chaque protéine pour chaque stade de F, ce qui représente 1220 (244 protéines x 5 stades) Q-Q plots à visualiser. Afin de simplifier l'évaluation graphique des Q-Q plots, nous avons centré-réduit les abondances pour chaque protéine, puis réalisé un Q-Q plot pour l'ensemble des données qui est représenté dans la Figure 7.9. Le Q-Q plot montre que la distribution des valeurs observées suit la première bissectrice pour une grande majorité des valeurs. Les valeurs les plus éloignées de la bissectrice correspondent aux valeurs extrêmes des données observées. Si l'on considère l'intervalle centré en la moyenne et de longueur égale à 2 fois l'écart-type, on constate que le pourcentage de valeurs observées comprises dans cet intervalle est de 95.45%. Il y a donc moins de 5% de valeurs extrêmes et nous pouvons donc considérer qu'il s'agit d'une distribution normale.

Concernant l'homoscédasticité des données qui est exigée par l'ANOVA, il existe de nombreux tests statistiques pour l'évaluer [221], les plus connus étant les tests de Bartlett [222] et de Levene [223]. Cependant, ces tests ont eux-mêmes comme condition d'application, l'hypothèse que les échantillons de chaque condition (ici les stades de F) suivent une loi normale. Le test de Bartlett est le moins robuste et le plus sensible à la non-normalité des données tandis que le test de Levene s'avère robuste en cas de violation de l'hypothèse de normalité des données [224, 225]. Une étude montre que les deux tests ont une puissance statistique relativement similaire selon les différents scénarii testés [226], avec une puissance supérieure à 0,90 pour un nombre de conditions égal à 5, comme c'est le cas dans notre analyse. Cette même étude met en évidence que l'ensemble des tests statistiques perdent beaucoup en puissance lorsque les variances sont peu différentes. Nous avons choisi de vérifier l'homoscédasticité des données pour chaque protéine en fonction des stades de F de fibrose à l'aide du test de Levene plutôt que d'utiliser une méthode graphique car le nombre de graphiques qu'il fallait générer aurait rendu l'interprétation délicate. Un test de Levene a été réalisé pour chacune des 244 protéines. Après ajustement des $p_{valeurs}$ par la méthode de Benjamini-Hochberg (en raison de la multiplicité des tests), aucun test ne s'est révélé être significatif ; par conséquent toutes les distributions peuvent être considérées comme homoscédastiques. Dans notre étude les effectifs pour

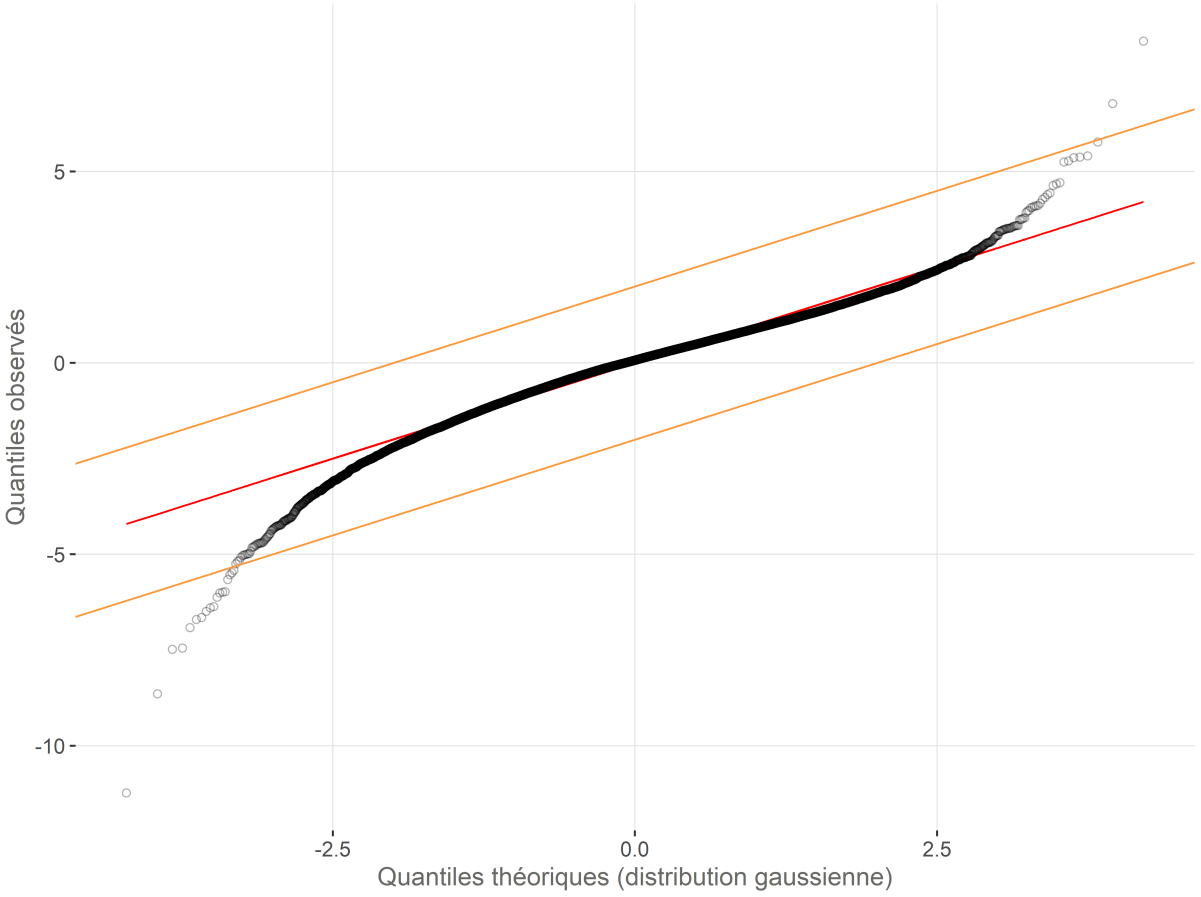



FIGURE 7.9 – Q-Q plot pour les données NAFLD centrées réduites. La première bissectrice (en rouge) représente le cas où les données observées alignées sur cette droite suivent une distribution normale à la perfection. Les deux droites orange représentent les bornes supérieure et inférieure de l'intervalle [-2 écart-type ; 2 écart-type]. Les points sont majoritairement alignés sur la droite et 95,45% des valeurs sont comprises dans l'intervalle, ce qui suggère que la distribution observée suit approximativement une loi normale.

chacun des stades F sont déséquilibrés (F0 : 24 échantillons, F1 : 30 échantillons, F2 : 46 échantillons, F3 : 33 échantillons et F4 : 26 échantillons). Or, l'ANOVA est un test statistique sensible à un plan d'expérience déséquilibré car cela accroit l'hétéroscédasticité des données [227, 129]. Bien que l'homoscédasticité ait été vérifiée avec le test de Levene, les effectifs sont tout de même très différents et il se peut que le test n'ait pas été suffisamment puissant. Nous avons donc décidé d'utiliser la méthode robuste de Welch [127] pour réaliser l'ANOVA à 1 facteur, qui, en cas de variances inégales, permet de réaliser une ANOVA sans perdre en puissance statistique. Nous n'avons pas utilisé la méthode *Limma* (méthode décrite dans le Chapitre 6) qui est généralement privilégiée dans *Prostar* pour deux raisons : d'une part à cause des effectifs à disposition dans chaque stade, plus

importants que dans la plupart des cas d'usage de *Prostar*, de sorte que la modération de la variance ne semble pas aussi nécessaire (et risquerait d'augmenter artificiellement la puissance statistique au détriment de la robustesse du test) ; et d'autre part, car aucune documentation ne fait référence au cas d'un plan déséquilibré, et par conséquent nous ne savons pas si *Limma* permet effectivement de réaliser une ANOVA en cas de plan d'expérience déséquilibré sans qu'il n'y ait perte de puissance statistique. Nous avons ensuite ajusté les $p_{valeurs}$ par la méthode de Benjamini-Hochberg [134] (notées dans la suite $q_{valeur}$) et nous avons obtenus 114 protéines statistiquement différentiellement abondantes (notées DA par la suite) pour un FDR de 1%, ce qui représente 47% des protéines initiales.

Nous avons ensuite réalisé le clustering de profils d'abondance des 114 protéines avec l'approche décrite au Chapitre 6, et en utilisant la méthode de clustering nommée « *Affinity propagation* »[194]. La Figure 7.10 présente les quatre clusters obtenus avec chaque profil d'abondance coloré en fonction de la $q_{valeur}$. De manière fort intéressante, ce graphique illustre parfaitement l'absence de lien entre $q_{valeur}$ et l'importance de l'effet qui avait été expliqué à la Section 3.2.4. En effet, nous pouvons remarquer que les profils colorés en rouge foncé ($q_{valeur} < 1.10^{-9}$), notamment dans les clusters 2 et 4, ne sont pas ceux dont le profil d'abondance présente la plus grande différence d'abondance entre les conditions. Par ailleurs, on observe que l'abondance des protéines s'effondre entre le stade F3 et le stade F4 dans les clusters 2, 3 et 4. Dans ces clusters, nous retrouvons des protéines telles que l'Albumine (ALBU) (cluster 4, $q_{valeur} = 8,47.10^{-12}$) qui est impliquée dans la régulation de la pression oncotique sanguine. Ce phénomène est en adéquation avec ce qui est connu pour cette pathologie dans les stades avancés, car au stade F4 qui correspond à la cirrhose, le foie n'est plus fonctionnel et sécrète moins de protéines plasmatiques (dont l'albumine). Le cluster 1 rassemble les protéines pour lesquelles le profil augmente progressivement depuis F0 jusqu'à F4. Il comprend notamment le biomarqueur A2MG ($q_{valeur} = 4,18.10^{-4}$), qui joue un rôle dans la régulation négative de la coagulation sanguine.

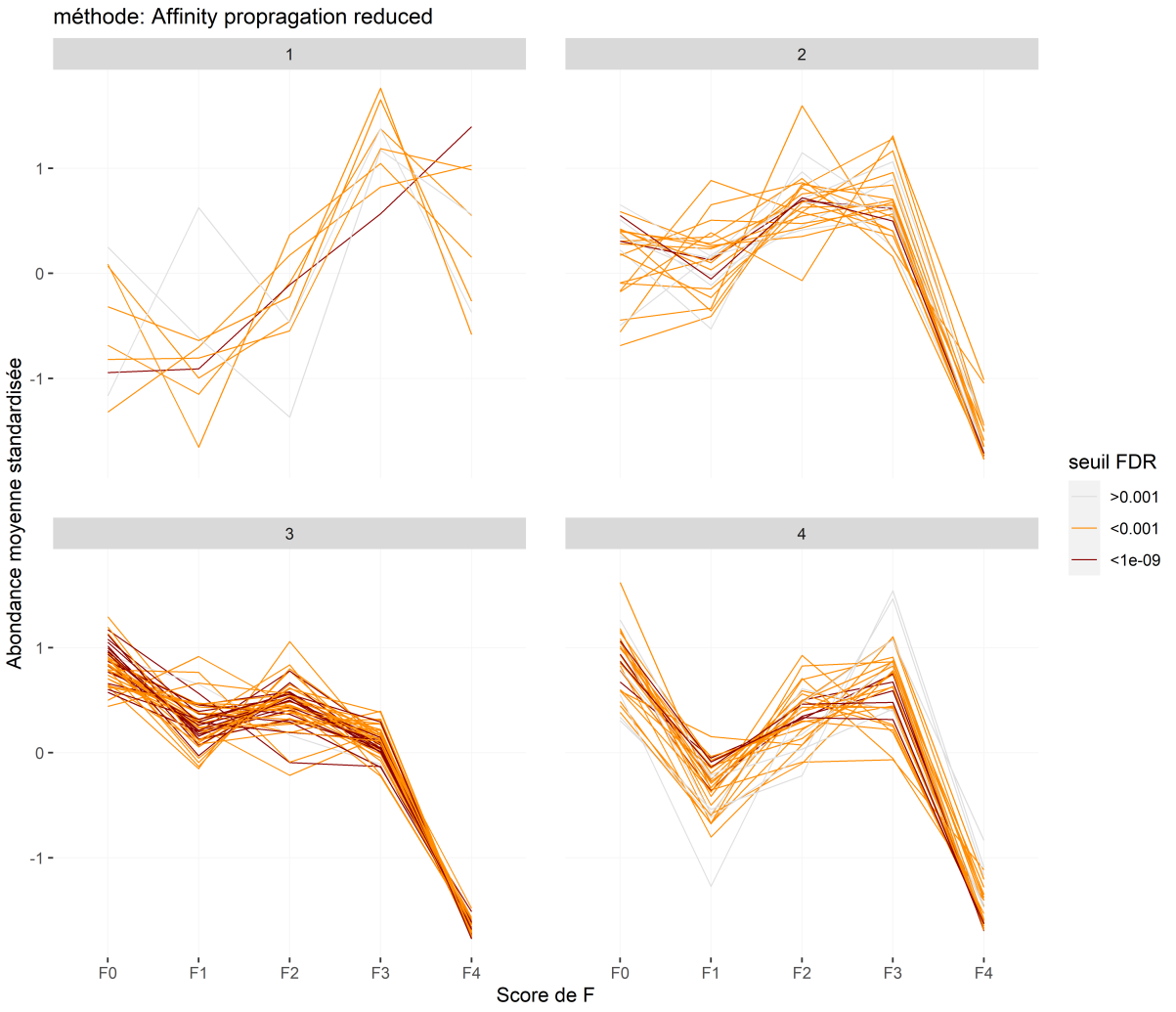



FIGURE 7.10 – Clustering des profils dabondance des 114 protéines statistiquement DA pour un FDR 1%. Chaque ligne dans un cluster représente le profil d'abondance moyen d'une protéine dans chaque condition et est coloré en fonction de la $q_{valeur}$ obtenue à l'ANOVA. Lecture : Le cluster 3 présente des profils qui diminuent entre F0 et F1, puis augmentent entre F1 et F2, diminuent entre F2 et F3 puis qui s'effondrent en F4. Nombre de protéines par cluster : cluster 1 (9), cluster 2 (25), cluster 3 (47), cluster 4 (33).

Une analyse d'enrichissement de « *Gene Ontology* » (GO) [228] a également été réalisée pour les protéines statistiquement différentielles, afin de compléter les résultats statistiques et de clustering. Les analyses d'enrichissement pour les processus biologiques (BP) et les fonctions moléculaires (MF) ont été réalisées à l'aide du système de classification PANTHER [229] sur le génome humain complet. Seuls les résultats ayant obtenu une $q_{valeur} < 0,05$ au test exact de Fisher avec un contrôle du FDR, ont été retenu pour les deux types d'analyses (« *GO biological process complete* » et « *GO molecular function*

*complete* »). Nous avons réalisé ces analyses pour chacun des 4 clusters obtenus, dont les résultats sont détaillés dans les Tableau 7.3 et 7.4. Parmi les 114 protéines statistiquement différentiellement abondantes, 16 sont des protéases, 13 sont des inhibiteurs de protéases, 31 sont impliquées dans l'immunité et la réponse inflammatoire, dont 28 sont impliquées dans le complément et son activation, 3 protéines sont des hormones, 13 sont impliquées dans le métabolisme des lipides, 12 sont impliquées dans la coagulation sanguine, 2 sont impliquées dans le stress oxydatif, et 4 sont impliquées dans l'adhésion cellulaire.

| | GO BP | |
|---|---|---|
| Cluster | Nom | $q_{valeur}$ |
| **1** | Régulation du complément d'activation | $2,28.10^{-3}$ |
| | Régulation de la réponse immunitaire humorale | $2,34.10^{-3}$ |
| **2** | Régulation négative de la clairance des particules de lipoprotéines de très basse densité | $5,05.10^{-3}$ |
| | Régulation négative du processus catabolique des triglycérides | $6,22.10^{-3}$ |
| | Régulation positive de l'activité de la lipoprotéine lipase | $1,26.10^{-2}$ |
| | Activation du complément, voie alternative | $6,08.10^{-4}$ |
| | Transport de lipides | $8,23.10^{-4}$ |
| | Coagulation sanguine, formation de caillots de fibrine | $1,66.10^{-3}$ |
| | Régulation négative de la coagulation sanguine | $7,93.10^{-3}$ |
| **3** | Régulation positive de la fibrinolyse | $1,29.10^{-2}$ |
| | Régulation négative de la coagulation sanguine | $4,42.10^{-8}$ |
| | Activation du complément, voie alternative | $3,68.10^{-7}$ |
| | Régulation positive de la clairance cellulaire apoptotique | $3,54.10^{-2}$ |
| | Fibrinolyse | $1,14.10^{-6}$ |
| | Coagulation sanguine, voie intrinsèque | $3,78.10^{-5}$ |
| | Réponse en phase aiguë | $3,69.10^{-10}$ |
| | Régulation de l'activation du complément | $1,78.10^{-17}$ |
| | Dégranulation plaquettaire | $1,11.10^{-9}$ |
| | Transport de lipides | $2,15.10^{-3}$ |
| | Processus métabolique des lipoprotéines | $3,33.10^{-2}$ |
| | Modification post-traductionnelle des protéines | $4,59.10^{-3}$ |
| **4** | Régulation de l'activité des transporteurs de lipides | $6,10.10^{-3}$ |
| | Régulation positive de la fibrinolyse | $6,04.10^{-3}$ |
| | Régulation négative de la coagulation sanguine | $9,46.10^{-7}$ |
| | Régulation négative du remodelage des particules de lipoprotéines de très basse densité | $5,99.10^{-3}$ |
| | Régulation positive de l'activité glycogène (amidon) synthase | $7,99.10^{-3}$ |
| | Régulation positive de l'estérification du cholestérol | $2,66.10^{-4}$ |
| | Régulation négative de la différenciation des macrophages | $1,20.10^{-2}$ |
| | Fibrinolyse | $1,05.10^{-6}$ |
| | Régulation de l'absorption intestinale du cholestérol | $1,81.10^{-2}$ |
| | Dégranulation plaquettaire | $5,76.10^{-16}$ |

Tableau 7.3 – Résultats de l'analyse d'enrichissement GO à l'aide de PANTHER pour les processus biologiques (« *GO biological process complete* » ; colonne GO BP) pour les protéines statistiquement DA dans chaque cluster.

| | **GO MF** | |
|---|---|---|
| Cluster | Nom | $q_{valeur}$ |
| **1** | Pas de résultats significatifs | |
| **2** | Activité activatrice de la lipoprotéine lipase | $7,61.10^{-3}$ |
| | Liaison à l'héparine | $3,01.10^{-6}$ |
| | Activité d'inhibiteur de l'endopeptidase | $7,91.10^{-5}$ |
| | Liaison aux lipides | $1,66.10^{-2}$ |
| **3** | Liaison au complément | $2,93.10^{-8}$ |
| | Liaison aux hormones thyroïdiennes | $4,85.10^{-2}$ |
| | Activité d'inhibiteur d'endopeptidase de type sérine | $3,86.10^{-9}$ |
| **4** | Liaison aux récepteurs de particules de lipoprotéines de haute densité | $9,83.10^{-3}$ |
| | Activité d'activateur de la phosphatidylcholine-stérol O-acyltransférase | $2,23.10^{-2}$ |
| | Liaison aux récepteurs des apolipoprotéines | $2,10.10^{-2}$ |
| | Activité d'inhibiteur d'endopeptidase de type sérine | $1,55.10^{-5}$ |
| | Liaison au cholestérol | $2,46.10^{-2}$ |
| | Activité endopeptidase de type sérine | $4,66.10^{-4}$ |

TABLEAU 7.4 – Résultats de l'analyse d'enrichissement GO à l'aide de PANTHER pour les fonctions moléculaires (« *GO molecular function complete* » ; colonne GO MF) pour les protéines statistiquement DA dans chaque cluster.

Les résultats de l'analyse GO sont en adéquation avec ce qui était attendu pour l'étude de la fibrose. En effet, ils correspondent à ce qui est connu de la physiopathologie de la NAFLD et de la NASH, ainsi qu'au type d'échantillon analysé (ici du plasma sanguin). Le Tableau 7.3 montre que les clusters sont globalement associés *i)* aux processus biologiques impliqués dans la fluidité sanguine (régulation positive de la fibrinolyse ($q_{valeur} = 1,29.10^{-2}$), régulation négative de la coagulation du sang ($q_{valeur} = 4,42.10^{-8}$) et dégranulation plaquettaire ($q_{valeur} = 5,76.10^{-16}$)) ; *ii)* aux processus impliqués dans le métabolisme des lipides (régulation négative du processus catabolique des triglycérides ($q_{valeur} = 6,22.10^{-3}$), transport de lipides ($q_{valeur} = 8,23.10^{-4}$), régulation positive de l'estérification du cholestérol ($q_{valeur} = 2,66.10^{-4}$), régulation de l'absorption intestinale du cholestérol ($q_{valeur} = 1,81.10^{-2}$)). Ces résultats sont en adéquation avec le fait que la NAFLD/NASH implique le stockage de lipides (la stéatose) et la présence d'un syndrome métabolique, bien que cette dernière information ne soit pas fournie par la base de données cliniques. En outre, les protéines identifiées sont également associées aux processus biologiques impliqués dans la régulation de l'activité lipoprotéique et la réaction inflammatoire et les fonctions métaboliques (cf. Tableau 7.4) sont également cohérentes, avec pour principales fonctions impliquées l'activité activatrice des lipoprotéines ($7,61.10^{-3}$) pour le cluster 2 et la liaison au complément ($2,93.10^{-8}$) pour le cluster 3.

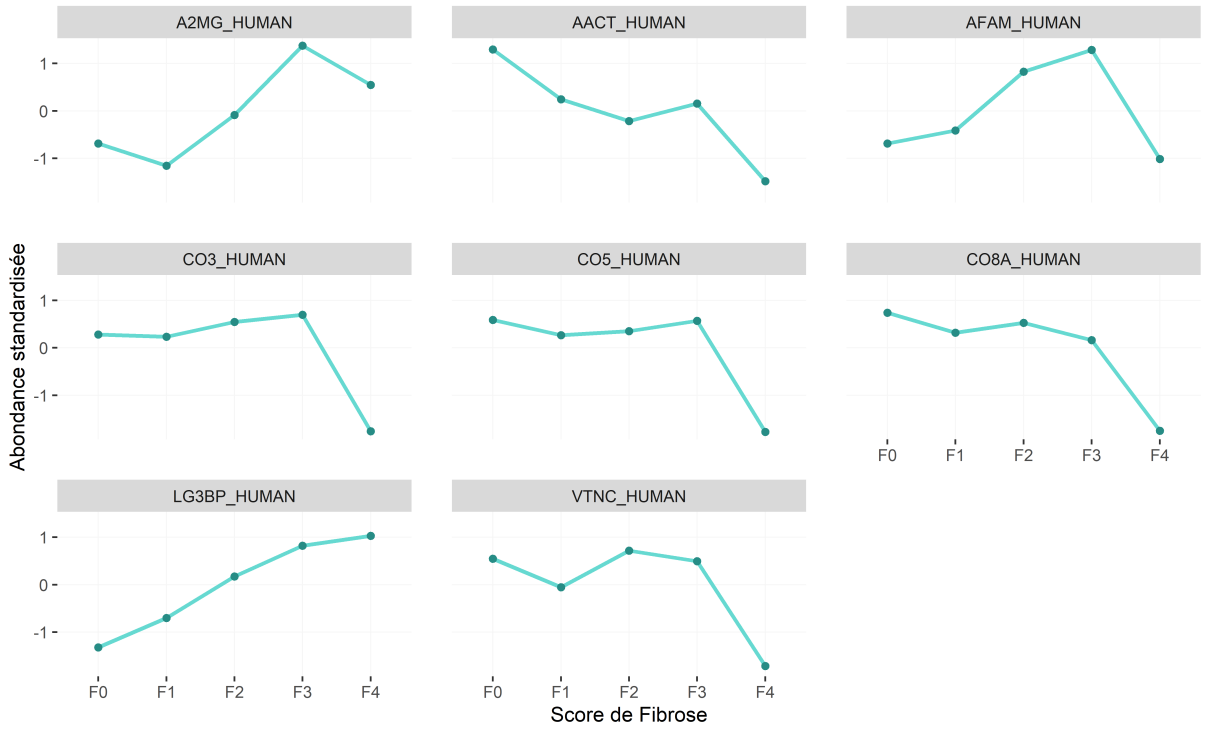



FIGURE 7.11 – Profil d'abondance en fonction du score de F de fibrose pour des protéines identifiées comme biomarqueurs candidats dans les études de Niu *et al.* [230] et Hou *et al.* [231]. On observe que l'abondance diminue fortement au stade F4 par rapport aux autres stades de fibrose, sauf pour LG3BP dont l'abondance augmente constamment depuis F0 jusqu'à F4.

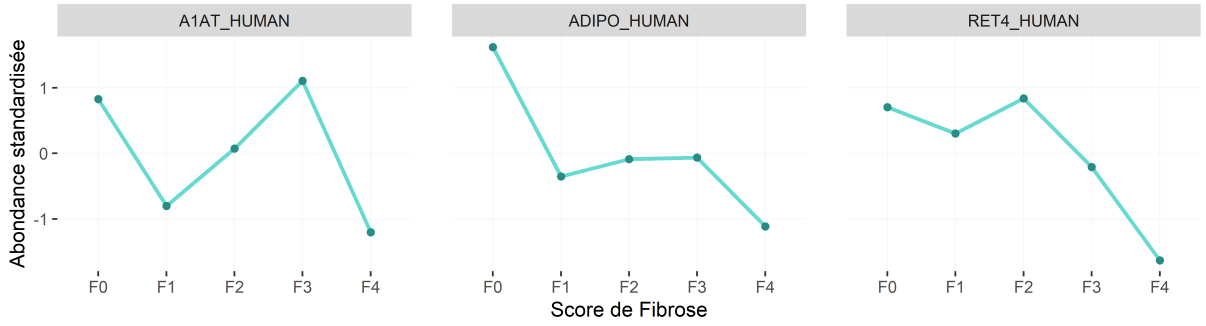



FIGURE 7.12 – Profil d'abondance en fonction du score de F de fibrose pour des biomarqueurs connus [206, 232]. On observe que l'abondance diminue entre le stade F0 et F1. En ce qui concerne A1AT cette diminution d'abondance est aussi observée entre les stades F3 et F4. RET4 présente une légère altération d'abondance sur les trois premiers stades puis s'effondre entre F2 et F4.


FIGURE 7.13 – Profil d'abondance des 13 biomarqueurs candidats identifiés dans l'analyse statistique. Par soucis de confidentialité, les protéines sont dénommées par des lettres.

Certaines des protéines considérées comme différentiellement abondantes à un FDR de1% sont déjà connues de l'état de l'art comme biomarqueurs de panels ou biomarqueurs potentiels, notamment A2MG (cluster 1,$q_{valeur} = 4,18.10^{-4}$) qui est utilisée dans le NASHtest et le STEATOtest. Les protéines LG3BP (cluster 1, $q_{valeur} = 1,48.10^{-7}$), VTNC (cluster 2, $q_{valeur} = 1,49.10^{-10}$) et AFAM (cluster 2, $q_{valeur} = 2,64.10^{-4}$) sont proposées comme biomarqueurs candidats de la NAFLD dans l'étude de Niu et al. [230]. Les protéines A2MG, « *Alpha-1-antichymotrypsin* » (AACT) (cluster 3, $q_{valeur} = 8,09.10^{-7}$), « *Complement component C8 alpha chain* » (CO8A) (cluster 3, $q_{valeur} = 1,82.10^{-10}$), « *Complement C3* » (CO3) (cluster 2, $q_{valeur} = 4,49.10^{-7}$) et « *Complement C5* » (CO5) (cluster 2, $q_{valeur} = 3,13.10^{-5}$) sont proposées comme biomarqueurs candidats de fibrose avancée dans l'étude de Hou et al. [231]. Leurs profils d'abondance respectifs sont présentés en Figure 7.11.

Nous avons également identifié des protéines spécifiques du foie, selon la définition donnée par le *Human Proteome Atlas* [233], autrement dit les protéines dont l'expression est enrichie au niveau du tissu hépatique (plus précisément les termes « *liver enriched* » et « *liver enhanced* » ont été utilisés pour la recherche). Il s'agit notamment de protéines sécrétées par le foie et dont l'abondance dans le plasma dépend de la fonction hépatique. Elles présentent donc un intérêt certain pour caractériser des états hépatiques dégradés comme c'est le cas pour les stades avancés de fibrose (F2, F3 et F4), un exemple qui est montré dans la Figure 7.12 est la protéine RET4 (cluster 3, $q_{valeur} = 1,87.10^{-5}$). Les protéines du cluster 4 présentent une altération de l'abondance marquée dans les stades précoces F0 et F1 (voir Figure 7.10). Dans ce cluster, il y a des biomarqueurs déjà connus tels que l'Adiponectine (ADIPO) ($q_{valeur} = 8,08.10^{-5}$) et l'« *Alpha-1-antitrypsin* » (A1AT) ($q_{valeur} = 1,90.10^{-5}$), dont les profils d'abondance sont présentés dans la Figure 7.12. L'adiponectine est une protéine impliquée dans la régulation du métabolisme des lipides qui est lié à l'obésité et au syndrome métabolique, ainsi que dans la sensibilité à l'insuline, phénomène lié au diabète. A1AT est impliquée dans la coagulation sanguine. Ces protéines sont déjà étudiées comme biomarqueurs candidats mais elles nécessitent des validations cliniques [206, 232].

Finalement, bien que des protéines caractérisées précédemment telles que K1C18 et TNR6 n'aient pas été détectées dans notre analyse protéomique, l'ensemble des résultats de cette analyse statistique confortent à grande échelle les conclusions des recherches plus focalisées menées dans l'étude de la NAFLD. En effet, concernant la K1C18 qui est une kératine, il est probable qu'elle ait été filtrée lors des étapes bioinformatiques. Notre analyse a

également permis de mettre en évidence 13 nouveaux biomarqueurs candidats dont les profils d'expression sont présentés dans la Figure 7.13, avec une convention de nommage arbitraire, afin d'en préserver la confidentialité (car possiblement brevetable). Une thèse spécifiquement consacrée à l'étude de ces candidats vient de débuter dans notre laboratoire, dans le but de valider certaines de ces protéines comme des biomarqueurs candidats de la fibrose dans le cas de la NAFLD. Ces 13 protéines ont été retenues notamment car elles présentaient soit une forte augmentation soit une forte diminution d'abondance entre au moins deux scores de fibrose. Dans la Figure 7.13, nous pouvons voir que la majorité sont des protéines pour lesquelles l'abondance est fortement diminuée entre les scores F3 et F4. La protéine E est intéressante car elle augmente constamment entre les différents scores et est constante entre les scores F0 et F1, elle pourrait permettre de bien dissocier stades précoces et tardifs. Les protéines F et K sont également intéressantes car elles présentent une forte augmentation d'abondance entre les scores F2 et F3 et pourraient permettre de marquer la transition entre stade moyen (F2) et stades avancés (F3 et F4). La protéine M est également intéressante car elle est constante sur les quatre premiers stades et son abondance s'effondre au dernier stade (F4) ; elle montre donc une réelle rupture entre le stade tardif F4 et les autres stades.

### 7.3.3 Limites de l'étude

La première limitation de cette étude est la puissance statistique potentiellement insuffisante. En effet, elle est impactée par le déséquilibre du plan d'expérience, comme cela a été souligné précédemment, ainsi que par le nombre relativement faible d'échantillons pour chaque condition, l'ampleur de l'effet biologique observable sur les protéines (difficile à évaluer), la variabilité inter-individuelle et l'impact des facteurs confondants (notamment diabète et obésité) qui n'ont pas pu être formellement pris en compte par manque de temps. La seconde limite est l'absence d'un groupe contrôle (état hépatique normal ou patients sains) clairement établi. La comparaison réalisée dans notre analyse ne permet pas de se prononcer vis-à-vis d'un état physiologique hépatique normal, alors qu'il aurait été intéressant de voir si l'on identifiait des protéines dont le profil se distingue bien entre état sain et les différents stades de fibrose.

Nous avons pu constater grâce à la visualisation des clusters et des profils d'abondances (Figures 7.10, 7.11 et 7.12) qu'il existe une différence d'abondance marquée entre le stade F4 et les autres stades de fibrose pour de nombreuses protéines. Ce phénomène peut expliquer en partie la proportion élevée (47%) de protéines trouvées comme statistique-

ment différentiellement abondantes avec l'ANOVA à 1 facteur. Des profils d'abondance qui l'illustrent sont ceux des protéines CO3, CO5 et CO8A (Figure 7.11), qui présentent une abondance relativement stable entre les stades F0, F1, F2 et F3 et qui diminue fortement au stade F4. Il y a également une explication biologique à cette différence marquée entre F4 et les autres stades de fibrose : le stade F4 correspondant à un état cirrhotique, la sécrétion de protéines plasmatiques par le foie est diminuée. Il serait intéressant de ne comparer que les quatre premiers stades entre eux, afin de découvrir si le nombre de protéines différentiellement abondantes serait aussi important que lorsque l'on intègre le stade F4 ; bien qu'en tant que tel, cela réponde à une tout autre question biologique. Nous aurions également pu nous attendre à ce que se présente le cas de figure où des protéines, spécifiques du foie et intracellulaires (comme l'ALAT1) en temps normal, soient relarguées dans la circulation sanguine lors de la mort cellulaire des hépatocytes, comme en cas de cirrhose ou de cancer. Nous aurions alors pu observer des protéines spécifiques du foie ayant une abondance accrue en F4 par rapport aux autres stades, mais ce n'est pas le cas suggérant que la mort cellulaire des hépatocytes reste limitée à ces stades. Les protéines spécifiques du foie, au sens du Human Proteome Atlas, qui ressortent comme statistiquement différentielles, présentent au contraire une abondance fortement diminuée entre F3 et F4. Les autres protéines connues comme spécifiques, étaient absentes dans le jeu de données fourni pour l'analyse. Cela peut être dû à tout simplement leur absence totale des échantillons, ou bien aux critères de sélection des protéines lors des étapes d'identification et de quantification qui étaient éventuellement trop stringentes. Alternativement, ces protéines sont en concentrations trop faibles dans le plasma pour être détectées par notre approche protéomique.

Enfin, notre plan expérimental et notre analyse statistique simplifient dans une certaine mesure la question biologique à laquelle les cliniciens du projet souhaitent initialement répondre. En effet, nous ne prenons pas en compte l'ensemble de la complexité de la pathologie (nous n'avons étudié que le stade de fibrose), en sachant que la pathologie en elle-même nécessite plus d'investigation pour mieux la caractériser.

## 7.4 Analyses complémentaires

L'analyse présentée dans la Section 7.3 a servi de preuve de concept pour proposer un cheminement méthodologique applicable aux autres facteurs à étudier, à savoir le score de Stéatose (S) et le score d'Activité (A) pour des données DDA. Ainsi, la méthodologie

décrite et justifiée dans la Section 7.3, ainsi que les outils décrits aux Chapitres 5 et 6 permettront de poursuivre le projet malgré la fin de ma thèse. L'application de cette méthodologie est possible car les motifs qui ont prévalus à certains choix pour l'analyse du score F sur des données DDA restent valables pour les autres scores. En effet, il s'agit du même ordre de grandeur de protéines quantifiées. En outre, la proportion de valeurs manquantes entre conditions et au sein d'une même condition est similaire. Par conséquent les mêmes méthodes de filtrage et d'imputation peuvent être appliquées pour l'analyse des scores S et A. Par ailleurs, la distribution des quantifications est similaire à celle des quantifications pour le score F, de sorte que la normalisation est également nécessaire. Enfin, le plan expérimental est également déséquilibré (S0 : 10, S1 :32, S2 :54, S3 :63 et A0 :19, A1 :31, A2 :41, A3 :54, A4 :14), de sorte que toute l'approche de vérification des conditions d'application de l'ANOVA doit également être réalisée. Cependant, cette méthodologie n'est valide que pour des données quantitatives obtenues par AC (cf. Section 7.3.1). En effet, ne pas réaliser d'AC impliquerait nécessairement une proportion de valeurs manquantes différente, et donc toutes les étapes de pré-traitement s'en trouveraient impactées. Quelques tests supplémentaires seraient donc nécessaires pour valider une chaîne de traitement statistique sur ces données. Concernant l'analyse des scores S, A et F pour les données en DIA, il est nécessaire d'évaluer la pertinence de l'approche que je propose dans ce travail de thèse et d'éventuellement l'adapter ou bien développer une méthodologie différente.

Enfin, l'ensemble des résultats de cette analyse statistique a été le point de départ d'exploration pour une collaboration réalisée avec la start-up Signet ainsi que pour la thèse de protéomique précédemment mentionnée, qui a débuté au laboratoire l'an dernier et qui s'intéresse aux 13 protéines anonymisées sur la Figure 7.13. Celles-ci feront l'objet de publications ultérieures avec potentiellement, la proposition d'un biomarqueur composite permettant de discriminer les différents scores de fibrose.

# Chapitre 8

# Conclusion et perspectives

Pour rappel, le travail de cette thèse a été guidé par deux objectifs principaux. Le premier était de mener une réflexion sur l'analyse statistique telle qu'elle est réalisée actuellement en protéomique quantitative afin d'en améliorer la robustesse et la reproductibilité dans un contexte de projets cliniques. Cette réflexion devait aboutir à une méthodologie à la fois adaptée aux données de protéomique quantitative *label-free*, aux plans d'expériences cliniques complexes, et être intégrée dans un outil logiciel permettant une utilisation facilitée pour les expérimentateurs. Le second objectif était de valider ce développement sur des données réelles issues d'une cohorte clinique (la cohorte NAFLD), afin de participer à l'identification de biomarqueurs candidats.

Afin de répondre à ces objectifs, plusieurs contributions ont été proposées. Tout d'abord, la librairie R *wpm*, a été réalisée afin de traiter la problématique des effets de lots en amont de l'expérimentation, afin d'anticiper des biais évitables et ainsi minimiser leur impact sur la suite de l'analyse. Le logiciel développé s'appuie largement sur des algorithmes stochastiques et constitue une alternative à ce qui est proposé dans la littérature. Notamment, il permet d'obtenir des plans de plaque d'aussi bonne qualité et de manière plus rapide que ceux proposés par d'autres auteurs, tout en proposant des options de placement plus diversifiées. Le logiciel *wpm* est à la disposition de la communauté sur le dépôt Bioconductor. Par ailleurs, la réflexion portée sur la pertinence de l'étape de normalisation (afin de réduire les effets de lots restant en aval de l'analyse) nous a amené à utiliser et à proposer le calibration-plot comme un outil utile dans l'évaluation des performances. Cette manière d'évaluer la pertinence d'une normalisation est à la fois différente et complémentaire de celles habituellement utilisées dans la littérature de protéomique.

Ce travail a conduit à l'ajout de deux méthodes de normalisation au logiciel *Prostar*. Nos observations confirment les arguments avancés par certains auteurs, à savoir qu'il est nécessaire, voire primordial, de prendre le temps de comparer plusieurs méthodes de normalisation afin d'identifier celle qui est la plus adaptée aux données étudiées. En effet, nous avons vu dans le Chapitre 5 que les procédures usuellement utilisées pour des données de protéomique *label-free*, se sont avérées inadaptées à nos données, notamment en raison de l'importante variabilité biologique inter-échantillon. Comme cela a été évoqué dans la Section 5.2.2, des points d'amélioration pourraient être apportés à *wpm* : modifier l'algorithme actuel en une version récursive plus proche du véritable backtracking ; ajouter une option de sérialisation des échantillons pour l'injection dans un spectromètre de masse ; et faciliter le formatage des facteurs confondants lors de la saisie des paramètres. De même, il aurait été pertinent de comparer notre stratégie d'évaluation des différentes normalisations à celles déjà utilisées classiquement, telles que notamment, le calcul du pourcentage de variance totale réduite avant et après application. En effet, il est souvent proposé d'adopter la méthode de normalisation qui obtient la plus grande diminution de variance totale, alors que cela peut se faire au détriment de la variabilité biologique réelle : A trop réduire la variance totale, on peut être amené à perdre de l'information pertinente d'un point de vue biologique.

En ce qui concerne notre réflexion sur la problématique du contrôle du FDR dans le cas de comparaison multiples (développée et discutée au Chapitre 6), nous avons expliqué pourquoi l'interprétation des résultats statistiques finaux n'est pas si aisée en raison de la superposition de deux types de correction de multiplicité des tests (entre protéines et entre conditions). En supposant qu'il n'est pas indispensable de réaliser systématiquement les deux, nous avons proposé de les prioriser, ce qui nous a amenés à discuter et comparer plusieurs scénarios pratiques. Nous pouvons conclure de ces comparaisons qu'aucun des scénarios classiquement utilisés dans la littérature ne permet de concilier l'ensemble des rôles du contrôle du FDR (contrôle statistique, rôle normatif et stringence expérimentale) avec la question biologique initiale. Il apparaît donc nécessaire de déterminer quel(s) sont les rôle(s) du FDR que l'on souhaite préserver en fonction de la question biologique d'intérêt, afin d'employer le scénario le plus adapté. Il s'avère que la méthode la plus communément appliquée (décrite à la Section 6.3.1.1) n'est pas adaptée à une vision globale des conditions biologiques étudiées, comme on se propose de le faire pour l'étude de la cohorte NAFLD. C'est notamment cela qui nous a amené à défendre un scénario différent (décrit à la Section 6.3.1.3 et consistant à combiner les résultats d'une MTC et ceux d'un clustering pour l'interprétation biologique finale). Concernant l'implémentation proposée,

celle-ci se restreint au cas de l'ANOVA à un facteur. Elle ne permet pas actuellement une généralisation à des modèles plus complexes comme l'ANOVA à plus d'un facteur ou encore l'analyse de covariance. Néanmoins, de tels développements futurs auraient un intérêt, puisque nous avons vu au Chapitre 7 que dans le cadre d'études cliniques, la pathologie est le plus souvent associée à des facteurs confondants avérés, qu'il serait pertinent de prendre en considération dans l'analyse statistique.

Ensuite, nous avons présenté dans le Chapitre 7 la mise en application concrète de la méthodologie développée aux Chapitres 5 et 6 sur des données cliniques afin d'identifier des biomarqueurs candidats. L'exploration des résultats a permis l'identification de 114 protéines statistiquement différentiellement abondantes avec quatre clusters de profils d'abondance. L'analyse d'enrichissement GO a permis de mettre en évidence que ces protéines sont impliquées dans des processus biologiques correspondants à ceux déjà relatés pour la NAFLD, à savoir le métabolisme des lipides, la coagulation sanguine et l'inflammation. Treize nouveaux biomarqueurs candidats ont été mis en évidence mais ont été anonymisés dans ce manuscrit, afin de préserver la confidentialité des travaux de thèse qui poursuivent mes investigations. Par ailleurs, parmi les protéines remarquées comme différentiellement abondantes, nous retrouvons celles déjà proposées dans la littérature et les tests pré-existants. Ces résultats positifs confortent également les biomarqueurs candidats proposés dans d'autres études de découverte (cf. Chapitre 7). Les principales limites de ce travail ont été discutées dans le Chapitre 7. Elles concernent la puissance statistique de l'étude et la simplification du plan expérimental vis-à-vis de la complexité de la NAFLD. Il serait par conséquent intéressant de poursuivre l'analyse statistique pour les facteurs d'Activité et de Stéatose, de confronter les différents résultats des trois scores S, A et F, tout en réalisant par exemple des analyses de covariance qui tiennent compte de facteurs confondants importants tels que l'obésité et le diabète.

Pour résumer, l'essentiel du travail réalisé pendant cette thèse est schématisé dans la Figure 8.1.

Plus globalement, ce travail de thèse s'est focalisé sur l'analyse statistique à l'échelle des protéines, mais progressivement, la communauté de protéomique accepte de regarder les données à un niveau plus fin (celui des peptides, c'est-à-dire, celui correspondant à la mesure instrumentale, plutôt qu'à celui correspondant à l'entité biologiquement pertinente). Il pourrait donc être profitable de transposer la méthodologie statistique proposée à cette échelle. De plus, l'étude de la cohorte NAFLD présentée dans ce manuscrit s'est limitée à identifier des protéines différentiellement abondantes indépendamment les unes

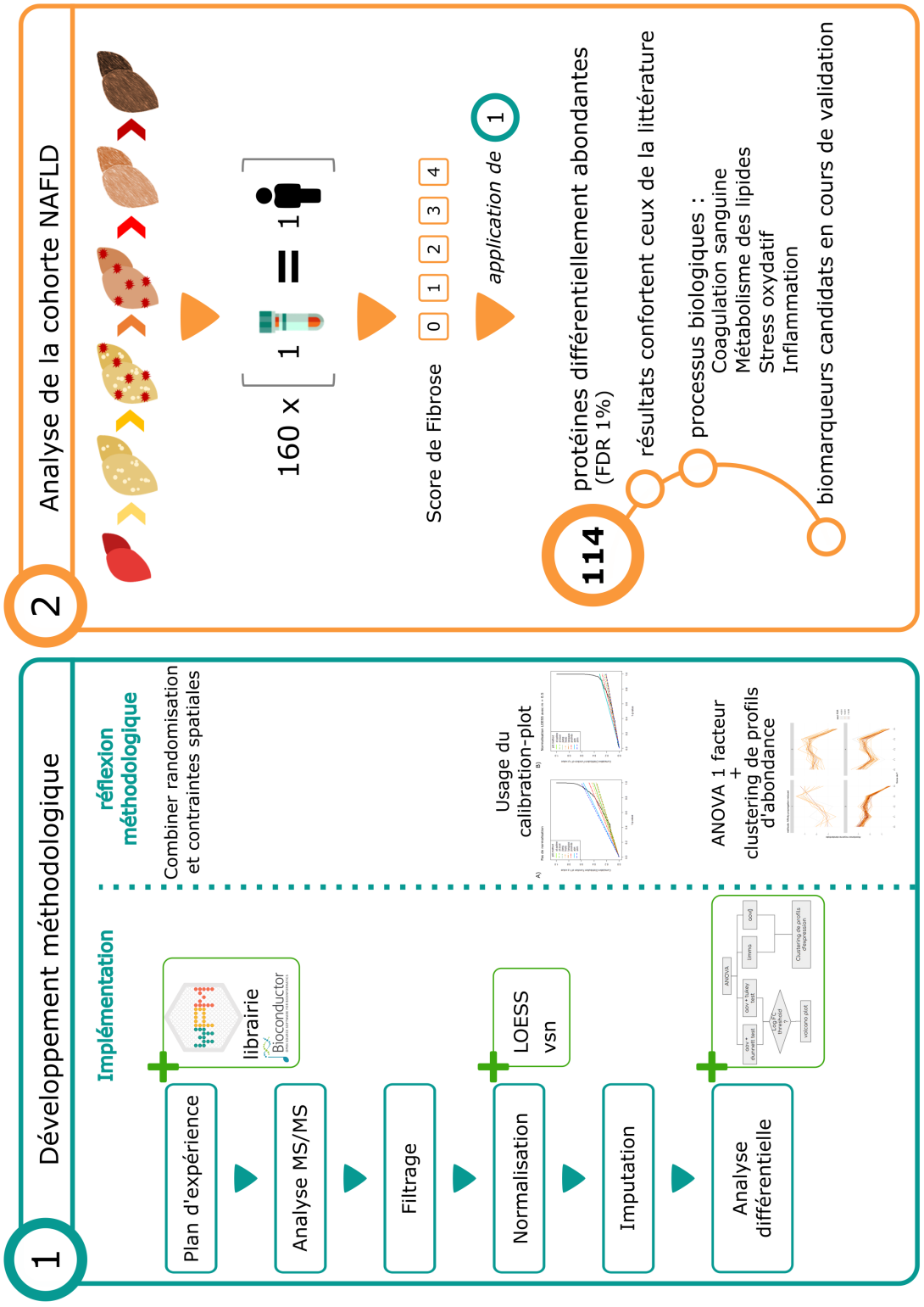



FIGURE 8.1 – Schéma récapitulatif des contributions de la thèse sur le plan méthodologique (volet 1) et applicatif (volet 2).

des autres. Mais il serait intéressant d'identifier des groupes de protéines, comme ce qui est fait par exemple dans les travaux de Enroth *et al.* [234], qui pris collectivement permettent à la fois une meilleure sensibilité et une meilleure spécificité que les biomarqueurs constitués d'une unique protéine, aussi pertinente soit-elle. Il serait également intéressant de mettre en place des modèles prédictifs de score de SAF par apprentissage supervisé sur des protéines quantifiées et identifiées dans des échantillons plasmatiques associés à un score SAF connu. En théorie, les conclusions de tels modèles seraient beaucoup plus proches de ce qu'attendent les médecins, car se baser sur les abondances de protéines pour déterminer le score de fibrose ou d'activité d'un patient permettrait *in fine* de se passer de la biopsie pour un diagnostic tout aussi fiable. Néanmoins, l'apprentissage supervisé ne devrait être utilisé que comme complément à d'autres techniques de diagnostic car il peut poser des problèmes d'interprétabilité. En effet, un diagnostic basé uniquement sur de l'apprentissage supervisé n'est pas exempt d'un risque d'erreur, qui peut être difficile à évaluer. Cette difficulté est principalement due à l'aspect « boîte noire » des algorithmes, c'est-à-dire qu'il est difficile d'expliquer comment ces algorithmes sont parvenus à leur conclusions. De plus, ces algorithmes peuvent être biaisés par un sur-apprentissage, c'est-à-dire qu'ils fourniront d'excellentes performances sur les données de test mais qui ne pourront être généralisées à de nouvelles données. En dépit du grand intérêt de ces approches, elles ne sauraient donc se substituer au cheminement plus classique qui a été suivi durant ce travail.

Finalement, même si mon travail reste dans la limite de la phase de découverte (cf. Chapitre 1), il vise à rendre l'identification de biomarqueurs candidats toujours plus fiable. Malheureusement, il s'avère que de nombreux biomarqueurs candidats identifiés en phase exploratoire (protéomique de découverte), ne présentent pas de résultats concluants lors des phases suivantes en particulier à l'étape de validation [27, 235]. Ceci conduit certains auteurs à remettre en question le modèle actuel de recherche de biomarqueurs (cf. Section 1.3.1). Pour répondre à cet échec lors de la phase de validation, un nombre croissant d'études montrent l'intérêt des signatures protéiques complexes que nous avons déjà évoqué dans le paragraphe précédent [236, 237, 238]. En plus de se limiter à la phase de découverte, mon travail reste dans un cadre de recherche car il semble difficile de transposer la MS à une application clinique quotidienne. Même si le biomarqueur est découvert et validé *via* la MS, celui-ci ne sera pas exploité en clinique grâce à cette dernière, car il s'agit d'une technologie extrêmement coûteuse demandant une grande expertise. En milieu hospitalier les biomarqueurs seront analysés par d'autres méthodes, si bien qu'ils n'auront pas la même sensibilité et spécificité pour le diagnostic de la pathologie ou le suivi des

patients. Un défi consiste donc à réussir cette transition entre la MS et la technologie qui est utilisée quotidiennement, tout en s'assurant que le biomarqueur reste efficace dans son rôle de pronostic, de diagnostic ou de suivi.

# Bibliographie


[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome," *science*, vol. 291, no. 5507, pp. 1304–1351, 2001. (Cité en page 1.)

[2] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003. (Cité en page 1.)

[3] T. Shu, W. Ning, D. Wu, J. Xu, Q. Han, M. Huang, X. Zou, Q. Yang, Y. Yuan, Y. Bie, *et al.*, "Plasma proteomics identify biomarkers and pathogenesis of covid-19," *Immunity*, vol. 53, no. 5, pp. 1108–1122, 2020. (Cité en page 2.)

[4] S. Srivastava, M. Merchant, A. Rai, and S. N. Rai, "Standardizing proteomics workflow for liquid chromatography-mass spectrometry : technical and statistical considerations," *Journal of proteomics & bioinformatics*, vol. 12, no. 3, p. 48, 2019. (Cité en page 2.)

[5] P. E. Geyer, L. M. Holdt, D. Teupser, and M. Mann, "Revisiting biomarker discovery by plasma proteomics," *Molecular systems biology*, vol. 13, no. 9, p. 942, 2017. (Cité en pages 2, 11, 12, 13 et 14.)

[6] M. R. Wilkins, J.-C. Sanchez, A. A. Gooley, R. D. Appel, I. Humphery-Smith, D. F. Hochstrasser, and K. L. Williams, "Progress with proteome projects : why all proteins expressed by a genome should be identified and how to do it," *Biotechnology and genetic engineering reviews*, vol. 13, no. 1, pp. 19–50, 1996. (Cité en pages 4 and 7.)

[7] A. Ambrogelly, S. Palioura, and D. Söll, "Natural expansion of the genetic code," *Nature chemical biology*, vol. 3, no. 1, pp. 29–35, 2007. (Cité en page 5.)

[8] A. Gutteridge and J. M. Thornton, "Understanding nature's catalytic toolkit," *Trends in biochemical sciences*, vol. 30, no. 11, pp. 622–629, 2005. (Cité en page 7.)

[9] P. C. Havugimana, G. T. Hart, T. Nepusz, H. Yang, A. L. Turinsky, Z. Li, P. I. Wang, D. R. Boutz, V. Fong, S. Phanse, *et al.*, "A census of human soluble protein complexes," *Cell*, vol. 150, no. 5, pp. 1068–1081, 2012. (Cité en page 7.)

[10] O. Gruzdeva, D. Borodkina, E. Uchasova, Y. Dyleva, and O. Barbarash, "Leptin resistance : underlying mechanisms and diagnosis," *Diabetes, metabolic syndrome and obesity : targets and therapy*, vol. 12, p. 191, 2019. (Cité en page 7.)

[11] D. M. Rosenbaum, S. G. Rasmussen, and B. K. Kobilka, "The structure and function of g-protein-coupled receptors," *Nature*, vol. 459, no. 7245, pp. 356–363, 2009. (Cité en page 7.)

[12] D. V. Fyodorov, B.-R. Zhou, A. I. Skoultchi, and Y. Bai, "Emerging roles of linker histones in regulating chromatin structure and function," *Nature reviews Molecular cell biology*, vol. 19, no. 3, p. 192, 2018. (Cité en page 7.)

[13] P. Kahn, "From genome to proteome : looking at a cell's proteins," 1995. (Cité en page 7.)

[14] M. R. Wilkins, C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J.-C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, *et al.*, "From proteins to proteomes : large scale protein identification by two-dimensional electrophoresis and arnino acid analysis," *Bio/technology*, vol. 14, no. 1, pp. 61–65, 1996. (Cité en page 7.)

[15] S.-E. Ong and M. Mann, "Mass spectrometry–based proteomics turns quantitative," *Nature chemical biology*, vol. 1, no. 5, pp. 252–262, 2005. (Cité en pages 7 and 8.)

[16] B. Cañas, D. López-Ferrer, A. Ramos-Fernández, E. Camafeita, and E. Calvo, "Mass spectrometry technologies for proteomics," *Briefings in Functional Genomics*, vol. 4, no. 4, pp. 295–320, 2006. (Cité en page 7.)

[17] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray ionization for mass spectrometry of large biomolecules," *Science*, vol. 246, no. 4926, pp. 64–71, 1989. (Cité en pages 7 and 20.)

[18] H. Thiellement, N. Bahrman, C. Damerval, C. Plomion, M. Rossignol, V. Santoni, D. De Vienne, and M. Zivy, "Proteomics for genetic and physiological studies in plants," *ELECTROPHORESIS : An International Journal*, vol. 20, no. 10, pp. 2013–2026, 1999. (Cité en page 7.)

[19] M. Bantscheff, S. Lemeer, M. M. Savitski, and B. Kuster, "Quantitative mass spectrometry in proteomics : critical review update from 2007 to the present," *Analytical and bioanalytical chemistry*, vol. 404, no. 4, pp. 939–965, 2012. (Cité en pages 8 and 28.)

[20] O. T. Schubert, H. L. Röst, B. C. Collins, G. Rosenberger, and R. Aebersold, "Quantitative proteomics : challenges and opportunities in basic and applied research," *Nature protocols*, vol. 12, no. 7, pp. 1289–1294, 2017. (Cité en pages 8, 15 et 16.)

[21] A. P. Diz, M. MARTÍNEZ-FERNÁNDEZ, and E. ROLÁN-ALVAREZ, "Proteomics in evolutionary ecology : linking the genotype with the phenotype," *Molecular ecology*, vol. 21, no. 5, pp. 1060–1080, 2012. (Cité en page 9.)

[22] G. A. Khoury, R. C. Baliban, and C. A. Floudas, "Proteome-wide post-translational modification statistics : frequency analysis and curation of the swiss-prot database," *Scientific reports*, vol. 1, no. 1, pp. 1–5, 2011. (Cité en page 9.)

[23] Y. Zhang, B. R. Fonslow, B. Shan, M.-C. Baek, and J. R. Yates III, "Protein analysis by shotgun/bottom-up proteomics," *Chemical reviews*, vol. 113, no. 4, pp. 2343–2394, 2013. (Cité en pages 9, 28 et 34.)

[24] K. Strimbu and J. A. Tavel, "What are biomarkers ?," *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010. (Cité en page 10.)

[25] F.-N. B. W. Group *et al.*, "Best (biomarkers, endpoints, and other tools) resource," 2016. (Cité en page 10.)

[26] R. M. Califf, "Biomarker definitions and their applications," *Experimental Biology and Medicine*, vol. 243, no. 3, pp. 213–221, 2018. (Cité en page 10.)

[27] C. A. Sobsey, S. Ibrahim, V. R. Richard, V. Gaspar, G. Mitsa, V. Lacasse, R. P. Zahedi, G. Batist, and C. H. Borchers, "Targeted and untargeted proteomics approaches in biomarker development," *Proteomics*, vol. 20, no. 9, p. 1900029, 2020. (Cité en pages 10 and 141.)

[28] T. V. Pham, S. R. Piersma, G. Oudgenoeg, and C. R. Jimenez, "Label-free mass spectrometry-based proteomics for biomarker discovery and validation," *Expert review of molecular diagnostics*, vol. 12, no. 4, pp. 343–359, 2012. (Cité en page 11.)

[29] M. Plebani, "Proteomics : the next revolution in laboratory medicine ?," *Clinica chimica acta*, vol. 357, no. 2, pp. 113–122, 2005. (Cité en page 12.)

[30] V. Ignjatovic, P. E. Geyer, K. K. Palaniappan, J. E. Chaaban, G. S. Omenn, M. S. Baker, E. W. Deutsch, and J. M. Schwenk, "Mass spectrometry-based plasma proteomics : considerations from sample collection to achieving translational data," *Journal of proteome research*, vol. 18, no. 12, pp. 4085–4097, 2019. (Cité en page 13.)

[31] P. L. Horvatovich and R. Bischoff, "Current technological challenges in biomarker discovery and validation," *European Journal of Mass Spectrometry*, vol. 16, no. 1, pp. 101–121, 2010. (Cité en page 13.)

[32] G. L. Hortin, S. A. Jortani, J. C. Ritchie Jr, R. Valdes Jr, and D. W. Chan, "Proteomics : a new diagnostic frontier," *Clinical chemistry*, vol. 52, no. 7, pp. 1218–1222, 2006. (Cité en page 13.)

[33] Z. R. Gregorich and Y. Ge, "Top-down proteomics in health and disease : Challenges and opportunities," *Proteomics*, vol. 14, no. 10, pp. 1195–1210, 2014. (Cité en page 16.)

[34] I. Eidhammer, K. Flikka, L. Martens, and S.-O. Mikalsen, *Computational methods for mass spectrometry proteomics*. Wiley Online Library, 2007. (Cité en page 18.)

[35] R. Matthiesen *et al.*, *Mass spectrometry data analysis in proteomics*, vol. 367. Springer, 2007. (Cité en page 18.)

[36] K. Dreisewerd, "The desorption process in maldi," *Chemical reviews*, vol. 103, no. 2, pp. 395–426, 2003. (Cité en page 20.)

[37] Q. Hu, R. J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks, "The orbitrap : a new mass spectrometer," *Journal of mass spectrometry*, vol. 40, no. 4, pp. 430–443, 2005. (Cité en page 22.)

[38] G. Zhang, R. S. Annan, S. A. Carr, and T. A. Neubert, "Overview of peptide and protein analysis by mass spectrometry," *Current protocols in protein science*, vol. 62, no. 1, pp. 16–1, 2010. (Cité en pages 22, 23 et 27.)

[39] J. R. Wiśniewski, A. Zougman, N. Nagaraj, and M. Mann, "Universal sample preparation method for proteome analysis," *Nature methods*, vol. 6, no. 5, pp. 359–362, 2009. (Cité en page 23.)

[40] J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, "Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra," *Nature methods*, vol. 1, no. 1, pp. 39–45, 2004. (Cité en page 26.)

[41] F. Zhang, W. Ge, G. Ruan, X. Cai, and T. Guo, "Data-independent acquisition mass spectrometry-based proteomics and software tools : A glimpse in 2020," *Proteomics*, vol. 20, no. 17-18, p. 1900276, 2020. (Cité en page 26.)

[42] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the ms/ms spectra generated by data-independent acquisition : a new concept for consistent and accurate proteome analysis," *Molecular & Cellular Proteomics*, vol. 11, no. 6, pp. O111–016717, 2012. (Cité en page 26.)

[43] S. J. Geromanos, C. Hughes, S. Ciavarini, J. P. Vissers, and J. I. Langridge, "Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples," *Analytical and bioanalytical chemistry*, vol. 404, no. 4, pp. 1127–1139, 2012. (Cité en page 26.)

[44] J. D. Egertson, A. Kuehn, G. E. Merrihew, N. W. Bateman, B. X. MacLean, Y. S. Ting, J. D. Canterbury, D. M. Marsh, M. Kellmann, V. Zabrouskov, *et al.*, "Multiplexed ms/ms for improved data-independent acquisition," *Nature methods*, vol. 10, no. 8, pp. 744–746, 2013. (Cité en page 26.)

[45] C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras, and A. I. Nesvizhskii, "Dia-umpire : comprehensive computational framework for data-independent acquisition proteomics," *Nature methods*, vol. 12, no. 3, pp. 258–264, 2015. (Cité en page 26.)

[46] J. Wang, M. Tucholska, J. D. Knight, J.-P. Lambert, S. Tate, B. Larsen, A.-C. Gingras, and N. Bandeira, "Msplit-dia : sensitive peptide identification for data-independent acquisition," *Nature methods*, vol. 12, no. 12, pp. 1106–1108, 2015. (Cité en page 26.)

[47] O. Permiakova, R. Guibert, A. Kraut, T. Fortin, A.-M. Hesse, and T. Burger, "Chickn : extraction of peptide chromatographic elution profiles from large scale mass spectrometry data by means of wasserstein compressive hierarchical cluster analysis," *BMC bioinformatics*, vol. 22, no. 1, pp. 1–30, 2021. (Cité en page 27.)

[48] J. Seidler, N. Zinn, M. E. Boehm, and W. D. Lehmann, "De novo sequencing of peptides by ms/ms," *Proteomics*, vol. 10, no. 4, pp. 634–649, 2010. (Cité en page 27.)

[49] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *Journal of proteomics*, vol. 73, no. 11, pp. 2092–2123, 2010. (Cité en page 27.)

[50] T. U. Consortium, "Uniprot : The universal protein knowledgebase in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D480–D489, 2021. (Cité en page 27.)

[51] D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *ELECTROPHORESIS : An International Journal*, vol. 20, no. 18, pp. 3551–3567, 1999. (Cité en page 27.)

[52] A. J. Link, J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik, and J. R. Yates, "Direct analysis of protein complexes using mass spectrometry," *Nature biotechnology*, vol. 17, no. 7, pp. 676–682, 1999. (Cité en page 27.)

[53] R. Craig and R. C. Beavis, "Tandem : matching proteins with tandem mass spectra," *Bioinformatics*, vol. 20, no. 9, pp. 1466–1467, 2004. (Cité en page 27.)

[54] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant, "Open mass spectrometry search algorithm," *Journal of proteome research*, vol. 3, no. 5, pp. 958–964, 2004. (Cité en page 27.)

[55] R. G. Sadygov, D. Cociorva, and J. R. Yates, "Large-scale database searching using tandem mass spectra : looking up the answer in the back of the book," *Nature methods*, vol. 1, no. 3, pp. 195–202, 2004. (Cité en page 27.)

[56] J. E. Elias and S. P. Gygi, "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry," *Nature methods*, vol. 4, no. 3, pp. 207–214, 2007. (Cité en pages 28 and 51.)

[57] P. Sinitcyn, J. D. Rudolph, and J. Cox, "Computational methods for understanding mass spectrometry–based shotgun proteomics data," *Annual Review of Biomedical Data Science*, vol. 1, pp. 207–234, 2018. (Cité en page 28.)

[58] S. A. Gerber, J. Rush, O. Stemman, M. W. Kirschner, and S. P. Gygi, "Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 6940–6945, 2003. (Cité en page 28.)

[59] B. Domon and R. Aebersold, "Options and considerations when selecting a quantitative proteomics strategy," *Nature biotechnology*, vol. 28, no. 7, pp. 710–721, 2010. (Cité en page 28.)

[60] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, "Quantitative mass spectrometry in proteomics : a critical review," *Analytical and bioanalytical chemistry*, vol. 389, no. 4, pp. 1017–1031, 2007. (Cité en page 28.)

[61] N. W. Bateman, S. P. Goulding, N. J. Shulman, A. K. Gadok, K. K. Szumlinski, M. J. MacCoss, and C. C. Wu, "Maximizing peptide identification events in proteomic workflows using data-dependent acquisition (dda)," *Molecular & Cellular Proteomics*, vol. 13, no. 1, pp. 329–338, 2014. (Cité en page 29.)

[62] M. E. Monroe, N. Tolić, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith, "Viper : an advanced software package to support high-throughput lc-ms peptide identification," *Bioinformatics*, vol. 23, no. 15, pp. 2021–2023, 2007. (Cité en page 29.)

[63] J. Cox and M. Mann, "Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification," *Nature biotechnology*, vol. 26, no. 12, pp. 1367–1372, 2008. (Cité en page 29.)

[64] S. Tyanova, T. Temu, and J. Cox, "The maxquant computational platform for mass spectrometry-based shotgun proteomics," *Nature protocols*, vol. 11, no. 12, p. 2301, 2016. (Cité en pages 29, 30 et 31.)

[65] D. Bouyssié, A.-M. Hesse, E. Mouton-Barbosa, M. Rompais, C. Macron, C. Carapito, A. Gonzalez de Peredo, Y. Couté, V. Dupierris, A. Burel, *et al.*, "Proline : an efficient and user-friendly software suite for large-scale proteomics," *Bioinformatics*, vol. 36, no. 10, pp. 3148–3155, 2020. (Cité en page 29.)

[66] Y. Zhang, Z. Wen, M. P. Washburn, and L. Florens, "Refinements to label free proteome quantitation : how to deal with peptides shared by multiple proteins," *Analytical chemistry*, vol. 82, no. 6, pp. 2272–2281, 2010. (Cité en page 30.)

[67] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, *et al.*, "The swiss-prot protein knowledgebase and its supplement trembl in 2003," *Nucleic acids research*, vol. 31, no. 1, pp. 365–370, 2003. (Cité en page 30.)

[68] M. Blein-Nicolas and M. Zivy, "Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1864, no. 8, pp. 883–895, 2016. (Cité en pages 30 and 31.)

[69] J. A. Bubis, L. I. Levitsky, M. V. Ivanov, I. A. Tarasova, and M. V. Gorshkov, "Comparative evaluation of label-free quantification methods for shotgun proteomics," *Rapid Communications in Mass Spectrometry*, vol. 31, no. 7, pp. 606–612, 2017. (Cité en page 30.)

[70] B. Dost, N. Bandeira, X. Li, Z. Shen, S. P. Briggs, and V. Bafna, "Accurate mass spectrometry based protein quantification via shared peptides," *Journal of Computational Biology*, vol. 19, no. 4, pp. 337–348, 2012. (Cité en page 30.)

[71] M. Wierer, M. Prestel, H. B. Schiller, G. Yan, C. Schaab, S. Azghandi, J. Werner, T. Kessler, R. Malik, M. Murgia, *et al.*, "Compartment-resolved proteomic analysis of mouse aorta during atherosclerotic plaque formation reveals osteoclast-specific protein expression," *Molecular & Cellular Proteomics*, vol. 17, no. 2, pp. 321–334, 2018. (Cité en page 33.)

[72] A. G. Grocin, R. A. Serwa, J. M. Sanfrutos, M. Ritzefeld, and E. W. Tate, "Whole proteome profiling of n-myristoyltransferase activity and inhibition using sortase a," *Molecular & Cellular Proteomics*, vol. 18, no. 1, pp. 115–126, 2019. (Cité en page 33.)

[73] S. Wieczorek, F. Combes, C. Lazar, Q. Giai Gianetto, L. Gatto, A. Dorffer, A.-M. Hesse, Y. Couté, M. Ferro, C. Bruley, *et al.*, "Dapar & prostar : software to perform statistical analyses in quantitative discovery proteomics," *Bioinformatics*, vol. 33, no. 1, pp. 135–136, 2017. (Cité en pages 34, 36, 43 et 121.)

[74] M. Choi, C.-Y. Chang, T. Clough, D. Broudy, T. Killeen, B. MacLean, and O. Vitek, "Msstats : an r package for statistical analysis of quantitative mass spectrometry-based proteomic experiments," *Bioinformatics*, vol. 30, no. 17, pp. 2524–2526, 2014. (Cité en page 34.)

[75] L. J. Goeminne, K. Gevaert, and L. Clement, "Experimental design and data-analysis in label-free quantitative lc/ms proteomics : A tutorial with msqrob," *Journal of proteomics*, vol. 171, pp. 23–36, 2018. (Cité en page 34.)

[76] A. M. Crowell, C. S. Greene, J. J. Loros, and J. C. Dunlap, "Learning and imputation for mass-spec bias reduction (limbr)," *Bioinformatics*, vol. 35, no. 9, pp. 1518–1526, 2019. (Cité en pages 34, 35 et 36.)

[77] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M. Y. Hein, T. Geiger, M. Mann, and J. Cox, "The perseus computational platform for comprehensive analysis of (prote) omics data," *Nature methods*, vol. 13, no. 9, p. 731, 2016. (Cité en page 34.)

[78] G. K. Smyth, "Limma : linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*, pp. 397–420, Springer, 2005. (Cité en pages 34, 44, 55 et 90.)

[79] W. Huber, A. Von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18, no. suppl_1, pp. S96–S104, 2002. (Cité en pages 34, 76, 81, 82 et 83.)

[80] N. Pavelka, M. L. Fournier, S. K. Swanson, M. Pelizzola, P. Ricciardi-Castagnoli, L. Florens, and M. P. Washburn, "Statistical similarities between transcriptomics and quantitative shotgun proteomics data," *Molecular & Cellular Proteomics*, vol. 7, no. 4, pp. 631–644, 2008. (Cité en page 34.)

[81] A. L. Oberg and O. Vitek, "Statistical design of quantitative mass spectrometry-based proteomic experiments," *Journal of proteome research*, vol. 8, no. 5, pp. 2144–2156, 2009. (Cité en page 34.)

[82] S. J. Callister, R. C. Barry, J. N. Adkins, E. T. Johnson, W.-j. Qian, B.-J. M. Webb-Robertson, R. D. Smith, and M. S. Lipton, "Normalization approaches for removing sys-

tematic biases associated with mass spectrometry and label-free proteomics," *Journal of proteome research*, vol. 5, no. 2, pp. 277–286, 2006. (Cité en pages 34, 36, 59, 61, 84 et 85.)

[83] T. Aittokallio, "Dealing with missing values in large-scale studies : microarray data imputation and beyond," *Briefings in bioinformatics*, vol. 11, no. 2, pp. 253–264, 2010. (Cité en page 34.)

[84] W. S. Noble, "How does multiple testing correction work ?," *Nature biotechnology*, vol. 27, no. 12, pp. 1135–1137, 2009. (Cité en page 34.)

[85] I. S. L. Zeng and T. Lumley, "Review of statistical learning methods in integrated omics studies (an integrated information science)," *Bioinformatics and biology insights*, vol. 12, p. 1177932218759292, 2018. (Cité en page 34.)

[86] S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, "Light-directed, spatially addressable parallel chemical synthesis," *science*, vol. 251, no. 4995, pp. 767–773, 1991. (Cité en page 34.)

[87] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995. (Cité en page 34.)

[88] Y. V. Karpievitch, A. R. Dabney, and R. D. Smith, "Normalization and missing value imputation for label-free lc-ms analysis," *BMC bioinformatics*, vol. 13, no. 16, pp. 1–9, 2012. (Cité en pages 35, 36, 38, 59 et 61.)

[89] W. W. B. Goh, W. Wang, and L. Wong, "Why batch effects matter in omics data, and how to avoid them," *Trends in biotechnology*, vol. 35, no. 6, pp. 498–507, 2017. (Cité en pages 35, 58, 59 et 61.)

[90] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010. (Cité en pages 35 and 57.)

[91] K. A. Baggerly, K. R. Coombes, and E. S. Neeley, "Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer," *Journal of Clinical Oncology*, vol. 26, no. 7, pp. 1186–1187, 2008. (Cité en pages 35 and 57.)

[92] T. Välikangas, T. Suomi, and L. L. Elo, "A systematic evaluation of normalization methods in quantitative label-free proteomics," *Briefings in bioinformatics*, vol. 19, no. 1, pp. 1–11, 2018. (Cité en pages 35, 36, 59, 61, 77, 84 et 85.)

[93] B.-J. M. Webb-Robertson, H. K. Wiberg, M. M. Matzke, J. N. Brown, J. Wang, J. E. McDermott, R. D. Smith, K. D. Rodland, T. O. Metz, J. G. Pounds, *et al.*, "Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics," *Journal of proteome research*, vol. 14, no. 5, pp. 1993–2001, 2015. (Cité en pages 35 and 36.)

[94] C. Lazar, L. Gatto, M. Ferro, C. Bruley, and T. Burger, "Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies," *Journal of proteome research*, vol. 15, no. 4, pp. 1116–1125, 2016. (Cité en pages 35, 36 et 38.)

[95] A. Davey *et al.*, *Statistical power analysis with missing data : A structural equation modeling approach*. Routledge, 2009. (Cité en page 35.)

[96] A. R. T. Donders, G. J. Van Der Heijden, T. Stijnen, and K. G. Moons, "A gentle introduction to imputation of missing values," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1087–1091, 2006. (Cité en page 36.)

[97] L. Jin, Y. Bi, C. Hu, J. Qu, S. Shen, X. Wang, and Y. Tian, "A comparative study of evaluating missing value imputation methods in label-free proteomics," *Scientific reports*, vol. 11, no. 1, pp. 1–11, 2021. (Cité en pages 36 and 38.)

[98] K. Kultima, A. Nilsson, B. Scholz, U. L. Rossbach, M. Fälth, and P. E. Andrén, "Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides," *Molecular & Cellular Proteomics*, vol. 8, no. 10, pp. 2285–2295, 2009. (Cité en pages 36, 84 et 85.)

[99] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003. (Cité en pages 36, 59 et 82.)

[100] M. B. ORourke, S. E. Town, P. V. Dalla, F. Bicknell, N. Koh Belic, J. P. Violi, J. R. Steele, and M. P. Padula, "What is normalization ? the strategies employed in top-down and bottom-up proteome analysis workflows," *Proteomes*, vol. 7, no. 3, p. 29, 2019. (Cité en page 36.)

[101] M. Liu and A. Dongre, "Proper imputation of missing values in proteomics datasets for differential expression analysis," *Briefings in Bioinformatics*, 2020. (Cité en page 38.)

[102] N. A. Karp and K. S. Lilley, "Design and analysis issues in quantitative proteomics studies," *Proteomics*, vol. 7, no. S1, pp. 42–50, 2007. (Cité en page 38.)

[103] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman, "Statistical tests, p values, confidence intervals, and power : a guide to misinterpretations," *European journal of epidemiology*, vol. 31, no. 4, pp. 337–350, 2016. (Cité en pages 38 and 42.)

[104] G. Saporta, *Probabilités, analyse des données et statistique*. Editions Technip, 2006. (Cité en page 39.)

[105] N. Rifai, M. A. Gillette, and S. A. Carr, "Protein biomarker discovery and validation : the long and uncertain path to clinical utility," *Nature biotechnology*, vol. 24, no. 8, pp. 971–983, 2006. (Cité en page 40.)

[106] J. A. Sterne and G. D. Smith, "Sifting the evidencewhat's wrong with significance tests ?," *Physical therapy*, vol. 81, no. 8, pp. 1464–1469, 2001. (Cité en page 41.)

[107] W. M. Shibasaki and R. P. Martins, "Simple randomization may lead to unequal group sizes. is that a problem ?," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 154, no. 4, pp. 600–605, 2018. (Cité en pages 41 and 59.)

[108] R. B. Bausell and Y.-F. Li, *Power analysis for experimental research : a practical guide for the biological, medical and social sciences*. Cambridge University Press, 2002. (Cité en page 41.)

[109] J. Cohen, *Statistical power analysis for the behavioral sciences*. Academic press, 2013. (Cité en page 41.)

[110] M. Krzywinski and N. Altman, "Points of significance : Significance, p values and t-tests," 2013. (Cité en page 42.)

[111] F. Dorey, "In brief : The p value : What is it and what does it tell you ?," 2010. (Cité en page 42.)

[112] R. Nuzzo, "Scientific method : statistical errors," *Nature News*, vol. 506, no. 7487, p. 150, 2014. (Cité en page 42.)

[113] R. L. Wasserstein and N. A. Lazar, "The asa statement on p-values : context, process, and purpose," 2016. (Cité en page 42.)

[114] J. I. Krueger and P. R. Heck, "Putting the p-value in its place," *The American Statistician*, vol. 73, no. sup1, pp. 122–128, 2019. (Cité en page 42.)

[115] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908. (Cité en page 43.)

[116] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics bulletin*, vol. 2, no. 6, pp. 110–114, 1946. (Cité en page 44.)

[117] N. J. Johnson, "Modified t tests and confidence intervals for asymmetrical populations," *Journal of the American Statistical Association*, vol. 73, no. 363, pp. 536–544, 1978. (Cité en page 44.)

[118] Q. F. Gronau, A. Ly, and E.-J. Wagenmakers, "Informed bayesian t-tests," *The American Statistician*, vol. 74, no. 2, pp. 137–143, 2020. (Cité en page 44.)

[119] B. L. Welch, "The significance of the difference between two means when the population variances are unequal," *Biometrika*, vol. 29, no. 3/4, pp. 350–362, 1938. (Cité en page 44.)

[120] S. Lee and D. K. Lee, "What is the proper way to apply the multiple comparison test ?," *Korean journal of anesthesiology*, vol. 71, no. 5, p. 353, 2018. (Cité en pages 44 and 48.)

[121] C. Tian, X. Manfei, T. Justin, W. Hongyue, and N. Xiaohui, "Relationship between omnibus and post-hoc tests : An investigation of performance of the f test in anova," *Shanghai archives of psychiatry*, vol. 30, no. 1, p. 60, 2018. (Cité en pages 44, 46, 48 et 90.)

[122] K. R. Tarlow, "Teaching principles of inference with anova," *Teaching Statistics*, vol. 38, no. 1, pp. 16–21, 2016. (Cité en page 44.)

[123] M. J. Crawley, *The R book*. John Wiley & Sons, 2012. (Cité en page 45.)

[124] R. A. Cribbie, L. Fiksenbaum, H. J. Keselman, and R. R. Wilcox, "Effect of non-normality on test statistics for one-way independent groups designs," *British Journal of Mathematical and Statistical Psychology*, vol. 65, no. 1, pp. 56–73, 2012. (Cité en page 46.)

[125] G. V. Glass, P. D. Peckham, and J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance," *Review of educational research*, vol. 42, no. 3, pp. 237–288, 1972. (Cité en page 46.)

[126] M. Delacre, C. Leys, Y. L. Mora, and D. Lakens, "Taking parametric assumptions seriously : Arguments for the use of welchs f-test instead of the classical f-test in one-way anova," *International Review of Social Psychology*, vol. 32, no. 1, 2019. (Cité en page 46.)

[127] B. L. Welch, "On the comparison of several mean values : an alternative approach," *Biometrika*, vol. 38, no. 3/4, pp. 330–336, 1951. (Cité en pages 46, 55 et 124.)

[128] P. Mair and R. Wilcox, "Robust statistical methods in r using the wrs2 package," *Behavior research methods*, pp. 1–25, 2019. (Cité en page 46.)

[129] K. Moder, "Alternatives to f-test in one way anova in case of heterogeneity of variances (a simulation study)," *Psychological Test and Assessment Modeling*, vol. 52, no. 4, pp. 343–353, 2010. (Cité en pages 46 and 124.)

[130] R. Bender and S. Lange, "Adjusting for multiple testingwhen and how ?," *Journal of clinical epidemiology*, vol. 54, no. 4, pp. 343–349, 2001. (Cité en pages 47 and 48.)

[131] A. P. Diz, A. Carvajal-Rodríguez, and D. O. Skibinski, "Multiple hypothesis testing in proteomics : a strategy for experimental work," *Molecular & Cellular Proteomics*, vol. 10, no. 3, pp. M110–004374, 2011. (Cité en pages 47 and 88.)

[132] K. L. Hassall and A. Mead, "Beyond the one-way anova foromics data," *BMC bioinformatics*, vol. 19, no. 7, pp. 109–126, 2018. (Cité en page 47.)

[133] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilita," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, vol. 8, pp. 3–62, 1936. (Cité en page 47.)

[134] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate : a practical and powerful approach to multiple testing," *Journal of the Royal statistical society : series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. (Cité en pages 47, 50, 51 et 125.)

[135] G. D. Ruxton and G. Beauchamp, "Time for some a priori thinking about post hoc testing," *Behavioral ecology*, vol. 19, no. 3, pp. 690–693, 2008. (Cité en pages 48 and 90.)

[136] H. Scheffé, "A method for judging all contrasts in the analysis of variance," *Biometrika*, vol. 40, no. 1-2, pp. 87–110, 1953. (Cité en page 48.)

[137] J. W. Tukey, "Comparing individual means in the analysis of variance," *Biometrics*, pp. 99–114, 1949. (Cité en page 48.)

[138] H. Abdi and L. J. Williams, "Tukeys honestly significant difference (hsd) test," *Encyclopedia of research design*, vol. 3, no. 1, pp. 1–5, 2010. (Cité en page 48.)

[139] C. W. Dunnett, "A multiple comparison procedure for comparing several treatments with a control," *Journal of the American Statistical Association*, vol. 50, no. 272, pp. 1096–1121, 1955. (Cité en page 48.)

[140] U. Kucuk, M. Eyuboglu, H. O. Kucuk, and G. Degirmencioglu, "Importance of using proper post hoc test with anova," *International journal of cardiology*, vol. 209, p. 346, 2016. (Cité en page 48.)

[141] K. Holm and N. J. Christman, "Post hoc tests following analysis of variance," *Research in nursing & health*, vol. 8, no. 2, pp. 207–210, 1985. (Cité en page 48.)

[142] R. F. Barber, E. J. Candès, *et al.*, "Controlling the false discovery rate via knockoffs," *Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015. (Cité en page 50.)

[143] Y. Benjamini, A. M. Krieger, and D. Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *Biometrika*, vol. 93, no. 3, pp. 491–507, 2006. (Cité en page 50.)

[144] S. Pounds and C. Cheng, "Robust estimation of the false discovery rate," *Bioinformatics*, vol. 22, no. 16, pp. 1979–1987, 2006. (Cité en pages 50 and 51.)

[145] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001. (Cité en page 51.)

[146] J. J. Goeman and A. Solari, "Multiple hypothesis testing in genomics," *Statistics in medicine*, vol. 33, no. 11, pp. 1946–1978, 2014. (Cité en page 51.)

[147] T. Burger, "Gentle introduction to the statistical foundations of false discovery rate in quantitative proteomics," *Journal of proteome research*, vol. 17, no. 1, pp. 12–22, 2018. (Cité en page 51.)

[148] N. Gupta, N. Bandeira, U. Keich, and P. A. Pevzner, "Target-decoy approach and false discovery rate : when things may go wrong," *Journal of the American Society for Mass Spectrometry*, vol. 22, no. 7, pp. 1111–1120, 2011. (Cité en page 51.)

[149] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by ms/ms and database search," *Analytical chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002. (Cité en page 51.)

[150] M. The, A. Tasnim, and L. Käll, "How to talk about protein-level false discovery rates in shotgun proteomics," *Proteomics*, vol. 16, no. 18, pp. 2461–2469, 2016. (Cité en page 51.)

[151] Y. Couté, C. Bruley, and T. Burger, "Beyond target–decoy competition : Stable validation of peptide and protein identifications in mass spectrometry-based discovery proteomics," *Analytical Chemistry*, vol. 92, no. 22, pp. 14898–14906, 2020. (Cité en page 52.)

[152] P. Bedossa, C. Poitou, N. Veyrie, J.-L. Bouillot, A. Basdevant, V. Paradis, J. Tordjman, and K. Clement, "Histopathological algorithm and scoring system for evaluation of liver lesions in morbidly obese patients," *Hepatology*, vol. 56, no. 5, pp. 1751–1759, 2012. (Cité en pages 53, 107 et 110.)

[153] H. Borges, A.-M. Hesse, A. Kraut, Y. Couté, V. Brun, and T. Burger, "Well plate maker : A user-friendly randomized block design application to limit batch effects in largescale biomedical studies," *Bioinformatics*, 2021. (Cité en pages 55, 64 et 161.)

[154] B. J. Mertens, "Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies," *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry*, pp. 1–21, 2017. (Cité en pages 58 and 59.)

[155] N. Pandis, A. Polychronopoulou, and T. Eliades, "Randomization in clinical trials in orthodontics : its significance in research design and methods to achieve it," *The European Journal of Orthodontics*, vol. 33, no. 6, pp. 684–690, 2011. (Cité en page 59.)

[156] J. A. Nelder, "The analysis of randomized experiments with orthogonal block structure. i. block structure and the null analysis of variance," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 283, no. 1393, pp. 147–162, 1965. (Cité en page 59.)

[157] B. Burger, M. Vaudel, and H. Barsnes, "Importance of block randomization when designing proteomics experiments," *Journal of Proteome Research*, vol. 20, no. 1, pp. 122–128, 2020. (Cité en pages 59, 60, 61 et 70.)

[158] O. D. Buhule, R. L. Minster, N. L. Hawley, M. Medvedovic, G. Sun, S. Viali, R. Deka, S. T. McGarvey, and D. E. Weeks, "Stratified randomization controls better for batch effects in 450k methylation analysis : a cautionary tale," *Frontiers in genetics*, vol. 5, p. 354, 2014. (Cité en page 59.)

[159] A. Chawade, E. Alexandersson, and F. Levander, "Normalyzer : a tool for rapid evaluation of normalization methods for omics data sets," *Journal of proteome research*, vol. 13, no. 6, pp. 3114–3120, 2014. (Cité en page 59.)

[160] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901. (Cité en page 59.)

[161] L. L. Thurstone, "Multiple factor analysis.," *Psychological review*, vol. 38, no. 5, p. 406, 1931. (Cité en page 60.)

[162] B. Everitt, "Latent variable models for categorical data," in *An Introduction to Latent Variable Models*, pp. 72–88, Springer, 1984. (Cité en page 60.)

[163] F. Hahne, W. Huber, M. Ruschhaupt, and J. Toedling, "prada : Data analysis for cell-based functional assays," 2008. (Cité en page 61.)

[164] S. M. Hughes, "plater : Read, tidy, and display data from microtiter plates," *Journal of Open Source Software*, vol. 1, no. 7, p. 106, 2016. (Cité en page 61.)

[165] MathWorks, "Bioinformatics toolbox : Users guide," 2020. r2020a, https://fr.mathworks.com/products/bioinfo.html. (Cité en page 61.)

[166] I. D. Shterev, C. Cliburn, and S. G. D., "highscreen : High-throughput screening for plate based essays," 2019. R package version 0.4, https://cran.r-project.org/web/packages/highSCREEN/index.html. (Cité en page 61.)

[167] W. Scott, "platetools : Tools and plots for multi-well plates." R package version 0.1.5, https://cran.r-project.org/web/packages/platetools/index.html. (Cité en page 61.)

[168] M. Suprun and M. Suárez-Fariñas, "Platedesigner : a web-based application for the design of microplate experiments," *Bioinformatics*, vol. 35, no. 9, pp. 1605–1607, 2019. (Cité en pages 61 and 70.)

[169] D. E. Knuth, *Art of computer programming*, vol. 4b(5). Addison Wesley Professional, 2018. (Cité en page 62.)

[170] L. Sinke, D. Cats, and B. T. Heijmans, "Omixer : multivariate and reproducible sample randomization to proactively counter batch effects in omics studies," *Bioinformatics*, 2021. (Cité en page 71.)

[171] M. Olaf, "microbenchmark : Accurate timing functions," 2019. R package version 1.4-7, https://CRAN.R-project.org/package=microbenchmark. (Cité en page 73.)

[172] W. S. Cleveland and S. J. Devlin, "Locally weighted regression : an approach to regression analysis by local fitting," *Journal of the American statistical association*, vol. 83, no. 403, pp. 596–610, 1988. (Cité en pages 76, 77, 81 et 84.)

[173] K. V. Ballman, D. E. Grill, A. L. Oberg, and T. M. Therneau, "Faster cyclic loess : normalizing rna arrays via linear models," *Bioinformatics*, vol. 20, no. 16, pp. 2778–2786, 2004. (Cité en pages 76 and 81.)

[174] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Parameter estimation for the calibration and variance stabilization of microarray data," *Statistical applications in genetics and molecular biology*, vol. 2, no. 1, 2003. (Cité en pages 76 and 82.)

[175] X. Zhang, A. H. Smits, G. B. van Tilburg, H. Ovaa, W. Huber, and M. Vermeulen, "Proteome-wide identification of ubiquitin interactions using ubia-ms," *Nature protocols*, vol. 13, no. 3, p. 530, 2018. (Cité en page 76.)

[176] W. G. Jacoby, "Loess: : a nonparametric, graphical tool for depicting relationships between variables," *Electoral Studies*, vol. 19, no. 4, pp. 577–613, 2000. (Cité en pages 77 and 81.)

[177] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979. (Cité en pages 77, 78 et 81.)

[178] W. S. Cleveland, "Lowess : A program for smoothing scatterplots by robust locally weighted regression," *American Statistician*, vol. 35, no. 1, p. 54, 1981. (Cité en page 77.)

[179] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments," *Statistica sinica*, pp. 111–139, 2002. (Cité en page 79.)

[180] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cdna microarray images," *Journal of Biomedical optics*, vol. 2, no. 4, pp. 364–374, 1997. (Cité en page 82.)

[181] B.-J. M. Webb-Robertson, M. M. Matzke, J. M. Jacobs, J. G. Pounds, and K. M. Waters, "A statistical selection strategy for normalization procedures in lc-ms proteomics experiments through dataset-dependent ranking of normalization scaling factors," *Proteomics*, vol. 11, no. 24, pp. 4736–4741, 2011. (Cité en page 83.)

[182] T. Schweder and E. Spjøtvoll, "Plots of p-values to evaluate many tests simultaneously," *Biometrika*, vol. 69, no. 3, pp. 493–502, 1982. (Cité en page 84.)

[183] Q. Giai Gianetto, F. Combes, C. Ramus, C. Bruley, Y. Couté, and T. Burger, "Calibration plot for proteomics : A graphical tool to visually check the assumptions underlying fdr control in quantitative experiments," *Proteomics*, vol. 16, no. 1, pp. 29–32, 2016. (Cité en pages 85 and 121.)

[184] G. K. Smyth, M. Ritchie, N. Thorne, J. Wettenhall, W. Shi, and Y. Hu, "limma : linear models for microarray and rna-seq data users guide," 2020. (Cité en pages 88, 96, 97 et 99.)

[185] J. J. Goeman and A. Solari, "Tutorial in biostatistics : multiple hypothesis testing in genomics," *Stat Med*, pp. 1–20, 2012. (Cité en page 88.)

[186] F. Emmert-Streib and M. Dehmer, "Large-scale simultaneous inference with hypothesis testing : Multiple testing procedures in practice," *Machine Learning and Knowledge Extraction*, vol. 1, no. 2, pp. 653–683, 2019. (Cité en page 88.)

[187] S. R. Austin, I. Dialsingh, and N. Altman, "Multiple hypothesis testing : a review," *J Indian Soc Agric Stat*, vol. 68, no. 2, pp. 303–14, 2014. (Cité en page 88.)

[188] B. Zhang, N. C. VerBerkmoes, M. A. Langston, E. Uberbacher, R. L. Hettich, and N. F. Samatova, "Detecting differential and correlated protein expression in label-free shotgun proteomics," *Journal of proteome research*, vol. 5, no. 11, pp. 2909–2918, 2006. (Cité en pages 95 and 105.)

[189] G. Domingo, F. Villa, C. Vannini, E. Garuglieri, E. Onelli, M. Bracale, and F. Cappitelli, "Label-free proteomic approach to study the non-lethal effects of silver nanoparticles on a gut bacterium," *Frontiers in microbiology*, vol. 10, p. 2709, 2019. (Cité en page 95.)

[190] J. Zhou, Y. Song, T. Xing, L. Ge, L. Ma, M. Lu, and L. Zhong, "Proteomic analysis reveals distinctive protein expression patterns of thrombus in clear cell renal cell carcinoma," *Translational Oncology*, vol. 14, no. 1, p. 100895, 2021. (Cité en page 95.)

[191] N. Casas-Vila, A. Bluhm, S. Sayols, N. Dinges, M. Dejung, T. Altenhein, D. Kappei, B. Altenhein, J.-Y. Roignant, and F. Butter, "The developmental proteome of drosophila melanogaster," *Genome research*, vol. 27, no. 7, pp. 1273–1285, 2017. (Cité en page 95.)

[192] Y.-J. Chen, T. I. Roumeliotis, Y.-H. Chang, C.-T. Chen, C.-L. Han, M.-H. Lin, H.-W. Chen, G.-C. Chang, Y.-L. Chang, C.-T. Wu, *et al.*, "Proteogenomics of non-smoking lung cancer in east asia delineates molecular signatures of pathogenesis and progression," *Cell*, vol. 182, no. 1, pp. 226–244, 2020. (Cité en page 101.)

[193] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967. (Cité en page 102.)

[194] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007. (Cité en pages 102, 103 et 125.)

[195] B. S. Everitt, "Unresolved problems in cluster analysis," *Biometrics*, pp. 169–181, 1979. (Cité en page 103.)

[196] C. Patil and I. Baidari, "Estimating the optimal number of clusters k in a dataset using data depth," *Data Science and Engineering*, vol. 4, no. 2, pp. 132–140, 2019. (Cité en page 103.)

[197] A. Adolfsson, M. Ackerman, and N. C. Brownstein, "To cluster, or not to cluster : An analysis of clusterability methods," *Pattern Recognition*, vol. 88, pp. 13–26, 2019. (Cité en page 104.)

[198] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *The annals of Statistics*, pp. 70–84, 1985. (Cité en page 104.)

[199] M. E. Rinella, F. Tacke, A. J. Sanyal, Q. M. Anstee, *et al.*, "Report on the aasld/easl joint workshop on clinical trial endpoints in nafld," *Journal of hepatology*, vol. 71, no. 4, pp. 823–833, 2019. (Cité en pages 107, 108, 109, 110, 112 et 113.)

[200] S. Singh, A. M. Allen, Z. Wang, L. J. Prokop, M. H. Murad, and R. Loomba, "Fibrosis progression in nonalcoholic fatty liver vs nonalcoholic steatohepatitis : a systematic review and meta-analysis of paired-biopsy studies," *Clinical gastroenterology and hepatology*, vol. 13, no. 4, pp. 643–654, 2015. (Cité en page 108.)

[201] Z. M. Younossi, "Non-alcoholic fatty liver disease–a global public health perspective," *Journal of hepatology*, vol. 70, no. 3, pp. 531–544, 2019. (Cité en pages 108, 109 et 110.)

[202] Z. M. Younossi, A. B. Koenig, D. Abdelatif, Y. Fazel, L. Henry, and M. Wymer, "Global epidemiology of nonalcoholic fatty liver diseasemeta-analytic assessment of prevalence, incidence, and outcomes," *Hepatology*, vol. 64, no. 1, pp. 73–84, 2016. (Cité en pages 108 and 109.)

[203] Z. Younossi, M. Reyes, A. Mishra, R. Mehta, and L. Henry, "Systematic review with meta-analysis : non-alcoholic steatohepatitis-a case for personalised treatment based on pathogenic targets," *Alimentary pharmacology & therapeutics*, vol. 39, no. 1, pp. 3–14, 2014. (Cité en page 108.)

[204] O. Nabi, K. Lacombe, J. Boursier, P. Mathurin, M. Zins, and L. Serfaty, "Prevalence and risk factors of nonalcoholic fatty liver disease and advanced fibrosis in general population : the french nationwide nash-co study gastroenterology. 2020 aug (epub 2020 may 4), 159 (2) : 791-793. doi : 10.1053/j. gastro. 2020.04. 048.," *Gastroenterology*, 2020. (Cité en page 109.)

[205] Z. Younossi, F. Tacke, M. Arrese, B. Chander Sharma, I. Mostafa, E. Bugianesi, V. Wai-Sun Wong, Y. Yilmaz, J. George, J. Fan, *et al.*, "Global perspectives on nonalcoholic fatty liver disease and nonalcoholic steatohepatitis," *Hepatology*, vol. 69, no. 6, pp. 2672–2682, 2019. (Cité en page 109.)

[206] V. W.-S. Wong, L. A. Adams, V. de Lédinghen, G. L.-H. Wong, and S. Sookoian, "Noninvasive biomarkers in nafld and nashcurrent progress and future promise," *Nature reviews*

*Gastroenterology & hepatology*, vol. 15, no. 8, pp. 461–478, 2018. (Cité en pages 109, 110, 111, 112, 130 et 132.)

[207] P. Bedossa and F. P. Consortium, "Utility and appropriateness of the fatty liver inhibition of progression (flip) algorithm and steatosis, activity, and fibrosis (saf) score in the evaluation of biopsies of nonalcoholic fatty liver disease," *Hepatology*, vol. 60, no. 2, pp. 565–575, 2014. (Cité en pages 110, 112, 113 et 114.)

[208] M. A. Karsdal, S. J. Daniels, S. Holm Nielsen, C. Bager, D. G. Rasmussen, R. Loomba, R. Surabattula, I. F. Villesen, Y. Luo, D. Shevell, *et al.*, "Collagen biology and non-invasive biomarkers of liver fibrosis," *Liver International*, vol. 40, no. 4, pp. 736–750, 2020. (Cité en pages 110, 111 et 112.)

[209] S. P. Singh and R. K. Barik, "Noninvasive biomarkers in nonalcoholic fatty liver disease : Are we there yet ?," *Journal of clinical and experimental hepatology*, vol. 10, no. 1, pp. 88–98, 2020. (Cité en pages 110 and 111.)

[210] D. E. Kleiner, E. M. Brunt, M. Van Natta, C. Behling, M. J. Contos, O. W. Cummings, L. D. Ferrell, Y.-C. Liu, M. S. Torbenson, A. Unalp-Arida, *et al.*, "Design and validation of a histological scoring system for nonalcoholic fatty liver disease," *Hepatology*, vol. 41, no. 6, pp. 1313–1321, 2005. (Cité en page 110.)

[211] C. D. Byrne and G. Targher, "Whats new in nafld pathogenesis, biomarkers and treatment ?," *Nature reviews gastroenterology & hepatology*, vol. 17, no. 2, pp. 70–71, 2020. (Cité en page 111.)

[212] G. Bedogni, S. Bellentani, L. Miglioli, F. Masutti, M. Passalacqua, A. Castiglione, and C. Tiribelli, "The fatty liver index : a simple and accurate predictor of hepatic steatosis in the general population," *BMC gastroenterology*, vol. 6, no. 1, pp. 1–7, 2006. (Cité en page 111.)

[213] J.-H. Lee, D. Kim, H. J. Kim, C.-H. Lee, J. I. Yang, W. Kim, Y. J. Kim, J.-H. Yoon, S.-H. Cho, M.-W. Sung, *et al.*, "Hepatic steatosis index : a simple screening tool reflecting nonalcoholic fatty liver disease," *Digestive and Liver Disease*, vol. 42, no. 7, pp. 503–508, 2010. (Cité en page 111.)

[214] T. Poynard, V. Ratziu, S. Naveau, D. Thabut, F. Charlotte, D. Messous, D. Capron, A. Abella, J. Massard, Y. Ngo, *et al.*, "The diagnostic value of biomarkers (steatotest) for the prediction of liver steatosis," *Comparative hepatology*, vol. 4, no. 1, pp. 1–14, 2005. (Cité en page 111.)

[215] V. Ratziu, J. Massard, F. Charlotte, D. Messous, F. Imbert-Bismut, L. Bonyhay, M. Tahiri, M. Munteanu, D. Thabut, J. F. Cadranel, *et al.*, "Diagnostic value of biochemical markers (fibrotest-fibrosure) for the prediction of liver fibrosis in patients with non-alcoholic fatty liver disease," *BMC gastroenterology*, vol. 6, no. 1, pp. 1–13, 2006. (Cité en page 111.)

[216] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis : a guide for non-statisticians," *International journal of endocrinology and metabolism*, vol. 10, no. 2, p. 486, 2012. (Cité en page 122.)

[217] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Annals of cardiac anaesthesia*, vol. 22, no. 1, p. 67, 2019. (Cité en page 123.)

[218] D. Öztuna, A. H. Elhan, and E. Tüccar, "Investigation of four different normality tests in terms of type 1 error rate and power under different distributions," *Turkish Journal of Medical Sciences*, vol. 36, no. 3, pp. 171–176, 2006. (Cité en page 123.)

[219] K. Le Boedec, "Sensitivity and specificity of normality tests and consequences on reference interval accuracy at small sample size : a computer-simulation study," *Veterinary clinical pathology*, vol. 45, no. 4, pp. 648–656, 2016. (Cité en page 123.)

[220] M. Kozak and H.-P. Piepho, "What's normal anyway ? residual plots are more telling than significance tests when checking anova assumptions," *Journal of Agronomy and Crop Science*, vol. 204, no. 1, pp. 86–98, 2018. (Cité en page 123.)

[221] D. Sharma and B. G. Kibria, "On some test statistics for testing homogeneity of variances : a comparative study," *Journal of Statistical Computation and Simulation*, vol. 83, no. 10, pp. 1944–1963, 2013. (Cité en page 123.)

[222] M. S. Bartlett, "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, vol. 160, no. 901, pp. 268–282, 1937. (Cité en page 123.)

[223] H. Levene, "Robust tests for equality of variances," *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, pp. 279–292, 1961. (Cité en page 123.)

[224] G. V. Glass, "Testing homogeneity of variances," *American Educational Research Journal*, vol. 3, no. 3, pp. 187–190, 1966. (Cité en page 123.)

[225] Z. Mu, *Comparing the statistical tests for homogeneity of variances.* PhD thesis, East Tennessee State University, 2006. (Cité en page 123.)

[226] B. Y. Lemeshko and A. Y. Novikova, "Application and power of tests for homogeneity of variances," in *2018 XIV International Scientific-Technical Conference on Actual Problems of Electronics Instrument Engineering (APEIE)*, pp. 146–152, IEEE, 2018. (Cité en page 123.)

[227] Y. J. Kim and R. A. Cribbie, "Anova and the variance homogeneity assumption : Exploring a better gatekeeper," *British Journal of Mathematical and Statistical Psychology*, vol. 71, no. 1, pp. 1–12, 2018. (Cité en page 124.)

[228] D. Botstein, J. M. Cherry, M. Ashburner, C. A. Ball, J. A. Blake, H. Butler, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology : tool for the unification of biology," *Nat genet*, vol. 25, no. 1, pp. 25–9, 2000. (Cité en page 126.)

[229] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, "Panther : a library of protein families and subfamilies indexed by function," *Genome research*, vol. 13, no. 9, pp. 2129–2141, 2003. (Cité en page 126.)

[230] L. Niu, P. E. Geyer, N. J. Wewer Albrechtsen, L. L. Gluud, A. Santos, S. Doll, P. V. Treit, J. J. Holst, F. K. Knop, T. Vilsbøll, *et al.*, "Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease," *Molecular systems biology*, vol. 15, no. 3, p. e8793, 2019. (Cité en pages 130 and 132.)

[231] W. Hou, M. G. Janech, P. M. Sobolesky, A. M. Bland, S. Samsuddin, W. Alazawi, and W.-K. Syn, "Proteomic screening of plasma identifies potential noninvasive biomarkers associated with significant/advanced fibrosis in patients with nonalcoholic fatty liver disease," *Bioscience reports*, vol. 40, no. 1, p. BSR20190395, 2020. (Cité en pages 130 and 132.)

[232] K. Ogawa, T. Kobayashi, J.-i. Furukawa, H. Hanamatsu, A. Nakamura, K. Suzuki, N. Kawagishi, M. Ohara, M. Umemura, M. Nakai, *et al.*, "Tri-antennary tri-sialylated monofucosylated glycan of alpha-1 antitrypsin as a non-invasive biomarker for non-alcoholic steatohepatitis : a novel glycobiomarker for non-alcoholic steatohepatitis," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020. (Cité en pages 130 and 132.)

[233] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, *et al.*, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, 2015. (Cité en page 132.)

[234] S. Enroth, M. Berggrund, M. Lycke, J. Broberg, M. Lundberg, E. Assarsson, M. Olovsson, K. Stålberg, K. Sundfeldt, and U. Gyllensten, "High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer," *Communications biology*, vol. 2, no. 1, pp. 1–12, 2019. (Cité en page 141.)

[235] N. L. Anderson, "The clinical plasma proteome : a survey of clinical assays for proteins in plasma and serum," *Clinical chemistry*, vol. 56, no. 2, pp. 177–185, 2010. (Cité en page 141.)

[236] A. C. Chatziioannou, J. C. Wolters, K. Sarafidis, A. Thomaidou, C. Agakidis, N. Govorukhina, J. A. Kuivenhoven, R. Bischoff, and G. Theodoridis, "Targeted lc-ms/ms for the evaluation of proteomics biomarkers in the blood of neonates with necrotizing enterocolitis and late-onset sepsis," *Analytical and bioanalytical chemistry*, vol. 410, no. 27, pp. 7163–7175, 2018. (Cité en page 141.)

[237] C. P. Tanase, E. Codrici, I. D. Popescu, S. Mihai, A.-M. Enciu, L. G. Necula, A. Preda, G. Ismail, and R. Albulescu, "Prostate cancer proteomics : Current trends and future perspectives for biomarker discovery," *Oncotarget*, vol. 8, no. 11, p. 18497, 2017. (Cité en page 141.)

[238] D. Z. Dieters-Castator, P. F. Rambau, L. E. Kelemen, G. M. Siegers, G. A. Lajoie, L.-M. Postovit, and M. Köbel, "Proteomics-derived biomarker panel improves diagnostic precision to classify endometrioid and high-grade serous ovarian carcinoma," *Clinical Cancer Research*, vol. 25, no. 14, pp. 4309–4319, 2019. (Cité en page 141.)

[239] H. Borges, R. Guibert, O. Permiakova, and T. Burger, "Distinguishing between spectral clustering and cluster analysis of mass spectra," *Journal of proteome research*, vol. 18, no. 1, pp. 571–573, 2018. (Cité en page 161.)

[240] S. Wieczorek, F. Combes, H. Borges, and T. Burger, "Protein-level statistical analysis of quantitative label-free proteomics data with prostar," 2019. (Cité en page 161.)

# Annexe A

# Livrables et responsabilités

## A.1 Liste des publications

— H. Borges, R. Guibert, O. Permiakova, and T. Burger, Distinguishing between spectral clustering and cluster analysis of mass spectra, Journal of proteome research, vol. 18, no. 1, pp. 571573, 2018.[239]
— S. Wieczorek, F. Combes, H. Borges, and T. Burger, Protein-level statistical analysis of quantitative label-free proteomics data with prostar,[240]
— H. Borges, A.-M. Hesse, A. Kraut, Y. Couté, V. Brun, and T. Burger, Well plate maker : A user-friendly randomized block design application to limit batch effects in largescale biomedical studies, Bioinformatics, 2021.[153]

## A.2 Responsabilités durant la thèse

— Webmaster du site de l'équipe EDyP
— Représentante des Doctorants et Post-doctorants au comité du laboratoire BGE pour 2018-2020
— Représentante des non-permanents au Conseil de l'IRIG 2019 et 2020

## A.3 Conférences, séminaires

— Les Doctoriales 2019 (avril 2019), *poster*
— Journée scientifique de l'ED ISCE (28 mai 2019), *poster (second prix du meilleur poster)*
— SMAP 2019 (septembre 2019), *poster*