Développements méthodologiques en protéomique quantitative pour mieux comprendre la biologie évolutive d'espèces non séquencées

Margaux Benhaïm

▶ **To cite this version:**

Margaux Benhaïm. Développements méthodologiques en protéomique quantitative pour mieux comprendre la biologie évolutive d'espèces non séquencées. Chimie analytique. Université de Strasbourg, 2017. Français. NNT : 2017STRAF032 . tel-01698589

HAL Id: tel-01698589
https://theses.hal.science/tel-01698589

Submitted on 1 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

UMR 7178

THÈSE présentée par :

Margaux BENHAÏM

soutenue le : **27 septembre 2017**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Chimie Analytique

Développements méthodologiques en protéomique quantitative pour mieux comprendre la biologie évolutive d'espèces non séquencées

THÈSE dirigée par :

M. BERTILE Fabrice	Chargé de recherche, CNRS, université de Strasbourg

RAPPORTEURS :

M. ZIVY Michel	Directeur de recherche, INRA/CNRS, Université Paris-Sud
M. THOLEY Andreas	Professeur, Université de Kiel, Allemagne

AUTRES MEMBRES DU JURY :

Mme. HABOLD Caroline	Chargé de recherche, CNRS, université de Strasbourg
M. BERTILE Fabrice	Chargé de recherche, CNRS, université de Strasbourg

A mes parents, à mon frère,

« An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer. »,

Max Planck

Remerciements

Dans un premier lieu, je tiens à remercier Alain Van Dorsselaer et Sarah Cianferani, pour m'avoir accueillie au sein du Laboratoire de Spectrométrie de Masse Bio-Organique.

Je remercie Mme Caroline Habold, M. Michel Zivy et M. Andreas Tholey d'avoir accepté d'évaluer mes travaux de thèse.

Je tiens à remercier Fabrice Bertile, pour m'avoir encadrée pendant plus de trois ans, d'abord en stage M2 puis tout au long de ma thèse. Merci pour tes nombreux conseils, pour ta disponibilité et pour tout ce que j'ai appris en travaillant sur ces projets. Merci pour ton soutien, et ton aide pour ce manuscrit. « Ça va être superbe ».

Je souhaiterais également remercier tous les collaborateurs avec qui j'ai eu la chance de travailler : François Criscuolo, Pierre Bize, Roberto Geremia et Jérémy Terrien. J'ai beaucoup appris grâce à tous ces projets. Merci également à Barbara et Martin, ce fut un plaisir de t'accompagner sur un petit bout de ton stage !

Merci à tout le LSMBO et plus particulièrement ;

Strubi, que ferait-on sans toi ? Que ferait la science sans ton expertise ? Je te remercie infiniment, pour ta disponibilité, ta bonne humeur et ton savoir que tu es toujours prêt à partager ! Tu es toujours prêt à nous venir en aide en cas de besoin, tu vas me manquer. J'espère que tu continueras à faire avancer la science !
Tatcher, c'est grâce à toi que j'ai le plus appris je crois ! Tu fus mon mentor au LSMBO dès mon arrivée en M2. Viel Dank für alle.
La p'tite Hirschler, je ne te remercie pas d'avoir tant écorché la langue française ! Mais merci pour le Nutella, et autres sucreries ; merci de m'avoir appris des expressions alsaciennes et merci pour les charades et les « quelques » moments d'incompréhension générale dans le bureau !
Marianne, merci pour les quelques mots en libanais que tu nous as appris (je n'en ai pas retenu beaucoup…), merci de m'avoir appris ce qu'était vraiment la science, et surtout MERCI pour les baklawas ! Je te souhaite plein de bonheur au Canada !
Grothier, alors voilà, contre toute attente, il semblerait que je gagne le pari (je touche du bois). Merci pour tous les mots doux qu'on s'est échangé, pour les discussions profondes de fin de journée, les appels anonymes ou autre blagues moustiquaires. Merci kidnappeur kidnappeur, pour le beignet salé. Merci pour ces trois années de buggage. Bref, un jour, peut-être, je paierais ma dette !

Un grand merci à Nina, d'avoir été là pendant ces trois années. La thèse aurait été bien différente sans toi je pense, tu as été un soutien moral essentiel. Merci pour les covoiturages/ragots, les sorties, les soirées, l'EVJ (« c'est correct ?! »), votre mariage, les séances sport ; bref merci pour tous ces bons moments passés ensemble ! J'ai hâte de vous revoir, sous le soleil, avec Martoune et de rencontrer le mini-vous !

Et à tous les autres que je n’ai pas cité mais que je n’oublie pas !

Merci enfin à mes parents, de m’avoir toujours soutenue et permis de faire ce que je voulais. Sans vous, je n’en serais pas là. Merci à Arthur, pour ton soutien de grand frère à distance. Merci pour vos encouragements.

Publications et communications

Publication

Plumel, M.I., Benhaim-Delarbre M, Rompais M, Thiersé D, Sorci G, van Dorsselaer A, Criscuolo F, Bertile F., *Differential proteomics reveals age-dependent liver oxidative costs of innate immune activation in mice.* J Proteomics, 2016. **135**: p. 181-90.

Communications orales

Margaux Benhaim-Delarbre, Marine Plumel, Alain van Dorsselaer, Gabriele Sorci, François Criscuolo, and Fabrice Bertile (2015). **Label-free-based quantification of splenic immunosenescence markers**, Mass Spectrometry in Biotechnology and Medicine Summer School, Dubrovnik, Croatie.

Margaux Benhaim-Delarbre, Georg Tascher, Mikko Lehto Hürlimann, Alain van Dorsselaer, Pierre Bize, and Fabrice Bertile (2015). **The role of brown adipose tissue against obesity: proteomics in a non-sequenced vole model**, Congrès de Spectrométrie de Masse et d'Analyse Protéomique, Ajaccio, France.

Margaux Benhaim-Delarbre, Blandine Chazarin, Morgane Maillard, Jérémy Terrien, and Fabrice Bertile (2016). **Liver proteome changes in a seasonal primate reveal possible mechanisms underlying the safe uncoupling of obesity and insulin resistance**, Congrès de la Société Française d'Electrophorèse et d'Analyse Protéomique, Chambéry, France.

Communications par affiche

Margaux Benhaim-Delarbre, Georg Tascher, Mikko Lehto Hürlimann, Alain Van Dorsselaer, Pierre Bize, and Fabrice Bertile (2014). **De novo sequencing and quantitative proteomics in a non-sequenced species as a way to discover new anti-obesity mechanisms**, Annual World Congress of Human Proteome Organization (HUPO), Madrid, Espagne.

Margaux Benhaim-Delarbre, Marine Plumel, Alain van Dorsselaer, Gabriele Sorci, François Criscuolo, and Fabrice Bertile (2015). **Label-free-based quantification of splenic immunosenescence markers**, Mass Spectrometry in Biotechnology and Medicine Summer School, Dubrovnik, Croatie.

Margaux Benhaim-Delarbre, Georg Tascher, Mikko Lehto Hürlimann, Alain van Dorsselaer, Pierre Bize, and Fabrice Bertile (2015***). The role of brown adipose tissue against obesity: proteomics in a non-sequenced vole model***, Congrès de Spectrométrie de Masse et d'Analyse Protéomique, Ajaccio, France.

Margaux Benhaim-Delarbre, Barbara Henning, Georg Tascher, Pierre Bize, and Fabrice Bertile (2017). **Quantitative proteomics in a non-sequenced vole model: insights into the role of brown adipose tissue against obesity**, American Society for Mass Spectrometry annual conference, Indianapolis, Etats-Unis d'Amérique.

Liste des principales abréviations

2D-DIGE: Differential In-Gel Electrophoresis
Ac: Anticorps
ACP: Analyse en Composante Principale
AUC: Area Under the Curve
BAT: Brown Adipose Tissue
BLAST: Basic Local Alignment Search Tool
CID: Collision Induced Dissociation
Da: Dalton
DDA: Data Dependent Acquisition
DIA: Data Independent Acquisition
ESI: ElectroSpray Ionisation
FDR: False Discovery Rate
GO: Gene Ontology
HPLC: High Performance Liquid Chromatography
IL10: Interleukine 10
KEGG: Kyoto Encyclopedia of Genes and Genomes
LC: Liquid Chromatography
LPS: LipoPolySaccharide
LV: Lande à Vaccinium
MS: Mass Spectrometry
MS/MS ou MS2: Tandem Mass Spectrometry
NST: Non-Shivering Thermogenesis
PBS: Phosphate Buffered Saline
PF: Prairie à Fétuques
pI: Point Isoélectrique
PRM: Parallel Reaction Monitoring
PSM: Peptide-Spectrum Match
QQQ: Triple Quadrupole
SRM: Selected Reaction Monitoring
TOF: Time Of Flight
UCP1/UCP2: Uncoupling protein 1/2
UPLC: Ultra high Performance Liquid Chromatography
VM: Valeur Manquante
XIC: eXtracted Ion Current

Sommaire

Introduction générale

Introduction générale

En écologie évolutive, la théorie des traits d'histoire de la vie stipule que l'évolution façonne la biologie des organismes à partir de facteurs individuels (e.g. l'âge, le genre ou la génétique et leurs interactions), environnementaux (e.g. température, humidité, accès aux ressources, exposition à es pathogènes), et écologiques (e.g. investissement dans la reproduction, activation du système immunitaire) qui conditionnent leur physiologie, survie et reproduction [1]. Face à son environnement, un organisme exprime donc un ensemble de traits adaptatifs spécifiques [2]. Or il n'est pas possible de maximiser tous ces traits, chacun étant contraint par l'énergie (disponible en quantité limitée) investie pour les autres.

Ainsi, la biologie évolutive s'intéresse à la diversité des organismes présents sur Terre et vise à comprendre ce qui cause une telle diversité. La biologie évolutive s'interroge sur les différences entre populations, entre espèces et entre individus pour comprendre quels mécanismes ont permis, au cours de l'évolution, de telles différences [1].

Une question centrale de l'écologie évolutive est de comprendre les contraintes qui ont mené aux compromis du cycle biologique, notamment les compromis énergétiques. La compétition entre les différentes fonctions de l'organisme (croissance, reproduction, système immunitaire, maintenance, etc.) pour des ressources limitées implique une variabilité en termes d'efficacité de ces fonctions [3]. Ces compromis réalisés par l'individu dépendent de son histoire de vie qui module alors la composante génétique. Dans ce contexte, un individu qui allouera beaucoup d'énergie à la reproduction, ou à la fuite de prédateurs, verra la contribution aux mécanismes de réparation diminuer, les erreurs s'accumuleront et deviendront délétères : on parle de « coûts » inhérents aux compromis énergétiques.

La protéomique consiste à analyser qualitativement et quantitativement le protéome, i.e. l'ensemble des protéines exprimées dans un tissu, une cellule ou un organisme à un instant donné et dans des conditions données. Ce domaine s'est largement imposée dans les sciences de la vie grâce à une forte évolution technique au cours des 20 dernières années [4]. La spectrométrie de masse, outil indispensable à la protéomique, s'est largement développée et permet aujourd'hui, grâce à des appareils toujours plus résolutifs, sensibles et précis, de détecter et quantifier des milliers de protéines dans des échantillons très complexes [5]. Le remplissage croissant des banques de données, essentielles à l'identification des protéines, ainsi que les outils bio-informatiques qui permettent le traitement des données ont fortement contribué au développement de la protéomique. Malgré ces nombreux développements, le traitement de données reste une difficulté majeure en protéomique, il n'existe pas une solution universelle à toutes les problématiques.

L'utilisation de la protéomique en biologie évolutive reste émergente [6], par rapport à la génomique et la transcriptomique ; environ 10 fois plus d'études d'écologie évolutive sont publiées avec de l'outil transcriptomique versus la protéomique [6]. L'intérêt pour la protéomique dans ce domaine est multiple, le protéome étant plus proche du phénotype que le génome ou le transcriptome, pourrait donc être plus sensible à la sélection naturelle et donc permettre de mieux comprendre les

mécanismes adaptatifs. De plus, les avancées technologiques dans ce domaine permettent aujourd'hui d'aborder des questions de biologie évolutive, notamment chez des organismes « exotiques », non modèles [7, 8]. C'est dans ce contexte que les méthodes de protéomique ont été optimisées/développées dans le cadre de cette thèse.

De l'organe à l'écosystème, nous étudions les variations protéomiques induites par des changements environnementaux. Il s'agit de mieux comprendre les mécanismes adaptatifs de certaines espèces, en allant du modèle « classique » de la souris de laboratoire à des modèles plus exotiques tels que le campagnol ou la fourmi, leur permettant de supporter des situations qui seraient délétères pour l'homme. A part pour la souris de laboratoire, le génome de certains modèles animaux étudiés (e.g. campagnol, microcèbe, écosystème) n'est pas encore séquencé. Ceci implique, outre la nécessité d'adapter les méthodologies analytiques à chaque type d'échantillon, de développer de nouveaux outils bio-informatiques pour pouvoir analyser les données.

Ce manuscrit s'articule en deux parties :

Dans une première partie sera présenté un état de l'art de la protéomique. Allant de la préparation d'échantillons à l'analyse des données, en passant par une description des spectromètres de masse, cette partie, organisée en trois chapitres, a pour but de présenter au lecteur les bases et outils de l'analyse protéomique.

Dans une seconde partie seront présentés les résultats obtenus au cours de ces travaux de thèse sur différentes problématiques biologiques. Chaque projet a été l'occasion de répondre à divers défis méthodologiques.

❖ Le premier chapitre de la partie « Résultats » présente l'analyse d'un organisme modèle : la souris. Trois projets sont présentés dans ce chapitre, qui concernent l'étude du coût de l'activation du système immunitaire. Deux projets prennent en compte l'âge des individus, et permettent donc d'étudier le phénomène d'immunosenescence ; tandis que pour le dernier projet, on impose un contrôle du système immunitaire. Ces projets ont permis de mettre en place différentes stratégies de préparation d'échantillons ainsi que différentes approches quantitatives en protéomique.

❖ Le deuxième chapitre décrit les développements mis au point pour le traitement de données issues de l'analyse d'organismes dont le génome n'est pas séquencé.

- Un premier projet a été l'occasion d'importants développements bio-informatiques pour le séquençage *de novo*. Ce chapitre présentera dans un premier temps l'amélioration apportée au paramétrage du logiciel « Pepnovo », qui est passée par des tests pour mieux comprendre son fonctionnement, dans le but d'obtenir des résultats plus fiables et robustes. Dans un second temps, on présentera le développement qui a permis la quantification de ces peptides, ce qui n'était pas possible auparavant. Enfin, nous appliquerons ces développements à l'étude du rôle du tissu adipeux bruns (BAT) dans le contrôle de l'obésité chez les campagnols (organisme dont le génome n'est pas séquencé). Ces rongeurs, lorsqu'ils possèdent une forte

activité du BAT résistent mieux à une diète obésogène que leurs congénères dont l'activité de BAT est faible. L'objectif est de comprendre, grâce à ce modèle unique, les mécanismes sous-jacents à la protection contre l'obésité, en étudiant le protéome de ces animaux soumis à différentes contraintes environnementales (régime alimentaire, température).

- Un second projet visait à déterminer l'apport du séquençage *de novo* par rapport à une dé complexification des échantillons en amont de l'analyse LC-MS/MS. Le but ici était d'obtenir un maximum de couverture de protéome chez une espèce non séquencée : le microcèbe. Ce petit primate est un modèle prometteur d'obésité réversible. Il arrive à s'adapter en fonctions des ressources disponibles, avec une phase obésogène suivie d'une phase diabétogène en hiver, sans jamais atteindre de seuils pathologiques.

- Un dernier projet nous a permis d'étudier les interactions entre socialité et mécanismes de vieillissement chez les fourmis. En effet, au sein d'une même espèce, on observe d'importants écarts de longévité en fonction de la caste : les reines peuvent vivre jusqu'à 20 ans, tandis que les ouvrières vivent entre 2 et 4 ans.

❖ Le troisième chapitre traite des développements mis au point pour l'analyse d'un métaprotéome : le sol alpin. Il s'agissait de comparer deux sols qui différent par leur couverture végétale dans le but de déterminer si cette différence peut être liée à une différence au niveau des communautés biotiques du sol, en en dressant un profil « méta-fonctionnel ». Pour cela, le premier défi présenté est l'extraction des protéines à partir d'une telle matrice. Le deuxième défi majeur résidait dans l'analyse des données, ce chapitre présentera donc les étapes de détermination de la banque de données la plus appropriée et de validation des identifications pour obtenir un maximum de couverture de ce métaprotéome.

En conclusion, nous ferons un bilan sur ces travaux et la protéomique en général.

Partie I : Introduction à la protéomique

Chapitre I : Analyse protéomique par spectrométrie de masse

Partie I : Introduction à la protéomique

Chapitre I : Analyse protéomique par spectrométrie de masse

La protéomique consiste à identifier, caractériser et quantifier l'ensemble des protéines exprimées dans une cellule ou un tissu à un moment donné et dans des conditions données. L'approche la plus couramment utilisée en protéomique est l'approche « bottom-up » [9]. Elle consiste à digérer à l'aide d'une enzyme les protéines et à analyser les peptides de digestion par chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS). L'utilisation de la trypsine, qui génère des peptides basiques, relativement petits (en moyenne une dizaine d'acides aminés) et facilement ionisables, et dont la digestion est reproductible, a rendu cette approche très populaire. Les peptides sont ensuite identifiés à partir de l'interprétation des spectres MS et MS/MS acquis grâce à des banques de données, ce qui permet ensuite de remonter à l'identification des protéines. C'est pourquoi cette approche est appelée « bottom-up », on part des fragments des peptides pour remonter aux peptides et ensuite aux protéines (du bas vers le haut).

Rapidement, il existe également une approche « middle-down » [10], qui diffère de l'approche précédente par l'enzyme utilisée, qui génère des peptides plus gros, ou polypeptides, dont la masse va jusqu'à 20kDa.
Et enfin, la dernière approche, appelée « top-down » [11] consiste à analyser les protéines entières. Celles-ci sont identifiées grâce aux spectres de masse MS et MS/MS, qui sont très complexes, étant donné le nombre d'états de charge que peuvent prendre les fragments de protéines. Cette méthode est donc plus adaptée à des échantillons non complexes, très purifiés.

Dans la suite de cette partie I seront décrites les différentes étapes de l'approche « bottom-up », qui a été utilisée dans les travaux présentés dans ce manuscrit. Ces étapes sont présentées en Figure I-1.

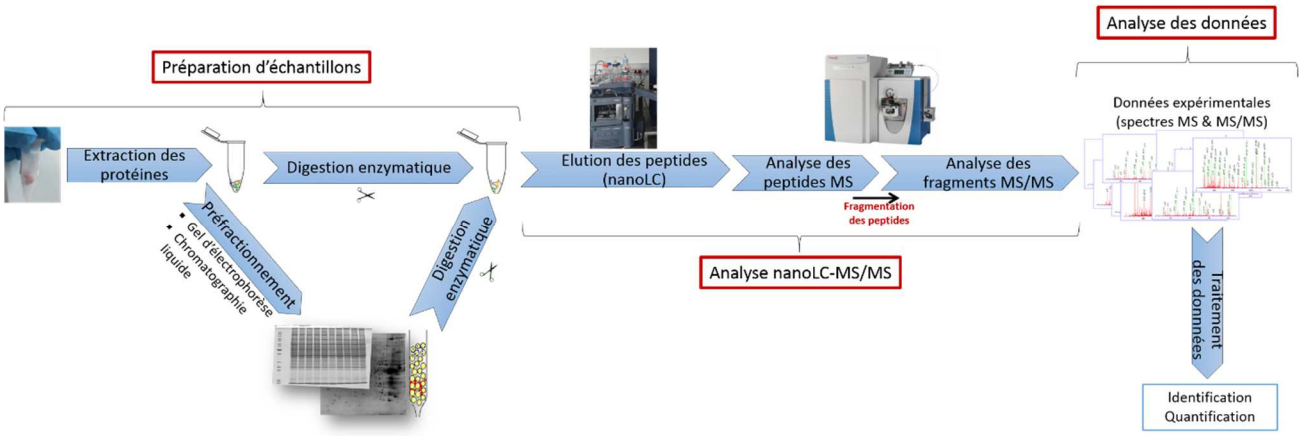


Figure I-1 : Représentation schématique des grandes étapes de l'analyse protéomique en « bottom-up ».
Les deux premières étapes (préparation d'échantillons et analyse nanoLC-MS/MS) sont présentées dans ce chapitre, la dernière étape (analyse des données) dans les chapitres II et III.

I. Préparation d'échantillons

A. Extractions des protéines

Les échantillons biologiques constituent une matrice très complexe, il est donc souvent essentiel de dé-complexifier les extraits protéiques après extraction. Les protéines sont extraites en provoquant la lyse de la cellule, mécaniquement (broyage, sonication, etc.) ou chimiquement (lyse enzymatique, choc osmotique, etc.), ce qui entraîne la libération des protéines mais aussi d'interférents (sels, lipides, etc.). Les protéines sont donc ensuite isolées, généralement par précipitation [12] (à l'acétone ou à l'acide trichloroacétique par exemple), et les protéases inhibées afin de préserver les protéines.
Il est possible de concentrer les espèces les moins abondantes, en éliminant les protéines majoritaires qui ne sont pas d'intérêt par rapport à la question posée, par immunodéplétion par exemple. De même, certaines protéoformes peuvent être enrichies, comme par exemple les protéines portant certaines modifications post-traductionnelles, grâce à leur affinité spécifique pour certains ligands.
Les protéines sont finalement digérées (à l'aide d'une enzyme, généralement la trypsine), soit directement après extraction et précipitation, soit après préfractionnement.
La dé-complexification des échantillons peut s'obtenir au niveau protéique et/ou au niveau peptidique, par électrophorèse ou chromatographie par exemple.

B. Séparation des protéines par électrophorèse

Par gel d'électrophorèse, les protéines peuvent être séparées selon une ou deux dimensions (i.e. selon deux propriétés physico-chimiques) suivant le niveau de dé complexification souhaité.

❖ Electrophorèse monodimensionnel (1D ou SDS-PAGE) : Les protéines sont séparées en fonction de leur masse moléculaire dans un gel de polyacrylamide. Après extraction, les protéines sont reprises dans un tampon contenant du détergent SDS (Sodium Dodecyl Sulfate) qui permet de conférer une charge négative aux protéines. Ainsi, sous l'effet d'un champ électrique, les protéines vont entrer dans le gel et avancer entre les mailles de polyacrylamide, ce qui permettra la séparation (les plus petites protéines migreront plus vite dans les mailles du gel). Les protéines sont ensuite colorés,

généralement au bleu de Coomassie [13, 14]. Suivant le degré de séparation souhaité, il est possible de faire varier la réticulation du gel (i.e. le pourcentage d'acrylamide utilisé, pour obtenir des mailles plus ou moins grandes) afin de mieux séparer des gammes de masse moléculaires spécifiques. La migration peut aussi être plus ou moins longue, allant d'une seule bande protéique (le but ici est surtout d'éliminer les détergents utilisés pour l'extraction des protéines, incompatibles avec la masse), jusqu'à une trentaine de bandes protéiques. Plus on sépare, plus l'analyse par spectrométrie de masse sera longue.

Un exemple d'image de gel 1D est présenté en Figure I-2.

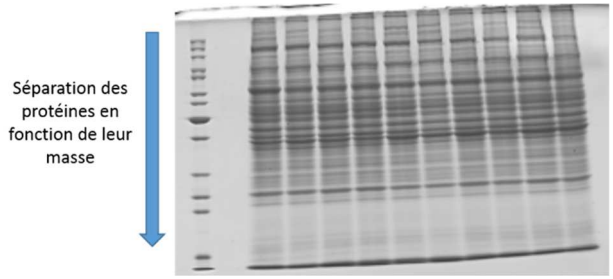


Figure I-2 : Image d'un gel d'électrophorèse monodimensionnel.
Les molécules de plus petit poids moléculaires se trouvent en bas du gel car elles migrent plus vite.

❖ Electrophorèse bidimensionnel (2D) [15, 16]: Les protéines sont séparées en fonction de leur point isoélectrique (pI) dans une première dimension puis selon leur masse dans une deuxième dimension.

Selon la première dimension, les protéines sont déposées sur une bande de gel d'acrylamide fonctionnalisée avec des immobilines (acides et bases faibles qui créent un gradient de pH sur la bande), sous l'effet d'un champ électrique les protéines migrent donc jusqu'à arriver à l'endroit où le pH est égal à leur pI, puisqu'alors elles sont globalement neutres [17]. La séparation sur la deuxième dimension est la même que pour le gel 1D.

L'intérêt du gel bidimensionnel est sa grande capacité résolutive, en effet, on retrouve généralement moins d'une dizaine de protéines dans chaque spot protéique ; tandis que sur un gel 1D les bandes protéiques restent encore très complexes (plusieurs centaines de protéines, suivant le niveau de séparation). En revanche, analyser chacun des spots peut se révéler très long, surtout si l'on a beaucoup d'échantillons. De plus, cette méthode donne accès à des informations physico-chimiques des protéines, comme leur masse moléculaire ou leur point isoélectrique [18]. Enfin, on peut réaliser une quantification relative (directement sur gel), appelé 2D-DIGE [19] dont le principe est expliqué en page 43.

Un exemple d'image de gel 2D est présenté en Figure I-3.

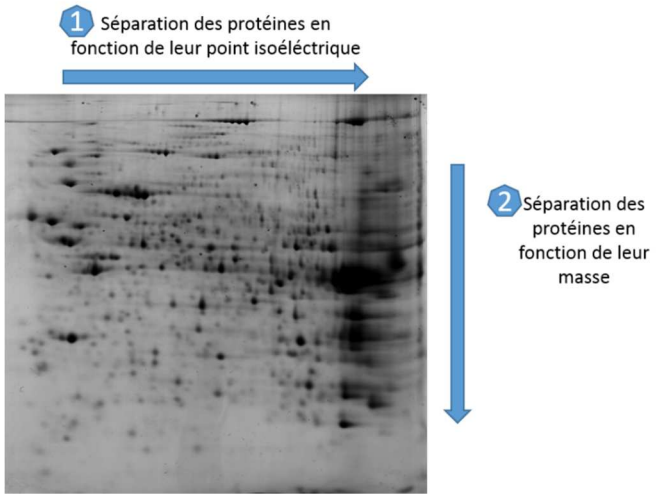


Figure I-3 : Image d'un gel d'électrophorèse bidimensionnel.

L'intérêt de préfractionner les échantillons est de pouvoir analyser des mélanges moins complexes en spectrométrie de masse et donc de pouvoir détecter plus de peptides en réduisant la compétition pour la fragmentation. Cependant, préfractionner revient à multiplier le nombre d'analyses pour un seul échantillon, ce qui peut rallonger considérablement le temps de l'analyse globale.

II. Séparation par chromatographie liquide

La chromatographie liquide à haute performance (HPLC) est utilisée pour dé complexifier les échantillons [20], qui contiennent des dizaines de milliers de peptides issus de la digestion enzymatique de milliers de protéines. La séparation a pour but de limiter la compétition en masse entre les peptides, ce qui permet d'augmenter la sensibilité, la sélectivité et la profondeur du protéome analysé. D'autres techniques sont utilisées, telle que l'électrophorèse capillaire.

La LC permet de séparer les composés en fonction de leur affinité avec une phase stationnaire et une phase mobile. En protéomique, la chromatographie en phase inverse est la plus couramment utilisée, c'est à dire une phase stationnaire de silice greffée au C18 (apolaire) et une phase mobile composée de deux solvants (généralement eau et acétonitrile), permettant de jouer sur l'hydrophobicité de la phase mobile et ainsi de séparer les peptides en fonction de leur hydrophobicité. Généralement pour une meilleure séparation on utilise un gradient de solvant ; les peptides les plus hydrophobes (plus retenus par la phase stationnaire) sont ainsi élués au fur et à mesure que le pourcentage d'acétonitrile est augmenté.

Plusieurs paramètres peuvent influencer la qualité de séparation (ou résolution): plus la colonne est longue, meilleure est la séparation (mais le temps d'analyse est également rallongé) ; la réduction du diamètre interne de la colonne permet de gagner en sensibilité en réduisant les volumes de solvants ; les phases stationnaires de faible granulométrie permettent de gagner en résolution. En protéomique, on utilise généralement des systèmes UPLC : chromatographie liquide ultra performance, avec des phases stationnaires de granulométrie inférieure à 2µm, des colonnes de 25cm de longueur et 75µm de diamètre interne. Enfin la composition de la phase mobile ainsi que la durée et pente du gradient permettent également d'améliorer la séparation des peptides.

Il est possible de réaliser une séparation par chromatographie bidimensionnelle en couplant deux types de chromatographie [21], ce qui permet de séparer les peptides selon deux propriétés physico-chimiques différentes, et donc d'obtenir une meilleure séparation. On peut notamment utiliser des phases échangeuses d'ions (IEX) qui sépare les peptides en fonction de leur charge ; une phase HILIC (*Hydrophilic Interaction Liquid Chromatography*) plutôt adapté à la séparation des peptides hydrophiles (comme les peptides glycosylés) ; ou encore des phases qui retiennent spécifiquement certains acides aminés par chélation de métaux (e.g. IMAC, *Immobilized Metal Affinity Chromatography*) plutôt adapté pour les peptides phosphorylés.

III. Analyse par spectrométrie de masse

En protéomique, le système chromatographique est couplé à un spectromètre de masse afin d'analyser en ligne les peptides séparés.
Un spectromètre de masse permet de déterminer le rapport masse sur charge (m/z) des analytes. Comme présenté en Figure I-4, il est composé d'une source d'ionisation, qui permet d'ioniser les peptides à l'entrée du spectromètre de masse pour leur permettre d'être détectés ; d'un ou plusieurs analyseurs qui vont permettre de séparer les ions et déterminer leur rapport m/z ; et d'un détecteur dont le rôle est d'enregistrer le signal des ions. Peut s'ajouter à ces éléments une cellule de collision, qui permet de fragmenter les ions ; les fragments seront alors analysés dans un deuxième temps (spectrométrie de masse en tandem ou MS/MS).

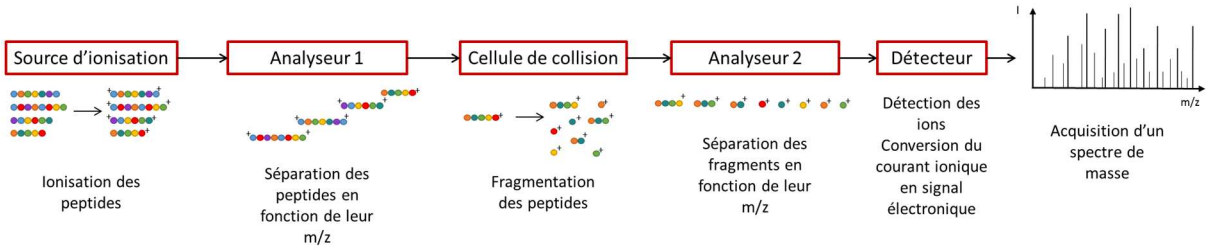


Figure I-4 : Représentation schématique des différentes parties composant un spectromètre de masse en tandem.

A. Les sources d'ionisation

En protéomique, on utilise des sources dites douces, l'énergie transmise aux molécules ne risquent pas de les détruire: la source MALDI (Matrix-Assisted Laser Desorption Ionisation) [22] et la source ESI (ElectroSpray Ionisation) [23]. Les peptides sont ionisés directement en sortie de colonne. Pour tous les projets présentés dans ce manuscrit, une source ESI a été utilisée.

La source ESI permet de générer des ions polychargés $[M + nH]^{n+}$ ou $[M - nH]^{n-}$ à pression atmosphérique et sous l'effet d'un champ électrique.
Les analytes sous forme liquide arrivent dans une aiguille sur laquelle est appliquée une forte tension. Sous l'effet du champ électrique, la goutte qui est émise va prendre la forme d'un cône (appelé cône de Taylor) à la surface duquel il y a un excès de charges positives (voir Figure I-5). Lorsque les forces coulombiennes atteignent la tension de surface (limite de Rayleigh) à la surface du cône, des gouttelettes chargées se forment. Au fur et à mesure que ces gouttelettes avancent vers l'entrée

chauffée du spectromètre de masse, le solvant s'évapore, augmentant encore une fois la tension de surface jusqu'à atteindre la limite de Rayleigh et générer de multiplies gouttelettes filles chargées.

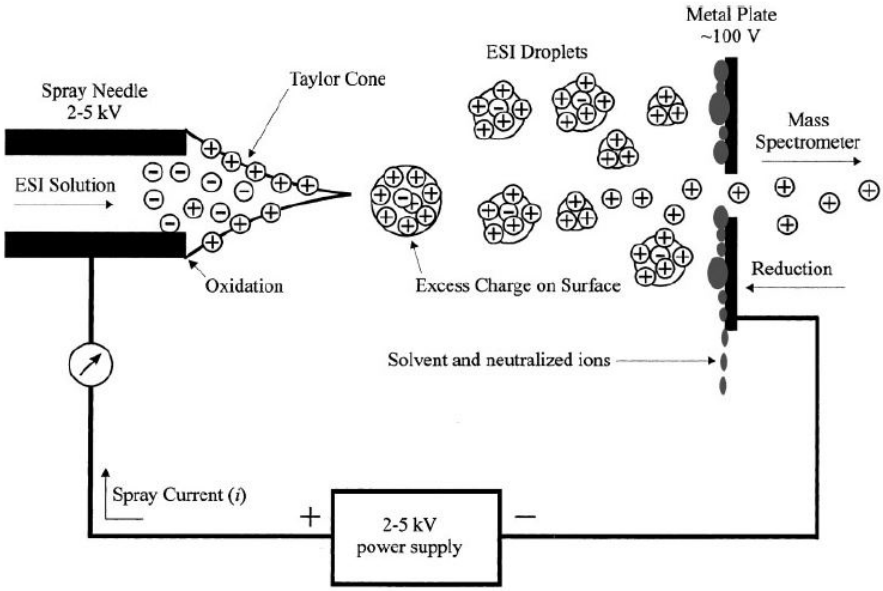


Figure I-5 : Représentation schématique de l'ionisation par electrospray [24].

B. Les analyseurs

Une fois les analytes ionisés, un système de focalisation va permettre de diriger les ions vers l'analyseur, ainsi que d'éliminer les molécules non chargées ou non désirées (certains états de charge par exemple). Ce système de focalisation diffère d'un instrument à un autre, il peut être constitué de lentilles ou de multipôles de différentes géométries.

Le but de l'analyseur est de séparer les ions et déterminer leur rapport masse sur charge (m/z). Il en existe différents types, les principaux utilisés aujourd'hui en protéomique étant le quadripôle (Q), le temps de vol (TOF), la trappe ionique (IT), l'analyseur à transformée de Fourier (FT-ICR) et la trappe orbitale (Orbitrap™). Les analyseurs sont caractérisés par :

- La gamme de masse, qui correspond à l'intervalle entre le plus petit m/z et le plus grand m/z détectable.
- La précision de masse, qui correspond à l'erreur sur la mesure du rapport m/z. Il est important de connaître la précision de masse de l'appareil utilisé pour l'identification des peptides. Bien sûr, plus l'appareil est précis en masse, plus l'identification d'un composé sera pertinente.
- La sensibilité, qui correspond à la plus petite quantité détectable. Etant donné la forte gamme dynamique des échantillons analysés en protéomique, une bonne sensibilité est indispensable.
- La résolution, qui correspond à la capacité de l'instrument à séparer deux peptides de masses voisines. La résolution est définie comme le rapport $\frac{M}{\Delta M}$; M étant le rapport masse sur charge (m/z) et ΔM la largeur à mi-hauteur du pic. Ainsi, une résolution de 1000 pour un m/z à 100 signifie que l'analyseur pourra séparer les composés à 100 et 100,1 m/z.

Pour un même peptide, la coexistence de différentes formes composées de différents isotopes (légers ou lourds) de certains atomes implique que le spectromètre ne détecte pas une seule masse mais un

massif isotopique. Les pics de ce massif sont appelés P, P+1, P+2, P+3… P est le pic monoisotopique, c'est-à-dire qu'il est composé des isotopes légers de chaque élément (^{1}H, ^{12}C, ^{14}N, ^{16}O…). Les pics suivants (P+1, P+2…) correspondent à l'incorporation d'au moins un isotope lourd. Plus l'appareil est résolutif, plus il sera capable de définir le massif isotopique des ions. A partir du massif isotopique, on peut retrouver la charge (z) d'un ion. La différence de masse (m) entre les isotopes est de 1Da, donc les pics du massif sont séparés de 1/z m/z, ainsi on peut retrouver la charge d'un ion par déconvolution ; pour un peptide chargé 2 fois (z=2), la différence entre les isotopes sera de 0,5 m/z (1/2), si z=3 la différence sera de 0,33 m/z (1/3) etc.

Dans les travaux présentés dans ce manuscrit, les analyseurs utilisés étaient le quadripôle, le temps de vol et la trappe orbitale, seuls ces analyseurs seront décrits ici.

B.1. Quadripôle

Le quadripôle [25] est composé de quatre électrodes cylindriques connectées deux à deux. Un potentiel électrique Φ est appliqué aux électrodes, composé d'une tension continu U et d'une tension alternative de radiofréquence Vcosωt, tel que Φ = ± (U – Vcosωt). Deux électrodes adjacentes sont de potentiels opposés (voir Figure I-6).

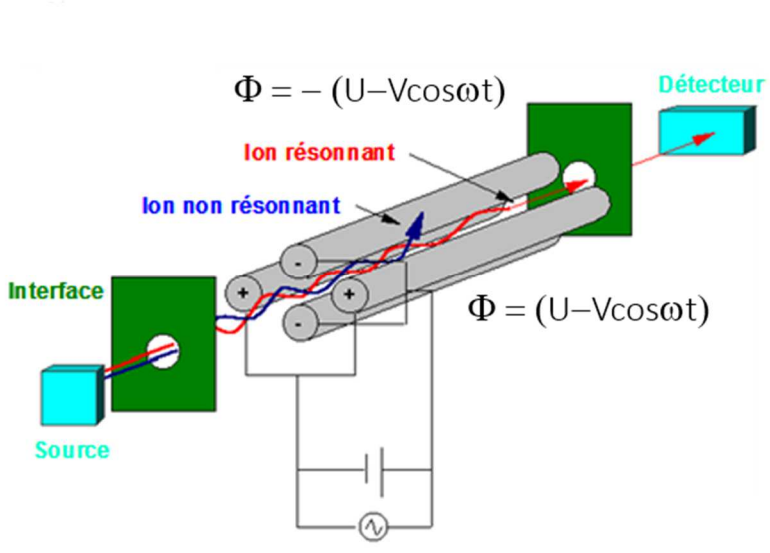


Figure I-6 : Représentation schématique du fonctionnement d'un analyseur quadripolaire. Figure emprunté de LSMBO.

Les ions résonnants, qui ont une trajectoire stable à des tensions U et V donnés traversent le quadripôle pour atteindre le détecteur, tandis que les ions instables sont éjectés.

Les ions pénétrant dans le quadripôle subissent ce champ électrique oscillant et ont une trajectoire plus ou moins stable, décrite par les équations de Mathieu, en fonction des potentiels U et V. Seuls les ions stables pourront traverser le quadripôle pour atteindre le détecteur [26]. La zone de stabilité pour un ion donné est présentée en Figure I-7.

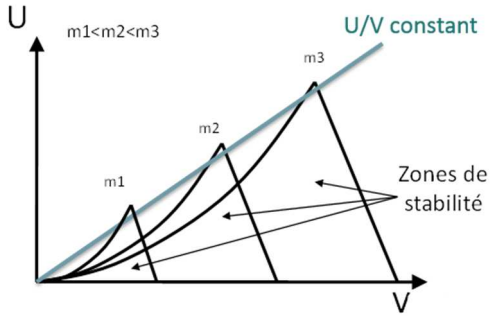


Figure I-7 : Zone de stabilité d'un m/z donné en fonction des potentiels U et V au sein d'un quadripôle.

En faisant varier les tensions U et V, et en maintenant le rapport U/V constant (droite bleue sur la Figure I-7), il est possible de stabiliser et donc de transmettre des masses de plus en plus élevées, dans le but de séparer et d'analyser un à un tous les ions qui arrivent à l'analyseur, il s'agit du mode « Full scan ». La pente de la droite définit la résolution : plus la pente est proche des apex des domaines de stabilité des masses, plus l'analyseur est résolutif (car la fenêtre de sélection est réduite), en revanche, si la pente est plus faible, plus d'ions seront sélectionnés, ce qui augmente la sensibilité. Un autre mode, « RF only », qui consiste à appliquer une tension U nulle, permet de transmettre tous les ions, ce mode est utilisé lorsque le quadripôle est couplé à un autre analyseur, il sert alors simplement à focaliser les ions. Il est également possible d'isoler des ions, c'est-à-dire d'appliquer des tensions telles qu'un seul m/z reste stable, dans le but d'analyser spécifiquement un peptide. Ce mode est utilisé en SRM (dont le principe est expliqué en page 42), lorsque l'on veut analyser spécifiquement quelques peptides.

Le quadripôle est généralement utilisé en tandem avec un autre analyseur, du fait de sa faible résolution et vitesse de scan, il sert principalement de focaliseur (« RF only ») en mode MS et de filtre pour la MS/MS.

B.2. Temps de vol

L'analyseur à temps de vol (*Time Of Flight* en anglais, TOF) [27] est composé d'un tube de vol soumis à un vide poussé (de l'ordre de 10^{-7} mbar), comme présenté en Figure I-8.

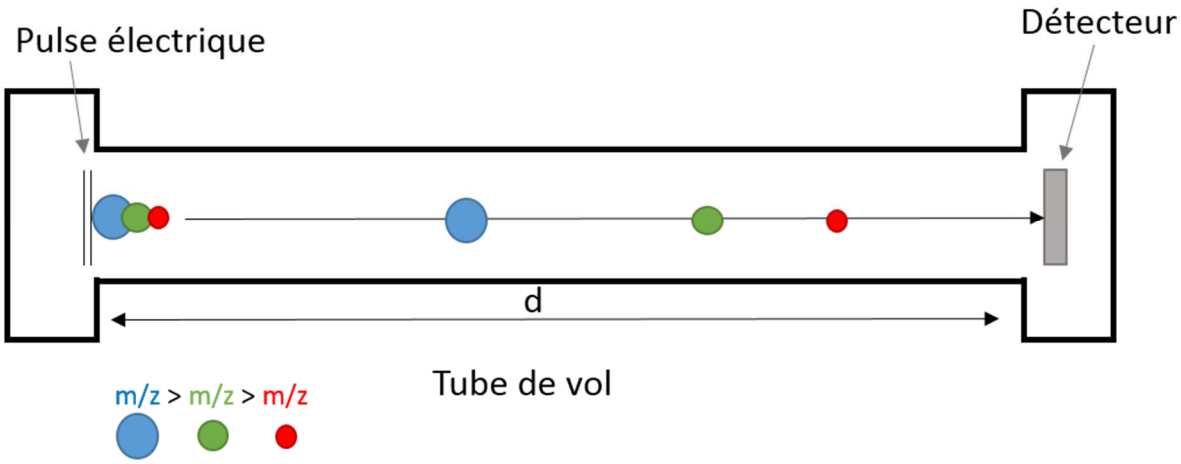


Figure I-8 : Représentation schématique d'un analyseur à temps de vol.
Les trois ions de rapport m/z différents acquièrent une énergie cinétique plus ou moins importante en fonction de leur masse. Les ions les plus rapides (les plus petits) vont arriver en premier au détecteur.

A l'entrée du tube de vol, les ions sont accélérés sous l'effet d'une impulsion électrique puis vont parcourir librement le tube de vol grâce à l'énergie cinétique acquise. Les ions sont ainsi séparés en fonction de leur masse, plus le ratio m/z sera petit, plus l'ion traversera le tube rapidement. On peut alors déterminer le rapport m/z d'un ion en fonction du temps qu'il met à parcourir le tube (dont la distance d est fixe) grâce à l'équation (dérivée de la définition de l'énergie cinétique) :

$$t^2 = \frac{m}{z} \times \frac{d^2}{2eV}$$

Avec t = temps de parcours ; m= masse du peptide ; z= charge du peptide ; d= distance de vol ; e= charge d'un électron ; V= potentiel auquel l'ion est soumis.

Il arrive que des ions de même rapport m/z acquièrent une énergie cinétique différente à l'entrée du tube de vol ; séparant ainsi ces ions et donc diminuant la résolution de l'analyseur. Pour pallier à ce problème, un réflecteur peut être utilisé en milieu de tube de vol [26]. Le réflecteur (ensemble de lentilles) permet de ralentir les ions les plus rapides. Les ions possédant une énergie cinétique plus importante ne vont pas avoir la même trajectoire, et vont pénétrer plus profondément dans le réflecteur avant d'être réfléchi, tandis que les ions de plus faible énergie, et donc moins rapide pénétreront moins. Ainsi tous les ions seront rassemblés, focalisés à l'issue du réflecteur pour atteindre le détecteur en même temps (voir Figure I-9).

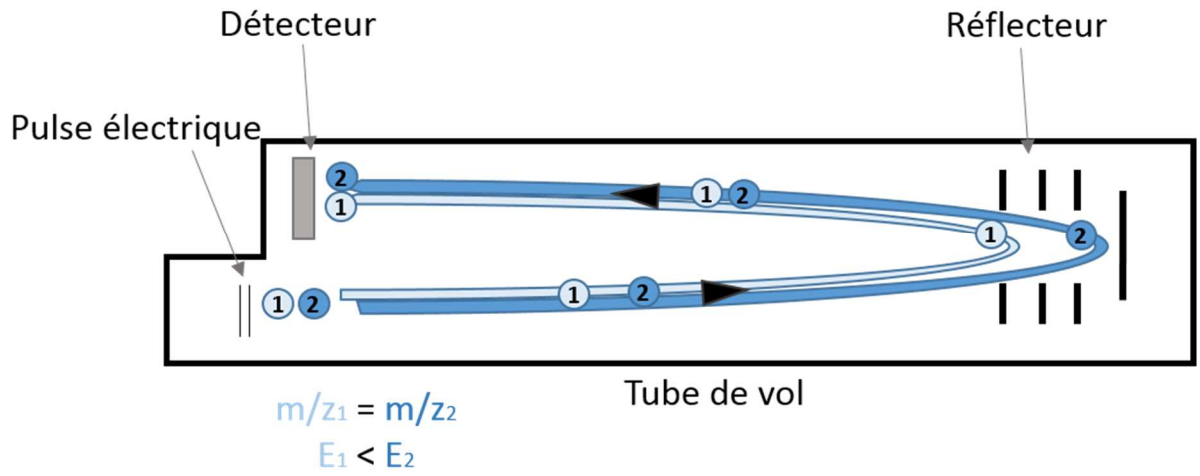


Figure I-9 : Représentation schématique d'un analyseur à temps de vol possédant un réflecteur.
Les deux ions de même rapport m/z ayant acquis une énergie cinétique légèrement différente à l'entrée du tube de vol sont focalisés par le réflecteur pour arriver en même temps au détecteur.

Les analyseurs TOF possèdent une bonne résolution et une vitesse de balayage importante, ils sont beaucoup utilisés en protéomique.

B.3. Trappe Orbitale

La trappe orbitale, plus connue sous son nom commercial Orbitrap™ est un analyseur breveté par la société Thermo Fisher Scientific.

La trappe orbitale [28, 29], présentée en Figure I-10, est composée d'une électrode centrale en forme de fuseau et de deux électrodes externes, le long desquelles est appliqué un champ électrique linéaire. Les ions sont introduits dans la trappe tangentiellement à l'électrode centrale, par paquets grâce à une accumulation en amont dans une trappe linéaire (C-trap). Sous l'effet du champ, les ions vont donc entrer en oscillation harmonique [30] et effectuer un mouvement de rotation autour de l'électrode centrale ainsi qu'un mouvement axial le long de cette même électrode, avec une fréquence qui dépend de leur rapport m/z. Le mouvement axial des ions génère un courant induit enregistré par les électrodes externes. La transformée de Fourier va permettra de calculer la fréquence d'oscillation axiale des ions, qui permet d'accéder à leur rapport m/z grâce à l'équation $\omega = \sqrt{\frac{k}{m/z}}$; avec k= courbure du champ électrique.

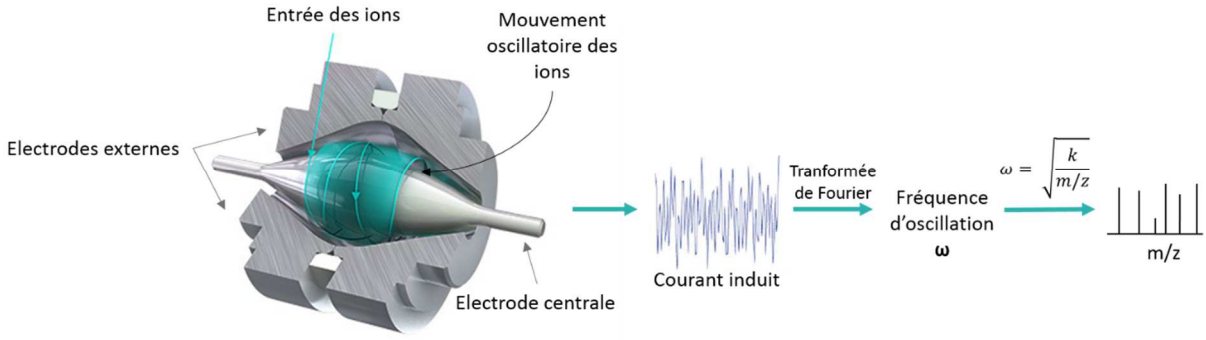


Figure I-10 : Représentation schématique d'une trappe orbitale. Figure adaptée de Thermo Fisher Scientific.
Sous l'effet d'un champ électrostatique, les ions entre en oscillation autour de l'électrode centrale, ce qui génère un courant induit enregistré par les électrodes externes. La fréquence d'oscillation est calculée à partir de ce courant par transformée de Fourier, ce qui permet d'accéder au rapport m/z des ions.

Cet analyseur est précis en masse, rapide et résolutif, son utilisation est donc très répandue en protéomique.

C. Les instruments hybrides

Les instruments hybrides combinent plusieurs analyseurs différents, afin de combiner les propriétés et avantages de chacun et de permettre la fragmentation des peptides. En mode MS, un premier analyseur va servir de focaliseur pour envoyer tous les ions au deuxième analyseur ; puis en mode MS/MS le premier analyseur sélectionne les ions qui seront fragmentés avant d'être analysés dans le deuxième analyseur. Le quadripôle est souvent utilisé comme premier analyseur. Il est souvent couplé à un TOF (Q-TOF) ou une trappe orbitale (Q-Orbitrap™).

❖ Les spectromètres de type Q-TOF [31] sont très répandus en protéomique haut débit, grâce à leur forte résolution (jusqu'à 60 000 pour les instruments dernière génération), leur vitesse de balayage (de 25 Hz ou 25 spectres par seconde) et leur précision de masse inférieure à 5 ppm.

❖ Les instruments de type Q-Orbitrap™ qui, à l'heure actuelle, ne sont commercialisés que par Thermo Fisher Scientific, sont également bien implantés en protéomique. Ils combinent un quadripôle, une trappe linéaire (C-trap) et un Orbitrap™. La C-trap permet de regrouper les ions avant de les envoyer à l'Orbitrap™ améliorant ainsi la résolution qui peut atteindre 140 000 à 200 m/z. Ces appareils, en plus d'être très résolutifs, sont précis en masse (erreur < 5ppm), très sensibles et possèdent une gamme dynamique de plus de 4 ordres de grandeur.

❖ Il existe également des appareils de type triple-quadripôle (QQQ) qui combinent trois quadripôles. Ils sont principalement utilisés pour des analyses ciblées en mode SRM (*Selected Reaction Monitoring*). Rapidement, en SRM on analyse uniquement des peptides d'intérêt et leurs fragments choisis. Le premier quadripôle sélectionne le peptide parent (ou précurseur), qui sera fragmenté dans le deuxième quadripôle (qui sert alors de cellule de collision) tandis que le troisième quadripôle sélectionne un ion fragment à son tour.

D. Les détecteurs

Le détecteur a pour rôle d'enregistrer le signal des ions et de le convertir en signal électrique. Après amplification du signal, ce dernier est digitalisé en valeur numérique par un système électronique codé en bits. Il existe plusieurs types de détecteurs :

❖ Détection par courant induit (voir page 18) comme avec la trappe orbitale ou autres instruments utilisant la transformée de Fourier. Dans ce cas, les ions conservent leur intégrité durant la détection.

❖ Détection par multiplicateur d'électrons, channeltron [32] ou galette de microcanaux (MCP pour *Micro Channel Plate* en anglais) [33]. Les microcanaux sont des surfaces semi-conductrices contre lesquelles les ions vont s'écraser et ainsi générer des électrons qui vont eux-mêmes, au contact d'une nouvelle surface, générer des électrons secondaires et ainsi de suite pour amplifier le signal et ainsi produire un courant électrique. Ce type de détecteur est utilisé dans la majorité des spectromètres de masse (Q-TOF, QQQ).

E. Les modes de fragmentation

La fragmentation des ions permet d'obtenir une information supplémentaire sur le peptide et ainsi obtenir une identification plus robuste. On parle d'ions précurseurs et d'ions fragments.

Il existe différents modes de fragmentation, qui permettent d'obtenir différents type d'ions : le CID (*Collision Induced Dissociation*) [34], l'ECD (*Electron Capture Dissociation*) [35] ou encore l'ETD (*Electron Transfer Dissociation*) [36].

En protéomique, le mode le plus utilisé est le CID car il est le plus adapté à la fragmentation des peptides trypsiques. Seul ce mode sera présenté ici.

Au sein d'une cellule de collision, les ions sont accélérés et entrent en collision avec des molécules de gaz inerte (Hélium, Azote, Argon…). Le processus de fragmentation est basé sur le modèle du proton mobile [37]. Sur les peptides, les charges se situent généralement en position N-terminale et sur les résidus arginine (R) ou lysine (K) (la trypsine clivant spécifiquement en position C-terminale de ces résidus). Suite à la collision, l'énergie cinétique des ions augmente, ce qui entraîne la délocalisation d'un proton mobile (qui est facilement transférable le long du squelette peptidique), et donc la rupture de la liaison amide (C – N qui est la plus faible). Cela est valable pour les peptides au moins deux fois chargés. Dans le cas des peptides monochargés, l'énergie nécessaire pour délocaliser la charge le long du squelette peptidique est trop importante, d'où la difficulté de fragmenter ce type de peptide.
La nomenclature des fragments a été décrite par Biemann [38] et est présentée en Figure I-11. La fragmentation CID permet de former des ions b et y. Tandis que les autres modes favorisent plutôt les ions a et x ou c et z.

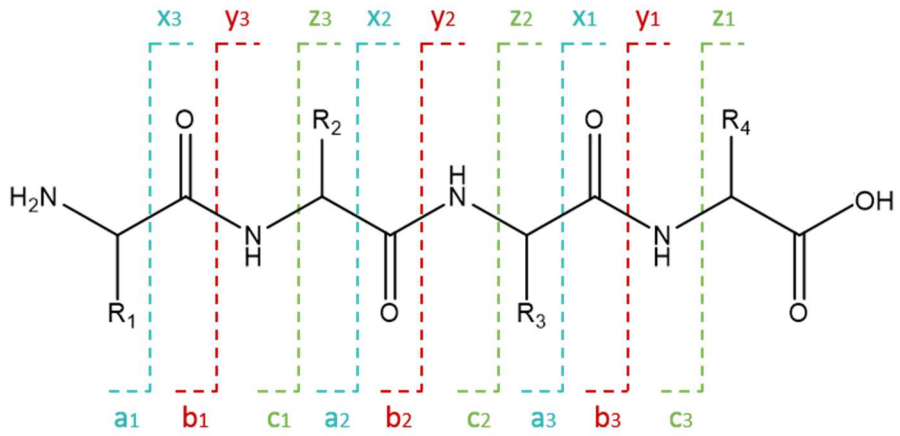


Figure I-11 : Nomenclature de Biemann [38].
Les ions a, b et c contiennent la partie N-terminale du peptide, la charge est portée du côté N-terminal ; les ions x, y et z contiennent la partie C-terminale et la charge est portée du côté C-terminal.

Il existe également un mode HCD (*Higher energy Collisional Dissociation*) [39], qui est un acronyme utilisé par Thermo Fisher Scientific pour décrire le mode de fragmentation utilisé dans les Orbitrap™. Le principe de la fragmentation est exactement le même que celui du CID. Le terme *higher energy* (énergie plus élevée) fait référence à l'énergie nécessaire pour piéger les ions avant leur entrée dans la cellule de collision.

Les spectres de fragmentation sont alors appelés spectres MS/MS ou MS2.

F. Les modes d'acquisition

En analyse protéomique non ciblée, il existe deux modes d'acquisition :

- La DDA (*Data Dependent Acquisition*) ou acquisition dépendant des données. Avec ce mode, tous les ions ne sont pas fragmentés, mais certains précurseurs sont sélectionnés pour la fragmentation. Cette sélection est basée sur l'intensité des ions (c'est pourquoi on dit que l'acquisition dépend des données), le plus souvent on utilise un « top N » c'est-à-dire qu'on sélectionne, les uns après les autres, les N ions les plus intenses d'un spectre MS donné. C'est-à-dire qu'au cours d'un cycle d'acquisition, il y aura 1 acquisition MS suivie de N acquisitions MS/MS, comme présenté en Figure I-12.

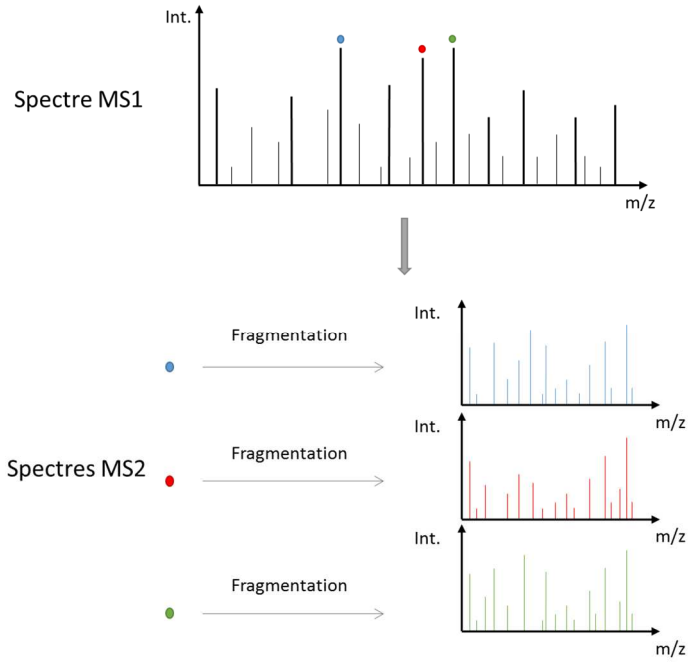


Figure I-12 : Représentation schématique d'une acquisition en mode DDA, avec un « top 3 ».
Les 3 ions les plus intenses sont sélectionnés pour être fragmentés les uns après les autres.

Afin de ne pas toujours fragmenter les mêmes ions (qui feront toujours partie des N plus intenses sur plusieurs spectres MS consécutifs), on peut paramétrer un temps d'exclusion dynamique. C'est-à-dire qu'à partir du moment où un peptide est fragmenté, il ne sera plus sélectionné pendant x secondes, même s'il fait partie du top N. Cela permet de fragmenter également les peptides moins intenses qui co-éluent.

L'inconvénient du mode DDA est de ne pouvoir fragmenter tous les peptides [40]. Cela est dû à la grande complexité des échantillons et donc à la co-élution d'un grand nombre de peptides, qui rentrent en compétition pour la fragmentation. C'est pourquoi il est avantageux de fractionner les protéines, en amont de l'analyse LC-MS/MS, cela permet de réduire cette compétition et donc de détecter et fragmenter davantage de peptides.

❖ La DIA (*Data Independent Acquisition*) ou acquisition indépendante des données. Ce mode d'acquisition permet de pallier aux limitations de la DDA. Il a été développé en 2004 par J.R Yates [41]. Plutôt que de sélectionner des peptides précurseurs, tous les ions présents sur une fenêtre de x m/z (classiquement de 25 Da) sont fragmentés en même temps. Au cours d'un cycle (généralement fixé à 3 sec), toute la gamme de masse est balayée par fenêtres de x Da, comme présenté en Figure I-13.

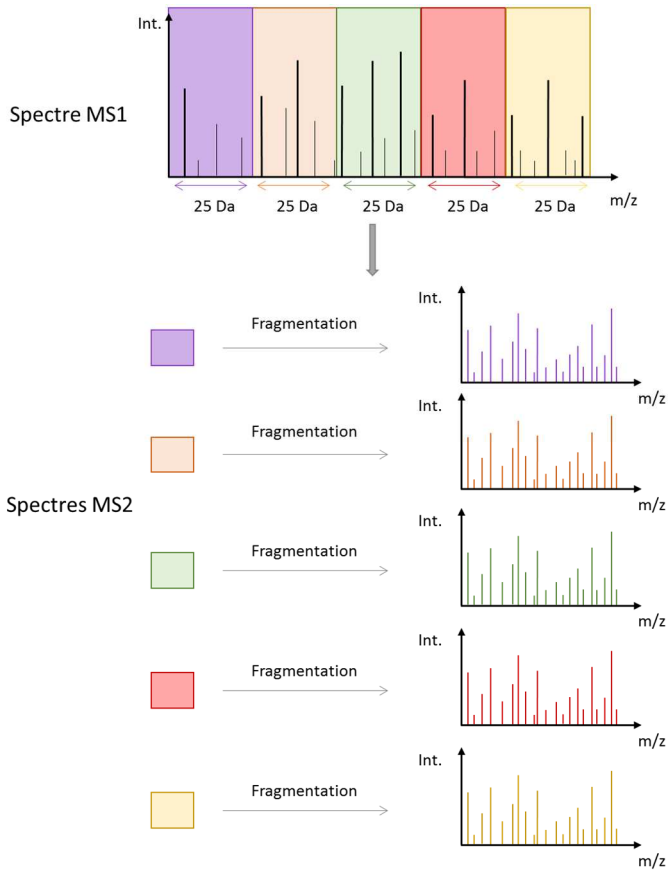


Figure I-13 : Représentation schématique d'une acquisition en mode DIA, avec des fenêtres de 25 Da.
Tous les ions de chaque fenêtre sont fragmentés en même temps. Tout le spectre est fragmenté.

Il est préférable de réaliser au préalable une analyse en mode DDA pour constituer une librairie de spectre et identifier les peptides. En effet, l'identification à partir de spectres issus de la fragmentation de plusieurs peptides reste un défi majeur, même si quelques logiciels commencent à se développer aujourd'hui. Ce mode d'acquisition est surtout utilisé pour la quantification.

Pour les projets présentés dans ce manuscrit, seul le mode DDA a été utilisé.

Partie I : Introduction à la protéomique

Chapitre II : Stratégies d'identification des protéines

Partie I : Introduction à la protéomique

Chapitre II : Stratégies d'identification des protéines

I. Stratégie d'identification chez les espèces séquencées

A. L'empreinte de fragments peptidiques

La stratégie d'identification par empreinte de fragments peptidiques [42] consiste à comparer les données expérimentales à des données théoriques qui se trouvent dans les banques de données. Les protéines contenues dans ces banques sont digérées *in silico* selon les règles de digestion de l'enzyme utilisée et les peptides sont ensuite fragmentés *in silico* selon le mode de fragmentation utilisé. Les masses des peptides et fragments expérimentaux sont ensuite comparés aux masses des peptides et fragments théoriques (voir Figure II-1) grâce à un algorithme de recherche [43]. Plus un peptide est identifié avec un grand nombre de fragments, plus l'identification est robuste.

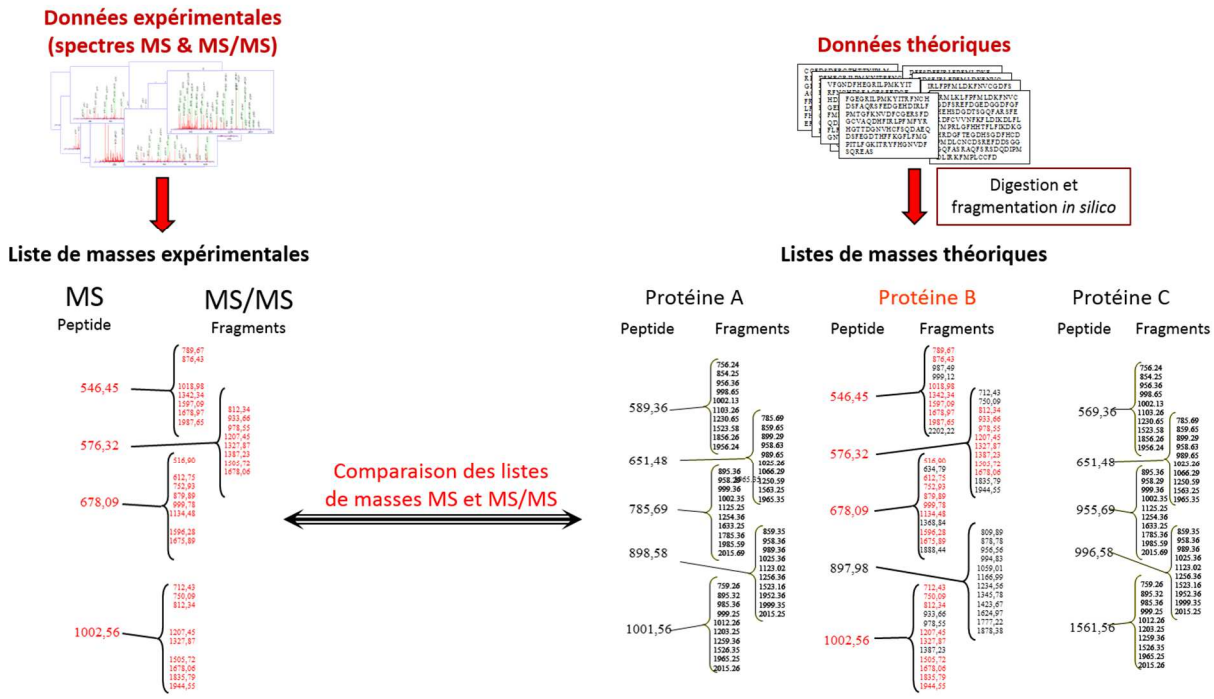


Figure II-1 : Représentation schématique de la recherche en banque de données par empreinte de fragments peptidiques.

Les masses expérimentales des peptides et de leurs fragments sont comparées aux masses théoriques des peptides et fragments issus de la digestion et fragmentation *in silico* des protéines contenues dans les banques de données.

Il est nécessaire de définir l'enzyme utilisée, le type de fragment générés (en fonction du mode de fragmentation utilisé), une tolérance de masse pour les peptides et leurs fragments, ainsi qu'une banque de données dans laquelle effectuer la recherche. On pourra, en effet, identifier uniquement

les peptides qui se trouvent dans la banque. Dans le cas où l'on étudie un organisme dont le génome est séquencé, on utilisera alors directement la banque de données contenant les séquences protéiques de cet organisme.

Il est également possible d'indiquer au moteur de recherche les modifications post-traductionnelles (MPT) recherchées. Très souvent, ce sont principalement les modifications induites par la préparation d'échantillons qui sont recherchés (telle que la carbamidométhylation des cystéines) mais plus de 500 sont actuellement répertoriées [44]. Les MPT correspondent à l'addition covalente d'un groupe chimique sur différents acides aminés ou la modification de la structure même d'un acide aminé, qui ont lieu après la traduction de l'ARNm. Elles jouent un rôle clé dans de nombreux processus biologiques, et sont donc importantes pour caractériser l'ensemble du protéome [9]. Rechercher l'ensemble des MPT possibles conduirait à des temps de recherche extrêmement longs. L'approche classique pour déterminer des MPT implique une étape d'enrichissement, car les MPT sont souvent peu abondantes [9]. Les MPT sont donc détectées grâce à la différence de masse qu'elles induisent au peptide et également à des fragments spécifiques.

Il existe plusieurs algorithmes de recherche : Mascot [45], Andromeda [46] (intégré au logiciel MaxQuant), X!Tandem [47], Sequest [43], OMSSA [48]… Afin d'évaluer la qualité de l'assignation d'une séquence à un spectre (PSM pour *Peptide-Spectrum Match*), ces algorithmes attribuent un score à chaque identification, qui reflète la probabilité p que l'assignation soit due au hasard. Pour Mascot, par exemple, ce score d'ion est égal à $-10 * \log(p)$. Plus le score est élevé, plus l'assignation est de bonne qualité. Mascot attribue également un deuxième score aux assignations : le score d'identité, qui permet d'ajuster le score à la taille de la banque de données. En effet, ce score prend en compte le nombre de peptides (théoriques) dans la banque de données dont la masse est égale à la masse du peptide expérimental.

Dans la suite de ce manuscrit, cette méthode sera appelée « approche classique » ou « recherche classique ».

B. Les banques de données protéiques

De la banque découle directement l'identification fiable des peptides et des protéines, seuls les peptides présents dans la banque pourront être identifiés. C'est pourquoi le choix de la banque de données la plus adaptée est primordial. Le degré de complétude de la banque est donc un facteur décisif dans les étapes d'identification et d'interprétation biologique des données. Les banques protéiques proviennent de l'annotation des banques génomiques, et la qualité des données est très variante. Il existe deux types de banques, les banques corrigées et les banques non corrigées.

La banque UniProt regroupe deux banques issues de l'annotation des banques de séquences nucléotidiques provenant du consortium GenBank/EMDL/DDBL [49] :

- Swiss-Prot [50] qui est la banque de données de référence aujourd'hui, pour laquelle les annotations ont été revues et validées.
- TrEMBL, dont les séquences ont été annotées automatiquement et non revues/non validées.

La banque NCBI est également composée de séquences protéiques traduites des banques nucléotidiques GenBank/EMDL/DDBL, et ne créée pas de nouvelles annotations mais reprend les informations contenues dans d'autres banques (telles que Swiss-Prot ou Protein Information Resource, RefSeq…) et contient donc de nombreuses redondances et est très volumineuse.

Si le but est d'identifier un maximum de variants de séquences, il est plus approprié d'utiliser une large banque, telle que NCBI. Mais ces larges banques ne contiennent pas que des variants biologiquement significatifs, mais également de nombreuses redondances artificielles provenant d'erreurs de séquençage. Lorsque la qualité de l'annotation des séquences est plus importante que l'identification d'un maximum de variants, il est préférable d'utiliser une banque corrigée telle que Swiss-Prot [51].

Toutes les banques de données utilisées dans ce manuscrit proviennent d'UniProt, et lorsque cela est possible, préférentiellement de Swiss-Prot.

C. Validation des identifications

Il est essentiel de valider les identifications à l'issue d'une recherche en banque de données, pour renforcer la confiance dans les résultats. Pour cela, on utilise les scores attribués à chaque assignation par les algorithmes de recherche.

La méthode la plus couramment utilisée est l'évaluation du taux de faux positif (FDR : *False Discovery Rate*) [52]. Une banque de données « decoy » est créée à partir de la banque originale (« target »), soit en inversant la séquence des protéines, soit en randomisant l'ordre des acides aminés des protéines ; créant ainsi des fausses séquences qui n'existent pas. La recherche en banque est alors effectuée soit dans les deux banques séparément, soit les deux banques sont réunies et la recherche se fait dans la banque concaténée. L'évaluation du calcul du taux de faux positifs se base sur l'hypothèse qu'une assignation aléatoire est aussi susceptible d'avoir lieu dans la banque « decoy » que dans la banque « target », ainsi le nombre d'assignations obtenues dans la banque « decoy » donne une estimation du nombre d'assignations aléatoires (faux-positifs) que l'on peut obtenir dans la banque « target » [53]. Les identifications sont donc filtrées grâce au score, jusqu'à obtenir un FDR inférieur au seuil défini (généralement 1%). Le FDR peut être calculé au niveau peptidique ou protéique, selon la formule suivante :

$$FDR = \frac{Nombre\ de\ decoy}{(Nombre\ de\ target + Nombre\ de\ decoy)}$$

Il existe plusieurs logiciels permettant d'estimer le FDR, parmi lesquels Scaffold (Proteome software), Andromeda (Maxquant), X!Tandem, Proline…

II. Stratégies d'identification chez les espèces non séquencées

A. L'empreinte de fragments peptidiques

La stratégie d'identification par empreinte de fragments peptidiques peut être utilisée lorsque l'on étudie une espèce dont le génome n'est pas séquencé. Les identifications peptidiques se feront alors par homologie de séquence. A défaut d'utiliser une banque de données protéique directement dérivée du génome de l'organisme étudié, on utilise une banque de données contenant les séquences

protéiques d'une espèce phylogénétiquement proche prise comme référence. Les peptides trypsiques dont la séquence est **strictement** conservée entre organismes pourront alors être identifiés.
Les critères pour déterminer l'organisme de référence sont : que son génome soit lui-même suffisamment annoté ; qu'il soit suffisamment proche de l'espèce étudié afin que le plus grand nombre de peptides trypsiques soient **strictement** conservées. Il est également possible d'utiliser une banque de données « multi-espèces », qui contient les séquences protéiques de la classe, de l'ordre, de la famille, ou du genre auquel appartient l'espèce étudiée. Le choix de cette banque est crucial pour identifier un maximum de peptides sans prendre le risque d'augmenter les identifications de faux-positifs.

En revanche, cette méthode limite considérablement le nombre de peptides identifiables, surtout pour les espèces les plus « exotiques ». On peut alors combiner cette stratégie avec le séquençage *de novo*, qui est présenté dans le paragraphe suivant.

B. Séquençage *de novo* et alignement de séquence

B.1. Séquençage *de novo*

Contrairement à la stratégie d'identification par empreinte de fragments peptidiques, qui consiste à comparer des masses expérimentales à des masses théoriques, le séquençage *de novo* consiste à interpréter directement un spectre de fragmentation pour en déduire la séquence primaire du peptide [8, 54].

Comme présenté en page 20, la fragmentation CID rompt les liaisons peptidiques (entre le C terminal d'un acide aminé et le N terminal de l'acide aminé (aa) suivant), formant ainsi des ions b et y. La différence de masse entre deux pics consécutifs sur le spectre est donc égale à la masse d'un acide aminé (Figure II-2 et Figure II-3).

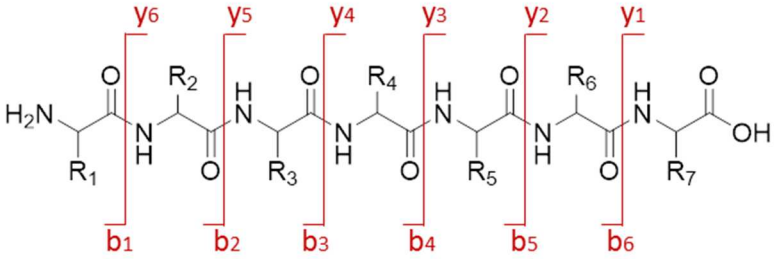


Figure II-2 : Nomenclature de Biemann, représentant uniquement les ions b et y.
Le fragment y1 contient l'aa 7 (aa portant le groupement fonctionnel « R_7 ») ; le fragment y2, les aa 7 et 6… et le fragment y6 les aa 7, 6, 5, 4, 3 et 2, comme présenté sur la figure suivante. La différence de masse entre les fragments y1 et y2 donne donc la masse de l'aa 6, entre y2 et y3 la masse de l'aa 5 etc.

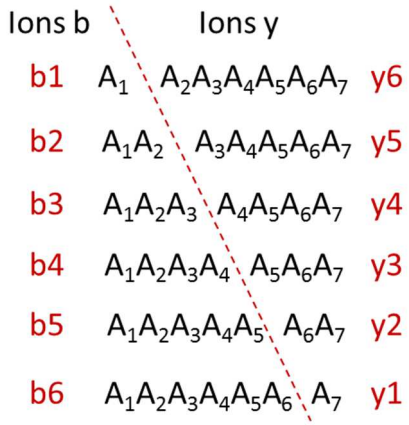


Figure II-3 : Représentation des ions fragments b et y du peptide de la figure précédente.

L'interprétation d'un spectre MS/MS permet donc d'obtenir les acides aminés qui constituent la séquence primaire du peptide, comme présenté en Figure II-4. Cependant, l'interprétation est rarement complète, en effet certains fragments peuvent être trop faibles ou tout simplement non détectés. Ainsi, on obtient souvent des « tags » de séquence, et non la séquence complète du peptide analysé.

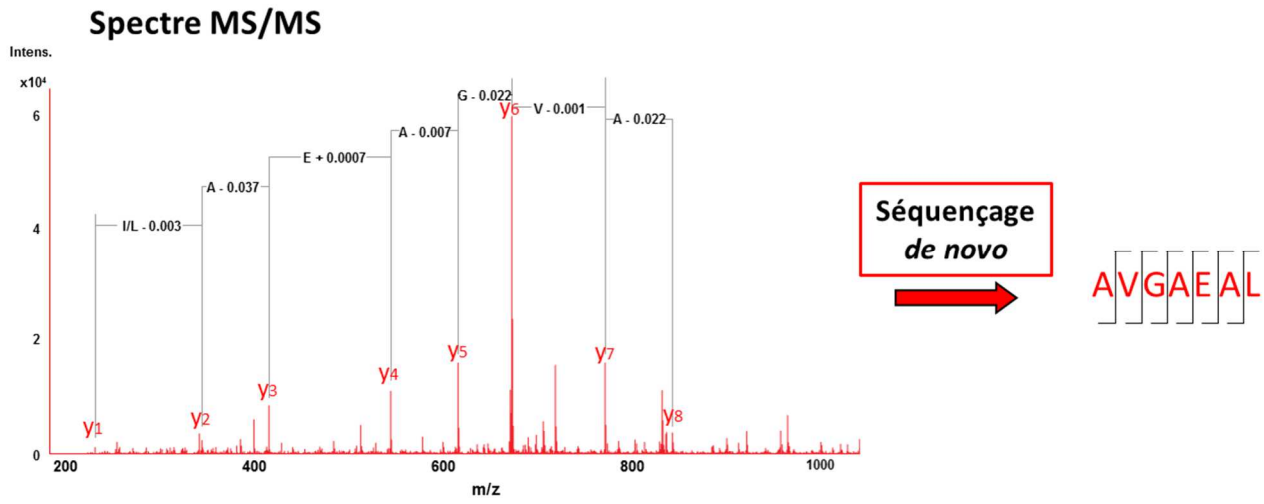


Figure II-4 : Interprétation d'un spectre MS/MS par séquençage *de novo*.

Cette approche nécessite d'utiliser des spectromètres de masse résolutifs et possédant une grande précision de masse pour limiter les erreurs et obtenir une interprétation fiable. De plus, il est nécessaire d'utiliser des spectres de haute qualité ; c'est-à-dire dont le rapport signal sur bruit est important, afin de pouvoir différencier les pics de fragmentation du bruit de fond. Le logiciel Recover [55] peut être utilisé pour filtrer les spectres en fonction de leur qualité. Pour cela, on définit un nombre de pics utiles (UPN pour *Useful Peak Number* en anglais) dont l'intensité doit être supérieure à un seuil d'intensité définit comme un multiple (E) du bruit de fond. Le bruit de fond est défini comme l'intensité médiane de tous les pics du spectre. L'exemple d'un spectre de bonne qualité selon des critères définis à Recover est présenté en Figure II-5.

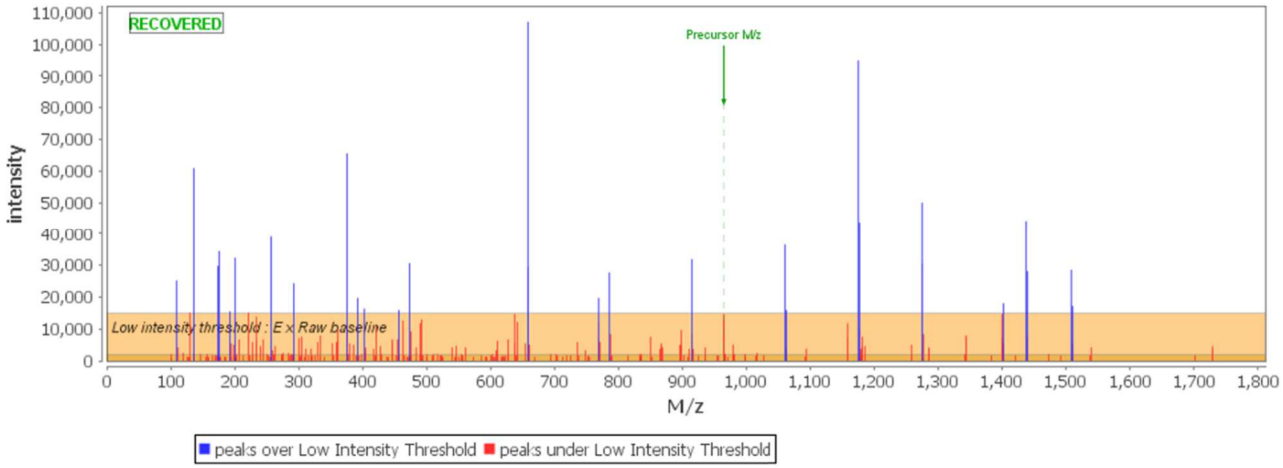


Figure II-5 : Exemple d'un spectre MS/MS « recovered », i.e. de bonne qualité selon les critères appliqués.
Ici les paramètres étaient E= 8 et UPN= 8. Un spectre est donc conservé (« recovered ») s'il présente au moins 8 pics (UPN) dont l'intensité est supérieure à 8 (E) fois le bruit de fond.

Avec Recover, il est également possible d'éliminer les spectres déjà assignés dans la recherche classique, pour ne pas avoir à les réinterpréter et ainsi limiter l'espace de recherche.

Enfin, le nombre de spectres générés par des appareils de plus en plus performants (plusieurs dizaines de milliers de spectres par analyse), rend l'interprétation manuelle non envisageable à grande échelle. On utilise donc des algorithmes qui font ces calculs de manière automatisé à grande échelle et rapidement. On peut citer notamment les logiciels PEAKS studio [56] et PepNovo [57]. Le fonctionnement du logiciel PepNovo sera expliqué en Partie II - Chapitre II - I. - C.1.2 (page 81).

B.2. Alignement de séquence

Suite au séquençage *de novo*, les tags de séquences peptidiques doivent être comparés aux séquences protéiques contenues dans une banque de données (voir Figure II-6). Cette comparaison s'effectue par une recherche par similarité de séquences grâce à l'algorithme MS BLAST (*Basic Local Alignment Search Tool*) [58] par exemple. Ainsi, si une protéine de la banque de données contient des peptides similaires aux tags de séquence *de novo,* la protéine étudiée est alors très probablement un homologue de cette protéine de la banque. Malgré le fait qu'on ne puisse obtenir la séquence complète de la protéine, on obtient ainsi une information fonctionnelle par homologie [8].

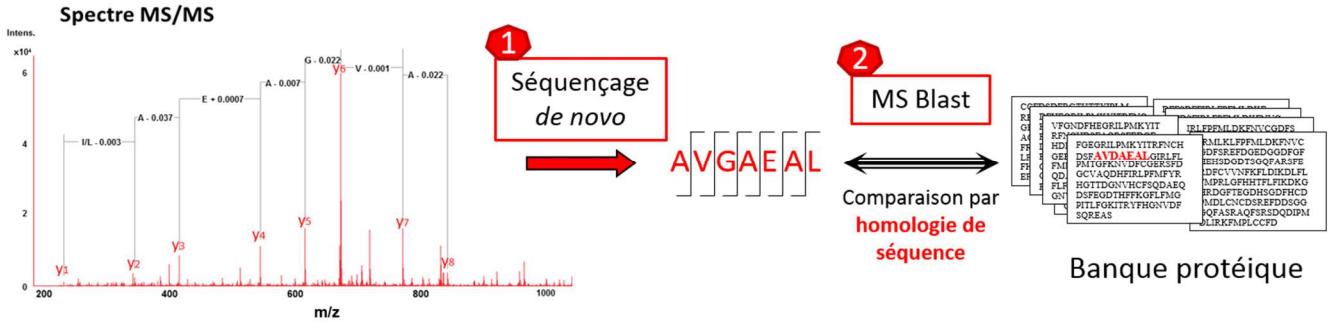


Figure II-6 : Interprétation d'un spectre MS/MS par séquençage *de novo* suivi de la recherche MS BLAST dans une banque de données protéiques.
Les peptides séquencés *de novo* sont ainsi assignés à des protéines par homologie de séquence.

Ici encore, il faudra utiliser la banque de données protéique d'une espèce proche (la même qui a été choisie pour l'approche classique). L'avantage ici, c'est que l'on compare des suites de caractères et non plus des masses ; la comparaison n'est donc plus stricte. On pourra autoriser des différences entre les séquences, des mutations d'un ou plusieurs acides aminés (en fonction de la stringence que l'on accorde ou de la proximité phylogénétique entre les deux organismes) et ainsi identifier des peptides qui ne sont **pas strictement** conservés et augmenter la couverture de protéome.

Afin d'obtenir des résultats fiables et de bonnes qualités par l'approche séquençage *de novo* suivi d'un BLAST, il est essentiel de travailler sur des données d'excellentes qualité mais les résultats doivent également être validés sur la base du score, de la longueur de séquence, de la longueur de la correspondance etc. Nous nous sommes intéressés aux paramètres du logiciel PepNovo afin d'en optimiser l'utilisation. Le fonctionnement du logiciel PepNovo, les optimisations et leurs résultats seront présentés en détail en Partie II - Chapitre II - I. - C.1.2 (page 81).

C. Autres approches

Pour les organismes non modèles, il est possible de séquencer le génome étudié afin d'obtenir une banque de données soit génomique soit protéique (par traduction dans les 6 cadres de lecture) [59]. L'annotation du génome permet d'obtenir des protéines prédites [60], ce qui permet d'interpréter les spectres MS/MS par l'approche classique (voir Figure II-7).

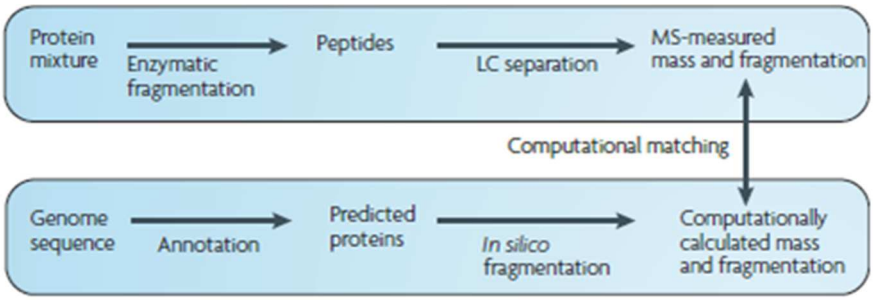


Figure II-7 : Schéma expérimental décrivant l'interprétation des spectres MS/MS dans une banque protéique issue d'une banque de génomique annotée. Figure provenant de l'article [60].

Il est également possible de séquencer l'ARN des échantillons étudiés [61-64], et d'assembler le transcriptome [65, 66]. Une banque de données protéique est ensuite constituée par traduction du transcriptome, comme le montre la Figure II-8.

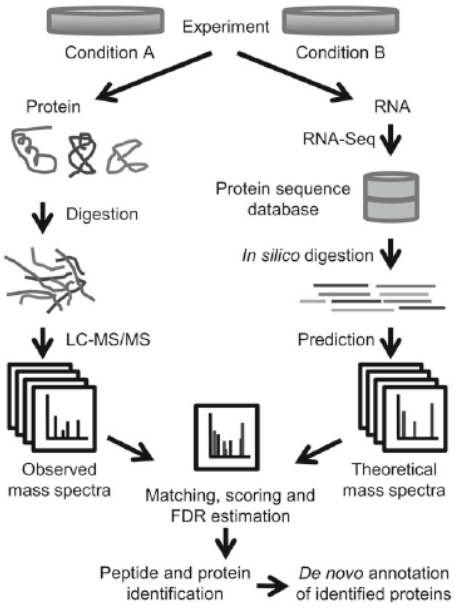


Figure II-8 : Schéma expérimental décrivant l'interprétation des spectres MS/MS dans une banque protéique issue d'une banque de transciptomique. Figure provenant de l'article [67].

Evans *et. al.* [63] ont évalué cette technique, qu'ils ont appelé « *Proteomics Informed by Transcriptomics* » (PIT) en l'appliquant à des organismes dont le génome est très bien annoté, tel que l'Homme. Avec la technique PIT, ils ont ainsi identifié près de 95% des protéines identifiées dans la banque Swiss-Prot *Homo Sapiens*.

Même si aujourd'hui les techniques de séquençage sont de plus en plus rapides et abordables, ces méthodes ne donnent pas accès à toutes les séquences protéiques étant donné la qualité du séquençage, de l'assemblage et de l'annotation du génome. L'annotation se fait sur la base de l'état actuel des connaissances sur le gène et est donc sujette aux erreurs et aux incohérences [7, 51, 68]. De plus, La prédiction d'une protéine hypothétique dans le mauvais cadre de lecture empêchera l'identification de la protéine. A l'inverse, le fait d'avoir une banque de données protéique contenant toutes les prédictions possibles (dans les six cadres de lecture) permet d'utiliser les résultats de protéomique pour valider la prédiction des gènes, ce qui se fait de plus en plus aujourd'hui [69-71].

III. Stratégies d'identification en métaprotéomique

La métaprotéomique est l'étude protéomique de communautés complexes d'organismes (tels que le microbiote intestinal, les communautés biotiques des sols…) [60]; l'analyse métaprotéomique correspond donc à l'analyse des protéines de tous les organismes présents dans un tel échantillon. Le but est de quantifier ces protéines afin de comprendre le fonctionnement des communautés biotiques au sein d'un environnement donné, et d'en définir les éventuelles conséquences sur cet environnement.

La méthode d'identification par empreinte de fragments peptidiques est utilisée pour l'analyse métaprotéomique. La difficulté majeure est de trouver une banque de données adaptée à l'étude d'un grand nombre d'organismes. Il faut déjà connaître ou déterminer les organismes présents, puis trouver/obtenir une banque de données suffisamment complète et représentative de la communauté étudiée.

Le séquençage *de novo* est également une possibilité pour interpréter les spectres MS/MS issus d'une analyse de métaprotéomique [72], et ne nécessite pas, dans un premier temps, de banque de données. Cependant, cette méthode n'est pas largement répandue aujourd'hui en métaprotéomique, la recherche en banque de données reste la plus utilisée [73, 74].

A. A la recherche d'une banque de données adaptée

A.1. Banques de données protéiques préexistantes

Lorsque les organismes présents dans l'échantillon sont connus, il est possible d'utiliser des banques de données protéiques existantes contenant les séquences de ces organismes. L'inconvénient reste la qualité et l'exhaustivité de ces banques préexistantes qui ne sont pas toujours au rendez-vous. En revanche, l'avantage est que la taille de ces banques est limitée par rapport à une banque directement dérivée du génome (voir paragraphe suivant), et sont plus spécifiques que des banques génériques (comme par exemple l'intégralité de Swiss-Prot ou Uniprot par exemple), ce qui permet d'obtenir une bonne couverture de protéome [75].

Pour caractériser la composition des communautés biotiques de l'échantillon étudié, il est possible d'utiliser la génomique, et plus précisément la méthode de metabarcoding [76]. L'objectif du metabarcoding est d'identifier les espèces présentes dans un milieu, grâce à un « code-barres ADN », qui est un fragment spécifique de l'ADN d'une espèce donnée, ce qui est particulièrement pratique pour des organismes que l'on ne peut caractériser morphologiquement (comme les micro-organismes) [77]. On peut également détecter ainsi des macro-organismes qui auraient laissé des traces dans le milieu. L'ADN est extrait de l'échantillon et le « code-barres » (fragment spécifique) est amplifié par PCR (*Polymerase Chain Reaction*). Les amplicons ainsi obtenus sont ensuite séquencés puis comparés à des séquences de référence pour identifier l'espèce dont l'ADN est issu.

A.2. Banques de données directement dérivées des échantillons

A.2.1. Méta-génomique

Il est possible de séquencer le métagénome étudié afin d'obtenir une banque de données soit génomique soit protéique [60, 72], comme présenté plus haut pour les organismes non séquencés (voir Figure II-7).

Comme exposé plus haut, cette méthode ne donne pas accès à toutes les séquences protéiques étant donné la qualité du séquençage, de l'assemblage et de l'annotation du génome, surtout pour des données aussi importantes (génome de plusieurs espèces) [73, 74]. De plus, les banques de données protéiques ainsi générées sont très volumineuses (jusqu'à des millions de séquences).

A défaut de pouvoir séquencer directement les génomes étudiés, il existe également de plus et plus de données de méta-génomique disponibles à la communauté (*repositories*) pour de nombreux éco-systèmes, qui peuvent être utilisés pour des études métaprotéomique. Ces ressources peuvent être utilisés en complément du génome (ou protéome) d'un organisme spécifique étudié, pour augmenter la couverture de protéome [73, 74, 78].

A.2.2. Méta-transcriptomique

Il est également possible de séquencer l'ARNm des échantillons étudiés, comme présenté plus haut pour les organismes non séquencés (voir Figure II-8) et créer une banque de données protéiques à partir de ces données. Si on retrouve de nombreux exemples de transcriptomique appliqué à la protéomique pour des espèces non modèles [67], les applications à la métaprotéomique sont plus rares [74, 79]. Les données méta-transcriptomiques sont plus souvent utilisées en complément des données de méta-protéomique [80] que pour générer des banques de données protéiques.

B. Validation des données

Les banques de données pouvant être très volumineuse pour des analyses de métaprotéomique (dû au grand nombre d'organismes présents), il peut être difficile d'évaluer un taux de faux positifs fiable pour la validation des identifications. En effet, plus l'espace de recherche et la similarité entre les séquences deviennent importants (ce qui est le cas pour des banques contenant les séquences protéiques de nombreux organismes proches), plus il est difficile de séparer les assignations peptide-spectre correctes des incorrectes [81]. En effet, l'avantage de la banque « decoy » est qu'elle ne partage pas de séquences avec la banque « target », ce qui devient difficile lorsque cette dernière dépasse les 6 milliards d'acides aminés [82].
La méthode de calcul du taux de faux positifs (FDR) est basée sur la génération d'une banque « decoy », contenant des protéines que l'on ne devrait pas retrouver dans les échantillons. Lorsque l'on utilise une banque de données contenant un nombre trop important d'entrées protéiques, la banque decoy est alors également importante. Statistiquement les chances d'identifier un peptide faux-positif sont donc largement augmentées, tandis que le nombre de « vraies » protéines présentes dans l'échantillon n'est pas plus important ; de ce fait le ratio $\frac{\text{faux positifs}}{\text{vrais positifs}}$ est augmenté [53]. Le FDR étant ainsi superficiellement augmenté, et obtenir un FDR convenable (communément inférieur à 1%) nécessite alors d'invalider un grand nombre de protéines « vrai-positifs ».

De ce fait, plusieurs solutions ont été envisagées dans la littérature pour pallier à ce problème, mais il n'existe aujourd'hui aucun consensus de la communauté quant à la façon d'évaluer le taux de faux positifs pour ce type de données [53]. Ces différentes solutions sont :

- Des méthodes de correction du FDR ont été décrites [53], qui consistent par exemple soit à calculer un facteur correctif [83] ; ou à modéliser les faux positifs par une distribution hypergéométrique [84] ; ou à utiliser une approche bayésienne pour calculer un FDR local [85, 86] ;

- Des stratégies de validation sans utiliser de decoy ont également été décrites [81, 87], qui consistent à se baser sur de l'apprentissage automatique (ou *machine learning*) ou sur la distribution des scores des PSMs pour modéliser les PSMs incorrects ;
- Diminuer la banque de données :
    - A partir d'une banque génomique : la traduction peut être restreinte aux séquences qui excèdent la taille moyenne d'un exon dans l'organisme étudié, ou à des régions possédant des hauts scores par prédictions des gènes *ab initio* [82].
    - A partir d'une banque transcriptomique : il est possible, à la suite de la quantification de transcrits, d'éliminer les gènes peu exprimés [88].
    - A partir des banques protéiques : utiliser les identifications d'une première recherche non stringente (non validé par FDR) pour créer une deuxième banque, restreinte, avec laquelle une deuxième recherche stringente (FDR < 1%) est effectuée [89].

Une difficulté de l'analyse métaprotéomique est l'identification sûre des organismes responsables de l'expression de chaque protéine détectée. En effet, on ne peut pas toujours identifier de peptides spécifiques pour un organisme donné, plus il y a d'espèces présentes dans la banque, plus il y a de risques d'identifier des peptides partagés (potentiellement entre protéines homologues entre espèces).
Il est en effet possible d'identifier un ou plusieurs peptides conservés/partagés entre 2 organismes. Mais comment savoir alors si les deux organismes sont effectivement présents si aucun peptide unique, propre à l'un ou l'autre de ces 2 organismes, n'est identifié ? Ainsi, il est quasiment impossible d'assigner toutes les indentifications à un organisme précis [60]. Les informations taxonomiques obtenues en métaprotéomique sont donc à traiter avec précaution [78], dans la mesure où seul un organisme est choisi « arbitrairement » parmi les différentes possibilités, ce qui peut induire une perte d'information [75]. Une technique a été développée en métagénomique pour inférer les séquences au plus petit ancêtre commun (LCA pour *Lowest Common Ancestor*) [90], qui peut être utilisée en métaprotéomique : si une séquence est assignée à plusieurs espèces appartenant au même genre, alors la séquence sera assignée à ce genre. Dans le cas où l'on s'intéresse à une réponse fonctionnelle, et non taxonomique, il n'est pas nécessaire d'utiliser cette technique pour déterminer la provenance taxonomique de chaque peptide/protéine identifié(e).

Les résultats de métaprotéomique apportent donc, en général, plutôt une **vue d'ensemble de l'activité métabolique d'une communauté**, mais ne permet pas de déterminer précisément quel membre de la communauté *spécifiquement* est responsable de ces fonctions [60, 78].

Partie I : Introduction à la protéomique

Chapitre III : Stratégies de quantification des protéines

Partie I : Introduction à la protéomique

Chapitre III : Stratégies de quantification des protéines

La protéomique permet non seulement d'identifier des peptides et des protéines, mais il est également possible d'extraire des informations quantitatives, soit par le biais de stratégies de marquage soit directement à partir des données spectrales MS et MS/MS.

Il existe différentes méthodes de quantification, présentées en Figure III-1 et dans la suite de ce chapitre, à choisir suivant la problématique, le nombre et la nature d'échantillons à comparer. La première question à se poser est : le but est-il de réaliser une analyse globale ou ciblée ? Autrement dit, a-t-on un a priori sur les échantillons et ce que l'on cherche ?

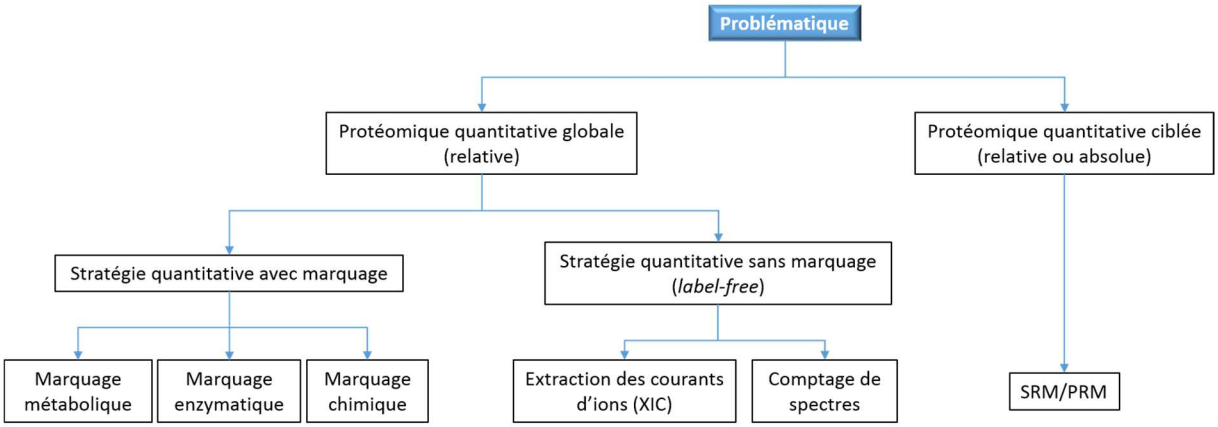


Figure III-1 : Principales stratégies quantitatives en protéomique

I. Stratégies de quantification ciblée

Les méthodes ciblées permettent de quantifier un « faible » nombre de protéines simultanément, en suivant plusieurs peptides d'intérêt pour chaque protéine. En ne suivant qu'un nombre restreint de peptides, ces méthodes sont plus spécifiques et sensibles que les méthodes globales car elles permettent de s'affranchir du problème de compétition pour la fragmentation. L'inconvénient de ces méthodes est le nombre de « cibles », on ne peut en effet pas quantifier un grand nombre de protéines (généralement pas plus d'une cinquantaine). Cependant, ces méthodes restent totalement adaptées si l'on désire avoir une information quantitative sur quelques protéines seulement, et que l'on n'a pas besoin d'information sur l'ensemble de l'échantillon. Ces méthodes sont très utilisées pour la recherche ou validation de biomarqueurs [91, 92].

A. SRM

La SRM (*Selected Reaction Monitoring*) [93] consiste à sélectionner certains peptides pour la fragmentation puis à sélectionner quelques fragments pour l'analyse. Un couple peptide-fragment s'appelle une transition. On sélectionne donc plusieurs transitions par peptide. Ce type d'analyse est réalisé sur un instrument triple quadripôle (QQQ) ; le premier quadripôle permet d'isoler le peptide, le deuxième de le fragmenter et le dernier d'isoler le ou les fragments d'intérêt, comme présenté en Figure III-2).

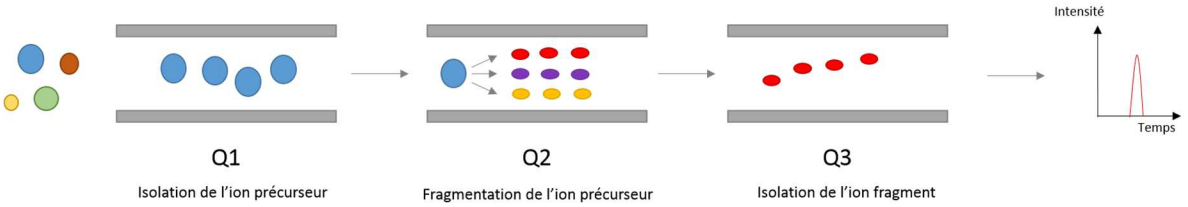


Figure III-2 : Représentation schématique du fonctionnement d'un triple quadripôle en mode SRM (*Selected Reaction Monitoring*).

La première étape consiste à réaliser une analyse globale pour choisir les transitions. Les protéines d'intérêt sont tout d'abord déterminées en fonction de la question biologique posée, puis les peptides sélectionnés selon plusieurs critères : il faut qu'ils soient protéotypiques (spécifiques de la protéine, non partagés), qu'ils s'ionisent et se fragmentent bien en spectrométrie de masse, ne possèdent pas de méthionines (qui peuvent être présentes sous forme oxydée ou non), ni de clivages manqués ; afin d'être sûr de voir au mieux et de façon reproductible les peptides cibles. Ensuite les fragments les plus intenses sont choisis, généralement entre 3 et 10 par peptide. Plus il y a de transitions, plus la quantification est spécifique. En effet, la co-élution de plusieurs transitions pour un peptide confirme sa présence et infirme la présence d'une interférence dont une transition aurait la même masse et le même temps de rétention (ce qui devient moins probable pour plusieurs transitions).

La quantification se fait ensuite par extraction des courants d'ions des fragments. Afin d'obtenir une quantification plus robuste, on ajoute des standards internes. Ces standards sont des peptides synthétiques, de même séquence que les peptides cibles (endogènes), marqués à l'aide d'isotopes stables (^{13}C et ^{15}N) pour les différencier des endogènes. Ces peptides synthétiques ont donc les mêmes propriétés physico-chimiques que les endogènes (donc même rétention chromatographique et efficacité d'ionisation), la seule différence étant la masse. On calcule un ratio entre l'abondance du peptide endogène et l'abondance du peptide marqué (« léger sur lourd »), ce qui permet de corriger les écarts entre les analyses. Suivant la pureté des peptides, on peut obtenir une quantification relative ou absolue.

Les limitations principales de la SRM sont le faible nombre de protéines analysables et le temps de développement de la méthode. Son avantage indéniable est sa grande spécificité et sa sensibilité (une dizaine d'amol en général) sur des grandes gammes dynamiques de concentration (entre 4 et 5 ordres de grandeur) [94].

B. PRM

La PRM (*Paralell Reaction Monitoring*) [95] est basée sur le même principe que la SRM mais permet de suivre tous les fragments des peptides, il n'est plus nécessaire de sélectionner des transitions. Le choix des peptides reste le même que pour la SRM. Les analyses PRM se font sur des instruments à haute résolution (Q-Orbitrap™ ou Triple-ToF).

L'intérêt de la PRM par rapport à la SRM est le fait de ne pas être limité en nombre de transitions par peptide, rendant la quantification plus spécifique et robuste, ce qui est rendu possible par l'utilisation de spectromètres de masse haute résolution. En revanche, cette méthode est moins sensible car tous les fragments sont analysés en même temps, et non un par un comme en SRM.

II. Stratégies de quantification globale

Les stratégies de quantification globale ont pour avantage de quantifier un grand nombre de protéines, et ne requièrent pas d'avoir un a priori sur les échantillons. Ces méthodes de quantification sont toujours relatives, et ne permettent pas d'accéder à une quantification précise ; elles permettent cependant d'émettre des hypothèses sur les régulations biologiques entre les différentes conditions comparées. Ces hypothèses peuvent être vérifiées par des méthodes complémentaires (Western blots, ELISA, SRM, dosages chimiques et enzymatiques, tests fonctionnels…).

Il existe deux types de quantification globale, avec ou sans marquage des protéines ou peptides, qui seront présentées dans la suite de ce chapitre.

A. Stratégies de quantification globale avec marquage

Les stratégies de quantification globale avec marquage consistent à marquer les protéines ou peptides avec des isotopes stables avant ou après extraction des protéines. Excepté pour la méthode 2D-DIGE (présentée plus bas), la quantification est basée sur l'extraction des courants d'ions et comparaison des aires sous la courbe (AUC pour *Area Under the Curve*) des courants d'ions. Certaines de ces méthodes permettent un multiplexage, c'est-à-dire de comparer plusieurs échantillons dans la même analyse LC-MS/MS. Les différents échantillons à comparer sont marqués avec différents isotopes stables. Les peptides ainsi marqués auront les mêmes propriétés physico-chimiques (même rétention chromatographique, même efficacité d'ionisation et de fragmentation), ce qui les rend tout à fait comparable, mais vont différer par leurs masses, ce qui permettra de reconnaître la provenance de chaque peptide.

Le marquage peut être réalisé à différentes étapes et de différentes façons. Il existe plusieurs stratégies, présentées ci-dessous :

- Le marquage métabolique : La méthode SILAC [96] (*Stable Isotope Labeling by Amino Acids in Cell culture*) consiste à marquer des protéines en culture par incorporation métabolique des isotopes stables. Généralement, les acides aminés marqués avec cette méthode sont l'Arginine et la Lysine, afin que les peptides issus d'une digestion trypsique possèdent au moins un acide aminé marqué. On peut

comparer jusqu'à trois échantillons, un « léger » non marqué et deux marqués différemment. Tous les échantillons sont mélangés et analysés en une fois (voir Figure III-3).

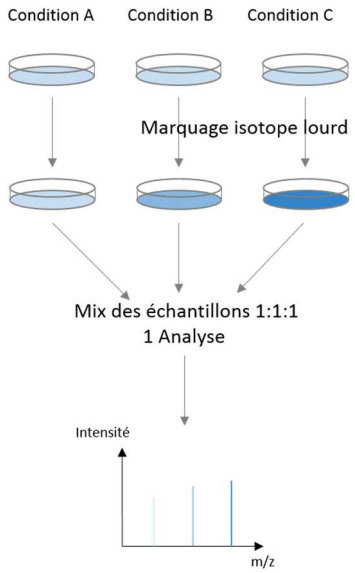


Figure III-3 : Représentation schématique de la méthode SILAC.

L'avantage de cette méthode est l'incorporation du marquage à un stade précoce, limitant les biais techniques qui peuvent survenir entre les échantillons, et ainsi obtenir une quantification plus juste. En revanche, le nombre d'échantillons comparables est limité et le type d'échantillons quantifiable est également un facteur limitant. En effet, cette méthode s'applique à des cultures cellulaires, et non à des échantillons tissulaires collectés sur des organismes entiers.

❖ Le marquage enzymatique consiste à introduire un isotope lourd lors de l'étape de digestion enzymatique dans un solvant enrichi en $H_2{}^{18}O$ [97], voir Figure III-4. L'isotope lourd ^{18}O est ainsi incorporé dans le groupe carboxylique C-terminal des peptides. On peut ainsi comparer deux échantillons : un « léger » non marqué et un « lourd » marqué. Le nombre d'échantillons comparables est ici encore une limitation. De plus, l'étape de marquage est réalisée au dernier stade de la préparation des échantillons, cette stratégie ne permet donc pas de réduire les biais de préparation (mais permet de réduire les variations instrumentales, étant donné que les échantillons sont analysés en même temps). Enfin, cette méthode est également limitée par les faibles degrés d'incorporation de l'isotope lourd.

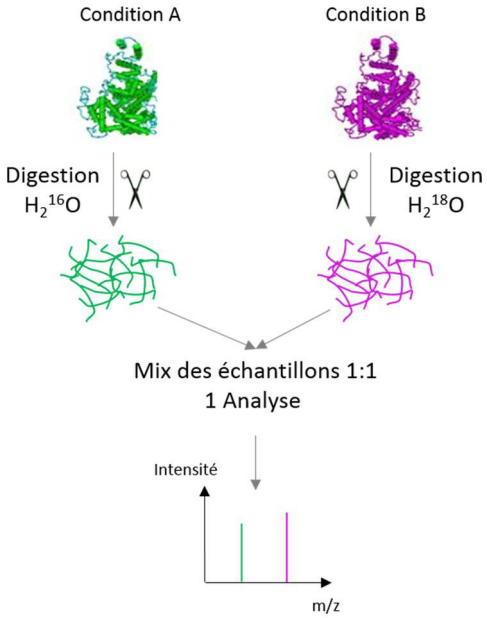


Figure III-4 : Représentation schématique de la méthode par marquage enzymatique ^{18}O.

❖ Le marquage chimique consiste à modifier le groupe réactif d'un peptide à l'aide d'isotopes légers et lourds. Il existe plusieurs méthodes pour cela : ICAT [98] (*Isotope Coded Affinity Tag*), iTRAQ [99] (*isobaric Tags for Relative and Absolute Quantification*), TMT™ [100] (*Tandem Mass Tags*), le marquage par diméthylation [101], le marquage chimique en gel 2D-DIGE [19] (*Differential In-Gel Electrophoresis*).

En 2D-DIGE, l'information quantitative est obtenue par analyse densitométrique de spots protéiques. Des cyanines sont utilisées pour marquer les groupements thiols des cystéines ou les amines primaires des protéines (extrémités N-terminales ou amines de lysines). Il existe trois types de cyanines (Cy2, Cy3, Cy5) qui diffèrent par leur longueur d'onde d'absorption et d'émission, on peut alors marquer chaque échantillon avec une cyanine différente, les mélanger et faire migrer les protéines sur un même gel 2D (dont le principe est expliqué en page 10) puis révéler le signal spécifique de chaque cyanine (et donc de chaque échantillon) dans chaque spot protéique (voir Figure III-5). Généralement, deux cyanines sont utilisées pour marquer des échantillons et la dernière pour un standard interne déposé sur tous les gels (en général un mélange de tous les échantillons) qui permettra de réaliser une normalisation inter-gel pour diminuer les biais. Des logiciels de traitement d'images (Samespot, Waters ; PD-Quest, Bio-Rad…) permettent de déterminer l'abondance relative de chaque spot protéique, pour chaque échantillon. Les spots sont ensuite découpés et analysés par spectrométrie de masse (après digestion des protéines et extraction des peptides du gel) pour identifier les protéines.

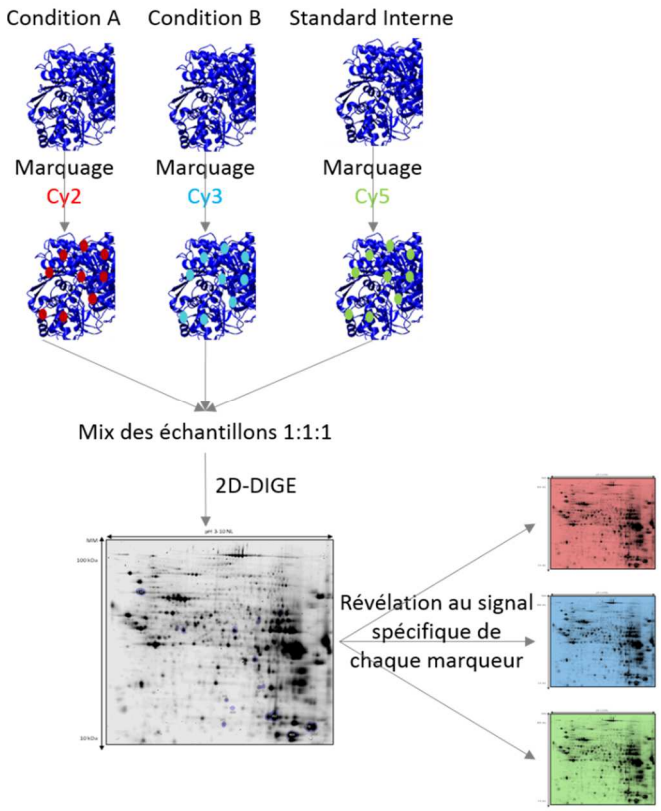


Figure III-5 : Représentation schématique de la méthode 2D-DIGE.

Les limitations de cette technique sont le manque d'automatisation et le fait qu'on puisse difficilement quantifier des protéines de masse moléculaire ou de point isoélectrique extrêmes. Cependant, l'utilisation de gels précoulés et de systèmes d'électrophorèse horizontaux (HPE tower, serva), simplifient et réduisent les temps de manipulations de l'utilisateur.

L'avantage majeur de cette technique est sa forte résolution. De ce fait, il n'est pas nécessaire d'avoir un spectromètre de masse hautement performant pour analyser les spots de gel car ils ne contiennent que quelques protéines et la quantification n'est pas basée sur l'analyse en masse.

B. Stratégies de quantification globale sans marquage

Les approches sans marquage ou « *label-free* » [102] en anglais, consistent à analyser en spectrométrie de masse les peptides issus de la digestion (trypsique) des protéines. L'information quantitative peut ensuite être obtenue à partir des courants d'ions (XIC) générés par les peptides (MS1) ou leurs fragments (MS2) ou à partir des spectres MS2 générés par peptide. Ces méthodes sont plus simples et rapides à mettre en place que les méthodes de marquage, et ne requièrent pas d'analyse préalable (sauf pour le XIC à partir de données acquises en DIA). Pour ces méthodes, les différents échantillons à comparer sont analysés séparément, les uns après les autres, il n'est pas possible de réaliser de multiplexage (sans marquage, il est en effet impossible de déterminer de quel échantillon provient chaque signal).

B.1. Méthode de comptage des spectres et des peptides

❖ La méthode de comptage de spectres (aussi appelée « quantification MS2 ») se base sur l'hypothèse que plus un peptide est abondant, plus il sera sélectionné pour la fragmentation et donc plus il générera de spectres MS2. Le principe de cette stratégie est donc de compter le nombre de spectres MS/MS acquis par protéine [103] (voir Figure III-6).

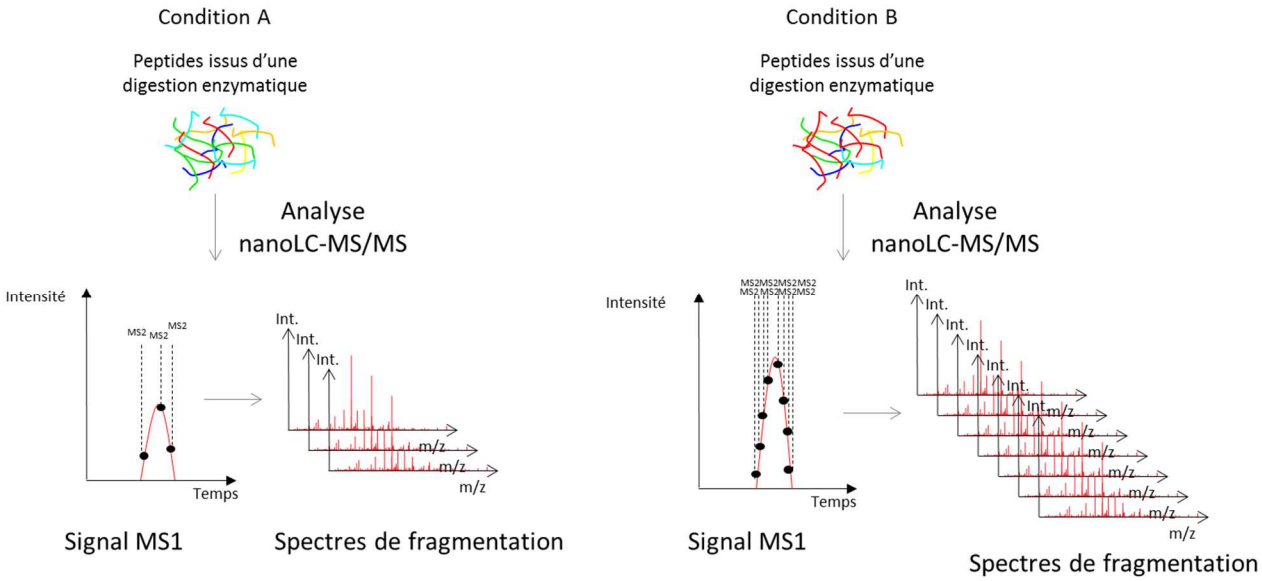


Figure III-6 : Représentation schématique de la méthode de quantification sans marquage de comptage de spectres.
Plus un peptide est abondant, plus il sera sélectionné pour être fragmenté et donc plus il générera de spectres MS/MS.

Les grandes protéines vont avoir tendance à donner lieu à plus de peptides issus de la digestion enzymatique et donc à d'autant plus de spectres (sans pour autant être plus abondantes). Afin de corriger ce biais, un indice peut être calculé : le NSAF (pour *Normalized Spectral Abundance Factor*) [104], qui consiste à diviser le nombre de spectres (SpC) acquis pour une protéine par la taille de cette protéine (L), divisé par la somme des SpC/L de toutes les protéines de l'expérience. L'intérêt de ce facteur de normalisation est de pouvoir comparer différentes protéines (et non la même protéine entre différents échantillons), comme dans l'étude [105], où les auteurs ont comparé l'abondance de différentes sous-unités d'un même complexe.

Pour cette stratégie, le temps d'exclusion dynamique doit être largement réduit pour que les peptides soient suffisamment sélectionnés afin d'observer des différences significatives. Autrement, tous les peptides seraient sélectionnés un faible nombre de fois, et le nombre de spectres MS2 par peptide ne serait pas représentatif de son abondance et la quantification serait biaisée. De ce fait, on ne peut quantifier un grand nombre de peptides/protéines, car les ions les moins abondants (dont le signal est donc peu intense) ne sont jamais sélectionnés.

De plus, l'effet de « sous-échantillonnage » est également une limitation à cette méthode. Le sous-échantillonnage est le fait que certains peptides, peu intenses, vont être sélectionnés dans une analyse et pas dans un autre, malgré leur présence dans l'échantillon. Dans ce cas, on conclurait alors, à tort, l'absence du peptide dans la deuxième analyse.

La méthode de comptage de peptides consiste à compter le nombre de peptides identifiés par protéine, en se basant sur l'hypothèse que plus une protéine est abondante, plus on en détectera de peptides. Afin de corriger pour la taille de la protéine, un indice d'abondance protéique (PAI pour *Protein Abundance Index*) [106] peut être calculé, en divisant le nombre de peptides identifiés par le nombre de peptides trypsiques théoriques. Ici encore, l'intérêt est de pouvoir comparer différentes protéines. C'est le cas par exemple en stratégie 2D-DIGE, le PAI des protéines identifiées dans un spot est calculé pour déterminer la/les protéines majoritaires.

B.2. Méthode d'extraction des courants d'ions (XIC)

B.2.1. En mode DDA

B.2.1.1. Principe

La méthode d'extraction des courants d'ions (ou XIC pour *eXtracted Ion Current*, aussi appelé « XIC MS1 ») [107] se base sur le fait que plus un peptide est abondant, plus son signal en masse sera important. L'information quantitative est obtenue pour chaque peptide par extraction des courants d'ions (voir Figure III-7), à la masse et au temps de rétention du peptide, suivi de l'intégration de l'aire sous la courbe (AUC) des ions P, P+1 et P+2.

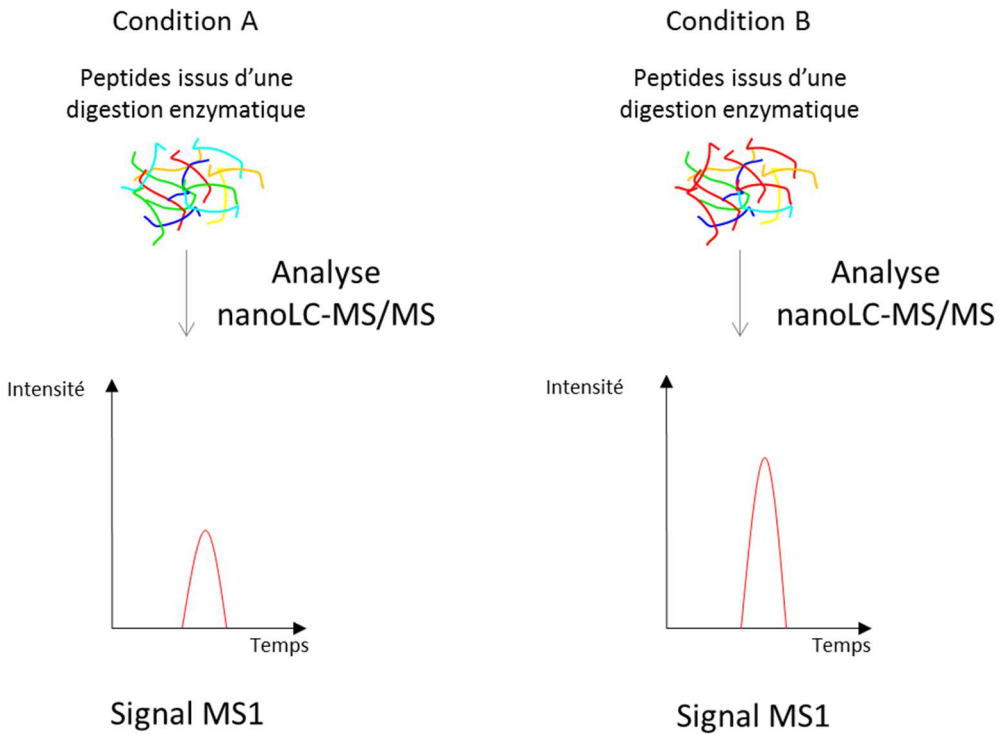


Figure III-7: Représentation schématique de la méthode de quantification sans marquage d'extraction des courants d'ions.
Plus un peptide est abondant, plus l'aire sous la courbe sera importante.

Il existe deux types d'approche pour la quantification sans marquage d'extraction des courants d'ions, qui dépendent du logiciel utilisé [108]:
- Une première méthode consiste à ne considérer que les peptides déjà identifiés et validés.

- Une deuxième méthode consiste à extraire les courants d'ions de tous les composés détectés qui présentent un profil isotopique peptidique. Ainsi, on obtient une information quantitative au niveau MS de tous les ions avant d'interpréter les spectres MS/MS pour obtenir une information qualitative.
Dans le paragraphe suivant, sont présentés deux logiciels qui utilisent l'une ou l'autre de ces approches.

B.2.1.2. Solutions logicielles

❖ De nombreuses solutions logicielles existent, comme par exemple MassChroQ [109] (qui peut quantifier une liste de peptides donnée ou travailler à partir de listes de masses et temps de rétention, il permet d'aligner les temps de rétention et est notamment capable de gérer des données préfractionnées), MFPaQ [110] (qui permet de quantifier les peptides préalablement identifiés, et de normaliser les données quantitatives issues d'un préfractionnement), Proline (en développement, pas encore publié), Skyline [111] (qui permet de quantifier les peptides identifiés, sans traitement postérieur des données quantitatives), ou encore Maxquant [112] (qui permet de quantifier les peptides identifiés, de normaliser les données quantitatives issues d'un préfractionnement). Tous ces logiciels possèdent des qualités indéniables. Notre laboratoire ayant pour l'instant décidé d'investir ses efforts dans l'utilisation Skyline et Maxquant (et très prochainement Proline), et ces deux logiciels permettant de répondre aux questions qui nous ont été posées par nos collaborateurs biologistes (moyennant le cas échéant quelques développements bioinformatiques), ce sont les logiciels Skyline et Maxquant qui ont été utilisés dans cette thèse et sont donc décrits plus précisément ci-après.

❖ Le logiciel Skyline, qui a originellement été développé pour la SRM [113] puis étendu au XIC MS1 [111], fonctionne selon la première approche présentée dans le paragraphe précédent ; seuls les peptides identifiées et validées sont quantifiés. Il faut donc d'abord réaliser une recherche dans les banques de données (avec Mascot par exemple) pour identifier les peptides puis les valider.
L'utilisateur renseigne donc Skyline avec :

- La liste de peptides à quantifier, dont il va calculer la masse (à partir de la séquence fournie) ;
- Une librairie de spectres qui contient les séquences des peptides, leur masse et le temps de rétention auquel le spectre MS/MS a été acquis (cette librairie est le plus souvent l'export du résultat Mascot, au format « .dat ») ;
- Les données brutes.

Skyline va donc faire le lien entre ces trois informations : il retrouve la séquence à quantifier dans la librairie et extrait le courant d'ions à la masse de ce peptide à partir données brutes dans une fenêtre (à définir) autour du temps de rétention. Skyline calcule un « idot product » (idotp) qui peut aider à vérifier les résultats ; il caractérise, pour un peptide donné, la proportion relative des aires P, P+1 et P+2 par rapport à la distribution théorique attendue (qui est calculée d'après l'abondance naturelle de chaque isotope). Cette valeur d'idotp est comprise entre 0 et 1 ; plus elle est proche de 1, plus la distribution isotopique est proche de la valeur théorique. Cela permet de vérifier que le signal intégré est correct.
L'avantage de ce logiciel est la possibilité de vérifier manuellement l'intégration des aires sous la courbes, et la possibilité de supprimer des signaux si on les estime peu robustes (faible intensité,

proche du bruit de fond…). En revanche, cette étape peut devenir très chronophage, étant donné le nombre d'analyses et de peptides généralement identifiés (plusieurs milliers) lors d'une analyse globale, mais elle reste indispensable pour la qualité et la fiabilité des résultats.
Enfin, Skyline permet seulement d'obtenir les valeurs d'aires sous la courbe des courants d'ions extraits des peptides préalablement identifiés et ne fait pas de traitement sur les données quantitatives (normalisation, rassemblement des valeurs d'abondance des peptides pour obtenir une valeur par protéine).

❖ Le logiciel MaxQuant [112] fonctionne selon la deuxième approche : tous les ions sont quantifiés.
Contrairement à Skyline, MaxQuant permet un traitement automatisé des données, de l'identification jusqu'à la normalisation des données. MaxQuant travaille directement à partir des données brutes et rend une valeur d'abondance relative, directement au niveau protéique (l'information quantitative au niveau peptidique est également disponible).
Dans MaxQuant est implémenté un algorithme de recherche appelé Andromeda, qui permet d'identifier les peptides selon la stratégie d'empreintes de fragments peptidiques (expliquée plus haut).
L'option LFQ (pour « *Label-Free Quantification* ») permet de normaliser les données quantitatives, et notamment issues d'un préfractionnement des protéines. Il s'agit de « rassembler » les abondances d'un même peptide, qui se retrouve en général réparti dans plusieurs fractions adjacentes, pour un même échantillon, puis des différents peptides assignés à une même protéine, avant de pouvoir comparer ensuite l'abondance de cette protéine entre différentes conditions. Or, la séparation en fractions peut être légèrement différentes entre les échantillons. De ce fait, un facteur de normalisation doit être défini pour chacune des fractions avant de pouvoir sommer les intensités de chaque peptide (entre les différentes fractions). MaxQuant, et plus particulièrement l'option LFQ, propose de pallier à ce problème en sommant les intensités de chacun des peptides tout en ajustant la valeur des facteurs de normalisation de manière à obtenir une variation minimale à l'échelle du protéome (modèle de régression des moindres carrés) [114].
Pour obtenir une information quantitative par protéine, l'option LFQ rassemble les valeurs d'abondance des peptides [114]. Pour ce faire, un rapport d'intensités des peptides est calculé entre tous les échantillons pour chaque comparaison deux à deux. Par la suite, le calcul de la médiane des rapports peptidiques (pour chaque peptide d'une protéine) permet d'obtenir les rapports protéiques pour chaque comparaison deux à deux. Un modèle de régression des moindres carrés est utilisé pour obtenir les valeurs d'abondance protéique.
La limitation majeure de MaxQuant réside dans le fait qu'on ne peut pas revoir et corriger les données. Aucune validation manuelle des intégrations n'est possible, contrairement à Skyline. De ce fait l'utilisateur est obligé de « faire confiance » au logiciel, sur toutes les étapes du traitement de données.

B.2.1.3. Avantages et limitations de la stratégie « XIC MS1 »

La méthode XIC MS1 requière l'utilisation d'un instrument haute-résolution pour obtenir une quantification précise. En effet, la détermination précise des courants d'ions de chaque peptide est

essentielle pour cette quantification. A faible résolution, les courants d'ions sont souvent interférés par les courants d'ions des peptides voisins, ce qui fausse l'information quantitative [115].

L'avantage de la méthode « label-free » d'extraction des courants d'ions (XIC) est l'absence de limite en termes de nombre d'échantillons à comparer, contrairement aux méthodes de marquage. En revanche, le fait d'analyser individuellement les échantillons à comparer peut entraîner des biais, d'autant plus s'il y a beaucoup d'analyses.. En effet, dans le cas de très grand nombres d'échantillons à analyser, les analyses peuvent durer plusieurs jours à plusieurs semaines. Plus ce temps est long, plus il y a de risques de « dérive instrumentale » en termes de temps de rétention ou de sensibilité des instruments. Pour « contrôler » ces possibles biais techniques, il convient donc généralement de restreindre le nombre d'échantillons à analyser au sein d'une expérience. Nous avons pu constater que nos instruments restaient stables sur des périodes d'environ 2 semaines. Des temps plus longs n'ont pas été testés. Pour vérifier cette stabilité instrumentale, des « contrôles qualité » sont généralement mis en place. Il est ainsi possible d'analyser un échantillon contrôle et de vérifier la stabilité des valeurs d'abondance des peptides entre plusieurs analyses de ce même échantillon ; ou encore d'ajouter à chaque échantillon un mélange de peptides synthétiques et de contrôler la stabilité de leurs temps de rétention au cours du temps.
Enfin, pour pallier à ces dérives et aux biais techniques, il est essentiel de normaliser les données quantitatives à l'issue d'une approche de quantification XIC. Certains logiciels (e.g. Maxquant [114], MFPaQ [110]) permettent une normalisation intégrée, il existe également des outils comme Normalyzer [116].

De plus, contrairement à la méthode de comptage de spectres, la méthode par XIC n'est pas limitée par l'effet de sous-échantillonnage. En effet, si un peptide est fragmenté dans une analyse et donc identifié, on peut « transférer » cette identification aux autres analyses grâce au ratio m/z et au temps de rétention, voir Figure III-8. On peut donc extraire le courant d'ions d'un peptide d'une analyse dans laquelle il n'a pas été fragmenté. Cependant, les temps de rétention peuvent être légèrement décalés entre les échantillons, il est alors nécessaire de les ré-aligner pour quantifier la même espèce dans tous les échantillons. Des fonctions de réalignement existent dans plusieurs logiciels (e.g. MassChroQ [109], MFPaQ [110], Maxquant [114]).

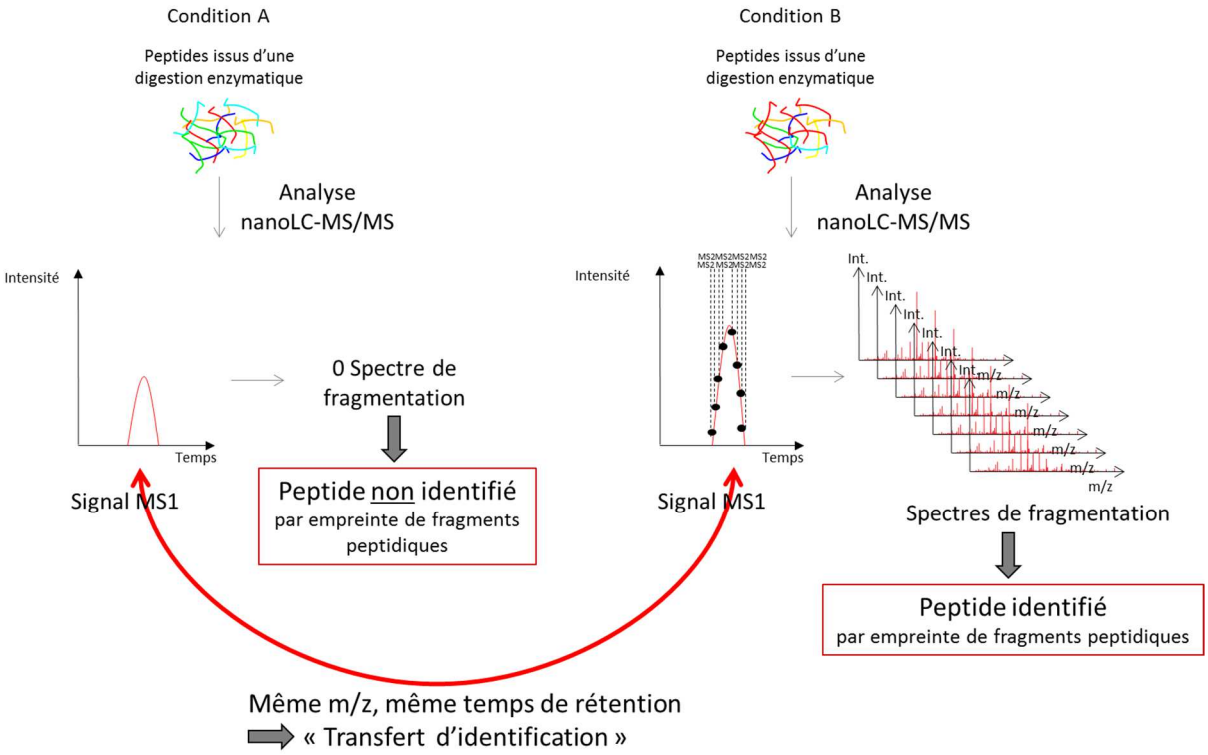


Figure III-8 : Représentation schématique de la possibilité de « transférer » l'identification d'un peptide, grâce au m/z au temps de rétention, lorsque le peptide n'a pas été fragmenté.

Cela permet de quantifier un peptide non fragmenté par la méthode XIC, tandis que c'est impossible par la méthode de comptage de spectres.

Enfin, cette méthode est facile à mettre en place et applicable à tous types d'échantillons. En revanche, le traitement des données reste un point complexe de cette stratégie, dans la mesure où il faut interpréter les données spectrales de milliers de peptides pour finalement obtenir une valeur d'abondance relative par protéine. Il existe plusieurs stratégies pour obtenir une valeur de quantification par protéine à partir des valeurs quantitatives des peptides. Les valeurs d'AUC de tous les peptides d'une protéine peuvent être sommées, moyennées ou la médiane peut être calculée [117], certains choisissent d'utiliser les valeurs d'abondance des 3 peptides les plus intenses [118].

Les peptides d'une protéine donnée peuvent avoir des profils d'abondance différents, c'est-à-dire qu'ils ne varient pas de la même façon entre les conditions. Cela peut être dû à la présence de modifications post-traductionnelles (MPT), ou bien à des peptides de très faible intensité, dont le signal peut être interféré avec le bruit de fond. Pour s'affranchir de ces disparités, il est utile de prendre en compte les rapports peptidiques individuels entre différents échantillons, ce qui est le cas de l'option LFQ qui est expliquée en Partie IChapitre IIIII.B.2.1.2. B.2.1.2Par ailleurs, il est également possible d'évaluer la corrélation des peptides d'une protéine et d'éliminer les peptides mal corrélés, c'est-à-dire qui présentent un profil différent. Un exemple de cette méthode est présentée en page 92.

Un dernier inconvénient pour le traitement des données réside dans le grand nombre de valeurs manquantes (VM) obtenues [119]. Les valeurs manquantes sont des valeurs quantitatives absentes d'un ou plusieurs échantillons pour une protéine donnée. Ces VM peuvent venir de peptides proches de la limite de détection et donc aléatoirement détectables. Il n'existe pas une méthode universelle

pour pallier à ces valeurs manquantes [120], certains les imputent, par exemple avec une valeur de bruit de fond, la médiane des abondances de toutes les protéines, la plus petite valeur d'abondance, ou encore sur la base des valeurs observées en transcriptomique [121, 122]. Il est également possible de filtrer les données pour ne conserver que les protéines qui ne présentent aucune valeur manquante (mais cela revient à réduire le jeu de données) [120], ou alors qui ont suffisamment de valeurs valides (i.e. non manquantes), ce qui réduit dans une moindre mesure le jeu de données.

B.2.2. En mode DIA

Comme présenté en page 21, le mode d'acquisition DIA permet de fragmenter tous les peptides sur une fenêtre de masse donnée. Par exemple, La méthode SWATH™ (*Sequential Windowed Acquisition of All Theoretical fragment ion mass spectra*), développée par SCIEX [123] consiste à fragmenter les peptides dans des fenêtres de masses successives, et à les quantifier par intégration des courants d'ions des fragments. Ces analyses sont généralement réalisées sur des instruments de type TripleTOF® de chez SCIEX.

Pour quantifier des peptides par SWATH™, il est nécessaire de disposer d'une librairie spectrale préalablement acquise en mode DDA (acquisition réalisée en amont sur les mêmes échantillons ou grâce aux données publiquement disponibles). Cette librairie permet de retrouver les signaux des fragments de chaque peptide identifié en DDA grâce à leur temps de rétention, la masse de chaque peptide et fragment ainsi que les rapports d'intensités entre les différents fragments d'un même peptide. Ces informations permettront alors d'extraire le signal des fragments. Si un gradient différent a été utilisé pour acquérir la librairie spectrale et les données en DIA, l'utilisation de peptides synthétiques permet de normaliser les temps de rétention entre les deux analyses et d'extraire les signaux de chaque peptide. Il existe plusieurs logiciels pour traiter ce type de données, tel que Skyline, Peakview, Spectronaut, OpenSWATH [124].

L'utilisation de signaux MS/MS pour la quantification rend cette technique plus spécifique et sensible par rapport à l'utilisation des signaux MS1 [123] (car les interférences sur les courants d'ions des fragments sont moindres grâce aux fenêtres d'isolement).

Une autre méthode, la MS^e développée par Waters [125] consiste à fragmenter l'ensemble de la gamme de masses à chaque instant, et à quantifier les peptides par intégration des courants d'ions (extraction à partir de spectres MS) des peptides.

III. Conclusion sur les stratégies de quantification

Finalement, il existe de nombreuses solutions pour la quantification des protéines, plus ou moins onéreuses, plus ou moins chronophages, dont la sensibilité et la spécificité varient. Il est essentiel de choisir la méthode la plus adaptée à son type d'échantillons, mais également à la question posée. Ce choix fait souvent l'objet de compromis, il n'existe pas de solution universelle [126].

Plusieurs critères sont à prendre en compte dans le choix d'une méthode de quantification :

- Le contexte de l'étude va diriger l'analyse vers une méthode ciblée ou globale. Plus l'idée de départ sera précise, plus la quantification le sera.

❖ Quantification absolue ou relative : certaines applications requièrent une quantification absolue des protéines, pour d'autres une comparaison relative de plusieurs conditions peut suffire à obtenir une réponse. Il faut tenir compte du fait que la quantification absolue a un coût non négligeable (les peptides standards utilisés étant très purs et quantifiés précisément, ils sont très onéreux).

❖ Le parc instrumental : suivant les spectromètres de masse dont l'on dispose, on va plutôt se diriger vers une stratégie en particulier. Par exemple, pour mettre en place une stratégie SRM, il est préférable de disposer d'un triple quadripôle. Si les spectromètres de masse disponibles sont peu performants (faible résolution, peu rapide…), on va préférer une méthode telle que la 2D-DIGE qui permet de décomplexifier les échantillons. Si l'on veut mettre en place une méthode de type « label-free », il est nécessaire d'avoir un instrument très performant (type Q-Orbitrap™, Q-ToF, Triple ToF…).

❖ Le nombre d'échantillons à comparer est un paramètre important à prendre en compte. Les méthodes de marquage qui permettent de multiplexer restent limitées en nombre de conditions comparables, étant donné le nombre de marqueurs disponibles [127]. Les méthodes de « label-free » sont plus adaptées, bien qu'il faille prendre en considération le temps d'analyse globale. Par exemple, il est difficile d'envisager un préfractionnement des protéines en amont de l'analyse LC-MS/MS si le nombre de conditions à comparer est déjà important.

Enfin, quelle que soit la méthode choisie, des tests statistiques sont généralement réalisés pour déterminer les différences significatives d'expressions protéiques entre les conditions à comparer. D'un point de vue statistique, plus il y a de réplicas par condition, plus les résultats ont du poids. Cependant, il n'est pas toujours possible d'obtenir un très grand nombre de réplicas, suivant le modèle d'étude utilisé.

Conclusion Partie I

Finalement, à chaque étape de l'analyse protéomique, différentes solutions sont envisageables, à choisir en fonction de diverses contraintes :

- La nature des échantillons peut imposer une méthode d'extraction spécifique, une décomplexification plus ou moins poussée ;

- Le nombre de réplicas peut limiter le degré de décomplexification ;

- La question posée, qui est le point le plus important, va imposer la stratégie globale : veut-on identifier un maximum de protéines présentes dans l'échantillon, obtenir un maximum de couverture de séquence, quantifier de manière ciblée ou globale, absolue ou relative ;

- Le parc instrumental disponible va également diriger vers une méthode plutôt qu'une autre, avec des instruments peu rapides, peu résolutifs, il est difficile d'envisager une méthode de quantification en « label-free » par exemple ;

- Les solutions logicielles disponibles : il faut être capable de traiter le très grand nombre de données générées (surtout par les appareils dernière génération) aussi bien pour l'interprétation des spectres et recherche dans les banques de données (elles-mêmes de plus en plus importantes) que pour l'étape de quantification, normalisation, statistiques…

Il est important de noter que **chacune des étapes de l'analyse protéomique sont liées**, et que chaque étape dépend de plusieurs contraintes ; du choix de la dernière étape découle le choix de la première.

Partie II : Résultats

Chapitre I : Analyse protéomique quantitative chez un organisme séquencé

Partie II : Résultats

Chapitre I : Analyse protéomique quantitative chez un organisme séquencé

Les trois projets présentés dans ce chapitre ont été réalisés en collaboration avec le Dr. François Criscuolo du DEPE, Université de Strasbourg.

En écologie évolutive, la théorie des traits d'histoire de vie précise que les processus physiologiques entrent en compétition pour une quantité limitée en ressources disponibles [128, 129]. Par conséquent, si deux processus partagent les mêmes ressources, alors toute augmentation de l'allocation desdites ressources à l'un de ces processus se fera au détriment de l'autre. C'est pourquoi de tels compromis peuvent entrainer des effets délétères, ci-après désignés comme des « coûts ».

Le maintien d'un système immunitaire efficace ou l'activation de celui-ci en cas d'infection est couteux [130-134], et peut même avoir un coût énergétique très important (jusqu'à plus de 50% du métabolisme de base chez un passereau comme le moineau domestique par exemple [130]). Par ailleurs, les coûts sont dépendants de l'âge de l'individu exposé [135-137]. En effet, l'efficacité du système immunitaire diminue avec l'âge, un phénomène connu sous le nom d'immunosenescence [135, 138-140]. Nous nous intéresserons donc ici aux réponses hépatiques et spléniques induites en fonction de l'âge chez la souris, lors d'une activation du système immunitaire. Nous aborderons aussi les fonctions régulatrices de l'immunité, en étudiant les coûts de l'activation du système immunitaire lorsque le contrôle de celui-ci est perturbé.

I. Etude des marqueurs spléniques et hépatiques de l'immunosénescence

A. Contexte biologique

L'étude des coûts liés à l'activation du système immunitaire a constitué le sujet de nombreux travaux au cours des dernières années. La plupart de ces études se sont consacrées à la dépense énergétique induite par la réponse immunitaire et aux compromis qui en découlent [141]. Par exemple, une réponse aigue (induite par une infection par exemple) efficace ou encore un entretien chronique important du système induisent un coût énergétique élevé ; ce qui pourrait être trop coûteux à entretenir et délétère pour d'autres fonctions.

Il existe cependant un autre mécanisme qui peut potentiellement contraindre l'activité du système immunitaire. En effet, une fois activé, celui-ci peut induire des coûts auto-immuns, notamment dus au défaut de détection entre le soi et le non-soi, à l'activation de cellules de la réponse innée non-spécifique (macrophages, neutrophiles…) et à la production d'espèces réactives de l'oxygène (ERO) pouvant induire des dommages oxydatifs [142]. Plusieurs études récentes ont pu ainsi mettre en

évidence l'augmentation du stress oxydatif en tant que coût de l'immunité. Cependant, ces études ont été réalisées à partir d'échantillons sanguins et sont quasi-exclusivement centrées sur des mesures systémiques globales. Nous n'avons que peu d'information sur la résultante d'une activation immunitaire sur les mécanismes enzymatiques antioxydants cellulaires, sur le fonctionnement mitochondrial, ou encore sur la façon dont le métabolisme énergétique global de l'organisme peut être modifié. De plus, la dégradation du fonctionnement du système immunitaire avec l'âge (immuno-sénescence) est sensée augmenter l'importance des dommages auto-immuns liés à la suractivation du système immunitaire [143, 144].

B. Etude des marqueurs hépatiques

B.1. Introduction

Nous nous sommes intéressés aux réponses **hépatiques** induites en fonction de l'âge chez la souris, lors d'une activation du système immunitaire innée avec un antigène non-pathogène, le lipopolysaccharide (LPS, élément de membranes bactériennes). Le foie est en effet un carrefour métabolique important et doit constamment gérer des charges antigéniques.

Le protocole complet est détaillé dans la publication présentée dans la suite de ce chapitre.

Schéma expérimental :

Nous avons sélectionné deux groupes de souris (voir Figure I-1) : un groupe « jeune » constitué d'individus de 3 mois, et un groupe « plus agé » dont les individus étaient âgés d'un an. La moitié des individus de chaque groupe a reçu une injection de LPS afin d'activer le système immunitaire tandis que l'autre moitié a reçu une injection contrôle de tampon phosphate salin (PBS), qui n'avait pas d'effet sur l'organisme. Chaque groupe était constitué de 4 individus.

L'hypothèse ici est que l'immunosenescence n'induit pas des coûts liés à une dépense énergétique augmentée mais que ces coûts pourraient passer par des mécanismes pro-inflammatoires et un déséquilibre de la balance oxydative. Dans ce cas, on s'attend à ce que les vieux individus recevant une injection de LPS payent un coût plus élevé que les jeunes. Si, en revanche, la dépense énergétique est le principal facteur régulateur, alors les individus plus âgés qui ont des capacités amoindries en termes de dépense vont avoir une réponse moins efficace et donc moins délétère que celle des jeunes.

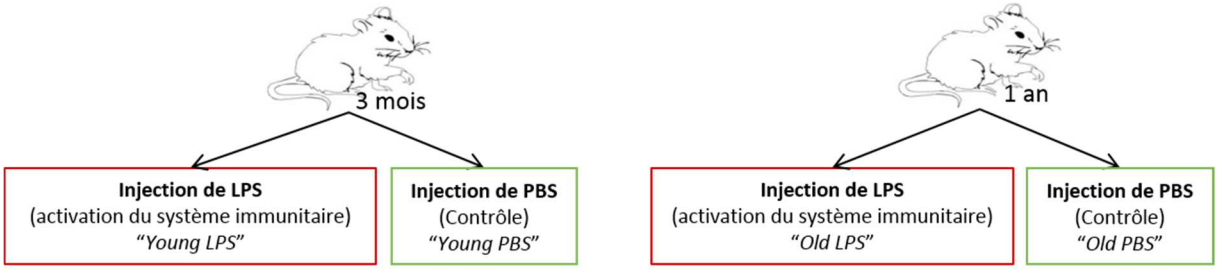


Figure I-1 : Schéma expérimental du traitement appliqué aux souris.

Pour cette comparaison, nous avons mis au point une stratégie 2D-DIGE, dont le principe est expliqué en page 43. A l'époque où ces analyses ont été réalisées, nous ne disposions pas d'instruments

suffisamment performants pour réaliser une analyse de type *label-free*, nous nous sommes donc orientés vers une stratégie 2D-DIGE.

B.2. Analyse des échantillons

Les extraits protéiques de foie ont été aléatoirement marqués au Cy3 ou Cy5 et un pool de tous les échantillons a été marqué au Cy2. Ensuite ont été mélangés un échantillon marqué au Cy3, un au Cy5 et le pool au Cy2 (50 µg chacun), puis cette mixture protéique a été séparée sur gel 2D. L'analyse d'image a permis d'obtenir les volumes densitométriques de chaque spot protéique. Puis des analyses statistiques (ANOVA) ont permis de mettre en évidence les spots différentiels. Ces derniers ont été excisés et analysés en spectrométrie de masse (HCT™Plus ion trap (Bruker Daltonics) pour identifier les protéines présentes et responsables de la différence.

B.3. Publication

Journal of Proteomics 135 (2016) 181–190

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Differential proteomics reveals age-dependent liver oxidative costs of innate immune activation in mice

Marine I. Plumel [a,c], Margaux Benhaim-Delarbre [a,c], Magali Rompais [a,c], Danièle Thiersé [a,c], Gabriele Sorci [d], Alain van Dorsselaer [a,c], François Criscuolo [b,c,1], Fabrice Bertile [a,c,*,1]

[a] Institut Pluridisciplinaire Hubert Curien, Département Sciences Analytiques, CNRS UMR7178, 25 rue Becquerel, 67087 Strasbourg Cedex 2, France
[b] Institut Pluridisciplinaire Hubert Curien, Département d'Ecologie, Physiologie et Ethologie, CNRS UMR7178, 23 rue Becquerel, 67087 Strasbourg Cedex 2, France
[c] Université de Strasbourg, 4 rue Blaise Pascal, F-67081 Strasbourg Cedex, France
[d] Biogéosciences, CNRS UMR6282, Université de Bourgogne, 6 boulevard Gabriel, F-21000 Dijon, France

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form 1 September 2015
Accepted 7 September 2015
Available online 12 September 2015

Keywords:
Immunosenescence
Liver
Proteomics
Oxidative stress
Ageing
Mice

ABSTRACT

Individual response to an immune challenge results from the optimization of a trade-off between benefits and costs of immune cell activation. Age-related immune disorders may have several mechanistic bases, from immune cell defects to chronic pro-inflammatory status and oxidative imbalance, but we are still lacking experimental data showing the relative importance of each of these mechanisms. Using a proteomic approach and subsequent biochemical validations of proteomics-derived hypotheses, we found age-dependent regulations in the liver of 3-months and 1-year old-mice in response to an acute innate immune activation. Old mice presented a chronic up-regulation of several proteins involved in pathways related to oxidative stress control. Interestingly, these pathways were weakly affected by the innate immune activation in old compared to young individuals. In addition, old mice suffered from lower glutathione-S-transferase activity and from higher oxidative damage at the end of the experiment, thus suggesting that they paid a higher immune-related cost than young individuals. On the whole, our data showed that a substantial fraction of the liver costs elicited by an activation of the innate immune response is effectively related to oxidative stress, and that ageing impairs the capacity of old individuals to control it.

Significance: Our paper tackles the open question of the cost of mounting an innate immune response. Evolutionary biologists are familiar since a long time with the concept of trade-offs among key traits of an organism, trade-offs that shape life history trajectories of species and individuals, ultimately in terms of reproduction and survival. On the other hand, medicine and molecular biologists study the intimate mechanisms of immune senescence and underline that oxidative imbalance is probably playing a key role in the progressive loss of immune function with age. This paper merges the two fields by exploring the nature of the cellular pathways that are mainly affected by age when the innate immunity is triggered. To this purpose, a proteomic approach was used to explore liver protein profiles and provide for the first time convincing data supporting the idea that oxidative stress constitutes a cost of innate immune response in old mice, possibly contributing to senescence. Proteomics-derived hypotheses were furthermore validated using biochemical assays. This paper therefore illustrates the added value of using proteomics to answer evolutionary biology questions, and opens a promising way to study the inter-specific variability in the rates of immune-senescence.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Senescence is a multi-level phenomenon that starts at the level of the cell, and has ultimate consequences on the entire organism. It has been defined as the accumulation of unrepaired damage on biomolecules and cells causing the progressive decline of both organismal functions and reproductive and survival rates of individuals over time [1]. Among the physiological functions shown to be affected by age, senescence of the immune system has been largely studied [2–4], and old individuals are often characterized by chronic or repeated infections, inflammatory diseases and autoimmune disorders [5]. In addition, the finding that senescent fibroblasts actually express inflammatory genes [6], suggests that even non-immune cells may contribute to mal-adaptive immunity in old-age individuals.

Interestingly, several studies from the molecular biology and medicine fields recently underlined that tackling the question of immune-senescence from an evolutionary point of view may help to better

* Corresponding author at: Institut Pluridisciplinaire Hubert Curien, Département Sciences Analytiques, CNRS UMR7178, 25 rue Becquerel, 67087 Strasbourg Cedex 2, France.
[1] Equal contributors, shared seniorship of the paper.

http://dx.doi.org/10.1016/j.jprot.2015.09.008
1874-3919/© 2015 Elsevier B.V. All rights reserved.

understand which mechanisms are important in the progressive impairment of the immune system [7, 8], and also whether the change in the immune functioning with age is mal-adaptive or not. More particularly, the well-known evolutionary concept of antagonistic pleiotropy [9] has been successfully applied to explain how some cell pathways may contribute to ageing. Those pathways have deleterious impact at old ages, but are still preserved by natural selection because of their beneficial effects on early life traits (e.g. TOR pathways, [10]), and ultimately on overall individual fitness. Antagonistic pleiotropy may also well explain why the control of the immune system is compromised with age, as strong responses are favoured by natural selection because of their role in fighting infectious diseases, but pro-inflammatory processes can also be harmful in terms of increased oxidative stress and immunopathology [11, 12].

This idea of progressive accumulation of damage while growing old has also been theorized in the context of evolutionary biology, leading to the *Disposable Soma* theory [13]. While taking over the global idea of the *Antagonistic Pleiotropy* theory, Kirkwood gave a mechanistic explanation related to metabolism and direct negative impact of one function on another through the production of reactive oxygen species (ROS). ROS are inevitably produced by mitochondria when processing reduced co-enzymes to ATP. Because functions compete for limited resources, this may reduce the investment an organism is currently doing in somatic maintenance, for example in buffering oxidants. This is even more critical when the activated physiological functions entail an increase in energy expenditure or produce reactive species, putatively modulating oxidative balance and ultimately ageing [14]. Such more or less clear relationships have been highlighted in model and non-model species [15–21], leading to the idea that oxidative stress could be one of the mediating mechanisms sustaining life-history trade-offs such as those among reproduction, growth or somatic maintenance and longevity [22].

Within somatic maintenance, the immune system has a preponderant role. It preserves the organism both from external (pathogens) and internal (cancer) threats. Because immune response potentially implies the production of ROS both *via* an increased energy expenditure [23, 24] and/or the activation of immune cells [25, 26], mounting an immune response is likely to be costly and to lead to energy and oxidative based trade-offs. For instance, innate immunity relies on production of nitric oxide and superoxide by macrophages [27, 28] which may have non-specific deleterious impact on host cells [29]. However, individuals are not paying an identical energy cost when responding to an immune challenge, and old individuals may have to face more critical immune-associated trade-offs [24]. Such an observation suggests an ultimate cost of mounting an immune response, an idea further characterized in several taxa where individuals challenged with an immune treatment exhibit a reduced survival rate [30, 31].

We challenged one specific arm of the host immune system of mice using bacterial lipopolysaccharide (LPS), the innate immunity. LPS is recognized by different innate cell receptors (macrophage scavenger receptors, MARCO receptors or toll-like receptor 4) [32], and triggers an innate inflammatory response. Timely innate response to LPS injection guarantees the early clearance of bacterial endotoxin and prevents excessive immune inflammation ultimately leading to a septic shock [33]. Previous studies have shown an increase of oxidative stress in LPS treated individuals using blood markers [34], and successfully characterized proteomic specific patterns induced by LPS exposure [35]. Based on this knowledge, one objective was therefore to decipher here liver regulations that are elicited in response to an immune challenge in mice. Indeed, the liver is a key organ for the maintenance of metabolic homeostasis, but it also has to constantly deal with antigenic loads [36, 37], and oxidative stress plays an important role in the pathogenesis of many liver diseases [38]. More particularly, the aim was to establish whether age-related changes in protein profiles could support the idea that old individuals actually pay a higher cost in terms of deleterious reactions triggered by activation of the immune system (*e.g.* like oxidative imbalance) than young individuals. To this purpose, we took advantage of the benefits a global analytical strategy like proteomics can bring to the evolutionary ecology field [39].

2. Materials and methods

2.1. Experimental procedures

The experiment was conducted using eight 3 month-old (young) and eight one year-old (old) C57BL/6J male mice, reared in our laboratory under constant temperature (24 ± 2 °C) and photoperiod (13:11 L:D cycle), with free access to a standard chow diet consisting (by weight) of 21.4% proteins, 51.7% carbohydrates and 5.1% fat (SAFE A03) and water. Each group was randomly divided in two subgroups (4 mice each kept in individual cages) where animals received a single intraperitoneal injection of either 25 μg/kg body mass of LPS from *Escherichia Coli* (serotype 055-B5, patch 011M400IV, Sigma Aldrich), or phosphate buffered solution (PBS). This LPS dose is largely below the lethal dose for young and 1 year-old mice (see [40]). None of the injected mice died after the injection. Body mass (±0.1 g) of individuals was measured just before the injection and animal death. Mice were sacrificed 24 h after injection by cervical dislocation and their liver was quickly collected, snap-frozen in liquid nitrogen, and several sample aliquots were stored at −80 °C until biochemical and proteomic analyses. The study complied with legislation (L87-848) on animal experimentation in France and was done under the DEPE license obtained from the French Department of Veterinary Service (number G67-482-18). Dr F. Criscuolo is the holder of an animal experimentation license (no. 67–78) delivered by French authorities.

2.2. Liver proteomics

Unless otherwise specified, all chemicals and reagents were purchased from Sigma Aldrich (St. Louis, MO, USA).

2.2.1. Protein extraction

Frozen liver samples were first pulverized using a laboratory ball mill (Mikrodismembrator, Sartorius). ~10 mg of the grinded powders were then dissolved in 400 μL of a buffer composed of 8 M Urea, 2 M Thiourea, 4% Chaps, 1% dithiothreitol, Triton X100 0.5%, TLCK 0.05% and 0.02 to 2 mM protease inhibitors. After sonication on ice (10 s, 135 watts), 9 volumes of cold acetone were added, and samples were kept at −20 °C during 16 h. Proteins were pelleted by centrifugation (14 min, 4 °C, 14000 g), vacuum-dried (Speedvac, Thermoscientific) after discarding supernatants, and then dissolved in a buffer composed of 7 M Urea, 2 M Thiourea, 30 mM Tris (pH 8.5) and 4% Chaps buffer. After adjustment of the pH to 8.5, homogenization was finally completed by sonication on ice (10 s, 135 watts).

After determination of total protein concentrations using the Bio-Rad Protein Assay [41] (BioRad, Hercules, CA, USA), protein integrity and similarity of electrophoretic protein profiles was checked prior to 2D-DIGE analysis [42–44]. To do so, proteins were electrophoresed on a 12% SDS-PAGE acrylamide gel (20 μg loaded; 50 V for 30 min and then 100 V to complete migration) and stained with colloidal Coomassie blue [45] (G 250, Fluka, Buchs, Switzerland).

2.2.2. 2D-DIGE experiment

Protein samples were first labelled using a CyDye DIGE Fluor Minimal Dye Labeling Kit (GE HealthCare, Uppsala, Sweden). More precisely, 400 pmol of Cy3 and Cy5 were used to randomly label 50 μg of protein samples from the different groups, and 3.2 nmol of Cy2 were used to label 400 μg of proteins after having mixed all the samples (25 μg each; internal standard). After incubation in the dark for 30 min on ice, protein labelling was quenched by addition of 10 mM lysine and incubation in the dark for 10 min on ice. Random distribution of samples from the 4 groups (i.e. young PBS, yound LPS, old PBS and old LPS) was

done by mixing and diluting 50 μg of Cy2, Cy3 and Cy5-labelled protein samples in 400 μL of a buffer composed of 7 M urea, 2 M thiourea, 2% Chaps, 2% DTT, 2% ampholytes (Amersham Pharmacia-Biotech, Uppsala, Sweden), and a trace of bromophenol blue. Loading onto 18 cm pH3-10 non-linear immobilized pH gradient strips (IPG Ready strip, Biorad, Hercules, CA, USA), was then followed by passive rehydration over 2 h 30 min in the dark prior to active rehydration overnight by applying a voltage of 50 V using a Protean IEF cell (Biorad, Hercules, CA, USA). Isoelectric focusing (IEF) was afterwards performed until reaching a total focusing time of 85000 V h, by applying voltage gradient steps (from 0 to 200 V in 1 h, from 200 to 1000 V in 4 h, from 1000 to 5000 V in 16 h, then 5000 V for 7 h). Focused proteins were then reduced and alkylated through a first incubation of IPG strips in a buffer composed of 1% DTT, 6 M Urea, 50 mM Tris pH 8.8, 30% glycerol and 2% SDS during 30 min, and a second incubation in a buffer composed of 2.5% iodoacetamide, 6 M Urea, 50 mM Tris pH 8.8, 30% glycerol and 2% SDS during 30 min. IPG strips were then sealed onto 10% polyacrylamide SDS-PAGE gels (20 × 20 cm) with 0.5% agarose, and focused proteins were electrophoresed using a Protean II xi Cell (Biorad Hercules, CA, USA) by application of 5 mA per gel for 1 h followed by 8 mA per gel for 8 h.

Another 2D-gel was run in parallel, on which a larger amount of proteins (i.e. 1 mg of the non-labelled internal standard) was loaded. Protein spots were visualized by a colloidal blue method (see above). It was used to specifically improve quality of mass spectrometry-based protein identifications.

2.2.3. Quantitative analysis from 2D-gel images

After electrophoresis, gels were washed with water and gel images were acquired at 100 μm resolution (Ettan DIGE Imager, Ge Healthcare Uppsala, Sweden). Using Progenesis Samespots (v4.5, Nonlinear dynamics, Newcastle, UK), image quality was first controlled and all images were automatically aligned, with subsequent minor "hand-made" adjustments to improve accuracy of alignments. Background subtraction was then followed by normalization of Cy3 and Cy5 spot volumes to those of corresponding Cy2 spots, and application of a correction based on 1) the calculation of the global distribution of all Cy3/Cy2 and Cy5/Cy2 ratios and 2) the determination of a global scaling factor for all gels. Hence, any possible inter-gel variations were eliminated and accurate quantitative data were obtained.

2.2.4. nanoLC-MS/MS analyses

After automatic excision of differential protein spots (see Statistics) using an automated gel cutter (PROTEINEER sp, Bruker Daltonics, Bremen, Germany), a Massprep Station (Waters, MicroMass, Manchester, UK) was used first to apply 3 wash cycles (10 min each) in 50 μL of 25 mM NH_4HCO_3 and 50 μL of acetonitrile, followed by a dehydration step (50 μL acetonitrile, 60 °C, 5 min). Destaining was then followed by in-gel reduction (incubation at 60 °C for 30 min in 50 μL of 10 mM DTT, 25 mM NH_4HCO_3) and in-gel alkylation (incubation 30 min in 55 mM iodoacetamide, 25 mM NH_4HCO_3) of proteins using the same Massprep Station. A last washing step (10 min) in 50 μL of 25 mM NH_4HCO_3 and 50 μL of acetonitrile, followed by gel spots dehydration during 15 min in 50 μL of acetonitrile were then carried out before in-gel protein digestion (5 h at 37 °C) using trypsin (Promega, Madison, WI, USA) diluted in 25 mM NH_4HCO_3. The resulting tryptic peptides were then extracted using 30 μL of a 60% acetonitrile solution containing 0.1% of formic acid. Acetonitrile was removed by vacuum drying using a speedvac.

A 1200 series nanoHPLC-Chip system (Agilent Technologies, Palo Alto, CA, USA) coupled to an HCT™ Plus ion trap (Bruker Daltonics, Bremen, Germany) was used to analyze tryptic peptides. The solvent system consisted of 2% acetonitrile, 0.1% HCOOH in water (solvent A) and 2% water, 0.1% formic acid in acetonitrile (solvent B). After loading of 3 μL of samples onto the enrichment column (ZORBAX 300SB-C18, 40 nL, 4 mm, with a 5 μm particle size) at a flow rate of 3.75 μL/min with solvent B, elution was performed on a separation column (ZORBAX 300SB-C18, 43 mm × 75 μm, with a 5 μm particle size) at a flow rate of 300 nL/min, according to the following gradient steps: From 8% to 40% B in 7 min, then from 40% to 70% B in one min, then 70% B during 2 min.

The mass spectrometer was operated with automatic switching between MS and MS/MS modes. The following voltages were set up: −1800 V (inlet), +147.3 V (outlet) and a skimmer voltage of +40 V. For mass spectrometry data acquisition, the scan speed was set at 8100 m/z per sec in the MS mode and 26000 m/z per sec in the MS/MS mode. Mass range was set at 250–2000 m/z in the MS mode and 50–2800 m/z in the MS/MS mode. The 3 most intense ions (doubly charged) were selected for CID-based fragmentation, and exclusion was set at 1 min or 2 spectra. The system was fully controlled by ChemStation (Rev B.01.035R1) and EsquireControl (v5.3) software (Agilent technologies and Bruker Daltonics, respectively).

2.2.5. MS/MS data analysis

Two different algorithms were used to analyze MS/MS data. The Mascot™ v2.3.02 program (Matrix Science, London, UK) was installed on a local server and the OMSSA v2.1.7 program (Open Mass Spectrometry Search Algorithm) [46] was run using the MSDA software suite [47]. Data were searched against a target-decoy version of the *Mus musculus* (Taxonomy 10090) protein database downloaded from NCBInr containing common contaminants like keratins and trypsin (July 2015, 410544 target + decoy entries), with a mass tolerance of 0.25 Da in MS and MS/MS modes, and allowing a maximum of one trypsin missed cleavage. Optional modifications were set as follows: carbamidomethylation of cysteine residues, oxidation of methionine residues, and acetylation of protein N-termini. Stringent filtering criteria based on probability-based scoring of the identified peptides were applied using Scaffold software v.3.0.7 (Proteome software Inc., Portland, OR, USA), to obtain a FDR < 1%. Hence, single peptide-based identifications were validated for MS/MS ion scores higher than 45 (Mascot) and –logE values higher than 5.3 (Omssa). Multiple peptide-based identifications were validated for MS/MS ion scores higher than 30 (Mascot) and –logE values higher than −0.05 (Omssa). Common contaminants such as keratin and trypsin were not considered.

When several different proteins were identified from analysis of a same protein spot, the so-called major ones (supposedly more abundant and responsible of possible spot intensity variation) were determined through a "peptide counting strategy" considering the percentage of experimentally detected peptides per protein (Mascot + Omssa) relative to the theoretical detectable number. To compute the theoretical number of detectable tryptic peptides, the possible presence of a Proline after a tryptic sites was considered, one missed cleavage was allowed, and the adequate size of peptides for their detection by mass spectrometry was determined directly from our data, which identified peptides composed of 5–33 amino acids. Hence, we calculated similar theoretical numbers of detectable tryptic peptides between major (69 ± 2) and minor (73 ± 3) proteins. Here, the proteins that were considered as major ones in a given protein spot were those to which about three times more peptides had been assigned versus minor ones (22 ± 1 % vs. only 8 ± 1 % of the possible tryptic peptides, respectively).

2.3. Body mass and biochemical validation of proteomic-derived hypothesis

Body mass loss (expressed in g) was recorded over the 24 hour post-injection to evaluate the energetic cost of the immune response to the LPS challenge. Biochemical measurements in liver samples were done in duplicate using assay kits purchased from Cayman Chemical (Ann Arbor, MI, USA). First, to confirm that variations in protein abundances were consistent with corresponding variations in protein activity, total glutathione S-transferase (GST) activity was assessed. To further validate the hypothesis on oxidative stress, reduced (GSH) and oxidized (GSSG) glutathione contents were measured in mice liver and liver

protein carbonyl content was also measured as it is a commonly used marker of ROS-induced protein oxidation.

2.4. Statistical analysis

All analyses were conducted on SPSS 18.0. Proteomic data were first checked for differences among groups using ANOVA with age (young and old) and treatment (LPS and PBS) as fixed factors. Post-hoc Tukey tests were then used for multiple pairwise comparisons. A multivariate analysis (Pillai's Trace's test followed by separated ANOVAs) was first used to compare body mass loss and biochemical variables among the four groups. In a second step, we then decided to run two Principal Component Analyses (PCA) with varimax rotation separately with a PCA conducted on the body mass loss and biochemical data (PCA1) and a PCA with only the restricted number of differential (i.e. significant) protein spots (PCA2). PCA resulted in two (PCA1) and three (PCA2) orthogonal variables (components, PCs) allowing easier comparisons of overall differences in liver biochemical and proteomic profiles among groups. Determinants of PCA1 and PCA2 were greater than 0.00001 (1.00×10^{-3} and 1.89×10^{-3}, respectively), Kaiser-Meyer-Olkin (KMO) measures showed sample adequacy for the analysis (0.71 and 0.61, respectively), individual items KMO values were > 0.66 (PCA1) and > 0.53 (PCA2) and Bartlett's test of sphericity showed sufficiently large correlation among variables for PCA (χ^2 (10) = 97.898, $P < 0.001$, χ^2 (66) = 135.684, $P < 0.001$, respectively). Only components with eigenvalues which met the Kaiser's criterion of 1 were conserved to explain total variance of the data. Subsequently, PCA scores of each individual were analysed by using Generalized Linear Model with Age, treatment and Age × Treatment interaction to check for differences in response to LPS. When the interaction was found significant, post-hoc comparisons were conducted using the same model but separated by Age or Treatment. Normality was tested for all models on residuals, using Kolmogorov–Smirnov test and checking linearity of QQ plots. Significance threshold is $P < 0.05$.

3. Results

3.1. Immune challenge-induced liver proteome changes

Multiple ANOVA analyses applied to relative intensities of 384 detected 2D-DIGE protein spots revealed that 17 of them exhibited significant differences among old and young mice treated with PBS or LPS ($P < 0.05$; Table 1). These 17 protein spots (Fig. 1) contained 18 different proteins, which were unambiguously identified on the basis of mass spectrometry data analysis (see Supplementary Table 1 for details). Most of these 18 liver proteins are known to play key roles in the response to oxidative stress, in energy metabolism and in the response

Table 1
Multiple ANOVA analyses applied to 2D-DIGE protein spot relative intensities among old and young mice treated with PBS or LPS.

Spot N°	Dependent variables	Acc. N°	Biological process
110	78 kDa glucose-regulated protein (GRP-78)	gi|2506545	Protein folding & Response to ER stress
ANOVA: F = 9.02; P = 0.0021/Tukey (LPS vs. PBS)young: FC = 1.3; P = 0.0034/Tukey (LPS vs. PBS)old: FC = 1.1; P = 0.7352			
180	Dihydrolipoyl dehydrogenase	gi|118572640	Carbohydrate metabolism
	+ Prolyl aminopeptidase	gi|124028616	+ Proteolysis
	+ Fibrinogen beta chain	gi|67460959	+ hemostasis & Immune response
ANOVA: F = 3.67; P = 0.0437/Tukey (LPS vs. PBS)young: FC = 1.2; P = 0.4008/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9389			
214	Mitoch. aldehyde dehydrogenase	gi|1352250	Response to LPS
ANOVA: F = 8.82; P = 0.0023/Tukey (LPS vs. PBS)young: FC = 1.3; P = 0.0188/Tukey (LPS vs. PBS)old: FC = 0.9; P = 0.2948			
271	Sorbitol dehydrogenase precursor	gi|152031591	Carbohydrate metabolism
ANOVA: F = 4.62; P = 0.0227/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0153/Tukey (LPS vs. PBS)old: FC = 0.9; P = 0.9036			
341	3-hydroxyanthranilate 3,4-dioxygenase	gi|61211578	Tryptophan metabolism
ANOVA: F = 4.24; P = 0.0292/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0284/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9574			
360	Carbonic anhydrase 3	gi|30581036	Response to oxidative stress
ANOVA: F = 4.57; P = 0.0234/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.4281/Tukey (LPS vs. PBS)old: FC = 1.0; P = 1.000			
362	Carbonic anhydrase 3	gi|30581036	Response to oxidative stress
ANOVA: F = 7.05; P = 0.0055/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.431/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9946			
375	Peroxiredoxin-6	gi|6671549	Response to oxidative stress
ANOVA: F = 11.93; P = 0.0007/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0073/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.997			
378	Glutathione-S-transferase	gi|121747	Response to oxidative stress
ANOVA: F = 5.63; P = 0.0121/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0242/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9549			
379	Glutathione-S-transferase	gi|121747	Response to oxidative stress
ANOVA: F = 5.75; P = 0.0112/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0364/Tukey (LPS vs. PBS)old: FC = 0.9; P = 0.7431			
452	Fibrinogen alpha isoform 2 precursor	gi|33563252	Hemostasis & Immune response
	+ Propanoyl-CoA C-acyltransferase	gi|32130432	+ bile acid biosynthesis & lipid oxidation
	+ NADP-dependent malic enzyme	gi|162139827	+ Carbohydrate metabolism
ANOVA: F = 5.76; P = 0.0112/Tukey (LPS vs. PBS)young: FC = 1.3; P = 0.0144/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9979			
459	Haloacid dehalogenase-like hydrolase domain-containing protein 3	gi|81904469	unknown
ANOVA: F = 10.31; P = 0.0043/Tukey (LPS vs. PBS)young: FC = 1.1; P = 0.3589/Tukey (LPS vs. PBS)old: FC = 1.1; P = 0.6768			
469	Glutathione-S-transferase	gi|121747	Response to oxidative stress
ANOVA: F = 4.97; P = 0.0181/Tukey (LPS vs. PBS)young: FC = 0.9; P = 0.8254/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9984			
491	Glutathione synthetase	gi|1708057	Response to oxidative stress
ANOVA: F = 3.78; P = 0.0403/Tukey (LPS vs. PBS)young: FC = 1.3; P = 0.0336/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9739			
492	S-adenosylmethionine synthase isoform type-1	gi|81902386	Resp. to oxidative stress & methylation & S-adenosylmethionine metabolism
ANOVA: F = 3.79; P = 0.0402/Tukey (LPS vs. PBS)young: FC = 1.3; P = 0.0397/Tukey (LPS vs. PBS)old: FC = 1.1; P = 0.7084			
507	Glycine N-methyltransferase	gi|55976615	Response to oxidative stress & tumor suppression & S-adenosylmethionine metabolism & immune response
ANOVA: F = 4.45; P = 0.0254/Tukey (LPS vs. PBS)young: FC = 0.8; P = 0.0769/Tukey (LPS vs. PBS)old: FC = 1.1; P = 0.9254			
510	Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic	gi|121557	Carbohydrate metabolism
ANOVA: F = 5.40; P = 0.0139/Tukey (LPS vs. PBS)young: FC = 0.7; P = 0.0197/Tukey (LPS vs. PBS)old: FC = 1.0; P = 0.9823			

Multiple ANOVA and post-hoc Tukey tests were conducted on proteomics-derived liver protein levels in old and young mice treated with PBS or LPS (n = 4 in each group). 17 differential protein spots were hence determined ($P < 0.05$), which were considered for subsequent PCA analyses. The biological processes in which each of the major proteins from the 17 differential protein spots is involved in were computed using annotation explorer module of the MSDA software suite [47] and complemented with literature examination. Spot N° refers to those reported in Fig. 1.

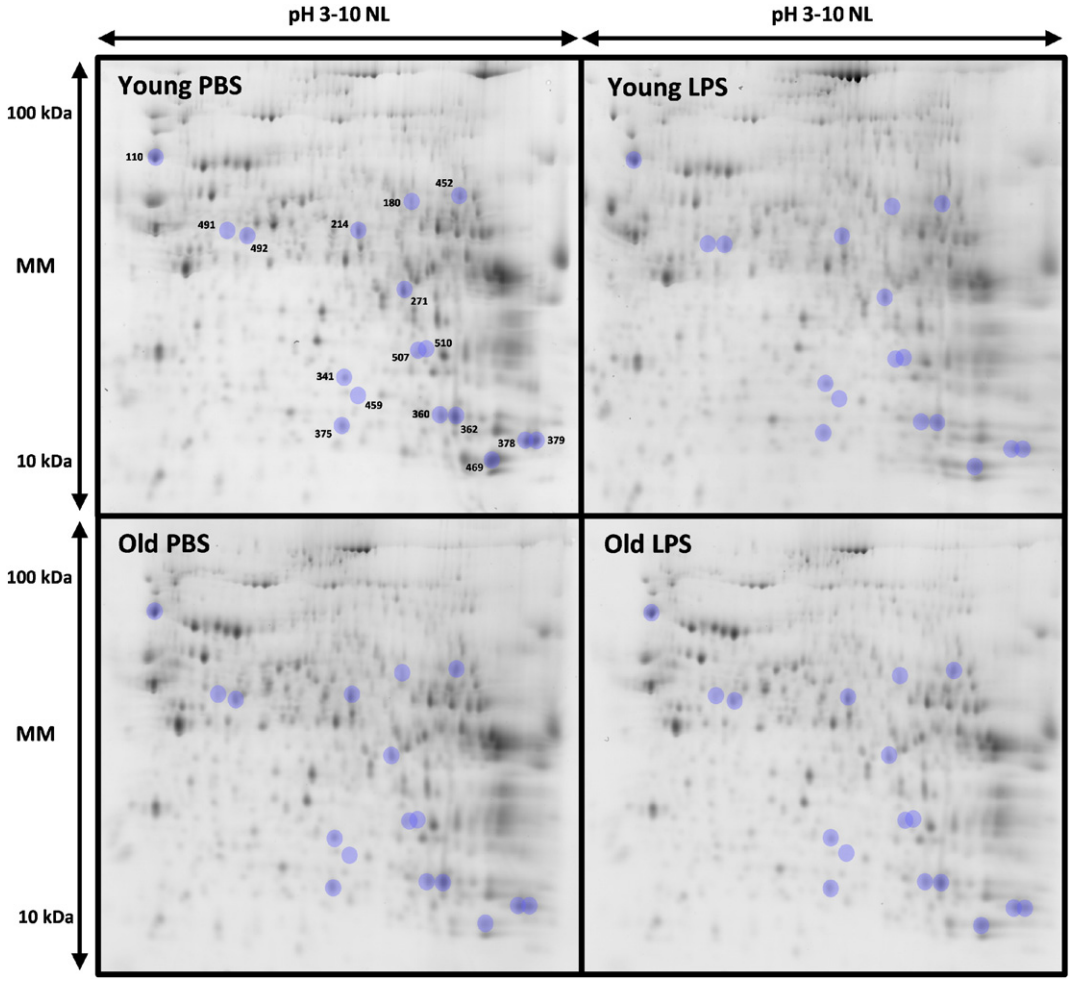


Fig. 1. Representative 2D-gel image of liver proteins in old and young mice treated with PBS or LPS. Significantly different protein spots according to multiple ANOVA analysis ($P < 0.05$) are shown.

to immune challenges, which suggest involvement of these biological processes when immune system is activated.

To further understand phenotypic responses in immune-challenged (LPS) old and young mice, PCA was run using only 15 of the 17 differential protein spots. Indeed, protein spots N°459 and N°510 were not considered here because they rendered the definite matrix non-positive. Three components (PC1, PC2, and PC3) with eigenvalues higher than 1, and explaining 79% of the total variance after rotation, were obtained (Table 3). Only protein loadings over 0.6 were considered due to relatively small sample size [48], which allowed changes in liver protein abundances to be attributed to either of the 3 principal components. Particular biological processes were then linked to each of the principal components, on the basis of GO ontologies (extracted using the MSDA software suite [47]) and literature examination. Hence, our results suggest that PC1 is especially related to response to oxidative stress, PC2 to both response to oxidative stress and energy metabolism, and PC3 to both response to oxidative stress and immune challenge.

Table 2
Multivariate analysis applied to body mass loss and biochemical measurements among old and young mice treated with PBS or LPS.

Dependent variables	Old PBS	Old LPS	Young PBS	Young LPS	$F_{1,3}$	P
	[Estimates]	[Estimates]	[Estimates]	[Estimates]		
Body mass loss (g)	−0.08 ± 0.14 [0.007 ± 0.339]	−2.61 ± 0.40 [−2.535 ± 0.339]	−0.08 ± 0.13 [0]	−1.81 ± 0.19 [−1.723 ± 0.339]	28.23	<0.001
GST activity ($nmol.min^{-1}.mL^{-1}$)	147.53 ± 18.60 [−77.5 ± 26.1]	160.93 ± 9.78 [−64.1 ± 26.1]	224.98 ± 22.86 [0]	165.45 ± 19.82 [−59.5 ± 26.1]	3.48	0.050
Total GSH ($\mu mol.L^{-1}$)	683.54 ± 104.28 [−354.8 ± 83.8]	109.82 ± 21.01 [218.9 ± 83.8]	464.66 ± 32.66 [0]	313.23 ± 40.8 [−151.4 ± 83.8]	16.72	<0.001
GSSG ($\mu mol.L^{-1}$)	441.28 ± 62.04 [141.5 ± 50.9]	85.56 ± 13.16 [−214.2 ± 50.9]	299.76 ± 21.02 [0]	220.93 ± 26.95 [−78.8 ± 50.9]	17.05	<0.001
protein carbonyl ($pmol.L^{-1}$)	0.46 ± 0.18 [−0.72 ± 0.39]	1.53 ± 0.15 [0.35 ± 0.39]	1.18 ± 0.43 [0]	2.04 ± 0.25 [0.86 ± 0.39]	5.93	0.010

Multivariate analysis (Pillai's Trace's test followed by separated ANOVAs) was conducted on body mass loss and biochemical values in old and young mice treated with PBS or LPS ($n = 4$ in each group) and showed a significant effect of the treatment and age class on the different variables $F_{15,30} = 4.4$, $P < 0.001$). Levene's tests showed that the assumption of the homogeneity of variances was met for all variables. Data are given as means ± SEM. Protein carbonyl, GST (glutathione-S-transferase) activity, total GSH and GSSG (reduced and oxidized glutathione) contents are expressed per μg of liver proteins.

Table 3
Phenotypic responses in immune-challenged (LPS) old and young mice.

			PC1	PC2	PC3
Liver oxidative proxies and body mass loss					
	Body mass loss		**0.855**	0.394	
	Glutathione-S-transferase activity		−0.114	**0.967**	
	Total GSH		**0.963**	0.023	
	GSSG		**0.961**	0.014	
	Protein carbonyl		**−0.760**	0.314	
	% variance explained		*0.63*	*0.24*	
Liver protein levels					
Resp. Ox. stress	3-hydroxyanthranilate 3,4-dioxygenase	spot 341	**0.862**	−0.153	−0.321
	Glutathione-S-transferase	spots 378 & 379 & 469	**0.832**	−0.349	−0.005
	Peroxiredoxin-6	spot 375	**0.794**	−0.430	0.190
	78 kDa glucose-regulated protein (GRP-78)	spot 110	**−0.775**	0.397	0.067
	Carbonic anhydrase 3	spots 360 & 362	**0.701**	−0.297	0.454
Resp. Ox. Stress & energy metab.	Glutathione synthetase	spot 491	−0.218	**0.921**	0.154
	S-adenosylmethionine synthase isoform type-1	spot 492	−0.296	**0.889**	0.101
	Sorbitol dehydrogenase precursor	spot 271	0.440	**−0.717**	−0.187
	Fibrinogen alpha polypeptide isoform 2 precursor + Propanoyl-CoA C-acyltransferase + NADP-dependent malic enzyme	spot 452	−0.567	**0.694**	0.101
Resp. Ox. Stress & immune chall.	Mitochondrial aldehyde dehydrogenase	spot 214	−0.101	0.023	**0.878**
	Glycine N-methyltransferase	spot 507	0.201	−0.212	**−0.752**
	Dihydrolipoyl dehydrogenase + Prolyl aminopeptidase + Fibrinogen beta chain	spot 180	0.324	0.162	**0.744**
	% variance explained		*0.33*	*0.27*	*0.19*

Principal Components Analyses were conducted on body mass loss values and liver oxidative proxies, and on proteomics-derived liver protein levels in old and young mice treated with PBS or LPS (n = 4 in each group). For protein levels, only differential 2D-DIGE protein spots (initially recognized on the basis of multiple ANOVA analysis; see Table 1) were considered. Nature of the variables forming the main axes (PC1, PC2 and PC3) allow phenotypic responses to an immune challenge (LPS) to be described compared between old and young mice. Spot N° refers to those reported in Fig. 1. Protein loadings over 0.6 (in bold) were considered and allowed principal components to be linked to particular biological processes. GSH and GSSG: reduced and oxidized glutathione

To better determine how the 3 components were altered by LPS immune challenge in old and young mice, generalized linear models were run (Table 4). A significant effect of Treatment (LPS or PBS) on PC1 and PC2 was found, while Age had a significant effect only on PC3. Interaction Age × Treatment was significant for PC1 and PC3. Thus, young and old mice responded differently to LPS injections with regard to PC1 (PC1 values significantly decreased in young LPS mice while did not change in old mice; Fig. 2). More precisely, levels of 3-hydroxyanthranilate 3,4-dioxygenase, GST, Peroxiredoxin-6 and Carbonic anhydrase 3 were significantly reduced while those of 78 kDa glucose-regulated protein (GRP-78) were increased only in LPS-injected young mice vs. PBS-injected young mice (Table 3 and Fig. 2). PC2 values were higher in both old and young LPS-injected animals (Fig. 2), thus suggesting that LPS injections had a comparable effect whatever the age of mice, levels of Glutathione synthetase, S-adenosylmethionine synthase isoform type-1, and Fibrinogen alpha polypeptide isoform 2 precursor + Propanoyl-CoA C-acyltransferase + NADP-dependent malic enzyme were increased while those of Sorbitol dehydrogenase precursor were decreased in LPS-injected vs. PBS-injected mice (Table 3 and Fig. 2). Finally, PC3 values were globally higher in older mice (Fig. 2), suggesting that levels of mitochondrial aldehyde dehydrogenase and dihydrolipoyl dehydrogenase + Prolyl aminopeptidase + Fibrinogen beta chain were higher while those of Glycine N-methyltransferase were lower in old vs. young mice (Table 3). In addition, the significant interaction Age × Treatment indicated that old individuals, despite preserving higher PC3 values, exhibited a decrease after injection while young individuals significantly increased their PC3 values (Table 4 and Fig. 2).

Thus, proteomics data analysis led to the hypothesis that oxidative balance is adjusted upon LPS treatment, but especially in young animals where response to endoplasmic reticulum (ER) stress would be higher. Proteomics data also indicate that a pro-oxidative status pre-exists in old mice before LPS injection. Altogether, these data strongly support that LPS treatment induces liver oxidative stress, but also that ageing is associated with different pre- and post-LPS injection oxidative status. To test this hypothesis, liver oxidative stress-related proxies were assessed.

4. Immune challenge-induced liver oxidative stress

For a given age, initial body masses were not different between LPS- and PBS-mice (GLM, $F_{1,\ 15} = 1.08$, $P = 0.319$). Old individuals were shown to have a higher body mass than young individuals, irrespectively of the immune treatment (GLM, $F_{1,\ 15} = 5.00$, $P < 0.001$; estimates 10.19 ± 2.04). A multivariate analysis revealed significant differences (Roy's greatest root, $P < 0.001$) regarding body mass loss and livre oxidative proxies (Table 2). PCA2 analysis using body mass loss and liver oxidative stress-related measurements produced two components (PC1 and PC2) with eigenvalues higher than 1, and explaining 87% of the total variance (Table 3). Considering only factor loadings over 0.6 (see above), we found that PC1 was related to body mass loss and levels of total GSH, GSSG and protein carbonyl. PC2 was here related to GST activity.

Generalized linear models revealed a significant effect of Treatment (LPS or PBS) on PC1, while Age had a significant effect only on PC2 (Table 5). Interaction Age × Treatment was significant only for PC1. Thus, young and old mice responded differently to LPS injections with regard to PC1, PC1 values being lower in LPS *vs.* PBS mice, but with a more marked drop in old LPS-treated animals (Fig. 3). Thus, body mass loss was more pronounced and levels of total GSH and GSSG were significantly more decreased in LPS-injected vs. PBS-injected old mice than in LPS-injected vs. PBS-injected young mice (Table 3 and Fig. 3). Oxidative damage on liver proteins (protein carbonyl contents) was more markedly increased in older mice after LPS injection. PC2 values were higher only in young PBS-injected mice (Fig. 3), suggesting that old mice have lower GST activity than younger individuals (Table 3 and Fig. 3).

Table 4
GLM applied to PC1, PC2 and PC3 determined for liver protein levels.

	Mean values	Estimates	D.F.	F	P
PC1. Response to oxidative stress					
Age (old *vs.* young)		0.026 ± 0.533	1.15	2.090	0.174
Treatment (LPS *vs.* PBS)		−1.467 ± 0.533	1.15	6.346	**0.027**
Age × treatment			1.15	5.997	**0.031**
Old PBS	0.286 ± 0.352		1.7	0.458	0.524
Old LPS	0.663 ± 1.059				
Young PBS	0.259 ± 0.913		1.7	8.407	**0.027**
Young LPS	−1.208 ± 0.436				
PC2. Response to oxidative stress and energy metabolism					
Age (old *vs.* young)		0.118 ± 0.646	1.15	0.311	0.588
Treatment (LPS *vs.* PBS)		1.395 ± 0.646	1.15	5.014	**0.045**
Age × treatment			1.15	0.667	0.430
PC3. Response to oxidative stress and immune challenge					
Age (old *vs.* young)		2.341 ± 0.369	1.15	25.831	<**0.001**
Treatment (LPS *vs.* PBS)		1.399 ± 0.369	1.15	2.164	0.167
Age × treatment			1.15	15.167	**0.002**
Old PBS	0.979 ± 0.547		1.7	1.853	0.222
Old LPS	0.347 ± 0.750				
Young PBS	−1.362 ± 0.357		1.7	34.376	**0.001**
Young LPS	0.037 ± 0.314				

Generalized linear models were conducted to explain the variance in PC1, PC2 and PC3 determined from proteomics-derived liver protein levels (see Fig. 3) in old and young mice treated with PBS or LPS (n = 4 in each group). Because both treatment effects and interaction Age × treatment were significant for PC1, and both age effects and interaction age × treatment were significant for PC3, we further conducted post-hoc tests to better characterize differences between groups, by comparing Old PBS *vs.* Old LPS and Young PBS *vs.* Young LPS. Estimates and mean values are given ± SE and significant values are indicated in bold. Residuals for each models followed a normal distribution (checked using Kolmogorov–Smirnov test and QQ plot).

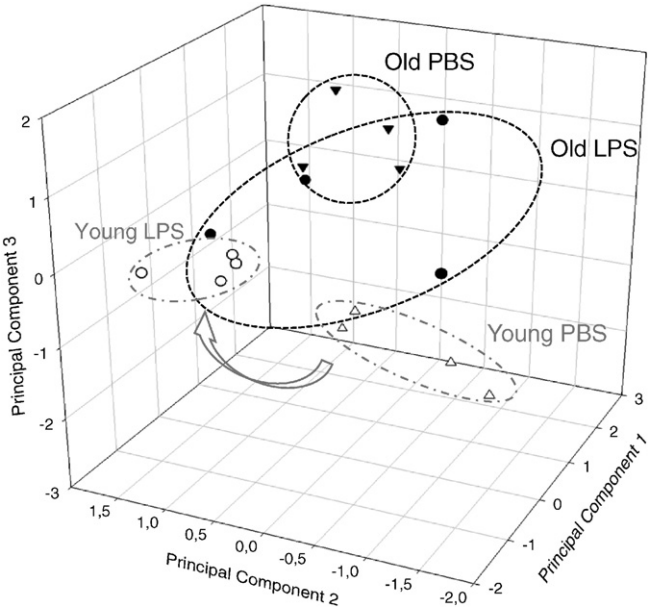


Fig. 2. PCA on proteomics-derived liver protein levels. PCA was conducted on relative protein abundance values for differential 2D-DIGE protein spots (initially recognized on the basis of multiple ANOVA analysis) in old and young mice treated with PBS or LPS (n = 4 in each group). Principal components (see Table 3) accounted for a total of 33% (PC1), 60% (PC1 + PC2) and 79.7% (PC1 + PC2 + PC3) of the total variance. Ellipses and arrows indicate how old (black) and young (grey) mice reacted in response to LPS vs. PBS treatment.

Table 5
GLM applied to PC1 and PC2 determined for liver oxidative proxies and body mass loss.

	Mean values	Estimates	D.F.	F	P
PC1					
Age (old *vs.* young)		0.709 ± 0.184	1.15	0.014	0.757
Treatment (LPS *vs.* PBS)		−1.057 ± 0.261	1.15	11.661	<**0.001**
Age × treatment			1.15	12.468	**0.004**
Old PBS	1.208 ± 0.538		1.7	4.126	0.089
Young PBS	0.499 ± 0.329				
Old LPS	−1.150 ± 0.131		1.7	4.998	**0.059**
Young LPS	−0.557 ± 0.358				
PC2					
Age (old *vs.* young)		−1.697 ± 0.563	1.15	7.292	**0.019**
Treatment (LPS *vs.* PBS)		−1.173 ± 0.563	1.15	1.921	0.191
Age × treatment			1.15	2.433	0.145

Generalized linear models were conducted to explain the variance in PC1 and PC2 determined from liver oxidative measurements and body mass loss (see Fig. 2) in old and young mice treated with PBS or LPS (n = 4 in each group). Because both treatment effects and interaction age × treatment were significant for PC1, we further conducted post-hoc tests to better characterize differences between groups, by comparing Old PBS *vs.* Young PBS and Old LPS *vs.* Young LPS. Estimates and mean values are given ± SE and significant values are indicated in bold. Residuals for each models followed a normal distribution (checked using Kolmogorov–Smirnov test and QQ plot).

Thus, our biochemical tests confirmed proteomics-driven hypothesis that LPS treatment induces liver oxidative stress, and that this effect is more marked in older mice due to a lack of antioxidant capacities.

5. Discussion

The present study strongly suggests that activation of the immune system in mice triggers several cellular and metabolic pathways that are all related to a response to oxidative stress. In addition, one other important conclusion is that 1 year-old individuals do not exhibit marked changes in their liver protein profiles 24-h after LPS injection, suggesting that they have lost part of their immune responsiveness. This is corroborated by the fact that old individuals had higher body mass loss after injection than young ones, which would indicate a slower clearance of LPS by their immune system. Consequently, LPS-

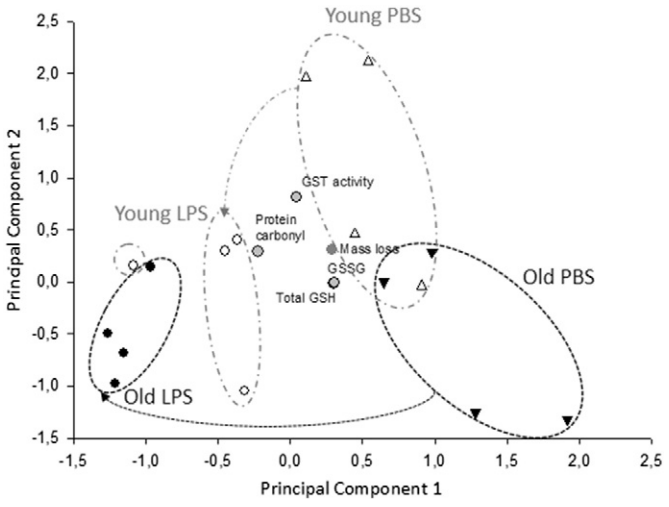


Fig. 3. PCA on liver oxidative proxies and body mass loss. PCA was conducted on liver oxidative measurements and body mass loss recorded in old and young mice treated with PBS or LPS (n = 4 in each group). Indicated by grey points, age and oxidative balance proxies (GST activity, protein carbonyl, total GSH and GSSG contents), as well as body mass loss are projected onto their first two principal components (see Table 3) accounting for a total of 63.6% (PC1) and 87.4% (PC1 + PC2) of the total variance. Ellipses and arrows indicate how old (black) and young (grey) mice reacted in response to LPS vs. PBS treatment.

challenged old-individuals pay a higher immune cost in terms of oxidative stress, as shown by higher protein carbonyl levels and lower GST activity in liver homogenates. We also found higher levels of proteins related to oxidative stress control in old individuals independently of immune treatment, suggesting constitutive costs associated to dysregulation of immune-related functions even in the absence of immune stimulation. Our study therefore brings to the fore proteomic proofs that preserving immune efficiency at old age in mice may trigger oxidative imbalance.

Our proteomic data revealed significant changes in the intensity of protein spots with rather low fold changes. This not only underlines the low animal-to-animal variations within groups, but also the fact that small variation in protein levels upon LPS treatment after only 24 h already reflects major effects on mouse liver (see below). More precisely, abundance of several proteins was changed independently of age following LPS injection, with proteins related to oxidative stress and energy metabolism (PCA1, PC2), rather suggesting that carbohydrate (and possibly fatty acid) metabolism is enhanced (Tables 1–3). It is interesting to note that fibrinogen alpha-2 precursor could also be positively affected by LPS injection, underlying that the preservation of hemostasis could be an important feature of the immune response. However, an experiment conducted on Blue Tit (*Parus caeruleus*) found rather weak energy cost of immune (antibody) response (<13% of the basal metabolic rate) and suggested that non-energetically driven trade-offs may be more constraining for the immune system [49]. Among different possibilities (see [50] for an alternative explanation), the production of ROS either because of the increase in metabolic rate or directly by immune cells has been previously proposed to be particularly deleterious [51]. For example, immune system regulation largely depends on the ability of a large number of cell types to produce nitric oxide (such as fibroblasts, macrophages, natural killer cells) either to regulate cell activation/proliferation (*e.g.* T lymphocytes) or to destroy infectious organisms [52]. The absence of tightly coordinated or well-balanced control of immune-induced ROS production may lead to damage accumulation, thereby accounting for immune trade-offs [53, 54]. Interestingly, this deleterious phenomenon may particularly take place when individuals are in poor conditions, either because of environmental energy constraints or of a decrease in organismal functionality with age. Indeed, autoimmune (oxidative-derived) damage remains one of the main consequences of immune-senescence in old-individuals [51]. In both young and old mice, proteomics data suggest that LPS treatment would trigger synthesis of glutathione, a major antioxidant [55], and of S-Adenosylmethionine, which can potentiate the activity of antioxidant enzymes [56, 57]. Therefore, improvement of antioxidant capacities seems to be an important regulating feature when mounting an innate immune response. However, young and old mice did not respond in the same way. Old mice showed very little changes in their protein profiles related to response to oxidative stress. For instance, proteins like mitochondrial aldehyde dehydrogenase, peroredoxin-6 or GST were all mobilized in young individuals (which would reflect the need to fight against oxidative stress upon LPS treatment, as reflect by increased protein carbonyl values), whereas they were not changed significantly in old LPS-treated animals. These proteins are either antioxidant enzymes or involved in mitochondrial protection against oxidative stress [58], suggesting that old mice have lost part of the antioxidant barrier that must come with immune response. This would contribute to explain why the increase in protein carbonyl values is higher upon LPS treatment in old vs. young mice. Oxidative cost of badly tuned immune response could damage irreversibly some key ageing markers, with particular deleterious impact for immune cells themselves. Shortened telomere ends of linear chromosomes may reduce T- and B-lymphocyte proliferative capacity, thereby contributing to defective immune response in old-individuals [59]. We reach here one limitation of our study since we did not detect any specific immune markers to vary among groups, or target immune tissue for proteomic analysis. Therefore, whether populations of macrophages, natural killers or lymphocytes have decreased cell proliferation in old mice because of enhanced senescence rate remains an open question, for which proteomic has clearly a role to play. To reach more depth in analysing the liver proteome, future studies should consider MS-based proteomics strategies such as the global label-free techniques which allow quantification of thousands of proteins simultaneously [60]. An analytical strategy based on LC-SRM analyses could also help to directly target immune markers [61]. Finally, one could have focused on tissues that play key roles in the immune system, i.e. the spleen, or analysed liver samples enriched in immune cells (like macrophages). Actually, the liver is a source of leucocyte lineages such as Kupffer cells and plays a predominant role in response to bacterial infection, producing interleukins and recruiting natural killer and T cells in response to LPS [62]. Therefore, working on Kupffer cell isolates or cultures may bring to the fore innovative proteomic data about the ageing cost of inflammation. In addition, hepatocytes themselves respond to Kupffer cells signalling by producing oxidants such as nitric oxide in the presence of LPS [63]. Evaluating how this domino cascade of activations is altered with age may be of prime importance in our understanding of the liver physiological processes leading to sepsis. Still, there are two interesting points to note coming out from our data. First, that old animals, in addition to their weak global protein response to LPS, also exhibited a basal pro-oxidant status compared to young individuals. These proteomic data coupled with higher basal antioxidant values suggest that old individuals were facing a chronic oxidative challenge even in the absence of pathogen. Such age-related cost could partly be attributable to a putative high constitutive immune cost in aging animals, *i.e.* related to the maintenance of the immune system in the absence of immune stimulation [64]. Secondly, we found a chaperon protein, GRP-78 likely to be more expressed in young animals after LPS injection. This protein is associated at the cell surface with major histocompatibility class 1 molecule, suggesting a role in cell-mediated immune response [65]. In addition, GRP-78 may also fulfil a protective role against endoplasmic reticulum stress-induced cell death [66]. This indicates an interesting mechanism linking immunity and body maintenance that should be more precisely studied in the context of trade-offs between immunity and other life history traits.

6. Conclusions

In the present study, we used a proteomic approach to characterize in mice liver the protein pathways that may be differently affected by the trade-off between body maintenance and an innate immune response in relation with age. We highlighted that most of the differential protein spots contain proteins that are related to the control of an individual's oxidative status. In old animals prior to the injection, those pathways appeared already up-regulated, and then were only weakly modified by the LPS challenge. This is likely to be associated to a higher oxidative stress paid as a cost of constitutive immune maintenance and/or functioning, as shown by our biochemical measurements. Interestingly, we also highlighted one new way through which immunity and lifespan may be traded-off, *via* the regulation of the GRP-78 chaperon protein, illustrating the added value of applying proteomics to evolutionary biology questions. A coming step will be to both explore immune trade-offs in other organisms with contrasting lifespan (i.e. birds) using the same methodology, to better determine the nature of the mechanisms on which are based immune costs [64, 67–70], and how they have been modified by evolutionary history.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2015.09.008.

Competing interest statement

The authors declare that they have no competing interests.

Acknowledgements

This work was supported by the CNRS and Strasbourg University (H2E project; IDEX UNISTRA), the French Proteomic Infrastructure (ProFI; ANR-10-INSB-08-03), and a CNRS "Projets Exploratoires Premier Soutien" (PEPS). We wish to thank Aurélie Hranitsky for her contribution with animal husbandry.

References

[1] T. Finkel, N.J. Holbrook, Oxidants, oxidative stress and the biology of ageing, Nature 408 (6809) (2000) 239–247.
[2] M.A. Blasco, Immunosenescence phenotypes in the telomerase knockout mouse, Springer Semin. Immunopathol. 24 (1) (2002) 75–85.
[3] A. Lasry, Y. Ben-Neriah, Senescence-associated inflammatory responses: aging and cancer perspectives, Trends Immunol. 36 (4) (2015) 217–228.
[4] R.A. Miller, The aging immune system: primer and prospectus, Science 273 (5271) (1996) 70–74.
[5] C. Caruso, S. Buffa, G. Candore, G. Colonna-Romano, D. Dunn-Walters, D. Kipling, G. Pawelec, Mechanisms of immunosenescence, Immun. Ageing 6 (2009) 10.
[6] D.N. Shelton, E. Chang, P.S. Whittier, D. Choi, W.D. Funk, Microarray analysis of replicative senescence, Curr. Biol. 9 (17) (1999) 939–945.
[7] N.P. Weng, Aging of the immune system: how much can the adaptive immune system adapt? Immunity 24 (5) (2006) 495–499.
[8] D.P. Shanley, D. Aw, N.R. Manley, D.B. Palmer, An evolutionary perspective on the mechanisms of immunosenescence, Trends Immunol. 30 (7) (2009) 374–381.
[9] G.C. Williams, Pleiotropy, natural selection, and the evolution of senescence, Evolution 11 (1957) 398–411.
[10] M.V. Blagosklonny, Revisiting the antagonistic pleiotropy theory of aging TOR-driven program and quasi-program, Cell Cycle 9 (16) (2010) 3151–3156.
[11] F. Licastro, G. Candore, D. Lio, E. Porcellini, G. Colonna-Romano, C. Franceschi, C. Caruso, Innate immunity and inflammation in ageing: a key for understanding age-related diseases, Immun. Ageing 2 (2005) 8.
[12] G. Sorci, B. Faivre, Inflammation and oxidative stress in vertebrate host-parasite systems, Philos. Trans. R. Soc. Lond. B Biol. Sci. 364 (1513) (2009) 71–83.
[13] T.B. Kirkwood, Evolution of ageing, Nature 270 (5635) (1977) 301–304.
[14] D. Harman, Aging: a theory based on free radical and radiation chemistry, J. Gerontol. 11 (3) (1956) 298–300.
[15] C. Alonso-Alvarez, S. Bertrand, G. devevey, J. Prost, B. Faivre, G. Sorci, Increased susceptibility to oxidative stress as a proximate cost of reproduction, Ecol. Lett. 7 (2004) 363–368.
[16] C. Alonso-Alvarez, S. Bertrand, B. Faivre, G. Sorci, Increased susceptibility to oxidative damage as a cost of accelerated somatic growth in zebra finches, Funct. Ecol. 21 (5) (2007) 873–879.
[17] D. Costantini, G. Dell'Omo, Effects of T-cell-mediated immune response on avian oxidative stress, Comp. Biochem. Physiol. A Mol. Integr. Physiol. 145 (1) (2006) 137–142.
[18] M.L. Hamilton, H. Van Remmen, J.A. Drake, H. Yang, Z.M. Guo, K. Kewitt, C.A. Walter, A. Richardson, Does oxidative damage to DNA increase with age? Proc. Natl. Acad. Sci. U. S. A. 98 (18) (2001) 10469–10474.
[19] H. Klandorf, D.S. Rathore, M. Iqbal, X. Shi, K. Van Dyke, Accelerated tissue aging and increased oxidative stress in broiler chickens fed allopurinol, Comp. Biochem. Physiol. C Toxicol. Pharmacol. 129 (2) (2001) 93–104.
[20] E. Le Bourg, Oxidative stress, aging and longevity in Drosophila melanogaster, FEBS Lett. 498 (2-3) (2001) 183–186.
[21] P. Wiersma, C. Selman, J.R. Speakman, S. Verhulst, Birds sacrifice oxidative protection for reproduction, Proc. Biol. Sci. 271 (Suppl. 5) (2004) S360–S363.
[22] P. Monaghan, N.B. Metcalfe, R. Torres, Oxidative stress as a mediator of life history trade-offs: mechanisms, measurements and interpretation, Ecol. Lett. 12 (1) (2009) 75–92.
[23] L.B. Martin, A. Scheuerlein, M. Wikelski, Immune activity elevates energy expenditure of house sparrows: a link between direct and indirect costs? Proc. R. Soc. B Biol. Sci. 270 (1511) (2003) 153–158.
[24] G.E. Demas, V. Chefer, M.I. Talan, R.J. Nelson, Metabolic costs of mounting an antigen-stimulated immune response in adult and aged C57BL/6J mice, Am. J. Physiol. 273 (5 Pt 2) (1997) R1631–R1637.
[25] M. De la Fuente, Effects of antioxidants on immune system ageing, Eur. J. Clin. Nutr. 56 (Suppl. 3) (2002) S5–S8.
[26] Y. Emre, C. Hurtaud, T. Nubel, F. Criscuolo, D. Ricquier, A.M. Cassard-Doulcier, Mitochondria contribute to LPS-induced MAPK activation via uncoupling protein UCP2 in macrophages, Biochem. J. 402 (2) (2007) 271–278.
[27] S. Gordon, Alternative activation of macrophages, Nat. Rev. Immunol. 3 (1) (2003) 23–35.
[28] J.D. Lambeth, NOX enzymes and the biology of reactive oxygen, Nat. Rev. Immunol. 4 (3) (2004) 181–189.
[29] J.D. Lambeth, K.H. Krause, R.A. Clark, NOX enzymes as novel targets for drug development, Semin. Immunopathol. 30 (3) (2008) 339–363.
[30] S.A. Hanssen, D. Hasselquist, I. Folstad, K.E. Erikstad, Costs of immunity: immune responsiveness reduces survival in a vertebrate, Proc. R. Soc. B Biol. Sci. 271 (1542) (2004) 925–930.
[31] Y. Moret, P. Schmid-Hempel, Survival for immunity: The price of immune system activation for bumblebee workers, Science 290 (5494) (2000) 1166–1168.
[32] C.A. Janeway Jr., R. Medzhitov, Innate immune recognition, Annu. Rev. Immunol. 20 (2002) 197–216.
[33] S. Sriskandan, D.M. Altmann, The immunology of sepsis, J. Pathol. 214 (2) (2008) 211–223.
[34] Y. Bai, H. Onuma, X. Bai, A.V. Medvedev, M. Misukonis, J.B. Weinberg, W. Cao, J. Robidoux, L.M. Floering, K.W. Daniel, S. Collins, Persistent nuclear factor-kappa B activation in Ucp2-/- mice leads to enhanced nitric oxide and inflammatory cytokine production, J. Biol. Chem. 280 (19) (2005) 19062–19069.
[35] E. Oveland, T.V. Karlsen, H. Haslene-Hox, E. Semaeva, B. Janaczyk, O. Tenstad, H. Wiig, Proteomic evaluation of inflammatory proteins in rat spleen interstitial fluid and lymph during LPS-induced systemic inflammation reveals increased levels of ADAMST1, J. Proteome Res. 11 (11) (2012) 5338–5349.
[36] E. Liaskou, D.V. Wilson, Y.H. Oo, Innate immune cells in liver inflammation, Mediat. Inflamm. 2012 (2012) 949157.
[37] V. Racanelli, B. Rehermann, The liver as an immunological organ, Hepatology 43 (2 Suppl. 1) (2006) S54–S62.
[38] K. Tanikawa, T. Torimura, Studies on oxidative stress in liver diseases: important future trends in liver research, Med. Mol. Morphol. 39 (1) (2006) 22–27.
[39] A.P. Diz, M. Martinez-Fernandez, E. Rolan-Alvarez, Proteomics in evolutionary ecology: linking the genotype with the phenotype, Mol. Ecol. 21 (5) (2012) 1060–1080.
[40] V. Belloni, B. Faivre, R. Guerreiro, E. Arnoux, J. Bellenger, G. Sorci, Suppressing an anti-inflammatory cytokine reveals a strong age-dependent survival cost in mice, PLoS One 5 (9) (2010), e12940.
[41] M.M.A. Bradford, Rapid and sensitive method for quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, Anal. Biochem. 7 (1976) 248–254.
[42] M. Unlu, M.E. Morgan, J.S. Minden, Difference gel electrophoresis: a single gel method for detecting changes in protein extracts, Electrophoresis 18 (11) (1997) 2071–2077.
[43] T. Rabilloud, Use of thiourea to increase the solubility of membrane proteins in two-dimensional electrophoresis, Electrophoresis 19 (5) (1998) 758–760.
[44] A. Gorg, W. Weiss, M.J. Dunn, Current two-dimensional electrophoresis technology for proteomics, Proteomics 4 (12) (2004) 3665–3685.
[45] V. Neuhoff, R. Stamm, I. Pardowitz, N. Arold, W. Ehrhardt, D. Taube, Essential problems in quantification of proteins following colloidal staining with coomassie brilliant blue dyes in polyacrylamide gels, and their solution, Electrophoresis 11 (2) (1990) 101–117.
[46] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, Open mass spectrometry search algorithm, J. Proteome Res. 3 (5) (2004) 958–964.
[47] C. Carapito, A. Burel, P. Guterl, A. Walter, F. Varrier, F. Bertile, A. Van Dorsselaer, MSDA, a proteomics software suite for in-depth Mass Spectrometry Data Analysis using grid computing, Proteomics 14 (9) (2014) 1014–1019.
[48] A. Field, Discovering statistics using SPSS, 2nd ed. Sage Publications, Ltd., 2009
[49] E. Svensson, L. Raberg, C. Koch, D. Hasselquist, Energetic stress, immunosuppression and the costs of an antibody response, Funct. Ecol. 12 (6) (1998) 912–919.
[50] S.A. Adamo, J.L. Roberts, R.H. Easy, N.W. Ross, Competition between immune function and lipid transport for the protein apolipophorin III leads to stress-induced immunosuppression in crickets, J. Exp. Biol. 211 (Pt 4) (2008) 531–538.
[51] B.M. Sadd, M.T. Siva-Jothy, Self-harm caused by an insect's innate immunity, Proc. Biol. Sci. 273 (1600) (2006) 2571–2574.
[52] J.W. Coleman, Nitric oxide in immunity and inflammation, Int. Immunopharmacol. 1 (8) (2001) 1397–1406.
[53] T. von Schantz, S. Bensch, M. Grahn, D. Hasselquist, H. Wittzell, Good genes, oxidative stress and condition-dependent sexual signals, Proc. Biol. Sci. 266 (1414) (1999) 1–12.
[54] D.K. Dowling, L.W. Simmons, Reactive oxygen species as universal constraints in life-history evolution, Proc. Biol. Sci. 276 (1663) (2009) 1737–1745.
[55] M. Mari, A. Morales, A. Colell, C. Garcia-Ruiz, J.C. Fernandez-Checa, Mitochondrial glutathione, a key survival antioxidant, Antioxid. Redox Signal. 11 (11) (2009) 2685–2700.
[56] R.A. Cavallaro, A. Fuso, V. Nicolia, S. Scarpa, S-adenosylmethionine prevents oxidative stress and modulates glutathione metabolism in TgCRND8 mice fed a B-vitamin deficient diet, J. Alzheimers Dis. 20 (4) (2010) 997–1002.
[57] A.A. Caro, A.I. Cederbaum, Antioxidant properties of S-adenosyl-L-methionine in Fe(2+)-initiated oxidations, Free Radic. Biol. Med. 36 (10) (2004) 1303–1316.
[58] I. Ohsawa, K. Nishimaki, C. Yasuda, K. Kamino, S. Ohta, Deficiency in a mitochondrial aldehyde dehydrogenase increases vulnerability to oxidative stress in PC12 cells, J. Neurochem. 84 (5) (2003) 1110–1117.
[59] J.J. Goronzy, H. Fujii, C.M. Weyand, Telomeres, immune aging and autoimmunity, Exp. Gerontol. 41 (3) (2006) 246–251.
[60] M. Bantscheff, S. Lemeer, M.M. Savitski, B. Kuster, Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present, Anal. Bioanal. Chem. 404 (4) (2012) 939–965.
[61] M. Rezeli, A. Vegvari, E. Silajdzic, M. Björkqvist, S.J. Tabrizi, T. Laurell, G. Marko-Varga, Inflammatory markers in Huntington's disease plasma-A robust nanoLC-MRM-MS assay development, Eupa Open Proteomics 3 (2014) 68–75.
[62] S. Seki, Y. Habu, T. Kawamura, K. Takeda, H. Dobashi, T. Ohkawa, H. Hiraide, The liver as a crucial organ in the first line of host defense: the roles of Kupffer cells, natural killer (NK) cells and NK1.1 Ag + T cells in T helper 1 immune responses, Immunol. Rev. 174 (2000) 35–46.
[63] R.D. Curran, T.R. Billiar, D.J. Stuehr, K. Hofmann, R.L. Simmons, Hepatocytes produce nitrogen oxides from L-arginine in response to inflammatory products of Kupffer cells, J. Exp. Med. 170 (5) (1989) 1769–1774.
[64] G.J. Sandland, D.J. Minchella, Costs of immune defense: an enigma wrapped in an environmental cloak? Trends Parasitol. 19 (12) (2003) 571–574.

[65] M. Triantafilou, D. Fradelizi, K. Triantafilou, Major histocompatibility class one molecule associates with glucose regulated protein (GRP) 78 on the cell surface, Hum. Immunol. 62 (8) (2001) 764–770.
[66] R.V. Rao, A. Peel, A. Logvinova, G. del Rio, E. Hermel, T. Yokota, P.C. Goldsmith, L.M. Ellerby, H.M. Ellerby, D.E. Bredesen, Coupling endoplasmic reticulum stress to the cell death program: role of the ER chaperone GRP78, FEBS Lett. 514 (2-3) (2002) 122–128.
[67] K.M. Fedorka, M. Zuk, T.A. Mousseau, Immune suppression and the cost of reproduction in the ground cricket, Allonemobius socius, Evolution 58 (11) (2004) 2478–2485.
[68] R.L. Lochmiller, C. Deerenberg, Trade-offs in evolutionary immunology: just what is the cost of immunity? Oikos 88 (1) (2000) 87–98.
[69] M.T. Siva-Jothy, Y. Tsubaki, R.E. Hooper, Decreased immune response as a proximate cost of copulation and oviposition in a damselfly, Physiol. Entomol. 23 (3) (1998) 274–277.
[70] P. Horak, I. Ots, L. Tegelmann, A. Moller, Health impact of phytohaemagglutinin-induced immune challenge on great tit (Parus major) nestlings, Can. J. Zool.-Rev. Can. Zool. 78 (6) (2000) 905–910.

B.4. Conclusion

Sur l'ensemble des gels 2D, 384 spots protéiques ont été détectés, parmi lesquels 17 présentaient une intensité différentielle entre les groupes (ANOVA unidirectionnelle + test posthoc de Tukey avec correction de Bonferroni pour comparaisons multiples ; $p<0.05$). Dans ces 17 spots, 18 protéines ont été identifiées, bien connues pour jouer un rôle dans la réponse au stress oxydatif, dans le métabolisme énergétique et la réponse au défi immunitaire.

Une analyse en composante principale (ACP) a été réalisée et a permis de mettre en évidence trois axes qui séparent les groupes, comme présenté en Figure I-2.

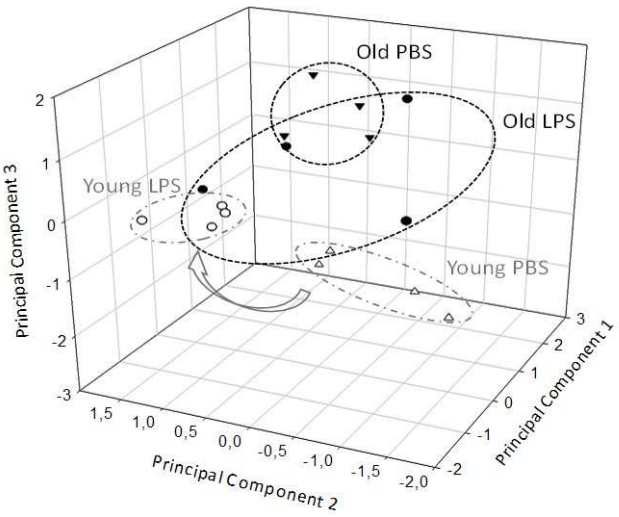


Figure I-2 : Résultat de l'analyse en composante principale (ACP), réalisée sur les 18 protéines différentiellement exprimées entre les groupes.
L'axe 1 explique 33% de la variance, l'axe 2 en explique 27% et l'axe 3 19%.

Pour les 3 composantes de l'ACP, des protéines de réponse au stress oxydatif étaient impliquées. La composante 1 expliquait 33% de la variance; la composante 2 expliquait 27% de la variance et était construite à partir de protéines également impliquées dans le métabolisme énergétique ; et enfin la composante 3 expliquait 19% de la variance et était constituée de protéines ayant également un rôle dans la réponse au défi immunitaire.
On remarque une nette séparation des jeunes individus en fonction du traitement selon les axes 1 et 2 principalement. Ceci correspond surtout à une augmentation des protéines répondant au stress oxydatif chez les jeunes souris ayant reçu une injection de LPS.
Chez les individus plus âgés, on ne voit pas de changements marqués suite à l'injection de LPS, en effet les groupes « Old PBS » et « Old LPS » ne sont pas vraiment séparés. On remarque en revanche une séparation des jeunes et des individus plus âgés, indépendamment du traitement, selon l'axe 3. Ceci correspond à un défaut de réponse des protéines antioxydantes et à des coûts oxydatifs pour les individus plus âgés, en pré et post- traitement. Tout cela semble suggérer une réduction des capacités anti-oxydantes chez ces individus.

Ainsi, nos résultats protéomiques montrent une augmentation chronique de l'abondance de protéines hépatiques répondant au stress oxydatif chez les souris plus âgées, d'où le fait qu'elles soient peu

affectées en réponse au LPS. En revanche, ces voies sont fortement mobilisées chez les souris plus jeunes. De plus, les souris plus âgées souffraient d'une activité plus faible de la glutathione-S-transferase et de dommages oxydatifs plus marqués, suggérant que le coût hépatique de l'immunité est plus fort que chez des souris plus jeunes. De manière générale, ces données démontrent donc que les coûts hépatiques d'une activation du système immunitaire sont liés au stress oxydatif que les souris plus âgées ont perdu la capacité de gérer.

C. Etude des marqueurs spléniques

C.1. Contexte analytique et objectifs

Nous nous sommes ensuite intéressés aux réponses **spléniques** induites en fonction de l'âge chez la souris, lors de la même activation du système immunitaire innée avec le LPS. La rate joue en effet un rôle important dans le système immunitaire ; c'est un organe lymphoïde secondaire.

Le traitement des souris était le même que pour l'étude des marqueurs hépatiques, présentés dans la paragraphe précédent, en Figure I-1. Cette fois-ci, nous disposions de 6 réplicas biologiques par groupe.

Cette étude a été réalisée alors que nous disposions d'instruments beaucoup plus performants et rapides au laboratoire, de ce fait il n'était plus nécessaire d'autant préfractionner les protéines qu'en 2D-DIGE pour obtenir une bonne couverture du protéome. Nous avons donc mis au point une stratégie de quantification « Label-free ». Nous avons également fait des tests quant à la préparation des échantillons, pour évaluer le bénéfice de préfractionner les protéines par gel mono dimensionnel SDS-PAGE. Il est connu que préfractionner les protéines permet d'augmenter la couverture de protéome analysé [145], le but de ce test était d'évaluer ce gain pour nos échantillons, et si celui-ci était suffisamment important.

C.2. Optimisations méthodologiques pour la préparation d'échantillons

Nous avons testé deux protocoles de préparation sur un même échantillon, qui sont détaillés dans la Partie Expérimentale (page 149) et rapidement résumés ici.

Protocole 1 (*sans préfractionnement*) : Après broyage du tissu et extraction des protéines, celles-ci ont été déposées sur gel SDS-PAGE et migrées jusqu'à l'entrée du gel de séparation (donc les protéines n'ont pas été séparées). Après analyse par spectrométrie de masse et recherche classique dans la banque de données Swiss-Prot *Mus Musculus*, nous avons pu identifier **596** protéines.

Protocole 2 (*avec préfractionnement*) : Après broyage du tissu et extraction des protéines, celles-ci ont été déposées sur gel SDS-PAGE et migrées sur 12mm. Ont ensuite été découpées 6 bandes de 2mm. Le contenu protéique de chaque bande a été analysé par spectrométrie de masse puis une recherche classique dans la banque de données Swiss-Prot *Mus Musculus* a permis d'identifier **1053** protéines, dans l'ensemble des 6 bandes.

Finalement, bien que le protocole 1 permette de minimiser le temps global d'analyse (car il n'y aurait qu'une seule analyse par échantillon, et non 6), il permet d'identifier beaucoup moins de protéines (presque deux fois moins), ce qui serait donc désavantageux. De plus, étant donné le faible nombre

d'échantillons à analyser (24), il est tout à fait envisageable de faire un préfractionnement sans pour autant que le temps d'analyse total ne nuise à la qualité des données.

C.3. Analyse des échantillons

Après extraction des protéines, nous avons donc réalisé un préfractionnement sur gel SDS-PAGE. Les données spectrales acquises sur un instrument de type Q-TOF (Impact HD ; Bruker) ont ensuite été interprétées par une recherche classique, dans la banque de données Swiss-Prot *Mus Musculus* grâce à l'algorithme de recherche Andromeda ; puis la quantification et normalisation des données ont été réalisées par MaxQuant.

Le protocole de l'ensemble de ces étapes est détaillé dans la Partie Expérimentale, en page 149.

C.4. Résultats

Sur l'ensemble des 24 échantillons, nous avons pu identifier **3022** protéines, et considérer **2026** protéines pour la quantification. En effet, les protéines qui présentaient plus de deux valeurs manquantes par groupe (soit moins de 4 valeurs valides) ont été écartées car leur quantification jugée pas assez robuste. En revanche, les cas où il n'y avait aucune valeur valide dans un groupe (cas absents/présents) ont été conservés. Les protéines « decoy » et les contaminants ont également été éliminés.

Ici, nous avons considéré les valeurs d'abondance protéique (« LFQ intensity ») fournies par Maxquant. Nous n'avons donc pas appliqué le critère de valeur manquante au niveau peptidique, ni calculeé un coefficient de corrélation entre peptides (comme cela sera présenté plus loin dans le manuscrit, Partie IIChapitre III.C.3, en page 92). De tels filtres auraient nécessité de travailler à partir des intensités peptidiques et donc de rassembler, à postériori, les peptides par protéines. Cela est tout à fait envisageable et a été récemment discuté par Goeminne *et al.* [146]. Nous avons préféré utiliser le résultat rendu par Maxquant et bénéficier de sa normalisation au niveau protéique, qui permet de corriger pour les peptides « aberrants » (i.e. dont le profil diffère des autres pour une protéine donnée) en calculant une médiane des rapports peptidiques (comme expliqué en Partie IChapitre IIIII.B.2.1.2).

Ici nous avons considéré les protéines quantifiées avec un peptide. Quantifier une protéine avec **un** peptide peut être dangereux car un seul peptide n'est pas forcément représentatif de l'abondance de la protéine entière, c'est pourquoi il est préférable d'avoir un maximum de peptides par protéine pour obtenir une quantification plus robuste et fiable. Ici, nous avons fait le choix de conserver les protéines quantifiées avec un seul peptide afin de conserver un maximum d'informations. A l'issue des interprétations, s'il s'avère qu'une de ces protéines à un intérêt particulier pour la question posée, il sera alors possible de confirmer les hypothèses par une autre méthode.

C.5. Suivi de la stabilité instrumentale

Au cours des 13 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 5 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de la 6ème bande de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit; Biognosys AG, Schlieren, Switzerland) en

quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure I-3 et Figure I-4.

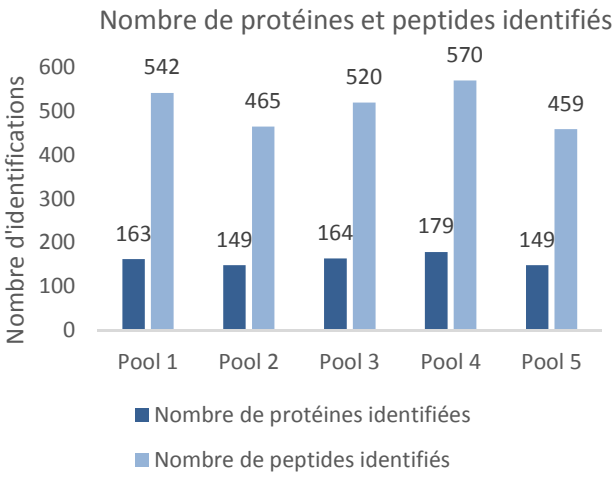


Figure I-3 : Nombre de protéines et peptides identifiés dans les 5 injections répétées de l'échantillon de référence (ou Pool).

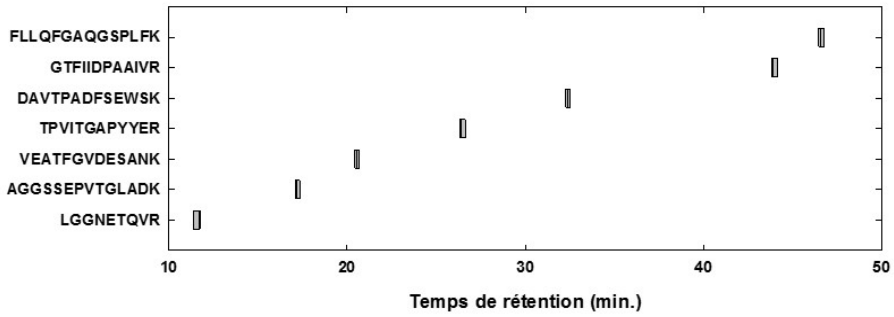


Figure I-4 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate que le nombre de protéines et peptides identifiés est resté stable au cours des 13 jours d'analyse, avec un CV de 7% et 9% respectivement, ce qui montre une bonne reproductibilité du système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 15,8%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation de temps de rétention des peptides iRT étaient en moyenne égaux à 0,7%.

C.6. Interprétation des résultats

Un test statistique (ANOVA unidirectionnelle + test posthoc de Tukey avec correction de Bonferroni pour comparaisons multiples ; $p<0.05$) a permis de mettre en évidence 294 protéines différentiellement exprimées entre les groupes. Une analyse en cluster a permis de déterminer que 7 d'entre elles participaient le plus fortement à la définition des axes de l'ACP. Nous sommes partis du postulat que ces variations protéomiques pouvaient être liées, pour le moins corrélées, à la masse corporelle et la taille des télomères mesurées par nos collaborateurs. L'analyse en composante principale (Figure I-5) a donc été réalisée sur les variations de masse corporelle, la taille des télomères et les 7 protéines qui contribuaient le plus significativement à la définition des axes de l'ACP (dont les

valeurs d'abondance sont présentées en Tableau Annexe 1, page 185). Ainsi, deux axes qui séparent les groupes ont été mis en évidence.

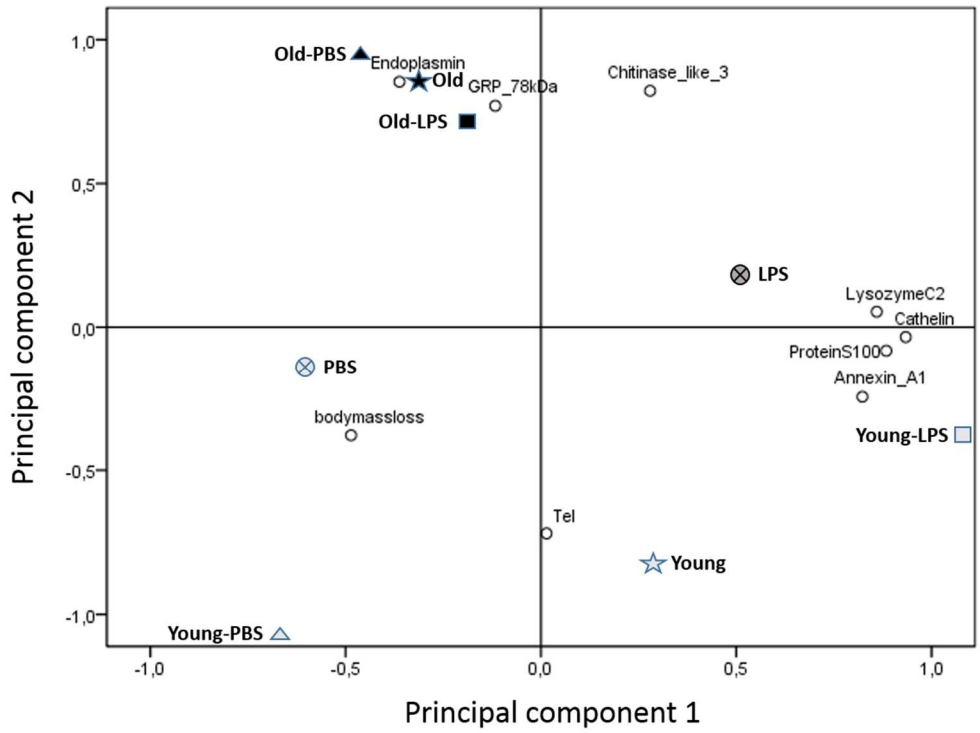


Figure I-5 : Analyse en composante principal réalisée sur les variations de masse corporelle (« bodymassloss »), la taille des télomères (« tel ») et 7 protéines les plus significatives (déterminées grâce à une analyse en cluster).
L'axe 1 explique 40% de la variance, l'axe 2 en explique 29,5%.

L'axe 1 sépare les « jeunes » individus en fonction du traitement (« Young-LPS » versus « Young-PBS »). En revanche, les individus plus âgés ne semblent pas répondre au LPS (« Old-LPS » versus « Old-PBS »), ce que l'on retrouvait dans le projet précédent sur les marqueurs hépatiques. Notamment, le LPS semble augmenter l'expression de la cathelicidin, chez les jeunes individus. Cette protéine est connue pour être exprimée dans le dysfonctionnement des télomères [147], et également pour répondre au LPS [148] (il s'agit d'un peptide anti microbien).
L'axe 2 permet de bien séparer les individus selon leur âge. La protéine chitinase-like protein 3, connue pour être associée à un dysfonctionnement des télomères [147, 149, 150], est surexprimée chez les individus plus âgés. Les régulations de cette protéine sont en accord avec le fait que les télomères sont plus courts chez les vieux individus. D'ailleurs, la taille des télomères participe fortement à définir l'axe 2 et sa variation est inversement liée à celle des trois autres protéines (endoplamsin, GRP78, et Chitinase-like protein 3). Ceci ouvre des voies d'étude sur leurs liens fonctionnels potentiels.

Finalement, nos résultats suggèrent i) un défaut de réponse au LPS chez les individus plus âgés qui pourrait être lié à une mal fonction de leur système immunitaire qui s'apparente à une induction chronique, reflétant une immunoscénéscence, ii) une possible dérégulation de la maintenance des télomères suite à l'injection du LPS chez les jeunes individus, indiquant des possibles « coûts » pro-vieillissement liés à l'activation du système immunitaire au niveau de la rate.

Ces régulations sont actuellement examinées en détail par nos collaborateurs. A l'issue de ces examens, un article sera préparé pour publier ces résultats.

II. Etude des coûts de l'immunité lorsque le contrôle du système immunitaire est altéré

A. Contexte biologique

La plupart des travaux dédiés à la problématique des coûts-bénéfices de la réponse immunitaire considèrent les coûts en termes de ressources consommées. Dans ce contexte, le paradigme de l'immuno-écologie a largement ignoré les coûts indépendants des ressources, ainsi que le rôle possible des parasites eux-mêmes sur le rapport coûts/bénéfices de l'immunité. Pourtant, les dommages causés par une réponse immunitaire mal ciblée ou surexprimée et leurs effets négatifs sont maintenant avérés. La réponse inflammatoire illustre parfaitement cela puisqu'elle est à l'origine de nombreuses atteintes responsables de pathologies courantes. Cependant, les hôtes sont dotés de mécanismes régulateurs qui contrôlent la réponse immunitaire et en atténuent les effets négatifs [151]. Or ces fonctions régulatrices méritent d'être considérées dans les scénarios évolutifs émis sur les défenses immunitaires. En effet, le niveau de régulation peut fortement moduler les effets négatifs de la réponse immunitaire avec des conséquences en termes d'aptitude phénotypique, c'est-à-dire de participation d'un individu à l'évolution de sa population. Ce projet vise à explorer ce thème dans un cadre évolutif (plusieurs espèces de mammifères et d'oiseaux seront à terme étudiées), en considérant certains effecteurs de l'inflammation et en se focalisant sur le rôle de la régulation de la réponse immunitaire. Dans le cadre de cette thèse, nous nous sommes intéressés à la souris. Le but ici est toujours d'étudier les coûts d'activation du système immunitaire (ou compromis immuns), non plus en fonction de l'âge comme précédemment, mais en absence de contrôle du système immunitaire. Pour cela, nous avons mimé une inflammation chronique croissante chez la souris.

Schéma expérimental :

L'expérience développée ici (voir Figure I-6) concerne la réduction de la fonction de régulation de l'inflammation par injection chronique d'un anticorps (Ac) anti-IL10 (IL10 étant une cytokine anti-inflammatoire), et ce à deux souches de souris différentes : des souris dont un gène inhibant l'activation des macrophages a été invalidé (UCP2-KO) et la souche témoin sauvage (UCP2). Enfin, une réaction inflammatoire a été induite par des injections de LPS.

Ici, on peut s'attendre à ce que plus les individus subissent une inflammation chronique, plus les coûts d'activation du système immunitaire seront importants. Les individus dont on contrôle la régulation de l'inflammation (Ac anti-IL10) et exposés au LPS devraient alors subir un coût plus important que les individus non régulé et exposé au LPS, eux-mêmes subissant un coût plus important que les individus non exposés. Enfin, l'activation des macrophages devrait entraîner une réponse plus efficace mais moins contrôlable.

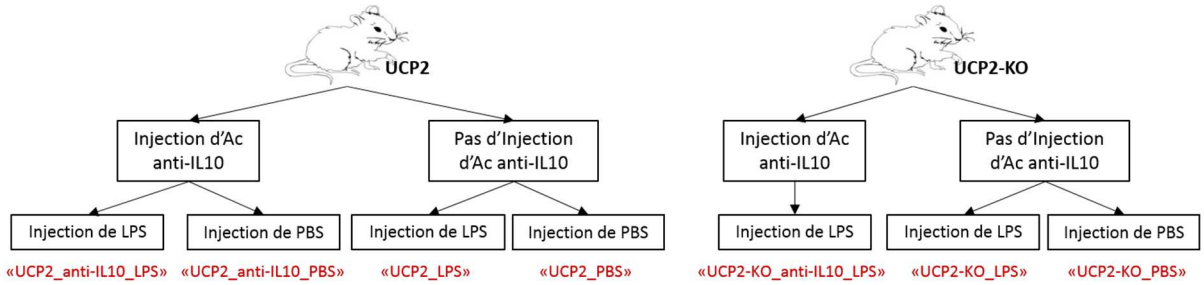


Figure I-6: Schéma expérimental du traitement des souris.

La condition « UCP2-KO_anti-IL10_PBS » est absente de l'expérience car il n'y avait qu'un individu dans ce groupe.
Le nombre de réplicas biologiques par groupe était de 4 ou 5.

B. Contexte analytique et objectifs

L'objectif ici était d'étudier les réponses spléniques à une réduction de la régulation de l'inflammation chez la souris. Nous avons donc mis en place une stratégie de quantification « Label-free » pour comparer les 7 conditions. Au vu du nombre d'échantillons à analyser (33), et malgré le bénéfice démontré dans la partie précédente d'un préfractionnement par SDS-PAGE des protéines, le temps d'analyse pour 33 échantillons préfractionnés dépasserait les 15 jours et pourrait induire des variations.

C. Analyse des échantillons

Après extraction des protéines, celles-ci ont été digérées en solution puis les peptides analysés sur un Q-Exactive+ (Thermo Fisher). Les données spectrales ont ensuite été interprétées par une recherche classique, dans la banque de données *Mus Musculus* grâce à l'algorithme de recherche Andromeda ; puis la quantification et normalisation des données ont été réalisées par MaxQuant.

Le protocole de l'ensemble de ces étapes est détaillé dans la Partie Expérimentale, en page 152.

D. Résultats

Sur l'ensemble des échantillons, nous avons identifié **4144** protéines. Pour la quantification, nous n'avons considéré que les protéines pour lesquelles il n'y avait pas plus d'une valeur manquante pour les groupes à n= 5, et aucune valeur manquante pour les groupes à n=4 (sauf pour les cas présents/absents) ; ont également été eliminées les protéines « decoy » et les contaminants ; finalement **3234** protéines ont été quantifiées.

Comme expliqué précédemment (Partie IIChapitre II.C.4), ce sont les valeurs d'abondance protéique (« LFQ intensity ») fournies par Maxquant qui ont été directement utilisées.

Ici aussi nous avons considéré les protéines quantifiées avec un peptide. Comme précisé précédemment (Partie IIChapitre II.C.4), ces protéines sont étudiées avec précaution.

E. Suivi de la stabilité instrumentale

Au cours des 7 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 4 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit ; Biognosys AG, Schlieren, Switzerland) en quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure I-7 et Figure I-8.

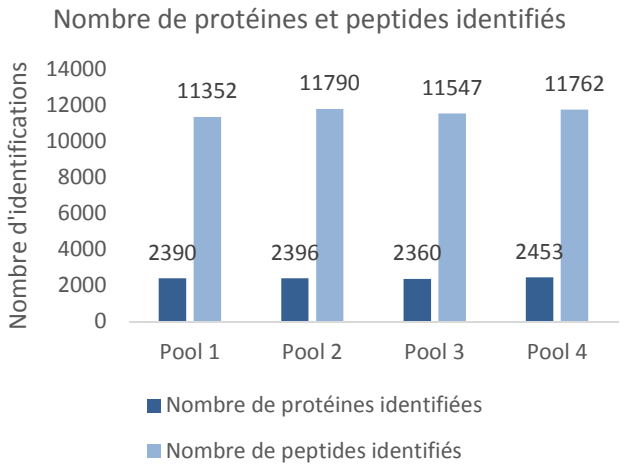


Figure I-7 : Nombre de protéines et peptides identifiés dans les 4 injections répétées de l'échantillon de référence (ou Pool).

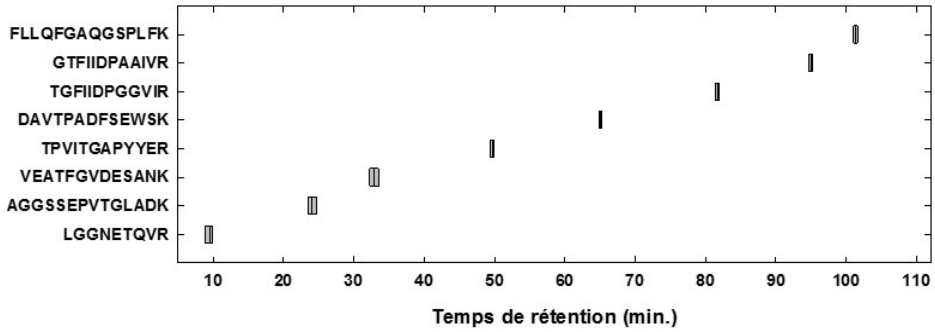


Figure I-8 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate, que le nombre de protéines et peptides identifiés est resté stable au cours des 7 jours d'analyse, avec un CV de 1,6% et 1,7% respectivement, ce qui montre une bonne reproductibilité du système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 10%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation de temps de rétention des peptides iRT étaient en moyenne égaux à 1,6%.

F. Interprétation des résultats

L'analyse de ces données reste très préliminaire. En effet, une analyse en composantes principales réalisée à partir des régulations des 292 protéines différentiellement abondantes (ANOVA unidirectionnelle + test posthoc de Tukey avec correction de Bonferroni pour comparaisons multiples ;

p<0.05) entre les groupes de souris permet de définir 2 composantes (2 axes) qui séparent les différents animaux mais n'expliquent qu'un total de 43% de la variance totale (Figure I-9 B). Nous nous sommes plus particulièrement intéressés aux protéines qui varient le plus, i.e. celles dont les coordonnées (Figure I-9 A) sont inférieures à -0.5 ou supérieures à 0.5 pour la dimension 1. Leurs coordonnées sont présentées dans les Tableau 1 et Tableau 2; et leur valeur d'abondance dans le Tableau Annexe 2, page 186.

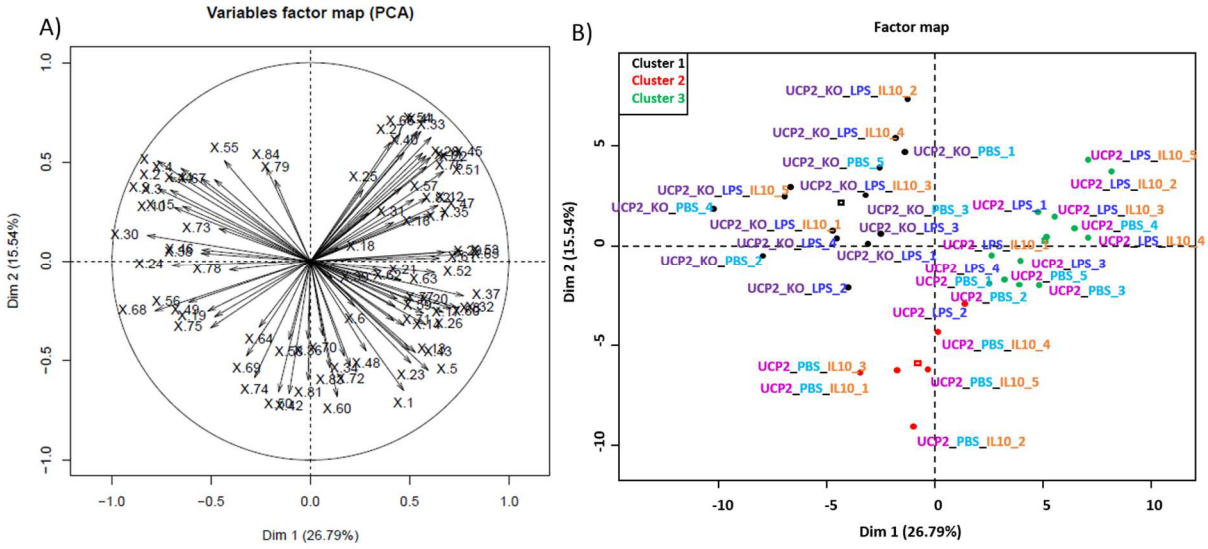


Figure I-9 : A) Coordonnées des variables de l'analyse en composante principale (ACP) ; B) ACP réalisée sur les protéines différentielles (ANOVA p-value < 0,05).

La nomenclature des groupes utilisée est celle de la Figure I-6. Une analyse en cluster a permis de séparer les individus en 3 groupes (ou 3 clusters) représentés ici en noir, vert et rouge.

En violet est indiqué le traitement UCP2-KO ; en rose UCP2 ; en bleu foncé LPS ; en bleu ciel PBS ; en orange IL10.

Ainsi, la première dimension (Axe 1) sépare toutes les souris UCP2-KO des autres animaux. On constate que l'abondance de protéines notamment impliquées dans la fonction mitochondriale est diminuée chez les souris UCP2-KO (protéines en bleu dans le Tableau 1). Ceci pourrait refléter un dérèglement mitochondrial chronique chez ces souris. Ce dérèglement pourrait témoigner des modifications dues à l'âge et pourraient traduire une trajectoire de vieillissement mitochondrial particulière des souris UCP2-KO. A l'inverse, l'abondance de protéines notamment impliquées dans la réponse immunitaire et l'hématopoïèse est augmentée chez les souris UCP2_anti-IL10_LPS (protéines en vert dans le Tableau 1). Ceci pourrait refléter une situation pro-inflammatoire chez ces souris, induite par l'absence du retro-contrôle négatif de l'IL10 sur la réponse immunitaire, qui agit notamment via une régulation de la différentiation des cellules intervenant dans l'immunité [152]. De plus, ces effets de l'inhibition de l'IL10 ne sont pas visibles chez les souris UCP2-KO, et ceci quel que soit le traitement (PBS ou LPS). Ces souris ne semblent donc pas répondre non plus au LPS. Tout ceci pourrait être dû au fait que les souris UCP2-KO ont des macrophages et phagocytes extrêmement actifs [153, 154], d'où un système immunitaire et anti-inflammatoire déjà activé au maximum et répondant très vite au défi immunitaire. Cette hyper-sensibilité du système immunitaire et la clairance rapide de l'antigène LPS injecté pourrait expliquer l'absence d'effet des injections chroniques pro-inflammatoires qui ont été conduite sur 1 an.

Concernant la seconde dimension, celle-ci sépare surtout les animaux UCP2_anti-IL10_PBS des autres animaux, notamment les UCP2-KO. Ceci pourrait refléter l'effet dé-régulateur de l'inhibition de l'IL10 sur les souris possédant une UCP2 fonctionnelle. Les souris UCP2-KO expriment en effet davantage des protéines impliquées dans l'activation du système immunitaire, et des protéines chaperonnes ou encore structurales, mais indifféremment selon le traitement PBS ou LPS (protéines en bleu dans le Tableau 2). Les animaux UCP2_anti-IL10_PBS expriment moins les protéines de synthèse ou encore certaines protéines immunitaires (protéines en rouge dans le Tableau 2), ce qui pourrait refléter une situation pro-inflammatoire.

Tableau 1 : Coordonnées des protéines sur l' axe 1 de l'ACP présentée en Figure I-9.
En vert sont représentées les protéines surexprimées chez les souris UCP2_anti-IL10_LPS ; en bleu les protéines sous-exprimées chez les UCP2-KO.

Numéro d'accession	Nom de la protéine	Coordonnées sur l'axe 1 de l'ACP
Q6ZQA0	Neurobeachin-like protein 2	0,771
O35639	Annexin A3	0,766
P32261	Antithrombin-III	0,765
P06909	Complement factor H	0,739
Q6GV12	3-ketodihydrosphingosine reductase	0,708
Q9JKR6	Hypoxia up-regulated protein 1	0,681
Q8BHN3	Neutral alpha-glucosidase AB	0,677
Q9CXI5	Mesencephalic astrocyte-derived neurotrophic factor	0,671
P62500	TSC22 domain family protein 1	0,669
P04186	Complement factor B	0,649
Q61703	Inter-alpha-trypsin inhibitor heavy chain H2	0,643
Q9D7N9	Adipocyte plasma membrane-associated protein	0,626
P27046	Alpha-mannosidase 2	0,611
Q9D0F3	Protein ERGIC-53	0,607
P20029	78 kDa glucose-regulated protein	0,605
Q9WV98	Mitochondrial import inner membrane translocase subunit Tim9	-0,611
Q78IK2	Up-regulated during skeletal muscle growth protein 5	-0,642
Q8BYB9	Protein O-glucosyltransferase 1	-0,65
Q9ES97	Reticulon-3	-0,686
Q9CX86	Heterogeneous nuclear ribonucleoprotein A0	-0,697
Q8K1R3	Polyribonucleotide nucleotidyltransferase 1, mitochondrial	-0,705
Q9ERF3	WD repeat-containing protein 61	-0,714
Q9D1L0	Coiled-coil-helix-coiled-coil-helix domain-containing protein 2	-0,766
P02089	Hemoglobin subunit beta-2	-0,76996
Q8R0F6	Integrin-linked kinase-associated serine/threonine phosphatase 2C	-0,782
Q8JZX4	Splicing factor 45	-0,821

Tableau 2 : Coordonnées des protéines sur l' axe 2 de l'ACP présentée en Figure I-9.
En bleu les protéines surexprimées chez les UCP2-KO ; en rouge les protéines sous-exprimées chez les UCP2_anti-IL10_PBS.

Numéro d'accession	Nom de la protéine	Coordonnées sur l'axe 2 de l'ACP
O54734	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kDa subunit	0,656
Q9DBG6	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2	0,654
Q9ESP1	Stromal cell-derived factor 2-like protein 1	0,644
Q9D0F3	Protein ERGIC-53	0,624
Q5XJY5	Coatomer subunit delta	0,5996
P61961	Ubiquitin-fold modifier 1	0,548
Q9JKR6	Hypoxia up-regulated protein 1	0,547
Q80WW9	DDRGK domain-containing protein 1	0,545
Q9D662	Protein transport protein Sec23B	0,532
P20029	78 kDa glucose-regulated protein	0,524
B2RXS4	Plexin-B2	0,507
P02089	Hemoglobin subunit beta-2	0,506
P48453	Serine/threonine-protein phosphatase 2B catalytic subunit beta isoform	-0,514
Q9CSN1	SNW domain-containing protein 1	-0,529
Q9JJF3	Bifunctional lysine-specific demethylase and histidyl-hydroxylase NO66	-0,538
Q9ERE7	LDLR chaperone MESD	-0,55
Q9D6Z1	Nucleolar protein 56	-0,587
Q8BFR5	Elongation factor Tu, mitochondrial	-0,599
Q9QZI9	Serine incorporator 3	-0,651
P62307	Small nuclear ribonucleoprotein F	-0,659
Q9QXA5	U6 snRNA-associated Sm-like protein LSm4	-0,683

Comme déjà rapporté plus haut, toutes ces analyses statistiques doivent encore être raffinées. Les données protéomiques définitives seront par la suite mises en relation avec d'autres mesures réalisées chez les mêmes animaux (stress oxydatif, variations de masse, courbes de survie…), pour mieux comprendre le rôle potentiel de l'UCP2 dans la réponse immunitaire et l'immunosénescence.

A l'issue de ces examens, un article sera préparé pour publier ces résultats.

III. Conclusion

Ces trois projets nous ont donc permis d'apporter des éléments de réponse quant aux réponses hépatiques et spléniques induites lors d'une activation du système immunitaire, ces réponses pouvant s'apparenter à des coûts plus ou moins délétères.

Du point de vue analytique, les deux projets traitant de l'immunosenescence ont été l'occasion d'observer l'impact de la méthode de quantification choisie. En effet, les contraintes techniques (absence d'instrument rapide et résolutif) ont imposé de mettre en place une stratégie de quantification 2D-DIGE (qui a donc permis de pallier aux limitations instrumentales) pour l'étude hépatique, tandis que l'étude splénique, grâce à l'arrivée d'instruments plus performants, a été réalisée selon une stratégie de quantification « label-free » XIC MS1. Ces deux études ne concernent pas le même tissu mais ont été réalisées sur le même organisme, on peut donc en tirer des conclusions : en 2D-DIGE, 384 spots protéiques ont été quantifiés ; tandis qu'en label-free nous avons quantifié plus de 2000 protéines. On peut ainsi obtenir une couverture de séquence plus profonde en label-free, malgré le fort pouvoir résolutif de la 2D-DIGE. Cette technique souffre effectivement de limitations connues, telles que la faible représentation des protéines peu abondantes, d'acidité, basicité ou poids moléculaires extrêmes et les difficultés d'automatisation [155, 156]. La méthode « label-free » souffre également de limitations (absence de standard interne pour limiter les biais techniques notamment, couverture du protéome non complète), mais permet tout de même d'obtenir une couverture plus profonde et est plus rapide à mettre en œuvre ; c'est pourquoi elle a progressivement remplacé la méthode DIGE au cours des dernières années [102]. Ainsi, lorsque les moyens techniques le permettent, il est préférable de mettre en place une stratégie « label-free ».

De plus, lorsque le nombre d'échantillons le permet, il est préférable de réaliser un préfractionnement des protéines. Bien sûr, il est difficile d'évaluer un nombre d'échantillons précis à partir duquel préfractionner devient nuisible à la reproductibilité des données. De plus, avec des appareils de plus en plus performants, on peut augmenter la profondeur de protéome analysé sans avoir à préfractionner. En effet, l'analyse des tissus de rate sur le projet « immunosenescence » a été réalisé en **préfractionnant les protéines** par gel SDS-PAGE sur un **Q-TOF** (Impact HD, Bruker) ; tandis que l'analyse des rates sur le projet « étude évolutive de l'inflammation », a été réalisé **sans préfractionnement** sur un appareil plus récent et plus performant : l'**Orbitrap™** (Q-Exactive+ Thermo Fisher Scientific). La deuxième analyse a permis d'identifier un plus grand nombre de protéines (4144 vs. 3022), ce qui confirme le bénéfice d'utiliser un appareil plus performant, ce qui permet de compenser le fait de ne pas préfractionner.

Partie II : Résultats

Chapitre II : Analyse protéomique quantitative chez des organismes non séquencés

Partie II : Résultats

Chapitre II : Analyse protéomique quantitative chez des organismes non séquencés

I. Développements méthodologiques pour l'identification et la quantification fiable des peptides séquencés *de novo*

A. Contexte biologique: importance du tissu adipeux brun dans le contrôle de la balance énergétique

Au niveau mondial, le nombre de personnes en surcharge pondérale (1.5 milliards) [157] a aujourd'hui dépassé le nombre de personnes sous-nutries (~900 millions). Alors que depuis des décennies les études réalisées chez l'Homme et les modèles rongeurs de laboratoire ont clarifié de nombreux mécanismes de l'obésité [158], les traitements restent essentiellement inefficaces. Or l'obésité est un trouble métabolique majeur qui peut entraîner de nombreux problèmes, dont le diabète [158]. Il est donc urgent de trouver des nouveaux modèles d'études. Au cours de l'évolution, certaines espèces ont développé des adaptations à des contraintes environnementales extrêmes, leur permettant de ne pas présenter de troubles métaboliques dans des environnements conduisant à l'obésité et au diabète chez l'Homme. C'est par l'étude d'une telle espèce, le campagnol des champs, et de ses réponses à une diète obésogène que ce projet vise à découvrir de nouvelles voies de lutte contre l'obésité induite par l'alimentation.

L'obésité est caractérisée par un bilan énergétique positif, l'organisme recevant plus d'énergie qu'il n'en dépense [158]. Dans ce cas, l'excès d'énergie est stocké sous forme de réserves lipidiques dans le tissu adipeux blanc. Les entrées d'énergie proviennent bien sûr de l'alimentation alors que la dépense d'énergie est multifactorielle : métabolisme de base, thermogenèse, travail musculaire, productions.
Chez les petits mammifères, la thermorégulation peut impliquer la thermogenèse dite de frisson, qui est une réponse immédiate au froid. Dans ce cas, ce sont les contractions musculaires qui produisent de la chaleur pour assurer l'homéostasie thermique, mais l'efficacité va rester limitée. La thermogenèse sans frisson (NST) peut prendre le relai. Celle-ci passe par la production de chaleur due à l'activation du tissu adipeux brun (BAT) [159, 160].
Les cellules du BAT sont très différentes des adipocytes blancs. Elles contiennent de nombreuses petites gouttelettes contenant des lipides (alors que les adipocytes blancs possèdent une vacuole lipidique occupant tout l'espace), mais surtout elles sont remplies de nombreuses mitochondries (voir Figure II-1) [161]. Au niveau des membranes internes de ces mitochondries, la synthèse d'ATP est couplée aux réactions d'oxydation de la chaine respiratoire. Brièvement, la chaine respiratoire permet une accumulation de protons dans l'espace inter-membranaire des mitochondries, et le retour de ces protons dans la matrice mitochondriale est réalisée par la synthase d'ATP qui utilise cette énergie pour générer de l'ATP. Or ce couplage n'est pas parfait, notamment du fait de la présence d'une protéine,

l'UCP1 (« *uncoupling protein 1* »), qui permet le retour des protons depuis l'espace inter-membranaire vers la matrice mitochondriale sans passer par l'ATP synthase et, dans ce cas, l'énergie est dissipée sous forme de chaleur [162] .

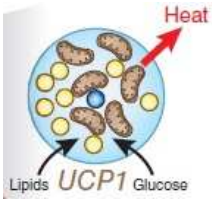


Figure II-1 : Représentation schématique d'une cellule adipeuse brune[161].

La présence de BAT chez les nouveaux nés est connue depuis longtemps, mais la récente découverte de leur présence chez l'homme adulte [163] a relancé l'intérêt porté à ce tissu dans la lutte contre l'obésité. En effet, Fridyland *et al.* ont émis l'hypothèse selon laquelle les troubles métaboliques chez l'homme (tels que l'obésité ou le diabète) pourraient être dus à un fonctionnement inadéquat de leur BAT en réponse à des changements environnementaux (température et régime alimentaire) [164]. Le possible rôle du BAT comme acteur essentiel du bilan énergétique de l'organisme réside dans le fait qu'il convertit l'énergie emmagasinée en chaleur (voir plus haut). Cette activité est très loin d'être anecdotique puisqu'il a été montré que chez l'adulte la production de chaleur due au BAT représente 15% de la dissipation moyenne d'énergie journalière [165]. De plus, le fait que le BAT soit « activé » en réponse à l'alimentation [166] renforce son potentiel rôle protecteur vis-à-vis de pathologies comme l'obésité ou le diabète [167] [168] [169, 170].

L'objectif de ce projet était donc d'étudier le rôle possible du BAT dans la protection contre l'obésité. Cette étude a été réalisée en collaboration avec le Dr Pierre Bize de l'Université d'Aberdeen (Ecosse) qui a artificiellement sélectionné 2 lignées de campagnols des champs (*Microtus arvalis*, Taxonomie 47230) pour leurs niveaux bas ou élevés de NST. Des données préliminaires, montrant que les campagnols présentant des bas niveaux de NST ont une masse plus élevée que les autres à l'age adulte (Bize et al., résultats non publiés), suggèrent que ce modèle animal unique pourrait permettre de mieux comprendre le possible rôle du BAT dans la protection contre l'obésité. Ces 2 lignées de campagnols ont donc été exposées soit à un régime alimentaire contrôle (4.5% de graisses) soit à un régime enrichi en graisses (23.6% de graisse) et maintenus dans un environnement froid (14°C) ou chaud (28°C).

Du point du vue analytqiue, ce travail était spécifiquement dédié au développement d'une stratégie de protéomique quantitative pour pouvoir ensuite analyser le protéome différentiel du BAT de ces animaux en fonction de leur lignée, de la température et du traitement nutritionnel subi. Les verrous à lever résidaient dans l'optimisation des méthodes et du traitement des données protéomiques afin de les adapter aux échantillons de Campagnol, une espèce dont le génome n'est pas connu.

Schéma expérimental :

Le schéma expérimental de sélection des campagnols est présenté sur la Figure II-2. Les analyses protéomiques consistaient à comparer 8 conditions, chaque condition composée de 6 réplicas.

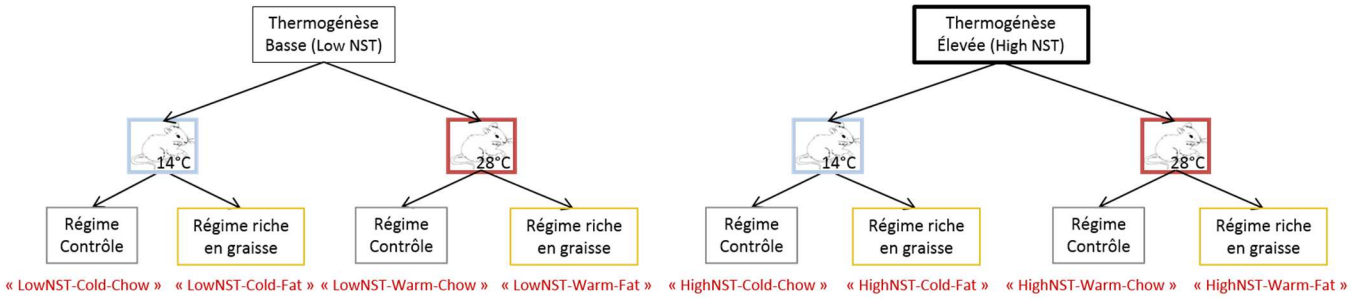


Figure II-2 : Schéma expérimental du traitement appliqué aux campagnols.
Des animaux des deux lignées (thermogénèse élevée ou non) ont été maintenus dans un environnement froid (14°C) ou chaud (28°C) puis soumis soit à un régime gras (23.6% de graisse) ou contrôle (4.5% de graisses).

B. Contexte analytique et objectifs

Le campagnol est un modèle de choix pour l'étude de l'obésité, cependant, le génome de cet organisme n'est pas séquencé, ce qui complexifie le traitement des données. En effet, nous ne disposons pas d'une banque de données qui contiendrait les séquences protéiques du campagnol, banque essentielle pour réaliser une recherche par empreinte de fragments peptidiques (grâce à Mascot notamment). La première étape a donc été de trouver une banque de données d'un (ou de plusieurs) organisme(s) suffisamment proche(s) du campagnol pour réaliser des identifications par homologie de séquences grâce aux peptides dont la séquence a été strictement conservée au cours de l'évolution entre le campagnol et cet(ces) autre(s) organisme(s).

On peut imaginer que, bien que cette méthode permette d'identifier un certain nombre de protéines, il restera un très grand nombre de spectres non interprétés car les peptides sont absents de la banque de données. Afin d'augmenter la couverture du protéome étudié, nous avons donc mis en place une stratégie de séquençage *de novo*. Cette stratégie, donc le principe est expliqué en page 30, permet d'interpréter les spectres issus de la fragmentation de peptides dont la séquence n'est pas strictement conservée. Pour ce faire, nous avons utilisé le logiciel PepNovo [57] puis l'algorithme MS BLAST (Basic Local Alignment Search Tool). Nous avons souhaité dans un premier temps évaluer le fonctionnement de PepNovo, pour en optimiser le paramétrage. Puis nous avons déterminé des paramètres de validation des assignations spectres-peptides et de filtre « manuel » des identifications.

Le but de ce projet étant de déterminer le protéome différentiel du BAT des campagnols selon la lignée, la diète ou encore la température, nous avons de plus mis en place une quantification de ces protéines. En effet, c'est l'information de l'abondance relative de chaque protéine entre les conditions qui nous permettra de répondre à la question posée. Vu le contexte, la méthode la plus appropriée ici est une méthode de quantification globale relative sans marquage par extraction des courants d'ions (dont le principe est expliqué en page 48).

L'objectif de ce projet était donc de mettre en place une méthode « Label-free » XIC MS1 à partir de l'identification fiable des peptides par Mascot et par séquençage *de novo*. Si la mise en place de cette stratégie (d'un point de vue bio-informatique) était aisée pour la première catégorie de peptides, il a en revanche été nécessaire de procéder à différents développements pour les peptides « *de novo* », que nous allons développer dans la suite de ce chapitre.

Enfin, Le grand nombre d'échantillons (n= 48) rendait difficile la mise en place d'un préfractionnement protéique car le temps d'analyse aurait alors été trop important, ce qui aurait pu nuire à la reproductibilité des données. Nous avons donc décidé de traiter les échantillons par digestion liquide, afin de n'avoir qu'une analyse par échantillon.

C. Développement méthodologique pour l'analyse des données

C.1. Stratégie d'identification

C.1.1. Identification grâce à l'algorithme de recherche Mascot (Empreintes de fragments peptidiques)

L'identification par recherche classique en banque de données a été effectuée grâce à l'algorithme de recherche Mascot. Dans un premier temps, nous avons effectué la recherche dans différentes banques de données, afin de déterminer la plus appropriée.

Le campagnol des champs appartient à l'ordre des rongeurs et à la classe des mammifères, c'est également une espèce proche de la souris (la souris et le campagnol font partie de la même super-famille des Muridés). Lors d'une étude préliminaire, en validant les identifications avec un FDR protéique de 1% (à l'aide du logiciel Scaffold v4.3), la recherche dans la banque de données restreinte aux séquences protéiques de souris (*Mus Musculus*, taxonomie 10090, 16 754 séquences) a permis d'identifier **1252** protéines, alors que **1433** protéines étaient identifiées si on utilisait une banque contenant les séquences protéiques de tous les rongeurs (*Rodentia*, taxonomie 9989, 26 379 séquences) et **1804** si on utilisait une banque contenant les séquences protéiques de tous les mammifères (*Mammalia*, taxonomie 40674, 66 414). Soit un gain avec la banque « Mammifères » de 20% et 30% respectivement par rapport aux banques « rongeurs » et « souris ».

Il n'est pas surprenant que l'on identifie plus de protéines avec la banque « Mammifères », étant donné qu'elle contient entre 60 et 75% de séquences protéiques en plus, respectivement versus les banques « rongeurs » et « souris ». Cela montre également que malgré la forte proximité phylogénétique entre la souris et le campagnol des champs, la banque restreinte à la souris ne suffit pas à expliquer tous les spectres de fragmentation des peptides du campagnol.

Nous avons donc utilisé la banque de données « Mammifères » pour identifier les protéines avec Mascot. Par contre, l'inconvénient de travailler avec des banques non restreintes à une seule taxonomie réside dans la « redondance » des protéines identifiées : on peut identifier plusieurs protéines d'espèces différentes, qui sont en fait des homologues d'une seule protéine (celle du campagnol ici). Pour « réunir » ces homologues, nous avons réalisé un BLAST de type fasta 36 [171] qui permet d'aligner les séquences des protéines identifiées à celles contenues dans une banque de données – idéalement restreinte à une taxonomie – pour obtenir un identifiant unique. Ici, nous avons aligné nos séquences à celles de la banque *Homo Sapiens* (taxonomie 9606) car c'est la mieux annotée et une grande partie des protéines étaient déjà identifiées dans cette banque. Les critères utilisés pour ces analyses BLAST respectaient les conditions suivantes : Score blast $\leq$ 0.00001, pourcentage de similarité > 50% et vérification que les noms et/ou symboles des homologues étaient identiques ou similaires.

C.1.2. Séquençage *de novo*

C.1.2.1. Fonctionnement du logiciel PepNovo combiné à l'algorithme MS BLAST

Le principe du séquençage *de novo* est expliqué en page 30, et celui du BLAST en page 32. Ici, nous allons développer plus spécifiquement le fonctionnement du logiciel PepNovo [57]. Ce logiciel fonctionne sur la base d'un entrainement automatique qui doit être réalisé sur les spectromètres de masse utilisés. Les concepteurs nous ont affirmé que certaines fonctionnalités de l'algorithme PepNovo ne sont plus disponibles lorsqu'on entraine soi-même le logiciel. Un entrainement par défaut a été réalisé par les concepteurs. Nous avons utilisé la version « par défaut » de PepNovo et vérifié qu'il donnait de bons résultats avec les données issues de nos spectromètres de masse.

Les spectres de fragmentation étant très complexes, il y a souvent plusieurs interprétations possibles, plusieurs séquences qui peuvent expliquer un spectre (voir ① de la Figure II-3). Ces solutions sont appelées « *PepNovo sequence tag* ». Le logiciel PepNovo assigne deux scores à ces différentes solutions :

- Le premier score, appelé « PepNovo score », donne la probabilité que la fragmentation du peptide ait créé le spectre.
- Le deuxième score, appelé « Rank score » [172], est basé sur un apprentissage automatique (*machine learning* en anglais), méthode qui consiste à établir un modèle à partir de données expérimentales (ici 300 000 spectres). Ce score est basé sur différents critères [172] :
    - **« Peak rank prediction »** : Prédiction d'un spectre théorique pour les peptides candidats et établissement d'un rang (basé sur l'intensité) pour chaque fragment du peptide et comparaison au rang observé sur le spectre expérimental.
    - **« Peak annotation »** : Examine la qualité du « peptide-spectrum match » (ou PSM ; il s'agit de l'assignation d'une séquence peptidique à un spectre) en utilisant des fonctions qui comptent le nombre de pics annotés parmi les 25 ou 50 plus intenses du spectre.
    - **« Peak offset »** : Différence entre la masse théorique et expérimentale d'un fragment.
    - **« Sequence composition »** : Se base sur le fait que certaines séquences sont plus susceptibles d'être identifiées par MS/MS.

Pour un spectre, le tag de séquence qui a le meilleur « PepNovo score » n'est pas forcément celui qui a le meilleur « Rank score ». Dans les paramètres par défaut, les séquences sont classées par ordre décroissant de score PepNovo.

Une première question se pose ici : **vaut-il mieux utiliser la séquence qui a le meilleur « PepNovo score » ou celle qui a le meilleur « Rank score » ?**

Ensuite le tag de séquence est modifié avant d'être envoyé à MS BLAST (voir ② de la Figure II-3): La lettre B est ajoutée du côté N-terminal; une Lysine ou une Arginine est ajouté du côté C-terminal; les acides aminés manquants sont remplacés par des X; certains acides aminés dont la probabilité est

faible sont remplacés par des X; dans le but d'augmenter la probabilité de la séquence d'être reconnue en BLAST. Ce tag modifié est appelé « *PepNovo extended sequence for Blast* ».

La deuxième question qui se pose est alors: **faut-il considérer ce tag modifié pour le BLAST (*PepNovo extended sequence for Blast*), ou bien le tag non modifié déduit directement du spectre de fragmentation (*PepNovo sequence tag*)?**

Une fois les séquences modifiées, celles-ci sont soumises à l'algorithme MS BLAST, de manière automatique dans la suite logicielle MSDA [55] ; il est possible d'envoyer plusieurs solutions par spectre.

La troisième question qui se pose est donc: **combien de solutions faut-il envoyer à MS BLAST? Et comment savoir laquelle conserver?**

MS BLAST va diviser le tag qui lui est soumis en plusieurs tags dont les séquences se recouvrent [173] (voir ③ de la Figure II-3) afin d'augmenter les probabilités d'identification. Ces tags sont appelés « *MSBlast query* ». Ces « *query* » vont ensuite être comparés aux séquences contenues dans la banque de données (séquences appelées *« MSBlast subject* », voir ④ de la Figure II-3), puis un score appelé « MS BLAST score » (voir ⑤ de la Figure II-3) est attribué à chaque correspondance tag – séquence théorique, reflétant la qualité de la correspondance. Ici encore, il peut y avoir plusieurs solutions par « query ». Enfin, chaque peptide identifié est assigné à la ou les protéines auxquelles il appartient ; ce qui ajoute un niveau de redondance et donc de complexité au traitement des résultats.

Ainsi, pour un spectre, on se retrouve avec une multitude de solutions, dont il faut déterminer laquelle est la meilleure.

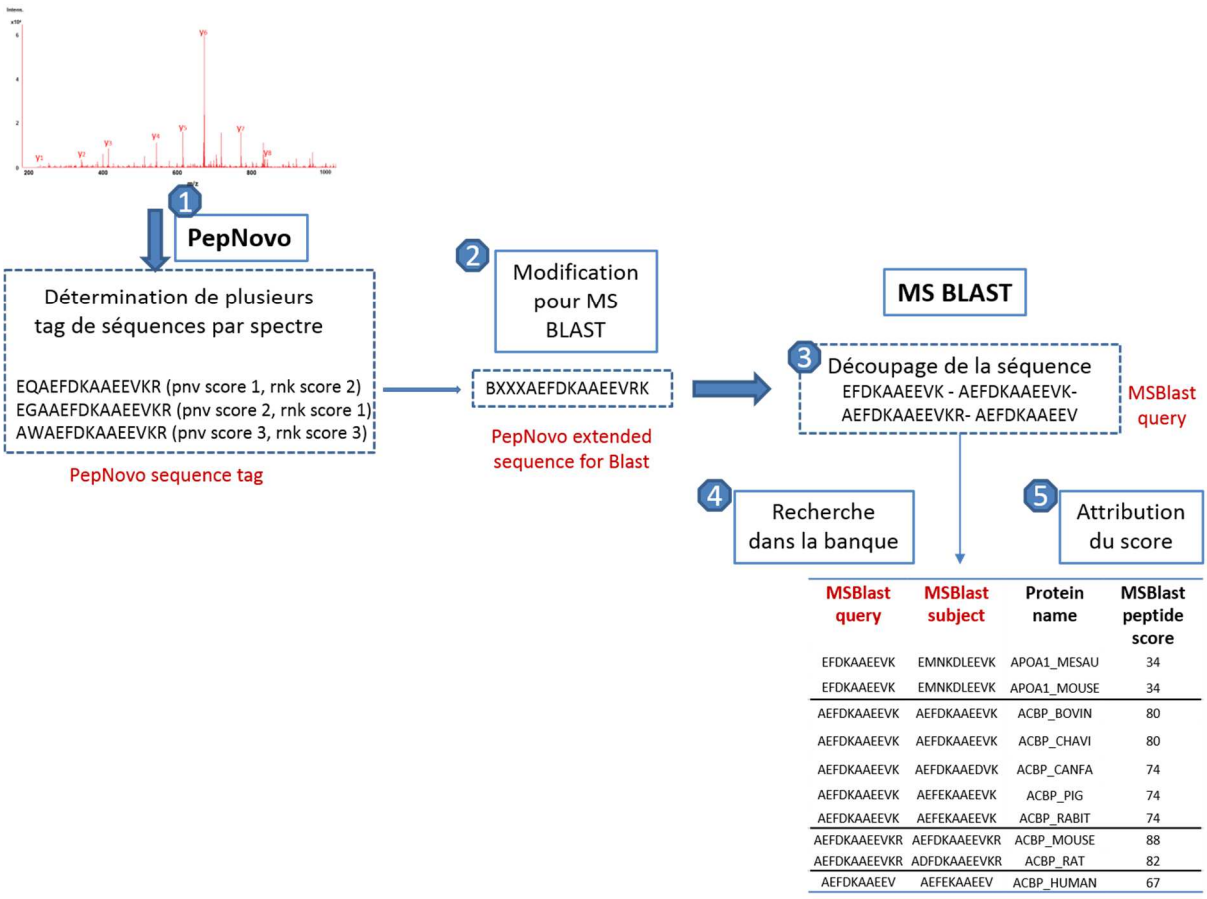


Figure II-3 : Exemple schématique de l'interprétation d'un spectre par PepNovo et recherche par homologie de séquence par MS BLAST.

① : Interprétation d'un spectre par PepNovo. A chaque tag est attribué deux scores : pnv score = « PepNovo score » et rnk score = « Rank score » ; ici les numéros attribués au score montrent le rang du peptide suivant le score choisi. Par défaut, les séquences sont classées selon le « PepNovo score » ; ② : Modification de la séquence pour MS BLAST ; ③ : Découpage de la séquence par MS BLAST ; ④ : Recherche dans la banque de données ; ⑤ : Attribution d'un score par MS BLAST.

Pour répondre aux interrogations posés dans ce paragraphe, nous avons utilisé un **mélange connu**: un digeste de levure dans lequel ont été ajoutées différentes quantités de protéines UPS1 (Universal Proteomics Standard, Sigma Aldrich ; mélange équimolaire de 48 protéines), qui a été analysé sur différents spectromètres de masse. Les données spectrales ont ensuite été interprétées par Mascot et par PepNovo avec différents critères, afin de déterminer avec quels paramètres le séquençage *de novo* donnait le même résultat que la recherche classique.

C.1.2.2. Analyse d'un mélange connu pour l'optimisation de PepNovo

Les données spectrales obtenues par l'analyse d'un mélange d'UPS1 ont été interprétées par Mascot (FDR 1% ; Scaffold v4.3) à l'aide d'une banque contenant les séquences protéiques des 48 protéines UPS1 ; puis le logiciel Recover a été utilisé pour conserver uniquement les spectres déjà interprétés par Mascot, qui ont alors été soumis à interprétation par PepNovo suivi de MS BLAST (recherche par homologie de séquence dans la même banque de données). Pour chaque spectre, on comparera donc le résultat obtenu par les deux méthodes, en considérant la recherche classique Mascot comme étant la référence (puisqu'on sait ce que l'on doit trouver ; voir Figure II-4).

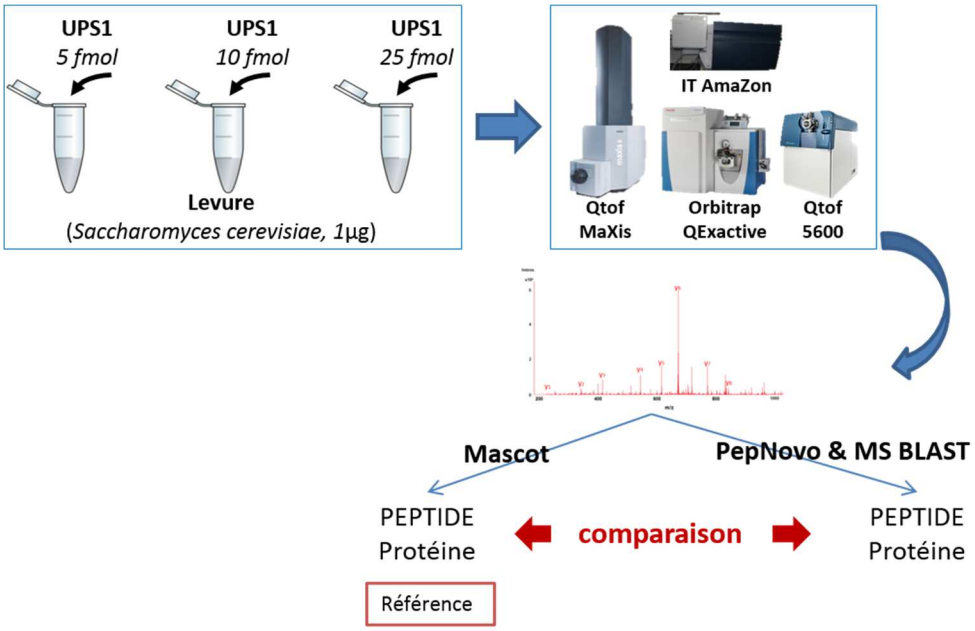


Figure II-4: Représentation schématique de la comparaison, spectre par spectre, de l'interprétation Mascot versus PepNovo & MS BLAST

Afin de déterminer les paramètres optimaux de PepNovo, et plus précisément quelle interprétation permet le mieux d'expliquer un spectre donné (i.e. donne la même solution que Mascot), à la sortie de PepNovo, nous avons soumis différentes séquences (suivant leur score ou le fait qu'elles soient modifiées) à MS BLAST (voir Figure II-5):

1. On soumet à MS BLAST **le** *PepNovo sequence tag* qui a le meilleur « PepNovo score », voir ① de la Figure II-5.
2. On soumet à MS BLAST **le** *PepNovo sequence tag* qui a le meilleur « Rank score », voir ② de la Figure II-5.
3. On soumet à MS BLAST **la** *PepNovo extended sequence for Blast* qui a le meilleur « PepNovo score », voir③ de la Figure II-5.
4. On soumet à MS BLAST les **7** *PepNovo sequence tag* qui ont les meilleurs « PepNovo score », voir ④ de la Figure II-5.
5. On soumet à MS BLAST les **7** *PepNovo sequence tag* qui ont les meilleurs « Rank score », voir⑤ de la Figure II-5.
6. On soumet à MS BLAST les **7** *PepNovo extended sequence for Blast* qui ont les meilleurs « PepNovo score », voir ⑥ de la Figure II-5.

Figure II-5 : Représentation schématique des différents tests effectués.

pnv = « PepNovo score » et rnk = « Rank score ». Ici les numéros attribués au score montrent le rang du peptide suivant le score choisi. Par défaut, les séquences sont classées selon le « PepNovo score ».

Puis, pour chacun de ces cas, nous avons cherché si une (et seulement une) des solutions données par MS BLAST était la même que celle proposée par Mascot (peptide et protéine), quel que soit son rang pour les cas où 7 séquences sont soumises.

En effet, on ne veut considérer qu'une séquence par spectre. Bien sûr, il est possible d'avoir des spectres chimères, issus de la fragmentation de plusieurs peptides, et ne conserver qu'une séquence par spectre reviendrait à perdre l'information du peptide co-fragmenté. Cependant, considérer deux séquences pourrait induire des erreurs : comment être sûr que les deux séquences dérivées correspondent bien à deux peptides co-fragmentés et non à une erreur d'interprétation car la lecture du spectre est difficile ? Ainsi, il est préférable de perdre de l'information que de risquer d'obtenir une information erronée.

- La première question que l'on peut se poser est le nombre de séquence à soumettre à MS BLAST. Autrement dit, la séquence qui a le meilleur score (que ce soit le « PepNovo score » ou le « Rank score ») est-elle forcément la meilleure ?

  La Figure II-6 montre le pourcentage de spectres « correctement assignés » (c'est-à-dire dont une des solutions peptide-protéine correspond à la solution Mascot) en noir ; le pourcentage de spectre « non assignés » (i.e. n'ayant pas donné de résultat par PepNovo et MS BLAST) en rouge, et enfin

le pourcentage de spectres « mal assignés » (i.e. dont aucune solution peptide-protéine ne correspond à la solution Mascot) en vert.

On observe que très peu de spectres sont mal assignés, quel que soit la solution choisie, ce qui montre que le séquençage *de novo* donne globalement des résultats justes.

Ensuite, on remarque que quel que soit la séquence (modifiée ou non) et la méthode de score (« PepNovo score » ou « Rank score ») choisie, envoyer 7 séquences à MS BLAST augmente les chances d'obtenir la bonne solution (qu'elle soit classée 1ère ou 7ème avec le score). Ce qui montre que les scores établis par PepNovo ne permettent pas, à eux seuls, de trouver la meilleure interprétation du spectre. En effet, ne soumettre que la meilleure séquence à MS BLAST (pnv 1, rnk 1, pnv ext sur la Figure II-6) entraîne plus de spectres mal assignés (en rouge).

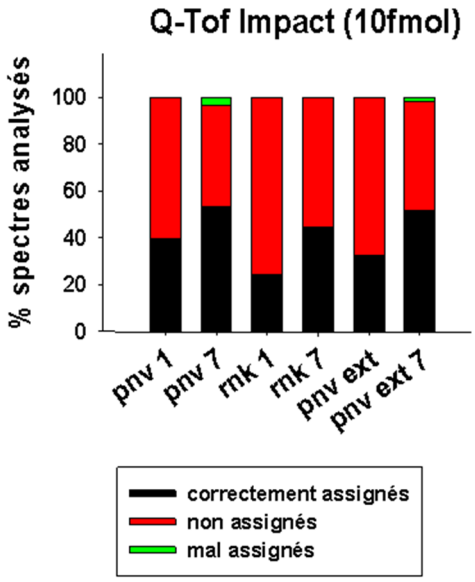


Figure II-6 : Pourcentage de spectres correctement assignés (noir), non assignés (rouge), mal assignés (vert) ; issus de l'analyse sur l'Impact HD (Bruker), à 10fmol d'UPS1.

pnv 1 (pnv 7) = on a soumis à MS Blast la (les 7) *PepNovo sequence tag* qui a (ont) le meilleur score « PepNovo score » ;
rnk 1 (rnk 7) = on a soumis à MS Blast la (les 7) *PepNovo sequence tag* qui a (ont) le meilleur score « Rank score » ;
pnv ext (pnv ext 7) = on a soumis à MS Blast la (les 7) *PepNovo extended sequence for Blast* qui a (ont) le meilleur score « PepNovo score ».

- Une deuxième question se pose alors : s'il est préférable de soumettre 7 tags de séquence à MS BLAST, comment « reconnaître » lequel des 7 est le bon ? Nous nous sommes alors intéressés au « MS Blast score ».

  Sur la Figure II-7, on observe que parmi les spectres correctement assignés (seul les noirs de la Figure II-6), presque tous (≈ 100%) ont été correctement assignés par le tag de séquence qui possède **le plus haut score MS BLAST**.

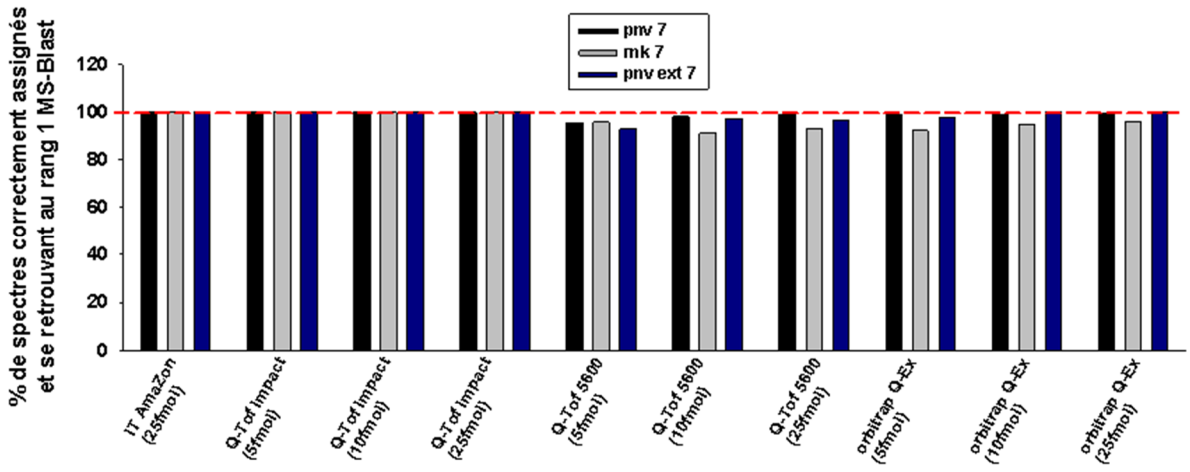


Figure II-7 : Pourcentage de spectres étant correctement assignés par la séquence possédant le plus haut score MS BLAST, parmi tous les spectres correctement assignés ; pour l'analyse sur chaque machine, aux différentes quantités d'UPS1 injectées.

pnv 7 = on a soumis à MS Blast les 7 *PepNovo sequence tag* qui ont le meilleur score « PepNovo score » ; rnk 7 = on a soumis à MS Blast les 7 *PepNovo sequence tag* qui ont le meilleur score « Rank score » ; pnv ext 7 = on a soumis à MS Blast les 7 *PepNovo extended sequence for Blast* qui ont le meilleur score « PepNovo score ».

Il est donc possible, et préférable, de soumettre 7 tags de séquence à MS BLAST puis de ne conserver, pour chaque spectre, que le résultat qui a le meilleur « MS BLAST score ».

- Enfin, la séquence modifiée pour MS BLAST est-elle vraiment la plus juste, et sur quel score se baser ?

  On voit, sur la Figure II-7 que lorsque l'on cherche la bonne solution parmi les 7 tags de séquence qui ont le meilleur « Rank score » (en gris sur la figure), la bonne solution est moins souvent trouvée avec le MS BLAST score le plus haut. Ainsi, si l'on analyse un mélange inconnu, il sera impossible de déterminer quelle est la bonne solution, si l'on ne peut pas se baser sur ce score. Il ne paraît donc pas judicieux de travailler sur les tags de séquence rangées selon le « Rank score ».

  En revanche, lorsque l'on cherche la bonne solution parmi les 7 tags de séquence qui ont le meilleur « PepNovo score », que ce soit la séquence non modifiée (*PepNovo sequence tag* en noir sur la figure), ou la séquence modifiée pour MS BLAST (*PepNovo Extended sequence for Blast* en bleu sur la figure), les résultats obtenus sont similaires. Cependant on voit sur la Figure II-6 qu'utiliser les tags de séquence non modifiées (*PepNovo sequence tag*) apporte plus de fausses identifications (i.e. de spectres mal assignés, en vert). Il est donc préférable d'utiliser la séquence modifiée. De plus, utiliser la séquence non modifiée entraînerait une modification du code source de PepNovo, ce qui ne semble pas nécessaire d'après nos résultats.

Finalement, nos tests ont permis de mettre en évidence que se baser uniquement sur le score donné par PepNovo ne suffit pas à obtenir le bon résultat. En effet, il est préférable de soumettre un plus grand nombre de tag de séquences à MS BLAST (jusqu'à 7) et de faire un deuxième tri à postériori, sur le résultat du BLAST. Nous avons également montré que la séquence modifiée pour MS BLAST présente un réel intérêt et permet de limiter les faux positifs. Ces différents paramètres ont donc été définis par défaut dans la version PepNovo qui est implémentée dans la suite MSDA.

Il aurait également été envisageable d'appliquer des seuils sur les scores (PepNovo ou MS-BLAST), plutôt que de « jouer » sur les rangs des interprétations. Bien que nous n'ayons pas appliqué de tels seuils, ce sont néanmoins les meilleurs scores MS-BLAST qui ont systématiquement rapportés les résultats « corrects ». Concernant PepNovo, il aurait effectivement pu être intéressant de jouer sur les seuils pour conserver des interprétations de meilleure qualité. En revanche, cela aurait permis de retenir un nombre x d'interprétations, variable selon les spectres. On aurait alors conservé 1, 3 ,7, 10… ou encore aucune interprétation par spectre. Dans la publication [174], les auteurs ont obtenu de bons résultats en considérant jusqu'à 7 interprétations par spectre. Nous avons donc décidé de nous baser sur cet article et fait ce choix « arbitraire » de considérer 7 interprétations par spectre.

Enfin, il aurait été intéressant d'évaluer les limites de la méthode sur les protéines minoritaires de la levure, afin de déterminer un seuil à partir duquel la stratégie PepNovo – MS BLAST n'est plus « efficace » pour identifier correctement un peptide. En revanche, pour déterminer ce seuil, il aurait été utile de connaître la concentration des protéines de levure, ce qui n'était pas le cas. Nous aurions éventuellement pu considérer des protéines de levure connues pour être peu ou très abondantes. Cela aurait donc pu être testé avec une gamme plus étendue de concentration des protéines UPS1.
De plus, pour réaliser un séquençage *de novo*, nous préférons travailler avec des spectres de haute qualité (dont les pics d'intérêt sont intenses et sortent bien du bruit de fond), d'où l'utilisation du module Recover pour éliminer les spectres de « mauvaise qualité ». On sait que les spectres de peptides de protéines très minoritaires sont moins « informatifs » (peu de pics intenses) et notre stringence sur Recover aurait probablement éliminé ces spectres. C'est pourquoi nous n'avons pas cherché à travailler sur les protéines de levures éventuellement minoritaires.

C.1.2.3. Validation des données et élimination des redondances

Nous nous sommes ensuite servis des résultats de ces optimisations pour traiter les données obtenues par l'analyse des tissus adipeux bruns de campagnol.
Le module Recover de la suite MSDA a été utilisé pour ne conserver que les spectres non assignés par Mascot et de très bonne qualité. Pour ce faire, on définit un nombre de pics utiles (UPN) dont l'intensité doit être supérieure à un seuil de bruit de fond défini comme un multiple (E) de l'intensité médiane de tous les pics du spectre. Ici, nous avons défini E=9 et UPN=9 afin d'être très stringents sur la qualité des spectres.
En moyenne, sur toutes les analyses, 23% des spectres étaient assignés par Mascot, et 60% des spectres ont ensuite été conservés pour le séquençage *de novo*. Compte-tenu de la stringence que nous avons paramétré dans le module Recover, ce fort pourcentage de spectres retenu atteste de la qualité des données spectrales obtenues. Selon les développements rapportés plus haut, 7 tag de séquences modifiées pour MS BLAST (*PepNovo extended sequence for blast*) par spectre ont été soumis à MS BLAST (pour comparaison aux séquences protéiques contenues dans la banque de données « Mammifères », comme pour la recherche Mascot) à l'issu du séquençage de novo, puis nous avons conservé le tag qui avait le meilleur score MS BLAST pour chaque spectre. Ensuite, nous avons validé les identifications obtenues par MS BLAST ; les tags de moins de 6 acides aminés ont été éliminés et la correspondance entre le tag de séquence et la séquence théorique devait être stricte pour au moins 5

résidus consécutif pour que le peptide correspondant soit conservé. Cela signifie que l'on autorisait une seule mutation pour 6 acides aminés, ce qui reste stringent pour éviter d'avoir de faux résultats, mais permet l'identification de peptides qui ne sont pas strictement conservés.

Une difficulté de notre stratégie *de novo* provenait une nouvelle fois de l'utilisation d'une banque de données comportant les séquences issues de plusieurs taxonomies (mammifères). En effet, un même tag de séquence sera potentiellement assigné à des peptides de la même protéine homologue chez plusieurs espèces du fait de la tolérance accordée aux identifications inter-espèces. Afin de s'affranchir de cette redondance, qui augmente considérablement le nombre de résultats obtenus, nous l'avons éliminé de la même façon que pour la recherche classique (BLAST fasta 36 de toutes les protéines identifiées contre la banque *Homo Sapiens*).

Les redondances inter-espèces ne sont pas les seules qui sont présentes dans les données. En effet, du fait de la multiplicité de résultats possibles à chaque étape (comme expliqué en page 81), l'analyse des résultats issus des recherches BLAST devient longue et fastidieuse. Les différents types de redondance et la façon dont nous les avons traitées sont présentés ci-après :

- Plusieurs spectres sont issus de la fragmentation du même peptide :
  - Ces spectres identifient le même *PepNovo sequence tag*:
    - MS BLAST identifie la même protéine → On élimine simplement la redondance (en conservant la séquence qui a le meilleur score).
    - MS BLAST identifie des protéines différentes → Il s'agit alors d'un peptide partagé, c'est un cas que l'on peut rencontrer également dans une recherche classique. On conserve l'information, ces peptides ne seront pas utilisés pour la quantification.

  - Ces spectres identifient différents *PepNovo sequence tag*. Nous émettons l'hypothèse qu'il s'agit du même peptide car ils ont le même temps de rétention ± 120sec et la même masse ± 15ppm :
    - MS BLAST identifie la même protéine → On élimine simplement la redondance (en conservant la séquence qui a le meilleur score).
    - MS BLAST identifie des protéines différentes → Il s'agit alors d'un peptide partagé. On conserve l'information, ces peptides ne seront pas utilisés pour la quantification.

- Plusieurs spectres sont issus de la fragmentation de différents peptides :
  - Ces spectres identifient différents *PepNovo sequence* tag mais le même *MS Blast query* ou *MS Blast subject* : On a alors des peptides *a priori* différents qui ont une correspondance par BLAST sur la même séquence donc sont identifiés comme identiques. Ces cas-là sont éliminés car ils sont trop ambigus.

C.1.3. Rassemblement des identifications Mascot et *de novo*

La dernière étape de l'identification est le rassemblement des données Mascot et *de novo*. Une solution aurait pu être d'utiliser l'information de l'emplacement des peptides identifiés dans la séquence de la protéine ; information disponible dans Mascot et également sur le résultat de l'alignement BLAST des peptides *de novo*. Le problème est que l'on identifie des protéines de différentes espèces, donc de tailles différentes, les positions ne correspondent donc pas (même pour des homologues), ce qui empêche leur utilisation. Pour une protéine donnée, l'idée serait donc plutôt d'aligner toutes les séquences de peptides identifiés avec les séquences de toutes les protéines homologues présentes dans les banques de données. Ceci permettrait de valider qu'aucun conflit de séquence entre le tag *de novo*, les séquences peptidiques Mascot, et les séquences des protéines homologues n'est présent dans nos données (voir [175] pour un exemple). Ceci permettrait de plus d'estimer plus justement la couverture des séquences protéiques que nous avons atteinte. Il faudrait donc développer un outil qui permette d'automatiser l'alignement massif d'un grand nombre de séquences, ce que nous n'avons pas encore réalisé.

En l'état actuel, l'imprécision qui subsiste réside donc dans la surestimation du nombre de peptides identifiés par protéine. En effet, il est possible qu'un peptide ait été identifié par les deux méthodes (par deux spectres différents) mais que le tag déterminé par séquençage *de novo* diffère sensiblement de la séquence Mascot. Ainsi, on comptabilisera deux fois le même peptide pour une protéine, mais ceci ne devrait pas avoir d'impact sur la qualité des données, ni qualitatives, ni quantitatives.

C.2. Stratégie de quantification

Pour la quantification, tous les peptides n'ont pas été considérés. Ont été éliminés :

- Les peptides partagés entre plusieurs protéines ;
- Les peptides contenant des méthionines ;
- Les peptides présents dans les blancs ;
- Les contaminants ;
- Les peptides « decoy » (les faux positifs).

La quantification « label-free » XIC MS1 des peptides a été réalisé avec le logiciel Skyline v3.1 (MacCoss Lab, UW).

C.2.1. Quantification des peptides identifiés par Mascot

Les courants d'ions ont été extraits pour les ions P, P+1 et P+2 de chaque peptide. L'intégration des aires sous la courbe a été vérifiée manuellement pour chaque peptide, dans chaque analyse. Cette étape était très chronophage mais nécessaire pour s'assurer de la qualité des données. En effet, le logiciel est prompt à des erreurs d'intégration ; de plus, certains peptides ont été éliminés car leur signal se trouvait dans le bruit de fond. Les aires sous la courbe des ions P, P+1 et P+2 ont ensuite été sommées pour chaque peptide.

C.2.2. Quantification des peptides issus du séquençage *de novo*

Tandis que la quantification des peptides « Mascot » était aisée et routinière, la quantification des peptides « *de novo* » à l'aide du logiciel Skyline a quant à elle nécessité des développements bio-informatiques.

Le tag de séquence est déduit de la fragmentation d'un peptide précurseur. Dans la majorité des cas, la détermination du tag n'est pas complète (voir explication page 30), donc la masse calculée du tag diffère de la masse expérimentale mesurée du précurseur. Comme expliqué en page 49, Skyline calcule la masse de la séquence qu'on lui soumet et extrait le courant d'ion correspondant à partir des données brutes. Si l'on soumettait une séquence incomplète (comme c'est le cas pour les tags *de novo*), le courant d'ion serait extrait à la mauvaise masse et on intègrerait un signal erroné. Un exemple est présenté en Figure II-8.
De plus, Skyline a besoin d'une librairie spectrale pour l'extraction des courants d'ions, qui est habituellement le résultat de la recherche Mascot (format « .dat ») ; librairie dont on ne dispose pas avec la stratégie *de novo*.

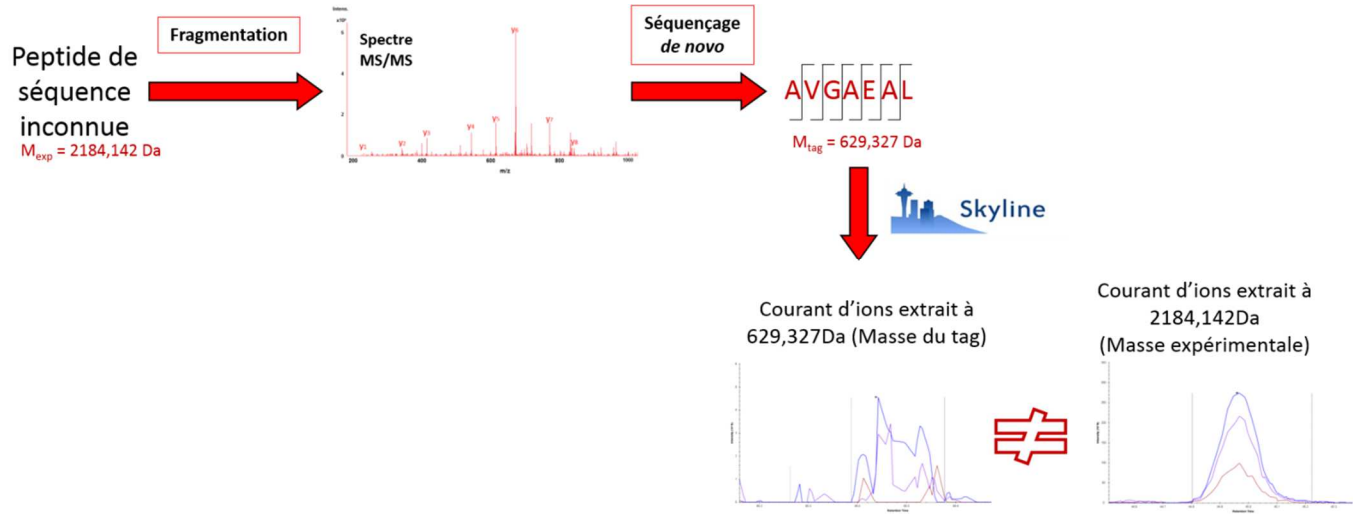


Figure II-8 : Schéma décrivant le problème pour quantifier les peptides issus du séquençage *de novo*.
Dans cet exemple, le tag (non complet) déduit par *de novo* a une masse de 629,327Da, Skyline extrait donc le courant d'ions à cette masse, qui est différente de la masse expérimentale du peptide précurseur (2184,142 Da).

N'ayant pas accès à la séquence complète du peptide, le seul moyen de soumettre à Skyline une séquence qui lui permettrait d'avoir accès au bon courant d'ions, était de reconstituer une « fausse » séquence, à partir de la masse du précurseur (information dont l'on dispose dans les données spectrales). Nous avons donc mis en place un calcul combinatoire (avec l'aide de bio-informaticiens), pour compléter les tags de séquence avec une combinaison d'acides aminés de sorte que la masse de cette fausse séquence soit égale à la masse du peptide précurseur (en tolérant une erreur de 15ppm). Un exemple est présenté en Figure II-9. Bien sûr, cette séquence a uniquement servi à leurrer le logiciel Skyline, elle n'a pas été considérée dans les traitements ultérieurs.

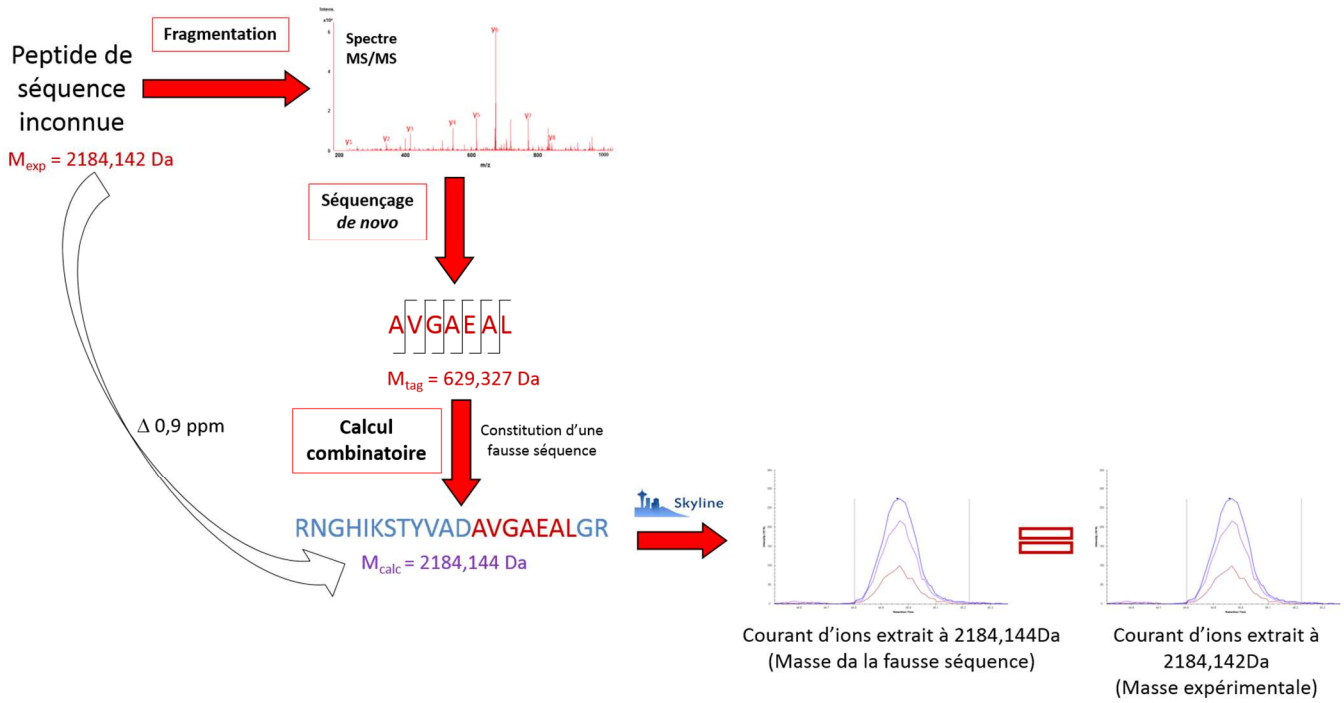


Figure II-9: Principe du calcul combinatoire mis en place pour la quantification des peptides « *de novo* ».
Dans cet exemple, le tag (incomplet) déduit par *de novo* a une masse de 629,327 Da, tandis que la masse expérimentale est de 2184,142 Da. Le calcul combinatoire a permis de déterminer une fausse séquence (RNGHIKSTYVADAVGAEALGR) dont la masse est très proche (à 0,9 ppm près) de la masse expérimentale ; permettant ainsi de leurrer Skyline et de quantifier le peptide.

Une fois cette liste de faux peptides établie, nous avons pu créer une librairie de spectres (avec l'aide de bio-informaticiens), contenant toutes les informations nécessaires à Skyline (« fausse séquence » du peptide, masse du précurseur, temps de rétention…).

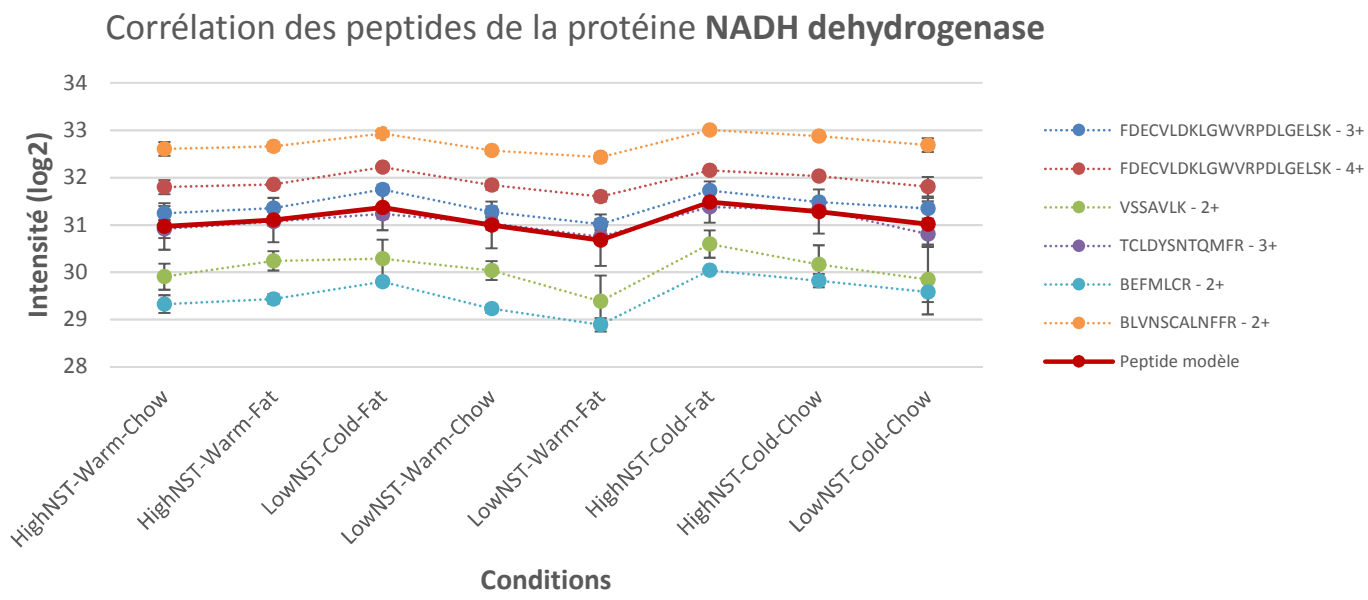
Ici aussi, une vérification manuelle de l'intégration des aires sous la courbe a été nécessaire. Et Les aires sous la courbe des ions P, P+1 et P+2 ont ensuite été sommées pour chaque peptide.

Grâce à ces faux peptides et à la librairie ainsi établie, nous avons donc pu « leurrer » le logiciel Skyline pour réaliser la quantification des peptides *de novo*.
Une autre solution pour quantifier ces peptides aurait été l'utilisation du logiciel MassChroQ [109], auquel il est possible de fournir des listes de masse (m/z) et de temps de rétention (RT) pour extraire les courants d'ions. Cela permet de s'affranchir du problème que présente Skyline. Une stratégie basée sur celle employée par MassChroQ aurait également pu être développée pour Skyline afin d'extraire le signal uniquement à partir de listes de m/z et RT. Enfin, le logiciel commercial PEAKS [176] permet également de traiter ces peptides.

C.3. Normalisation des données

A la sortie de Skyline, les données quantitatives des peptides Mascot et *de novo* ont été rassemblées et normalisées par la méthode des quantiles, à l'aide du package R (v3.0.3 ; http://www.R-project.org) nommé Normalyzer [116]. La méthode de normalisation utilisée impose une transformation logarithmique (log_2) des données avant normalisation proprement dite.

Après normalisation, nous avons vérifié la corrélation des profils d'abondance des peptides pour chaque protéine quantifiée. C'est-à-dire que nous avons voulu vérifier si tous les peptides d'une protéine variaient de la même façon entre les conditions. Pour cette évaluation, les 6 réplicas de chaque condition ont été regroupés (en moyennant leur valeur d'abondance) afin d'obtenir une seule valeur par condition. A l'aide d'une corrélation de Pearson [177], nous avons comparé le profil de chaque peptide à un peptide modèle. Ce peptide modèle a été créé en moyennant les valeurs d'abondance (somme des aires sous la courbe des ions P, P+1, P+2) de tous les peptides d'une protéine. L'exemple de la protéine NADH déhydrogénase (sp|P51970) est présenté en Figure II-10.



Peptides	Coefficient de corrélation/peptide modèle
FDECVLDKLGWVRPDLGELSK - 3+	0,96
FDECVLDKLGWVRPDLGELSK - 4+	0,96
VSSAVLK - 2+	0,94
TCLDYSNTQMFR - 3+	0,95
BEFMLCR - 2+	0,95
BLVNSCALNFFR - 2+	0,98

Figure II-10 : Evaluation de la corrélation des 6 peptides de la protéine NADH déshydrogénase.
L'abondance relative du peptide modèle est présentée en rouge. On voit sur le graphique une très bonne corrélation des 6 peptides, ce qui se retrouve dans les valeurs de coefficients de corrélation, supérieurs à 0,9.
La nomenclature utilisée pour les conditions est la même qu'en **Figure II-2**.

Nous avons ensuite conservé uniquement les peptides dont la corrélation était satisfaisante :

- Pour les protéines identifiées avec 1 seul peptide, celui-ci a été conservé (car on ne peut pas calculer de corrélation avec 1 seul peptide).

- Les protéines identifiées avec 2 peptides ont été conservées si les profils d'abondance des deux peptides étaient corrélés avec un coefficient de corrélation de Pearson supérieur à 0,7.

- Pour les protéines identifiées avec 3 peptides et plus: seuls les peptides qui avaient un coefficient de corrélation de Pearson supérieur à 0,6 ont été conservés.

Enfin, pour chaque échantillon, une moyenne des peptides retenus par protéine a été calculée, afin d'obtenir une valeur d'abondance relative par protéine.

Pour l'instant, nous n'avons pas éliminé les protéines quantifiées avec un seul peptide. Comme précisé précédemment (Partie IIChapitre II.C.4), ces protéines sont étudiées avec précaution.

D. Analyse des échantillons

Pour résumer ; après extraction des protéines, celles-ci ont été digérées en solution puis les peptides analysés sur un Q-Exactive+ (Thermo Fisher). Les données spectrales ont ensuite été interprétées selon la recherche classique par l'algorithme de recherche Mascot dans la banque de données Swiss-Prot *Mammalia* ; ainsi que par séquençage *de novo* grâce au logiciel PepNovo puis une recherche par similarité de séquence a été réalisée par l'algorithme MS BLAST contre la banque Swiss-Prot *Mammalia*. Enfin, la quantification a été réalisée avec le logiciel Skyline et l'intégration de tous les signaux a été manuellement vérifiée. Finalement, les données ont été normalisées par la méthode des quantiles et la corrélation des peptides a été vérifiée, comme expliqué dans le paragraphe précédent.

Le protocole de l'ensemble de ces étapes est détaillé dans la Partie Expérimentale, en page 155.

E. Résultats

Nous avons ainsi pu identifier un total de **3044** protéines uniques (voir Figure II-11). La stratégie *de novo* a permis d'identifier spécifiquement 144 protéines ; et pour près de 800 protéines (soit 25.5% des identifications), le séquençage *de novo* a permis d'augmenter la couverture de séquence.

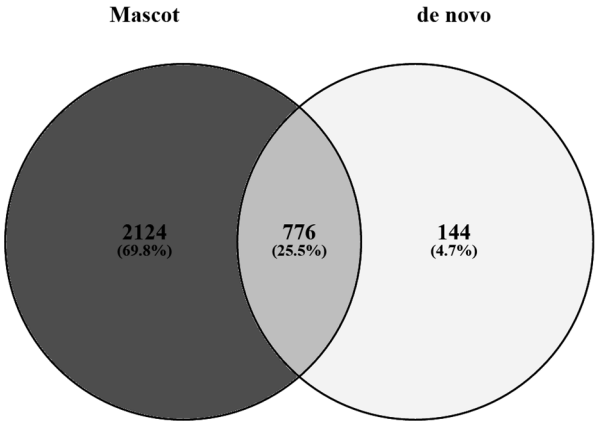


Figure II-11 : Diagramme de Venn représentant le recouvrement des protéines identifiées par Mascot et *de novo*.

Sur les 3044 protéines, 2518 ont été utilisées pour la quantification (voir page 90 pour les critères). Suite à la vérification manuelle sur Skyline, **937** protéines ont été quantifiées (dont 826 par mascot et 412 par *de novo*) avec **5135** peptides (dont 2962 par mascot et 2173 par *de novo*). La forte perte de protéines est due à la stringence de la vérification manuelle. De nombreux peptides ont été éliminés

car leur signal se trouvait dans le bruit de fond. Cette forte stringence, bien que coûteuse en termes de protéines quantifiées, nous permet de garder uniquement des données indiscutables.

A l'issue de l'étape de filtre (corrélation de Pearson), seulement 49 protéines et 1289 peptides ont été éliminés soit 5% et 25 % respectivement. Ce qui montre que ce tri était nécessaire pour obtenir des données plus robustes, mais que la majorité des peptides étaient tout de même bien corrélés et très peu de protéines ont été perdues.

Finalement, **888** protéines ont donc été quantifiées (voir Figure II-12), avec **3846** peptides (2608 issus de mascot et 1238 de *de novo*). Nous observons, sur la Figure II-12, que la stratégie *de novo* permet de quantifier près de 12% de protéines en plus, et que pour 28,5% des protéines déjà quantifiées par la stratégie classique, le *de novo* permet de gagner en robustesse (voir Figure II-13).

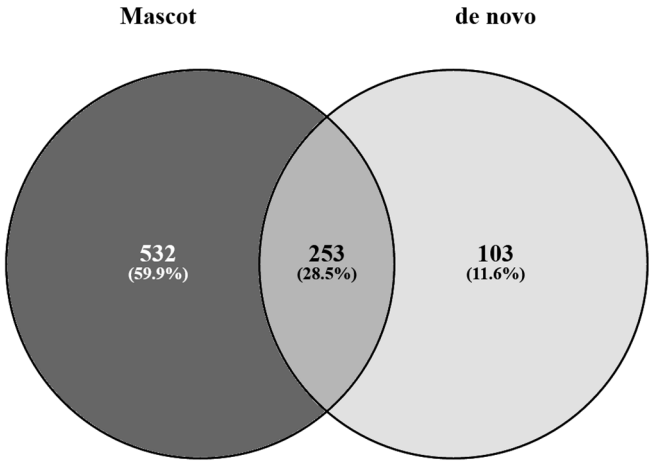


Figure II-12 : Diagramme de Venn représentant le recouvrement des protéines quantifiées avec des peptides issus de Mascot et *de novo*.

Sur la Figure II-13, nous nous sommes intéressés à l'origine des peptides qui ont été assignés à chacune des 253 protéines en commun (Figure II-12). Pour chaque protéine, est représentée la somme des peptides quantifiés, en gris issu de l'identification par mascot et en rouge (au-dessus) issus du séquençage *de novo*. Le séquençage *de novo* permet d'augmenter la couverture de séquence de 40% en moyenne (de 6 à 95%) pour ces 253 protéines, pour lesquelles la quantification est donc plus robuste. Cela montre le bénéfice d'avoir utilisé la stratégie *de novo* et tout l'intérêt des développements mises au point au cours de ces travaux.

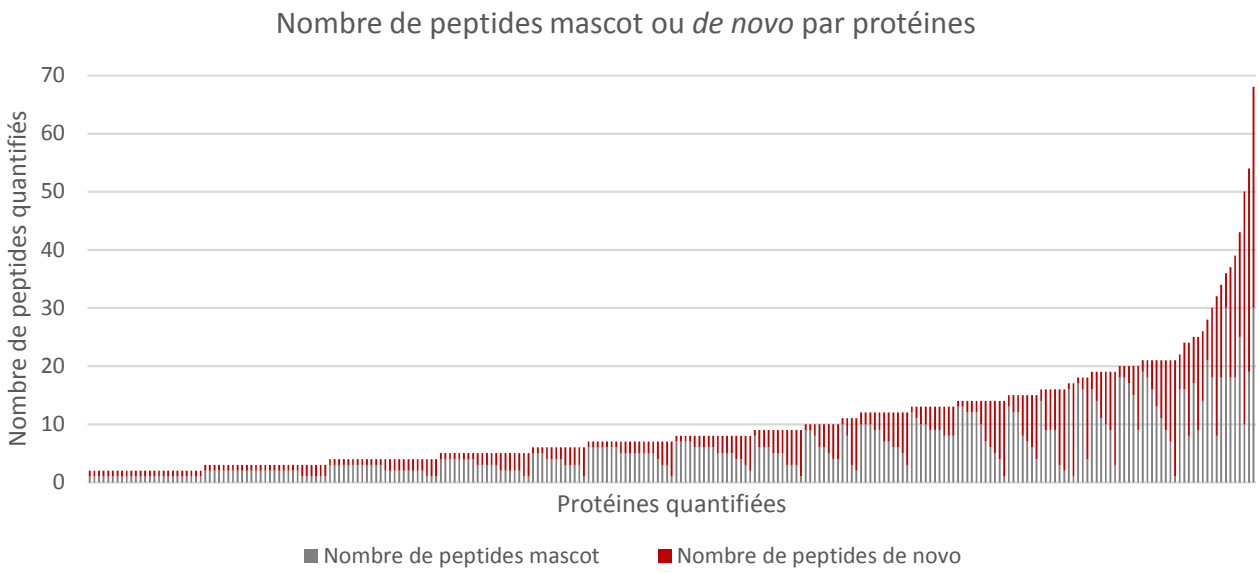


Figure II-13 : Répartition des peptides des 253 protéines quantifiées grâce aux deux stratégies.
Chaque barre de l'histogramme représente une protéine. Pour chaque protéine, le nombre de peptides issus de mascot est représenté en gris, et le nombre de peptides issus du séquençage *de novo* en rouge.

F. Suivi de la stabilité instrumentale

Au cours des 8 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 5 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit ; Biognosys AG, Schlieren, Switzerland) en quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure II-14 et Figure II-15.

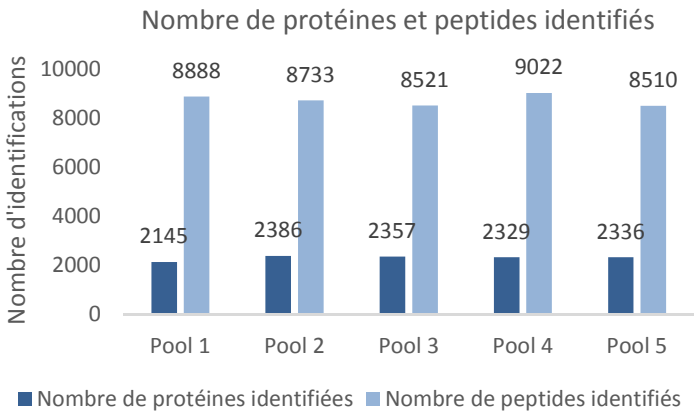


Figure II-14 : Nombre protéines et peptides identifiés dans les 5 injections répétées de l'échantillon de référence (ou Pool).

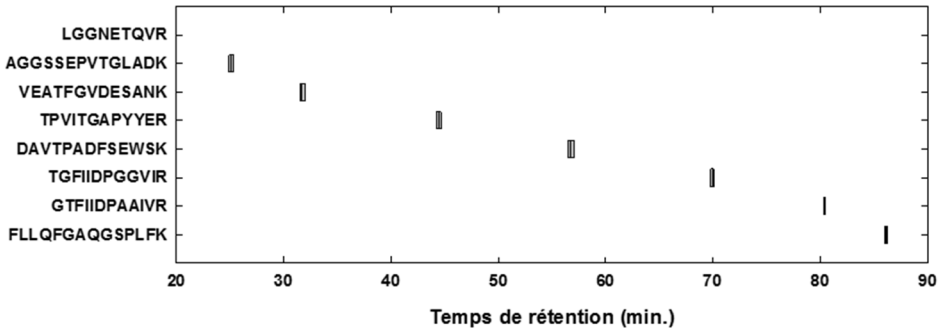


Figure II-15 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate, que le nombre de protéines et peptides identifiés est resté stable au cours des 8 jours d'analyse, avec un CV de 4% et 3% respectivement, ce qui montre une bonne reproductibilité du système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 11,6%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation de temps de rétention des peptides iRT étaient en moyenne égaux à 1%.

G. Interprétation des résultats

Le « package limma » (modèles linéaires) utilisé dans l'environnement R a permis de déterminer les protéines dont l'abondance varie significativement, dans le cadre d'une analyse multifactorielle prenant en compte les effets dus à la lignée des campagnols (lineage), la température environnementale, et le régime alimentaire (diet). Une correction selon la méthode de Benjamini et Hochberg a été utilisée pour corriger les valeurs de P pour les comparaisons multiples. Ainsi, les protéines différentiellement exprimées entre les groupes ont été mises en évidence (p-value < 0.05). La Figure II-16 montre le nombre de protéines différentielles en fonction du traitement.

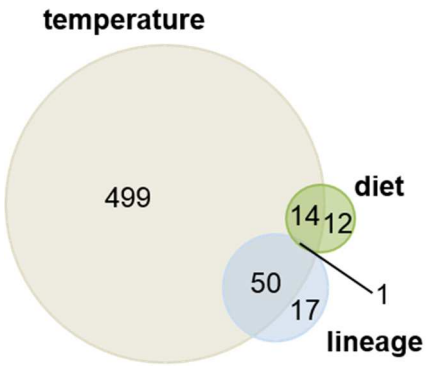


Figure II-16 : Recouvrement des protéines différentiellement exprimées selon les trois conditions température, régime alimentaire (« diet ») et lignée (« lineage »).
Les différences majeures sont dues à la température.

Les annotations fonctionnelles ont été extraites de la base de données Gene Ontology (termes GO) pour toutes ces protéines différentielles. Puis nous avons réalisé des calculs d'enrichissement fonctionnel (grâce aux algorithmes de DAVID (Ease v2.0) bioinformatics [178, 179]), les termes enrichis ont été filtrés pour ne considérer que ceux qui avaient un score Ease inférieur à 0.1, une p-value

Benjamini < 0.05 et un facteur d'enrichissement > 2. Les termes GO significativement enrichis ont ensuite été regroupé en « fonctions » significativement altérées par chacune des conditions (Figure II-17, A, B et C) ou leurs interactions (Figure II-17 D et E).

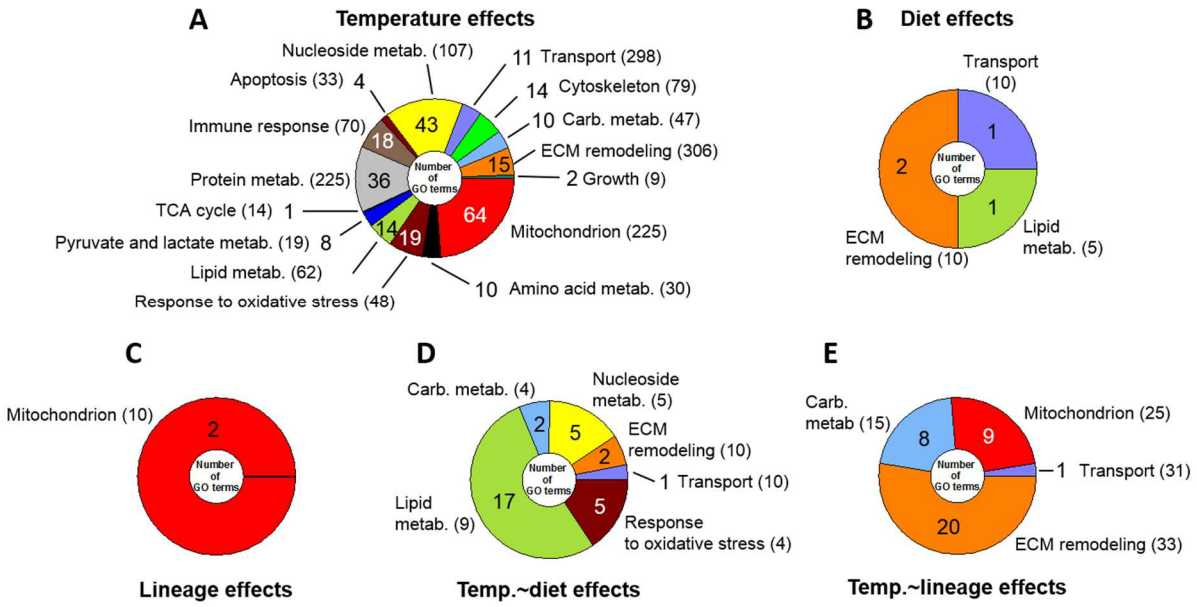


Figure II-17 : Répartition des grandes fonctions regroupant les termes GO significativement enrichis par la température (A), le régime alimentaire (B), la lignée (C), l'interaction entre température et régime (D) et entre température et lignée (E).

Les nombres de termes GO par grande fonction est indiqué dans les diagrammes, et le nombre de protéines impliquées dans chaque fonction est indiqué entre parenthèse.

Les différences entre les deux lignées (faible ou forte thermogénèse sans frisson) étaient dues uniquement à des protéines mitochondriales (Figure II-17 C), comme on s'y attendait.
On retrouve principalement des effets dues à l'exposition des campagnols à une faible température ambiante, affectant en grande partie le protéome mitochondrial (Figure II-17 A). De plus, les changements métaboliques observés sont en accord avec une activation du BAT en réponse au froid.
L'interaction entre température et régime alimentaire (Figure II-17 D) ou température et lignée (Figure II-17 E), et à moindre mesure les effets du régime seulement (Figure II-17 B), révèlent un programme de traduction spécifique qui serait en accord avec des changements métaboliques et un remodelage tissulaire. Le métabolisme du BAT serait orienté vers une oxydation accrue des substrats lipidiques quand un régime gras est associé à de faibles températures. Dans ce cadre, le stress oxydatif pourrait constituer une conséquence coûteuse de l'activation du BAT.

Sur les Figure II-18 et Figure II-19 , on peut observer les abondances relatives, dans les différentes conditions, de certaines protéines différentiellement exprimées en fonction du régime, de la lignée ou de la température. Les valeurs d'abondance relative de ces protéines sont présentée dans le Tableau Annexe 3, page 189. Pour l'instant, des analyses « globales » ont été réalisées, en considérant les effets d'un facteur donné (régime, lignée ou température) sur toutes les conditions. On observe ainsi une augmentation globale de l'abondance des protéines de transport du cholestérol, de dégradation et de synthèse des acides gras chez les animaux soumis à un régime hyperlipidique (Figure II-19 « Diet

Effect »). Ces résultats ne sont pas surprenant et reflètent l'induction globale du métabolisme des substrats lipidiques dont la disponibilité est fortement augmentée par le régime enrichi en graisses. On peut cependant remarquer que le protéome du tissu adipeux brun des campagnols varie finalement assez peu en réponse au régime hyperlipidique. On constate également une diminution globale de l'abondance des protéines mitochondriales (sauf pour celles qui sont impliquées dans le métabolisme des acides aminés) chez les animaux à forte thermogénèse (Figure II-19 « Lineage Effect »). Or l'activité thermogène du tissu adipeux brun de ces animaux étant plus forte à l'état basal, on aurait pu s'attendre à des différences plus marquées. Ceci pourrait refléter le fait que les niveaux d'activité du BAT chez les 2 lignées de campagnol sont davantage dépendants de l'activité de certaines protéines que de leur abondance. Enfin, on observe une augmentation globale de l'abondance des protéines mitochondriales chez les animaux exposés au froid (Figure II-18 « Temperature Effect »). Ce résultat est tout à fait cohérent avec la fonction thermogène du BAT, avec des variations reflétant l'induction globale du métabolisme des substrats glucidiques et lipidiques, et de la fonction respiratoire mitochondriale. De même, les niveaux de l'UCP1, protéine principale contrôlant la production de chaleur était fortement augmentés par l'exposition au froid. En réponse au froid, toute la « cascade métabolique » est donc activée, depuis l'oxydation des substrats énergétiques jusqu'à la production de chaleur.

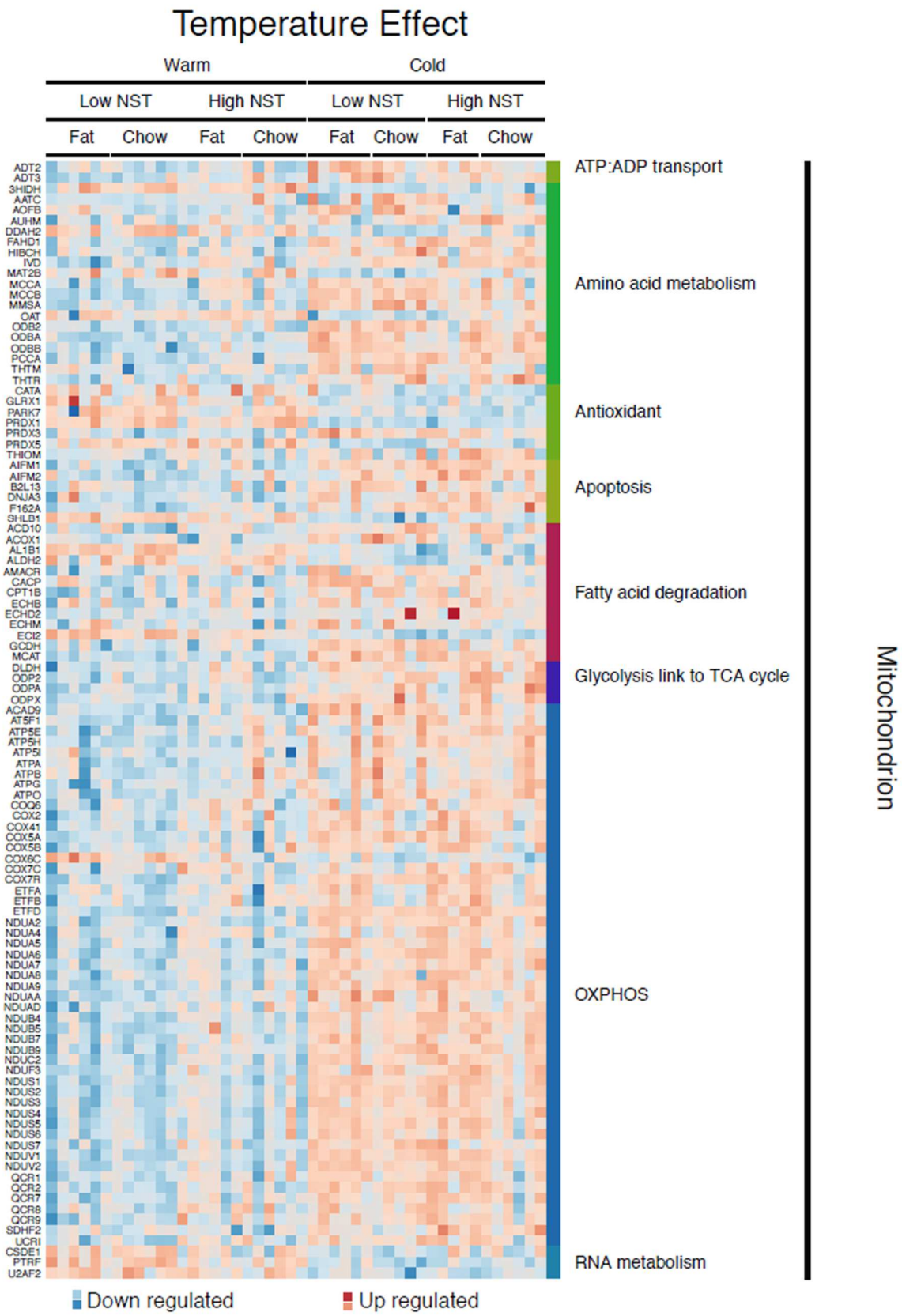


Figure II-18 : Carte (« *Heatmap* ») représentant les abondances, pour chaque condition, des protéines mitochondriales significativement altérées par la température.

En rouge, les protéines sont sur exprimées ; en bleu elles sont sous-exprimées. « OXPHOS » = oxidative phosphorylation.

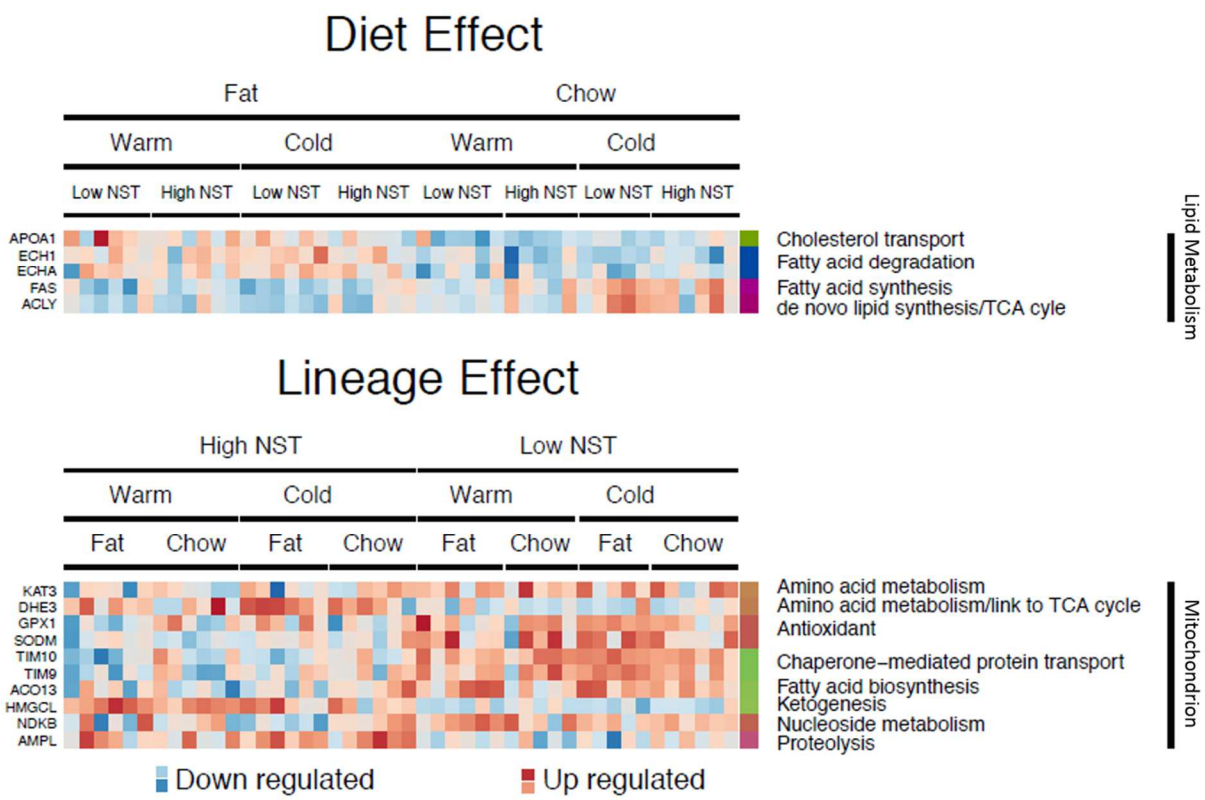


Figure II-19 : Carte (« *Heatmap* ») représentant les abondances, pour chaque condition, des protéines significativement altérées par le régime alimentaire (« Diet ») (métabolisme des lipides) et par la lignée (« Lineage ») (mitochondriales).

En rouge, les protéines sont sur exprimées ; en bleu elles sont sous-exprimées.

Pour conclure, le protéome du BAT des campagnols est fortement affecté par l'exposition au froid, mais de manière bien moindre par le régime gras. Ces données vont donc permettre maintenant de comparer dans le détail la réponse au froid en fonction de la lignée. Nos collaborateurs ont de plus prévu des analyses comparant le protéome du BAT des campagnols indépendamment de la température pour définir dans quelle mesure le niveau d'activité du BAT influe sur la réponse au régime gras. En effet, des analyses préliminaires avaient montré que l'adiposité des animaux à forte thermogenèse n'est pas altérée en réponse au régime hyperlipidique contrairement à celle des animaux à faible thermogenèse qui était augmentée.

Ces régulations sont actuellement examinées en détail par nos collaborateurs. A l'issue de ces examens, un article sera préparé pour publier ces résultats.

II. Détermination de l'apport du séquençage *de novo* vs. celui d'un préfractionnement protéique

A. Etude de la modification du protéome chez une espèce saisonnière

A.1. Contexte biologique

L'obésité et ses conséquences sont des problèmes majeurs dans nos sociétés modernes [180]. De grandes avancées ont été réalisées dans la compréhension des mécanismes impliqués dans l'apparition de tels désordres métaboliques, en particulier au niveau des réseaux neuronaux impliqués dans le contrôle de la faim. Pourtant, il n'existe pas encore de thérapie efficace aux effets durables. Seuls une hygiène alimentaire irréprochable et un niveau suffisant d'activité physique peuvent permettre de prévenir/limiter le développement d'une obésité.

Dans ce contexte, le microcèbe (*Microcebus Murinus*, taxonomie 30608) est un modèle de choix puisqu'il pourrait s'apparenter à un modèle « d'obésité réversible ». En effet, le microcèbe est un lémurien (primate prosimien) dont le cycle de vie, en conditions naturelles, est rythmé par la variabilité saisonnière en ressources alimentaires. Afin de survivre à des périodes de forte disette alimentaire, le microcèbe engraisse fortement au début de l'hiver, au cours d'une première phase, qualifiée d'obésogène. Par l'utilisation extensive d'un mécanisme d'hypométabolisme appelé torpeur, l'énergie emmagasinée sera économisée au maximum. Au milieu de l'hiver, on observe une inversion spontanée de l'engraissement vers une perte de poids, associée à une diminution de la prise alimentaire et une augmentation de l'activité métabolique (voir Figure II-20-A). Cette phase d'amaigrissement est caractérisée par le développement d'une intolérance au glucose (voir Figure II-20-B) associée à des taux d'insuline basale élevés, suggérant l'établissement d'un état insulino-résistant [181] (voir Figure II-20-C). Les phases obésogène et diabétogène s'enchainent donc spontanément chez le microcèbe, sans jamais atteindre le seuil pathologique, ce qui est paradoxal vis-à-vis des modèles classiques d'obésité, notamment humains.

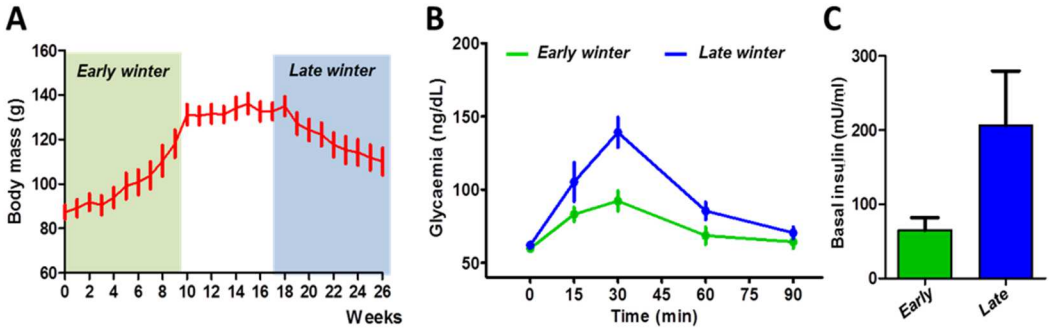


Figure II-20 : Caractérisation (réalisée par notre collaborateur) des phases dites obésogène (en début d'hiver, « early winter » en vert), et diabétogène (en fin d'hiver, « late winter » en bleu) du microcèbe.

A) Représentation de l'évolution de la masse corporelle (body mass) des individus au cours de l'hiver. En début d'hiver, les microcèbes prennent du poids (phase dite obésogène), et ne commencent à le perdre que sur la fin de l'hiver.

B) Glycémie des individus en début et fin d'hiver. En fin d'hiver, les individus régulent moins bien et moins vite leur glycémie, ce qui suggère une intolérance au glucose.

C) Taux d'insuline basal des individus en début et fin d'hiver. En fin d'hiver, le niveau d'insuline est beaucoup plus élevé, ce qui suggère une insulino-résistance (phase dite diabétogène).

Les régulations du phénotype saisonnier des microcèbes impliquent différents tissus métaboliquement actifs, comme le tissu adipeux brun qui va régler la dépense énergétique, le tissu adipeux blanc qui va stocker ou libérer les substrats énergétiques lipidiques, ou encore le foie qui est un carrefour métabolique essentiel. L'objectif de ce projet était de produire des données originales pour une meilleure compréhension des mécanismes qui sous-tendent la capacité du microcèbe à inverser spontanément le processus d'engraissement. Dans ce but, une analyse différentielle du protéome hépatique de microcèbes en phase obésogène *vs*. diabétogène a été réalisée par une méthode de quantification globale relative sans marquage.

Du point de vue analytique, les verrous à lever résidaient surtout dans l'optimisation de la préparation des échantillons et le choix de la stratégie de traitement des données spectrales générées par l'analyse d'échantillons collectés chez un organisme dont le génome n'est pas séquencé.

Ce projet a été réalisé en collaboration avec le docteur Jérémy Terrien de la Team BIOADAPT (UMR CNRS/MNHN 7179), Paris.

Schéma expérimental :

Deux groupes de microcèbe ont été comparés : un groupe en phase diabétogène et l'autre en phase obésogène. Une biopsie de foie a été réalisée sur chaque individu, et chaque groupe comportait 5 réplicas. Nous avions donc un total de 10 échantillons.

A.2. Contexte analytique et objectif

Le génome du microcèbe n'étant pas séquencé ; une première étape était donc de déterminer la banque de données la plus appropriée dans laquelle effectuer la recherche.

Ensuite, pour les mêmes raisons, il pourrait être judicieux de réaliser un séquençage *de novo*, en complément d'une recherche classique. De plus, étant donné le faible nombre d'échantillons à analyser (10), il est tout à fait envisageable de réaliser un préfractionnement, afin d'augmenter la couverture des protéomes étudiés ; et ce sans augmenter considérablement le temps d'analyse.
En revanche, il n'est pas possible avec les logiciels envisagés de combiner les deux approches, en tous cas pour l'étape de quantification. En effet, si l'on préfractionne les protéines, notre logiciel de prédilection est Maxquant qui permet de normaliser de telles données quantitatives (comme expliqué en page 49). Cependant, la quantification des peptides issus du séquençage *de novo* ne peut pas être réalisée avec ce logiciel. Pour ce cas, la stratégie que nous avons mise en place (Partie IIChapitre III.C.2.2) utilise le logiciel Skyline avec lequel la normalisation des données quantitatives issues d'échantillons fractionnés n'est pas géré. Nous étions donc face à une inadaptation logicielle entre séquençage *de novo* et préfractionnement protéique, pour la quantification. Cela met en avant le fait que toutes les possibilités de chaque étape sont à réfléchir dès le début du projet, car toutes les étapes sont intimement liées, et la façon dont on prépare les échantillons va influencer la façon dont on analysera les données spectrales.

L'un des objectifs était donc de répondre à la question suivante : est-il plus bénéfique de décomplexifier un échantillon en mettant de côté le séquençage de novo, ou au contraire de favoriser

le séquençage *de novo* à partir d'échantillons non préfractionnés ? Un deuxième objectif était de mettre en place une méthode de « label-free » XIC MS1 (dont le principe est expliqué en page 48) pour comparer les protéomes des individus en phase obésogène et diabétogène.

A.3. Développements méthodologiques

A.3.1. Recherche d'une banque de données adaptée à l'étude d'une espèce non séquencée

Il existe une banque de données du *Microcebus murinus* (taxonomie 30608) dans UniProt, mais cette banque étant très peu renseignée (703 séquences, dont 688 dans TrEMBL), il n'est pas envisageable de l'utiliser. Ainsi, il est nécessaire de trouver une espèce phylogénétiquement proche du microcèbe. Sachant que le microcèbe est un primate, nous avons tout d'abord considéré l'utilisation d'une banque de données contenant les séquences protéiques de tous les primates (taxonomie 9443, *Primates*, 26 928 séquences, Swiss-Prot). Or nous avons constaté qu'elle était constituée majoritairement (à 75%) de séquences de protéines humaines (taxonomie 9606, *Homo Sapiens*, 20 195 séquences Swiss-Prot). Nous avons donc fait une recherche dans chacune de ces banques. Nous avons identifié **925** protéines dans la banque *Primates* (en éliminant la redondance inter-espèce, comme expliqué en page 80) et **956** dans la banque *Homo Sapiens* ; soit un gain de 3% avec cette dernière. On identifie plus de protéines avec cette banque, malgré le fait qu'elle contienne moins de protéines. Il peut y avoir plusieurs explications à cela : la banque *Primates* regroupe les séquences protéiques de plusieurs espèces, donc beaucoup de protéines homologues. Si on les rassemblait comme on a rassemblé les protéines *identifiées*, le nombre d'entrées dans la banque se rapprocherait de celui de la banque *Homo Sapiens*, ce qui explique la faible différence entre les nombre de protéines identifiées dans les deux banques. Le fait d'en identifier tout de même plus chez *Homo Sapiens* peut venir du fait que la banque de recherche est plus petite, ce qui limite le nombre de faux-positifs (ou *decoy*), et donc réduit le FDR et ainsi augmente le nombre de vrai positifs.

Par ailleurs, le logiciel MaxQuant a été choisi pour normaliser les données quantitatives issues d'un préfractionnement des protéines, et donc son moteur de recherche Andromeda pour l'identification. Avec Maxquant, il est préférable d'utiliser une banque qui contient les séquences protéiques d'**une** seule espèce. En effet, MaxQuant effectuant directement un regroupement des peptides identifiés par protéine, sur lequel il n'est pas possible d'agir, nous ne pourrions pas regrouper les identifications et quantifications de protéines homologues chez diverses espèces de primates.

Finalement, la banque de données contenant les séquences protéiques des humains est donc la plus adaptée ici.

A.3.2. Séquençage *de novo* ou préfractionnement

Afin de tester l'apport du séquençage *de novo* par rapport à un préfractionnement protéique, des tests ont été réalisés sur **un** échantillon de foie de microcèbe, dont les protocoles sont détaillés dans la Partie Expérimentale (en page 157) et rapidement résumés ici.

Protocole 1 (sans préfractionnement) : Après broyage du tissu et extraction des protéines, celles-ci ont été déposées sur gel SDS-PAGE et migrées jusqu'à l'entrée du gel de séparation (donc les protéines n'ont pas été séparées). Après analyse par spectrométrie de masse et recherche classique dans la

banque de données *Homo Sapiens*, et un séquençage *de novo* (suivi d'un BLAST dans la même banque), nous avons pu identifier **1128** protéines (dont 24% par séquençage *de novo*).

Protocole 2 (avec préfractionnement) : Après broyage du tissu et extraction des protéines, celles-ci ont été déposées sur gel SDS-PAGE et migrées sur 12mm. Ont ensuite été découpées 6 bandes de 2mm. Le contenu protéique de chaque bande a été analysé par spectrométrie de masse puis une recherche classique dans la banque de données *Homo Sapiens* a permis d'identifier **1479** protéines, dans l'ensemble des 6 bandes.

Nous observons donc que, malgré l'apport indéniable du séquençage *de novo*, décomplexifier le mélange protéique est ici plus bénéfique en termes d'identification. De ce fait, nous utiliserons Maxquant pour la quantification car ce logiciel permet de normaliser les données quantitatives issues d'un préfractionnement ; et donc son moteur de recherche implémenté (Andromeda) pour l'identification.

A.4. Analyse des échantillons

Après extraction des protéines, nous avons donc réalisé un préfractionnement des protéines sur gel SDS-PAGE. Les données spectrales acquises sur un Q-Exactive + (Thermo Fisher) ont ensuite été interprétées par une recherche classique, dans la banque de données Swiss-Prot *Homo Sapiens* grâce à l'algorithme de recherche Andromeda ; puis la quantification et normalisation des données ont été réalisées par MaxQuant.

Le protocole de l'ensemble de ces étapes est détaillé dans la Partie Expérimentale, en page 157.

A.5. Résultats

Nous avons finalement identifié **4100** protéines sur l'ensemble des échantillons. Parmi celles-ci, **3611** ont été considérées pour la quantification. En effet, les protéines « decoy » et les contaminants ont été éliminés, ainsi que les protéines pour lesquels il y avait plus d'une valeur manquante par groupe.

Comme expliqué précédemment (Partie IIChapitre II.C.4), ce sont les valeurs d'abondance protéique (« LFQ intensity ») fournies par Maxquant qui ont été directement utilisées.

Pour l'instant, nous n'avons pas éliminé les protéines quantifiées avec un seul peptide. Comme précisé précédemment (Partie IIChapitre II.C.4), ces protéines sont étudiées avec précaution.

A.6. Suivi de la stabilité instrumentale

Au cours des 6 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 3 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit ; Biognosys AG, Schlieren, Switzerland) en quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure II-21 et Figure II-22.

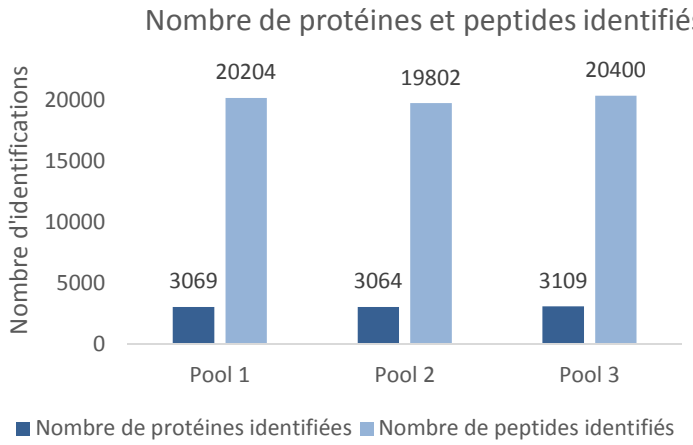


Figure II-21 : Nombre protéines et peptides identifiés dans les 3 injections répétées de l'échantillon de référence (ou Pool).

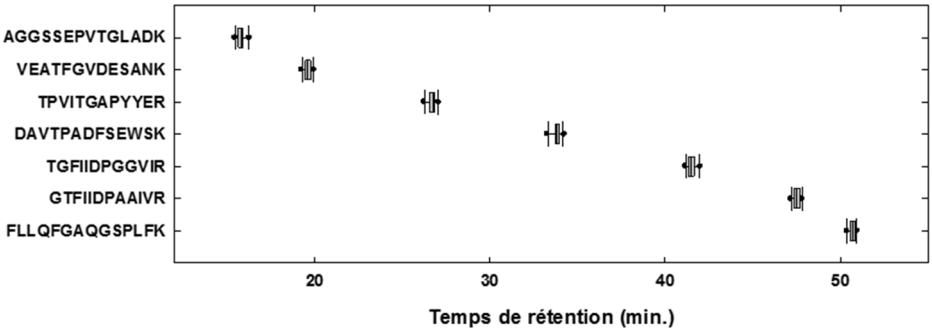


Figure II-22 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate que le nombre de protéines et peptides identifiés est resté stable au cours des 6 jours d'analyse, avec un CV de 2% et 1% respectivement, ce qui montre une bonne reproductibilité du système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 10,5%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation des temps de rétention des peptides iRT étaient en moyenne égaux à 0,8%.

A.7. Interprétation des résultats

Un test statistique t-test a été réalisé pour comparer les deux groupes d'individus (phase obésogène vs. phase diabétogène). Ce test a mis en évidence 373 protéines dont l'abondance variait significativement (p-value < 0,05) entre les deux phases. Les annotations fonctionnelles ont été extraites de la base de données Gene Ontology (termes GO) pour toutes ces protéines. Puis nous avons réalisé des calculs d'enrichissement fonctionnel (grâce aux algorithmes de DAVID (Ease v2.0) bioinformatics [178, 179]), les termes enrichis ont été filtrés pour ne considérer que ceux qui avaient un score Ease inférieur à 0.1, une p-value Benjamini < 0.05 et un facteur d'enrichissement > 2. Les 167 termes GO significativement enrichis ont ensuite été manuellement classés en 14 catégories fonctionnelles présentées en Figure II-23.

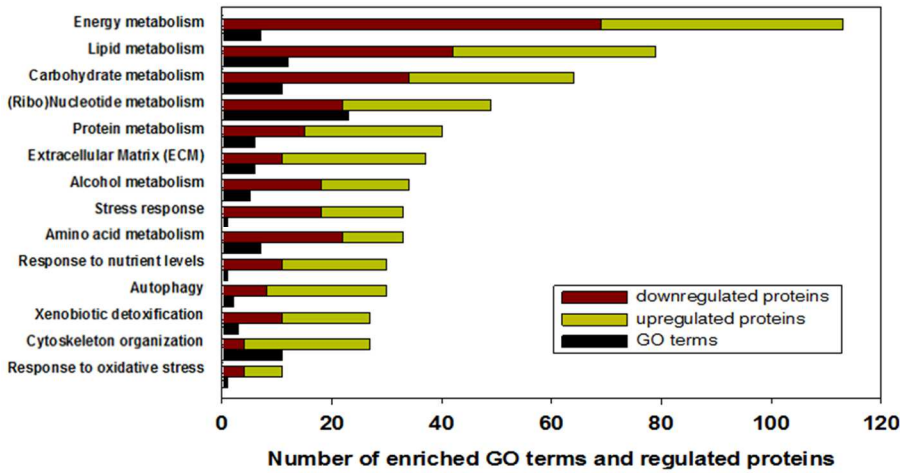


Figure II-23 : Catégories fonctionnelles significativement enrichies.
En vert sont représentées les protéines surexprimées (« downregulated » et en rouge les protéines sous-exprimées (« upregulated ») dans la phase diabétogène par rapport à la phase obésogène. En noir sont représentés les termes GO impliqués dans chaque catégorie.

Les principales différences entre les deux phases impliquent des protéines des métabolismes énergétique, lipidique et glucidique. Afin d'aller plus loin dans l'interprétation, nous avons reporté automatiquement (développement réalisé avec des bioinformaticiens) les régulations de ces protéines sur des cartes KEGG (Kyoto Encyclopedia of Genes and Genomes [182]); qui montrent les protéines impliquées dans un processus et leurs liens. Un exemple pour les voies de la glycolyse et néoglucogenèse est présenté en Figure II-24. Les valeurs d'abondance de ces protéines sont présentées en Tableau Annexe 4, page 193.

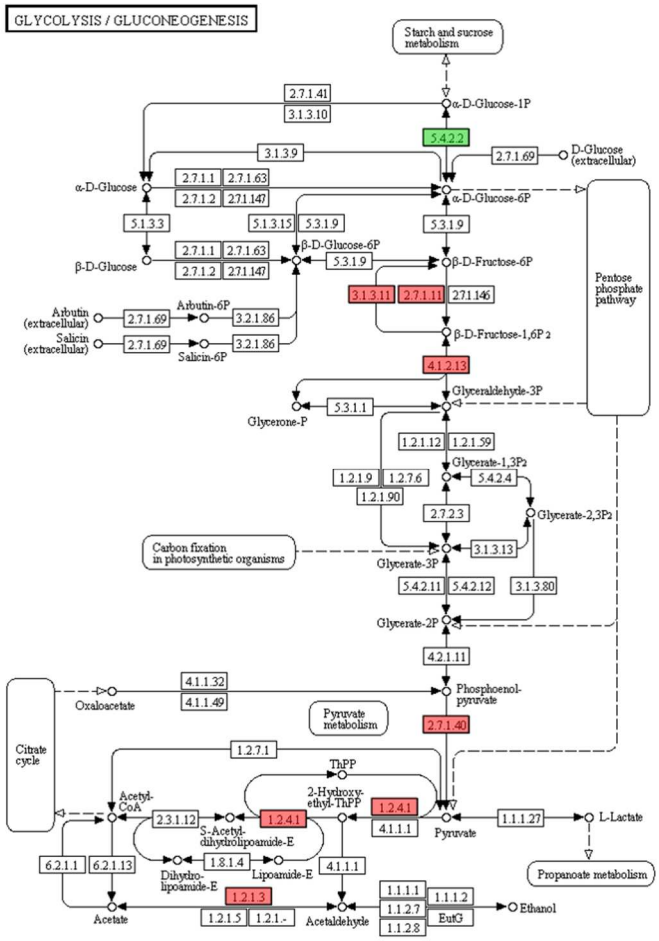


Figure II-24 : Carte KEGG représentant la variation des protéines différentiellement exprimées entre les deux phases et impliquées dans la glycolyse et la glyconéogenèse.

Les protéines que nous considérons sont colorées ; les cases non colorées correspondent aux autres protéines impliquées dans ce processus mais non identifiés par l'analyse protéomique. En vert sont représentées les protéines surexprimées et en rouge les protéines sous-exprimées dans la phase diabétogène par rapport à la phase obésogène.

Nos données suggèrent donc une diminution globale du métabolisme des glucides en phase diabétogène, ce qui serait en accord avec une intolérance au glucose. Les cartes KEGG établies pour les autres métabolismes susmentionnés suggèrent également une augmentation globale de l'entrée des acides gras lors de la phase diabétogène, ce qui serait en accord avec une stimulation du métabolisme des lipides, et donc une mobilisation des réserves accumulées dans la première partie de l'hiver, ce qui se confirme par la perte de masse corporelle durant la deuxième partie de l'hiver. Enfin, les variations de protéines impliquées dans la chaîne respiratoire suggèrent une diminution globale de son fonctionnement, ce qui concorderait avec une baisse de la prise alimentaire pendant cette phase diabétogène, et donc une baisse de la consommation d'énergie.

Ces interprétations préliminaires sont en cours d'étude par notre collaborateur. A l'issue de ces examens, un article sera préparé pour publier ces résultats.

B. Etude des variations du protéome de la fourmi *Lasius niger* en fonction de la caste

B.1. Contexte biologique

La vieillesse et la mort ont toujours été au centre des préoccupations humaines. Cependant, le questionnement à leur sujet évolue avec l'avancée des connaissances. Ainsi, on définit aujourd'hui le vieillissement comme l'évolution des processus biologiques dans le temps et la sénescence se limite aux processus entraînant une diminution de la fitness (ou aptitude phénotypique), autrement dit la capacité à transmettre son patrimoine génétique [183]. La sénescence constitue une véritable énigme évolutive car elle subsiste malgré l'effet de la sélection naturelle. Une manière de trouver une réponse à cette question est d'explorer pourquoi des individus placés dans un environnement semblable montrent parfois une importante diversité de longévités ?

Le statut social – place de l'individu dans son réseau social – fait partie des facteurs individuels impactant, entre autres, la longévité. Les facteurs sociaux ayant co-évolués avec une plus grande longévité sont divers : par exemple chez l'Homme l'appartenance à une classe aisée [184] et la quantité ainsi que la qualité des relations sociales au cours de la vie [185] ont un effet positif sur la durée de vie. A l'inverse, un haut rang hiérarchique peut aussi être source de stress [186, 187] et un environnement social stressant peut induire des modifications de mécanismes cellulaires associés au vieillissement comme la réduction de la longueur des télomères ou l'augmentation des dommages oxydatifs chez l'Homme [188].
Les relations entre l'âge et le vieillissement sont donc multiples, dépendantes de l'espèce étudiées et bilatérales. Mais il existe également un lien indéniable entre la socialité et le vieillissement.

Ces éléments décrivent l'influence mutuelle à explorer entre la socialité et le vieillissement et soulèvent de nombreuses questions. Comment expliquer la sélection d'individus à courte et longue durée de vie au sein de la même espèce ? Les effets de la sénescence sont-ils liés plus à l'âge ou au rôle social ? Et quels mécanismes moléculaires permettent cette variabilité en partant d'un patrimoine génétique quasiment identique ? Pour tenter de répondre à ces interrogations, nous nous sommes intéressés à l'influence de la caste sur le protéome de la fourmi noire des jardins (*Lasius niger*). Cette approche repose sur une comparaison entre castes différentes, permettant de tester les effets sociaux sur le profil protéique de *L. niger*. C'est donc une approche exploratoire qui permettra de déterminer la nature des mécanismes cellulaires et/ou physiologiques distinguant les individus en lien avec le rôle social et le vieillissement.

Nous avons étudié trois castes de *Lasius Niger* :
- Les ouvrières fourrageuses, qui passent la majeure partie de leur temps en dehors du nid et sont responsables de l'approvisionnement de l'ensemble de la colonie[189].
- Les ouvrières domestiques, qui prennent soin du couvain (ensemble des œufs) et passent donc la plupart du temps au sein du nid.
- Les reines qui s'occupent de la reproduction.

L'utilisation de *Lasius niger* dans le cadre de ce projet est pertinente pour plusieurs raisons. Tout d'abord, les colonies utilisées sont monogynes à fécondation unique (une seule reine fécondée par un seul mâle). Il existe donc un fort degré de parenté des individus formant la colonie, ce qui permet, pour un environnement similaire, de réduire la variabilité génétique dans l'explication des différences de longévité entre individus. En outre, *L. niger* pousse à l'extrême l'écart de longévité entre les ouvrières (2 à 4 ans) et la reine dont la longévité atteint souvent plus de 20 ans. Il semble prometteur de s'attarder sur les mécanismes qui permettent cette différence pour comprendre comment la sénescence est retardée à ce point chez les reines (ou accélérée chez les ouvrières). L'intérêt du modèle vient aussi du fait qu'il contredit le consensus de compromis entre reproduction et maintenance de l'organisme. En effet, les reines qui sont les individus reproducteurs sont aussi les plus longévives [190].

Ce projet a été réalisé en collaboration avec les Dr F. Criscuolo et C.sueur IPHC-DEPE, Strasbourg.

Schéma expérimental :

Le schéma expérimental de la sélection des fourmis est présenté en Figure II-25. Pour les reines, un échantillon a été constitué de 3 individus, tandis que pour les ouvrières (fourrageuses et domestiques), il a fallu 10 individus. Pour chaque caste, nous disposions de 5 réplicas, sauf pour les domestiques, où un échantillon a été perdu. Nous avions donc 14 réplicas. Pour ces analyses, l'intégralité de l'individu a été analysée.

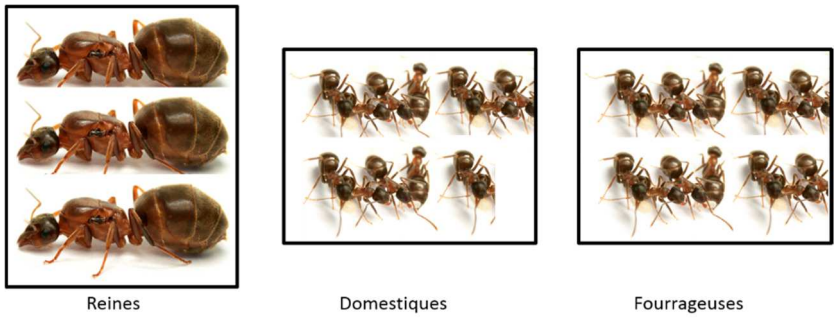


Figure II-25 : Schéma expérimental de la sélection des fourmis.

Ici, nous nous attendons à observer des différences liées à la fonction sociale des individus. Des études chez l'abeille européenne, qui visait à comparer différentes castes [191, 192] ont mis en avant des différences entre castes liées au métabolisme des glucides et acides gras, de la défense contre le stress oxydatif, de la production d'énergie ou encore de la division cellulaire. Nous pouvons donc nous attendre à retrouver les mêmes voies modifiées chez la fourmi. On s'attend surtout à voir des différences marquées entre reines et ouvrières, cependant la différence de rôle sociale entre domestiques et fourrageuses devrait également induire des différences.

B.2. Contexte analytique et objectifs

Les objectifs de ce projet étaient les mêmes que pour l'étude des microcèbes. Il a en effet été nécessaire de déterminer une banque de données appropriée ; de choisir entre préfractionnement et

séquençage *de novo*, et de mettre en place une méthode de quantification « label-free » XIC MS1 pour comparer les trois castes de fourmis.

B.3. Développements méthodologiques

Le cas de la fourmi est différent du microcèbe. En effet, il existe une banque de données de *Lasius Niger* (taxonomie 67767) dans TrEMBL, qui contient 18 075 séquences protéiques et pourrait donc être utilisée ici. Cependant, la banque de données TrEMBL contient des séquences annotées automatiquement et non vérifiées, il faut donc être vigilant quant à l'utilisation de cette banque.

Afin de savoir si la cette banque est tout de même la plus adaptée, nous avons effectué une recherche classique dans plusieurs banques : celle de *Lasius Niger* (TrEMBL, 18 075 séquences) à partir de laquelle nous avons identifié **951** protéines ; celle de tous les insectes, limitée à Swiss-Prot (taxonomie 50557 *Inscecta*, 8947 séquences) : **374** protéines identifiées ; et celle de tous les protostomiens, également limitée à Swiss-Prot (taxonomie 33317 *Protostomia*, 18 628 séquences) : **491** protéines identifiées.
On identifie donc moins de protéines avec les deux banques non restreintes à la fourmi. Il faut cependant noter que la banque *Insecta* n'est pas très fournie, ce qui pourrait expliquer le faible nombre d'identifications. Mais lorsque l'on s'intéresse à l'infra-règne, beaucoup plus large, que constitue les *Protostomia*, là encore le recouvrement est très faible. Il paraît donc évident que la banque *Lasius Niger* est la plus appropriée pour nos recherches. En effet, bien que les annotations ne soient pas vérifiées, cette banque de données reste tout de même directement dérivée du génome de l'espèce étudiée.

Même en utilisant une banque de données protéiques de l'espèce étudiée, le séquençage de novo aurait pu être bénéfique, par exemple dans le cas de l'occurrence de variants de séquences. Or le gain aurait probablement été très mineur ici. De ce fait, comme montré dans le projet précédent, nous avons décidé de bénéficier du préfractionnement des protéines, puis de les identifier par la seule recherche classique, dans la banque de données de *Lasius Niger*, et enfin de les quantifier par MaxQuant.

Enfin, la banque TrEMBL étant annotée automatiquement, il arrive que des protéines soient non annotées, et donc nommées « Uncharacterized protein » (pour « protéine non caractérisée »). Ainsi, nous n'avons pas accès à la fonction de ces protéines. Or, cette information reste essentielle pour interpréter les données et leur donner un sens du point de vue de la biologie, et bien sûr pour répondre à la question posée. Nous avons donc réalisé un BLAST de type fasta 36 (comme expliqué en page 80) contre la banque *Protostomia* (limitée à Swiss-Prot) pour les protéines non annotées. Cela nous a permis, pour chacune de ces protéines, d'identifier un homologue chez d'autres protostomiens, dans une banque bien annotée (*Protostomia* taxonomie 33317, Swiss-Prot, 18 628 séquences) et ainsi « transférer » cette annotation.

B.4. Analyse des échantillons

Après extraction des protéines, nous avons donc réalisé un préfractionnement sur gel SDS-PAGE. Les données spectrales acquises sur un Q-Exactive + (Thermo Fisher) ont ensuite été interprétées par une

recherche classique, dans la banque de données UniProt *Lasius Niger* grâce à l'algorithme de recherche Andromeda ; puis la quantification et normalisation des données ont été réalisés par MaxQuant.

Le protocole de l'ensemble de ces étapes est détaillé dans la Partie Expérimentale, en page 161.

B.5. Résultats

Comme dans tous les projets de ce type, afin de vérifier l'homogénéité des profils protéiques des échantillons, les protéines ont été séparées dans un premier temps sur un gel d'électrophorèse monodimensionnel (voir Figure II-26). Les profils obtenus n'étant pas totalement reproductible dans ce projet, nous montrons ce gel.

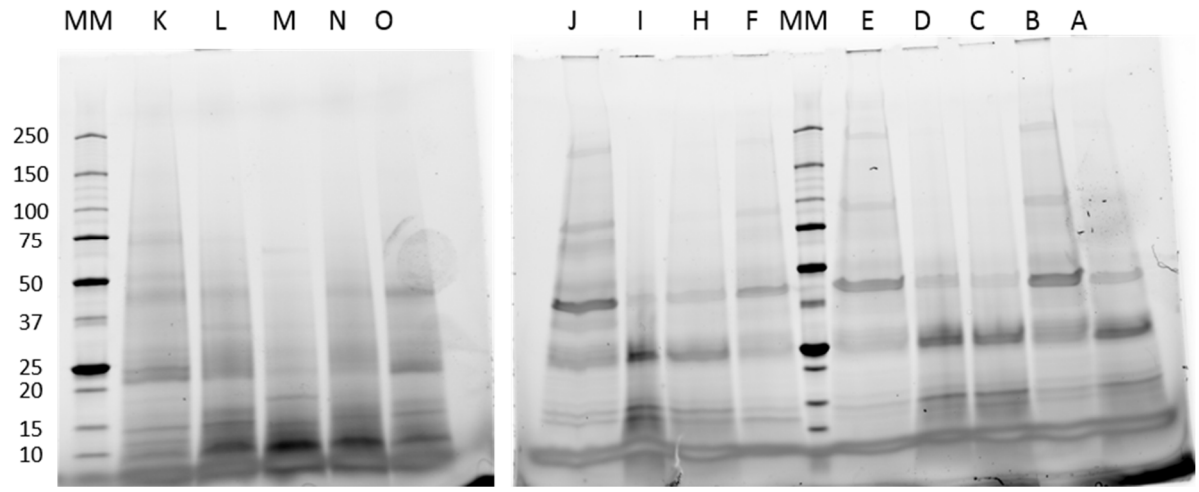


Figure II-26 : Gels d'électrophorèse de tous les échantillons.
MM = Marqueur Moléculaire (en kDa), le même marqueur a été utilisé pour les deux gels ; Echantillons K, L, M, N et O : Reines ; Echantillons F, H, I et J : Domestiques ; Echantillons A, B, C, D, E : Fourrageuses.

On remarque une forte disparité des profils électrophorétiques entre les différents échantillons, et même entre les réplicas d'un même groupe. On peut voir que ce sont vraiment les profils qui diffèrent, et que ça ne vient pas d'une différence de quantité de protéines chargées. En effet, par exemple pour les fourrageuses, la bande protéique la plus intense des échantillons B et E se trouve entre 37 et 50 kDa, bande moins intense chez A, C et D, échantillons pour lesquels la bande la plus intense se trouve à 25 kDa, moins intense pour B et surtout E. S'il y avait un problème de quantité chargée, toutes les bandes de la piste E seraient moins intenses que toutes les bandes de la piste D. Cela pourrait être dû au fait que chaque échantillon provenait de colonies différentes. Il est donc possible que la variabilité génétique entre colonies soit trop importante. Une explication à ces observations est proposée dans le paragraphe « Interprétations des résultats », page 114.

Nous avons finalement identifié **2764** protéines sur l'ensemble des échantillons. Parmi celles-ci, **1380** ont été considérées pour la quantification. En effet, les protéines « decoy » et les contaminants ont été éliminés, ainsi que les protéines pour lesquels il y avait plus de deux valeurs manquantes par groupe (sauf pour le groupe à n=4, pour lequel nous avons considéré un maximum d'une valeur manquante).

Comme expliqué précédemment (Partie IIChapitre II.C.4Partie IIChapitre IIII.A.5), ce sont les valeurs d'abondance protéique (« LFQ intensity ») fournies par Maxquant qui ont été directement utilisées.

Pour l'instant, nous n'avons pas éliminé les protéines quantifiées avec un seul peptide. Comme précisé précédemment (Partie IIChapitre II.C.4), ces protéines sont étudiées avec précaution.

Un test statistique (ANOVA unidirectionnelle + test posthoc de Tukey avec correction de Bonferroni pour comparaisons multiples ; $p<0.05$) a mis en évidence 643 protéines différentiellement exprimées entre les groupes. Ainsi, malgré les différences intra-groupes discutées plus haut, les différences intergroupes semblent bien plus importantes, permettant ainsi à ces 643 protéines d'avoir une significativité statistique. Cependant, pour prendre en compte les effets de la caste et de la colonie d'origine, la significativité a été évaluée sur les coordonnées de l'ACP ; soit par un test de Kruskal-Wallis (nb de modalités > 2, hétéroscédasticité), soit par des tests de la somme des rangs de Wilcoxon (2 modalités, homoscédasticité). Les tests de Kruskal-Wallis significatifs étaient suivis d'un test post-hoc de Conover-Iman avec correction de Bonferroni. L'égalité des variances a été testée par un test de Bartlett.

B.6. Suivi de la stabilité instrumentale

Au cours des 7 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 4 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit ; Biognosys AG, Schlieren, Switzerland) en quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure II-27 et Figure II-28.

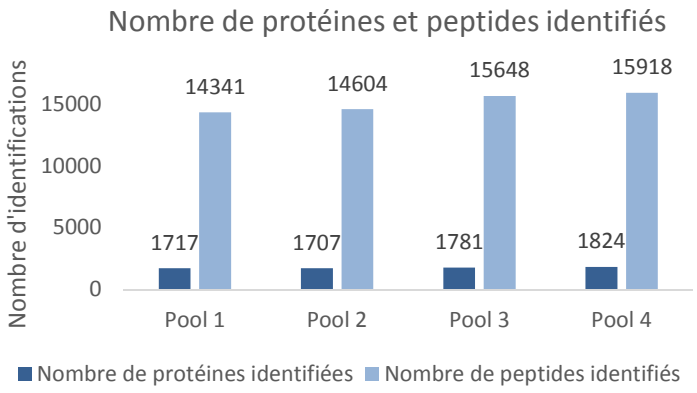


Figure II-27 : Nombre protéines et peptides identifiés dans les 4 injections répétées de l'échantillon de référence (ou Pool).

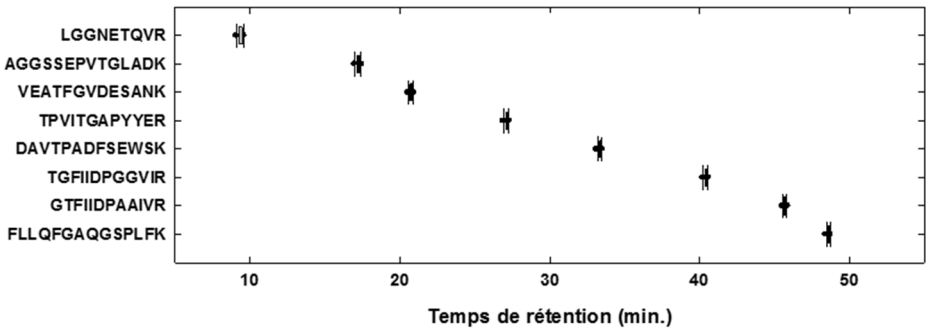


Figure II-28 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate que le nombre de protéines et peptides identifiés est resté stable au cours des 7 jours d'analyse, avec un CV de 5% et 3% respectivement, ce qui montre une bonne reproductibilité du système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 18,5%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation de temps de rétention des peptides iRT étaient en moyenne égaux à 0,6%.

B.7. Interprétation des résultats

Une analyse en composante principale (ACP) a été réalisée sur l'ensemble des protéines quantifiées (voir Figure II-29) et a permis de mettre en évidence un axe qui sépare les castes. L'axe 1, qui explique 62% de la variance, permet de séparer significativement les reines des ouvrières. L'axe 2 (expliquant 20% de la variance), quant à lui, ne sépare pas significativement les castes. Les valeurs d'abondance des protéines contribuant le plus à l'axe 1 (cos^2 projection > 0.9) sont présentées dans le Tableau Annexe 5, page 194.

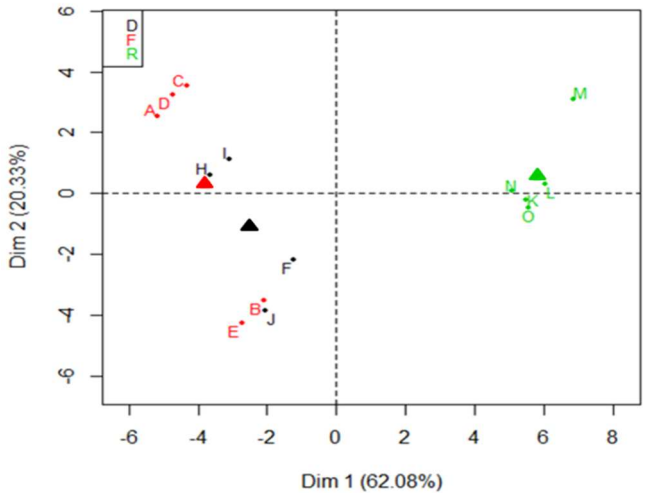


Figure II-29 : Résultat de l'analyse en composante principale réalisée sur les trois castes (domestiques en noir, fourrageuses en rouge et reines en vert).

Individus K, L, M, N et O : Reines (en vert) ; Individus F, H, I et J : Domestiques (en noir) ; Individus A, B, C, D, E : Fourrageuses (en rouge). Les coordonnées moyenne de chaque caste sont représentées par un triangle (dans leur couleur respective).

L'axe 1 est construit à partir de protéines impliquées dans la maintenance de l'organisme (processus de réparation, entretien, cycle cellulaire). Ce résultat suggère que la longévité de la reine est due à un investissement des ressources énergétiques pour empêcher les dommages aux macromolécules de s'accumuler. Cette allocation énergétique conduit à une longévité plus grande mais se fait au détriment d'autres fonctions, ici nos résultats suggèrent que les reines investiraient par exemple moins dans les protéines de résistance aux pathogènes. En effet, la reine est très peu exposée aux pathogènes (elle reste au sein du nid), c'est ce qui s'appelle « l'immunité sociale », contrairement aux ouvrières (qui sont en contact avec le milieu extérieur), dont le système immunitaire dirigé contre les pathogènes semble surexprimé.
Les reines et ouvrières sont également séparées par le métabolisme des lipides, surexprimé chez les reines ; ce qui pourrait être dû à la construction des membranes des œufs en formation (les reines étant les seuls individus reproducteurs). Dans ce sens, la dynamique cellulaire, plus importante chez les reines, peut également être expliquée par le développement embryonnaire des œufs.
Les ouvrières présentent des métabolismes (excepté celui des lipides) plus actifs que les reines. Leur système nerveux sensitif semble également surexprimé, ce qui serait en accord avec un besoin accru de ces fonctions chez les ouvrières pour détecter les phéromones déposées par la reine ou celles laissées entre fourrageuses pour indiquer une source alimentaire.
Les ouvrières concentrent majoritairement les protéines liées à la voie de signalisation ToR, qui est connue pour être stimulée par la présence de nutriments et réduire la longévité par de multiples biais [193]. Cela n'est pas étonnant, dans la mesure où les ouvrières prennent davantage en charge la nourriture que les reines (les fourrageuses approvisionnent la colonie et les domestiques nourrissent les larves). La séparation des tâches conduirait donc ici à la stimulation d'une voie diminuant la longévité chez les ouvrières, tandis que l'investissement des reines dans la maintenance serait associé à une plus grande longévité.

La première ACP ne permettant pas de d'observer des différences marquées entre les ouvrières, une seconde une deuxième ACP a été réalisée, uniquement avec les ouvrières (domestiques et fourrageuses), voir Figure II-30. L'axe 1 ne permet pas de séparer significativement les castes, en revanche l'axe 2 (qui explique 10% de la variance), permet de séparer significativement les domestiques et les fourrageuses.

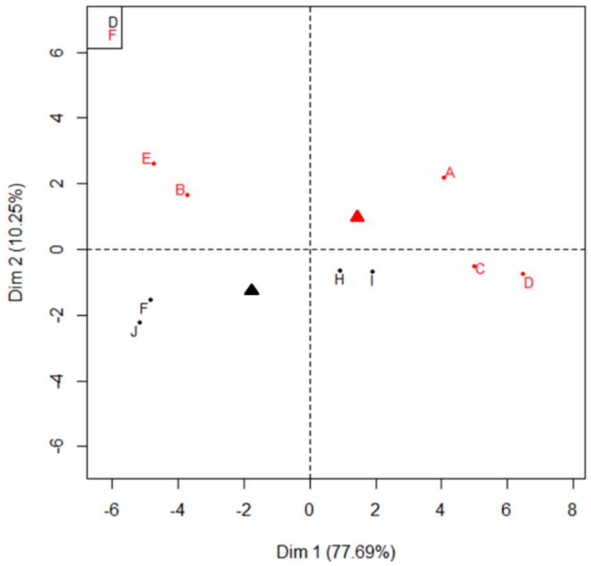


Figure II-30 : Résultat de l'analyse en composante principale réalisée sur les deux castes d'ouvrières (domestiques en noir, et fourrageuses en rouge).

Individus F, H, I et J : Domestiques (en noir) ; Individus A, B, C, D, E : Fourrageuses (en rouge). Les coordonnées moyennes de chaque caste sont représentées par un triangle (dans leur couleur respective).

Cette ACP montre que les fourrageuses se caractérisent par plus de protéines impliquées dans la défense immunitaire. Ceci suggère un gradient immunitaire dicté par l'organisation sociale : des fourrageuses (qui sortent le plus) avec beaucoup de protéines immunitaires jusqu'aux reines qui n'en n'ont que très peu, voire pas du tout (comme précisé plus haut).
Les domestiques se caractérisent par des protéines impliquées dans la digestion (principalement de l'amidon), qui pourrait être dû à un phénomène de prédigestion pour les larves.

Enfin, Sur les deux ACP réalisées, on note que l'axe qui ne sépare pas les individus selon la caste (axe 2 de la Figure II-29 ; axe 1 de la Figure II-30), sépare systématiquement deux groupes : les pools B, E, F, J d'un côté et A, C, D, H, I de l'autre ; ce qui est cohérent avec les différences observées sur les gels d'électrophorèse (Figure II-26). Il semble que, grâce à l'ACP en Figure II-30**,** cette variation interindividuelle non attribuable à la caste puisse provenir de la **plasticité phénotypique** des ouvrières. En effet, selon les besoins de la colonie, la proportion en fourrageuses et en domestiques est variable [194], et le changement de caste de certains individus peut avoir lieu, indépendamment de leur âge. L'état sénescent au sein des castes ne serait donc pas homogène et pourrait expliquer l'hétérogénéité des profils protéiques au sein d'une même caste.
De plus, *Lasius Niger* réalise généralement des diapauses en hiver (diminution du métabolisme basal) et l'axe 1 de la Figure II-30 oppose des protéines impliquées dans le métabolisme, à celles du routage RER Golgi, aux chaperones ou encore au cytosquelette ; fonctions connues pour être régulées de façon opposée chez d'autres espèces en diapause [195]. Il est alors possible que les individus prélevés ne soient pas au même stade métabolique, ce qui pourrait expliquer ces différences interindividuelles observées.

En conclusion, l'approche protéomique a permis de montrer que la socialité de L. Niger semble façonner son protéome. La séparation des castes se fait selon des critères que l'on peut directement

attribuer à leur rôle au sein de la colonie (e.g. immunité, reproduction, digestion, métabolisme élevé). Par exemple, nos résultats suggèrent un compromis évolutif en faveur de la reproduction et de la longévité, et au détriment de la résistance aux pathogènes chez la Reine. Ceci n'est rendu possible que grâce à la structure sociale de la colonie – en cercles concentriques [196] – qui isole la reine des pathogènes. Par ailleurs, l'activation de la voie ToR par la présence de nutriments dépend directement du rôle social de l'individu. La sénescence apparait donc dans ce contexte comme la conséquence de la spécialisation comportementale des individus.

Ces régulations sont actuellement examinées en détail par nos collaborateurs. A l'issue de ces examens, un article sera préparé pour publier ces résultats.

III. Conclusion

Nous avons ainsi montré, au cours de ce chapitre, qu'il est essentiel d'évaluer l'apport des différentes méthodes. En effet, il n'y a pas une solution universelle pour toutes les études. Notamment, nous avons vu ici que, pour l'étude d'une espèce non séquencée, le séquençage *de novo* n'est pas *toujours* la meilleure stratégie. Malgré l'apport indéniable du *de novo*, le préfractionnement des protéines permet d'identifier davantage de protéines. Dans le cas où le nombre d'échantillons devient trop important pour envisager un préfractionnement, alors le séquençage *de novo* permet de compenser au moins en partie la perte.
Le séquençage *de novo* est, en effet, une stratégie très bénéfique qui permet de compléter une recherche classique. En plus d'un gain en couverture de protéome, la méthode que nous avons mise au point au cours de ces travaux de thèse permet un gain en termes de données quantitatives, ce qui présente un fort intérêt pour une étude comparative différentielle, comme celles présentées ici.

L'idéal serait de pouvoir combiner préfractionnement et séquençage *de novo* afin de gagner en profondeur de protéome analysé. Cela est tout à fait envisageable si l'on se contente de l'étape d'identification des protéines, mais reste tout de même chronophage. En revanche, pour l'étape de quantification, ces deux stratégies ne sont pas « compatibles » avec les solutions logicielles envisagées ici, comme expliquées dans ce chapitre (voir Figure II-31).

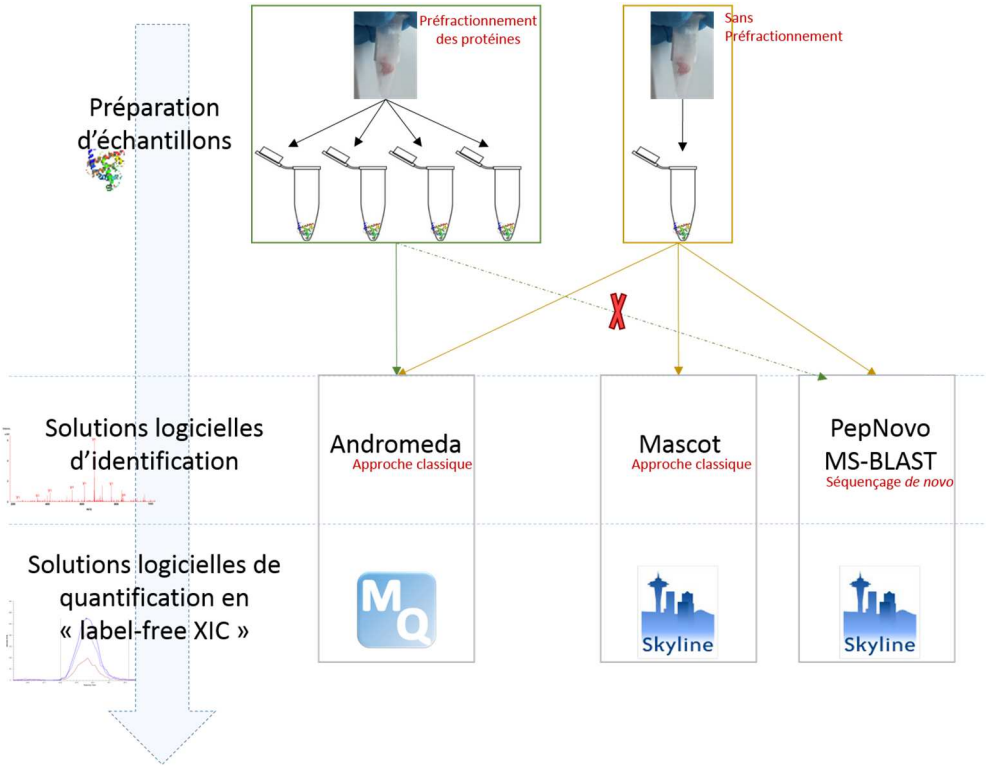


Figure II-31 : Schéma (non exhaustif) représentant différentes solutions logicielles envisageables en fonction de la préparation d'échantillons, dans le but d'une quantification en « label-free » XIC MS1.
Le fait de préfractionner impose d'utiliser MaxQuant (*MQ*) pour normaliser les données quantitatives ; tandis que la quantification des peptides *de novo* ne peut pas se faire avec ce logiciel; d'où une inadaptation des solutions logicielles entre préfractionnement et séquençage *de novo* pour la quantification en « label-free ».
Ce schéma n'est pas exhaustif, il ne présente pas *toutes* les solutions existantes, seules celles qui ont été traitées au cours de ces travaux.

Nous avons également vu que le logiciel MaxQuant permet de normaliser les données quantitatives issues de l'analyse d'échantillons préfractionnés, mais d'autres logiciels peuvent être envisagés (par exemple MFPaQ [110], ou prochainement Proline). Contrairement à Skyline, MaxQuant permet d'automatiser les étapes d'identification et de quantification. En revanche, une limitation de ce logiciel reste le manque de contrôle sur l'intégration des signaux ; tandis que l'avantage principal de Skyline réside dans le fait de pouvoir vérifier, corriger ou éliminer des intégrations. Au cours du traitement des données de BAT, nous avons dû corriger de nombreuses intégrations, ce qui nous démontre la nécessité de passer par cette étape qui, malheureusement, est très chronophage. En passant par Skyline, il est également nécessaire de réaliser tout le traitement en amont et en aval, contrairement à MaxQuant. Il faut ainsi passer par une première étape d'identification et validation des peptides et protéines, puis par une étape de normalisation des données quantitatives ; ce qui peut également être chronophage. Ce paramètre « temps », non négligeable, est à prendre en compte. Il a été montré que MaxQuant performe mieux que Skyline sans révision manuelle [197]. Si l'on manque de temps, il est alors préférable d'utiliser MaxQuant, même sur des données non préfractionnées.
En revanche, d'après la démarche expérimentale que nous avons mise en place, notamment pour les peptides *de novo*, l'idéal reste d'utiliser Skyline et de vérifier les intégrations. Cette étape cruciale permet d'être inattaquable quant à la qualité des données obtenues. De plus, les étapes de normalisation des données quantitatives et de vérification de la corrélation des peptides de chaque protéine (par la corrélation de Pearson) sont un excellent moyen de donner plus de robustesse aux données lorsque le logiciel utilisé ne permet pas directement de normaliser et d'obtenir une information quantitative par protéine.

Enfin, un dernier point très important pour l'étude d'un organisme non séquencé est le choix de la banque de données. L'identification des protéines passant par une correspondance exacte (pour la recherche classique) des peptides analysés et des peptides de la banque ; ou bien par une correspondance moins stricte (pour le séquençage *de novo*), il est essentiel que la banque utilisée soit celle de l'organisme le plus proche, afin d'avoir un maximum de séquence (strictement) conservée.
Il y a également la possibilité, non-étudiée au cours de ces travaux, de séquencer le génome (ce qui est de plus en plus réalisé pour les micro-organismes [7]) ou le transcriptome de l'organisme étudié. Le problème de ces stratégies est la taille de la banque de données générée et les erreurs d'annotation, comme expliqué en page 33.

Partie II : Résultats

Chapitre III : Analyse méta-protéomique quantitative chez une communauté complexe d'organismes

Partie II : Résultats

Chapitre III : Analyse méta-protéomique quantitative chez une communauté complexe d'organismes

I. Contexte biologique

Du point de vue des écosystèmes, la composition des sols et leur localisation géographique pourraient influencer leur couverture végétale, mais il n'y a pas de données moléculaires permettant d'étayer cette hypothèse pour ainsi mieux comprendre l'évolution écologique des sols terrestres. Ce projet s'intéresse donc au fonctionnement global des communautés biotiques des sols, et a pour but de dresser un profil « méta-fonctionnel » pour caractériser la composition des communautés biotiques du sol et leur métabolisme.

Les méthodes moléculaires sur les communautés microbiennes des sols ont permis d'apercevoir la richesse génétique des sols et les patrons de distribution des micro-organismes. La question des fonctionnalités métaboliques et biochimiques des sols, leur distribution et les effecteurs restent encore à explorer. Les méthodes « omics » ont révélé l'ampleur de la complexité génique et de l'expression de ces gènes chez les organismes du sol et sont très prometteurs pour mener des études fonctionnelles.

L'objectif était de réaliser une analyse méta-protéomique de deux sols qui diffèrent par leur couverture végétale et leur exposition, et qui ont donc des traits de vies contrastés : l'un est une prairie dominée par *Festuca alpina* (Pairie à Fétuques, « PF »), l'autre une lande dominée par *Vaccinium myrtillus* (Lande à Vaccinium, « LV »).

Ce projet a été réalisé en collaboration avec les Dr Roberto Geremia et Jean-Marc Bonneville, Laboratoire d'Ecologie Alpine (LECA), Grenoble.

Schéma expérimental :

Les échantillons ont été prélevés au Col du Lautaret, dans les Alpes. Les prélèvements ont été effectués sur trois sites différents, d'altitudes et d'expositions différentes, pour les deux types de sol (LV et PF). Sur chaque site, un échantillon était constitué de 3 prélèvements (afin d'obtenir un échantillon le plus représentatif possible du sol étudié). Trois échantillons par site ont ainsi été constitués afin d'avoir des triplicas. Ainsi, nous disposions de 18 échantillons (9 par type de sol), comme présenté en Figure III-1.

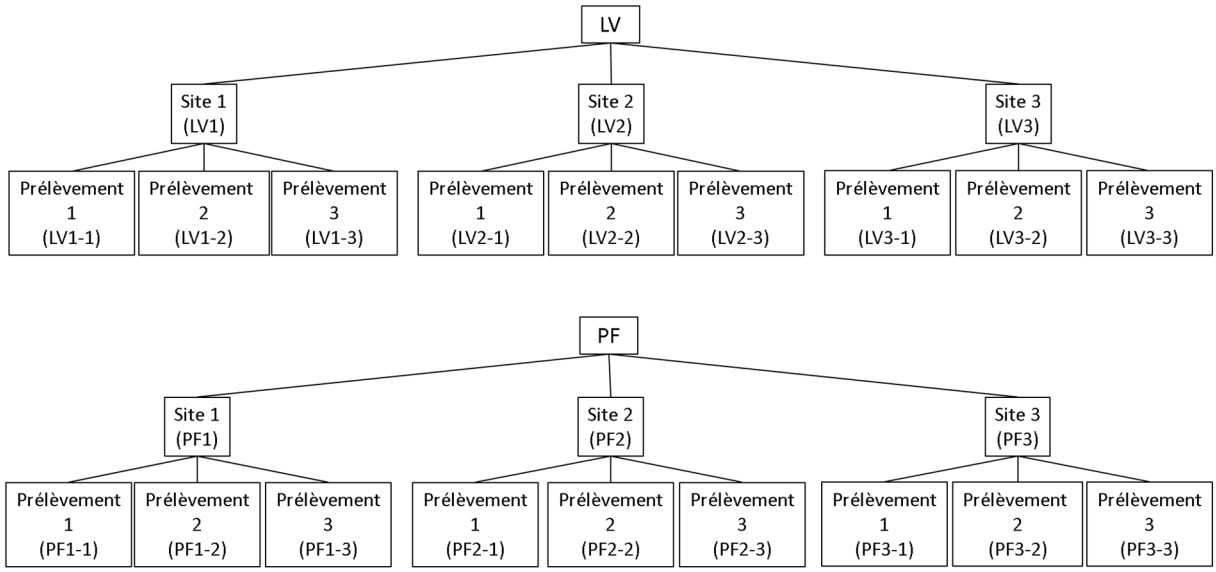


Figure III-1 : Schéma expérimental des 9 prélèvements de sol.
LV= Lande à Vaccinium ; PF= Prairie à Fétuques

Les échantillons ont été prélevés à l'aide d'une tarière, puis les végétaux éliminés et la terre tamisée, afin de s'assurer d'analyser uniquement les protéines du sol, et non des végétaux (voir Figure III-2).

Figure III-2 : Photos du déroulement du prélèvement des échantillons de sol.
L'échantillon est tout d'abord extrait du sol, puis les végétaux sont éliminés, la terre tamisée et finalement congelée pour préserver l'intégrité de l'échantillon.

II. Contexte analytique et objectifs

La difficulté majeure de ce projet résidait dans la singularité des échantillons à analyser, et ce pour plusieurs raisons. Tout d'abord la nature même des échantillons (essentiellement minéraux), qui diffère beaucoup des tissus, plus courant en protéomique, a nécessité des développements pour l'extraction des protéines.

Ensuite, une analyse métaprotéomique consiste à analyser l'expression protéique de toutes les communautés présentes au sein d'un même échantillon [60], ce qui diffère des autres projets présentés dans ce manuscrit, où l'on analysait le protéome d'un tissu donné d'**une** espèce donnée. Ainsi, un deuxième niveau de complexité est ajouté, en plus du grand nombre de protéines à détecter, il y a également un grand nombre d'organismes. En effet, le fait d'analyser plusieurs organismes à la fois augmente grandement la complexité du traitement bio-informatique qui suit les analyses. Lorsque l'on analyse un seul organisme, dont le génome n'est pas séquencé, il « suffit » de trouver la banque

de données d'un organisme ou d'un groupe d'organismes proches et l'on sait au final que toutes les protéines identifiées appartiennent à l'organisme étudié. Les études métaprotéomiques requièrent donc en général au préalable des analyses métagénomiques dont les résultats sont utilisés pour créer une banque de données, soit génomique soit protéique, qui est spécifique de l'échantillon et la plus complète possible [7]. Cependant, cette méthode ne donne pas accès à toutes les séquences protéiques étant donné la qualité du séquençage, de l'assemblage et de l'annotation du génome [73]. De plus, ces analyses n'étaient pas envisagées par notre collaborateur et nous avons donc dû mettre en place des solutions pour déterminer les organismes présents dans les échantillons à analyser, en déduire la banque de données protéique la plus représentative des organismes présents (sachant qu'une grande majorité d'entre eux ne sont pas ou mal séquencés), et enfin « valider » nos données de protéomique.

Pour déterminer les organismes les plus présents dans l'échantillon et en déduire quelle banque de données protéique utiliser, nous nous sommes orientés avec notre collaborateur vers une analyse en MetaBarCoding [76], dont le principe est expliqué en page 35.

Enfin, pour renforcer la confiance dans nos données protéomiques, nous les avons confrontées à des données incomplètes de métatranscriptomique. Généralement, ces données sont utilisées pour établir un « pseudo-métagénome » [75] et ainsi obtenir une banque de données protéiques personnalisées par traduction dans les 6 cadres de lecture [67]. Seulement, les banques ainsi générées sont très volumineuses et les séquences ainsi obtenues sont incomplètes et ne recouvrent pas tout le génome. En concertation avec nos collaborateurs, nous avons considéré que les données transcriptomiques obtenues étaient « trop » incomplètes et que leur utilisation aurait conduit à un nombre d'identification plus restreint que via la construction d'une banque protéique basée sur la détermination par MetaBarcoding des taxonomies représentées dans les échantillons. C'est pourquoi nous avons préféré utiliser ces données, non pas pour interpréter les données spectrales, mais pour confirmer la présence (en protéomique et en transcriptomique) des peptides identifiés dans une banque de données préexistante.

A partir de ces diverses stratégies, l'objectif ici était de réaliser une étude comparative du protéome des deux types de sols. Nous avons donc choisi de mettre en place une méthode de quantification « label-free » XIC MS1.

III. Développements méthodologiques

A. Préparation d'échantillons

La première étape d'optimisation consistait à trouver une méthode de préparation d'échantillons optimale pour l'extraction des protéines d'échantillons de sol. Nous nous sommes intéressés à deux protocoles, que l'on appellera *Protocole 1* [198] et *Protocole 2* [199] (voir Figure III-3), qui sont détaillés dans la Partie Expérimentale en page 164, et rapidement résumés ici :

- *Protocole 1* : Les protéines ont été extraites dans le tampon 1 (5% SDS, 50mM Tris pH 8.5, 0.1mM EDTA, 1mM $MgCl_2$, 50mM DTT). Après une légère centrifugation, le culot a été éliminé et à partir du

surnageant les protéines ont été précipitées à l'acide trichloroacétique (TCA). Le culot de précipitation a été récupéré et suspendu dans un tampon compatible avec l'électrophorèse (Laemmli: 10mM Tris pH 6.8, 1mM EDTA, 5% βME, 5% SDS, 10% glycérol), les protéines ont ensuite été dosées.

- *Protocole 2* : Les protéines ont été extraites dans un tampon 2a (0.25M citrate pH8). Après une légère centrifugation, le culot 2 et le surnageant 2 ont été récupérés.

- *Protocole 2a* : A partir du surnageant 2, les protéines ont été précipitées au TCA, le culot a été repris dans un tampon Laemmli et les protéines dosées.

- *Protocole 2b* : le culot 2 a été repris dans le tampon 2b (1% SDS, 0.1M TrisHCl pH 6.8, 20mM DTT, 50mM NH_4HCO_3). Après une deuxième centrifugation, le culot a été éliminé et à partir du surnageant les protéines ont été précipitées au TCA. Le culot protéique a ensuite été repris dans du tampon Laemmli et les protéines dosées.

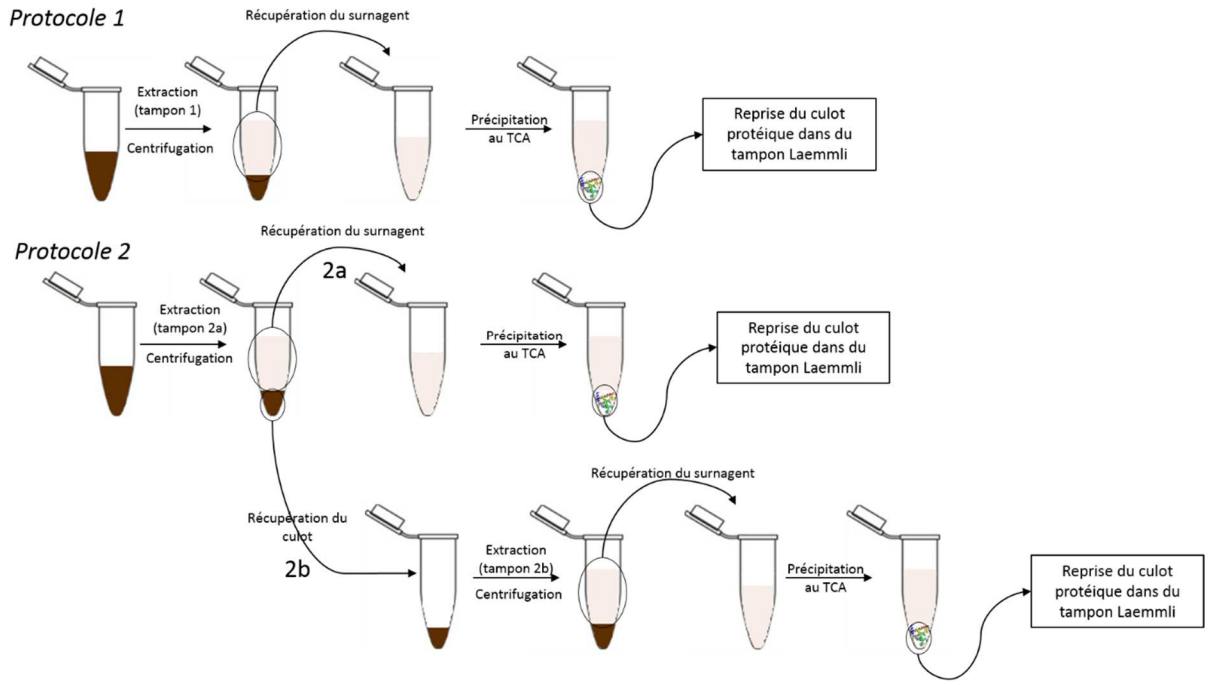


Figure III-3 : Représentation schématique des tests de préparation d'échantillons de sol.

Ces protocoles ont été comparés sur un échantillon test pour les deux types de sol (LV et PF).

Après dosage, les échantillons ont été déposés sur gel SDS-PAGE, voir Figure III-4.

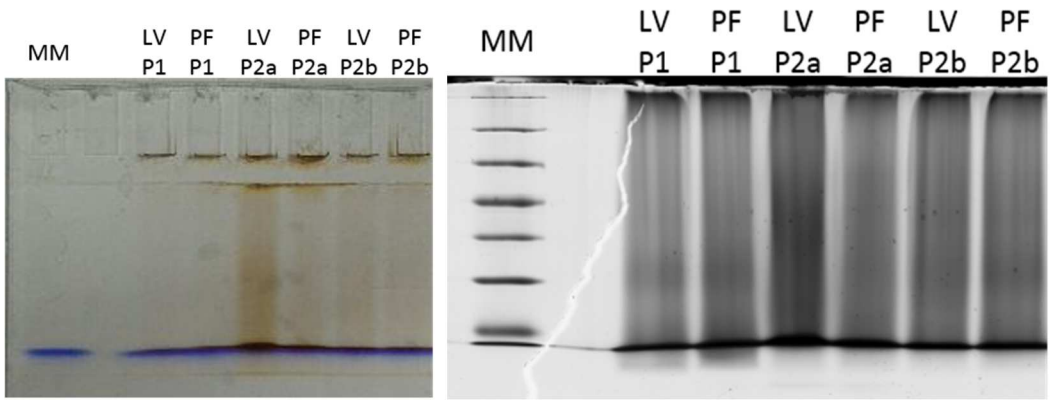


Figure III-4 : Image des gels SDS-PAGE avant et après coloration au Bleu de Coomassie.
P1= protocole 1 ; P2a= protocole 2a ; P2b= protocole 2b.

On observe que, contrairement à un extrait protéique de tissu, on ne distingue pas clairement de bandes protéiques sur le gel après coloration mais plutôt une « traînée » (« smear »), ce qui a déjà été observé dans la littérature [200].

On remarque également que le *Protocole 2* ne permet pas d'éliminer les acides humiques, responsables de la couleur de la terre. En effet, on voit nettement une traînée de couleur marron sur toutes les pistes « P2 » avant coloration au Bleu de Coomassie, et plus particulièrement sur la piste « LV P2a » (dû au fait que LV est un terre plus foncée). Cela pose deux problèmes :

- Tout d'abord, il y a un risque d'incompatibilité et d'encrassement avec la spectrométrie de masse.
- Ensuite, cette couleur peut interférer avec le dosage (colorimétrique) des protéines, et fausser les résultats. Pour une terre plus foncée (qui contient plus d'acides humiques), la quantité de protéines sera surestimée. Dans la mesure où l'on analyse des terres de couleur différente, cela posera problème pour l'analyse quantitative. On peut observer cet effet sur les Figure III-5 et Figure III-6. Avec le *Protocole 1*, il ne semble pas y avoir d'interférences des acides humiques dans le dosage, comme le montre l'intensité globale de l'analyse des deux sols, qui est comparable (Figure III-5). En revanche, avec le *Protocole 2a*, l'intensité globale de l'analyse de l'échantillon PF est supérieure à celle de LV (Figure III-6), ce qui va dans le sens d'une surestimation de la quantité de peptides injectées pour LV. En effet, d'après la Figure III-4, LV est plus foncée et conserve donc plus d'acides humiques après extraction et migration des protéines sur gel ; il est donc plus interféré que PF, ce que l'on retrouve dans les intensités.

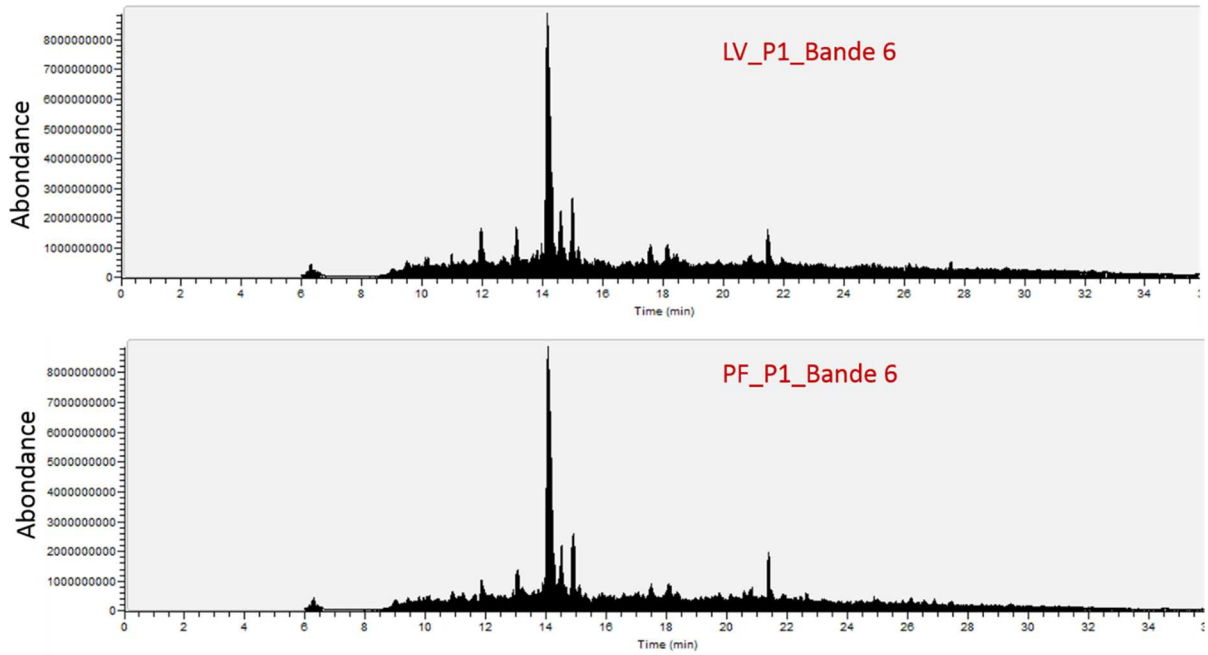


Figure III-5 : Comparaison des chromatogrammes de la fraction 6 des échantillons tests LV et PF, préparés avec le *Protocole 1* (« P1 »).

L'intensité globale des peptides est comparable entre les deux sols, ce qui montre l'élimination effective des acides humiques avec le *Protocole 1*.

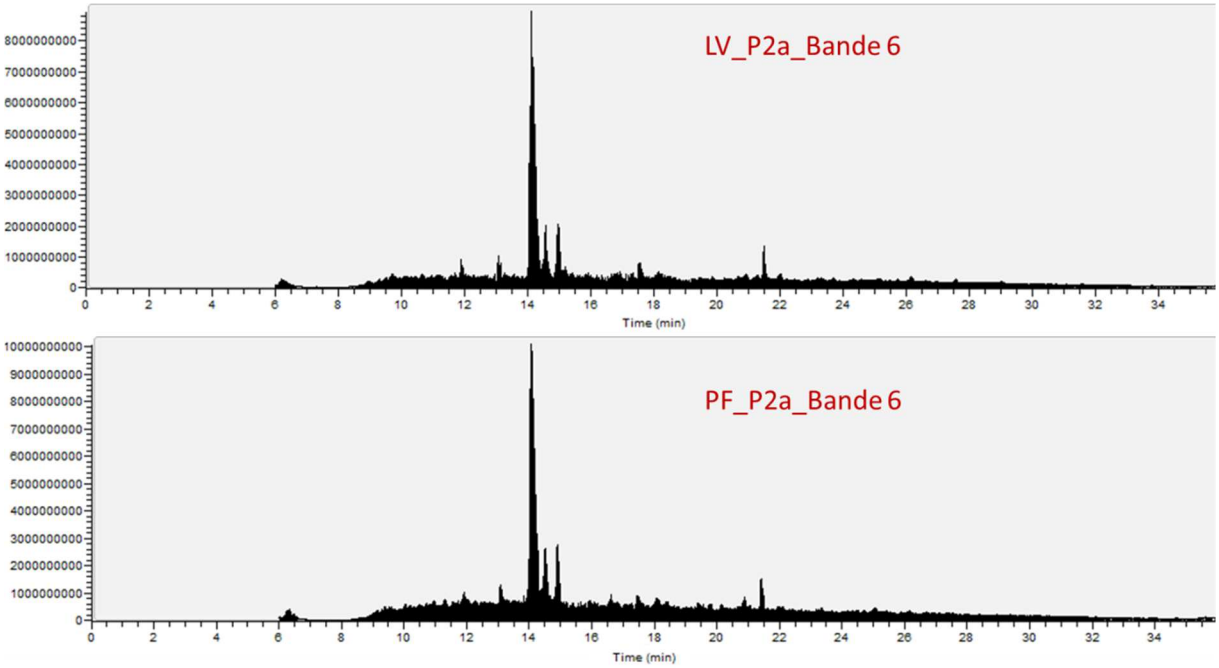


Figure III-6 : Comparaison des chromatogrammes de la fraction 6 des échantillons tests LV et PF, préparés avec le *Protocole 2a* (P2a).

L'intensité globale des peptides est supérieure pour le sol PF, ce qui concorde avec une interférence des acides humiques avec le *Protocole 2a* et donc une surestimation de la quantité de protéines pour l'échantillon LV (qui entraîne une plus faible quantité de peptides injectés), qui est plus foncé, donc plus interféré.

A ce stade, la banque de données adaptée n'avait pas encore été déterminée. En effet, l'étape de MetaBarCoding a été réalisée ultérieurement, sur les échantillons de l'expérience finale, qui n'avaient pas encore été prélevés au moment de ces tests.

Nous avons donc réalisé une recherche classique dans tout Swiss-Prot (Mai 2015, 548 454 séquences) pour identifier les protéines à partir des données spectrales acquises pour ces tests à l'aide de Mascot. Dans l'échantillon **LV** : nous avons ainsi identifié **1260** protéines avec le *protocole 1* et **735** avec le *protocole 2* (2a et 2b réunis). Dans l'échantillon **PF** : **714** protéines avec le *protocole 1* et **663** avec le *protocole 2*.

Pour l'échantillon LV, il y a une grande différence entre les deux protocoles. Comme expliqué plus haut, après extraction, la terre LV semble contenir beaucoup d'acides humiques, ainsi, la quantité de protéines est certainement surestimée pour LV avec le *protocole 2*, et au lieu d'avoir déposé 50µg sur le gel, nous avons en réalité déposé et donc analysé beaucoup moins, ce qui expliquerait que l'on identifie moins de protéines. Cet écart est beaucoup moins important entre les deux protocoles pour PF car la terre est moins foncée et donc le dosage moins faussé, ce que l'on peut observer sur la Figure III-4.

Ainsi, même si la grande différence entre les deux protocoles pour la terre LV semble davantage due à une surestimation de la quantité de protéines qu'à une réelle différence d'extraction des protéines (comme suggéré par les résultats obtenus avec PF) ; il n'en reste pas moins qu'il est essentiel de pouvoir estimer de manière juste et fiable la quantité de protéines extraites d'un échantillon, surtout pour réaliser une analyse quantitative comparative.

Nous avons donc choisi d'utiliser le *Protocole 1* pour extraire les protéines des échantillons de sol.

B. Apport du metabarcoding pour la construction de la banque de données

A partir des mêmes 18 échantillons, nos collaborateurs ont réalisé des analyses de « metabarcoding » [76] (dont le principe est expliqué en page 35), ce qui nous a permis de caractériser la composition des communautés biotiques des sols étudiés.

Dans l'ensemble des 18 échantillons de sol, nous avons ainsi mis en évidence les organismes les plus présents : il y avait beaucoup de champignons, parmi lesquels les *Agaricomycetidae* étaient particulièrement abondants ; nous avons également identifié des bactéries : *Rhizobiales*, *Bacteroides*, *Actinobacteria*, *Acidobacteria*, *Verrucomicrobia*, *Chloroflexi*. A partir des banques de données protéiques existantes (Uniprot), nous avons donc extrait les séquences protéiques pour ces organismes et ainsi constitué une banque personnalisée. Il a cependant fallu déterminer si l'on extrayait ces séquences de tout Uniprot (en tenant compte de la banque TrEMBL, dont les annotations sont automatiques et non vérifiées) ou uniquement de Swiss-Prot. Ce choix s'est basé sur la taille des banques de données. Il valait mieux utiliser Uniprot lorsque Swiss-Prot était trop incomplète et si TrEMBL n'était pas trop conséquente. C'était notamment le cas pour les organismes *Agaricomycetidae*, *Acidobacteria*, *Verrucomicrobia*, *Chloroflexi* (voir Tableau 3). En revanche, pour les *Fungi*, *Rhizobiales*, *Bacteroides* et *Actinobacteria*, la banque TrEMBL devenait trop importante pour envisager de l'utiliser, d'où le choix de Swiss-Prot. En effet, plus on augmente l'espace de recherche, et donc le nombre de candidat pour expliquer un spectre MS/MS, moins l'identification est spécifique et plus le risque d'identifier de faux positifs est augmenté. De plus, les moteurs de recherche actuels gèrent plus difficilement les recherches en un temps raisonnable avec des banques trop larges [201]. Ici, si on avait choisi UniProt pour chaque taxon étudié, la banque aurait dépassé les 14 millions de

séquences, ce qui aurait été difficilement gérable d'un point de vue computationnel. Nous avons donc opté pour un compromis entre Swissprot et Swissprot+Trembl pour retenir un maximum de séquences sans dépasser le million au total (Voir tableau 3).

Tableau 3 : Nombre de séquences protéiques contenues dans Swiss-Prot et TrEMBL pour chacun des organismes « sélectionnés ».

Les nombres en gras correspondent aux séquences (Swiss-Prot ou UniProt) retenues.

Taxon (taxon id)	*Nombre de séquences dans Swiss-Prot*	*Nombre de séquences dans UniProt (TrEMBL + Swiss-Prot)*
Fungi *(4751)*	***31 763***	*562 798 2*
Agaricomycetidae *(452333)*	*344*	***441 355***
Rhizobiales *(356)*	***18 596***	*235 499 3*
Bacteroides *(816)*	***1318***	*660 729*
Acidobacteria *(57723)*	*644*	***42 640***
Actinobacteria *(201174)*	***23 983***	*519 828 1*
Verrucomicrobia *(74201)*	*385*	***65 769***
Chloroflexi *(32061)*	*1271*	***28 732***

La banque ainsi créée, que l'on appellera **« db1 »** contenait **644 391** séquences protéiques.

C. Validation par FDR

Comme expliqué en page 36, la méthode de calcul du taux de faux positif FDR, classiquement utilisée en protéomique, n'est pas adaptée aux recherches dans les banques de données trop importantes. Il est difficile d'évaluer cette taille limite de banque à partir de laquelle le FDR n'est plus valable. Afin de tester cela sur notre banque de données « db1 », nous avons validé les identifications à l'aide de plusieurs seuils de score et regarder l'influence sur le FDR et le nombre d'identifications.

Nous avons réalisé une recherche classique (à l'aide de l'algorithme de recherche Mascot v2.5.1 Matrix Science) sur un échantillon test dans la « db1 ». Ensuite, à l'aide du logiciel Scaffold v4.3 (Proteome software inc.), différents filtres ont été appliqués pour valider les identifications peptidiques, basés sur le score d'ion donné par Mascot (« ion score »), la différence entre le score d'ion et le score d'identité de Mascot (« ion-ID score »), et sur la longueur de la séquence du peptide (« sequence length ») :

- Filtre 1 : (ion-ID score) > -5 ; ion score > 35 ; sequence length >7 ; nous avons identifié **1071** protéines, dont 191 « decoy », soit un FDR de **21,7%**.
- Filtre 2 : (ion-ID score) > 0 ; ion score > 45 ; sequence length >7 ; nous avons identifié **569** protéines, dont 42 « decoy », soit un FDR de **8%**.
- Filtre 3 : (ion-ID score) > 20 ; ion score > 50 ; sequence length >7 ; nous avons identifié **354** protéines, dont 3 « decoy », soit un FDR de **0,9%**.

On remarque que pour atteindre un FDR acceptable (inférieur à 1%), il faut appliquer des filtres très stringents et le nombre de protéines identifiées est largement réduit.

A titre de comparaison, sur un projet plus classique, comme celui présenté dans le Partie II -Chapitre II -II. -A, en page 102, en appliquant le filtre 1 on obtient un FDR de 0,9% pour près de 3000 protéines

identifiées (pour un échantillon analysé sur le même instrument, avec le même gradient, les mêmes quantités injectées).

On constate effectivement, sur ces données de métaprotéomique, l'effet d'une large banque de données sur le FDR. C'est pourquoi nous avons essayé de réduire la taille de la banque de données protéiques utilisée.

D. Réduction de la banque de données

D.1. Taxonomies identifiées

Pour réduire la taille de la banque de données, nous avons pensé utiliser uniquement les organismes permettant d'identifier des protéines. Nous avons identifié 505 taxons différents, sur 225 409 dans la « db1 », les identifications étant validées avec le filtre 1 (présenté dans le paragraphe précédent).

Nous avons donc constitué une nouvelle banque, contenant uniquement les données protéiques de ces 505 taxons. Les séquences protéiques ont été extraites d'Uniprot pour les espèces appartenant aux familles *Agaricomycetidae*, *Acidobacteria*, *Verrucomicrobia* et *Chloroflexia*, et de Swiss-Prot pour les autres (comme pour la banque originale « db1 »). Cette nouvelle banque contenait **631 533** séquences, contre **644 391** pour la banque originale. On constate donc que la banque de données n'a pas été significativement diminuée (de 2% seulement) et que la nouvelle banque reste tout aussi large, et ce malgré la réduction drastique du nombre d'organismes. Notre explication est la suivante :
A chaque protéine est assigné un taxon, nous avons donc compté le nombre de taxon identifié grâce à la recherche dans la « db1 ». Seulement, ce taxon ne se trouve pas toujours en bas de classification et regroupe en réalité plusieurs espèces. Par exemple, nous avons identifié la bactérie *Rhizobium meliloti*, taxon n°382, qui regroupe 65 sous-espèces (différentes souches de la bactérie) ; nous avons donc en fait identifié 65 taxons et non un seul. On a donc extrait les séquences protéiques, non pas d'une espèce, mais de plusieurs (65 dans l'exemple donné de *Rhizobium meliloti*) pour chacun des 505 taxons. On se rapproche finalement beaucoup de la banque de données originale « db1 », ce qui explique que la taille de la nouvelle banque de données ne soit quasiment pas réduite.

D.2. Protéines « decoy »

Une autre idée était d'éliminer les organismes qui n'identifiaient que -ou majoritairement- des protéines « decoy », c'est-à-dire des organismes « peu informatifs ». Ici aussi, la première recherche a été effectuée dans la « db1 » et validée avec le filtre 1.

Dans l'ensemble des échantillons, 2245 protéines « decoy » ont été identifiées, grâce à 117 organismes. Parmi ceux-ci, 31 ne permettent d'identifier que des protéines « decoy », ce qui représente uniquement 34 protéines « decoy », donc les éliminer ne permettrait pas un gain.

D.3. Recherches consécutives

Une dernière idée était de constituer une nouvelle banque de données à partir des protéines déjà identifiées [89]. La méthode décrite dans l'article [89] consiste à faire une première recherche dans une large banque de données, sans évaluer de FDR (pas de « decoy » dans la banque), puis d'utiliser

dans un second temps toutes les protéines identifiées par cette première recherche. Une seconde banque de donnée (dont la taille est inférieure à 1% de la banque originale dans l'article) est construite à partir de ces premières identifications, en version « target-decoy », permettant d'effectuer une deuxième recherche en évaluant un FDR acceptable (< 5% dans cet article).

Nous avons donc appliqué cette méthode à nos données. Après une première interprétation, sur l'ensemble des échantillons, effectuée par Andromeda (MaxQuant v1.5.3.30) dans la « db1 » avec un FDR accepté de 70%, nous avons conservé uniquement les protéines identifiées (pas uniquement les peptides identifiés, mais les séquences protéiques complètes). La nouvelle banque ainsi constituée, que l'on appellera **« db2 »** contenait **13 771** séquences, soit une réduction à 2% de la taille originale. Ainsi, la nouvelle banque de données, de taille beaucoup plus « raisonnable », permet de valider les données grâce au calcul du taux de faux positifs.

Par la suite, le logiciel MaxQuant v1.5.3.30 a été utilisé pour identifier, quantifier, et normaliser les données ; dans la mesure où nous avons choisi de préfractionner les protéines.

E. Double validation grâce à la méta-transcriptomique

Un séquençage d'ARN très incomplet (« RNA seq ») [202] a été réalisé sur les mêmes échantillons par nos collaborateurs, puis les séquences d'ARN ont été assemblées en contig. Nous nous sommes servis de ces données transcriptomiques pour valider nos identifications.

Les séquences peptidiques identifiées ont été comparées aux contigs grâce à l'algorithme BLASTx, ce qui nous a permis d'avoir une double confirmation de la présence du peptide : par protéomique d'une part et par transcriptomique d'autre part. Bien sûr, la méthode de RNAseq ne permet de séquencer **toutes** les séquences d'ARN, tout comme la protéomique ne permet pas d'identifier **tous** les peptides, il n'est donc pas attendu de valider l'ensemble des peptides identifiés par cette méthode. Cela n'invalide pas pour autant ces peptides non vus en transcriptomique.

Sur **4551** peptides identifiés en protéomique, il y avait une correspondance par BLASTx pour **2741** peptides sur **29 336** contigs. Il y avait donc plus de 10 fois plus de contigs que de peptides, ce qui montre qu'un peptide peut correspondre à plusieurs contigs ; il s'agirait alors de contigs issus d'ARN codant des protéines homologues entre différents organismes.

Il y a également des cas où plusieurs peptides différents correspondent à un même contig. En effet, un contig (plus long qu'un peptide) peut contenir plusieurs peptides. Or, un contig traduit pour une seule protéine. Tous les peptides alignés sur un même contig, par BLAST, devraient donc appartenir à la même protéine en protéomique (ou homologues, car il peut s'agir de peptides conservés entre plusieurs espèces). Pour les autres cas, il est difficile de savoir s'il s'agit de peptides partagés entre différentes protéines ou bien tout simplement une erreur de séquençage ou encore de BLAST.

Ainsi, nous avons considéré comme « doublement-validés » les peptides qui étaient assignés à la même protéine lorsqu'ils correspondaient au même contig (voir Figure III-7). Ensuite nous avons propagé la validation : un peptide validé par contig validait également la/les protéines auxquelles il

était assigné car il apporte une preuve que la protéine est effectivement présente. Cela valide par conséquent tous les peptides de cette protéine, même ceux n'ayant pas eu de correspondance par BLASTx contre les contigs. Au final, **4005** peptides sur 4551 ont été « doublement-validés », soit **1476** protéines sur 1898.

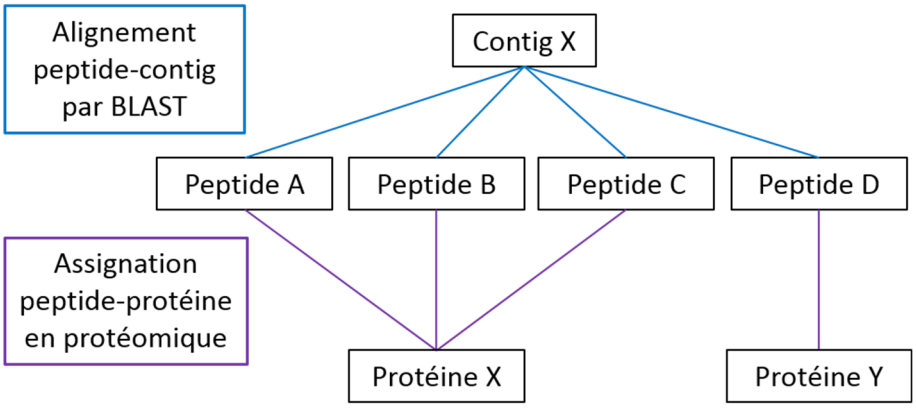


Figure III-7 : Exemple d'un contig X aligné par BLAST contre 4 peptides identifiés en protéomique (peptides A, B, C et D).

3 des peptides (A, B et C) ont été assignés à la même protéine X en protéomique, tandis que le peptide D appartient à une autre protéine Y. A, B et C sont donc « doublement-validés » par transcriptomique, et D reste simplement validé par protéomique.

IV. Analyse des échantillons

Pour résumer, les protéines ont été extraites des échantillons de sol selon le *protocole 1* [198], puis préfractionnées sur un gel SDS-PAGE. Les données spectrales acquises sur un Q-Exactive + (Thermo Fisher) ont ensuite été interprétées grâce à l'algorithme de recherche Andromeda, dans la banque de données que nous avons personnalisée (« db2 ») grâce au metabarcoding et à une première recherche non stringente. Nous avons ensuite utilisé les données de RNAseq pour doublement valider nos identifications.

Le protocole de ces étapes est détaillé dans la Partie Expérimentale, en page 164.

V. Résultats

L'intensité totale des chromatogrammes est comparable entre les deux sols, comme le montre la Figure III-8, ce qui suggère un dosage des protéines totales fiable et reproductible. Nous avons donc bel et bien injecté la même quantité de protéines pour les deux sols, ce qui montre qu'il n'y a pas d'interférences d'acides humiques, comme expliqué en page 125.

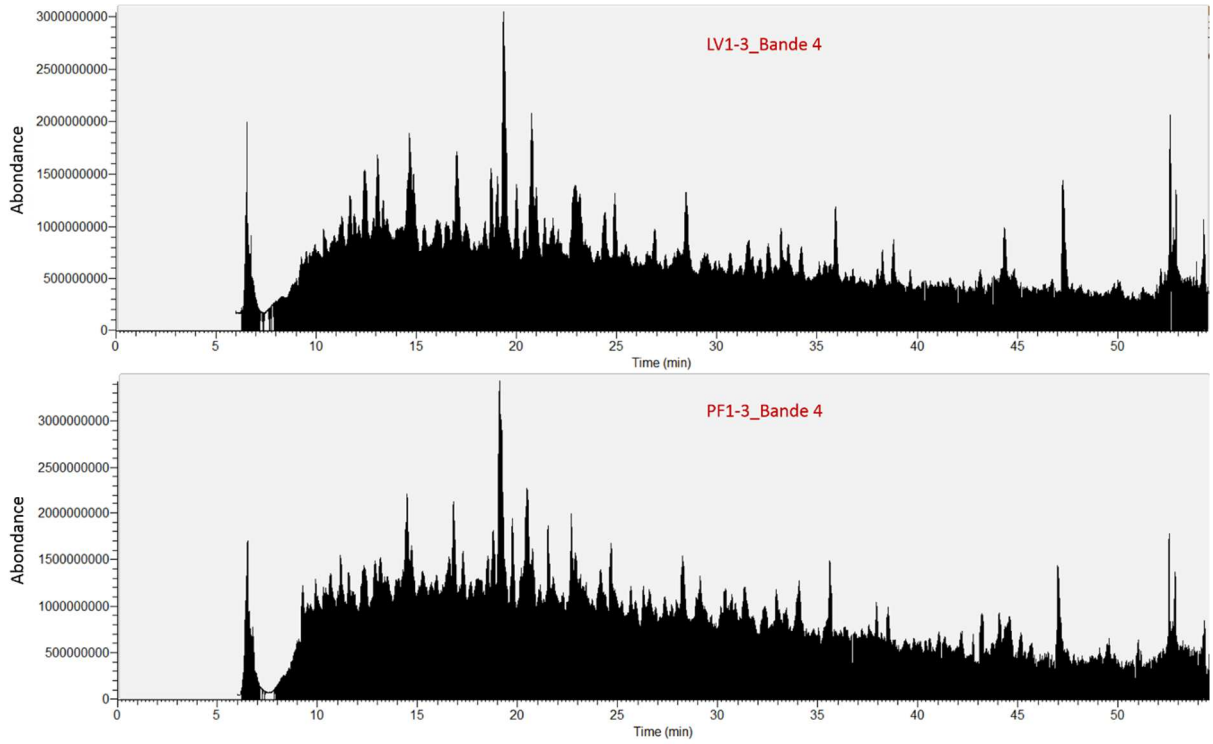


Figure III-8 : Comparaison des chromatogrammes de la fraction 4 d'un échantillon LV et un échantillon PF. L'intensité globale des peptides est comparable entre les deux sols, ce qui montre l'élimination effective des acides humiques.

Sur toutes les analyses, seuls **5,6%** des spectres MS/MS ont été identifiés et validés. A titre de comparaison, sur le même instrument (donc avec des spectres de qualité comparable), les analyses de foie de souris ont permis d'identifier 55% des spectres. Toujours sur le même instrument mais chez une espèce non séquencé (le campagnol), 32% des spectres étaient identifiés ; 27% chez le microcèbe (également non séquencé). Le fait d'étudier un organisme dont le génome n'est pas séquencé réduit le pourcentage de spectres assignés, ce qui est normal étant donné le manque de spécificité de la banque de données de recherche. En revanche, étudier un métaprotéome diminue drastiquement le nombre de spectres identifiés. Cela est sans nul doute dû au fait que la banque de données est loin d'être complète. Etant donné la grande complexité de tels échantillons, il est effectivement impossible d'obtenir une banque complète pour *chacun* des organismes présents. De ce fait, il y a beaucoup de spectres issus de la fragmentation de peptides absents de la banque, donc non identifiables [73].

Malgré cela, nous avons tout de même pu identifier **1898** protéines, avec **4551** peptides. Sur ceux-là, **2583** peptides ont été directement doublement validés par transcriptomique, puis **1422** peptides supplémentaires ont été validés par « propagation » au niveau protéique, soit **4005** peptides doublement validés, ce qui correspond à **88%** de tous les peptides identifiés. Finalement, seuls 56% des peptides avaient leur correspondance en transcriptomique. Cela peut s'expliquer par le fait que le séquençage de l'ARN n'est pas complet, surtout pour des échantillons aussi complexes (transcriptome de plusieurs organismes). De plus, l'assemblage en contig peut être sujet aux erreurs : par exemple des séquences d'ARN provenant de différents organismes ont pu être assemblées en un même contig « chimère », rendant impossible la correspondance avec des peptides trypsiques issus de l'analyse

protéomique. Enfin, les contigs ne correspondent pas forcément à des peptides trypsiques ; des contigs peuvent avoir été obtenu qui pourraient bien être assignées aux protéines identifiées mais ne correspondent pas aux séquences peptidiques auxquelles nous accédons. L'utilisation de plusieurs enzymes, autres que la trypsine, aurait sans doute pu permettre de réduire le nombre de peptides « orphelins ». Finalement, certains peptides trypsiques identifiés en protéomique chevauchent potentiellement deux contigs.
Malgré cela, le recouvrement avec les peptides identifiés reste très satisfaisant et cette méthode nous a permis de confirmer par deux méthodes différentes la présence d'un très grand nombre de peptides.

Sur les 1898 protéines, 1784 étaient identifiées dans les échantillons LV et 1864 dans les échantillons PF. Dans nos tests, on identifiait près de deux fois plus de protéines dans l'échantillon LV (voir page 125). Cette différence peut s'expliquer par un effet de sous-échantillonnage. En effet, dans les tests un seul échantillon a été analysé, et non 9 par sol ; il est possible que ce prélèvement n'était pas représentatif du sol PF, qu'il ait été prélevé dans un endroit peu riche. Le fait d'avoir fait 9 prélèvements dans le schéma expérimental final permet d'éliminer cet effet de sous-échantillonnage.

Nous avons donc identifié et validé (1% de FDR) 1898 protéines dans la « db2 ». Avec les mêmes critères (FDR 1%, identifications Andromeda sur l'ensemble des échantillons), seuls 681 protéines étaient identifiées dans la « db1 ». Finalement, le fait d'avoir réduit la banque de données en faisant deux recherches consécutives a permis de gagner 1217 protéines, soit 64% des identifications. Dans l'étude présenté dans l'article [203], cette technique a permis un gain de 63% d'identifications de protéines microbiennes.
La Figure III-9 montre le recouvrement des peptides identifiés avec les deux banques. Pour les peptides aussi, le gain est très important : près de 75%. Dans l'étude [89], la méthode de recherches consécutives leur avait permis de doubler le nombre de peptides microbiens identifiés.
Enfin, la recherche dans la « db1 » ne permettait d'interpréter que 1% des spectres sur l'ensemble des échantillons, contre près de 6% avec la « db2 », ce qui montre la très faible spécificité de la « db1 ».
Ces résultats montrent l'intérêt d'avoir réduit la banque de données, cela nous a effectivement permis d'interpréter davantage de spectres et de gagner en couverture de protéome et de séquence protéique grâce à des identifications fiables et validées.

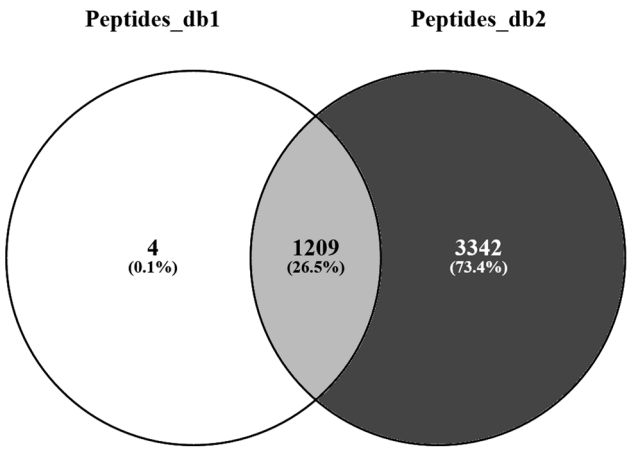


Figure III-9 : Recouvrement des peptides identifiés dans la « db1 » et la « db2 » (validation à 1% FDR).

Pour la quantification, nous avons considéré les protéines ayant au moins 6 valeurs valides sur 9 réplicas, il reste **839 protéines quantifiables**.

Comme expliqué précédemment (Partie IIChapitre II.C.4), ce sont les valeurs d'abondance protéique (« LFQ intensity ») fournies par Maxquant qui ont été directement utilisées.

Pour l'instant, nous n'avons pas éliminé les protéines quantifiées avec un seul peptide. Comme précisé précédemment (Partie IIChapitre II.C.4), ces protéines sont étudiées avec précaution.

VI. Suivi de la stabilité instrumentale

Au cours des 11 jours d'analyses, nous avons évalué la stabilité du système grâce à l'analyse de contrôles qualité. Nous avons analysé 5 fois, à intervalle régulier, un échantillon de référence. Cet échantillon, appelé « Pool », contenait un mélange de tous les échantillons ainsi qu'un mélange de peptides trypsiques synthétiques (iRT kit ; Biognosys AG, Schlieren, Switzerland) en quantité équivalente. Ces peptides iRT ont également été ajoutés à tous les échantillons, pour suivre la stabilité chromatographique. Les résultats sont présentés en Figure II-14 et Figure II-15.

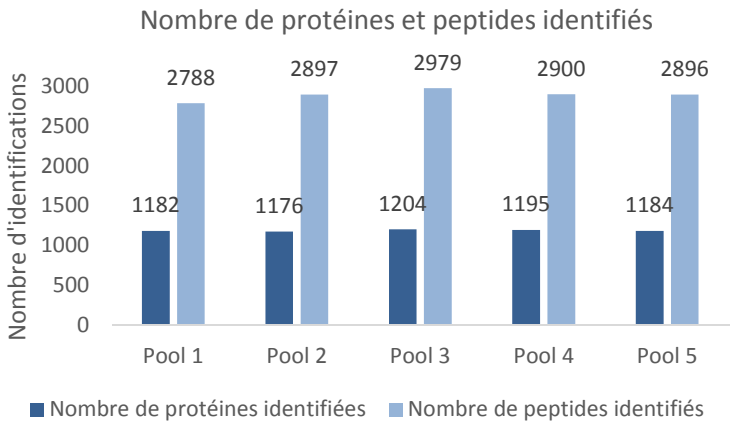


Figure III-10 : Nombre protéines et peptides identifiés dans les 5 injections répétées de l'échantillon de référence (ou Pool).

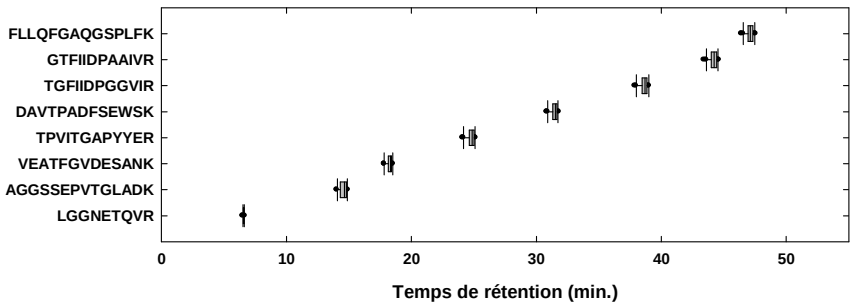


Figure III-11 : Représentation de la variation des temps de rétention des peptides iRT entre tous les échantillons.

On constate, que le nombre de protéines et peptides identifiés est resté stable au cours des 8 jours d'analyse, avec un CV de 1% et 2% respectivement, ce qui montre une bonne reproductibilité du

système. De plus, le coefficient de variation moyen d'abondance de toutes les protéines entre les 5 injections de l'échantillon de référence était de 20%, ce qui montre une bonne reproductibilité des données quantitatives. Enfin, le système chromatographique a également montré une très bonne stabilité, puisque les coefficients de variation de temps de rétention des peptides iRT étaient en moyenne égaux à 1%.

VII. Interprétation des résultats

Une analyse en composante principale réalisée sur l'ensemble des protéines quantifiées a mis en évidence deux axes qui séparent les groupes, comme présenté en Figure III-12 A. L'axe 1 (qui explique 52% de la variance) sépare nettement les deux types de sol, tandis que l'axe 2 (12% de la variance) sépare plutôt les échantillons par site, et particulièrement le site LV1 de LV2 et LV3. La nette différence entre les données correspondant au site LV1 par rapport aux 2 autres sites LV, ce qui se retrouvait dans les données de metabarcoding, nous a conduit à l'éliminer pour réaliser une nouvelle ACP (Figure III-12 B). Ceci pourrait s'expliquer par le fait que le site LV1 est plus accessible et proche des activités humaines. Ainsi, il est possiblement beaucoup plus exposé à certains stress (pollution automobile, piétinement des randonneurs). Une nouvelle ACP a donc été réalisée sans le site LV1. L'axe 1 de la nouvelle ACP (expliquant 53% de la variance) permet toujours de séparer nettement les deux types de sol.

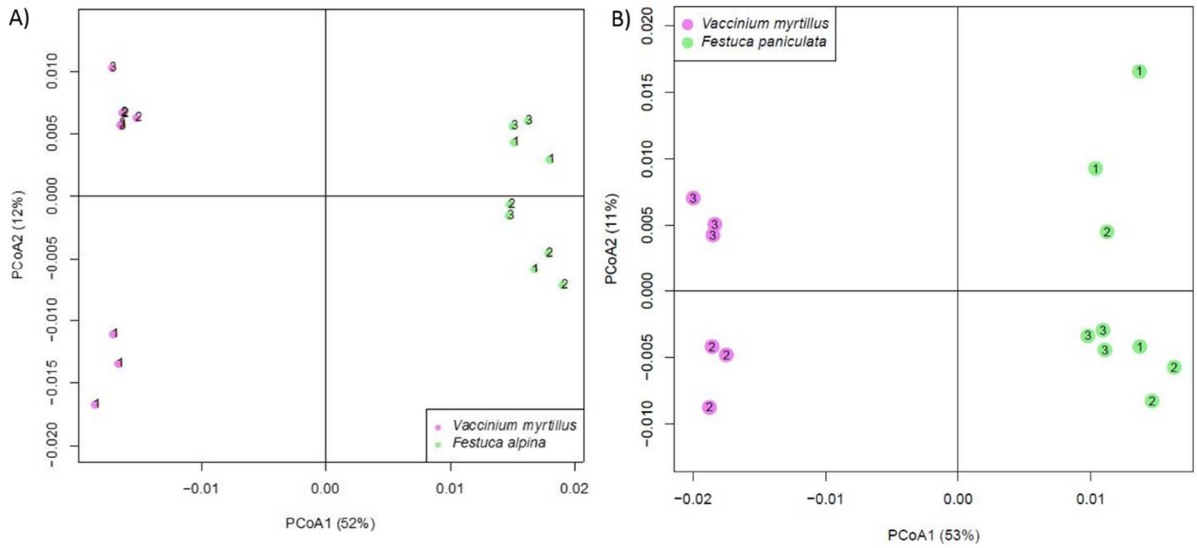


Figure III-12 : Analyse en composante principale réalisée sur l'ensemble des protéines ; A) avec tous les sites ; B) après élimination du site LV1

En vert, les échantillons PF ; en rose les échantillons LV. Les points « 1 » correspondent aux trois réplicas du site 1 (pour chacune des sols), les point « 2 » aux trois réplicas du site 2 et « 3 » aux trois réplicas du site 3.

Afin d'éviter de considérer des protéines pour lesquelles le site LV1 aurait possiblement un « poids » important dans la définition des axes de la première ACP, nous avons décidé de retenir les protéines définissant les axes de la deuxième ACP (sans LV1). Nos collaborateurs examinent actuellement cette liste de protéines.

Par ailleurs, toutes les protéines du jeu des données ont été annotées manuellement à partir des données BioCyc (https://biocyc.org/) et regroupées en 66 catégories fonctionnelles (en moyennant

leurs intensités). Des analyses statistiques ont ensuite été réalisées (en éliminant LV1) à l'aide du package « DESeq2 » et ont permis de mettre en évidence 10 catégories différentiellement exprimées entre les groupes, dont les variations sont présentées en Figure III-13 et en Tableau Annexe 6, page 195.

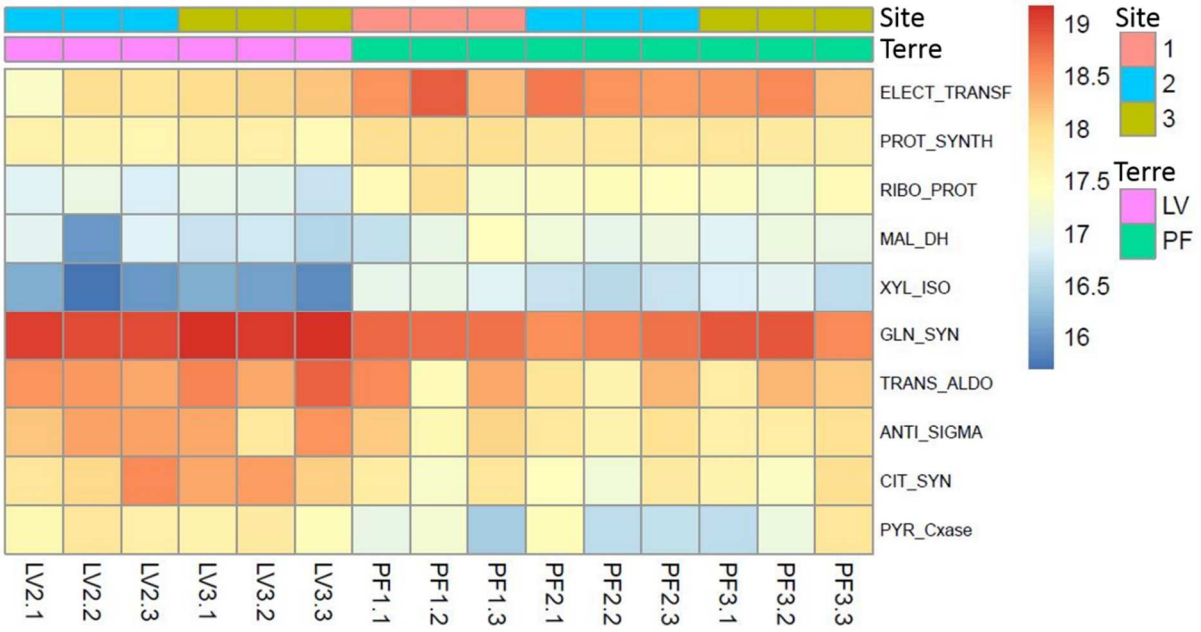


Figure III-13 : Carte (« *Heatmap* ») représentant les abondances moyennes des fonctions différentiellement exprimées entre les groupes.

Les niveaux d'expression sont présentés dans la barré dégradée : rouge (plus forte expression)-blanc-bleue (plus faible expression).

Les fonctions « Chaine respiratoire » (ELECT_TRANSF), « Synthèse protéique » (PROT_SYNTH), « protéines ribosomales » (RIBO_PROT), « Malate dehydrogenase » (MAL_DH), et « Xylose Isomerase » (XYL_ISO) sont globalement plus représentées dans les sols PF. Tandis que les fonctions « Glutamine Synthase » (GLN_SYN), « Transaldolase » (TRANS_ALDO), « Facteurs anti-sigma » (ANTI_SIGMA), « Citrate Synthase » (CIT_SYN), et « Pyruvate carboxylase » (PYR_Cxase) sont globalement plus représentées dans le sol LV.

Ces régulations au niveau des communautés des deux sols semblent refléter la différence de métabolisme des plantes les recouvrant :

- la Fétuque est plus « compétitive » que la Vaccinium, et a donc tendance à avoir un métabolisme accéléré. Les données protéomiques suggèrent que la communauté du sol PF est caractérisée par une chaîne respiratoire plus active.

-La Vaccinium, quant à elle, a tendance à faire de la biomasse avec le peu de ressources prélevées (NH_4). Les données protéomiques suggèrent que la communauté du sol LV semble adopter un fonctionnement similaire : le cycle de Krebs semble détourné en faveur du cétoglutarate qui est un précurseur de l'oxaloacetate, substrat de la glutamine synthase et « porte d'entrée » de l'ammonium dans les cellules.

Ces résultats vont par la suite permettre à nos collaborateurs de retraiter les données de méta-transcriptomique pour mieux comparer les deux types de sols, en focalisant leur attention sur les ARNm annotés pour ces mêmes 10 catégories fonctionnelles. Ces données vont également être complétées par de nouveaux prélèvements pour tester l'activité enzymatique des sols, en lien avec les 10 catégories fonctionnelles significativement différentes entre LV et PF.

Ces régulations sont actuellement examinées en détail par nos collaborateurs. A l'issue de ces examens, un article sera préparé pour publier ces résultats.

VIII. Conclusion

L'étude d'un métaprotéome est l'objet de nombreux développements, de par la singularité de tels échantillons.

Nous avons vu que, malgré nos efforts pour déterminer la banque de données la plus adaptée à notre échantillon, nous identifions un faible pourcentage de spectres (5,6% contre 55% chez la souris). En effet, le metabarcoding a permis d'identifier un grand nombre d'organismes, mais peut-être pas tous ; et parmi ces organismes détectés, de nombreux ne sont probablement pas encore séquencés ou s'ils le sont, la banque de données protéiques et certainement encore très incomplète. Cependant, c'est un phénomène que l'on retrouve également, bien que moindre, chez des espèces séquencées ; de manière générale, les banques de données restent incomplètes.
Dans une autre étude métaprotéomique de sols [200], plus de 30% des spectres étaient assignés mais cette expérience consistait à ajouter (« *spiker* ») aux échantillons de terre des extraits protéiques de cultures bactériennes et fongiques *connues* et *séquencées*. Dans le même article, sur des échantillons de terre non *spikés*, entre 6 et 24% seulement des spectres étaient assignés, suivant la méthode d'extraction utilisée (sachant que lorsque 24% sont assignés, en absolu il n'y a que très peu de spectres acquis). Dans une autre étude sur le microbiome intestinal [204], d'avantage de spectres sont assignés (≈ 30%) mais les auteurs montrent que la majorité des spectres assignés sont assignés à des protéines humaines, une minorité à des protéines bactériennes.
On retrouve donc les mêmes proportions faibles de spectres assignés que dans notre étude.

La Figure III-14, qui reprend toutes les données déposées dans PRIDE entre 2006 et 2015, montre que de manière générale le pourcentage de spectres identifiés est de plus en plus faible avec le temps. On voit que le nombre de spectres totaux acquis augmente avec une pente plus importante que le nombre de spectres identifiés. Cela s'explique par des instruments toujours plus rapides et performants qui peuvent acquérir plus de données. Il est possible que plus d'interférences soit analysées, mais une grande partie de ces spectres doivent tout de même provenir de la fragmentation de peptides. Dans un même temps, les banques de données se sont remplies mais pas suffisamment pour réduire l'écart, laissant un grand nombre de spectres non interprétés. Une autre explication aux spectres non interprétés est la présence possible de modifications post-traductionnelles (MPT). Si celles-ci ne sont pas spécifiées (composition, masse, acide aminé concerné) lors de la recherche, les peptides portant une MPT ne pourront être identifiés, du fait de la modification de masse qu'elles apportent. La plupart

du temps, seules quelques MPT sont recherchées (la plupart étant dues à la préparation d'échantillons, telle que la carbamidométhylation des cystéines), alors que plus de 500 sont connues [44].

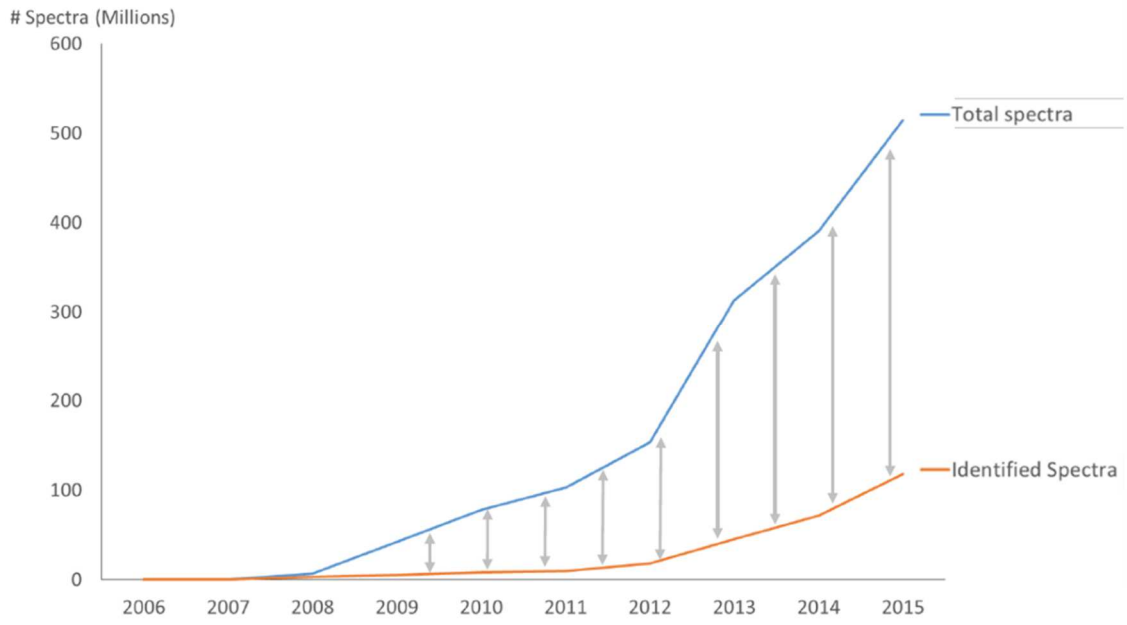


Figure III-14 : Evolution du nombre de spectres totaux acquis (en bleu) et du nombre de spectres identifiés (orange) obtenus entre 2006 et 2015 sur l'ensemble des données déposées dans PRIDE. Figure empruntée au Professeur Lennart Martens, Université de Gent.

Le pourcentage de spectres non identifiés devient de plus en plus important au cours du temps, ce qui est dû au double effet des spectromètres très performants et des banques de données incomplètes.

Cette étude met donc en évidence l'importance d'avoir une banque de données adaptée mais aussi leurs limitations. En effet, il reste encore beaucoup à faire en termes de remplissage de banque de données.

L'idéal aurait été de séquencer le métagénome étudié, puis de l'annoter pour en déduire une banque de données protéiques complète et personnalisée [75]. Malheureusement, au moment du projet, nos collaborateurs ne disposaient pas des ressources nécessaires. Malgré ces inconvénients, nos développements ont tout de même permis d'identifier un grand nombre de protéines. En effet, le fait d'avoir fait deux recherches consécutives, en ne conservant que les protéines identifiées et validées avec un très grand FDR de la première recherche pour la deuxième nous a permis d'augmenter considérablement le nombre de protéines identifiées avec un FDR communément accepté. Bien que cette méthode ne soit pas grandement utilisée dans la littérature, nous avons montré son bénéfice, comme décrit par Jagtap *et. al* [203]. Autrement dit, nous avons pu identifier de manière fiable près de 2000 protéines. Cela n'aurait pas été possible sans l'utilisation combinée du metabarcoding, de la métaprotéomique et de la métatranscriptomique. Ce projet met en avant l'avantage d'une étude multi-omique.

Conclusion générale et Perspectives

Conclusion générale

Au cours de cette thèse, nous avons mis en place des stratégies d'analyse protéomique quantitative appliquées à divers tissus clés du métabolisme, qui ont permis de mieux comprendre certains mécanismes adaptatifs de diverses espèces ; en allant du modèle « classique » de la souris de laboratoire à des modèles plus exotiques tels que le campagnol ou la fourmi, qui, de par un trait particulier, nous ont aidé à répondre à des questions de biologie évolutive non résolues par la seule étude de modèles classiques.

L'analyse d'un nouveau type d'échantillons, ou d'un nouvel organisme nécessite des développements de la préparation des échantillons au traitement bio-informatique des données. Nous avons confirmé que préfractionner les protéines permet d'augmenter (2 fois chez la souris) la couverture du protéome analysé. Il faut néanmoins rester vigilant et considérer que lorsque le nombre d'échantillons à analyser est très grand (projet campagnol) ceci risque d'introduire des biais techniques. Dans ce cas, et en particulier lorsque le génome de l'organisme étudié n'est pas encore séquencé, ces travaux de thèse ont également été l'occasion de montrer que la stratégie de séquençage *de novo* permet de compenser au moins en partie l'absence de décomplexification, en augmentant le nombre de protéines analysées.

L'optimisation du paramétrage du logiciel PepNovo et de l'alignement de séquence MS BLAST sur un mélange protéique connu nous a permis de transposer les résultats sur les échantillons inconnus de BAT (campagnol) afin d'obtenir des identifications fiables et justes. Nous avons également mis en place une stratégie quantitative sur ces peptides avec le logiciel Skyline, avec par exemple dans le projet campagnol 12% de protéines quantifiées supplémentaires et une quantification plus robuste (davantage de peptides quantifiés par protéine) pour 28% des protéines quantifiées.

Enfin, en adaptant l'étape d'extraction des protéines nous avons analysé un très grand nombre de protéines issues des sols alpins (métaprotéome). La multitude d'organismes présents a requis l'utilisation combinée de la génomique et de la protéomique pour établir une banque de données « personnalisée », la mieux adaptée possible aux échantillons ; la validation des données a pu être réalisée grâce à une double recherche par empreinte de fragments peptidiques puis la transcriptomique a permis de renforcer la confiance dans nos identifications.

Le schéma ci-dessous résume les différentes étapes de l'analyse protéomique quantitative (en « label-free » XIC MS1) développées au cours de ces travaux de thèse, en fonction de l'organisme étudié.

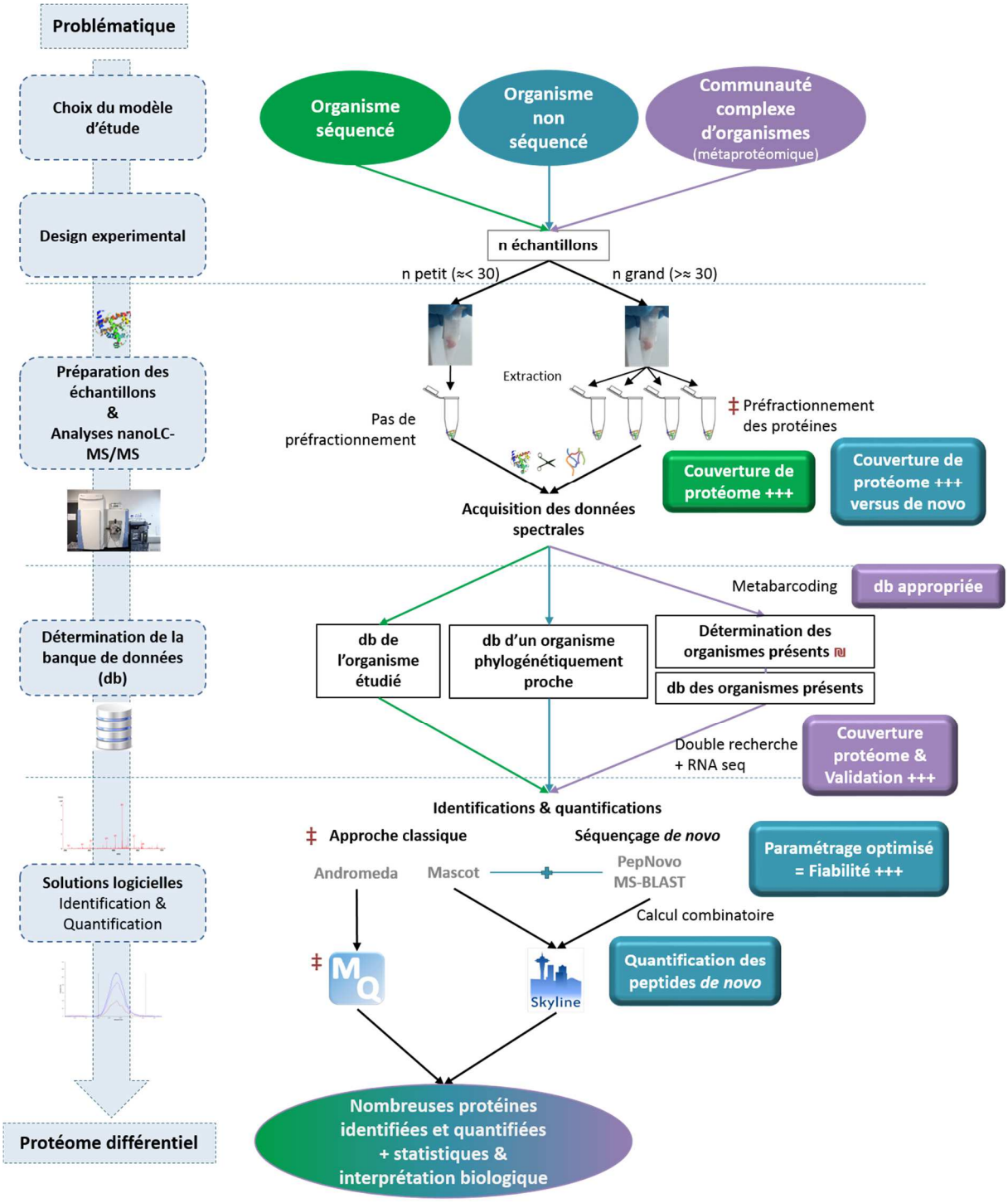


Schéma résumant les différentes étapes de l'analyse protéomique quantitative (« label-free » XIC MS1) développées au cours de ces travaux de thèse.

Les apports et optimisations liés à chaque chapitre (vert pour le chapitre I, bleu pour le chapitre II et violet pour le chapitre III) sont présentés dans les cadres en relief sur la droite de la figure.

Les flèches vertes sont relatives aux organismes séquencés ; les bleues aux organismes non séquencés ; et les violettes aux communautés complexes d'organismes.

₪ : Pour les organismes non séquencées et les communautés complexes (métaprotéomes), il est également possible de séquencer le génome ou transcriptome, ce qui n'a pas été étudié ici.

‡ : préfractionner les protéines impose d'utiliser le logiciel MaxQuant.

Perspectives

Nous avons donc vu que l'approche protéomique permet de répondre à diverses questions biologiques en permettant d'identifier et quantifier un grand nombre de protéines de manière fiable et robuste. Il existe de nombreuses méthodes, permettant de traiter différentes problématiques (quantification ciblée ou globale, absolue ou relative…). Malgré les fortes avancées des dernières années, la protéomique souffre toujours de certaines limitations :

- Il n'existe pas de méthode universelle qui combine tous les avantages. L'analyse protéomique est question de compromis : compromis entre rapidité, sensibilité et résolution ; compromis entre profondeur du protéome analysé et nombre d'analyses ; compromis entre nombre d'analyses et nombre de réplicas ; compromis entre nombre de protéines quantifiés et sélectivité/sensibilité de la méthode de quantification [205] …

- Le traitement des données reste un point complexe et pas toujours automatisable de l'analyse protéomique. Il est essentiel de toujours adapter et si nécessaire développer des solutions bio-informatiques à chaque cas particulier ; les outils bio-informatique sont indispensables en protéomique [206]. Cette étape peut s'avérer chronophage mais est indispensable pour obtenir des données de qualité et fiables. Acquérir des données d'excellente qualité avec des instruments très performants ne sert à rien si la bonne stratégie n'est pas utilisée pour les traiter. De plus, ces données sont conséquentes et nécessitent des ressources informatiques importantes ainsi que des logiciels adaptés pour les manipuler.

- Enfin, le remplissage des banques de données reste également aujourd'hui un point limitant de l'analyse protéomique. Les milliers de spectres acquis ne peuvent pas tous être assignés, même chez une espèce dont le génome est séquencé (comme la souris), bien que le taux de spectres identifiés et validés soit largement plus satisfaisant que pour une espèce non séquencée ou encore un métaprotéome, à peine plus de la moitié des spectres sont interprétés [40]. Il reste donc un gros effort à fournir pour remplir les banques de données protéiques, même si elles sont régulièrement mise à jour pour être corrigées ou complétées [51].

Finalement, ce travail illustre la nécessité de développer le traitement bio-informatiques des données pour qu'il soit le plus adapté à la question posée, tout en tenant compte des contraintes individuelles de chaque échantillon/organisme étudié/application.

Partie expérimentale

Partie expérimentale

Chapitre I : Analyse protéomique quantitative chez un organisme séquencé

I. Etude des marqueurs spléniques de l'immunosénescence

A. Préparation d'échantillons

Les biopsies de rates congelées ont été broyées mécaniquement (MM400, Retsch) et les protéines extraites dans un tampon d'extraction (7M urée, 2M thiourée, 4% CHAPS, 0.005% TLCK, inhibiteur de protéases). Après sonication à une amplitude de 10%, 3*10 sec sur glace (Digital Cell Disruptor), les protéines ont été précipitées dans 7 volumes d'acétone toute la nuit à -20°C. Le culot de protéines a ensuite été solubilisé dans le tampon Laemmli (10mM Tris pH 6.8, 1mM EDTA, 5% βME, 5% SDS, 10% glycérol) pour le gel. Un dosage RC-DC a été réalisé pour déterminer la concentration protéique.

A.1. Protocole 1 sans préfractionnement

20 µg de protéines ont été déposés sur gel SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. La migration a eu lieu pendant 39 min à 50V, les protéines ont ainsi été concentrées à l'entrée du gel de séparation. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). La bande protéique où les protéines se sont retrouvées concentrées a ensuite été découpée.

A.2. Protocole 2 avec préfractionnement

20 µg de protéines ont été déposés sur un SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. La migration a eu lieu pendant 75 min à 50V, les protéines ont ainsi été séparées sur 12mm. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). Six bandes protéiques de 2mm ont ensuite été découpées.

Pour les deux protocoles, la suite était identique :
Les bandes protéiques ont été traitées automatiquement à l'aide d'une station « MassPrep» (Waters). Une première étape a permis de décolorer les bandes (3 cycles de lavage avec 50µL de bicarbonate d'ammonium 25mM et 50µL d'acétonitrile, 10 min) ; puis les bandes ont été déshydratées (50µL d'acétonitrile, 60°C, 5min) ; les ponts disulfures des protéines ont été réduits (50µL dithiotréithol – DTT

– 10mM, 30min, 60°C) puis alkylés (50µL iodoacétamide – IAA –25mM, 30mM) ; puis un dernier cycle de lavage et enfin les bandes ont été déshydratées avec 50µL d'acétonitrile 15min.

Finalement, les protéines ont été digérées dans le gel par la trypsine (1%). A l'issue d'une incubation à 37°C toute la nuit, une solution à 60% acétonitrile, 0.1% d'acide formique a été ajoutée à chaque bande pour stopper la digestion et extraire les peptides du gel. Une deuxième extraction a été réalisée avec de l'acétonitrile 100%. Le solvant a été évaporé et les peptides repris dans une solution à 1% acétonitrile, 0,1% acide formique, de sorte à obtenir une concentration de 100 ng/µL.

Le protocole 2 a finalement été retenu pour l'analyse des 24 échantillons.

B. Analyse nanoLC-MS/MS

B.1. Analyse des échantillons tests

Les échantillons « tests » (i.e. avec ou sans préfractionnement) ont été analysés sur un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un analyseur de type temps de vol (maXis 4G, Bruker daltonics). Le couplage était contrôlé par les logiciels Bruker compass Hystar (v3.2) et OtofControl (Rev 3.4, Bruker daltonics GmbH).

1µL de chaque échantillon a été chargé sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450nL/min avec les gradients suivants : 1-5% de B en 1min puis 5-35% de B en 179min pour l'échantillon préparé sans préfractionnement ; 1-5% de B en 1min puis 5-35% de B en 59min pour chaque bande de l'échantillon préfractionné en 6.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 1.

Tableau Partie Exp. 1 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	100 – 2500 m/z
Fragmentation	Sélection	5 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	100 – 2500 m/z

B.2. Analyse des échantillons

Les échantillons ont été analysés sur un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un analyseur de type temps de vol (Impact HD, Bruker daltonics). Le couplage était contrôlé par les logiciels Bruker compass Hystar (v3.2) et OtofControl (Rev 3.4, Bruker daltonics GmbH).

3µL de chaque échantillon ont été chargés sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450nL/min avec le gradient suivant : 1-5% de B en 1min puis 5-35% de B en 59min.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 2.

Tableau Partie Exp. 2 : Paramètres d'acquisition du spectromètre de masse.

MS	Vitesse de balayage	500ms (2Hz)
	Gamme de masse	150 – 2200 m/z
Fragmentation CID	Temps de Cycle	3s
	Etats de charge	+2 à +5
	Cycle d'exclusion	1 min ou 1 spectre
MS/MS	Vitesse de balayage	Entre 4Hz et 25Hz, en fonction de l'intensité du précurseur
	Gamme de masse	300 – 2200 m/z

C. Analyse des données

C.1. Analyse des données des échantillons tests

L'algorithme de recherche mascot v2.5.1 (Matrix Science) [45] a été utilisé pour identifier les protéines à partir des données MS/MS, en utilisant une banque de données contenant les séquences protéiques des souris (*Mus Musculus* taxonomie 10090, Swissprot Mars 2014, 16 786 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a été générée afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications fixes et variables respectivement, l'erreur de masse tolérée était de 15 ppm en mode MS et 0.02 Da en mode MS/MS. Le logiciel Scaffold v3 (Proteome Software) a été utilisé pour valider les identifications avec les paramètres suivants : score d'ion supérieur à 25, différence entre le score d'ions et le score d'identité supérieur à 5 ; de sorte à obtenir un FDR inférieur à 1%.

C.2. Analyse des données des échantillons

Identification des protéines :

L'algorithme de recherche Andromeda (Maxquant v1.5.2.8) [46] a été utilisé pour identifier des protéines à partir des données MS/MS. La recherche a été réalisée dans une banque de données protéique restreinte à la taxonomie « souris » (Swissprot novembre 2014, taxonomie 10090 *Mus musculus*, 16 688 entrées contaminants inclus). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a alors été générée par Andromeda, afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche Andromeda étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications fixes et variables respectivement, l'erreur de masse tolérée était de 10 ppm en mode MS et 40ppm en mode MS/MS, les taux de faux positif (FDR) peptidique et protéique ont été fixés à 1%.

Quantification des protéines :

A partir des données spectrales (MS) d'un peptide donné, sa quantification a été obtenue après extraction puis intégration de la somme des courants d'ions générés. Pour cela, le logiciel Maxquant v1.5.2.8 a été utilisé. Seuls les peptides uniques ont été considérés pour la quantification, l'option « Label-Free Quantification » (LFQ) de Maxquant [114] a également permis de normaliser les intensités peptidiques et protéiques.

II. Etude évolutive de l'inflammation

A. Préparation d'échantillons

Les biopsies de rates congelées ont été broyées mécaniquement (MM400, Retsch) et les protéines extraites dans un tampon d'extraction (8M urée, 2M thiourée, 0,1M bicarbonate d'ammonium, 1% DTT, inhibiteur de protéases). Après sonication à une amplitude de 10%, 3*10 sec sur glace (Digital Cell Disruptor), les protéines ont été précipitées dans 5 volumes d'acétone toute la nuit à -20°C. Après centrifugation (13 500g, 20min, 4°C), le surnageant a été éliminé et le culot de protéines solubilisé dans un tampon contenant 8M urée, 0,1M bicarbonate d'ammonium (compatible avec la digestion liquide). Un dosage de Bradford a été réalisé pour déterminer la concentration protéique de chaque échantillon.

A ce stade, 20µg de chaque échantillon ont été regroupés en un pool (contrôle qualité), ensuite fractionné en quatre réplicas identiques. Ce pool a ensuite été utilisé comme contrôle qualité des analyses nanoLC-MS/MS (injections répétées du même échantillon tout au long de l'analyse).

Les ponts disulfures des protéines (l'équivalent de 50µg) ont été réduits au DTT (700mM, 30min, 37°C) puis alkylés à l'IAA (700mM, 1h dans l'obscurité, température ambiante), les échantillons ont alors été dilués dans du bicarbonate d'ammonium pour réduire la concentration en urée à 1M. Enfin, pour 500ng de trypsine ont été ajoutés à chacun des échantillons. A l'issue d'une incubation à 37°C toute la nuit, la digestion a été stoppée avec 10µL d'acide formique. Pour éliminer l'urée et concentrer les peptides, une purification par extraction en phase solide (SPE) sur colonne C18 (Sep-Pak Vac 1cc, Waters) a été réalisée : lavage 2mL méthanol, 2mL d'acétonitrile ; conditionnement 3mL d'eau 0,1%

acide formique ; lavage des peptides 3mL eau 0,1% acide formique ; élution des peptides 600µL acétonitrile 60%, acide formique 0,1%. Le solvant a ensuite été évaporé et les peptides repris dans 334µl d'une solution à 1% acétonitrile et 0,1% acide formique.

B. Analyse nanoLC-MS/MS

L'analyse a été réalisée en utilisant un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un quadripole en tandem avec un analyseur Orbitrap (Q-Exactive plus, Thermo Scientific). Le couplage était contrôlé par Thermo Xcalibur v3.0.63.

3µL de chaque échantillon ont été chargés sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450nL/min avec le gradient suivant : 1-5% de B en 2min puis 5-29% de B en 113min et enfin 29-35% de B en 5min.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 3.

Tableau Partie Exp. 3 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	300 – 1800 m/z
	AGC (Auto Gain Control)	3×10^6
	Maximum injection time	50ms
Fragmentation	Sélection	10 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	17 500
	Gamme de masse	A partir de 100m/z jusqu'à 15 fois la plus petite masse observée
	AGC (Auto Gain Control)	1×10^5
	Maximum injection time	100ms

C. Analyse des données

Identification des protéines :

L'algorithme de recherche Andromeda (Maxquant v1.5.3.30) [46] a été utilisé pour identifier des protéines à partir des données MS/MS. La recherche a été réalisée dans une banque de données protéique restreinte à la taxonomie « souris » (Swissprot janvier 2017, taxonomie 10090 « Mus musculus », 16 754 entrées). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a alors été générée par Andromeda, afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche

Andromeda étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications fixes et variables respectivement, l'erreur de masse tolérée était de 4,5 ppm en mode MS et 40ppm en mode MS/MS, les taux de faux positif (FDR) peptidique et protéique ont été fixés à 1%.

Quantification des protéines :

A partir des données spectrales (MS) d'un peptide donné, sa quantification a été obtenue après extraction puis intégration de la somme des courants d'ions générés. Pour cela, le logiciel Maxquant v1.5.3.30 a été utilisé. Seuls les peptides uniques ont été considérés pour la quantification, l'option « Label-Free Quantification » (LFQ) de Maxquant [114] a également permis de normaliser les intensités peptidiques et protéiques.

Chapitre II : Analyse protéomique quantitative chez des organismes non séquencés

I. Développements méthodologiques pour l'identification et la quantification fiable des peptides séquencés de novo

A. Préparation d'échantillons

Les biopsies tissus adipeux bruns congelés ont été broyées mécaniquement (MM400, Retsch) et les protéines extraites dans un tampon d'extraction (8M urée, 0,1M bicarbonate d'ammonium, 1% DTT, inhibiteur de protéases). Après sonication à une amplitude de 10%, 3*10 sec sur glace (Digital Cell Disruptor), les protéines ont été précipitées dans 6 volumes d'acétone toute la nuit à -20°C. Après centrifugation (13 500g, 20min, 4°C), le surnageant a été éliminé et le culot de protéines solubilisé dans un tampon 8M urée, 0,1M bicarbonate d'ammonium. Un dosage Bradford a été réalisé pour déterminer la concentration protéique.

A ce stade, 20µg de chaque échantillon ont été regroupés en un pool (contrôle qualité), ensuite fractionné en quatre réplicas identiques. Ce pool a ensuite été utilisé comme contrôle qualité des analyses nanoLC-MS/MS (injections répétées du même échantillon tout au long de l'analyse).

Les ponts disulfures des protéines (l'équivalent de 50µg) ont été réduits au DTT (700mM, 30min, 37°C) puis alkylés à l'IAA (700mM, 1h dans l'obscurité, température ambiante), les échantillons ont alors été dilués dans du bicarbonate d'ammonium pour réduire la concentration en urée à 1M. Enfin, pour 500ng de trypsine ont été ajoutés à chacun des échantillons. A l'issue d'une incubation à 37°C toute la nuit, la digestion a été stoppée avec 10µL d'acide formique. Pour éliminer l'urée et concentrer les peptides, une purification par extraction en phase solide (SPE) sur colonne C18 (Sep-Pak Vac 1cc, Waters) a été réalisée : lavage 2mL méthanol, 2mL d'acétonitrile ; conditionnement 3mL d'eau 0,1% acide formique ; lavage des peptides 3mL eau 0,1% acide formique ; élution des peptides 600µL acétonitrile 60%, acide formique 0,1%. Le solvant a ensuite été évaporé et les peptides repris dans 166,7 µL d'une solution à 1% acétonitrile et 0,1% acide formique.

B. Analyse nanoLC-MS/MS

L'analyse a été réalisée en utilisant un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un quadripole en tandem avec un analyseur Orbitrap (Q-Exactive plus, Thermo Scientific). Le couplage était contrôlé par Thermo Xcalibur v3.0.63.

2µL de chaque échantillon ont été chargés sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters)

maintenue à 60°C, à un débit de 450nL/min avec le gradient suivant : 1-5% de B en 2min puis 5-35% de B en 118min.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 4.

Tableau Partie Exp. 4 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	300 – 1800 m/z
	AGC (Auto Gain Control)	3×10^6
	Maximum injection time	50ms
Fragmentation	Sélection	10 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	17 500
	Gamme de masse	A partir de 100m/z jusqu'à 15 fois la plus petite masse observée
	AGC (Auto Gain Control)	1×10^5
	Maximum injection time	100ms

C. Analyse des données

Identification des protéines :
L'algorithme de recherche Mascot v2.5.1 (Matrix Science) [45] a été utilisé pour identifier les protéines à partir des données MS/MS, en utilisant une banque de données contenant les séquences protéiques des mammifères (*Mammalia*, taxonomie 40674, SwissProt août 2015, 66 448 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Les paramètres de recherche Mascot étaient les suivants: un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 5ppm en mode MS et de 0.05Da en mode MS/MS. les identifications validées avec un taux de faux positif inférieur à 1% grâce au logiciel Scaffold v.4.3 (Proteome Software).

Avant de réaliser le séquençage *de novo*, les données ont été triées grâce au module Recover de la suite logicielle MSDA [55] : tous les spectres déjà identifiés par Mascot ont été éliminés puis les filtres suivants ont été appliqués aux spectres restants : E = 9 et UPN = 9. Le logiciel PepNovo+ v3.1 [57] (intégré à la suite MSDA) a ensuite été utilisé pour interpréter les spectres par séquençage *de novo* avec les paramètres suivants : l'erreur de masse tolérée était de 0,05 Da pour le précurseur et 0,02 Da pour les fragments, l'enzyme spécifiée était la trypsine, la carbamidomethylation des cystéines et l'oxydation des méthionines spécifiées comme modifications variables, 7 séquences envoyées automatiquement à MS Blast grâce à la suite logicielle MSDA. La même banque de données (*Mammalia*) que pour la recherche classique a été utilisée pour la comparaison par homologie de

séquence. Seuls les peptides pour lesquels on retrouvait une similarité de séquence pour 6 résidus consécutifs (avec une erreur tolérée) ont été considérés.

Quantification des protéines :
Le logiciel Skyline v3.1 (MacCoss Lab, UW) [111] a été utilisé pour l'extraction et l'intégration des courants d'ions générés par les trois isotopes de l'ion précurseur (P, P+1, P+2).
Les données ont ensuite été normalisées par la méthode des quantiles à l'aide du package R (v3.0.3 ; http://www.R-project.org) nommé Normalyzer [116]. Finalement, la corrélation des profils d'abondance des peptides par protéine a été vérifié en calculant un coefficient de Pearson, comme expliqué en page 92.

II. Détermination de l'apport du séquençage de novo vs. celui d'un préfractionnement protéique

A. Etude de la modification du protéome chez une espèce saisonnière

A.1. Préparation d'échantillons

Les biopsies de foie congelé ont été broyées mécaniquement (MM400, Retsch) et les protéines extraites dans un tampon d'extraction (8M urée, 2M thiourée, 0,1M bicarbonate d'ammonium, 2% DTT, 2% CHAPS inhibiteur de protéases). Après sonication à une amplitude de 10%, 3*10 sec sur glace (Digital Cell Disruptor), les protéines ont été précipitées dans 6 volumes d'acétone toute la nuit à -20°C. Après centrifugation (13 500g, 20min, 4°C), le surnageant a été éliminé et le culot de protéines solubilisé dans un tampon Laemmli (10mM Tris pH 6.8, 1mM EDTA, 5% βME, 5% SDS, 10% glycérol) pour le gel). Un dosage RC-DC a été réalisé pour déterminer la concentration protéique.

A.1.1. Protocole sans préfractionnement

50 µg de protéines ont été chargés sur gel SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. La migration a eu lieu pendant 34 min à 50V, les protéines ont ainsi été concentrées à l'entrée du gel de séparation. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). La bande protéique où les protéines se sont retrouvées concentrées a ensuite été découpée.

A.1.2. Protocole avec préfractionnement

50 µg de protéines ont été déposés sur gel SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. La migration a eu lieu pendant 23 min à 50V, puis 24 min à 100V, les protéines ont ainsi été séparées sur 12mm. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). Six bandes protéiques de 2mm ont ensuite été découpées.

Pour les deux protocoles, la suite était identique :
Les bandes protéiques ont été traitées automatiquement à l'aide d'une station « MassPrep» (Waters). Une première étape a permis de décolorer les bandes (3 cycles de lavage avec 50µL de bicarbonate d'ammonium 25mM et 50µL d'acétonitrile, 10 min) ; puis les bandes ont été déshydratées (50µL d'acétonitrile, 60°C, 5min) ; les ponts disulfures des protéines ont été réduits (50µL DTT 10mM, 30min, 60°C) puis alkylés (50µL IAA 25mM, 30mM) ; puis un dernier cycle de lavage et enfin les bandes ont été déshydratées avec 50µL d'acétonitrile 15min.

Finalement, les protéines ont été digérées dans le gel par la trypsine (1%). A l'issue d'une incubation à 37°C toute la nuit, une solution à 60% acétonitrile, 0.1% d'acide formique a été ajoutée à chaque bande pour stopper la digestion et extraire les peptides du gel. Une deuxième extraction a été réalisée avec de l'acétonitrile 100%. Le solvant a été évaporé et les peptides repris dans une solution à 1% acétonitrile, 0,1% acide formique, de sorte à obtenir une concentration de 200 ng/µL.

Le protocole avec préfractionnement a finalement été retenu pour l'analyse des 10 échantillons.

A.2. Analyse nanoLC-MS/MS

A.2.1. Analyse des échantillons tests

Les échantillons tests (i.e. avec ou sans préfractionnement) ont été analysés sur un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un analyseur de type temps de vol (Impact HD, Bruker daltonics). Le couplage était contrôlé par les logiciels Bruker compass Hystar (v3.2) et OtofControl (Rev 3.4, Bruker daltonics GmbH).

1,5 µL de chaque échantillon a été chargé sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450nL/min avec les gradients suivants : 1-5% de B en 1min puis 5-30% de B en 175min en enfin 30-35% de B en 3min pour l'échantillon préparé sans préfractionnement ; et 1-5% de B en 1min puis 5-30% de B en 54min et enfin 30-35% de B en 5min pour chaque bande de l'échantillon préfractionné en 6.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 5.

Tableau Partie Exp. 5 : Paramètres d'acquisition du spectromètre de masse.

MS	Vitesse de balayage	500ms (2Hz)
	Gamme de masse	150 – 2200 m/z
Fragmentation CID	Temps de Cycle	3s
	Etats de charge	+2 à +5

	Cycle d'exclusion	1 min ou 1 spectre
MS/MS	Vitesse de balayage	Entre 4Hz et 25Hz, en fonction de l'intensité du précurseur
	Gamme de masse	300 – 2200 m/z

A.2.2. Analyse des échantillons

Les échantillons ont été analysés sur un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source d'ionisation de type électrospray et d'un quadripole en tandem avec un analyseur Orbitrap (Q-Exactive plus, Thermo Scientific). Le couplage était contrôlé par Thermo Xcalibur v3.0.63.

1,5 µL de chaque échantillon ont été chargés sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450nL/min avec le gradient suivant : 1-5% de B en 1min puis 5-30% de B en 54min et enfin 30-35% de B en 5min.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 6.

Tableau Partie Exp. 6 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	300 – 1800 m/z
	AGC (Auto Gain Control)	3×10^6
	Maximum injection time	50ms
Fragmentation	Sélection	10 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	17 500
	Gamme de masse	A partir de 100m/z jusqu'à 15 fois la plus petite masse observée
	AGC (Auto Gain Control)	1×10^5
	Maximum injection time	100ms

A.3. Analyse des données

A.3.1. Analyse des données des échantillons tests

A.3.1.1. Protocole sans préfractionnement

Identification des protéines :
L'algorithme de recherche Mascot v2.5.1 (Matrix Science) [45] a été utilisé pour identifier les protéines à partir des données MS/MS, en utilisant une banque de données contenant les séquences protéiques

de la taxonomie *Homo Sapiens* (taxonomie 9606, SwissProt décembre 2015, 20 195 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Les paramètres de recherche Mascot étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 5ppm en mode MS et de 0.05Da en mode MS/MS. Les identifications validées avec un taux de faux positif inférieur à 1% grâce au logiciel Scaffold v.4.3 (Proteome software).

Avant de réaliser le séquençage *de novo*, les données ont été triées grâce au module Recover de la suite logicielle MSDA [55]: tous les spectres déjà identifiés par Mascot ont été éliminés puis les filtres suivants ont été appliqués aux spectres restants : E = 4 et UPN = 5. Le logiciel PepNovo+ v3.1 [57] intégré dans MSDA a ensuite été utilisé pour interpréter les spectres par séquençage *de novo* avec les paramètres suivants : l'erreur de masse tolérée était de 0,05 Da pour le précurseur et 0,02 Da pour les fragments, l'enzyme spécifiée était la trypsine, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, 7 séquences envoyées automatiquement à MS Blast grâce à la suite logicielle MSDA. La même banque de données *Homo Sapiens* que pour la recherche classique spécifiée pour la comparaison par homologie de séquence a été utilisée ici, et seuls les peptides pour lesquels on retrouvait une similarité de séquence pour 6 résidus consécutifs au minimum (avec une erreur tolérée) ont été considérés.

A.3.1.2. Protocole avec préfractionnement

Identification des protéines :
L'algorithme de recherche mascot v2.5.1 (Matrix Science) [45] a été utilisé pour identifier les protéines à partir des données MS/MS, dans une banque de données contenant les séquences protéiques de la taxonomie *Homo Sapiens* (taxonomie 9606, SwissProt décembre 2015, 20 195 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Les paramètres de recherche Mascot étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 5ppm en mode MS et de 0.05Da en mode MS/MS. Grâce au logiciel Scaffold v.4.3 (Proteome Software), les résultats de la recherche des 6 fractions ont été fusionnés et les identifications validées avec un taux de faux positif inférieur à 1%.

A.3.2. Analyse des données des échantillons

Identification des protéines :
L'algorithme de recherche Andromeda (Maxquant v1.5.3.30) [46] a été utilisé pour identifier des protéines à partir des données MS/MS. La recherche a été réalisée dans une banque de données protéiques restreinte aux séquences *Homo Sapiens* (taxonomie 9606, SwissProt avril 2016, 20 195 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a alors été générée par Andromeda, afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche Andromeda étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par

peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 4,5 ppm en mode MS et 20ppm en mode MS/MS, les taux de faux positif (FDR) peptidique et protéique ont été fixés à 1%.

Quantification des protéines :
A partir des données spectrales (MS) d'un peptide donné, sa quantification a été obtenue après extraction puis intégration de la somme des courants d'ions générés. Pour cela, le logiciel Maxquant v1.5.3.30 a été utilisé. Seuls les peptides uniques ont été considérés pour la quantification, l'option « Label-Free Quantification » (LFQ) de Maxquant [114] a également permis de normaliser les intensités peptidiques et protéiques.

B. Etude des interactions entre rôle social et variabilité de l'espérance de vie chez la fourmi

B.1. Préparation d'échantillons

Les fourmis congelées ont été broyées mécaniquement (MM400, Retsch) et les protéines extraites dans un tampon d'extraction (8M urée, 2M thiourée, 0,1 M bicarbonate d'ammonium, 1% DTT, inhibiteur de protéases). Les échantillons ont été soniqués à une amplitude de 10%, 3*10 sec sur glace (Digital Cell Disruptor), puis centrifugés à 2000g pendant 2min pour éliminer les cuticules (insolubles), le surnageant a ensuite été récupéré et les protéines précipitées dans 8 volumes d'acétone, toute la nuit à -20°C. Après centrifugation (13 500g, 20min, 4°C), le surnageant a été éliminé et le culot de protéines solubilisé dans le tampon Laemmli (10mM Tris pH 6.8, 1mM EDTA, 5% βME, 5% SDS, 10% glycérol) en vue de la migration sur gel. Un dosage RC-DC a été réalisé pour déterminer la concentration protéique.

A ce stade, 12µg de protéines de chaque échantillon ont été regroupés en un pool (contrôle qualité), ensuite fractionné en quatre réplicas identiques. Ce pool a ensuite été utilisé comme contrôle qualité des analyses nanoLC-MS/MS (injection répétées du même échantillon tout au long de l'analyse).

20µg de protéines ont été chargées sur gel SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. La migration a eu lieu pendant 60 min à 50V, puis 15min à 100V, les protéines ont ainsi migré sur 10mm. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). Cinq bandes protéiques de 2mm ont ensuite été découpées. Les bandes ont été traitées automatiquement à l'aide d'une station « MassPrep» (Waters). Une première étape a permis de décolorer les bandes (3 cycles de lavage avec 50µL de bicarbonate d'ammonium 25mM et 50µL d'acétonitrile, 10 min) ; puis les bandes ont été déshydratées (50µL d'acétonitrile, 60°C, 5min) ; les ponts disulfures des protéines ont été réduits (50µL DTT 10mM, 30min, 60°C) puis alkylés (50µL IAA 25mM, 30mM) ; puis un dernier cycle de lavage a été réalisé et enfin les bandes ont été déshydratées avec 50µL d'acétonitrile pendant 15min.

Finalement, les protéines ont été digérées dans le gel par la trypsine (1%). A l'issue d'une incubation à 37°C toute la nuit, une solution à 60% acétonitrile, 0.1% d'acide formique a été ajoutée à chaque bande pour stopper la digestion et extraire les peptides du gel. Une deuxième extraction a été réalisée avec de l'acétonitrile 100%. Le solvant a été évaporé et les peptides repris dans 13,3µL d'une solution contenant 1% acétonitrile, 0,1% acide formique avant d'être analysés en spectrométrie de masse.

B.2. Analyse nanoLC-MS/MS

L'analyse a été réalisée en utilisant un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source nanoSpray et d'un quadripole en tandem avec un analyseur orbitrap (Q Exactive+, Thermo Scientific). Le couplage était contrôlé par Thermo Xcalibur v3.0.63.

1µL de chaque échantillon a été chargé sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5 µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450 nL/min avec le gradient suivant : 1-6% de B en 0,5 min puis 6-35% de B en 60 min.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 7.

Tableau Partie Exp. 7 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	300 – 1800 m/z
	AGC (Auto Gain Control)	3×10^6
	Maximum injection time	50ms
Fragmentation	Sélection	10 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	17500
	Gamme de masse	A partir de 100m/z jusqu'à 15 fois la plus petite masse observée
	AGC (Auto Gain Control)	1×10^5
	Maximum injection time	100ms

B.3. Analyse des données

Identification des protéines :

L'algorithme de recherche Andromeda (Maxquant v1.5.3.30) [46] a été utilisé pour identifier des protéines à partir des données MS/MS. La recherche a été réalisée dans une banque de données protéiques restreinte aux séquences de fourmis noires des jardins (taxonomie 67767 « Lasius Niger », Uniprot mars 2017, 18075 entrées). Les séquences des contaminants usuels, incluant la trypsine et la

kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a alors été générée par Andromeda, afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche Andromeda étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 4,5 ppm en mode MS et 20ppm en mode MS/MS, les taux de faux positif (FDR) peptidique et protéique ont été fixés à 1%.

Quantification des protéines :

A partir des données spectrales (MS) d'un peptide donné, sa quantification a été obtenue après extraction puis intégration de la somme des courants d'ions générés. Pour cela, le logiciel Maxquant v1.5.3.30 a été utilisé. Seuls les peptides uniques ont été considérés pour la quantification, l'option « Label-Free Quantification » (LFQ) de Maxquant [114] a également permis de normaliser les intensités peptidiques et protéiques.

Chapitre III : Analyse méta-protéomique quantitative chez une communauté complexe d'organismes

I. Préparation d'échantillons

A. Protocole 1

Les échantillons de terre congelée ont été broyés mécaniquement (MM400, Retsch) et les protéines ont été extraites dans un tampon d'extraction (5% SDS, 50mM Tris pH 8.5, 0.1mM EDTA, 1mM $MgCl_2$, 50mM DTT). Après centrifugation à 2000g pendant 10min, le culot a été éliminé et à partir du surnageant les protéines ont été précipitées avec 25% d'acide trichloroacétique (TCA) toute la nuit à 4°C. Après centrifugation (13 500g, 20min, 4°C), le culot de précipitation a été récupéré et suspendu dans un tampon Laemmli (10mM Tris pH 6.8, 1mM EDTA, 5% βME, 5% SDS, 10% glycérol). Les protéines ont ensuite été dosées par RC-DC.

B. Protocole 2

Les échantillons de terre congelée ont été broyés mécaniquement (MM400, Retsch) et les protéines ont été extraites dans un tampon d'extraction (0.25M citrate pH8). Après centrifugation à 2000 g pendant 10min, le culot 1 et le surnageant 1 ont été récupérés.
a) A partir du surnageant 1, les protéines ont été précipitées au TCA toute la nuit à 4°C, et après centrifugation (13 500g, 20min, 4°C), le culot a été repris dans un tampon Laemmli et les protéines dosées par RC-DC.
b) le culot 1 a été repris dans un deuxième tampon d'extraction (1% SDS, 0.1M TrisHCl pH 6.8, 20mM DTT, 50mM NH_4HCO_3). Après une centrifugation à 2000g pendant 10min, le culot a été éliminé et les protéines du surnageant ont été précipitées au TCA toute la nuit à 4°C. Après centrifugation (13 500g, 20min, 4°C), le culot protéique a ensuite été repris dans du tampon Laemmli et les protéines dosées par RC-DC.

La suite de la préparation d'échantillons était similaire pour les deux protocoles.

50µg de protéines ont été chargées sur gel SDS-PAGE. Le gel était composé d'un gel de concentration réticulé à 5% de polyacrylamide et un d'un gel de séparation réticulé à 12% de polyacrylamide. Pour les échantillons « tests », la migration a eu lieu pendant 35 min à 50V, puis 45min à 50V, les protéines ont ainsi migré sur 32mm. Pour les échantillons de l'expérience, la migration a eu lieu pendant 35 min à 50V, puis 25min à 150V, les protéines ont ainsi migré sur 12mm. Après fixation des protéines dans le gel (à l'aide d'une solution 50% éthanol, 3% acide phosphorique), les gels ont été colorés au bleu de Coomassie (0,1% Coomassie R250, 10% acide acétique, 40% méthanol). Les bandes protéiques de gel ont ensuite été découpées (2mm chacune). Les bandes ont été traitées automatiquement à l'aide

d'une station « MassPrep» (Waters). Une première étape a permis de décolorer les bandes (3 cycles de lavage avec 50µL de bicarbonate d'ammonium 25mM et 50µL d'acétonitrile, 10 min) ; puis les bandes ont été déshydratées (50µL d'acétonitrile, 60°C, 5min) ; les ponts disulfures des protéines ont été réduits (50µL DTT 10mM, 30min, 60°C) puis alkylés (50µL IAA 25mM, 30mM) ; puis un dernier cycle de lavage a été réalisé et enfin les bandes ont été déshydratées avec 50µL d'acétonitrile 15min.

Finalement, les protéines ont été digérées dans le gel par la trypsine (1%). A l'issue d'une incubation à 37°C toute la nuit, une solution à 60% acétonitrile, 0.1% d'acide formique a été ajoutée à chaque bande pour stopper la digestion et extraire les peptides du gel. Une deuxième extraction a été réalisée avec de l'acetonitrile 100%. Le solvant a été évaporé et les peptides repris dans 10µL d'une solution contenant 1% acétonitrile, 0,1% acide formique avant d'être analysés en spectrométrie de masse.

Le protocole 1 a finalement été retenu pour l'analyse des 18 échantillons.

II. Analyse nanoLC-MS/MS

L'analyse a été réalisée en utilisant un système nanoUPLC (nanoAcquity UPLC, Waters) couplé à un spectromètre de masse composé d'une source nanoSpray et d'un quadripole en tandem avec un analyseur orbitrap (Q Exactive+, Thermo Scientific). Le couplage était contrôlé par Thermo Xcalibur v3.0.63.

1µL de chaque échantillon a été chargé sur une pré-colonne d'enrichissement (Symmetry C18 180µm*20mm, 5µm ; Waters) en utilisant 99% de solvant A (0,1% acide formique dans l'eau) et 1% de B (0,1% acide formique dans l'acétonitrile) à un débit de 5 µL/min pendant 3min. L'élution des peptides a ensuite été effectuée sur une colonne de séparation (BEH 130 C18 75µm*250mm, 1,7µm ; Waters) maintenue à 60°C, à un débit de 450 nL/min avec les gradients suivants : 1-6% de B en 0,5 min puis 6-28% de B en 44,5 min et enfin de 28-35% de B en 3min pour les 18 échantillons de l'expérience ; 1-6% de B en 0,5 min puis 6-35% de B en 29,5 min pour les échantillons « tests ». Le gradient chromatographique était plus court pour les échantillons test car ils étaient davantage préfractionnés (16 vs. 6 bandes), donc nécessitaient une moindre séparation chromatographique.

Le spectromètre de masse a été utilisé en mode positif, les paramètres d'acquisition des spectres de masse sont présentés dans le Tableau Partie Exp. 8.

Tableau Partie Exp. 8 : Paramètres d'acquisition du spectromètre de masse.

MS	Résolution (à 200 m/z)	70 000
	Gamme de masse	300 – 1800 m/z
	AGC (Auto Gain Control)	3×10^6
	Maximum injection time	50ms
Fragmentation	Sélection	10 ions les plus intenses
	Temps d'exclusion	1 min
MS/MS	Résolution (à 200 m/z)	17500
	Gamme de masse	A partir de 100m/z jusqu'à 15 fois la plus petite masse observée

	AGC (Auto Gain Control)	1×10^5
	Maximum injection time	100ms

III. Analyse des données

A. Analyse des données des tests de protocole

Identification des protéines :
L'algorithme de recherche Mascot v2.5.1 (Matrix Science) [45] a été utilisé pour identifier les protéines à partir des données MS/MS, dans la banque de données SwissProt complète (avril 2015, 546 238 séquences). Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Les paramètres de recherche Mascot étaient les suivants : un site de clivage manqué par l'enzyme (la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 5ppm en mode MS et de 0.05Da en mode MS/MS. Grâce au logiciel Scaffold v4.3 (Proteome software inc.), les résultats de la recherche des 16 fractions ont été fusionnés et les identifications validées avec les critères suivants : (score d'ion – score d'identité) > -5 ; score d'ion > 35 ; longueur de séquence >7.

B. Analyse des données des échantillons

Une première banque de données, « db1 », a été constituée à partir des données de metabarcoding, elle contenait les séquences protéiques de tous les **champignons** (Swissprot mars 2016, taxonomie 4751) ; des **Agaricomycetidae** (UniProt mars 2016, taxonomie 452333) ; des **Rhizobiales** (Swissprot mars 2016, taxonomie 356) ; des **Bacteroides** (Swissprot mars 2016, taxonomie 816) ; des **Actinobacteria** (Swissprot mars 2016, taxonomie 201174) ; des **Acidobacteria** (UniProt mars 2016, taxonomie 57723) ; des **Verrucomicrobia** (UniProt mars 2016, taxonomie 74201) ; et des **Chloroflexia** (UniProt mars 2016, taxonomie 32061).

Pour les « tests de réduction de la banque de données », la suite logicielle MSDA [55] a été utilisée pour constituer les banques de données « intermédiaires », les données spectrales ont été interprétées par l'algorithme de recherche Mascot v2.5.1 (Matrix Science) [45] et les filtres ont été appliqués grâce au logiciel Scaffold v4.3 (Proteome Software). Finalement, à l'issue de ces tests, c'est l'algorithme de recherche Andromeda qui a été utilisé pour effectuer la première recherche non stringente (FDR 70%).

Identification des protéines :
L'algorithme de recherche Andromeda (Maxquant v1.5.3.30) [46] a été utilisé pour identifier des protéines à partir des données MS/MS. La recherche a été réalisée dans la banque de données protéiques personnalisée « db2 », contenant uniquement les séquences des protéines identifiées par la première recherche non stringente dans la « db1 ». Les séquences des contaminants usuels, incluant la trypsine et la kératine, ont été ajoutées à la banque de données. Une version target-decoy de la banque a alors été générée par Andromeda, afin de permettre d'évaluer le taux de faux positif (FDR). Les paramètres de recherche Andromeda étaient les suivants : un site de clivage manqué par l'enzyme

(la trypsine) toléré par peptide, la carbamidomethylation des cystéines et l'oxydation des methionines spécifiées comme modifications variables, l'erreur de masse tolérée était de 4,5 ppm en mode MS et 20ppm en mode MS/MS, les taux de faux positif (FDR) peptidique et protéique ont été fixés à 1%.

Quantification des protéines :

A partir des données spectrales (MS) d'un peptide donné, sa quantification a été obtenue après extraction puis intégration de la somme des courants d'ions générés. Pour cela, le logiciel Maxquant v1.5.3.30 a été utilisé. Seuls les peptides uniques ont été considérés pour la quantification, l'option « Label-Free Quantification » (LFQ) de Maxquant [114] a également permis de normaliser les intensités peptidiques et protéiques.

Bibliographie

Bibliographie

1. Stearns, S.C., *Life history evolution: successes, limitations, and prospects.* Naturwissenschaften, 2000. **87**(11): p. 476-86.
2. Gebhardt, M.D. and S.C. Stearns, *Phenotypic plasticity for life-history traits in Drosophila melanogaster. III. Effect of the environment on genetic parameters.* Genet Res, 1992. **60**(2): p. 87-101.
3. Kirkwood, T.B., *Understanding the odd science of aging.* Cell, 2005. **120**(4): p. 437-47.
4. Nilsson, T., M. Mann, R. Aebersold, J.R. Yates, 3rd, A. Bairoch, and J.J. Bergeron, *Mass spectrometry in high-throughput proteomics: ready for the big time.* Nat Methods, 2010. **7**(9): p. 681-5.
5. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
6. Diz, A.P., M. Martinez-Fernandez, and E. Rolan-Alvarez, *Proteomics in evolutionary ecology: linking the genotype with the phenotype.* Mol Ecol, 2012. **21**(5): p. 1060-80.
7. Armengaud, J., J. Trapp, O. Pible, O. Geffard, A. Chaumot, and E.M. Hartmann, *Non-model organisms, a species endangered by proteogenomics.* J Proteomics, 2014. **105**: p. 5-18.
8. Ma, B. and R. Johnson, *De novo sequencing and homology searching.* Mol Cell Proteomics, 2012. **11**(2): p. O111.014902.
9. Zhang, Y., B.R. Fonslow, B. Shan, M.C. Baek, and J.R. Yates, 3rd, *Protein analysis by shotgun/bottom-up proteomics.* Chem Rev, 2013. **113**(4): p. 2343-94.
10. Wu, C., J.C. Tran, L. Zamdborg, K.R. Durbin, M. Li, D.R. Ahlf, B.P. Early, P.M. Thomas, J.V. Sweedler, and N.L. Kelleher, *A protease for 'middle-down' proteomics.* Nat Methods, 2012. **9**(8): p. 822-4.
11. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top Down proteomics: facts and perspectives.* Biochem Biophys Res Commun, 2014. **445**(4): p. 683-93.
12. Chevallet, M., H. Diemer, A. Van Dorssealer, C. Villiers, and T. Rabilloud, *Toward a better analysis of secreted proteins: the example of the myeloid cells secretome.* Proteomics, 2007. **7**(11): p. 1757-70.
13. Candiano, G., M. Bruschi, L. Musante, L. Santucci, G.M. Ghiggeri, B. Carnemolla, P. Orecchia, L. Zardi, and P.G. Righetti, *Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis.* Electrophoresis, 2004. **25**(9): p. 1327-33.
14. Neuhoff, V., N. Arold, D. Taube, and W. Ehrhardt, *Improved staining of proteins in polyacrylamide gels including isoelectric focusing gels with clear background at nanogram sensitivity using Coomassie Brilliant Blue G-250 and R-250.* Electrophoresis, 1988. **9**(6): p. 255-62.
15. Rabilloud, T. and C. Lelong, *Two-dimensional gel electrophoresis in proteomics: a tutorial.* J Proteomics, 2011. **74**(10): p. 1829-41.
16. O'Farrell, P.H., *High resolution two-dimensional electrophoresis of proteins.* J Biol Chem, 1975. **250**(10): p. 4007-21.
17. Gorg, A., W. Weiss, and M.J. Dunn, *Current two-dimensional electrophoresis technology for proteomics.* Proteomics, 2004. **4**(12): p. 3665-85.
18. Rabilloud, T., M. Chevallet, S. Luche, and C. Lelong, *Two-dimensional gel electrophoresis in proteomics: Past, present and future.* J Proteomics, 2010. **73**(11): p. 2064-77.
19. Unlu, M., M.E. Morgan, and J.S. Minden, *Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.* Electrophoresis, 1997. **18**(11): p. 2071-7.

20. Sandra, K., M. Moshir, F. D'Hondt, K. Verleysen, K. Kas, and P. Sandra, *Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography.* J Chromatogr B Analyt Technol Biomed Life Sci, 2008. **866**(1-2): p. 48-63.
21. Sandra, K., M. Moshir, F. D'Hondt, R. Tuytten, K. Verleysen, K. Kas, I. Francois, and P. Sandra, *Highly efficient peptide separations in proteomics. Part 2: bi- and multidimensional liquid-based separation techniques.* J Chromatogr B Analyt Technol Biomed Life Sci, 2009. **877**(11-12): p. 1019-39.
22. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, 1988. **60**(20): p. 2299-301.
23. Fenn, J.B., M. Mann, C.K. Meng, S.F. Wong, and C.M. Whitehouse, *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.
24. Cech, N.B. and C.G. Enke, *Practical implications of some recent studies in electrospray ionization fundamentals.* Mass Spectrom Rev, 2001. **20**(6): p. 362-87.
25. Paul, W. and H. Steinwedel, *Notizen: Ein neues Massenspektrometer ohne Magnetfeld*, in *Zeitschrift für Naturforschung A*. 1953. p. 448.
26. de Hoffmann, E. and V. Stroobant, *Mass Spectrometry: Principles and Applications*. 2007: Wiley.
27. Wolff, M.M. and W.E. Stephens, *A Pulsed Mass Spectrometer with Time Dispersion.* Review of Scientific Instruments, 1953. **24**(8): p. 616-617.
28. Hardman, M. and A.A. Makarov, *Interfacing the orbitrap mass analyzer to an electrospray ion source.* Anal Chem, 2003. **75**(7): p. 1699-705.
29. Hu, Q., R.J. Noll, H. Li, A. Makarov, M. Hardman, and R. Graham Cooks, *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.
30. Makarov, A., *Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis.* Anal Chem, 2000. **72**(6): p. 1156-62.
31. Chernushevich, I.V., A.V. Loboda, and B.A. Thomson, *An introduction to quadrupole-time-of-flight mass spectrometry.* J Mass Spectrom, 2001. **36**(8): p. 849-65.
32. Kurz, E.A., *Effects of high-current input pulses upon Channeltrons.* Rev Sci Instrum, 1979. **50**(11): p. 1492.
33. Audier, M., J.C. Delmotte, and J.P. Boutot, *Multiplicateur à galette de microcanaux : amélioration des performances de gain et de dynamique de détection.* Rev. Phys. Appl. (Paris), 1978. **13**(4): p. 188-194.
34. Wells, J.M. and S.A. McLuckey, *Collision-induced dissociation (CID) of peptides and proteins.* Methods Enzymol, 2005. **402**: p. 148-85.
35. Zubarev, R.A., D.M. Horn, E.K. Fridriksson, N.L. Kelleher, N.A. Kruger, M.A. Lewis, B.K. Carpenter, and F.W. McLafferty, *Electron capture dissociation for structural characterization of multiply charged protein cations.* Anal Chem, 2000. **72**(3): p. 563-73.
36. Syka, J.E.P., J.J. Coon, M.J. Schroeder, J. Shabanowitz, and D.F. Hunt, *Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(26): p. 9528-9533.
37. Dongré, A.R., J.L. Jones, Á. Somogyi, and V.H. Wysocki, *Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model.* Journal of the American Chemical Society, 1996. **118**(35): p. 8365-8374.
38. Biemann, K., *Appendix 5. Nomenclature for peptide fragment ions (positive ions).* Methods Enzymol, 1990. **193**: p. 886-7.
39. Olsen, J.V., B. Macek, O. Lange, A. Makarov, S. Horning, and M. Mann, *Higher-energy C-trap dissociation for peptide modification analysis.* Nat Methods, 2007. **4**(9): p. 709-12.
40. Michalski, A., J. Cox, and M. Mann, *More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC-MS/MS.* J Proteome Res, 2011. **10**(4): p. 1785-93.

41. Venable, J.D., M.Q. Dong, J. Wohlschlegel, A. Dillin, and J.R. Yates, *Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra.* Nat Methods, 2004. **1**(1): p. 39-45.
42. Blueggel, M., D. Chamrad, and H.E. Meyer, *Bioinformatics in proteomics.* Curr Pharm Biotechnol, 2004. **5**(1): p. 79-88.
43. Eng, J.K., A.L. McCormack, and J.R. Yates, *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database.* J Am Soc Mass Spectrom, 1994. **5**(11): p. 976-89.
44. Guthals, A., J.D. Watrous, P.C. Dorrestein, and N. Bandeira, *The spectral networks paradigm in high throughput mass spectrometry.* Molecular bioSystems, 2012. **8**(10): p. 2535-2544.
45. Perkins, D.N., D.J. Pappin, D.M. Creasy, and J.S. Cottrell, *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
46. Cox, J., N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, and M. Mann, *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
47. Craig, R., J.P. Cortens, and R.C. Beavis, *Open source system for analyzing, validating, and storing protein identification data.* J Proteome Res, 2004. **3**(6): p. 1234-42.
48. Geer, L.Y., 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.* J Proteome Res, 2004. **3**(5): p. 958-64.
49. O'Donovan, C., M.J. Martin, A. Gattiker, E. Gasteiger, A. Bairoch, and R. Apweiler, *High-quality protein knowledge resource: SWISS-PROT and TrEMBL.* Brief Bioinform, 2002. **3**(3): p. 275-84.
50. Boeckmann, B., A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
51. Nesvizhskii, A.I. and R. Aebersold, *Interpretation of shotgun proteomic data: the protein inference problem.* Mol Cell Proteomics, 2005. **4**(10): p. 1419-40.
52. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.
53. Savitski, M.M., W.I. M, H. Hahne, B. Kuster, and M. Bantscheff, *A scalable approach for protein false discovery rate estimation in large proteomic data sets.* Mol Cell Proteomics, 2015. **17**: p. 046995.
54. Seidler, J., N. Zinn, M.E. Boehm, and W.D. Lehmann, *De novo sequencing of peptides by MS/MS.* Proteomics, 2010. **10**(4): p. 634-49.
55. Carapito, C., A. Burel, P. Guterl, A. Walter, F. Varrier, F. Bertile, and A. Van Dorsselaer, *MSDA, a proteomics software suite for in-depth Mass Spectrometry Data Analysis using grid computing.* Proteomics, 2014. **14**(9): p. 1014-9.
56. Ma, B., K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, *PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry.* Rapid Commun Mass Spectrom, 2003. **17**(20): p. 2337-42.
57. Frank, A. and P. Pevzner, *PepNovo: de novo peptide sequencing via probabilistic network modeling.* Anal Chem, 2005. **77**(4): p. 964-73.
58. Shevchenko, A., S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens, and K.G. Standing, *Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching.* Anal Chem, 2001. **73**(9): p. 1917-26.
59. !!! INVALID CITATION !!! (Armengaud, et al. 2014; Yates, et al. 1995).
60. VerBerkmoes, N.C., V.J. Denef, R.L. Hettich, and J.F. Banfield, *Systems biology: Functional analysis of natural microbial consortia using community proteomics.* Nat Rev Microbiol, 2009. **7**(3): p. 196-205.

61. Lopez-Casado, G., P.A. Covey, P.A. Bedinger, L.A. Mueller, T.W. Thannhauser, S. Zhang, Z. Fei, J.J. Giovannoni, and J.K. Rose, *Enabling proteomic studies with RNA-Seq: The proteome of tomato pollen as a test case.* Proteomics, 2012. **12**(6): p. 761-74.
62. He, R., M.J. Kim, W. Nelson, T.S. Balbuena, R. Kim, R. Kramer, J.A. Crow, G.D. May, J.J. Thelen, C.A. Soderlund, and D.R. Gang, *Next-generation sequencing-based transcriptomic and proteomic analysis of the common reed, Phragmites australis (Poaceae), reveals genes involved in invasiveness and rhizome specificity.* Am J Bot, 2012. **99**(2): p. 232-47.
63. Evans, V.C., G. Barker, K.J. Heesom, J. Fan, C. Bessant, and D.A. Matthews, *De novo derivation of proteomes from transcriptomes for transcript and protein identification.* Nat Methods, 2012. **9**(12): p. 1207-11.
64. Song, J., R. Sun, D. Li, F. Tan, X. Li, P. Jiang, X. Huang, L. Lin, Z. Deng, and Y. Zhang, *An improvement of shotgun proteomics analysis by adding next-generation sequencing transcriptome data in orange.* PLoS One, 2012. **7**(6): p. e39494.
65. Grabherr, M.G., B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, *Full-length transcriptome assembly from RNA-Seq data without a reference genome.* Nat Biotech, 2011. **29**(7): p. 644-652.
66. Haas, B.J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, and A. Regev, *De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis.* Nat Protoc, 2013. **8**(8): p. 1494-512.
67. Luge, T. and S. Sauer, *Generating Sample-Specific Databases for Mass Spectrometry-Based Proteomic Analysis by Using RNA Sequencing.* Methods Mol Biol, 2016. **1394**: p. 219-232.
68. Poptsova, M.S. and J.P. Gogarten, *Using comparative genome analysis to identify problems in annotated microbial genomes.* Microbiology, 2010. **156**(Pt 7): p. 1909-17.
69. Jaffe, J.D., H.C. Berg, and G.M. Church, *Proteogenomic mapping as a complementary method to perform genome annotation.* Proteomics, 2004. **4**(1): p. 59-77.
70. Savidor, A., R.S. Donahoo, O. Hurtado-Gonzales, N.C. Verberkmoes, M.B. Shah, K.H. Lamour, and W.H. McDonald, *Expressed peptide tags: an additional layer of data for genome annotation.* J Proteome Res, 2006. **5**(11): p. 3048-58.
71. Ansong, C., S.O. Purvine, J.N. Adkins, M.S. Lipton, and R.D. Smith, *Proteogenomics: needs and roles to be filled by proteomics in genome annotation.* Brief Funct Genomic Proteomic, 2008. **7**(1): p. 50-62.
72. Cantarel, B.L., A.R. Erickson, N.C. VerBerkmoes, B.K. Erickson, P.A. Carey, C. Pan, M. Shah, E.F. Mongodin, J.K. Jansson, C.M. Fraser-Liggett, and R.L. Hettich, *Strategies for metagenomic-guided whole-community proteomics of complex microbial environments.* PLoS One, 2011. **6**(11): p. e27173.
73. Muth, T., C.A. Kolmeder, J. Salojarvi, S. Keskitalo, M. Varjosalo, F.J. Verdam, S.S. Rensen, U. Reichl, W.M. de Vos, E. Rapp, and L. Martens, *Navigating through metaproteomics data: A logbook of database searching.* Proteomics, 2015. **16**(10): p. 201400560.
74. Muth, T., B.Y. Renard, and L. Martens, *Metaproteomic data analysis at a glance: advances in computational microbial community proteomics.* Expert Rev Proteomics, 2016. **13**(8): p. 757-69.
75. Tanca, A., A. Palomba, M. Deligios, T. Cubeddu, C. Fraumene, G. Biosa, D. Pagnozzi, M.F. Addis, and S. Uzzau, *Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture.* PLoS One, 2013. **8**(12): p. e82981.
76. Taberlet, P., S.M. Prud'Homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, J.C. Clement, C. Melodelima, F. Pompanon, and E. Coissac, *Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies.* Mol Ecol, 2012. **21**(8): p. 1816-20.

77. Pompanon, F., E. Coissac, and P. Taberlet, *Metabarcoding, une nouvelle façon d'analyser la biodiversité.* Biofutur, 2011. **319**: p. 30-32.
78. Kolmeder, C.A. and W.M. de Vos, *Metaproteomics of our microbiome - developing insight in function and activity in man and model systems.* J Proteomics, 2014. **97**: p. 3-16.
79. Wohlbrand, L., B. Wemheuer, C. Feenders, H.S. Ruppersberg, C. Hinrichs, B. Blasius, R. Daniel, and R. Rabus, *Complementary Metaproteomic Approaches to Assess the Bacterioplankton Response toward a Phytoplankton Spring Bloom in the Southern North Sea.* Front Microbiol, 2017. **8**: p. 442.
80. Morgan, X.C. and C. Huttenhower, *Meta'omic analytic techniques for studying the intestinal microbiome.* Gastroenterology, 2014. **146**(6): p. 1437-1448.e1.
81. Gonnelli, G., M. Stock, J. Verwaeren, D. Maddelein, B. De Baets, L. Martens, and S. Degroeve, *A decoy-free approach to the identification of peptides.* J Proteome Res, 2015. **14**(4): p. 1792-8.
82. Castellana, N. and V. Bafna, *Proteogenomics to discover the full coding content of genomes: a computational perspective.* Journal of proteomics, 2010. **73**(11): p. 2124-2135.
83. Shteynberg, D., E.W. Deutsch, H. Lam, J.K. Eng, Z. Sun, N. Tasman, L. Mendoza, R.L. Moritz, R. Aebersold, and A.I. Nesvizhskii, *iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates.* Mol Cell Proteomics, 2011. **10**(12): p. M111.007690.
84. Reiter, L., M. Claassen, S.P. Schrimpf, M. Jovanovic, A. Schmidt, J.M. Buhmann, M.O. Hengartner, and R. Aebersold, *Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry.* Mol Cell Proteomics, 2009. **8**(11): p. 2405-17.
85. Castellana, N. and V. Bafna, *Proteogenomics to discover the full coding content of genomes: a computational perspective.* J Proteomics, 2010. **73**(11): p. 2124-35.
86. Castellana, N.E., S.H. Payne, Z. Shen, M. Stanke, V. Bafna, and S.P. Briggs, *Discovery and revision of Arabidopsis genes by proteogenomics.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(52): p. 21034-21038.
87. Kim, S., N. Gupta, and P.A. Pevzner, *Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases.* J Proteome Res, 2008. **7**(8): p. 3354-63.
88. Wang, X., R.J. Slebos, D. Wang, P.J. Halvey, D.L. Tabb, D.C. Liebler, and B. Zhang, *Protein identification using customized protein sequence databases derived from RNA-Seq data.* J Proteome Res, 2012. **11**(2): p. 1009-17.
89. Jagtap, P., J. Goslinga, J.A. Kooren, T. McGowan, M.S. Wroblewski, S.L. Seymour, and T.J. Griffin, *A two-step database search method improves sensitivity in peptide sequence matches for metaproteomics and proteogenomics studies.* Proteomics, 2013. **13**(8): p. 1352-7.
90. Huson, D.H., A.F. Auch, J. Qi, and S.C. Schuster, *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.
91. Cima, I., R. Schiess, P. Wild, M. Kaelin, P. Schuffler, V. Lange, P. Picotti, R. Ossola, A. Templeton, O. Schubert, T. Fuchs, T. Leippold, S. Wyler, J. Zehetner, W. Jochum, J. Buhmann, T. Cerny, H. Moch, S. Gillessen, R. Aebersold, and W. Krek, *Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer.* Proc Natl Acad Sci U S A, 2011. **108**(8): p. 3342-7.
92. Iuga, C., A. Seicean, C. Iancu, R. Buiga, P.K. Sappa, U. Volker, and E. Hammer, *Proteomic identification of potential prognostic biomarkers in resectable pancreatic ductal adenocarcinoma.* Proteomics, 2014. **14**(7-8): p. 945-55.
93. Picotti, P. and R. Aebersold, *Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.* Nat Methods, 2012. **9**(6): p. 555-66.
94. Picotti, P., B. Bodenmiller, L.N. Mueller, B. Domon, and R. Aebersold, *Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics.* Cell, 2009. **138**(4): p. 795-806.

95. Gallien, S., E. Duriez, C. Crone, M. Kellmann, T. Moehring, and B. Domon, *Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer.* Mol Cell Proteomics, 2012. **11**(12): p. 1709-23.
96. Ong, S.E., B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, and M. Mann, *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
97. Miyagi, M. and K.C. Rao, *Proteolytic 18O-labeling strategies for quantitative proteomics.* Mass Spectrom Rev, 2007. **26**(1): p. 121-36.
98. Gygi, S.P., B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold, *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.
99. Ross, P.L., Y.N. Huang, J.N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D.J. Pappin, *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
100. Thompson, A., J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A.K. Mohammed, and C. Hamon, *Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS.* Anal Chem, 2003. **75**(8): p. 1895-904.
101. Hsu, J.L., S.Y. Huang, N.H. Chow, and S.H. Chen, *Stable-isotope dimethyl labeling for quantitative proteomics.* Anal Chem, 2003. **75**(24): p. 6843-52.
102. Neilson, K.A., N.A. Ali, S. Muralidharan, M. Mirzaei, M. Mariani, G. Assadourian, A. Lee, S.C. van Sluyter, and P.A. Haynes, *Less label, more free: approaches in label-free quantitative mass spectrometry.* Proteomics, 2011. **11**(4): p. 535-53.
103. Liu, H., R.G. Sadygov, and J.R. Yates, 3rd, *A model for random sampling and estimation of relative protein abundance in shotgun proteomics.* Anal Chem, 2004. **76**(14): p. 4193-201.
104. Zhu, W., J.W. Smith, and C.M. Huang, *Mass spectrometry-based label-free quantitative proteomics.* J Biomed Biotechnol, 2010. **2010**: p. 840518.
105. Florens, L., M.J. Carozza, S.K. Swanson, M. Fournier, M.K. Coleman, J.L. Workman, and M.P. Washburn, *Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors.* Methods, 2006. **40**(4): p. 303-11.
106. Rappsilber, J., U. Ryder, A.I. Lamond, and M. Mann, *Large-scale proteomic analysis of the human spliceosome.* Genome Res, 2002. **12**(8): p. 1231-45.
107. Higgs, R.E., J.P. Butler, B. Han, and M.D. Knierman, *Quantitative Proteomics via High Resolution MS Quantification: Capabilities and Limitations.* Int J Proteomics, 2013. **2013**: p. 674282.
108. Sandin, M., J. Teleman, J. Malmstrom, and F. Levander, *Data processing methods and quality control strategies for label-free LC-MS protein quantification.* Biochim Biophys Acta, 2014. **1844**(1 Pt A): p. 29-41.
109. Valot, B., O. Langella, E. Nano, and M. Zivy, *MassChroQ: a versatile tool for mass spectrometry quantification.* Proteomics, 2011. **11**(17): p. 3572-7.
110. Gautier, V., E. Mouton-Barbosa, D. Bouyssie, N. Delcourt, M. Beau, J.P. Girard, C. Cayrol, O. Burlet-Schiltz, B. Monsarrat, and A. Gonzalez de Peredo, *Label-free quantification and shotgun analysis of complex proteomes by one-dimensional SDS-PAGE/NanoLC-MS: evaluation for the large scale analysis of inflammatory human endothelial cells.* Mol Cell Proteomics, 2012. **11**(8): p. 527-39.
111. Schilling, B., M.J. Rardin, B.X. MacLean, A.M. Zawadzka, B.E. Frewen, M.P. Cusack, D.J. Sorensen, M.S. Bereman, E. Jing, C.C. Wu, E. Verdin, C.R. Kahn, M.J. Maccoss, and B.W. Gibson, *Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation.* Mol Cell Proteomics, 2012. **11**(5): p. 202-14.

112. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
113. MacLean, B., D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, R. Kern, D.L. Tabb, D.C. Liebler, and M.J. MacCoss, *Skyline: an open source document editor for creating and analyzing targeted proteomics experiments.* Bioinformatics, 2010. **26**(7): p. 966-8.
114. Cox, J., M.Y. Hein, C.A. Luber, I. Paron, N. Nagaraj, and M. Mann, *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
115. Andersen, J.S., C.J. Wilkinson, T. Mayor, P. Mortensen, E.A. Nigg, and M. Mann, *Proteomic characterization of the human centrosome by protein correlation profiling.* Nature, 2003. **426**(6966): p. 570-4.
116. Chawade, A., E. Alexandersson, and F. Levander, *Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets.* J Proteome Res, 2014. **13**(6): p. 3114-20.
117. Carrillo, B., C. Yanofsky, S. Laboissiere, R. Nadon, and R.E. Kearney, *Methods for combining peptide intensities to estimate relative protein abundance.* Bioinformatics, 2010. **26**(1): p. 98-103.
118. Ahrne, E., L. Molzahn, T. Glatter, and A. Schmidt, *Critical assessment of proteome-wide label-free absolute abundance estimation strategies.* Proteomics, 2013. **13**(17): p. 2567-78.
119. Lazar, C., 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.* J Proteome Res, 2016. **15**(4): p. 1116-25.
120. Webb-Robertson, B.J., 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, and K.M. Waters, *Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics.* J Proteome Res, 2015. **14**(5): p. 1993-2001.
121. Nie, L., G. Wu, F.J. Brockman, and W. Zhang, *Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins.* Bioinformatics, 2006. **22**(13): p. 1641-7.
122. Torres-Garcia, W., S.D. Brown, R.H. Johnson, W. Zhang, G.C. Runger, and D.R. Meldrum, *Integrative analysis of transcriptomic and proteomic data of Shewanella oneidensis: missing value imputation using temporal datasets.* Mol Biosyst, 2011. **7**(4): p. 1093-104.
123. Gillet, L.C., P. Navarro, S. Tate, H. Rost, 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.* Mol Cell Proteomics, 2012. **11**(6): p. O111.016717.
124. Navarro, P., J. Kuharev, L.C. Gillet, O.M. Bernhardt, B. MacLean, H.L. Rost, S.A. Tate, C.C. Tsou, L. Reiter, U. Distler, G. Rosenberger, Y. Perez-Riverol, A.I. Nesvizhskii, R. Aebersold, and S. Tenzer, *A multicenter study benchmarks software tools for label-free proteome quantification.* Nat Biotechnol, 2016. **34**(11): p. 1130-1136.
125. Silva, J.C., R. Denny, C.A. Dorschel, M. Gorenstein, I.J. Kass, G.Z. Li, T. McKenna, M.J. Nold, K. Richardson, P. Young, and S. Geromanos, *Quantitative proteomic analysis by accurate mass retention time pairs.* Anal Chem, 2005. **77**(7): p. 2187-200.
126. Domon, B. and R. Aebersold, *Options and considerations when selecting a quantitative proteomics strategy.* Nat Biotechnol, 2010. **28**(7): p. 710-21.
127. Dephoure, N. and S.P. Gygi, *Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast.* Sci Signal, 2012. **5**(217): p. rs2.
128. Stearns, S.C., *The evolution of life-histories. Oxford: Oxford University Press.* 1992.
129. Roff, D.A., *Evolution of life-histories: theory and analysis. Springer New York.* 1992.

130. Martin, L.B., A. Scheuerlein, and M. Wikelski, *Immune activity elevates energy expenditure of house sparrows: a link between direct and indirect costs?* Proceedings of the Royal Society B-Biological Sciences, 2003. **270**(1511): p. 153-158.
131. Lochmiller, R.L. and C. Deerenberg, *Trade-offs in evolutionary immunology: just what is the cost of immunity?* Oikos, 2000. **88**(1): p. 87-98.
132. Sheldon, B.C. and S. Verhulst, *Ecological immunology: costly parasite defences and trade-offs in evolutionary ecology.* Trends Ecol Evol, 1996. **11**(8): p. 317-21.
133. Norris, K. and M.R. Evans, *Ecological immunology: life history trade-offs and immune defense in birds.* Behavioral Ecology, 2000. **11**(1): p. 19-26.
134. Muehlenbein, M.P., J.L. Hirschtick, J.Z. Bonner, and A.M. Swartz, *Toward Quantifying the Usage Costs of Human Immunity: Altered Metabolic Rates and Hormone Levels During Acute Immune Activation in Men.* American Journal of Human Biology, 2010. **22**(4): p. 546-556.
135. Cao, Z., S. Yende, J.A. Kellum, D.C. Angus, and R.A. Robinson, *Proteomics reveals age-related differences in the host immune response to sepsis.* J Proteome Res, 2014. **13**(2): p. 422-32.
136. Franceschi, C., M. Bonafe, S. Valensin, F. Olivieri, M. De Luca, E. Ottaviani, and G. De Benedictis, *Inflamm-aging - An evolutionary perspective on immunosenescence.* Molecular and Cellular Gerontology, 2000. **908**: p. 244-254.
137. O'Connor, J.E., G. Herrera, A. Martinez-Romero, F.S. de Oyanguren, L. Diaz, A. Gomes, S. Balaguer, and R.C. Callaghan, *Systems Biology and immune aging.* Immunol Lett, 2014. **162**(1 Pt B): p. 334-45.
138. Bokov, A., A. Chaudhuri, and A. Richardson, *The role of oxidative damage and stress in aging.* Mechanisms of Ageing and Development, 2004. **125**(10-11): p. 811-26.
139. Shaw, A.C., D.R. Goldstein, and R.R. Montgomery, *Age-dependent dysregulation of innate immunity.* Nat Rev Immunol, 2013. **13**(12): p. 875-87.
140. Montecino-Rodriguez, E., B. Berent-Maoz, and K. Dorshkind, *Causes, consequences, and reversal of immune system aging.* Journal of Clinical Investigation, 2013. **123**(3): p. 958-965.
141. Martin, L.B., 2nd, A. Scheuerlein, and M. Wikelski, *Immune activity elevates energy expenditure of house sparrows: a link between direct and indirect costs?* Proc Biol Sci, 2003. **270**(1511): p. 153-8.
142. Sorci, G. and B. Faivre, *Inflammation and oxidative stress in vertebrate host-parasite systems.* Philos Trans R Soc Lond B Biol Sci, 2009. **364**(1513): p. 71-83.
143. Belloni, V., B. Faivre, R. Guerreiro, E. Arnoux, J. Bellenger, and G. Sorci, *Suppressing an anti-inflammatory cytokine reveals a strong age-dependent survival cost in mice.* PLoS One, 2010. **5**(9): p. e12940.
144. Demas, G.E., V. Chefer, M.I. Talan, and R.J. Nelson, *Metabolic costs of mounting an antigen-stimulated immune response in adult and aged C57BL/6J mice.* Am J Physiol, 1997. **273**(5 Pt 2): p. R1631-7.
145. Mostovenko, E., C. Hassan, J. Rattke, A.M. Deelder, P.A. van Veelen, and M. Palmblad, *Comparison of peptide and protein fractionation methods in proteomics.* EuPA Open Proteomics, 2013. **1**: p. 30-37.
146. Goeminne, L.J., K. Gevaert, and L. Clement, *Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics.* Mol Cell Proteomics, 2016. **15**(2): p. 657-68.
147. Jiang, H., E. Schiffer, Z. Song, J. Wang, P. Zurbig, K. Thedieck, S. Moes, H. Bantel, N. Saal, J. Jantos, M. Brecht, P. Jeno, M.N. Hall, K. Hager, M.P. Manns, H. Hecker, A. Ganser, K. Dohner, A. Bartke, C. Meissner, H. Mischak, Z. Ju, and K.L. Rudolph, *Proteins induced by telomere dysfunction and DNA damage represent biomarkers of human aging and disease.* Proc Natl Acad Sci U S A, 2008. **105**(32): p. 11299-304.
148. Guani-Guerra, E., T. Santos-Mendoza, S.O. Lugo-Reyes, and L.M. Teran, *Antimicrobial peptides: general overview and clinical implications in human health and disease.* Clin Immunol, 2010. **135**(1): p. 1-11.

149. Watabe-Rudolph, M., Z. Song, L. Lausser, C. Schnack, Y. Begus-Nahrmann, M.O. Scheithauer, G. Rettinger, M. Otto, H. Tumani, D.R. Thal, J. Attems, K.A. Jellinger, H.A. Kestler, C.A. von Arnim, and K.L. Rudolph, *Chitinase enzyme activity in CSF is a powerful biomarker of Alzheimer disease.* Neurology, 2012. **78**(8): p. 569-77.
150. von Figura, G., D. Hartmann, Z. Song, and K.L. Rudolph, *Role of telomere dysfunction in aging and its detection by biomarkers.* J Mol Med (Berl), 2009. **87**(12): p. 1165-71.
151. Belloni, V., G. Sorci, E. Paccagnini, R. Guerreiro, J. Bellenger, and B. Faivre, *Disrupting immune regulation incurs transient costs in male reproductive function.* PLoS One, 2014. **9**(1): p. e84606.
152. Moore, K.W., R. de Waal Malefyt, R.L. Coffman, and A. O'Garra, *Interleukin-10 and the interleukin-10 receptor.* Annu Rev Immunol, 2001. **19**: p. 683-765.
153. Rousset, S., Y. Emre, O. Join-Lambert, C. Hurtaud, D. Ricquier, and A.-M. Cassard-Doulcier, *The uncoupling protein 2 modulates the cytokine balance in innate immunity.* Cytokine, 2006. **35**(3-4): p. 135-42.
154. Emre, Y., C. Hurtaud, T. Nubel, F. Criscuolo, D. Ricquier, and A.M. Cassard-Doulcier, *Mitochondria contribute to LPS-induced MAPK activation via uncoupling protein UCP2 in macrophages.* Biochem J, 2007. **402**(2): p. 271-8.
155. Gygi, S.P., G.L. Corthals, Y. Zhang, Y. Rochon, and R. Aebersold, *Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.* Proc Natl Acad Sci U S A, 2000. **97**(17): p. 9390-5.
156. Ong, S.E. and A. Pandey, *An evaluation of the use of two-dimensional gel electrophoresis in proteomics.* Biomol Eng, 2001. **18**(5): p. 195-205.
157. Finucane, M.M. and al., *National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants.* Lancet, 2011. **377**: p. 557-567.
158. Spiegelman, B.M. and J.S. Flier, *Obesity and the regulation of energy balance.* Cell, 2001. **104**: p. 531-543.
159. Cypess, A.M., *Identification and importance of brown adipose tissue in adult humans.* The New England Journal of Medicine, 2009. **360**: p. 1509-1517.
160. Ouellet, V., *Brown adipose tissue oxidative metabolism contributes to energy expenditure during acute cold exposure in humans.* The Journal of Clinical Investigation, 2012. **122**: p. 545-552.
161. Nedergaard, J. and B. Cannon, *How brown is brown fat? It depends where you look.* Nature Medicine, 2013. **19**: p. 540-541.
162. Ricquier, D. and F. Bouillaud, *The uncoupling protein homologues: UCP1, UCP2, UCP3, StUCP and AtUCP.* Biochemical Journal, 2000. **345**: p. 161-179.
163. Nedergaard, J., T. Bengtsson, and B. Cannon, *Unexpected evidence for active brown adipose tissue in adult humans.* American Journal of Physiology - Endocrinology and Metabolism, 2007. **293**: p. 444-452.
164. Fridlyand, L.E. and L.H. Philipson, *Cold climate genes and the prevalence of type 2 diabetes mellitus.* Medical Hypothesis, 2006. **67**: p. 1034-1041.
165. van Marken Lichtenbelt, W. and P. Schrauwen, *Implications of nonshivering thermogenesis for energy balance regulation.* American Journal of Physiology - Regulatory, Integrative and Comparative Physiology, 2011. **301**: p. 285-296.
166. Cannon, B. and J. Nedergaard, *Yes, even human brown fat is on fire!* Journal of Clinical Investigation, 2012. **122**: p. 486-489.
167. Mattson, M.P., *Perspective: Does brown fat protect against diseases of aging?* Ageing Research Reviews, 2010. **9**: p. 69-76.
168. Stephens, M. and al., *Brown fat and obesity, the next big thing?* Clinical Endocrinology, 2011. **74 (6)**: p. 661-670.
169. Ortega-Molina, A., *Pten positively regulates brown adipose function, energy expenditure, and longevity.* Cell Metabolism, 2012. **15**: p. 382-394.

170. Whittle, A.J. and al., *Using brown adipose tissue to treat obesity - the central issue.* Trends in Molecular Medicine, 2011. **17**: p. 405-411.
171. Pearson, W.R., *Effective protein sequence comparison.* Methods Enzymol, 1996. **266**: p. 227-58.
172. Frank, A.M., *A ranking-based scoring function for peptide-spectrum matches.* J Proteome Res, 2009. **8**(5): p. 2241-52.
173. Eidhammer, I., K. Flikka, L. Martens, and S.-O. Mikalsen, *Database Searching for de novosequences*, in *Computational Methods for Mass Spectrometry Proteomics*. 2007, John Wiley & Sons, Ltd. p. 193-210.
174. Wielsch, N., H. Thomas, V. Surendranath, P. Waridel, A. Frank, P. Pevzner, and A. Shevchenko, *Rapid validation of protein identifications with the borderline statistical confidence via de novo sequencing and MS BLAST searches.* J Proteome Res, 2006. **5**(9): p. 2448-56.
175. Plumel, M.I., T. Wasselin, V. Plot, J.M. Strub, A. Van Dorsselaer, C. Carapito, J.Y. Georges, and F. Bertile, *Mass spectrometry-based sequencing and SRM-based quantitation of two novel vitellogenin isoforms in the leatherback sea turtle (Dermochelys coriacea).* J Proteome Res, 2013. **12**(9): p. 4122-35.
176. Zhang, J., L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G.A. Lajoie, and B. Ma, *PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification.* Mol Cell Proteomics, 2012. **11**(4): p. M111.010587.
177. Forshed, J., H.J. Johansson, M. Pernemalm, R.M. Branca, A. Sandberg, and J. Lehtio, *Enhanced information output from shotgun proteomics data by protein quantification and peptide quality control (PQPQ).* Mol Cell Proteomics, 2011. **10**(10): p. M111.010264.
178. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
179. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
180. Zhang, Y., J. Liu, J.L. Yao, G. Ji, L. Qian, J. Wang, G.S. Zhang, J. Tian, Y.Z. Nie, Y.E. Zhang, M.S. Gold, and Y.J. Liu, *Obesity: Pathophysiology and Intervention.* Nutrients, 2014. **6**(11): p. 5153-5183.
181. Morrison, C.D., P. Huypens, L.K. Stewart, and T.W. Gettys, *Implications of crosstalk between leptin and insulin signaling during the development of diet-induced obesity.* Biochim Biophys Acta, 2009. **1792**(5): p. 409-16.
182. Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-d361.
183. Medawar, P.B., *The Definition and Measurement of Senescence*, in *Ciba Foundation Symposium - General Aspects (Colloquia on Ageing)*. 2008, John Wiley & Sons, Ltd. p. 4-15.
184. Marmot, M.G., M.J. Shipley, and G. Rose, *Inequalities in death--specific explanations of a general pattern?* Lancet, 1984. **1**(8384): p. 1003-6.
185. Holt-Lunstad, J., T.B. Smith, and J.B. Layton, *Social relationships and mortality risk: a meta-analytic review.* PLoS Med, 2010. **7**(7): p. e1000316.
186. DeVries, A.C., E.R. Glasper, and C.E. Detillion, *Social modulation of stress responses.* Physiol Behav, 2003. **79**(3): p. 399-407.
187. Sapolsky, R.M., *The influence of social hierarchy on primate health.* Science, 2005. **308**(5722): p. 648-52.
188. Epel, E.S., E.H. Blackburn, J. Lin, F.S. Dhabhar, N.E. Adler, J.D. Morrow, and R.M. Cawthon, *Accelerated telomere shortening in response to life stress.* Proc Natl Acad Sci U S A, 2004. **101**(49): p. 17312-5.
189. Mersch, D.P., A. Crespi, and L. Keller, *Tracking individuals shows spatial fidelity is a key regulator of ant social organization.* Science, 2013. **340**(6136): p. 1090-3.
190. Hartmann, A. and J. Heinze, *Lay eggs, live longer: division of labor and life span in a clonal ant species.* Evolution, 2003. **57**(10): p. 2424-9.

191. Begna, D., Y. Fang, M. Feng, and J. Li, *Mitochondrial proteins differential expression during honeybee (Apis mellifera L.) queen and worker larvae caste determination.* J Proteome Res, 2011. **10**(9): p. 4263-80.
192. Fang, Y., F. Song, L. Zhang, D.W. Aleku, B. Han, M. Feng, and J. Li, *Differential antennal proteome comparison of adult honeybee drone, worker and queen (Apis mellifera L.).* J Proteomics, 2012. **75**(3): p. 756-73.
193. Blagosklonny, M.V., *Aging: ROS or TOR.* Cell Cycle, 2008. **7**(21): p. 3344-54.
194. Robinson, G.E., *Regulation of division of labor in insect societies.* Annu Rev Entomol, 1992. **37**: p. 637-65.
195. Tan, Q.Q., W. Liu, F. Zhu, C.L. Lei, D.A. Hahn, and X.P. Wang, *Describing the Diapause-Preparatory Proteome of the Beetle Colaphellus bowringi and Identifying Candidates Affecting Lipid Accumulation Using Isobaric Tags for Mass Spectrometry-Based Proteome Quantification (iTRAQ).* Front Physiol, 2017. **8**: p. 251.
196. Cremer, S., S.A. Armitage, and P. Schmid-Hempel, *Social immunity.* Curr Biol, 2007. **17**(16): p. R693-702.
197. Ramus, C., A. Hovasse, M. Marcellin, A.M. Hesse, E. Mouton-Barbosa, D. Bouyssie, S. Vaca, C. Carapito, K. Chaoui, C. Bruley, J. Garin, S. Cianferani, M. Ferro, A. Van Dorssaeler, O. Burlet-Schiltz, C. Schaeffer, Y. Coute, and A. Gonzalez de Peredo, *Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset.* J Proteomics, 2016. **132**: p. 51-62.
198. Chourey, K., J. Jansson, N. VerBerkmoes, M. Shah, K.L. Chavarria, L.M. Tom, E.L. Brodie, and R.L. Hettich, *Direct cellular lysis/protein extraction protocol for soil metaproteomics.* J Proteome Res, 2010. **9**(12): p. 6615-22.
199. Leary, D.H., W.J.t. Hervey, J.R. Deschamps, A.W. Kusterbeck, and G.J. Vora, *Which metaproteome? The impact of protein extraction bias on metaproteomic analyses.* Mol Cell Probes, 2013. **27**(5-6): p. 193-9.
200. Keiblinger, K.M., I.C. Wilhartitz, T. Schneider, B. Roschitzki, E. Schmid, L. Eberl, K. Riedel, and S. Zechmeister-Boltenstern, *Soil metaproteomics – Comparative evaluation of protein extraction protocols.* Soil Biology & Biochemistry, 2012. **54**(15-10): p. 14-24.
201. Chatterjee, S., G.S. Stupp, S.K. Park, J.C. Ducom, J.R. Yates, 3rd, A.I. Su, and D.W. Wolan, *A comprehensive and scalable database search system for metaproteomics.* BMC Genomics, 2016. **17**(1): p. 642.
202. Wolf, J.B., *Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial.* Mol Ecol Resour, 2013. **13**(4): p. 559-72.
203. Jagtap, P., T. McGowan, S. Bandhakavi, Z.J. Tu, S. Seymour, T.J. Griffin, and J.D. Rudney, *Deep metaproteomic analysis of human salivary supernatant.* Proteomics, 2012. **12**(7): p. 992-1001.
204. Xiong, W., R.J. Giannone, M.J. Morowitz, J.F. Banfield, and R.L. Hettich, *Development of an enhanced metaproteomic approach for deepening the microbiome characterization of the human infant gut.* J Proteome Res, 2015. **14**(1): p. 133-41.
205. Bourmaud, A., S. Gallien, and B. Domon, *CHAPTER 2 High Resolution/Accurate Mass Targeted Proteomics*, in *Quantitative Proteomics*. 2014, The Royal Society of Chemistry. p. 26-47.
206. Haoudi, A. and H. Bensmail, *Bioinformatics and data mining in proteomics.* Expert Rev Proteomics, 2006. **3**(3): p. 333-43.

Annexes

Annexe 1

Tableau Annexe 1 : Valeurs d'abondance relative des 7 protéines qui contribuent le plus significativement aux différences (présentées dans la Figure I-5) dans l'étude des marqueurs spléniques de l'immunosénescence.

			Abondance moyenne par groupe				Statistiques
Numéro d'accession	Nom de la protéine	Nom de la protéine dans l'ACP	**Young PBS**	**Young LPS**	**Old PBS**	**Old LPS**	Anova p-value
P20029	78 kDa glucose-regulated protein	GRP_78kDa	1,12E+08	1,09E+08	1,58E+08	1,72E+08	1,78E-03
P10107	Annexin A1	Annexin_A1	5,17E+07	1,26E+08	4,84E+07	5,78E+07	1,07E-03
P51437	Cathelin-related antimicrobial peptide	Cathelin	2,47E+07	8,40E+07	2,35E+07	3,18E+07	8,96E-03
O35744	Chitinase-like protein 3	Chitinase_like_3	4,67E+07	1,25E+08	2,17E+08	1,74E+08	2,82E-04
P08113	Endoplasmin	Endoplasmin	7,98E+07	7,88E+07	1,62E+08	1,39E+08	1,38E-08
P08905	Lysozyme C-2	LysozymeC2	4,67E+07	9,79E+07	4,47E+07	5,53E+07	2,55E-03
P31725	Protein S100-A9	ProteinS100	3,59E+07	1,04E+08	4,69E+07	4,64E+07	1,12E-03

Annexe 2

Tableau Annexe 2 : Valeurs d'abondance relative des protéines qui varient le plus dans l'ACP présentée en Figure I-9, et dont les coordonnées de l'ACP sont présentées en Tableau 1 et Tableau 2, dans l'étude évolutive de l'inflammation.

		Abondance moyenne par groupe							Statistiques
Numéro d'accession	Nom de la protéine	UCP2_PBS	UCP2_LPS	UCP2-KO_PBS	UCP2-KO_LPS	UCP2_antiIL10_PBS	UCP2_antiIL10_LPS	UCP2-KO_antiIL10_LPS	Anova
B2RXS4	Plexin-B2	3,20E+07	3,67E+07	4,30E+07	4,07E+07	2,81E+07	3,38E+07	3,97E+07	8,14E-03
O35639	Annexin A3	3,98E+08	3,44E+08	3,04E+08	3,49E+08	3,33E+08	4,30E+08	3,07E+08	1,10E-02
O54734	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase 48 kDa subunit	3,48E+08	3,27E+08	3,35E+08	3,15E+08	2,95E+08	3,92E+08	3,48E+08	7,88E-03
P02089	Hemoglobin subunit beta-2	1,97E+08	1,20E+08	5,56E+10	6,00E+10	2,26E+08	1,84E+08	6,33E+10	1,02E-18
P04186	Complement factor B	9,53E+07	1,01E+08	6,75E+07	7,96E+07	8,23E+07	1,00E+08	7,78E+07	1,00E-02
P06909	Complement factor H	2,40E+08	2,27E+08	1,51E+08	2,06E+08	2,19E+08	2,54E+08	1,57E+08	3,26E-03
P20029	78 kDa glucose-regulated protein	7,55E+09	7,41E+09	7,30E+09	6,74E+09	6,26E+09	1,06E+10	7,48E+09	1,36E-03
P27046	Alpha-mannosidase 2	4,03E+07	3,25E+07	3,36E+07	3,24E+07	3,18E+07	3,89E+07	3,34E+07	3,75E-03
P32261	Antithrombin-III	3,67E+08	3,43E+08	2,64E+08	2,73E+08	2,79E+08	3,80E+08	2,72E+08	7,61E-03
P48453	Serine/threonine-protein phosphatase 2B catalytic subunit beta isoform	1,15E+07	1,09E+07	8,70E+06	9,52E+06	1,19E+07	1,17E+07	7,06E+06	1,63E-03
P61961	Ubiquitin-fold modifier 1	5,30E+07	4,91E+07	5,57E+07	4,17E+07	4,19E+07	6,05E+07	4,79E+07	4,71E-03
P62307	Small nuclear ribonucleoprotein F	8,72E+07	8,97E+07	8,32E+07	8,30E+07	9,82E+07	7,56E+07	8,01E+07	7,30E-03
P62500	TSC22 domain family protein 1	1,33E+07	1,40E+07	7,45E+06	8,97E+06	1,17E+07	1,25E+07	9,11E+06	1,42E-02

Q5XJY5	Coatomer subunit delta	4,98E+08	4,46E+08	4,83E+08	4,48E+08	4,39E+08	5,16E+08	4,85E+08	1,86E-03
Q61703	Inter-alpha-trypsin inhibitor heavy chain H2	6,44E+07	6,52E+07	5,33E+07	6,07E+07	5,07E+07	7,68E+07	5,50E+07	7,00E-03
Q6GV12	3-ketodihydrosphingosine reductase	1,49E+07	1,30E+07	9,05E+06	8,57E+06	1,20E+07	1,42E+07	8,90E+06	1,07E-04
Q6ZQA0	Neurobeachin-like protein 2	1,54E+07	1,38E+07	1,04E+07	1,31E+07	1,39E+07	1,55E+07	1,11E+07	4,18E-03
Q78IK2	Up-regulated during skeletal muscle growth protein 5	1,59E+07	1,86E+07	2,39E+07	2,66E+07	1,94E+07	1,76E+07	3,10E+07	6,96E-04
Q80WW9	DDRGK domain-containing protein 1	1,80E+07	2,01E+07	1,74E+07	1,32E+07	1,23E+07	2,33E+07	1,86E+07	2,11E-03
Q8BFR5	Elongation factor Tu, mitochondrial	6,75E+08	6,11E+08	6,51E+08	6,68E+08	7,57E+08	6,61E+08	6,44E+08	1,49E-02
Q8BHN3	Neutral alpha-glucosidase AB	9,21E+08	8,43E+08	8,29E+08	8,40E+08	8,85E+08	9,94E+08	8,61E+08	2,39E-03
Q8BYB9	Protein O-glucosyltransferase 1	1,29E+07	1,19E+07	1,82E+07	1,74E+07	1,31E+07	1,28E+07	1,85E+07	2,94E-05
Q8JZX4	Splicing factor 45	1,75E+07	1,70E+07	2,56E+07	2,36E+07	1,96E+07	1,78E+07	2,32E+07	2,75E-03
Q8K1R3	Polyribonucleotide nucleotidyltransferase 1, mitochondrial	1,24E+07	1,44E+07	2,63E+07	2,29E+07	1,43E+07	1,19E+07	2,43E+07	2,29E-05
Q8R0F6	Integrin-linked kinase-associated serine/threonine phosphatase 2C	1,82E+07	1,97E+07	2,57E+07	2,32E+07	2,62E+07	1,80E+07	2,52E+07	1,15E-02
Q9CSN1	SNW domain-containing protein 1	6,85E+07	6,46E+07	5,87E+07	6,14E+07	7,91E+07	7,49E+07	6,46E+07	1,22E-02
Q9CX86	Heterogeneous nuclear ribonucleoprotein A0	5,76E+08	5,26E+08	6,76E+08	6,80E+08	6,44E+08	5,90E+08	6,49E+08	1,67E-03
Q9CXI5	Mesencephalic astrocyte-derived neurotrophic factor	7,13E+08	6,78E+08	6,20E+08	5,99E+08	5,57E+08	9,26E+08	6,62E+08	7,31E-03

Q9D0F3	Protein ERGIC-53	3,47E+08	3,42E+08	3,03E+08	2,81E+08	2,43E+08	4,21E+08	3,40E+08	3,51E-03
Q9D1L0	Coiled-coil-helix-coiled-coil-helix domain-containing protein 2	3,36E+07	3,83E+07	5,04E+07	4,94E+07	3,49E+07	3,25E+07	5,33E+07	1,98E-04
Q9D662	Protein transport protein Sec23B	1,40E+08	1,25E+08	1,31E+08	1,07E+08	1,12E+08	1,47E+08	1,26E+08	2,57E-04
Q9D6Z1	Nucleolar protein 56	3,87E+08	3,55E+08	4,25E+08	4,21E+08	5,26E+08	4,21E+08	4,09E+08	1,28E-02
Q9D7N9	Adipocyte plasma membrane-associated protein	6,91E+07	6,86E+07	5,70E+07	6,77E+07	6,28E+07	7,52E+07	6,26E+07	7,41E-03
Q9DBG6	Dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 2	4,42E+08	4,19E+08	4,12E+08	3,74E+08	3,64E+08	4,94E+08	4,44E+08	5,34E-03
Q9ERE7	LDLR chaperone MESD	3,65E+07	3,64E+07	2,76E+07	2,92E+07	3,58E+07	3,53E+07	2,84E+07	3,34E-05
Q9ERF3	WD repeat-containing protein 61	4,13E+07	4,53E+07	6,55E+07	5,88E+07	4,52E+07	4,52E+07	6,18E+07	1,31E-05
Q9ES97	Reticulon-3	2,92E+07	3,14E+07	3,99E+07	4,20E+07	3,34E+07	3,06E+07	4,11E+07	2,39E-04
Q9ESP1	Stromal cell-derived factor 2-like protein 1	1,38E+08	1,23E+08	1,30E+08	1,21E+08	1,11E+08	1,68E+08	1,46E+08	1,11E-02
Q9JJF3	Bifunctional lysine-specific demethylase and histidyl-hydroxylase NO66	4,75E+06	3,59E+06	3,74E+06	3,29E+06	6,28E+06	4,88E+06	3,62E+06	1,50E-02
Q9JKR6	Hypoxia up-regulated protein 1	1,02E+09	1,03E+09	9,21E+08	8,73E+08	8,27E+08	1,27E+09	1,02E+09	6,49E-03
Q9QXA5	U6 snRNA-associated Sm-like protein LSm4	8,61E+07	7,99E+07	7,33E+07	7,70E+07	9,75E+07	7,69E+07	7,03E+07	9,90E-03
Q9QZI9	Serine incorporator 3	5,32E+06	7,23E+06	0	0	8,53E+06	4,33E+06	0,00E+00	5,06E-09
Q9WV98	Mitochondrial import inner membrane translocase subunit Tim9	2,54E+07	2,59E+07	3,36E+07	2,79E+07	3,33E+07	2,79E+07	3,23E+07	8,24E-03

Annexe 3

Tableau Annexe 3 : Valeurs d'abondance relative (log2) des protéines différentielles et présentées sur les cartes en Figure II-18 et Figure II-19 dans l'étude du rôle du BAT dans le contrôle de la balance énergétique.

En rouge sont présentées les protéines altérées par la température (seule la fonction « OXPHOS » est présentée ici) ; en vert les protéines altérées par le régime alimentaire et en violet par la lignée. Dans les noms des groupes, H= HighNST ; L= LowNST ; W= Warm ; C= Cold ; F= Fat ; c= chow.

			Abondance moyenne par groupe (valeur en log2)								Statistiques
Numéro d'accession	Nom de la protéine	Nom; fonction sur la « heatmap »	HWc	HWF	LCF	LWc	LWF	HCF	HCc	LCc	p-value
Q8JZN5	Acyl-CoA dehydrogenase family member 9, mitochondrial	ACAD9;OXPHOS	27,811	27,949	28,547	27,700	27,887	28,346	28,252	28,309	9,95E-09
P13619	ATP synthase F(0) complex subunit B1, mitochondrial	AT5F1;OXPHOS	29,419	29,232	29,765	28,875	29,120	29,723	29,642	29,908	3,61E-07
P29418	ATP synthase subunit epsilon, mitochondrial	ATP5E;OXPHOS	29,382	29,203	29,245	29,026	28,884	29,456	29,613	29,511	3,02E-04
O75947	ATP synthase subunit d, mitochondrial	ATP5H;OXPHOS	29,042	28,682	29,258	28,603	28,590	29,292	29,318	29,387	1,95E-08
Q00361	ATP synthase subunit e, mitochondrial	ATP5I;OXPHOS	30,203	30,330	30,635	30,249	30,363	30,481	30,620	30,515	1,23E-03
Q03265	ATP synthase subunit alpha, mitochondrial	ATPA;OXPHOS	30,884	30,633	30,968	30,480	30,565	30,903	31,032	31,156	4,83E-05
P10719	ATP synthase subunit beta, mitochondrial	ATPB;OXPHOS	31,991	31,737	31,903	31,657	31,745	31,862	31,960	32,110	2,90E-02
Q91VR2	ATP synthase subunit gamma, mitochondrial	ATPG;OXPHOS	29,187	28,803	29,178	28,842	28,384	29,271	29,494	29,519	1,86E-05
Q9DB20	ATP synthase subunit O, mitochondrial	ATPO;OXPHOS	29,556	29,441	29,629	29,173	29,157	29,741	29,908	29,976	5,87E-06
Q8R1S0	Ubiquinone biosynthesis monooxygenase COQ6, mitochondrial	COQ6;OXPHOS	29,078	29,297	29,333	28,958	28,786	29,339	29,386	29,092	4,14E-03
P00405	Cytochrome c oxidase subunit 2	COX2;OXPHOS	33,342	33,427	33,646	33,190	33,031	33,704	33,462	33,214	8,54E-04
P10888	Cytochrome c oxidase subunit 4 isoform 1, mitochondrial	COX41;OXPHOS	31,182	31,331	31,568	31,007	31,076	31,670	31,388	31,119	5,86E-05
B0VYY2	Cytochrome c oxidase subunit 5A, mitochondrial	COX5A;OXPHOS	28,372	28,428	28,890	28,332	28,314	28,929	28,684	28,773	4,88E-08
Q5S3G4	Cytochrome c oxidase subunit 5B, mitochondrial	COX5B;OXPHOS	31,712	31,735	31,881	31,591	31,472	31,999	31,721	31,521	1,54E-02
A1XQT2	Cytochrome c oxidase subunit 6C	COX6C;OXPHOS	29,955	30,139	30,030	30,292	30,542	30,038	29,832	29,795	1,43E-04

P17665	Cytochrome c oxidase subunit 7C, mitochondrial	COX7C;OXPHOS	30,240	30,874	30,623	30,357	30,071	30,855	30,740	30,287	2,14E-02
O14548	Cytochrome c oxidase subunit 7A-related protein, mitochondrial	COX7R;OXPHOS	28,318	28,441	28,976	27,959	28,098	29,076	28,741	28,692	1,16E-09
Q99LC5	Electron transfer flavoprotein subunit alpha, mitochondrial	ETFA;OXPHOS	31,364	31,499	31,643	31,309	31,399	31,657	31,502	31,234	4,66E-03
Q9DCW4	Electron transfer flavoprotein subunit beta	ETFB;OXPHOS	32,297	32,400	32,479	32,263	32,239	32,565	32,452	32,148	6,32E-03
Q921G7	Electron transfer flavoprotein-ubiquinone oxidoreductase, mitochondrial	ETFD;OXPHOS	29,927	30,241	30,685	29,699	29,906	30,653	30,571	30,354	2,96E-13
Q9CQ75	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 2	NDUA2;OXPHOS	32,276	32,435	32,868	32,372	32,224	32,864	32,769	32,567	1,13E-08
Q62425	Cytochrome c oxidase subunit NDUFA4	NDUA4;OXPHOS	31,239	31,524	31,523	31,059	31,157	31,731	31,560	31,390	3,48E-04
Q63362	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 5	NDUA5;OXPHOS	31,544	31,678	32,104	31,594	31,419	32,043	31,822	31,726	4,76E-07
Q9CQZ5	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 6	NDUA6;OXPHOS	29,288	29,531	29,907	29,065	29,005	30,183	30,049	29,862	7,73E-12
Q9Z1P6	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 7	NDUA7;OXPHOS	30,602	30,821	31,280	30,601	30,576	31,246	31,220	30,932	4,40E-08
P51970	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 8	NDUA8;OXPHOS	30,968	31,104	31,370	30,998	30,679	31,484	31,284	31,014	4,04E-05
Q5BK63	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 9, mitochondrial	NDUA9;OXPHOS	29,951	30,131	30,778	29,768	29,630	30,882	30,636	30,512	1,86E-10
Q561S0	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 10, mitochondrial	NDUAA;OXPHOS	29,729	29,624	30,182	29,528	29,335	30,018	29,899	30,033	7,13E-07
Q0MQ88	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 13	NDUAD;OXPHOS	30,509	30,622	30,935	30,412	29,987	30,979	30,763	30,573	1,34E-04
Q9CQC7	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 4	NDUB4;OXPHOS	30,313	30,613	30,977	30,148	30,142	31,103	31,020	30,745	8,05E-10
Q9CQH3	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 5, mitochondrial	NDUB5;OXPHOS	28,021	28,300	28,658	27,674	27,721	28,815	28,701	28,462	4,92E-09
Q02368	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 7	NDUB7;OXPHOS	28,931	29,464	30,160	28,575	28,836	29,956	29,894	29,396	1,25E-07
Q9CQJ8	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 9	NDUB9;OXPHOS	30,200	30,471	30,755	29,777	29,768	30,872	30,786	30,425	3,52E-08

Q0MQF8	NADH dehydrogenase [ubiquinone] 1 subunit C2	NDUC2;OXPHOS	30,456	30,748	31,144	30,293	30,365	31,236	31,033	30,799	2,66E-09
Q9JKL4	NADH dehydrogenase [ubiquinone] 1 alpha subcomplex assembly factor 3	NDUF3;OXPHOS	28,724	29,000	29,309	28,777	28,649	29,381	29,279	29,058	1,41E-08
Q91VD9	NADH-ubiquinone oxidoreductase 75 kDa subunit, mitochondrial	NDUS1;OXPHOS	30,458	30,586	31,021	30,478	30,322	31,079	30,888	30,710	5,95E-10
P17694	NADH dehydrogenase [ubiquinone] iron-sulfur protein 2, mitochondrial	NDUS2;OXPHOS	30,176	30,441	30,980	30,017	29,912	31,016	30,843	30,614	3,25E-12
P23709	NADH dehydrogenase [ubiquinone] iron-sulfur protein 3, mitochondrial	NDUS3;OXPHOS	31,409	31,566	31,906	31,393	31,227	31,986	31,788	31,646	1,34E-07
O43181	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial	NDUS4;OXPHOS	31,821	31,917	32,340	31,862	31,690	32,351	32,254	32,016	2,74E-07
Q99LY9	NADH dehydrogenase [ubiquinone] iron-sulfur protein 5	NDUS5;OXPHOS	29,019	29,083	29,563	28,967	28,807	29,608	29,353	29,301	1,23E-07
P52503	NADH dehydrogenase [ubiquinone] iron-sulfur protein 6, mitochondrial	NDUS6;OXPHOS	31,240	31,276	31,714	31,250	31,105	31,649	31,533	31,374	1,58E-05
Q9DC70	NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial	NDUS7;OXPHOS	28,067	28,865	29,356	28,154	27,821	29,197	28,962	29,096	6,21E-06
Q0MQI6	NADH dehydrogenase [ubiquinone] flavoprotein 1, mitochondrial	NDUV1;OXPHOS	30,157	30,450	30,885	30,050	30,007	30,960	30,758	30,538	1,65E-12
P19404	NADH dehydrogenase [ubiquinone] flavoprotein 2, mitochondrial	NDUV2;OXPHOS	30,597	30,677	31,196	30,619	30,323	31,175	30,942	30,808	6,34E-08
Q9CZ13	Cytochrome b-c1 complex subunit 1, mitochondrial	QCR1;OXPHOS	32,220	32,236	32,473	32,081	32,058	32,478	32,269	32,183	7,07E-04
Q9DB77	Cytochrome b-c1 complex subunit 2, mitochondrial	QCR2;OXPHOS	30,600	30,764	31,040	30,518	30,449	31,086	30,923	30,862	1,67E-07
P00129	Cytochrome b-c1 complex subunit 7	QCR7;OXPHOS	31,743	31,682	31,887	31,591	31,478	32,032	31,854	31,700	9,08E-05
O14949	Cytochrome b-c1 complex subunit 8	QCR8;OXPHOS	31,900	32,042	32,080	31,611	31,708	32,056	31,951	31,966	3,04E-02
Q9UDW1	Cytochrome b-c1 complex subunit 9	QCR9;OXPHOS	28,930	29,061	29,294	28,528	28,124	29,531	29,192	29,142	7,80E-05
Q5RJQ7	Succinate dehydrogenase assembly factor 2, mitochondrial	SDHF2;OXPHOS	29,455	29,511	29,827	29,621	29,614	30,082	29,951	29,430	3,89E-04
P04575	Mitochondrial brown fat uncoupling protein 1	UCP1;OXPHOS	31,927	32,276	33,181	31,511	31,519	33,522	33,243	33,117	1,71E-14
Q9CR68	Cytochrome b-c1 complex subunit Rieske, mitochondrial	UCRI;OXPHOS	31,872	32,073	32,194	31,650	31,768	32,301	32,072	32,069	2,66E-04

P09809	Apolipoprotein A-I	APOA1;Cholesterol transport	30,670	31,109	31,179	30,634	31,360	31,008	30,914	30,929	5,94E-05
Q91V92	ATP-citrate synthase	ACLY;de novo lipid synthesis/TCA cyle	28,766	28,177	27,611	28,483	28,221	28,271	29,319	29,592	3,52E-04
Q13011	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial	ECH1;Fatty acid degradation	31,095	31,418	31,532	31,247	31,419	31,422	31,187	31,071	1,54E-05
Q64428	Trifunctional enzyme subunit alpha, mitochondrial	ECHA;Fatty acid degradation	31,452	31,632	31,721	31,394	31,630	31,700	31,527	31,313	3,35E-04
Q71SP7	Fatty acid synthase	FAS;Fatty acid synthesis	30,581	30,142	29,565	30,358	29,816	30,154	31,116	31,154	1,16E-05
Q71RI9	Kynurenine--oxoglutarate transaminase 3	KAT3;Amino acid metabolism	27,945	27,868	27,161	27,120	26,963	27,925	27,956	27,394	1,07E-03
P26443	Glutamate dehydrogenase 1, mitochondrial	DHE3;Amino acid metabolism/link to TCA cycle	30,716	30,676	31,182	30,880	30,996	30,725	30,747	31,117	7,55E-06
P11352	Glutathione peroxidase 1	GPX1;Antioxidant	30,377	30,414	30,372	30,289	30,076	30,547	30,395	30,089	3,58E-03
Q8HXP0	Superoxide dismutase [Mn], mitochondrial	SODM;Antioxidant	30,381	30,536	30,114	29,825	29,953	30,697	30,363	29,856	1,74E-05
P62072	Mitochondrial import inner membrane translocase subunit Tim10	TIM10;Chaperone-mediated protein transport	29,183	29,069	28,549	28,785	28,358	29,318	29,140	28,667	6,94E-09
Q9WV97	Mitochondrial import inner membrane translocase subunit Tim9	TIM9;Chaperone-mediated protein transport	29,066	28,975	28,769	28,635	28,515	29,239	29,016	28,706	7,16E-06
Q9CQR4	Acyl-coenzyme A thioesterase 13	ACO13;Fatty acid biosynthesis	28,963	29,342	28,548	28,469	28,611	29,374	29,117	28,745	4,28E-05
Q29448	Hydroxymethylglutaryl-CoA lyase, mitochondrial	HMGCL;Ketogenesis	28,772	28,778	29,323	29,427	29,554	28,585	28,716	29,215	8,51E-10
P22392	Nucleoside diphosphate kinase B	NDKB;Nucleoside metabolism	29,878	29,983	29,371	29,637	29,608	29,793	29,852	29,871	1,83E-03
Q68FS4	Cytosol aminopeptidase	AMPL;Proteolysis	28,548	28,760	29,206	28,954	29,032	28,506	28,819	29,143	2,93E-04

Annexe 4

Tableau Annexe 4 : Valeurs d'abondance relative des protéines impliquées dans la glycolyse et la néoglucogenèse (présentées sur la carte KEGG en Figure II-24) dans l'étude de la modification du protéome chez une espèce saisonnière.

			Abondance moyenne par groupe		Statistiques
Numéro d'accession	Nom de la protéine	Nom de la protéine sur la carte KEGG	**Obésogène**	**Diabétogène**	T-test p-value
P36871	Phosphoglucomutase-1	5.4.2.2	9,83E+09	1,50E+10	1,16E-02
P17858	ATP-dependent 6-phosphofructokinase, liver type	2.7.1.11	4,74E+08	3,48E+08	7,90E-03
P09467	Fructose-1,6-bisphosphatase 1	3.1.3.11	2,23E+10	1,71E+10	1,30E-02
P05062	Fructose-bisphosphate aldolase B	4.1.2.13	1,24E+11	8,45E+10	7,42E-04
P30613	Pyruvate kinase PKLR	2.7.1.40	6,13E+09	3,55E+09	8,09E-06
P08559	Pyruvate dehydrogenase E1 component subunit alpha, somatic form, mitochondrial	1.2.4.1	2,54E+09	1,69E+09	2,71E-03
P49189	4-trimethylaminobutyraldehyde dehydrogenase	1.2.1.3	2,77E+09	2,28E+09	4,29E-02

Annexe 5

Tableau Annexe 5 : Valeurs d'abondance relative des protéines qui contribuent le plus (cos² projection > 0,9) à l'axe 1 de l'ACP présentée en Figure II-29, qui permettent donc de séparer les Reines des Ouvrières, dans l'étude des interactions entre rôle social et variabilité de l'espérance de vie chez la fourmi.

		Abondance moyenne par groupe			Statistiques
Numéro d'accession	Nom de la protéine	**Fourrageuses**	**Domestiques**	**Reines**	Anova p-value
A0A0J7JZB2	Oxysterol-binding protein 1-like protein	1,08E+06	7,70E+05	0,00E+00	3,06E-07
A0A0J7K0B0	Thioredoxin-related transmembrane protein	3,42E+06	3,95E+06	8,92E+06	4,37E-05
A0A0J7K6T3	Cytochrome p450 9e2-like protein	9,18E+06	1,14E+07	2,33E+07	8,31E-04
A0A0J7KLG9	Triosephosphate isomerase	1,82E+09	1,43E+09	2,98E+08	1,89E-05
A0A0J7KLV6	Peptidyl-prolyl cis-trans isomerase nima-interacting 1-like protein	1,29E+07	1,16E+07	0,00E+00	1,71E-06
A0A0J7KQ60	Cklf-like marvel transmembrane domain-containing protein 4-like protein	1,05E+07	2,19E+07	4,82E+07	2,59E-04
A0A0J7KSL9	Nadh dehydrogenase	1,94E+07	1,32E+07	0,00E+00	5,76E-06
A0A0J7KVW5	Protein-l-isoaspartate-d-aspartate-o-methyltransferase	5,51E+06	4,33E+06	0,00E+00	3,10E-06
A0A0J7KWR2	Cytosolic purine 5-nucleotidase isoform x3	2,44E+07	3,02E+07	7,79E+07	2,90E-06
A0A0J7L3C0	Nucleolar protein 56	8,76E+06	8,95E+06	3,10E+07	1,47E-09
A0A0J7L4N1	Cd109 antigen	1,04E+08	1,56E+08	3,94E+08	5,59E-05
A0A0J7L648	Reticulon-like protein	1,83E+09	2,45E+09	5,81E+09	7,41E-07
A0A0J7L6E6	L-2-hydroxyglutarate mitochondrial	1,31E+07	1,71E+07	9,47E+07	1,22E-07
A0A0J7N0M1	Vacuolar protein-sorting-associated protein 25	3,88E+06	2,66E+06	0,00E+00	1,70E-04
A0A0J7NEN5	Striatin-3 isoformx2	5,97E+06	4,57E+06	0,00E+00	1,55E-05
A0A0J7NQ97	Leucine-rich repeat-containing protein 20	2,03E+06	1,97E+06	0,00E+00	3,64E-05
A0A0J7P466	Ferritin	0,00E+00	0,00E+00	4,31E+06	2,32E-12

Annexe 6

Tableau Annexe 6 : Valeurs d'abondance relative moyenne par site et par fonction (présentées dans la Figure III-13) dans l'étude métaprotéomique des deux sols alpins.
Les numéros d'accession des protéines impliquées dans chaque fonction sont données.

		Abondance moyenne par fonction et par site		Statistiques
Fonction	Numéro d'accession	**LV**	**PF**	p-value
ELECT_TRANSF	B1ZQF3;B4D2U5;F8PRQ0;G8NRQ1;P51131;P53573;Q01YB7	2,91E+07	3,93E+07	4,05E-05
PROT_SYNTH	A0A062XKM5;O94083;F8NDN0;A0A067Q4R4;A0A075E8T2;A0A0C2XA60;A0A0C2YDS1;A0A0C9U3Z9;B0D8U9;A0A0D0CAQ1;A0A0E3YV98;A0A0H4TAN6;Q877B9;Q1BE87;A1USL2;A5ETE1;A5DPE3;A7NS01;A6KYK9;A6L3G7;A6U7M8;A7HWP7;A8HTW6;A8PB71;A9Q1C8;Q4JT41;B4CUH1;B4CY10;B4D9P4;B5JK07;B6JH82;B8ELG5;B9XH65;B9XK85;D8PPR4;E1IEV2;E8WYT1;G2LK47;I6B0W6;L7X4P8;O14460;O59949;P10255;Q53871;P33334;P34825;P41752;Q01QC0;Q01SX2;Q01W31;Q01WB2;Q02CH0;Q3SVW2;Q3SRH5;Q1AU14;Q1AU26;Q89J81;Q211E6;Q47LJ1;Q6C024;Q89WA9;Q96X45;Q981F7;Q9Y713;V2XH67	2,33E+07	2,39E+07	2,26E-03
RIBO_PROT	A8N9E4;B0D5Q4;D8QA67;Q9HE74;V2XJ47;A0A067TWV7;A0A0C3C0H7;A0A0C3C5R2;A0A0C3C6H3;A0A0C3E9D5;A0A0C3G3M1;A0A0D0DJL3;A0A0G3EJH5;A0A0H4T7L8;Q01W96;A0A0H4T6M5;A0A0H4TSJ2;A0PM75;A0QYY6;A8M515;A4YJF2;Q3SSY4;Q98N61;Q3SSW0;Q89J94;Q211G0;A5ELK5;Q6N4V6;Q89MW4;Q6AD05;Q92L39;A8L6C4;A9WH60;B8ELG4;B8ELF9;Q89JA6;Q89J80;B9XFN2;B4CUY4;B4CUZ0;Q1QN05;Q211G5;B8ELF8;Q89J95;C0NKW7;Q1IU82;C5C0I5;D2JY95;D8PL16;E8UXE3;G8NQ77;G8NV81;I3ZBR9;I3ZL06;Q9P3T6;O43105;P05750;P19115;Q01W87;Q01W91;Q01W92;Q01W98;Q01WA2;Q01WA7;Q022G4;Q07KL8;Q3SSW1;Q21CT9;Q2IXS2;Q2IXP8;Q1IHI9;Q1QMN7;Q3SSW2;Q2J3K2;Q6NDP1;Q7RV75;Q7RVI1;Q7RVN0;Q89J70;Q8X034	1,41E+07	1,81E+07	1,24E-04
MAL_DH	Q6G1M0;Q1QQR2	1,12E+07	1,33E+07	4,19E-02
XYL_ISO	B0RIF1;B9XH30;Q022S9	7,14E+06	1,23E+07	1,06E-10
GLN_SYN	A0A067Q3S9;A0A0C3A1Z7;I0K0K5;A0A0D7ABK1;A0A0G3EH69;A8P5F5;B8GAV3;R7E3F9;B9XG68;C1F1J4;O93934;Q05542;P15623;P18819;P54388;Q027C2;Q02CY3;Q1ILV7;Q1IT55;Q59747;S5I8Z8	6,05E+07	4,51E+07	2,47E-04
TRANS_ALDO	A0LTY8;E8WXG1;J9MJK9;Q8U7I5	4,19E+07	2,65E+07	1,26E-02
ANTI_SIGMA	C1F4D1;Q02CU4;Q1INN3	3,54E+07	2,35E+07	1,18E-03
CIT_SYN	O00098;P94325	3,40E+07	1,94E+07	7,93E-05
PYR_Cxase	A0A0C3CM91;Q9HES8;Q9UUE1	2,28E+07	1,35E+07	1,30E-02

Margaux BENHAIM

Développements méthodologiques en protéomique quantitative pour mieux comprendre la biologie évolutive d'espèces non séquencées

L'analyse protéomique consiste en l'analyse qualitative et quantitative de l'ensemble des protéines exprimées dans une cellule ou tissu dans des conditions données (protéome). Les progrès instrumentaux en spectrométrie de masse et les avancées bioinformatiques des dernières années ont permis d'imposer ce domaine dans les sciences de la vie. Diverses stratégies protéomiques permettent ainsi, aujourd'hui, d'identifier et quantifier plusieurs centaines/milliers de protéines dans un échantillon complexe, ce qui permet classiquement de caractériser les états physiopathologiques. En revanche, la protéomique est un outil émergent en biologie évolutive. Ce domaine vise à comprendre les déterminants de la diversité des organismes présents sur Terre et de leur « fonctionnement », notamment leurs adaptations à certaines contraintes environnementales.

L'objectif de cette thèse était d'étudier, de l'organe à l'écosystème, les variations protéomiques induites par des changements environnementaux, tout en adaptant les différentes étapes de l'analyse à chaque type d'échantillons, à chaque organisme, de la préparation d'échantillons à l'analyse des données. Grâce à la mise en place d'une stratégie de séquençage *de novo* quantitative originale, ces travaux de thèse ont été l'occasion d'étudier le rôle du tissu adipeux brun dans la protection contre l'obésité chez le campagnol, espèce dont le génome n'est pas séquencé. D'autres traits particuliers ont été explorés, tels que l'obésité réversible du microcèbe, ou encore les interactions entre socialité et longévité chez la fourmi. Les solutions logicielles envisagées ne permettant de quantifier de manière robuste des peptides identifiés par séquençage de novo à partir d'échantillons fractionnés, nous avons ainsi établi que le préfractionnement permet d'obtenir une meilleure couverture de protéome. En revanche, sans préfractionnement, le séquençage *de novo* produit un gain indéniable. Enfin, en étudiant le métaprotéome de communautés biotiques des sols alpins, nous avons mis en évidence l'intérêt de combiner protéomique et génomique, afin d'établir la banque de données protéiques la plus appropriée, mais aussi pour « valider » les données protéomiques.

Mots-clés : Protéomique quantitative – Spectrométrie de masse – Séquençage de novo – Organismes non séquencés – Bio-informatique.

Proteomics analysis corresponds to the qualitative and quantitative analysis of all proteins expressed in a cell or tissue under given conditions (proteome). Instrumental progresses in mass spectrometry and bioinformatics advances in recent years have allowed its establishment in life sciences. Diverse proteomics strategies thus allow identification and quantification of hundreds/thousands of proteins in complex samples, which classically allows physiopathological states to be characterized. However, proteomics is only emerging in the evolutionary biology field. This field aims at understanding the determinants of the diversity of organisms present on Earth and their "functioning", including their adaptations to certain environmental constraints.

The objective of this thesis was to study, from the organ to the eco-system, the proteomic variations induced by environmental changes, while adapting the different steps of the analysis to each type of sample, each organism, from sample preparation to data analysis. Through the introduction of an original quantitative *de novo* sequencing strategy, we studied the role of brown adipose tissue against obesity in a non-sequenced species: the vole. Other particular traits were explored, such as the reversible obesity of the grey mouse lemur or the interactions between sociality and longevity in the ant. The considered software solutions did not allow to robustly quantify peptides identified by *de novo* sequencing from fractionated samples, we thus determined that prefractionation allows for better proteome coverage. On the other hand, without prefractionation, *de novo* sequencing produces an undeniable gain. Finally, by studying the metaproteome of alpine soil biotic communities, we have highlighted the advantage of combining proteomics and genomics, in order to establish the most appropriate protein database and to "validate" proteomics data.

Key words: Quantitative proteomics – Mass spectrometry – *de novo* sequencing – Non-sequenced organisms – Bio-informatics.