UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES CHIMIQUES

IPHC – UMR 7178

THÈSE présentée par :

Leslie MULLER

soutenue le : **21 décembre 2017**

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

Discipline/Spécialité : Chimie – Chimie Analytique

Développements de méthodes de préparation d'échantillons pour l'analyse protéomique quantitative : application à la recherche de biomarqueurs de pathologies

THÈSE dirigée par :

Mme Sarah CIANFERANI	Directrice de recherche, Université de Strasbourg - CNRS
Mme Christine CARAPITO co-encadrant	Chargée de recherche, Université de Strasbourg – CNRS

RAPPORTEURS :

M. Thierry RABILLOUD	Directeur de recherche, CEA Grenoble
Mme Odile SCHILTZ	Directrice de recherche, Université de Toulouse - CNRS

A ceux que j'aime,

Mes parents,

Mon frère,

Eric,

Ma famille,

« Il faut savoir douter où il faut, se soumettre où il faut, croire où il faut »

Blaise Pascal

« L'éducation est l'arme la plus puissante qu'on puisse utiliser pour changer le monde »

Nelson Mendela

REMERCIEMENTS

Avant toute chose, je tiens à remercier les Drs Alain VAN DORSSELAER et Sarah CIANFERANI de m'avoir permis d'intégrer le laboratoire en tant que stagiaire M2, et de m'avoir permis de poursuivre en thèse au sein de cette structure qui offre de très bonnes conditions de travail et de bons outils pour faire une belle thèse.

Un grand merci à mes directrices de thèse, Sarah CIANFERANI et Christine CARAPITO de m'avoir encadrée pendant ces 3 années et quelques mois. Merci de m'avoir accordée de votre temps pour discuter science ou répondre à mes questions. Merci Sarah pour le bonbon magique qui m'a sauvée lors de notre dernier voyage en train. Christine, merci d'avoir pris soin de moi lors de notre périple (je ne pense pas que ce soit un gros mot !) transatlantique.

Je tiens à remercier l'ensemble de l'équipe du LSMBO d'avoir contribué de près ou de loin à ces travaux. Plus particulièrement, merci à Jean-Marc pour ses conseils précieux et son aide sur les Nano-Acquity. Merci à Hélène de m'avoir aidée à faire mes premiers pas en tant que responsable machine sur le TTOF 5600, et à Christine S. pour l'Impact-II. Merci à Fabrice d'avoir redémarré ma session x fois – je pense qu'au vu de ces nombreux redémarrages on peut raisonnablement dire que j'ai eu la poisse ! Merci à Danièle d'avoir pris le temps pour me former aux gels d'électrophorèse. Un grand merci à ceux avec qui je n'ai pas pu travailler directement mais qui ont su me prodiguer des conseils par-ci par-là et qui ont toujours fait preuve de sympathie lorsque j'avais des questions : Agnès, Fabrice (le deuxième), Alfred, François, Martine et Stella. Je n'oublie pas l'équipe des bio-informaticiens, toujours prête à solutionner nos problèmes : merci à Alex, Aymen et Patrick.

Merci à Véronique et Jérémy pour leur extrême gentillesse et leur bonne humeur qui m'ont permis de trouver et prendre mes marques en TP. Merci à toute l'équipe d'enseignement des TP de chimie analytique de l'ECPM pour leur aide et leur sympathie.

Merci à Laurence d'avoir pris le temps de répondre à mes questions aussi bien de sciences que celles concernant le BDD de l'IPHC. Merci également pour le soutien que vous m'avez accordé lors de la préparation de l'oral de la MRT, mais aussi pour les pistes d'après-thèse.

Passons aux « étudiants » du LSMBO. Merci à Marianne, madame « j'aime pas les saucisses », Aurélie, Blandine, Ludivine, Magali, Steve, Margaux, Nina et Charlotte. Merci à Georg de m'avoir

transmis (presque toutes) ses connaissances sur le fameux Impact-II… qui aura eu raison de mon optimisme à certains moments. Merci à Gauthier, monsieur Schoko-Bons, pour sa gentillesse et nos nombreuses discussions philosophiques des jeudis/vendredis soirs.

Un grand merci à Luc de m'avoir suivie dans l'aventure des « *Tube-Gel* ». Merci aussi pour les nombreux fous rires en salle bio (oups…) et pour les prises de tête avec ces traitements statistiques et les calculs de mise au point de cette fameuse gamme UPS ! J'ai vraiment pris plaisir à travailler avec toi ! Un grand merci d'avoir pris le temps de relire des parties de cette thèse.

David, merci d'avoir apporté ce beau projet au laboratoire… même si je me demande encore si je dois te remercier de travailler avec du pipi ! On a encore du pain sur la planche pour la poursuite de ce projet.

Changeons de bâtiment, et direction le bureau des garçons dans lequel je suis souvent allée me ressourcer. Merci à Maxime de m'avoir piqué mes chaussures et m'avoir pêchée avec LA règle (qui te transforme en tortue Ninja, et que, pour une raison obscure, tu adores mettre dans ta bouche) et le scotch le vendredi soir. Il faut bien admettre que ça fait du bien de craquer. Merci aussi pour nos conversations sur des sujets existentiels. Même si tu as du mal à le voir, je t'apprécie beaucoup ! Anthony, toi qui as refusé ma proposition d'aller manger un MacDo, je te remercie pour tes petits mots sympas et ton petit rire qui remonte le moral. Malgré ton dos imposant, tu restes une personne attachante que l'on a envie de soutenir. Au passage, désolée d'avoir mis mon grain de sel dans nombre de tes powerpoints… Oscar, tu tenais absolument à ce que je te site dans ma thèse[1], voilà c'est chose faite ! Merci pour ton humour (un peu pourri faut dire), tes histoires rocambolesques et ton accent qui fait voyager ! Merci à Thomas, pour les fois où tu étais de bonne humeur (kalach, kalach !) ! Je ne sais pas si je dois dire merci à Stéphane qui est venu jouer l'intrus – ou apporter un équilibre – dans le bureau des filles... Je rigole ! Merci Stéphane pour ton écoute (quand tu n'étais pas ultra-concentré sur ton ordinateur), ta gentillesse et les photos de ta puce qui redonnent le sourire. Merci à Guillaume pour sa sympathie (et bon courage avec les étudiants en TP !).

Enfin, à mes collègues filles avec qui je partage le bureau… J'ai apprécié et apprécie d'ailleurs toujours de pouvoir discuter aussi bien de sciences que de gâteaux, famille, politique, etc. avec toi Justine. Merci à vous toutes, Pauline, Joanna et Paola de m'avoir divertie quand j'en avais besoin ! Merci à Pauline et Justine pour vos débats – auxquels on sait qu'il n'y aura pas d'issue parce-que vous avez souvent des idées divergentes. Merci aussi pour les fous rires. Merci à

Joanna pour son calme et ses conseils de langue française, et merci à Paola pour ses histoires exotiques que l'on aime tant. A croire qu'il n'y a qu'aux gens qui parlent l'espagnol qu'il arrive des trucs improbables ! En tout cas, merci les filles de m'avoir supportée avec mon esprit critique.

Je n'oublie pas les anciens… Marion, Maxime (#Coton), Mathieu, Gilles (qui a été pour moi la personne qui explique le mieux au labo), Sarah (qui m'a tout appris), Diego, Johann, Guillaume (avec qui j'ai fait mon baptême de l'air) Kévin, Marine, Sébastian et Benoîît. Benoît, Benoît… Difficile de résumer ces nombreuses années passées dans le même bureau. Ça a été un vrai plaisir d'être dans ton dos, de pouvoir discuter librement et de pouvoir se confier sans être jugée. Même si la fin a été plus difficile (il fallait bien trouver un moyen pour que la séparation soit moins difficile à vivre), j'en garde de très bons souvenirs.

Je remercie mes amies d'enfance, Anne, Perrine, Sophie, leurs compagnons et petit bout. On ne se voit pas beaucoup mais on peut toujours discuter librement et avec simplicité. Malgré les nombreuses années qui ont passé, on partage toujours la même vision de la vie !

Non en derniers, je tiens à adresser un énorme merci à mes parents qui ont consenti de nombreux sacrifices pour que j'arrive jusqu'ici. Merci pour votre amour, votre dévouement. Merci de m'avoir facilité la tâche pendant ces nombreuses années d'études. Merci de m'avoir soutenue et d'avoir cru en moi. Merci à ma famille qui a toujours su m'apporter la déconnexion dont j'avais besoin avec les repas du dimanche midi/soir et avec l'arrivée de deux petits « frères », Litchy et Mowgly, et deux petits « neveux », Newton et Papuche qui ont agrandi la famille pendant ces 3 années. Merci à mon frère, même si tu ne le montres pas beaucoup, je sais que tu m'as toujours soutenue. Un grand merci à Eric, avec qui je partage mon quotidien et avec qui j'ai accueilli notre petite poupette à griffes au début de nos thèses et qui a apporté de la vie et de nombreux fous rires dans notre petit cocon. Merci de m'avoir supportée, d'avoir été mon confident. Merci pour ton calme, ton amour, ta complicité et ton soutien sans faille.

Finalement, un grand merci à toutes les personnes qui m'ont soutenue, de quelque manière que ce soit, pendant toutes ces années d'études.

[1] Hernandez, O., Pulay, P., Maitre, P. & Paizs, B. Zundel-type H-bonding in biomolecular ions. J Am Soc Mass Spectrom 25, 1511-1514 (2014).

SOMMAIRE

LISTE DES ABREVIATIONS

1D SDS-PAGE - Gel d'électrophorèse monodimensionnel en conditions dénaturantes

2D SDS-PAGE - Gel d'électrophorèse bidimensionnel

2D-LC-MS/MS - Analyse par chromatographie liquide bidimensionnelle couplée à la spectrométrie de masse en tandem

ACN – Acétonitrile

APEX - « *Absolute Protein Expression* »

APS – Persulfate d'Ammonium

BS^3 – Sulfo-DSS

BSA – Albumine de Sérum Bovin

CD – Clusters de Différentiation

CHOP – Chimiothérapie (cyclophosphamide, doxorubicine, vincristine et prednisone)

CID – « *Collision Induced Dissociation* »

CMC – Concentration Micellaire Critique

CSC – Cellules Souches Cancéreuses

CTAC – Chlorure de Cétrimonium

C-ter – Extrémité C-terminale de la protéine

CV – Coefficients de variation

DDA – « *Data-Dependent Acquisition* »

DIA - « *Data-Independent Acquisition* »

DL – Digestion Liquide

DSS – Disuccinimidyl suberate

ECD – « *Electron Capture Dissociation* »

EMBL-EBI – Institut européen de bio-informatique ou « *The European Bioinformatics Institute* »

emPAI – « *exponentially modified Protein Abundance Index* »

ESI – « *ElectroSpray Ionization* »

ETD – « *Electron-Transfer Dissociation* »

FA – Acide Formique

FASP – « *Filter-Aided Sample Preparation* »

FDR – Taux de faux-positifs ou « *False Discovery Rate* »

FFPE – « *Formalin-Fixed and Paraffin-Embedded* »

FMN – Riboflavine

GO – « *Gene Ontology* »

$H_2{}^{18}O$ – Eau lourde

HCD – « *Higher-energy Collision Dissociation* »

HUS – Hôpitaux Universitaires de Strasbourg

iBAQ – « *intensity-Based Absolute Quantification* »

ICAT – « *Isotope Coded Affinity Tag* »

IEF – Focalisation isoélectrique

IRMA – Institut de Recherche Mathématique Avancée

iRT – « *index Retention Time* »

iTRAQ – « *isobaric Tags for Relative and Absolute Quantification* »

KEGG – « *Kyoto Encyclopedia of Genes and Genomes* »

LBDGC – Lymphomes B diffus à grandes cellules

LC-MS/MS – Analyse par chromatographie liquide couplée à la spectrométrie de masse en tandem

m/z – Ratio masse sur charge

MALDI – « *Matrix-Assisted Laser Desorption Ionization* »

MED-FASP – « *MultiEnzyme Digestion FASP* »

MS – Spectrométrie de Masse

MudPIT – « *Multidimensionnal Protein Identification Technology* »

NCBI – Centre National pour l'information biotechnologique ou « *National Center for Biotechnologic Information* »

NHC – Nouvel Hôpital Civil

N-ter – Extrémité N-terminale de la protéine

PAI - « *Protein Abundance Index* »

PDB – « *Protein DataBank* »

PFF – Empreinte de fragments peptidiques ou « *Peptide Fragment Fingerprinting* »

pI – Point Isoélectrique

PIR – « *Protein Identification Resource* »

PRIDE – « *PRoteomics IDEntification database* »

PSM – « *Peptide Spectrum Match* »

PVDF – « *PolyVinyliDene Fluoride* »

Q – Analyseur quadripôlaire

RCPG – Récepteurs Couplés aux Protéines G

SDC - Déoxycholate de sodium

SDS – Dodécylsulfate de sodium

SG – Gel « *Stacking* » ou de concentration

SIB – Institut suisse de bio-informatique ou « *Swiss Institute of Bioinformatics* »

SILAC – « *Stable Isotope Labeling with Amino acids in Cell cultures* »

SPE – Extraction sur phase solide ou « *Solid-Phase Extraction* »

SRM – « *Selected-Reaction Monitoring* »

TCA – Acide trichloroacétique

TCEP – Tris(2-carboxyethyl)phosphine

TEMED – Tétraméthyléthylènediamine

TFA – Acide trifluoroacétique

TG – « *Tube-Gel* »

TMT – « *Tandem Mass Tags* »

TOF – Analyseur à temps de vol – « *Time-Of-Flight* »

Tr – Temps de rétention

PUBLICATIONS ET COMMUNICATIONS

Publications

Muller, L., Fornecker, L., Van Dorsselaer, A., Cianferani, S. & Carapito, C. **Benchmarking sample preparation/digestion protocols reveals tube-gel being a fast and repeatable method for quantitative proteomics.** Proteomics 16, 2953-2961 (2016).DOI: 10.1002/pmic.201600288

Bouguenina H., Salaun D., Mangon A., Muller L., Baudelet E., Camoin L., Tachibana T., Cianferani S., Audebert S., Verdier-Pinard P. & Badache A. **EB1-binding-myomegalin protein complex promotes centrosomal microtubules functions.** PNAS, (2017). DOI : 10.1073/pnas.1705682114

Communications orales

Leslie MULLER, Luc-Matthieu FORNECKER, Alain VAN DORSESELAER, Sarah CIANFERANI, Christine CARAPITO **: Methodological developments for proteomic analysis identification of proteins associated with chemo-refractoriness in diffuse large B-cell lymphoma** – Journée des Doctorants de l'Ecole Doctorale 222 (2 novembre 2016) – Strasbourg, FRANCE

Leslie MULLER, Luc-Matthieu FORNECKER, Alain VAN DORSSELAER, Sarah CIANFERANI, Christine CARAPITO : **Tube-gel digestion: a fast and reproducible sample preparation protocol for quantitative proteomics–** SFEAP (Société Française d'Electrophorèse et d'Analyse Protéomique) (10-12 septembre 2016) - Chambéry, FRANCE

Communications par affiche

Leslie MULLER, Noémie ROBIL, Jihu DONG, Fabien PETEL, Hervé CHNEIWEISS, Marie-Claude KILHOFFER, Jacques HAIECH, Christine CARAPITO, Sarah CIANFERANI : **"Label-free" differential membrane proteomics for the search of novel Glioblastoma cancer stem cell biomarkers** – Congrès EuPA (European Proteomics Association) (23-25 juin 2015) - Milan, ITALIE

Leslie MULLER, Noémie ROBIL, Jihu DONG, Fabien PETEL, Hervé CHNEIWEISS, Marie-Claude KILHOFFER, Jacques HAIECH, Christine CARAPITO, Sarah CIANFERANI : **"Label-free" differential membrane proteomics for the search of novel Glioblastoma cancer stem cell biomarker .–** Ecole d'été européenne de protéomique (2-8 août 2015) - Brixen, Italie

Leslie MULLER, Luc-Matthieu FORNECKER, Alain VAN DORSSELAER, Sarah CIANFERANI, Christine CARAPITO : **Evaluation de la préparation d'échantillons « tube-gel » pour l'analyse quantitative de mélanges protéiques complexes** – SMAP (Congrès de Spectrométrie de Masse et Analyse Protéomique) (15-18 septembre 2015) - Ajaccio, FRANCE

Leslie MULLER, Luc-Matthieu FORNECKER, Alain VAN DORSSELAER, Sarah CIANFERANI, Christine CARAPITO *:* **Optimization of a promising sample preparation protocol, Tube-Gel, for high throughput quantitative proteomics** – ASMS (American Society for Mass Spectrometry) (4-8 juin 2017) – Indianapolis, ETATS-UNIS

INTRODUCTION GENERALE

Les protéines sont des macromolécules biologiques qui résultent de la traduction du génome. Elles assurent de nombreux rôles au sein de l'organisme et constituent ce que l'on nomme le protéome. Cette entité est hautement dynamique et complexe de par les nombreuses versions protéiques pouvant résulter d'un même gène. Ces versions, nommées protéoformes, sont conséquentes de modifications post-traductionnelles, d'épissage alternatif ou encore de variants de séquence[1].

Le protéome dépend et reflète l'état physiologique d'un tissu ou d'une cellule. Son analyse permet par conséquent d'étudier les processus biologiques sous-jacents à des perturbations, telles que les maladies ou les conditions environnementales[2]. C'est pour cette raison que l'analyse protéomique occupe de nos jours une place grandissante dans le domaine de la recherche de biomarqueurs de pathologies. En effet, elle offre des informations complémentaires à d'autres stratégies « omiques », telles que la transcriptomique ou la génomique, qui sont essentielles à la compréhension de processus biochimiques et de maladies[3].

A ses débuts dans les années 1980, l'analyse de protéines s'effectuait au travers de gels d'électrophorèse bidimensionnelle, et leur identification était obtenue par dégradation d'Edman. Depuis les années 2000, la spectrométrie de masse occupe une place prépondérante dans ce domaine[4]. Les efforts fournis par la communauté scientifique pour amener les stratégies de protéomique « *Bottom-up* » à maturité permettent aujourd'hui d'analyser des mélanges protéiques complexes de manière quasi-routinière, et de répondre à diverses problématiques biologiques. En effet, les développements aussi bien instrumentaux (en spectrométrie de masse) que d'outils bio-informatiques et de banques de séquences protéiques, offrent à l'heure actuelle la possibilité d'identifier des milliers de protéines au sein d'un mélange protéique complexe[5-8]. Ces développements permettent également d'obtenir des informations quantitatives, souvent essentielles pour répondre à des questions biologiques, notamment dans le domaine de la recherche de biomarqueurs de pathologies[9]. Pour de tels projets, des approches différentielles sont mises en œuvre, qui opposent par exemple des échantillons protéiques issus de patients sains à ceux de patients atteints de la maladie étudiée, de manière à détecter les protéines variant entre les deux conditions qui reflèteraient la pathologie[10]. Le laboratoire a fait le choix d'employer des stratégies de quantification sans marquage pour répondre à ce type de problématiques, notamment parce qu'elles offrent la possibilité d'analyser un nombre théoriquement illimité d'échantillons. Ces techniques requièrent cependant la pleine maîtrise de la répétabilité du schéma analytique afin d'assurer une quantification robuste[11]. Plus

particulièrement, la préparation d'échantillons, qui intervient au début du schéma analytique et qui est intimement liée à l'extraction des protéines du fait de l'incompatibilité de certains tampons d'extraction avec certaines préparations d'échantillons, doit être pleinement maîtrisée pour ne pas impacter la répétabilité de manière précoce[12].

C'est dans ce contexte que mes travaux de thèse s'inscrivent, avec le développement de méthodes de préparations d'échantillons pour la quantification sans marquage, adaptées à diverses matrices biologiques. Les différents développements effectués ont été appliqués à divers projets de recherche de biomarqueurs de pathologies, pour lesquels des stratégies de quantification sans marquage ont été employées.

Ainsi, le **Chapitre I** du manuscrit présentera les différentes étapes de l'analyse protéomique par spectrométrie de masse. Il s'attardera particulièrement sur les étapes d'extraction de protéines et de préparation d'échantillons, avant de décrire les différentes stratégies et outils disponibles à l'heure actuelle permettant l'identification et la quantification des protéines, et devant être adaptés de manière à pouvoir répondre à diverses problématiques. Ce chapitre apporte quelques lumières sur les améliorations et efforts à apporter pour tirer davantages d'informations des données générées par les protéomistes.

Le **Chapitre II** décrit l'ensemble des développements méthodologiques de préparation d'échantillons réalisés au cours de ces travaux de thèse pour la protéomique quantitative sans marquage.

- Dans un premier temps, les optimisations de préparation d'échantillons adaptée à l'étude du protéome membranaire de cellules souches cancéreuses de glioblastomes seront développées. Les difficultés inhérentes à l'extraction et la préparation de protéines membranaires seront discutées. Il s'agit de développements effectués dans le cadre du projet de recherche de biomarqueurs de glioblastomes.

- Une seconde partie traitera à la fois de l'évaluation d'une préparation d'échantillons innovante sans fractionnement pour la protéomique quantitative sans marquage, le « *Tube-Gel* », ainsi que de l'optimisation de ce protocole afin de le rendre universel à la diversité des échantillons biologiques. Au cours de ma thèse j'ai pu implémenter cette nouvelle technique de préparation d'échantillons rapide et répétable au laboratoire.

- Une troisième partie traitera des développements de méthodes d'extraction de protéines et de préparation d'échantillons à partir de tissus ganglionnaires à la fois frais et inclus en

paraffine. L'enjeu de tels développements, effectués dans l'objectif d'une étude de recherche de biomarqueurs de lymphome, sera discuté.

- Enfin, une quatrième et dernière partie traitera de l'optimisation d'une préparation d'échantillons permettant l'analyse du protéome urinaire. Les difficultés inhérentes à ce type de matrice seront exposées. L'objectif de ces optimisations réside en la mise au point d'une méthode simple, rapide et robuste pouvant être implémentée en clinique pour le diagnostic en routine de biomarqueurs protéiques urinaires. En effet, ces développements ont été effectués dans le cadre de la recherche de biomarqueurs de suivi de l'état de greffons rénaux.

Le **Chapitre III** portera quant à lui sur les projets de recherche de biomarqueurs à proprement parler. Ces études ont tiré profit des développements et optimisations décrits au Chapitre II. Ce chapitre se divise en trois parties :

- Une première partie traitera de la recherche de biomarqueurs membranaires de glioblastomes. L'objectif étant de trouver des cibles de thérapie pouvant améliorer le traitement et la survie des patients atteints de glioblastome. Une attention particulière sera portée sur les contrôles permettant d'attester la qualité des données.

- Une seconde partie abordera le projet de recherche de biomarqueurs de résistance des lymphomes B diffus à grandes cellules. L'objectif premier étant de mieux comprendre les processus biologiques liés à la chimiorésistance, et de pouvoir éventuellement la prévenir afin d'adapter les traitements de première ligne. Le traitement statistique original effectué à partir de données de quantification peptidique sera exposé, et les résultats d'une approche multi-« omiques » seront présentés.

- Une troisième et dernière partie explicitera le projet de recherche de biomarqueurs de suivi de greffons rénaux. Ce projet s'inscrit dans le développement d'une méthode permettant le suivi et le diagnostic précoce de complications au niveau des greffons rénaux. Cette étude a été menée de manière à être robuste vis-à-vis des contraintes imposées par le diagnostic de routine à l'hôpital, c'est-à-dire robuste à un traitement, depuis le recueil des urines jusqu'à l'analyse, simple et rapide.

Pour finir, quelques **Travaux complémentaires** seront brièvement résumés. Il s'agit notamment d'identification de partenaires de complexes protéiques, de la caractérisation d'une protéine et de projets de recherche de zones d'interactions par une technique de pontage chimique.

CHAPITRE I

L'ANALYSE PROTEOMIQUE PAR SPECTROMETRIE DE MASSE

CHAPITRE I - L'ANALYSE PROTEOMIQUE PAR SPECTROMETRIE DE MASSE

La protéomique consiste en la description, aussi bien qualitative que quantitative, de l'ensemble des protéines exprimées dans une cellule ou un tissu à un moment donné, dans des conditions données[13]. La spectrométrie de masse est devenue une méthode clé de l'analyse protéomique suite à deux avancées datant de la fin des années 1980, portant sur le développement de deux méthodes d'ionisation : l'ionisation « MALDI » (« *Matrix-Assisted Laser Desorption Ionization* »)[14] et l'électrospray (ESI pour « *ElectroSpray Ionization* »)[15], qui ont rendu l'analyse de larges molécules peu, voire non volatiles telles que les protéines, possible. Par ailleurs, l'augmentation croissante et la curation des banques de données protéiques provenant de banques génomiques, les progrès instrumentaux de ces dernières années améliorant la sensibilité, la résolution, la vitesse d'acquisition, l'exactitude de masse, et en ce sens la qualité et la quantité des données, ainsi que l'évolution constante des outils bio-informatiques permettant l'automatisation de l'interprétation des résultats ont également joué un rôle important dans l'instauration de la spectrométrie de masse comme élément central dans le domaine de l'analyse protéomique[2, 4, 5, 16, 17].

I- Les différentes approches en analyse protéomique

L'analyse protéomique par spectrométrie de masse permet aussi bien de caractériser une protéine et ses partenaires d'interaction, que d'étudier le niveau d'expression d'un ensemble de protéines. Selon la nature de l'information souhaitée, trois types de stratégies, schématisées en Figure I-1, peuvent être envisagées :

- **L'approche « *Top-Down* »** consiste à analyser les protéines entières, de manière à identifier des protéoformes ou les modifications portées par les protéines[17, 18]. Dans ce type d'approche, la masse de la protéine est dans un premier temps déterminée, puis sa séquence peut être déduite de sa fragmentation. Au vu de la complexité des données générées, le « *Top-Down* » nécessitait jusqu'il y a peu de temps d'analyser la protéine seule, et impliquait de ce fait de disposer d'échantillons très peu complexes, voire de la protéine purifiée[19]. En effet, les difficultés rencontrées lors du fractionnement des protéines en amont, ou encore le manque d'outils bio-informatiques permettant de traiter ce type de données en aval, limitaient l'utilisation du « *Top-Down* » à haut débit sur des mélanges complexes[17-19]. Depuis, des études identifiant quelques centaines de protéines par des méthodes de « *Top-Down* » ont progressivement fait leur apparition dans la littérature[20, 21].

- **L'approche « *Bottom-Up* »** consiste quant à elle à analyser les protéines par l'intermédiaire de peptides (d'une taille inférieure à 3 kDa) résultant d'une digestion enzymatique, généralement effectuée à l'aide de la trypsine. Les milliers de peptides générés par la digestion des protéines au sein d'un mélange complexe sont généralement séparés par chromatographie liquide, avant de faire l'objet d'une analyse par spectrométrie de masse. Au cours de cette dernière, la masse de chaque peptide peut être déterminée de manière à être par la suite isolé, fragmenté, puis identifié[17, 18]. A partir de l'identité des peptides, il est possible de remonter à l'identité des protéines présentes dans le mélange complexe de départ. Cette approche est aujourd'hui largement utilisée, notamment pour l'analyse de mélanges complexes, du fait des récents développements instrumentaux et d'outils bio-informatiques adaptés qui permettent d'effectuer des analyses à haut débit.

- **L'approche « *Middle-Down* »** est une approche hybride des deux précédentes. Elle consiste à analyser des peptides, tout comme la stratégie « *Bottom-Up* », mais de plus grande taille (entre 3 et 20 kDa), générés par une digestion avec des enzymes telles que l'Asp-N, la Glu-C[22] ou encore l'OmpT[23]. Elle génère des données de type « *Top-Down* » sur de grandes régions de protéines tout en contournant les problèmes inhérents à l'analyse de protéines intactes[19].

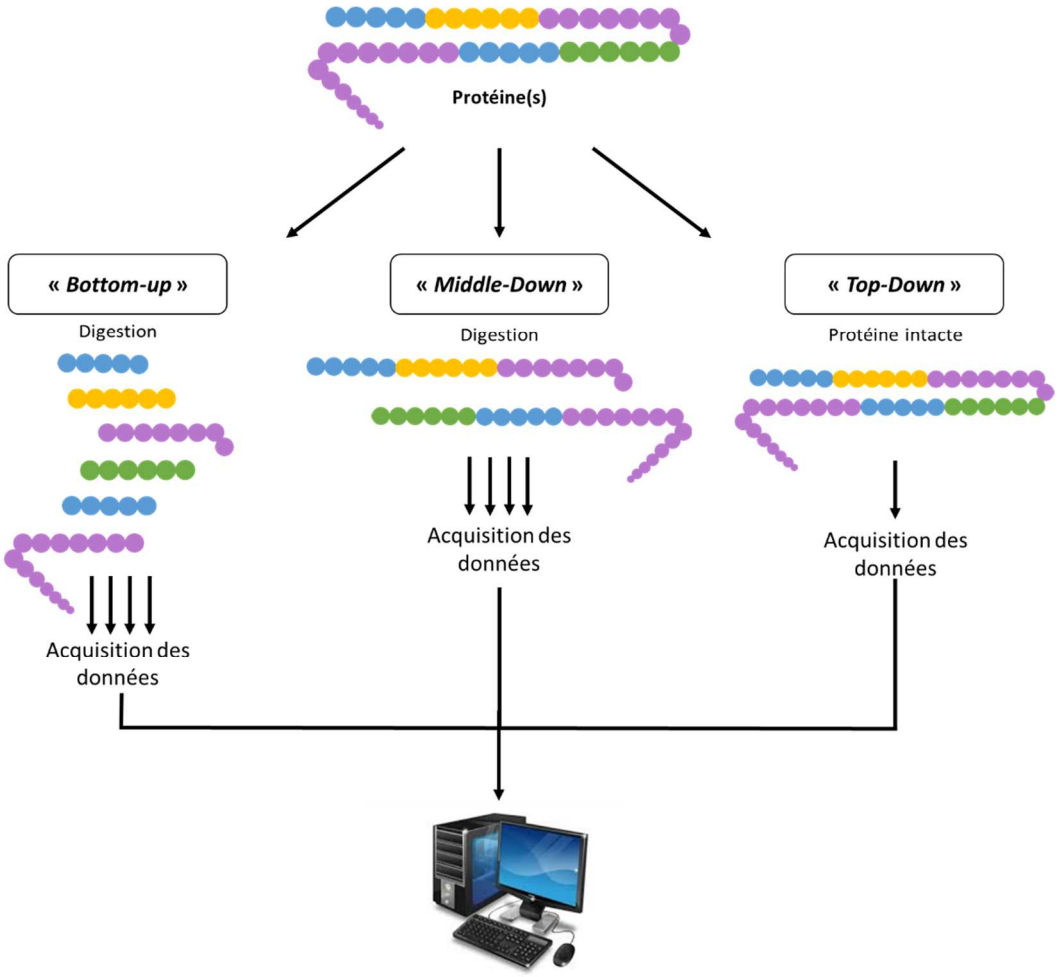


Figure I-1 - Les différentes approches de l'analyse protéomique par spectrométrie de masse, adaptée de [22]
L'analyse des protéines par l'approche « *Bottom-Up* » s'effectue après digestion enzymatique de celles-ci, générant des peptides < 2 kDa. L'analyse des protéines par approche « *Top-Down* » s'effectue sur la protéine entière. L'analyse des protéines par l'approche « *Middle-Down* », intermédiaire entre l'approche « *Top-Down* » et « *Bottom-Up* » s'effectue après digestion enzymatique des protéines, générant des peptides de 3 à 20 kDa.

Seule l'approche « *Bottom-Up* » a été utilisée au cours de ces travaux de thèse et sera détaillée étape par étape dans la suite de ce chapitre.

II- Les étapes de la stratégie « Bottom-Up »

L'analyse protéomique par approche « *Bottom-Up* » peut être résumée en quatre étapes, décrites en Figure II-1 et détaillées dans la suite de ce paragraphe, à savoir : l'extraction des protéines, la préparation d'échantillons, l'analyse par chromatographie liquide couplée à la spectrométrie de masse en tandem (LC-MS/MS) et le traitement des données. Chacune de ces étapes devra être adaptée de manière à ce que le schéma analytique permette de répondre au mieux aux questions biologiques posées.

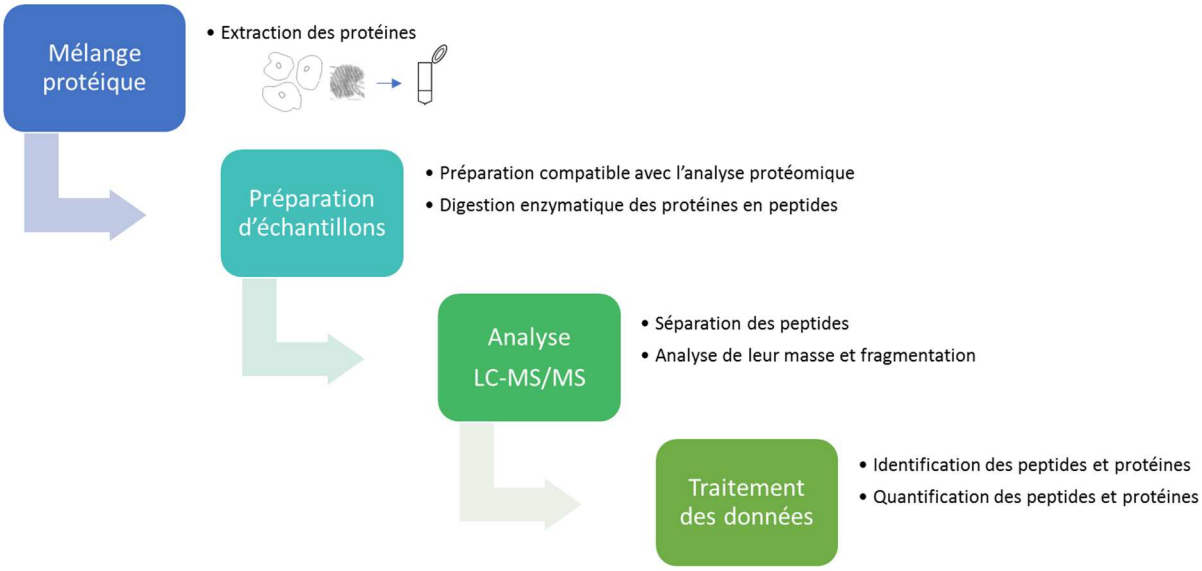


Figure II-1 - Les étapes de l'approche « *Bottom-Up* »

1- L'extraction des protéines

La réussite des analyses protéomiques est fortement dépendante de la qualité de l'échantillon de départ[24]. Ainsi, l'extraction des protéines est une étape clé du processus qui contribue largement à la variabilité des données si elle n'est pas maîtrisée[12]. Le type d'échantillon étudié (sang, tissu, etc.), la quantité de matériel disponible, le type de protéines auxquelles nous nous intéressons, mais aussi la méthode de préparation de l'échantillon souhaitée en aval, sont autant d'éléments à prendre en considération avant d'opter pour une stratégie d'extraction de protéines[24].

L'extraction des protéines s'effectue au travers d'un tampon de lyse et d'extraction, souvent accompagné de stimuli mécaniques. Le choix du tampon, qui a pour rôle de lyser, extraire, solubiliser les protéines et faciliter l'accès de la protéase aux protéines[25], est primordial. Il peut être basé sur l'utilisation :

- De **chaotropes** telles que l'urée ou la thiourée qui, en stabilisant l'état déplié des protéines, les dénaturent[26],
- De mélanges **solvant organique-eau** tels que acétonitrile (ACN)-eau ou méthanol-eau qui sont compatibles avec une analyse LC-MS/MS[26],
- D'**acides organiques** qui cassent les membranes et solubilisent bien les protéines membranaires, tels que l'acide formique (FA) ou l'acide trifluoroacétique (TFA),

- De **détergents** qui contiennent à la fois un domaine hydrophobe et un domaine hydrophile, et qui sont capables de s'auto-assembler et de former spontanément des micelles à partir d'une certaine concentration seuil nommée concentration micellaire critique (CMC), propre à chaque détergent. C'est à cette concentration que les micelles sont présentes en quantité suffisante pour solubiliser les protéines[26, 27]. Nous distinguons quatre types de détergents :

    - Les **détergents ioniques**, tels que le dodécylsulfate de sodium (SDS), qui est très largement utilisé et souvent considéré comme une référence, notamment pour l'extraction de protéines membranaires. Du fait de sa longue chaîne carbonée et sa tête anionique (Figure II-2), il casse les interactions inter- et intraprotéiques. Ainsi, il solubilise, prévient de l'adsorption des protéines sur les parois des contenants utilisés, de l'agrégation, et dénature les protéines. Cependant, ce type de détergent n'est pas compatible avec une digestion enzymatique ainsi qu'avec une analyse LC-MS/MS consécutive, étant donné qu'il interfère l'analyse chromatographique et perturbe l'ionisation des peptides[12, 26, 27].

    - Les **détergents non-ioniques** qui sont des agents de solubilisation doux tels que le Triton X-100, le NP-40 ou le Dodécyl-ß-D-maltoside (DDM). Tout comme les détergents ioniques, ils possèdent une chaîne carbonée, mais leur tête polaire ne présentant pas de charge, ils ne sont pas capables de casser les interactions protéines-protéines. Certains de ces détergents sont compatibles à faible concentration avec une analyse LC-MS/MS (DDM), alors que d'autres peuvent provoquer de fortes suppressions de signal et former des adduits (Triton X-100, NP-40).

    - Les **sels d'acides biliaires**, tels que le déoxycholate de sodium (SDC), qui possèdent une face polaire et une face apolaire et présentent des capacités de solubilisation et de dénaturation plus faibles. L'avantage de ces sels est la facilité avec laquelle ils peuvent être retirés d'un échantillon pour rendre l'échantillon compatible avec une analyse LC-MS/MS.

    - Les **détergents zwitterioniques**, tels que le CHAPS qui ont des capacités de solubilisation intermédiaires entre les détergents non-ioniques, les sels d'acides biliaires et les détergents ioniques. A très faible concentration, ils peuvent être compatibles avec une analyse LC-MS/MS, cependant ils peuvent former des adduits et générer des suppressions de signal à plus haute concentration.

- Les **détergents compatibles avec la spectrométrie de masse** (MS). Il s'agit de détergents qui contiennent un groupement labile vis-à-vis des acides, situé entre la tête hydrophile et la queue hydrophobe, qui permet de générer des sous-produits compatibles avec la MS suite à l'ajout d'acide dans le milieu. De plus, à faible concentration, ils sont tolérés par les protéases classiques[12, 26]. Ces détergents offrent la possibilité de traiter rapidement les échantillons, puisqu'ils permettent de s'affranchir d'étapes ayant pour but de les retirer avant digestion et analyse LC-MS/MS, mais restent cependant très coûteux. Il s'agit notamment du RapiGest™ (WATERS), ProteaseMAX™ (Promega), PPS™ (Protein Discovery)[28] et MaSDeS[27]. La Figure II-2 décrit le mécanisme d'hydrolyse du MaSDeS permettant de le rendre compatible avec une analyse par MS.

Par ailleurs, il a également été décrit que l'ajout d'alkylamines ou de polyamines comme la spermine et la spermidine à un tampon de lyse à base de détergent, permet d'améliorer la solubilisation ainsi que la stabilisation des protéines membranaires[29].

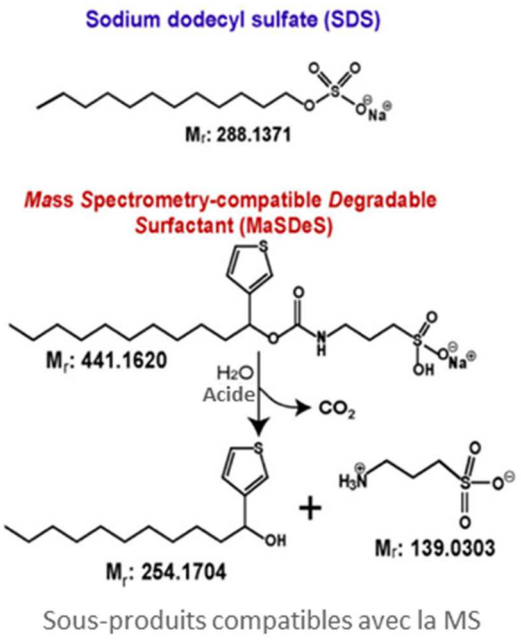


Figure II-2 - Molécule de SDS et mécanisme d'hydrolyse du MaSDeS, générant des sous-produits du détergent compatibles avec la MS (adaptée de [27])

Enfin, il est parfois nécessaire de retirer des contaminants tels que les lipides ou l'acide désoxyribonucléique (ADN) qui interfèrent en chromatographie liquide et dégradent le signal en MS. Généralement, ces composés sont retirés par une étape de précipitation des protéines (à l'acétone glacial, éthanol, acide trichloroacétique (TCA), …). Cependant, il faut garder à l'esprit qu'une étape de

précipitation peut engendrer une perte d'échantillon, en plus d'une resolubilisation du culot de protéines souvent difficile[24].

En résumé, il n'existe pas de méthode générique applicable à l'ensemble des échantillons. C'est pourquoi, il est nécessaire d'optimiser cette étape d'extraction des protéines à l'échantillon étudié afin d'en tirer un maximum d'informations de qualité[12].

2- La préparation d'échantillons

La préparation d'échantillons est intimement liée à l'étape d'extraction des protéines. Par conséquent, celle-ci doit être adaptée de manière à rendre l'échantillon compatible à la fois avec une digestion enzymatique et une analyse LC-MS/MS, notamment si le tampon utilisé n'est pas directement compatible avec ces étapes ultérieures. Nous distinguons deux types d'approches de préparation d'échantillons :

- Les **approches en solution** qui consistent à dénaturer les protéines, réduire les ponts disulfures, alkyler les cystéines et digérer les protéines en phase liquide. Ces approches ont l'avantage d'être simples, mais nécessitent l'utilisation d'un tampon de lyse compatible avec la digestion enzymatique et l'analyse LC-MS/MS[24]. La digestion liquide est l'une de ces approches, qui emploie le plus souvent un tampon d'extraction à base d'urée ou de SDC, compatible avec une digestion enzymatique à faible concentration, mais qui nécessite d'être retiré avant analyse LC-MS/MS par des étapes d'extraction sur phase solide (SPE) ou d'acidification dans le cas du SDC[12, 25]. Certaines approches en solution permettent l'utilisation de détergents pour l'extraction des protéines, tels que le SDS, puisqu'elles impliquent l'utilisation de filtres ou de membranes permettant de se débarrasser de ces composés avant protéolyse, entrainant ainsi un risque de perte d'échantillon[24]. Il s'agit notamment du « *Filter-Aided Sample Preparation* » (FASP)[30, 31]. Dans cette préparation, les protéines contenues dans un tampon à base de SDS sont déposées sur un filtre ayant un seuil de coupure de 10 ou 30 kDa. Ce filtre permet de retenir les protéines et d'effectuer un échange de tampon par des étapes de centrifugation, afin de rendre l'échantillon compatible avec une digestion enzymatique qui va générer des peptides de plus petite taille qui ne seront plus retenus par le filtre. Enfin, les sels présents dans l'extrait peptidique résultant sont retirés par une étape de SPE. D'autres techniques permettent de retirer le SDS tout en restant en phase liquide, comme la précipitation[25], ou encore l'utilisation de réseaux métallo-organiques MOF pour « *Metal-organic frameworks* »[32], qui n'ont à notre connaissance pas encore été employés dans une étude protéomique. En plus de ces méthodes, de nombreuses solutions commerciales

existent, souvent « tout-en-un », comme le *IST Sample Preparation Kit* (PreOmics)[33], *S-trap* (ProtiFi)[34], *Smart Digestion Kit* (Thermo Scientific), *Thermo Scientific Pierce Mass Spec Sample Prep Kit for Cultured Cells* ou encore *Rapid Digestion-Trypsin for Mass Spectrometry Analysis Kit* (Promega).

- Les **approches en gel** sont généralement employées lorsque les protéines ont été extraites à l'aide d'un tampon Laemmli[35] à base de SDS et d'agent réducteur, ou lorsque les échantillons sont susceptibles de contenir des contaminants. Un gel d'électrophorèse monodimensionnel dénaturant (1D SDS-PAGE) est alors préférentiellement réalisé.

  L'electrophorèse est une technique qui consiste à séparer des analytes sous l'effet d'un champ électrique[36]. Les analytes chargés migrent vers le pôle de signe opposé, et la vitesse de migration dépend du rapport charge/taille. Dans le cas de l'utilisation d'un tampon Laemmli, l'agent réducteur couplé au SDS va permettre de dérouler les protéines en plus de leur conférer une charge négative nette. Les protéines ayant ainsi la même densité de charge, migreront toutes vers le pôle positif. Le gel de polyacrylamide au travers duquel elles migreront, forme un tamis moléculaire qui permet d'effectuer une séparation en fonction de la masse moléculaire des protéines. La dimension des pores du gel, dictée par la concentration en acrylamide et bis-acrylamide, contrôle cette séparation. Ainsi, plus la proportion d'acrylamide sera élevée, plus le maillage sera serré et permettra de séparer des petites protéines. Le gel d'électrophorèse 1D constitue un système discontinu, schématisé en Figure II-3, avec deux parties distinctes : le gel « de concentration » constitué d'un faible pourcentage d'acrylamide/bis-acrylamide (4 à 5 %), et le gel « de séparation » constitué de 8 à 20 % d'acrylamide/bis-acrylamide. Du fait de la composition du tampon de migration à base de Tris-HCl et de glycine, la vitesse des ions sera influencée par le pH. Ainsi, dans le gel de concentration à pH 6,8, les ions chlorures (ions pilotes qui imposent la vitesse de migration) auront une grande mobilité, alors qu'à l'inverse les glycinates auront une mobilité plus faible puisque le pH est proche du point isoélectrique (pI) de la glycine, créant ainsi un vide ionique de faible conductivité qui ralentit les protéines et permet de les concentrer en une bande avant d'arriver à l'interface entre le gel de concentration et de séparation. Ce dernier, ayant un pH plus élevé (de 8,8), permettra aux ions glycinates d'accélérer et aux protéines de se séparer en fonction de leur masse moléculaire[36]. Une fois séparées, les protéines sont fixées dans le gel et leur présence est généralement révélée à l'aide d'une coloration au bleu de Coomassie[37, 38]. Les bandes de gel sont alors découpées et peuvent être décolorées, lavées pour retirer le SDS, et les protéines sont réduites et alkylées directement dans le gel avant digestion enzymatique.

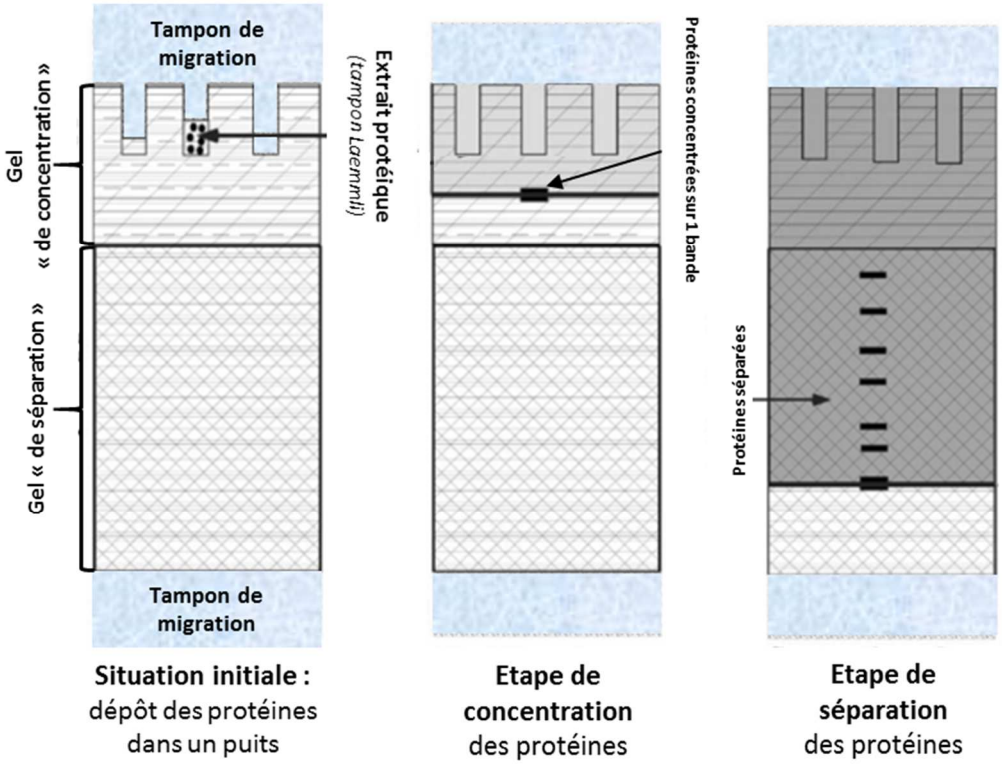


Figure II-3 - Schématisation d'une séparation de protéines sur un système de gel d'électrophorèse discontinu 1D SDS-PAGE, adaptée de [36]

Dans la situation initiale, l'extrait protéique est déposé dans un puits d'un gel d'électrophorèse composé d'un gel « de séparation » et d'un gel « de concentration ». Les protéines sont d'abord concentrées sur une bande dans le gel « de concentration » avant d'être séparées dans le gel « de séparation ».

Etant donnée la très grande gamme dynamique de certains échantillons, pouvant aller jusqu'à 12 ordres de magnitude dans un fluide biologique complexe comme le plasma[39], il est souvent nécessaire de décomplexifier l'échantillon avant analyse LC-MS/MS en le fractionnant, comme cela est le cas lors de l'utilisation d'un gel 1D SDS-PAGE. Le fractionnement en gel peut également s'effectuer à l'aide d'un gel bidimensionnel 2D SDS-PAGE, qui, dans une première dimension sépare les protéines selon leur pI (focalisation isoélectrique IEF), et dans une seconde selon leur masse moléculaire comme dans un gel 1D SDS-PAGE[40]. Cependant, cette approche est relativement chronophage[24, 41], et la première dimension n'étant pas compatible avec l'utilisation d'un tampon d'extraction à base de SDS, ne permet pas l'étude des protéines membranaires[26]. Notons que les approches en phase liquide permettent aussi le fractionnement des échantillons, soit au niveau protéique par des techniques d'IEF, soit au niveau peptidique après protéolyse par des techniques de chromatographie, comme le couplage d'une première séparation, par échange d'ions par exemple, à l'analyse LC-MS/MS classique que l'on nomme souvent 2D-LC-MS/MS (ou MudPIT pour « *Multidimensionnal Protein Identification Technology* » dans le cas précis d'une séparation par échange d'ions suivie d'une phase inverse)[26].

Les performances instrumentales accrues résultant des récentes avancées, permettent aujourd'hui d'envisager, selon le type d'étude et la question biologique posée, de ne pas fractionner l'échantillon avant analyse LC-MS/MS. Dans ce cas, en dehors des approches en phase liquide sans technique de fractionnement, il est possible d'envisager des approches en gel, comme l'utilisation du gel « *Stacking* », qui est une approche détournée du gel 1D SDS-PAGE et qui consiste à faire migrer les protéines uniquement dans le gel « de concentration » dans le but de les concentrer sur une seule bande.

De manière générale, il faut garder à l'esprit que des préparations d'échantillons multi-étapes sont à éviter pour limiter les biais techniques et garantir une bonne répétabilité des résultats. C'est pour ces raisons que le fractionnement de l'échantillon est souvent évité de nos jours, en plus du fait qu'il augmente la durée d'analyse pour un échantillon limitant ainsi le nombre d'échantillons à analyser, et qu'il requiert souvent une plus grande quantité de matériel de départ.

L'ensemble des approches décrites ici font systématiquement l'objet d'une digestion enzymatique avant analyse LC-MS/MS, que ce soit directement en solution ou en gel. Cette protéolyse s'effectue le plus souvent à l'aide de la trypsine, qui clive de manière spécifique en partie C-terminale (C-ter) des lysines et des arginines, acides-aminés fréquents, générant ainsi des mélanges complexes (d'une centaine de milliers) de peptides de tailles compatibles avec une séparation chromatographique et une analyse par spectrométrie de masse[42].

3- L'analyse par LC-MS/MS

a. La séparation des peptides par chromatographie liquide

Une étape de séparation par chromatographie liquide est systématiquement réalisée avant une analyse par MS en protéomique « *Bottom-up* ». Elle permet de réduire le nombre d'analytes qui entrent dans le spectromètre de masse, diminuant de ce fait le phénomène de compétition à l'ionisation, ce qui a pour finalité d'améliorer la sensibilité et la profondeur d'analyse[1, 43].

La chromatographie liquide en phase inverse est la plus communément utilisée en protéomique haut-débit, étant donné que la composition de la phase mobile est directement compatible avec une analyse par MS, rendant le couplage en ligne de ces deux techniques possible. La phase inverse permettant de séparer les peptides est constituée de silice greffée de chaînes de 18 carbones (C18). Du fait de son caractère hydrophobe, ce type de phase permet de retenir et de séparer les peptides en fonction de leur hydrophobicité. Ainsi les peptides hydrophobes seront davantage retenus que les peptides

hydrophiles. La composition de la phase mobile, basée sur l'utilisation d'un mélange d'acétonitrile et d'eau, va évoluer de manière à ce que la polarité diminue au cours du temps, rendant ainsi les interactions analytes-phase mobile plus favorables. Ceci va permettre le transfert des peptides de la phase stationnaire à la phase mobile qui va emmener les peptides dans le spectromètre de masse.

Au cours de ces travaux de thèse, seuls des systèmes UPLC Nano-Acquity WATERS ont été utilisés avec des colonnes de 25 cm de long, d'un diamètre interne de 75 µm et des particules de 1,7 µm de diamètre, générant de très hautes pressions (>500 bars) à de faibles débits (300 à 450 nl/min). Ces systèmes améliorent la résolution et la vitesse de séparation tout en offrant une quantité de chargement diminuée par rapport à des systèmes de micro-HPLC[1, 24]. Les échantillons de protéomique étant souvent précieux et disponibles en faibles quantités, ces caractéristiques rendent ces systèmes particulièrement intéressants pour les analyses protéomiques.

b. La spectrométrie de masse en tandem

Une fois les peptides élués de la chromatographie liquide en phase inverse, il sont ionisés grâce à la source ESI, utilisée en configuration nano-ESI qui montre à l'heure actuelle encore des faiblesses comme des instabilités de spray. Les ions peptides multi-chargés alors générés entrent dans le spectromètre de masse placé sous-vide, dans lequel ils vont être guidés et focalisés par des champs électriques. L'instrument va dans un premier temps mesurer le rapport masse/charge (m/z) et l'intensité de chaque ion afin de générer un spectre MS, puis, dans un deuxième temps, va séquencer chaque peptide en l'isolant, le fragmentant, et en mesurant le rapport m/z ainsi que l'intensité de l'ensemble des fragments générés, résultant en un spectre MS/MS[44]. La génération de spectres MS suivis de spectres MS/MS est nommée spectrométrie de masse en tandem, et s'effectue avec des instruments hybrides qui couplent différents analyseurs de masse.

Le mode d'acquisition « Data-Dependent Acquisition »

Le mode d'acquisition le plus largement utilisé en spectrométrie de masse en tandem pour l'analyse protéomique est le mode « *Data-Dependent Acquisition* » (DDA). Dans ce type d'acquisition, le spectromètre de masse effectue une succession de cycles tout au long de l'analyse. Au cours d'un cycle, un spectre MS est acquis, sur lequel les N ions peptides les plus intenses (appelés ions précurseurs) sont repérés pour être soumis à une fragmentation afin de générer N spectres MS/MS (Figure II-4). De ce fait, un peptide sera séquencé uniquement s'il fait partie du groupe des N plus intenses au moment où le spectre MS est acquis. Ce caractère stochastique de la sélection des ions peptides à fragmenter limite le nombre de peptides séquencés au cours d'une analyse, et donc le

nombre de peptides et de protéines identifiés. Malgré les récentes avancées instrumentales, notamment en termes de sensibilité et de vitesse d'acquisition permettant d'effectuer davantage de cycles afin de diminuer le phénomène de sous-échantillonnage, seuls 15 à 20 % des peptides détectables en MS pouvaient être séquencés en 2011[45]. Bien que ces performances se soient nettement améliorées ces dernières années, elles ne permettent toujours pas de séquencer la totalité des peptides détectables en MS par DDA.

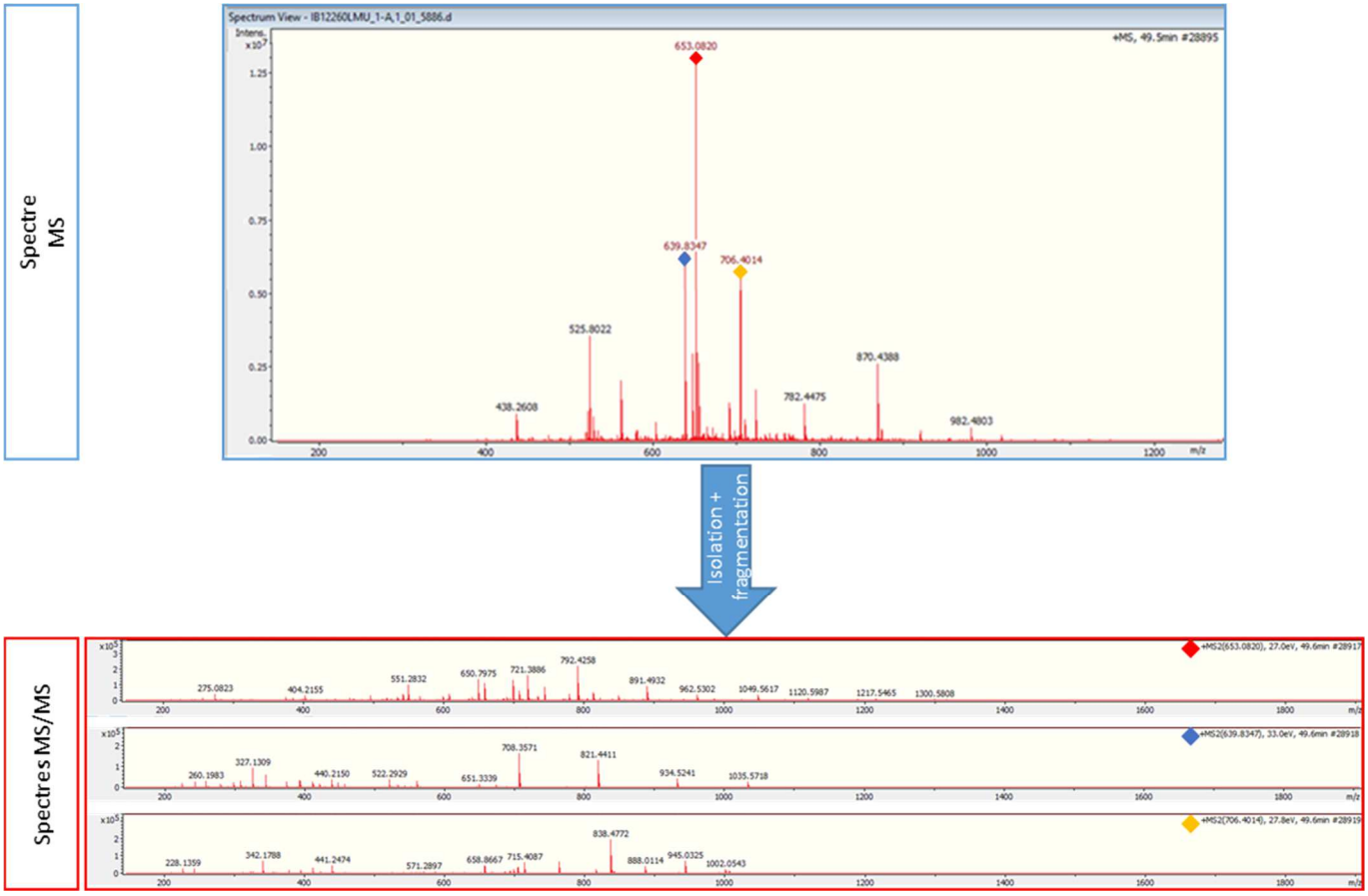


Figure II-4 - Schéma du mode d'acquisition DDA
(exemple d'un Top 3 : le spectromètre de masse sélectionne les 3 ions peptides précurseurs les plus intenses sur le spectre MS, et les fragmente pour générer les 3 spectres MS/MS correspondants)

Pour pallier à ce problème de stochastique des peptides sélectionnés pour leur séquençage MS/MS et ainsi au phénomène de sous-échantillonnage, un nouveau mode d'acquisition est apparu ces dernières années : le mode « *Data-Independent Acquisition* » (DIA). Il consiste à isoler et à fragmenter de manière séquentielle l'ensemble des peptides contenus dans une certaine gamme de masse. Cette approche, très prometteuse, permet d'augmenter drastiquement la gamme dynamique mais est à l'heure actuelle très peu utilisée pour identifier des peptides et des protéines, mais davantage pour les quantifier. Ce mode d'acquisition sera plus amplement détaillé au *Chapitre I –III-2-c*.

La fragmentation des peptides

Le mode de fragmentation le plus largement utilisé en protéomique est la fragmentation « *Collision Induced Dissociation* » (**CID**)[46]. Elle consiste à induire la fragmentation des peptides suite à la collision avec des molécules de gaz inerte (He, Ar, N_2). Celle-ci permet une augmentation de l'énergie interne des ions et provoque le déplacement de la charge du proton le long du squelette du peptide, rompant ainsi la liaison peptidique. Ceci a pour conséquence de générer des ions y et b, selon la nomenclature de fragmentation peptidique de Biemann[47, 48] (Figure II-5). Ce type de fragmentation, dictée par la charge, suit le modèle du proton mobile[49, 50]. Comme évoqué précédemment, les peptides sont majoritairement issus d'une digestion à la trypsine qui clive après les arginines et les lysines. Ils présentent ainsi des résidus basiques en partie C-ter, qui ont pour particularité de séquestrer une charge au niveau de leur chaîne latérale. De ce fait, le processus de déplacement de la charge séquestrée au sein d'un peptide trypsique monochargé requiert un apport d'énergie important, rendant leur fragmentation difficile. Cette caractéristique pousse à exclure les peptides trypsiques monochargés du processus d'isolation et de fragmentation. En revanche, ce type de fragmentation est particulièrement adapté aux peptides trypsiques doublement chargés, car ceux-ci portent, en plus de la charge séquestrée en C-ter, une charge sur l'amine en partie N-terminale (N-ter) dont le déplacement le long du squelette peptidique requiert moins d'énergie.

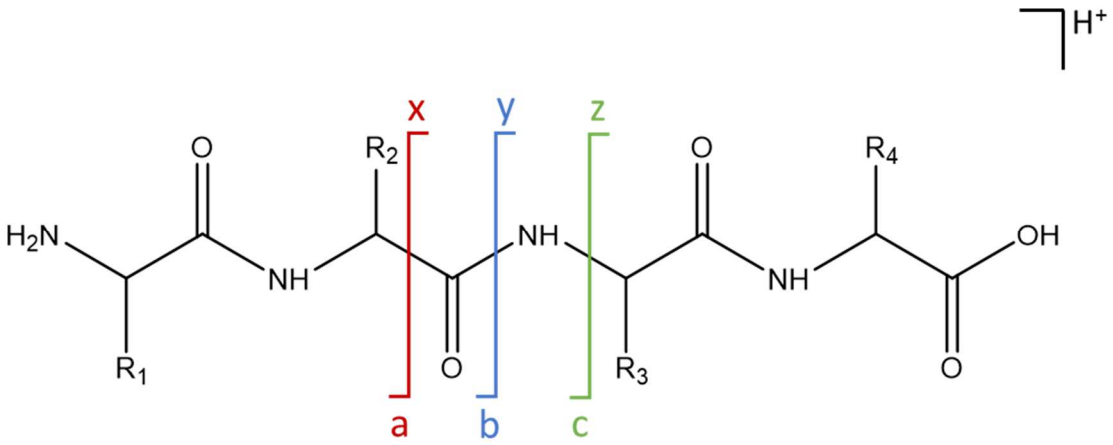


Figure II-5 - Nomenclature de Biemann pour la fragmentation peptidique

La fragmentation CID génère de cette façon des spectres MS/MS sur lesquels il est possible de lire la séquence des peptides fragmentés dans le sens C-ter vers N-ter à l'aide des ions y, qui se superposent à la série b dans la direction opposée (Figure II-6). Aussi, il apparaît important de connaître les mécanismes de fragmentation mis en jeu lors d'une analyse, afin de tirer des informations d'identification à partir des spectres MS/MS.

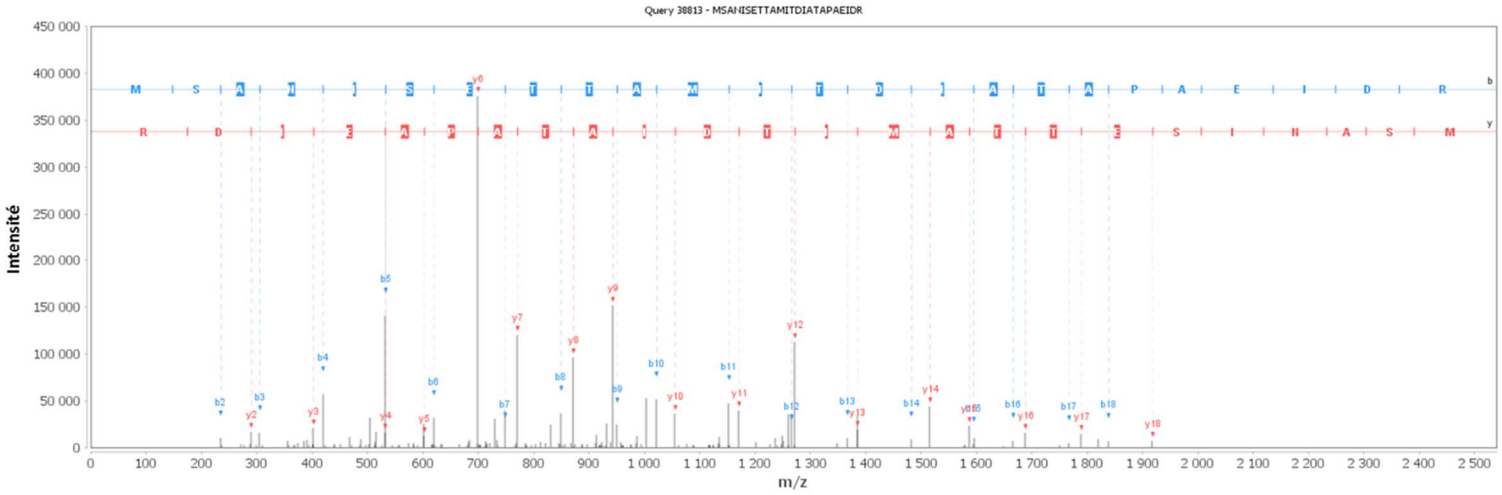


Figure II-6 - Exemple de spectre MS/MS issu de l'isolation et de la fragmentation CID d'un peptide

Notons que d'autres méthodes de fragmentation existent : **HCD**[51] (« *Higher-energy Collision Dissociation* ») qui génèrent des ions y et b tout comme la CID, mais qui s'effectue à plus haute fréquence, mais aussi l'**ETD**[52] (« *Electron-Transfer Dissociation* ») et l'**ECD**[53] (« *Electron Capture Dissociation* »), ou encore la combinaison de deux modes de fragmentation comme l'**EThcD** qui combine ETD et HCD[54].

Au cours de ces travaux de thèse, seuls les modes de fragmentation HCD et CID ont été utilisés.

c. Instrumentation

Durant ces travaux, quatre spectromètres de masse couplés à une chromatographie liquide UPLC ont été utilisés. Il s'agit d'instruments hybrides qui combinent soit un analyseur quadripôlaire (Q) et un analyseur TOF (« *Time-Of-Flight* »), soit un analyseur Q et un analyseur Orbitrap. Leurs caractéristiques sont détaillées dans le Tableau II-1 suivant.

Machine	Impact-HD	Impact II	TripleTOF 5600	Q-Exactive +
Constructeur	Bruker	Bruker	Sciex	Thermo Fisher Scientific
Type d'analyseur	Q-TOF	Q-TOF	Q-TOF	Q-Orbitrap
Résolution	40 000	60 000	30 000 (MS) 15 000 (MS/MS)	17 500 à 140 000 (à 200 m/z)
Exactitude de masse	10 ppm	5 ppm	15 ppm	5 ppm
Vitesse d'acquisition	17 Hz	50 Hz	100 Hz	13 Hz (17 500 résolution à 200 m/z)
Gamme de masse	20 000 m/z	20 000 m/z	40 000 m/z	6 000 m/z
Type de fragmentation	CID	CID	CID	HCD
Année d'installation	2013	2015	2014	2014

Tableau II-1 - Description des spectromètres de masse utilisés durant ces travaux de thèse

4- Le traitement des données

a. L'identification des protéines

Les moteurs de recherche

Afin d'identifier les peptides qui ont été analysés, les données brutes générées sont transformées en un fichier qui contient la masse du précurseur, des fragments associés, ainsi que leurs intensités respectives. A l'aide de ce fichier appelé « liste de masses », il est possible d'effectuer une recherche de type « *Peptide Fragmentation Fingerprinting* » (PFF)[55]. Cette recherche consiste à comparer les listes de masses expérimentales à des listes de masses théoriques issues d'une digestion trypsique suivie d'une fragmentation *in silico* de l'ensemble de la banque de séquences protéiques cible. Cette comparaison permet d'identifier les peptides, et par leur assemblage, d'inférer l'identité des protéines présentes dans l'échantillon analysé. Pour ce faire, de nombreux moteurs de recherche sont disponibles, parmi lesquels Mascot[56], OMSSA[57], X!Tandem[58], SEQUEST[59] et Andromeda[60]. Ces outils requièrent des informations concernant la façon dont les données expérimentales ont été générées, comme :

- La tolérance de masse sur le précurseur et les fragments qui dépend de l'exactitude de la mesure de masse de l'instrument utilisé,
- Leur charge,
- L'enzyme utilisée et le nombre de coupures manquantes autorisées,
- Le type d'ions générés qui découle du type de fragmentation utilisé.

Au cours de ces travaux de thèse, seuls Mascot et Andromeda ont été utilisés. Il s'agit de moteurs de recherche basés sur une approche probabilistique. Cela signifie qu'ils comparent les listes de masses théoriques et expérimentales, et attribuent un score qui reflète la probabilité que la concordance entre ces deux listes ne soit pas due au hasard. Le score étant égal à moins dix fois le logarithme en base dix de la probabilité, la meilleure corrélation entre masses théoriques et expérimentales aura une probabilité faible d'être due au hasard et présentera donc un score élevé. L'algorithme exact de calcul de la probabilité n'est pas connu pour Mascot, étant donné qu'il s'agit d'une solution commerciale, contrairement à Andromeda qui est disponible gratuitement sur internet. Malgré des échelles de scores différentes, ces deux moteurs de recherche semblent produire des résultats similaires, sans peptide exclusivement identifié par l'un ou par l'autre, selon Jürgen COX *et collaborateurs*[60]. Par ailleurs, le moteur de recherche Andromeda peut aussi bien fonctionner seul que de manière intégrée au logiciel MaxQuant qui va notamment permettre de recalibrer les spectres. L'utilisation du logiciel MaxQuant sera développée au *Chapitre I – III-2-b-Logiciels disponibles*.

Bien que ces moteurs de recherche permettent l'automatisation de l'identification des peptides et des protéines, seul un faible pourcentage (25 %) de spectres MS/MS donne réellement lieu à des identifications[61]. Ceci est dû à plusieurs phénomènes comme :

- La mauvaise qualité de certains spectres MS/MS, notamment lorsque les peptides trypsiques monochargés n'ont pas été correctement exclus du processus d'isolation et de fragmentation. Il est à noter que ce phénomène n'a pas été observé sur les spectromètres de masse de type Q-Orbitrap Q-Exactive + et Q-TOF SCIEX TripleTOF 5600, mais surtout sur le Q-TOF BRUKER Impact-II. Le Tableau II-2 reflète ce phénomène en comparant les résultats d'identification issus de l'analyse d'un même digeste de levure avec un gradient chromatographique de 79 minutes et des méthodes MS optimisées pour chaque instrument.

- La génération de spectres chimères qui résultent de la co-fragmentation de deux ou plusieurs peptides. En effet, même si des spectromètres de masse de haute résolution sont utilisés pour générer les données, l'isolation d'un précurseur est généralement effectuée au sein de l'instrument par un analyseur de type quadripôle qui présente une résolution inférieure, et ne permet pas toujours d'isoler sélectivement qu'un seul ion lorsque de nombreuses espèces présentes dans la même région de m/z co-éluent. Pour pallier à ce problème, le moteur de recherche Andromeda effectue une seconde recherche lorsque sur une carte LC-MS, représentant les temps de rétention (Tr) de l'ensemble des espèces mesurées par MS en fonction de leur ratio m/z, l'isolation d'un ion a été effectuée alors qu'une seconde espèce

était présente à un ratio m/z très proche ne permettant pas de les dissocier au moment de la sélection par le quadripôle. Les spectres MS/MS soumis à cette seconde recherche sont nettoyés des fragments assignés au peptide identifié lors de la première recherche.[60]

- La sous-évaluation des modifications lors des recherches qui est estimée à environ un tiers des spectres non-assignés[62, 63]. En effet, étant nombreuses, diverses, et souvent inattendues, l'ensemble des modifications portées par les peptides ne sont pas recherchées de manière systématique, contrairement aux modifications induites par le traitement de l'échantillon. Or, ces espèces modifiées existent et donnent lieu à des spectres de fragmentation qui ne mèneront alors pas à une identification. MSFragger est un outil qui propose de prendre en compte l'ensemble des modifications qu'un peptide peut porter au moment de la recherche, qui est effectuée avec une fenêtre de tolérance sur la masse du précurseur très large (500 Da)[64].

- L'utilisation de banques de séquences protéiques inadaptées ou incomplètes. C'est le cas par exemple lors de l'étude de maladies comme certains cancers chez l'homme qui peuvent générer des mutations au niveau de l'ADN dont la résultante protéique, nommée variant de séquence, n'est pas présente dans les banques de séquences humaines actuelles. Ce point sera davantage discuté dans le paragraphe suivant.

Machine	Impact II	TripleTOF 5600	Q-Exactive +
Quantité de lysat de levure analysée	200 ng	500 ng	100 ng
Nombre de spectres MS/MS acquis	40 992	48 697	23 032
Spectres MS/MS assignés à une identification	35 %	55 %	63 %
Spectres MS/MS provenant d'ions peptides monochargés	11 %	0 %	0 %
Nombre de spectres MS/MS validés (FDR < 1 %)	8824 (22 % des spectres acquis)	13 625 (28 % des spectres acquis)	10 667 (46 % des spectres acquis)
Nombre de peptides identifiés et validés (FDR < 1 %)	6110	8073	8694
Nombre de protéines identifiées et validées (FDR < 1 %)	1440	1371	1562

Tableau II-2 - Comparaison des résultats d'identification suite à l'analyse d'un digeste de levure avec un gradient chromatographique de 79 minutes et des méthodes MS optimisées pour chaque instrument
En noir les données issues des résultats non validés. En bleu les résultats validés avec un taux de faux-positifs (FDR) inférieur à 1 % (voir *Chapitre I – II – 4 b*)

D'une manière générale, l'ensemble des données générées par spectrométrie de masse ne sont aujourd'hui pas pleinement exploitées, en plus du fait du sous-échantillonnage lié au mode d'acquisition DDA. Le domaine de la bio-informatique est à l'heure actuelle un acteur clé de l'analyse

protéomique par MS, qui est en plein développement dans le but de tirer davantage d'informations des données de DDA, ou encore d'améliorer les outils dédiés aux nouveaux modes d'acquisition comme la DIA.

Les banques protéiques

L'assignement des peptides est limité aux séquences présentes dans la banque de séquences protéiques. C'est pourquoi il est important de bien choisir la banque de référence avant d'effectuer une recherche, afin d'extraire un maximum d'informations pertinentes et de qualité. Les **banques de séquences** proviennent de l'annotation des banques génomiques qui sont constamment en évolution et mises à jour du fait de la découverte de variants de séquence, d'épissage, etc. Une grande quantité de banques est actuellement disponible, mais leur degré de complètement, de redondance et de qualité d'annotation diffère. Nous distinguons notamment[7] :

- *NCBI Entrez*[65, 66] qui a été créée par le Centre National pour l'Information Biotechnologique (NCBI). Cette banque contient l'ensemble des banques protéiques NCBI provenant notamment de banques de séquences nucléotidiques (GenBank, RefSeq[67], « *Protein Information Ressource* » (PIR)[68], « *Protein Databank* » (PDB)[69] et UniProtKB/SwissProt[70]). Il s'agit d'une collection de séquences dont le niveau d'annotations varie beaucoup et qui présente beaucoup de redondance car elle n'est pas « nettoyée ».

- *RefSeq*[67] qui résulte d'un projet du NCBI de nettoyage de banques génomiques publiques annotées, de transcrits et de séquences protéiques. Il s'agit ainsi d'une banque non redondante qui a été manuellement nettoyée et supplémentée d'informations telles que des informations fonctionnelles par les équipes du NCBI.

- *UniProtKB*[70] qui découle de la collaboration entre l'Institut Européen de Bio-informatique (EMBL-EBI), l'Institut Suisse de Bioinformatique (SIB) et le PIR. Cette banque contient la banque PIR, TrEMBL et SwissProt.

    - La banque *UniProtKB/TrEMBL* est une banque contenant des séquences qui n'ont pas été vérifiées mais qui ont été annotées de manière informatique. Le 10 juillet 2017 elle contenait 88 032 926 séquences protéiques.

    - La banque *UniProtKB/SwissProt* qui résulte d'un travail considérable de vérification et d'annotations de séquences basé sur la littérature, débuté en 1986 et réalisé par une centaine de personnes expertes, afin d'améliorer le complètement de la banque et de

fournir une annotation des séquences de haute qualité. Le 10 juillet 2017 elle contenait 555 100 séquences protéiques.

Connaître et comprendre comment ces banques sont générées est primordial, de manière à choisir la plus adaptée à la question biologique posée lors d'une étude. Pour la recherche de biomarqueurs chez l'homme par exemple, une banque nettoyée sera privilégiée afin d'obtenir des identifications de haute qualité. Ainsi, seule la banque *UniProtKB/SwissProt* a été utilisée au cours de ce travail de thèse.

Nous distinguons, en parallèle des banques de séquence, les **banques de dépôts**[71]. Il s'agit du portail ProteomeXchange[72, 73] qui regroupe notamment PRIDE (« *PRoteomics IDEntificaiton database* »)[74], qui recueille les données des chercheurs associées à une publication scientifique, et le PeptideAtlas[75], qui avait pour objectif initial d'annoter les génomes d'eucaryotes à partir de données peptidiques obtenues par analyses protéomiques. Ces banques peuvent fournir des jeux de données pouvant servir à l'amélioration des outils bio-informatiques, ainsi qu'à l'élaboration et l'optimisation des banques de séquences.

En plus de ces banques, des plateformes existent, comme NeXtProt[76, 77], qui utilisent les séquences humaines de la banque *UniProtKB/SwissProt* et y ajoute des outils tels que des annotations fonctionnelles ou le « *peptide uniqueness checker* »[78] qui permet de vérifier l'unicité des peptides. Elle fournit des informations complètes, mises à jour, de haute qualité, et ce de manière organisée dans le but de rendre service à la communauté des scientifiques. Cette plateforme possède également un portail contenant des variants de génome dans les cancers héréditaires, ce qui peut être particulièrement utile pour les projets qui étudient ce type de maladies.

Comme évoqué précédemment, la non assignation de certains spectres MS/MS peut être due à des variants de séquence ou encore à des variants d'épissage (survenant au moment de la transcription de l'ADN en ARN mature) qui ne sont pas présents dans la banque utilisée. Même si la plateforme NeXtProt propose un portail contenant des variants de cancers, il est possible de combiner des études de transcriptomique aux études de protéomique sur de mêmes échantillons afin d'obtenir des banques spécifiques à l'échantillon. Nous entrons alors dans le domaine de la **protéogénomique**. Ce concept a été introduit pour la première fois en 2004[79]. Il définit l'ensemble des études qui lient la protéomique, la transcriptomique et la génomique. En dehors de la génération de banques spécialisées, voire même personnalisées à partir de données de transcriptomique et de séquençage du génome, la protéogénomique peut également tirer profit des données de protéomique pour améliorer l'annotation du génome et les algorithmes de prédiction des gènes[79-81]. De manière générale, les études de protéogénomique sont susceptibles d'identifier des espèces qui peuvent être

biologiquement pertinentes et qui expliqueraient le phénotype de certaines anormalités génétiques. Cependant, ces approches multi-« omiques » sont relativement récentes, et l'intégration des données demande souvent un travail manuel considérable. Il y a donc un réel besoin de solutions bio-informatiques qui permettraient de faciliter le traitement de ce genre de données, afin de rendre ces approches de protéogénomique plus routinières, car très prometteuses[82].

b. La validation des résultats d'identification

L'identification des peptides et des protéines de manière automatisée peut mener à des erreurs. En effet, les moteurs de recherche assignent souvent une identification incorrecte à un spectre MS/MS. Il n'est cependant pas envisageable de vérifier manuellement la véracité des identifications sur des jeux de données de plusieurs milliers de spectres de fragmentation, d'autant que l'annotation et la validation manuelle des spectres MS/MS sont largement subjectives. La stratégie « cible-leurre » est aujourd'hui la plus utilisée, la plus robuste et efficace pour distinguer les identifications correctes des identifications incorrectes, et ainsi garantir la qualité des résultats d'identification. Cette stratégie consiste à effectuer les recherches dans une banque contenant des séquences leurres, pour lesquelles il ne devrait pas y avoir de concordance avec un spectre MS/MS dans le jeu de données, généralement concaténée à une banque contenant les séquences protéiques cibles. Le nombre de fois qu'une séquence leurre sera assignée à un spectre MS/MS permettra d'estimer le taux d'assignements incorrects ou de faux-positifs (FDR). En effet, l'hypothèse sur laquelle repose cette stratégie consiste à dire qu'un vrai peptide ne peut être trouvé à la fois dans la banque de cibles et de leurres (ou tout du moins est censé concorder avec une séquence cible, et ce, avec un score plus élevé qu'avec une séquence leurre), et que le nombre d'assignements incorrects à partir d'une séquence cible est équivalent au nombre d'assignements incorrects à partir d'une séquence leurre[83, 84]. Ceci permet de calculer le FDR selon l'équation suivante[85] :

$$\mathrm{FDR} = 2 \times \frac{\text{Nombre de séquences leurres assignées}}{\text{Nombre de séquences leurres} + \text{Nombre de séquences cibles assignées}} \times 100$$

Les séquences leurres peuvent être générées, soit de manière stochastique, soit en inversant les séquences des protéines cibles afin de préserver la fréquence des acides aminés, la taille des peptides et des protéines, et ainsi leur masse[83]. Il n'y a à l'heure actuelle pas de consensus sur la manière de générer ces leurres ainsi que sur la manière de calculer le FDR[85]. De plus, cette stratégie est encore soumise à controverse du fait qu'il n'est pas possible de savoir combien d'identifications sont réellement incorrectes et que des informations potentiellement intéressantes peuvent par conséquent être omises.

Au cours de cette étape de traitement des données, il est également possible d'obtenir des informations quantitatives. Cependant, les stratégies de quantification des protéines étant nombreuses, elles font l'objet du paragraphe suivant.

III-Approches de quantification des protéines par spectrométrie de masse

Au-delà de l'identité des protéines, il est souvent utile d'obtenir des informations sur l'abondance des protéines présentes dans les échantillons, étant donné que beaucoup de variations protéiques résultent d'une perturbation du système biologique étudié, et ne sont détectables qu'avec une mesure quantitative. C'est la raison pour laquelle les approches quantitatives sont aujourd'hui au cœur de la plupart des études de protéomique[86].

Deux manières d'aborder la protéomique quantitative existent : la **protéomique quantitative ciblée** qui vise à quantifier uniquement certaines protéines, et la **protéomique quantitative globale** qui consiste à quantifier l'ensemble des protéines présentes dans l'échantillon. Ces travaux de thèse n'ont fait l'objet que de stratégies de quantification globale, c'est pourquoi nous nous attarderons à détailler uniquement ces approches. Parmi celles-ci, nous distinguons les approches avec marquage et sans marquage. Selon la question biologique initialement posée, une technique va être préférée à une autre, du fait de leur différence au niveau de leur polyvalence, leur coût, leur difficulté, leur degré de maturité et particulièrement sur le niveau d'expertise de la technique employée par le laboratoire qui mène la recherche[11, 86]. Etant donné qu'il n'existe pas de technique adaptable à toutes les études, chacune d'entre-elles aura une influence sur les quatre étapes du schéma d'analyse « *Bottom-up* », qui devront alors être adaptées en conséquence.

1- Les stratégies de quantification avec marquage

Les stratégies de quantification avec marquage permettent le mélange ou le multiplexage des échantillons à comparer, de manière à ce qu'ils soient analysés au cours d'une seule et même analyse. Les échantillons, qui auront été marqués par des isotopes de masses différentes avant multiplexage, pourront être distingués au moment de la mesure par MS. Ces approches ont l'avantage de ne pas souffrir de biais induits par un manque de répétabilité de l'analyse LC-MS/MS, puisque tous les échantillons sont analysés au cours d'une même analyse. Parmi les stratégies avec marquage, nous distinguons trois types d'approches :

- Le **marquage métabolique** qui consiste comparer l'intensité MS d'un peptide au sein d'un échantillon avec celle du même peptide standard marqué dans un second échantillon, pour lequel les isotopes marqués ont été introduits au moment de la croissance et de la division des cellules. Même si les avantages majeurs de cette technique sont l'introduction des isotopes marqués à une étape précoce du schéma d'analyse, limitant l'effet des biais introduits par les différentes étapes, cette méthode est limitée à un certain type d'échantillon (cultures cellulaires) et ne permet de multiplexer que deux à trois échantillons[11]. Il s'agit notamment du SILAC[87] (« *Stable Isotope Labeling with amino acids in cell cultures* ») et du Super SILAC[88].

- Le **marquage chimique** des peptides par des réactifs isotopiques, comme le marquage ICAT[89] (« *Isotope-Coded Affinity Tag* »), ou des réactifs isobariques, comme l'iTRAQ[90] (« *isobaric Tag for Relative and Absolute Quantification* ») et le TMT[91] (« *Tandem Mass Tags* »). Ces derniers sont les plus populaires puisqu'ils permettent de multiplexer jusqu'à 11 échantillons pour le TMT, contrairement à l'ICAT qui est limité à deux échantillons. L'ensemble de ces approches présentent l'avantage d'être adaptables à tous les types d'échantillon, contrairement au marquage métabolique[92]. En ce qui concerne les réactifs isobariques, ils sont conçus de manière à ce qu'un même peptide présente un même comportement en LC, le même rapport m/z et la même efficacité d'ionisation dans l'ensemble des échantillons comparés, mais qu'il soit distinguable au moment de sa fragmentation au travers d'ions rapporteurs différents selon l'échantillon et dont l'intensité va permettre l'étude quantitative différentielle. Malgré le multiplexage de plusieurs échantillons, l'échantillon résultant ne présente pas de complexité supplémentaire vis-à-vis de la séparation chromatographique du fait de ces réactifs particuliers. Cependant, la quantification s'effectuant au niveau MS/MS, ces approches sont sensibles aux contaminations survenant au moment de la fragmentation[11, 93]. Pour ce qui est des réactifs isotopiques de première génération, la forme lourde se distingue de la forme légère par l'insertion de deutériums qui modifient un tant soit peu la rétention des peptides des deux échantillons multiplexés. Ce problème a pu être contourné par l'insertion de ^{13}C dans la forme lourde des réactifs de seconde génération. La quantification s'effectue ici par l'intensité MS du peptide dans chaque échantillon.

- Le **marquage enzymatique** qui consiste à introduire la différence de masse au moment de la digestion enzymatique à l'aide d'eau lourde ($H_2{}^{18}O$), et à comparer l'intensité MS de l'échantillon marqué et non marqué[94-96]. Cette approche est peu employée du fait du taux d'incorporation variable de ^{18}O.

2- Les stratégies de quantification sans marquage

Ces dernières années, des techniques de quantification ne faisant pas intervenir de marquage isotopique, ou « sans marquage » ont fait leur arrivée, notamment du fait du développement de nouveaux spectromètres de masse plus sensibles avec des vitesses d'acquisition élevées. Ces approches présentent l'avantage d'une préparation d'échantillons facilitée et d'un nombre de conditions à comparer non défini. Elles sont particulièrement attractives lorsqu'un grand nombre d'échantillons est à comparer[77, 83], même si le multiplexage en 11-plex des approches TMT est aujourd'hui intéressant pour l'étude de larges cohortes d'échantillons. Contrairement aux techniques avec marquage, les protéomes à comparer ne sont pas multiplexés mais analysés séparément, les uns après les autres, dans les mêmes conditions[93]. Chaque étape, depuis l'extraction des protéines jusqu'à l'analyse LC-MS/MS, peut avoir un impact significatif sur la précision et l'exactitude de la quantification, et doit être maîtrisée (voir Figure III-1). C'est pourquoi la répétabilité et la robustesse de l'ensemble du schéma analytique est un prérequis majeur dans ces approches. L'ensemble des échantillons à comparer est d'ailleurs souvent analysé au sein d'une même séquence sur un même instrument[11, 93]. Pour pallier à d'éventuels biais techniques survenant au cours du traitement ou de l'analyse des échantillons, une étape de normalisation des données est souvent nécessaire.

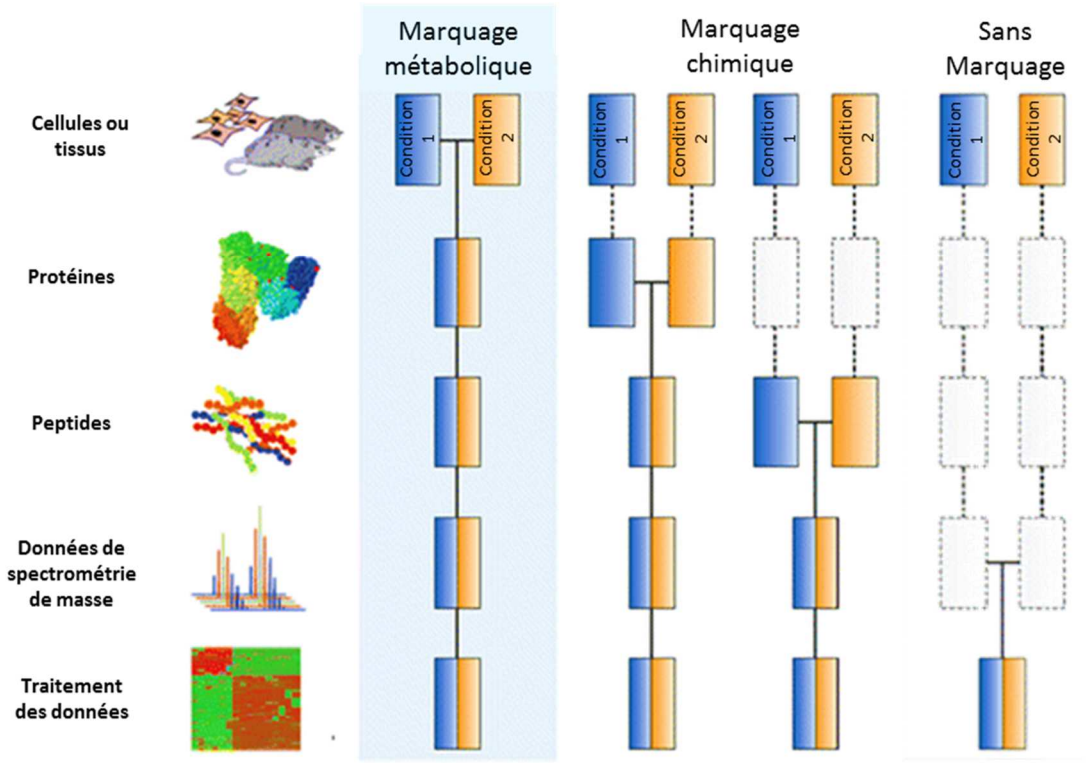


Figure III-1 - Résumé des approches de quantification par spectrométrie de masse (adapté de [11])
Le rectangle bleu et le rectangle orange représentent les 2 conditions à comparer. Le rectangle à la fois bleu et orange représente le multiplexage des 2 conditions. Quant aux rectangles en pointillés, ils représentent les étapes durant lesquelles les variations et les erreurs de quantification peuvent survenir

Ces approches permettent d'obtenir une estimation grossière de la quantité de protéine présente dans un échantillon, information largement suffisante pour des études de découvertes permettant d'émettre des hypothèses qui devront être validées subséquemment.

Ces travaux de thèse ont fait uniquement l'objet de ces techniques sans marquage qui seront développées ci-après. Nous distinguons : le comptage de spectres MS/MS (ou « *spectral count* »), l'extraction des courants d'ions (quantification MS1 ou encore XIC pour « *eXtracted Ion Current* »), et la DIA. Parmi ces stratégies, seules les deux premières approches ont été employées au cours de ces travaux de thèse.

a. Le comptage de spectres MS/MS

La quantification sans marquage par comptage de spectres MS/MS est basée sur la corrélation entre l'abondance d'une protéine et le nombre de spectres MS/MS (ou PSM pour « *Peptide Spectrum Match* ») ayant conduit à l'identification de cette protéine [97]. Cette approche a l'avantage de faciliter le traitement de données, puisque les résultats sont directement obtenus des outils utilisés pour l'identification et la validation des protéines. Cependant, cette stratégie est fortement impactée par la

stochastique de la DDA qui rend les données d'un échantillon à l'autre peu répétables. De plus, la linéarité de la méthode est influencée par le paramétrage du temps d'exclusion dynamique[92] : un temps d'exclusion de l'ion peptide une fois fragmenté trop long sous–évaluera l'abondance du peptide, alors qu'un temps d'exclusion trop court diminuerait drastiquement la couverture de séquence du protéome. De plus, l'absence de PSM dans une condition n'est pas forcément synonyme d'absence de la protéine, mais peut résulter du sous-échantillonnage inhérent au mode d'acquisition.

Une des considérations à prendre en compte lors de l'emploi de cette stratégie est le poids à donner à un PSM qui est attribué à un peptide non unique à une protéine, mais qui est partagé entre deux ou plusieurs protéines. Il est alors souvent d'usage de pondérer ces PSM de manière à les distribuer proportionnellement aux protéines concernées en fonction de leur nombre de PSM uniques[11]. Par ailleurs, comme évoqué plus tôt, il est prudent de normaliser les données de quantification sans marquage afin de prendre en compte les éventuelles variations pouvant survenir lors du schéma analytique. Diviser chaque valeur par le nombre total de PSM pour chaque condition est un moyen simple de normaliser les données de comptage de spectres.

Notons qu'il est possible d'estimer une quantification « absolue » en divisant le nombre de PSM pour une protéine par sa masse moléculaire[98], le nombre d'acides aminés[99] ou encore en utilisant l'index PAI (« *Protein Abundance Index* »)[100], l'emPAI (« *exponentially modified Protein Abundance Index* »)[101] ou encore l'APEX (« *Absolute Protein Expression* »)[102, 103]. Ces méthodes restent cependant peu précises et justes.

b. L'extraction des courants d'ions

Principe

La quantification par extraction des courants d'ions repose sur l'observation que la réponse en MS d'un peptide dans des conditions expérimentales similaires est linéairement corrélée à sa concentration[104, 105]. Ainsi, cette stratégie emploie le signal MS des peptides, soit en intégrant l'intensité de chaque ion sur son profil d'élution chromatographique, soit en considérant l'intensité à l'apex du pic d'élution du peptide pour inférer une quantité protéique[11, 92]. Cette inférence est faite en agrégeant les différences observées au niveau peptidique en sommant, en calculant la moyenne pondérée ou encore la médiane des valeurs peptidiques. Notons cependant que la moyenne et la médiane sont souvent préférées car plus robustes vis-à-vis des valeurs extrêmes[92].

L'avantage majeur de cette stratégie par rapport à celle du comptage de spectres MS/MS est qu'elle limite les effets de la stochastique de la DDA de par le report de l'identification effectuée dans une condition à toutes les conditions. Plus clairement, si un peptide a été séquencé et donc identifié uniquement dans une condition, sa masse pourra tout de même être extraite au sein de toutes les autres conditions à comparer pour y être quantifié. Attention cependant à ce que l'extraction du signal soit assignée de manière non ambiguë au bon peptide par le logiciel de traitement des données XIC. Cet assignement est facilité par l'utilisation de spectromètres de masse de haute résolution avec une bonne exactitude de masse, mais aussi par l'utilisation d'un couplage robuste qui fournit des Tr stables qui vont grandement aider à la distinction non ambiguë du bon signal[11, 93]. Cet avantage rend la quantification XIC plus exacte que le comptage de spectres MS/MS puisqu'elle fournit moins de données manquantes, et ce, sur une gamme dynamique plus large. En effet, le temps d'exclusion peut ici être largement augmenté (à condition d'avoir un seuil de sélection adapté pour générer des spectres MS/MS de suffisamment bonne qualité) pour améliorer la couverture du protéome, étant donné qu'un spectre MS/MS sert ici uniquement à identifier le peptide, limitant ainsi le phénomène de saturation par les espèces abondantes. Il faut toutefois rester vigilant à ne pas trop maximiser la couverture du protéome en augmentant le nombre de MS/MS effectués au sein d'un même cycle, et ainsi le temps de cycle, et par conséquent perdre en précision de quantification en diminuant le nombre de points obtenus pour un pic, soit la résolution (Figure III-2). Trouver la bonne balance est une considération importante de cette stratégie. Cependant, celle-ci peut être contournée en effectuant deux analyses pour une même condition : le paramétrage de la première analyse permettrait d'identifier un maximum de peptides/protéines (en favorisant un nombre de MS/MS élevé), et le paramétrage de la seconde analyse permettrait quant à lui de les quantifier (en favorisant le nombre de points par pic).

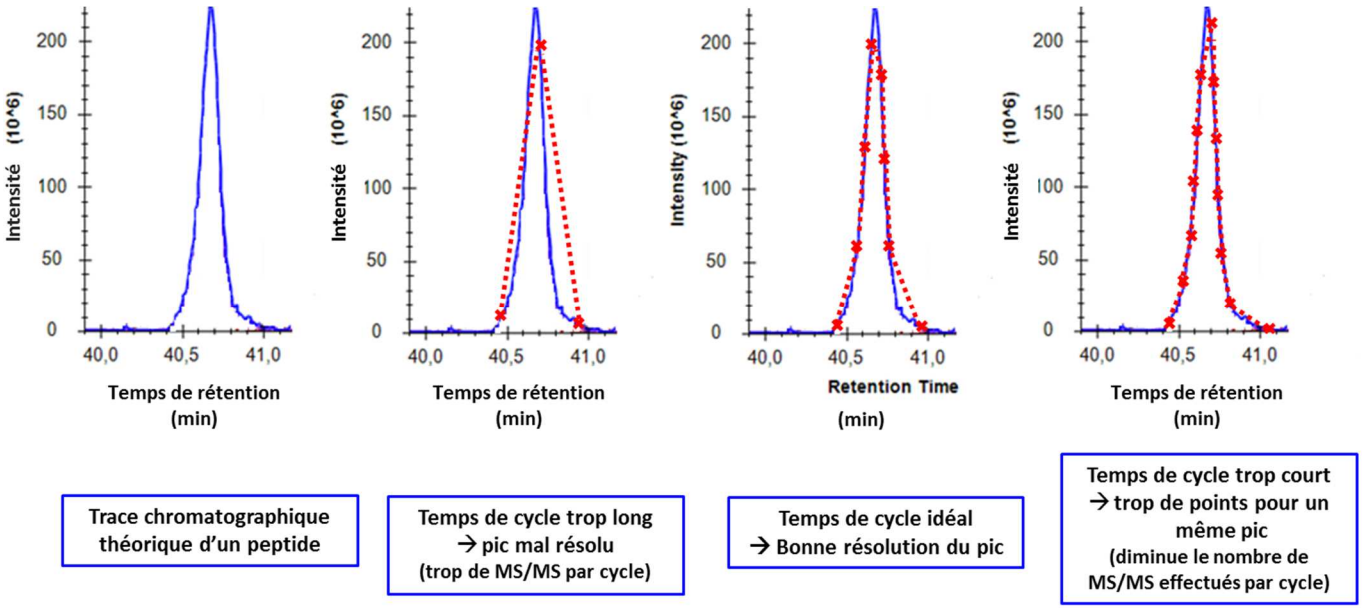


Figure III-2 - Schéma de la trace chromatographique d'un peptide après l'extraction de son courant d'ion
Le pic théorique est représenté en bleu. Les croix indiquent l'acquisition d'un spectre MS permettant de définir la résolution MS du pic expérimental (en pointillés rouges). Ainsi, la distance entre deux croix rouges reflète le temps de cycle.

Un changement dans l'abondance d'un peptide peut être le reflet d'une conséquence biologique, mais pas seulement. Il peut également provenir d'un biais provenant de la préparation d'échantillons, voire de l'instrumentation[93]. En effet, l'introduction de substances interférentes au même Tr et à un ratio m/z très proche dans une des conditions peut fausser la quantification du peptide considéré. C'est pourquoi une attention particulière doit être accordée au moment de la manipulation des échantillons pour que celle-ci soit le plus répétable possible. Concernant la préparation d'échantillons, la décomplexification est souvent écartée dans ce type d'approche. Certes, elle permettrait d'augmenter la couverture du protéome, mais aurait malheureusement un impact non négligeable sur le coût, puisqu'elle résulte en un temps machine plus long, en plus d'impacter la répétabilité d'un échantillon à l'autre et de compliquer le traitement de données. De plus, la vitesse d'acquisition et la résolution des nouveaux spectromètres de masse permettent d'identifier et de quantifier un nombre suffisant de protéines au sein d'une seule analyse pour un échantillon complexe par des approches de découverte, limitant ainsi le temps machine. Afin de pallier aux biais pouvant survenir au cours des analyses, il est souvent d'usage de normaliser les données de quantification XIC. Ce point sera d'ailleurs abordé dans le paragraphe *III-2-b-Logiciels disponibles* suivant.

Notons que tout comme le comptage de spectres MS/MS, il est possible d'estimer des quantités « absolues » avec cette stratégie, soit en divisant la somme de toutes les intensités peptidiques par le nombre théorique de peptides observables, stratégie souvent appellée iBAQ pour « *intensity-Based Absolute Quantification* »[106], soit en utilisant la moyenne des trois peptides les plus intenses par protéine[107].

Malgré les avantages que présentent cette approche par rapport au comptage de spectres MS/MS, elle nécessite un temps plus important pour le traitement de données qui est effectué à l'aide d'outils dédiés.

Logiciels disponibles

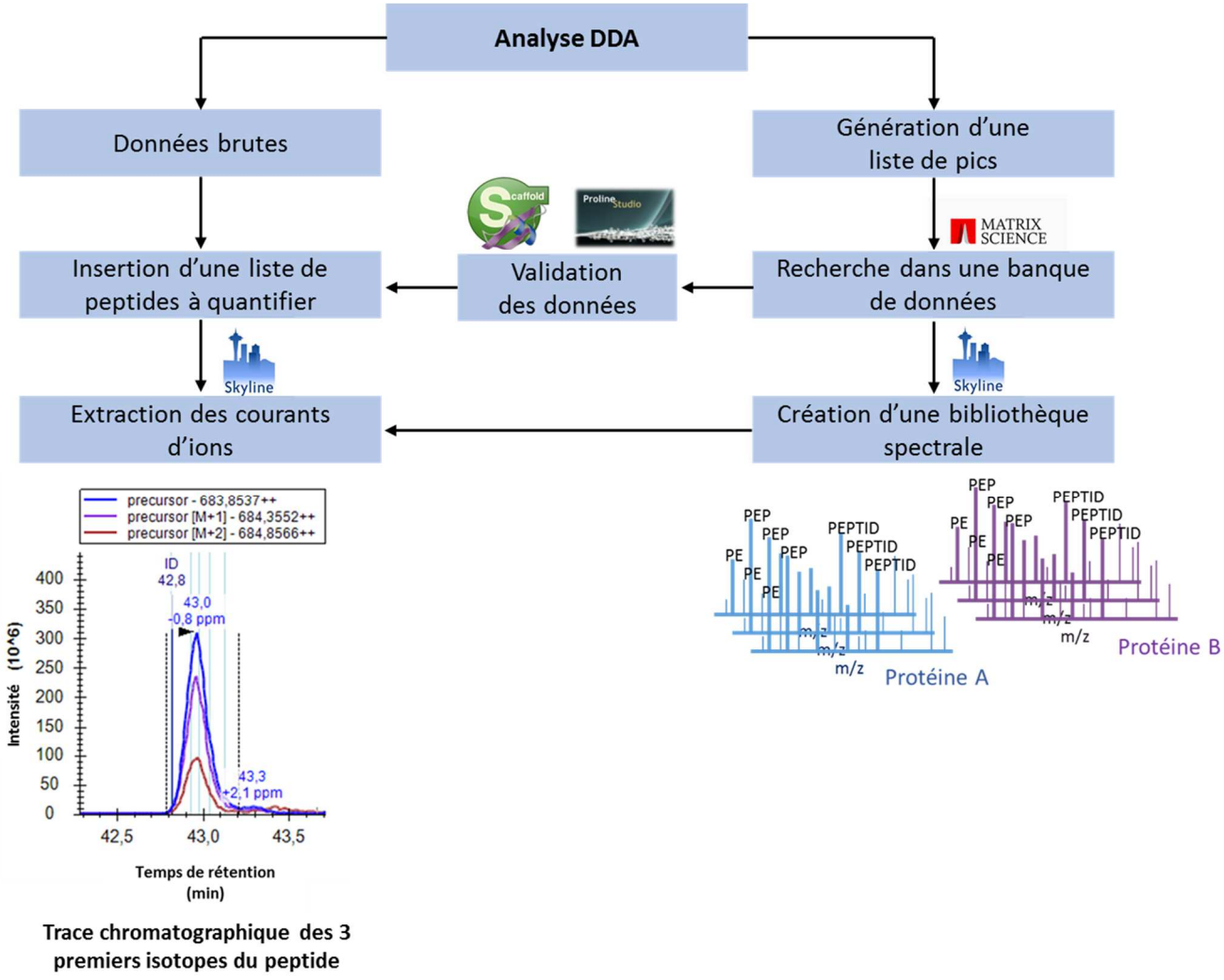
Les données de quantification XIC nécessitent l'utilisation de logiciels dédiés et robustes qui permettent l'alignement des analyses à l'aide du Tr et des ratios m/z afin de faire concorder des distributions isotopiques identiques. Actuellement, de nombreux logiciels, aussi bien commerciaux que publics sont disponibles, parmi lesquels certains sont implémentés dans des pipelines[11]. Généralement, les utilisateurs attendent de ces logiciels de bonnes performances avec un faible nombre de faux-positifs, tout en offrant une interface graphique conviviale qui permette notamment de visualiser, voire de modifier les extractions du signal MS effectuées[108]. La facilité d'installation, ainsi que la disponibilité de documentations, la possibilité de communiquer avec les utilisateurs et les développeurs sont également des paramètres qui entrent en compte lors du choix d'un logiciel[92]. Nous distinguons parmi les outils commerciaux Progenesis LC-MS et Mascot Distiller, et parmi les outils publics disponibles gratuitement Proline[109], OpenMS[110], Viper[111], MFPaQ[112], Skyline[113] et Maxquant[114]. Ces cinq derniers ont été employés par RAMUS *et collaborateurs* et ont montré une haute sensibilité et une capacité à détecter de fortes variations, tout comme des ratios entre deux conditions modérés[108].

Ces outils proposent généralement d'effectuer la détection des enveloppes isotopiques des peptides en les dissociant des interférences afin de générer des cartes LC-MS à partir des ratios m/z, des Tr, de la charge et de l'intensité du pic. Il s'agit là de l'étape la plus importante du traitement de données qui devrait être précédée d'étapes de réduction de la taille des données, de filtration et soustraction du bruit, et surtout de calibration en masse et d'alignement des Tr. En effet, un mauvais un alignement des Tr peut mener à des erreurs d'extraction du signal. Cependant, les algorithmes de nombreux logiciels sont souvent peu décrits et ces outils souffrent d'un manque de documentation et d'interfaces graphiques conviviales[92].

Les logiciels les plus couramment utilisés aujourd'hui sont Skyline et MaxQuant[108], pour lesquels les algorithmes sont décrits et disponibles, ainsi que des documentations et des groupes de discussions. Ces deux outils ont été utilisés au cours de ces travaux de thèse. Sachant que la qualité des résultats est influencée à la fois par le paramétrage et l'expertise de l'utilisateur du programme, et que la documentation sur l'utilisation de MaxQuant pour la quantification XIC a tardé à être publiée, le laboratoire a fait le choix au début de ces travaux de thèse de traiter les données avec Skyline. Ce choix

a progressivement évolué au cours de ces 3 années, et s'est tourné vers MaxQuant, notamment du fait des publications de Jürgen COX[115] et Stefka TYANOVA[116].

En ce qui concerne le logiciel **Skyline**, il a été originellement développé pour le traitement de données de protéomique quantitative ciblée de type « *Selected-Reaction Monitoring* » (SRM), mais son utilisation a été étendue à la quantification XIC dans le but de fournir une plateforme constructeur-indépendante pouvant être utilisée par tous les laboratoires. Contrairement à ce qui a été décrit plus haut, cet outil n'effectue pas de réalignement des analyses en fonction du Tr, ni de recalibration en masse et de détection des enveloppes isotopiques. Il effectue à partir d'une bibliothèque qui contient les résultats d'identification MS/MS, une extraction directe du signal MS à partir des ratios m/z et Tr associés à un PSM de la bibliothèque, dans un intervalle de temps défini. Le signal d'une espèce présente dans la bibliothèque va être extrait sur l'ensemble des analyses à partir des mêmes ratios m/z et Tr, même si elle n'a pas été séquencée dans l'ensemble des analyses (Figure III-3). Cependant, cette extraction mène souvent à des erreurs puisque ce logiciel ne propose pas de réaligner les Tr. Ces erreurs peuvent être facilement repérées et corrigées, notamment à l'aide d'un score (« *isotope dot product* ») qui compare la distribution de l'intensité relative de chaque isotope attendue et observée, ce qui présente un certain avantage. Malgré cet avantage, l'utilisation de Skyline est relativement chronophage du fait que les analyses brutes doivent dans un premier temps être transformées en listes de pics pour être soumises à une recherche dont les résultats doivent ensuite être validés, mais aussi parce que les erreurs d'extraction nécessitent une vérification manuelle longue qui permet de réduire le nombre de faux-positifs, de faux-négatifs et d'améliorer la sensibilité. De plus, si seuls les peptides uniques ou ne contenant qu'un certain type de modifications doivent être quantifiés, ils seront ajoutés manuellement de manière à n'extraire que le signal des peptides d'intérêt. Ces interventions humaines peuvent être sources d'erreurs qui peuvent influencer le résultat final.



Trace chromatographique des 3 premiers isotopes du peptide

Figure III-3 - Schéma de fonctionnement de Skyline pour la quantification XIC

Le logiciel **MaxQuant** est quant à lui un ensemble d'algorithmes intégrés, développé initialement pour les données de SILAC, mais qui est aujourd'hui utilisable pour des données provenant de la majorité des techniques de quantification globale, et générées par des spectromètres de masse de haute résolution. Ce logiciel a été conçu pour être facilement utilisable (il est d'ailleurs conseillé de ne modifier qu'une dizaine de paramètres pour des études simples). En effet, MaxQuant nécessite comme fichier d'entrée les données brutes de quatre marques théoriquement, et effectue à la fois l'identification, la validation et la quantification des protéines notamment grâce à l'implémentation d'Andromeda. Les étapes s'effectuent selon l'ordre décrit dans la Figure III-4. Cet outil a été testé au laboratoire sur des données issues d'instruments commercialisés par THERMO FISHER SCIENTIFIC, BRUKER et SCIEX, et s'est montré performant uniquement pour le traitement des données BRUKER et THERMO FISHER SCIENTIFIC.

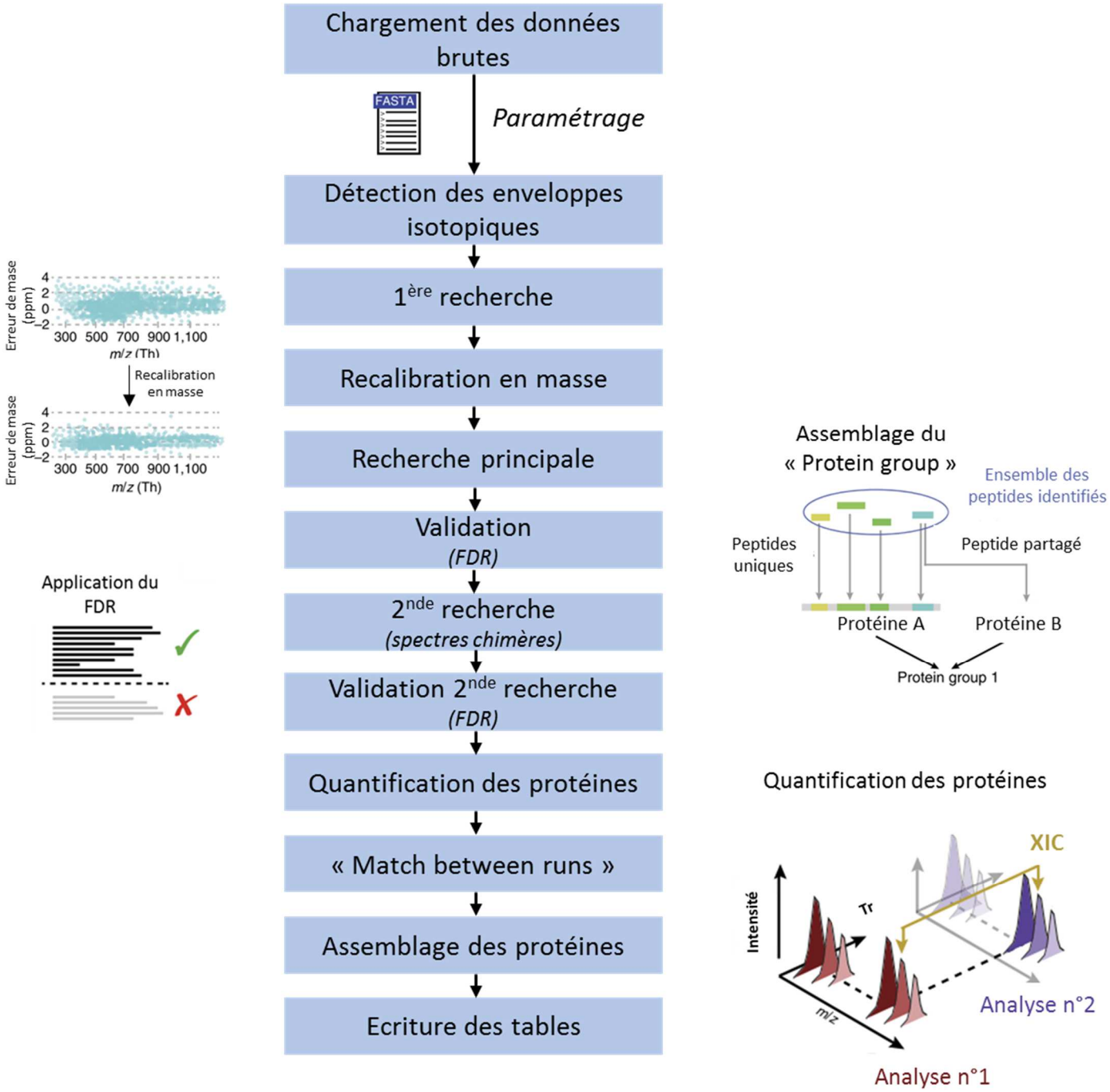


Figure III-4 - Résumé de la pipeline informatique de MaxQuant (adapté de [116])

L'avantage majeur de cet outil est la rapidité avec laquelle le traitement de données est effectué, sans intervention humaine une fois le paramétrage initial réglé. Les développeurs ont fourni un gros travail sur la détection des enveloppes isotopiques ainsi que sur le réalignement des cartes LC-MS, le transfert d'identification grâce à l'option « *Match between runs* » qui fonctionne particulièrement bien du fait du réalignement des cartes, et l'implémentation d'une étape de normalisation des données de quantification protéique. La méthode de normalisation générant des données « LFQ » s'effectue en comparant deux à deux les analyses, en partant du principe que la majorité des protéines ne varient

pas entre deux protéomes (hypothèse partagée par de nombreuses méthodes de normalisation). Ainsi, le comportement de la moyenne est considéré comme standard relatif. Les valeurs LFQ se sont d'ailleurs avérées être un bon compromis entre sensibilité et taux de faux-positifs[108]. Malgré ces avantages, MaxQuant ne propose pas de visualisation de l'intégration du signal, et donc pas de moyen de la corriger, ce qui peut être particulièrement handicapant lorsque le signal de protéines différentiellement exprimées nécessite d'être vérifié avant de se diriger vers une étape de validation. C'est pourquoi Skyline est encore utilisé pour la visualisation des données intéressantes après un traitement MaxQuant. Par ailleurs, l'infrastructure française de protéomique ProFI dont le laboratoire fait partie, a développé un outil nommé Proline qui permettrait d'éviter l'utilisation successive de ces deux logiciels. En effet, en plus de la validation des résultats d'identification, cet outil permet la quantification XIC en combinant les avantages de Skyline et MaxQuant, à savoir le même principe d'extraction du signal que MaxQuant, et la possibilité de visualiser et réintégrer le signal, si nécessaire, comme Skyline.

c. « *Data-Independent Acquisition* » ou DIA

Le mode d'acquisition DIA permet de générer des données qui apportent davantage d'informations, puisqu'il consiste à séquencer l'ensemble des peptides détectables dans un échantillon. En effet, les spectres d'ions fragments sont collectés de manière systématique et indépendante de toute information concernant le précurseur[117]. L'acquisition des spectres MS/MS peut s'effectuer de plusieurs manières :

- Suite à l'isolation et la fragmentation séquentielles d'ions précurseurs contenus dans des fenêtres d'isolation de quelques m/z, comme c'est le cas pour le **SWATH™**[118]. Ainsi, au cours d'un même cycle, le spectromètre de masse va effectuer autant de fenêtres d'isolation que nécessaire, dans lesquelles il va fragmenter l'ensemble des précurseurs présents, de manière à couvrir l'ensemble de la gamme de masse (Figure III-5). Ce type d'acquisition segmentée génère des cartes complexes d'ions fragments dans un espace bidimensionnel (Tr et m/z). L'avantage de cette stratégie est qu'un précurseur abondant va uniquement impacter la gamme dynamique de la fenêtre d'isolation dans laquelle il est contenu en la limitant, et non pas les précédentes et suivantes. Il est possible dans ce type d'approches de paramétrer soit des fenêtres fixes, soit des fenêtres variables qui permettent de diminuer la complexité des spectres MS/MS et d'augmenter la spécificité en s'adaptant à la densité des précurseurs[119].

- Suite à l'alternance de spectres à basse et à haute énergie acquis sur l'ensemble de la gamme de masse. Il s'agit du concept développé par WATERS nommé **MS^E** qui va, à l'aide d'un spectre

de basse énergie, cartographier les précurseurs, puis à l'aide d'un spectre à haute énergie, fragmenter l'ensemble des précurseurs[120]. Ce mode d'acquisition a été amélioré par l'ajout d'une cellule de mobilité ionique au sein du spectromètre de masse, qui permet de séparer des espèces qui co-éluent au même Tr en jouant sur leur temps de dérive au sein de cette cellule, qui dépend lui-même de la charge et de la conformation de l'espèce, et que l'on nomme HDMS^{E} [121]. Pour les spectromètres ne contenant pas de cellules de mobilité ionique, WATERS a récemment introduit la technique SONAR™ qui permet de séparer des espèces co-éluées à l'aide de rampes de m/z effectuées par le quadripôle. Ce type de données ne peut généralement être traité qu'à l'aide du logiciel constructeur, c'est pourquoi les stratégies de type SWATH™ sont souvent préférées.

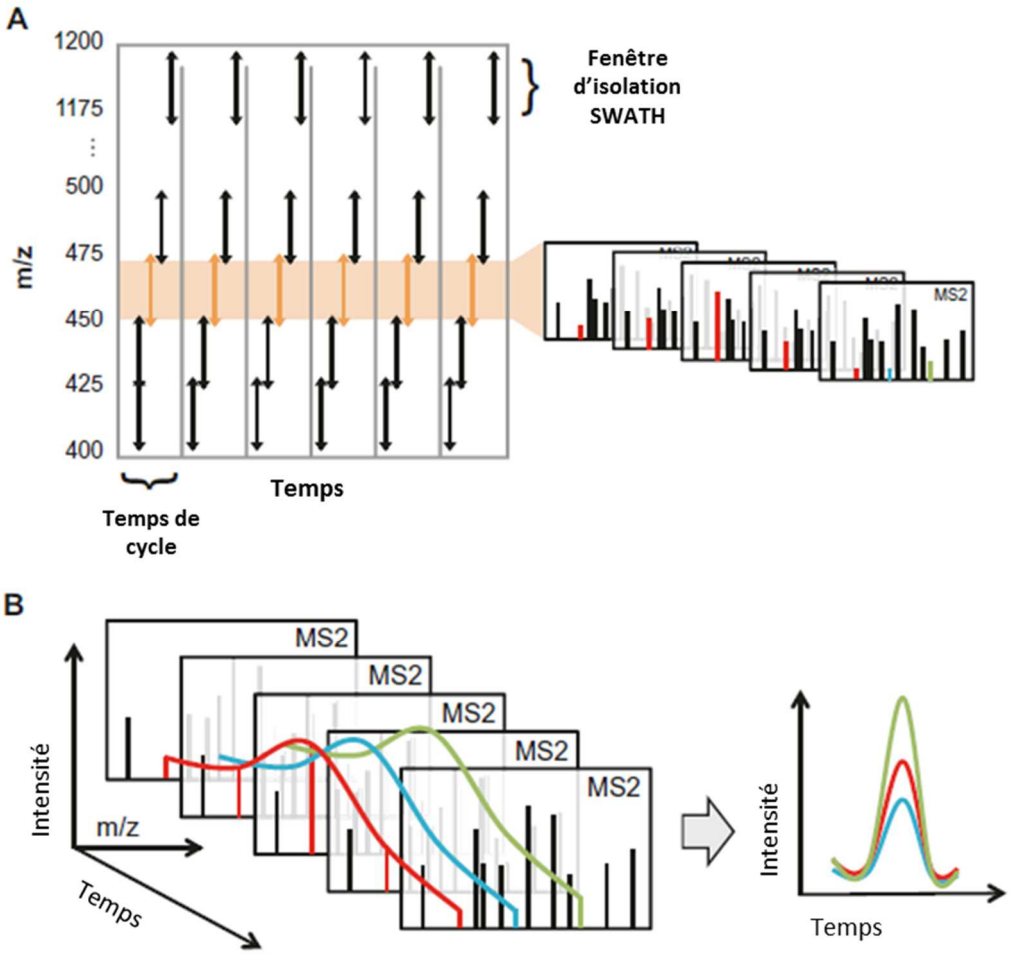


Figure III-5 – Schéma d'acquisition de données SWATH™ - adapté de [119]
A) Explication du cycle qui contient plusieurs fenêtres d'isolation (en m/z)
B) Extraction des intensités des ions fragments à partir des spectres MS/MS (ou MS2) pour donner lieu à un groupe de pics XIC

Quelle que soit la stratégie adoptée, les spectres de fragmentation générés présentent une complexité considérable par rapport à ceux générés par le mode DDA, et ne peuvent faire l'objet tels quels d'une recherche classique dans les banques de données pour l'identification des peptides et des protéines[122].

Ce mode d'acquisition a été pour la première fois introduit par PURVINE[123] pour contourner les problèmes liés à la stochastique du mode DDA (couverture du protéome limité et répétabilité insuffisante de détection entre réplicats). Ainsi, une analyse effectuée à faible voltage permettait d'obtenir les informations des précurseurs, et une seconde à haut voltage induisait une fragmentation. C'est seulement en 2004 qu'une notion supplémentaire de quantification par DIA est apparue suite à l'isolation et la fragmentation séquentielle des précurseurs[124]. La quantification s'effectue alors au niveau MS/MS, ce qui présente l'avantage d'être plus sensible du fait d'un meilleur rapport signal sur bruit, plus spécifique, et avec une grande gamme dynamique.

Afin d'obtenir des informations quantitatives, il est nécessaire d'extraire correctement le signal des peptides à partir de ces données DIA. Jusqu'ici, l'extraction était majoritairement effectuée de manière ciblée, ce type de données se substituant alors à la SRM puisqu'il permet de quantifier les peptides avec une bonne répétabilité et avec la même exactitude que la SRM. La DIA offre en théorie la possibilité tant espérée des protéomistes de quantifier l'ensemble des peptides présents dans un échantillon, mais cette possibilité est limitée par l'extraction des données qui s'effectue à l'aide de bibliothèques spectrales obtenues par une analyse DDA préalable, ou d'autres données DDA disponibles dans les banques de dépôt par exemple. Le nombre de peptides quantifiés est alors limité au nombre maximum de peptides identifiés dans la librairie, cependant cette quantification tire tout de même avantage d'une meilleure sensibilité et spécificité offertes par la quantification au niveau MS/MS. A l'heure actuelle, cette extraction des données reste un défi et confine malheureusement souvent la DIA à la quantification ciblée. Cependant, des espoirs reposent sur de nouveaux outils, comme DIA-Umpire[125] ou Group-DIA[126] qui visent à retrouver le lien précurseur-fragments au sein des données de DIA et à générer des pseudo-spectres MS/MS afin de les soumettre à une recherche classique d'identification dans les banques de données, dans le but ultime d'identifier l'ensemble des peptides et des protéines afin d'extraire les données correspondantes permettant de les quantifier.

3- Conclusion

Au travers des différentes stratégies évoquées dans ce chapitre, nous avons pu nous rendre compte que la boîte à outils de la protéomique « *Bottom-up* » est vaste et contient de nombreuses approches qui présentent chacune des forces mais aussi des faiblesses. Le travail du protéomiste consiste ainsi à choisir le spectromètre de masse (en fonction de sa résolution, de son exactitude de masse et de sa

gamme dynamique), la stratégie d'acquisition, ainsi que le traitement de données les plus adaptés aux échantillons à analyser[127].

CHAPITRE II

DEVELOPPEMENT ET OPTIMISATION DE METHODES DE PREPARATION D'ECHANTILLONS POUR LA PROTEOMIQUE QUANTITATIVE SANS MARQUAGE

CHAPITRE II – DEVELOPPEMENT ET OPTIMISATION DE METHODES DE PREPARATION D'ECHANTILLONS POUR LA PROTEOMIQUE QUANTITATIVE SANS MARQUAGE

Les approches de protéomique quantitative sans marquage ne permettant pas de multiplexer les différentes conditions à comparer, elles nécessitent d'analyser de manière séparée et successive l'ensemble des échantillons d'une étude. Dès lors, une répétabilité élevée et une maîtrise de l'ensemble du schéma analytique sont indispensables afin de garantir la similarité des conditions d'analyse de chacun des échantillons, dans le but d'éviter l'introduction de biais techniques pouvant mener à une quantification, voire une conclusion biologique erronées. Comme évoqué au cours du *Chapitre I*, les premières étapes du schéma analytique, intimement liées, que sont l'extraction de protéines et la préparation d'échantillons sont à l'origine de la majorité des variabilités pouvant intervenir dans les études de protéomique quantitative.

Afin d'assurer un maximum de répétabilité au cours des études de protéomique quantitative sans marquage, ces travaux de thèse ont principalement porté sur le développement et l'optimisation de protocoles de préparation d'échantillons qui impliquent peu, voire pas de fractionnement de l'échantillon, dans l'objectif :

- De **limiter les biais inhérents à la répétabilité du fractionnement**.

- De **réduire le temps machine** nécessaire à l'analyse d'un échantillon, puisque les spectromètres de masse de dernière génération présentent des vitesses d'acquisition rapides et des gammes dynamiques plus larges, compatibles avec l'analyse de mélanges complexes. Cela permet en outre d'assurer la stabilité du couplage LC-MS tout au long des analyses de manière plus aisée, mais également de minimiser les pertes de peptides hydrophobes qui peuvent être adsorbés sur les parois du contenant lorsque les échantillons d'une séquence ne sont pas analysés de suite.

- De **limiter la manipulation de l'échantillon** par le biais de préparations d'échantillons simples et rapides. La perte de protéines et de peptides est minimisée lorsque les protocoles impliquent peu d'étapes.

L'ensemble de ces objectifs permet **d'augmenter le nombre de conditions à comparer** tout en gardant un temps d'analyse raisonnable.

Etant donné qu'il n'existe pas de préparation d'échantillons universelle, adaptable à la diversité des échantillons à analyser ainsi qu'aux nombreuses questions biologiques pouvant faire l'objet d'une analyse protéomique, le développement et l'optimisation de préparations d'échantillons pour différents types d'échantillons ont été effectués.

I- Optimisation d'une préparation d'échantillons pour l'analyse protéomique d'un enrichissement membranaire de type « *ghosts* » *membranaires*

Les protéines membranaires sont encodées par 20 à 30 % du génome[128]. Elles se situent à un endroit clé de la cellule, soit à la jonction entre le compartiment intracellulaire et l'environnement extracellulaire. De ce fait, elles sont impliquées dans de nombreux processus biologiques fondamentaux de la cellule, comme les échanges de matériel et d'énergie entre la cellule et son environnement, la communication entre les cellules, et le transport de signaux[129]. Leur rôle majeur dans ces processus en font des cibles thérapeutiques préférentielles[26, 130]. Ces protéines sont d'ailleurs aujourd'hui la cible de 50 % des médicaments sur le marché. De surcroît, les récents développements de thérapies à base d'anticorps monoclonaux dirigés contre la surface des cellules augmentent l'intérêt des scientifiques pour les protéines membranaires[130].

Ces protéines représentent cependant un défi analytique du fait qu'elles sont :

- **Difficiles à extraire** car insérées dans la bicouche lipidique de la membrane des cellules, et que leur caractère hydrophobe les rend très peu solubles,
- **Faiblement abondantes**,
- Présentent **peu de sites de coupures** à la trypsine,
- Et ont **tendance à s'agréger**.

Afin de faciliter leur analyse et d'augmenter leur nombre d'identifications, il est souvent nécessaire d'effectuer un enrichissement en protéines membranaires, en plus d'employer une préparation d'échantillons compatible avec l'utilisation de détergents pour permettre leur solubilisation[26, 129-131]. Les stratégies d'enrichissement permettent d'effectuer un sous-fractionnement de l'échantillon en augmentant la proportion de protéines membranaires. Parmi ces stratégies, nous distinguons les méthodes de centrifugation différentielle, de purification par affinité entre deux phases (aqueuse et polymérique), des méthodes qui emploient l'interaction forte entre la biotine et la streptavidine[26, 129, 130, 132], etc.

Par ailleurs, d'autres tentatives d'enrichissement en protéines membranaires ont été menées au laboratoire, parmi lesquelles :

- L'induction de microparticules sur des cellules en culture[133, 134]. Il s'agit d'une méthode chronophage, ardue, faiblement répétable, difficile à implémenter sur un grand nombre d'échantillons, et qui nécessite un nombre important de cellules,

- Le recueil de microparticules circulantes[135] au sein du plasma humain,

- Une approche de type « *ghosts* » *membranaires* qui pourrait être une alternative lorsque l'étude porte sur des cellules.

1- Les « *ghosts* » *membranaires*

Dans le cadre du projet de recherche de biomarqueurs de glioblastomes (voir *Chapitre III-I*), mené en collaboration avec l'équipe du Dr Jacques HAIECH, un enrichissement de protéines membranaires à partir de cellules souches cancéreuses (CSC) par une approche de « *ghosts* » *membranaires* a été envisagée à l'initiative du collaborateur. Cet enrichissement consiste à lyser mécaniquement les cellules, afin d'éliminer le contenu intracellulaire et de récupérer les « *ghosts* » *membranaires* ainsi générés par des étapes de centrifugation successives. Le protocole utilisé dans le cadre de ce projet est détaillé dans la Figure I-1.

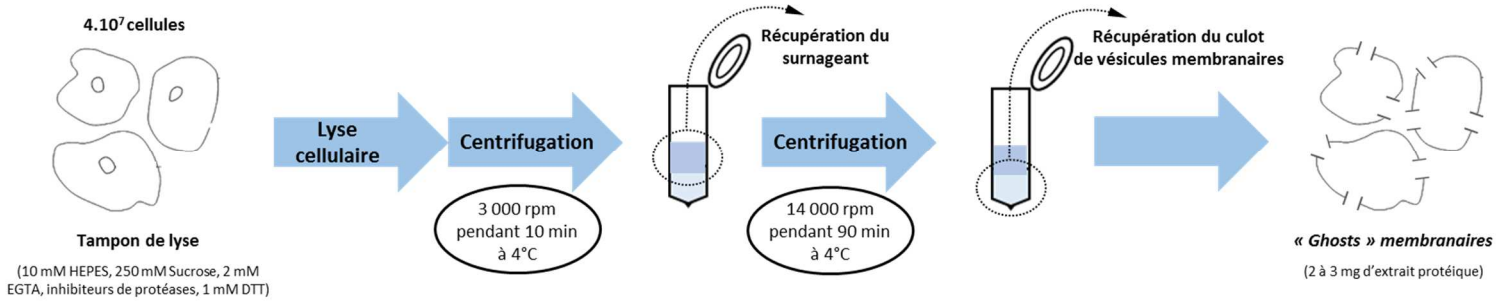


Figure I-1 - Schéma de préparation de l'enrichissement en protéines membranaires de type « *ghosts* » *membranaires* (adapté de la thèse d'Emilie Audran[136])

2- Evaluation des « *ghosts* » *membranaires* pour la protéomique quantitative sans marquage XIC

L'identification des protéines membranaires est dépendante non seulement de la méthode d'enrichissement utilisée, mais aussi de la solubilisation des protéines et de la préparation d'échantillons. Une étude préliminaire a été menée au laboratoire sur des « *ghosts* » *membranaires* de CSC, qui impliquait la décomplexification de l'échantillon en 90 bandes de gel 1D SDS-PAGE et une quantification par comptage de spectres MS/MS. Ce choix découlait de la disponibilité du parc instrumental au moment de l'étude (en 2011), et résultait en un temps d'analyse d'une semaine par échantillon analysé, limitant ainsi le nombre de conditions à comparer et pouvant être analysées au cours d'une même séquence d'analyses. L'arrivée de spectromètres de masse de nouvelle génération (de type Impact-HD de BRUKER avec des vitesses d'acquisition, une exactitude en masse et une résolution accrues) concomitante avec mon arrivée au laboratoire a permis d'envisager une optimisation de la préparation d'échantillons. Cette dernière devait être adaptée à la nouvelle instrumentation, qui offrait la possibilité non seulement d'analyser des mélanges plus complexes permettant de réduire le fractionnement de l'échantillon et ainsi la durée d'analyse pour un échantillon, et permettre d'implémenter une stratégie de quantification par extraction des courants d'ions (XIC), qui dépend moins du mode d'acquisition DDA que le comptage de spectres MS/MS. Le Tableau I-1 montre les performances du couplage LC-MS/MS utilisé au cours de de l'étude préliminaire (nanoLC-Chip couplée à un spectromètre de masse de type trappe à ions, amaZon BRUKER) ainsi que celles du couplage utilisé pour l'optimisation de la méthode (nanoLC couplée à un spectromètre de masse de type Q-TOF, Impact-HD BRUKER) obtenues suite à l'analyse d'un digeste de levure.

	NanoLC-Chip (AGILENT - C18, 150mm x 75µm, 5µm) **Trappe à ions** (AmaZon, BRUKER)	**NanoLC** (Nano-Acquity, WATERS - C18, 250mm x 75µm, 1,7µm) **Q-TOF** (Impact-HD, BRUKER)
Nombre de protéines identifiées Avec FDR < 1%	147	1237
Nombre de spectres assignés à une identification	3235	58369
Nombre de spectres assignés à une identification avec un score > 20 *(Score Mascot)*	2644	52348

Tableau I-1 - Résumé des performances des deux couplages sur l'analyse de 1 µg d'un digeste de levure avec un gradient chromatographique de 79min

a. Réduction du fractionnement : comparaison de trois protocoles de préparation d'échantillons

La possibilité de réduire le fractionnement a dans un premier temps été évaluée en comparant trois protocoles de préparation d'échantillons. L'objectif étant de trouver une préparation qui permet :

- D'obtenir une **profondeur d'analyse équivalente** aux 90 bandes de gel 1D SDS-PAGE,
- De **réduire la durée d'analyse** pour augmenter le nombre de conditions à comparer au sein d'une même séquence,
- De **limiter la manipulation d'échantillon** pour assurer un protocole répétable qui permet l'implémentation d'une quantification sans marquage de type XIC.

L'utilisation de détergents pour solubiliser les protéines membranaires est souvent requise car ils miment la membrane lipidique. Parmi les plus connus et les plus performants vis-à-vis de ce type de protéines, nous distinguons le SDS et le SDC[137, 138]. Néanmoins, ces détergents ne sont pas directement compatibles avec une analyse LC-MS/MS. Le SDC est utilisable en solution car il est compatible à une concentration plus élevée que le SDS avec une digestion à la trypsine, et peut être facilement retiré par précipitation avant analyse. Le SDS, doit quant à lui être retiré avant digestion trypsique et nécessite de ce fait souvent de passer par une étape d'électrophorèse 1D SDS-PAGE. Ainsi, trois préparations d'échantillons mettant en jeu ces deux détergents ont été testées. Il s'agit :

- D'un **gel d'électrophorèse 1D SDS-PAGE avec une migration sur 6 bandes** afin de réduire le fractionnement.
- D'un **gel 1D SDS-PAGE « *Stacking* » (SG) sans fractionnement, découpé en deux bandes** : la première bande contient les protéines concentrées (bande d'intérêt), et la seconde (traînées) résulte des traînées effectuées par l'échantillon lors de sa migration et se situe au-dessus de la bande d'intérêt.
- D'une **digestion liquide (DL) au SDC**, sans fractionnement, qui a l'avantage de s'affranchir des étapes de migration, fixation, coloration, et de découpe, liées au gel 1D SDS-PAGE.

La comparaison de ces trois protocoles a été effectuée sur deux échantillons de « *ghosts* » *membranaires* provenant de deux lignées cellulaires différentes de CSC issues de biopsies de glioblastomes (TG01 et TG16). Chaque préparation a fait l'objet de deux analyses par nanoLC-MS/MS sur un couplage entre une chromatographie liquide de type Nano-Acquity (WATERS) et un

spectromètre de masse de type Q-TOF (Impact-HD, BRUKER). Le schéma analytique utilisé est représenté en Figure I-2.

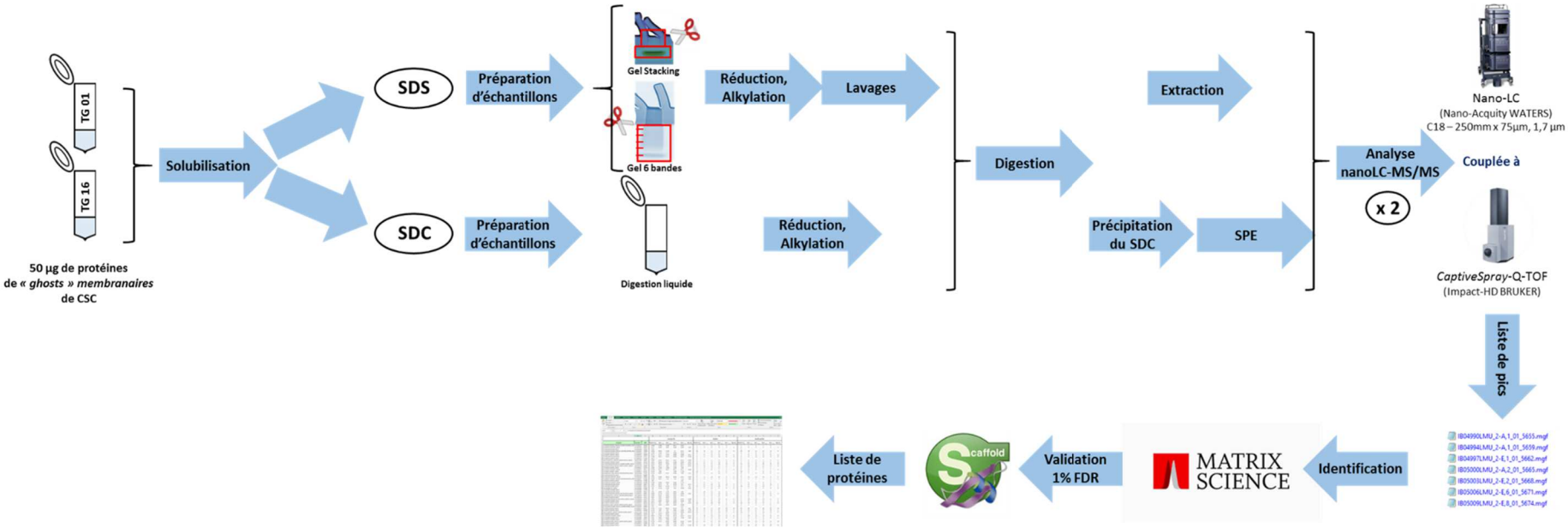


Figure I-2 - Schéma analytique employé pour la comparaison des trois protocoles de préparation permettant de réduire le fractionnement des « *ghosts* » *membranaires* de CSC issues de glioblastomes

Les gradients chromatographiques ont été adaptés à la complexité de chaque préparation d'échantillons. La durée d'analyse résultante pour un échantillon est donnée pour chaque préparation d'échantillons dans le Tableau I-2.

	Etude préliminaire	**Gel 1D SDS-PAGE 6 bandes**	**Gel 1D SDS-PAGE « *Stacking* »**	**Digestion Liquide SDC**
Durée d'analyse totale	1 semaine	9h *(6 bandes x 90min)*	6h *(2 bandes x 3h)*	3h

Tableau I-2 - Résumé du temps d'analyse requis pour un échantillon pour chaque protocole de préparation testé ainsi que pour l'étude préliminaire

Le nombre total moyen de protéines identifiées sur les deux réplicats d'injection après validation avec un FDR inférieur à 1 % pour chaque protocole et chaque échantillon est donné dans la Figure I- 3. Les annotations « *Gene Ontology* » (GO) ont été utilisées pour distinguer les protéines liées à la membrane plasmique (annotées « *plasma membrane* »). Notons que pour l'analyse du SG de l'échantillon TG16, une suppression de signal est survenue au moment de l'analyse, probablement du fait d'un nombre de lavages du gel insuffisants, permettant le retrait du SDS. Ainsi, ces valeurs n'ont pas été prises en compte lors des comparaisons de protocoles permettant la sélection de la préparation d'échantillons à adopter. Dans l'étude de recherche de biomarqueurs de glioblastomes, les collaborateurs ont été particulièrement intéressés par les protéines membranaires « Clusters de Différenciation » (CD) pour lesquelles des anticorps sont disponibles, facilitant ainsi leur ciblage par des thérapies et leur validation par des méthodes de cytométrie en flux ou ELISA par exemple. Ainsi, le nombre de CD identifiés par chaque méthode est donné par la Figure I-4.

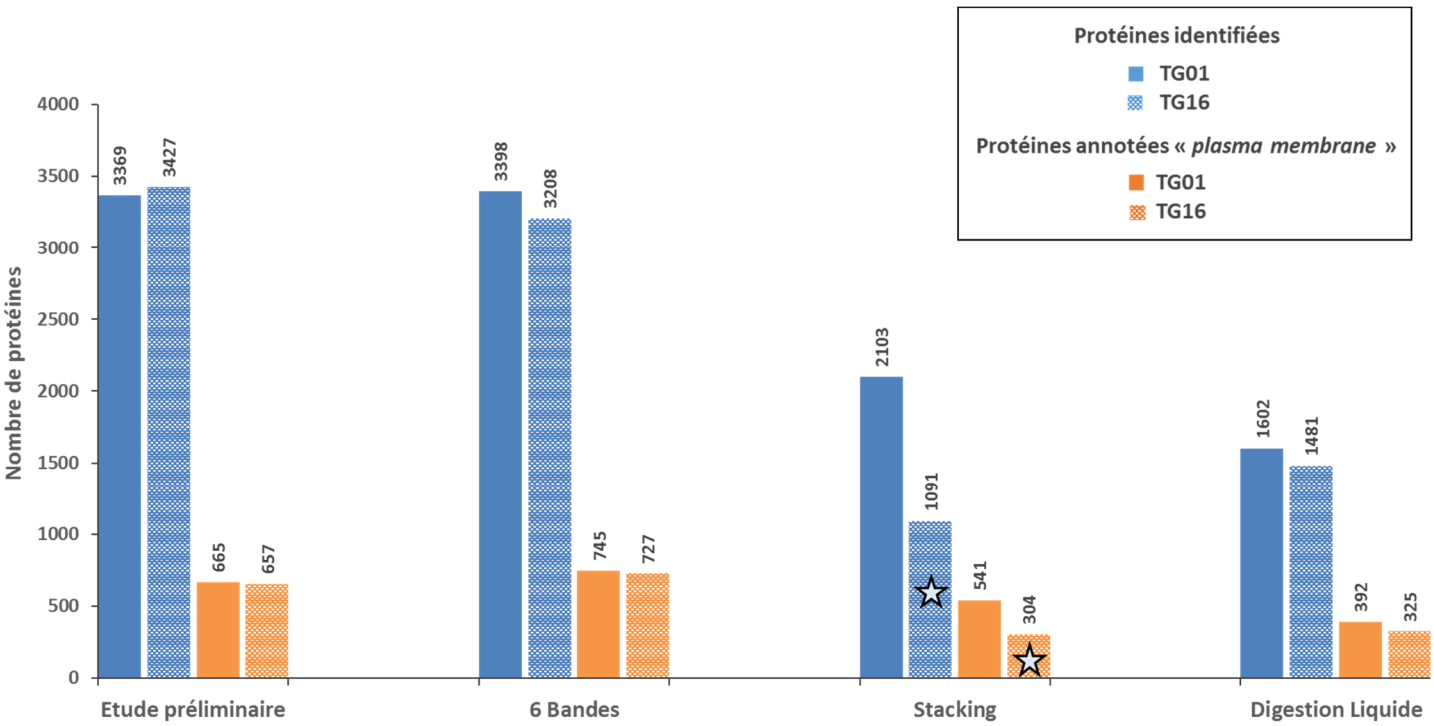


Figure I-3 - Nombre moyen de protéines identifiées et de protéines annotées « *plasma membrane* » sur les deux réplicats pour l'étude préliminaire et les trois protocoles testés pour l'échantillon TG01 et TG16.
Les étoiles indiquent les échantillons pour lesquels une suppression de signal est survenue lors des analyses

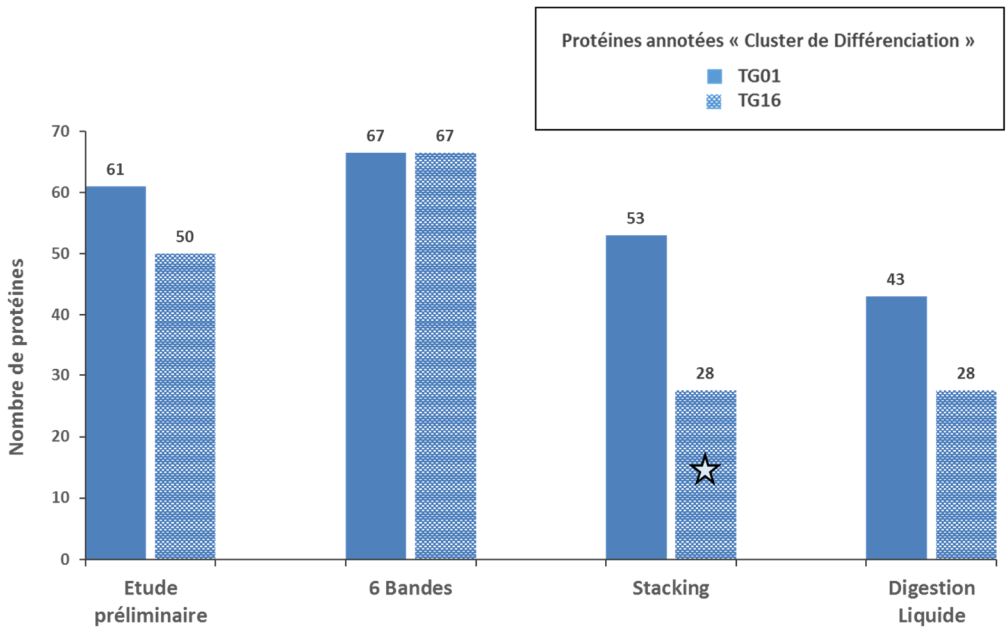


Figure I-4 - Nombre moyen de protéines CD identifiées sur les deux réplicats pour chaque protocole, pour l'échantillon TG01 et TG16.
L'étoile indique les échantillons pour lesquels une suppression de signal est survenue lors des analyses

Globalement, le nombre de protéines identifiées est très variable d'un protocole à l'autre, avec des valeurs s'étalant de 1481 à 3398 pour les trois protocoles testés (en ne prenant pas en compte l'échantillon pour lequel une suppression de signal est survenue au moment des analyses). En ce qui concerne les protéines membranaires, ces protocoles de préparation de « *ghosts* » *membranaires* ont permis d'identifier des proportions de protéines annotées « *plasma membrane* » en accord avec un enrichissement membranaire, soit 24, 26 et 22 % pour la DL, le SG et le gel 6 bandes respectivement pour l'échantillon TG01. Enfin, le nombre de protéines CD identifiées est moins variable d'un protocole à l'autre avec des valeurs allant de 43 à 67 pour l'échantillon TG01.

Par rapport à l'étude préliminaire, le gel 6 bandes permet de réduire la durée d'analyse, tout en offrant la même profondeur d'analyse avec davantage de protéines annotées « *plasma membrane* ». La DL avec utilisation du SDC s'avère ici être la préparation d'échantillons la moins adaptée à ce type d'échantillon, car elle fournit les résultats les plus faibles en termes d'identification de protéines de tout type. Le SG, permet quant à lui d'identifier en peu de temps un nombre satisfaisant de protéines, notamment de CD. Notons que pour le SG, l'intégralité des CD n'est pas uniquement identifiée dans la bande d'intérêt, mais aussi dans la bande des traînées comme le montre le diagramme de Venn pour la première analyse de l'échantillon TG01 (Figure I-5).

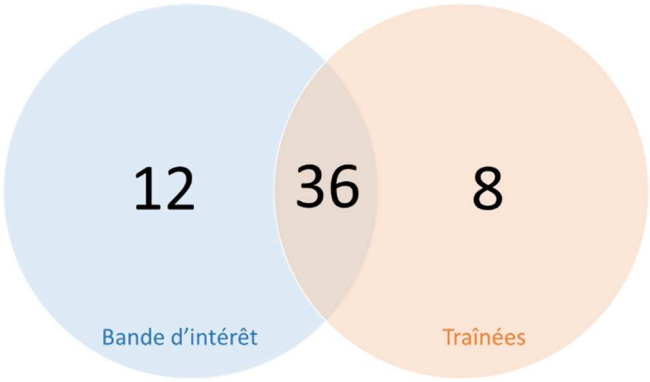


Figure I-5 - Diagramme de Venn représentant la répartition des protéines CD entre la bande d'intérêt et les traînées pour le gel « *Stacking* » TG01, analyse n°1

Par l'absence de fractionnement, le SG offre un gain de temps par rapport au gel 6 bandes, que ce soit en termes de manipulation ou en termes de durée d'analyse, tout en fournissant des résultats d'identification satisfaisants, particulièrement en termes de CD (80 à 85 % des CD identifiés lors de l'étude préliminaire). C'est pourquoi ce protocole a été sélectionné pour être implémenté dans le schéma analytique de la recherche de biomarqueurs de glioblastomes avec une quantification de type XIC, d'autant que le rassemblement des deux bandes avant analyse permet de réduire encore la durée d'analyse, sans perte d'informations, en plus de faciliter la quantification XIC.

b. Optimisation du gradient chromatographique pour l'analyse d'un « *ghost* » *membranaire* préparé en gel « *Stacking* »

Le SG étant le meilleur compromis pour l'analyse de « *ghosts* » *membranaires* de CSC de glioblastomes, une optimisation du gradient chromatographique a été réalisée sur un réplicat d'analyse de TG01 afin de trouver le meilleur compromis entre nombre de protéines identifiées et durée d'analyse. Ainsi, trois gradients ont été testés : 120, 180 et 240 minutes.

Afin d'évaluer chacune de ces conditions, les résultats provenant de la recherche dans la banque de données ont fait l'objet d'une analyse par un outil développé au laboratoire et nommé « MS Diag », qui permet d'avoir un aperçu visuel du nombre de spectres assignés ou non à une identification, ainsi que le score associé, et ce, en fonction du temps. Les résultats moyens des deux réplicats d'injection de cette analyse « MS Diag » ainsi que d'identifications validées avec un FDR inférieur à 1 % sont donnés par le Tableau I-3.

Gradient	**Nombre de protéines** (FDR < 1 %)	**Nombre de peptides avec score > 20** *(non redondants)*	**Spectres MS/MS dont score > 20**
120 min	1974	10699	37 %
180 min	2061	11434	37 %
240 min	2085	11888	38 %

Tableau I-3- Valeurs moyennes des deux réplicats d'analyse du nombre de protéines identifiées et validées avec un FDR <1 %, ainsi que du nombre de peptides et de spectres avec un score Mascot correct (> 20)

Le gradient optimal est celui qui permet :

- **D'identifier un maximum de protéines** dans un temps d'analyse raisonnable,
- Avec le **maximum de peptides présentant un score Mascot supérieur à 20**, qui reflète une concordance entre le spectre théorique et expérimental acceptable,
- Et générant un **maximum de spectres MS/MS assignés à une identification avec un score supérieur à 20**.

Ainsi, le gradient chromatographique qui semble optimal pour l'analyse d'un SG de « *ghosts* » *membranaires* de CSC de glioblastomes est le gradient de 180 minutes qui permet d'identifier davantage de peptides et de protéines que le gradient de 120 minutes, et un nombre quasiment équivalent au gradient de 240 minutes. C'est donc celui-ci qui sera utilisé pour les prochaines analyses

effectuées dans le cadre du projet de recherche de biomarqueurs de CSC de glioblastomes. Ce gradient de 180 minutes est détaillé dans la *Partie expérimentale-II*.

c. Répétabilité du schéma analytique avec la préparation d'échantillons « *Stacking* »

La répétabilité du schéma analytique impliquant une préparation d'échantillons SG et une quantification XIC a été évaluée en déposant trois fois l'échantillon TG01 au sein d'un même gel, ainsi que sur trois gels différents permettant ainsi d'établir plusieurs niveaux de répétabilité : la répétabilité inter-gels (Gel 1_1, Gel 2_1 et Gel 3_1) et intra-gel (Gel 1_1, Gel 1_2 et Gel 1_3) (Figure I-6). Les analyses pour chaque échantillon ont été effectuées trois fois.

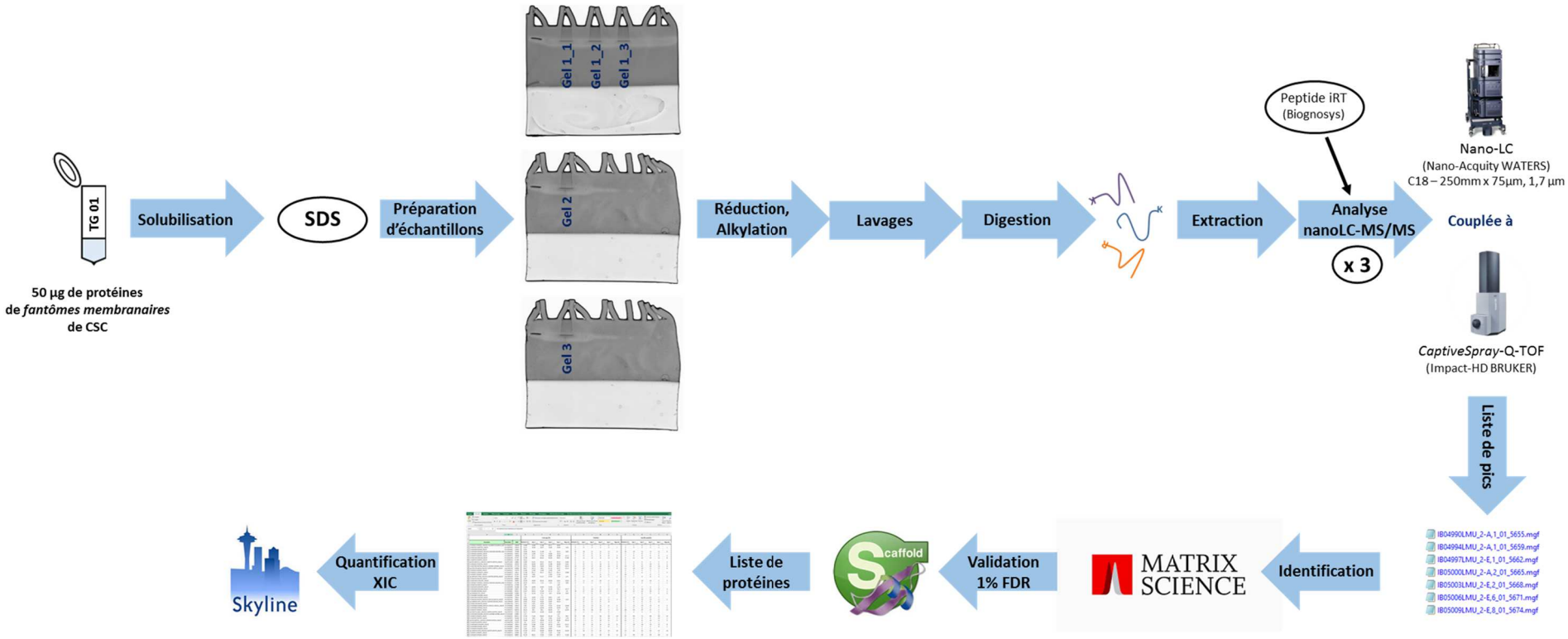


Figure I-6 - Schéma analytique impliquant la préparation d'échantillons gel « *Stacking* » des « *ghosts* » *membranaires* de CSC de glioblastomes et la quantification XIC pour l'étude de sa répétabilité

Dans un premier temps, la répétabilité a pu être évaluée au niveau qualitatif en comparant le nombre de protéines identifiées pour chaque réplicat technique de TG01 (Figure I-7). Ainsi, nous pouvons observer que le nombre de protéines, et notamment de protéines membranaires et CD est équivalent et stable sur l'ensemble des réplicats techniques.

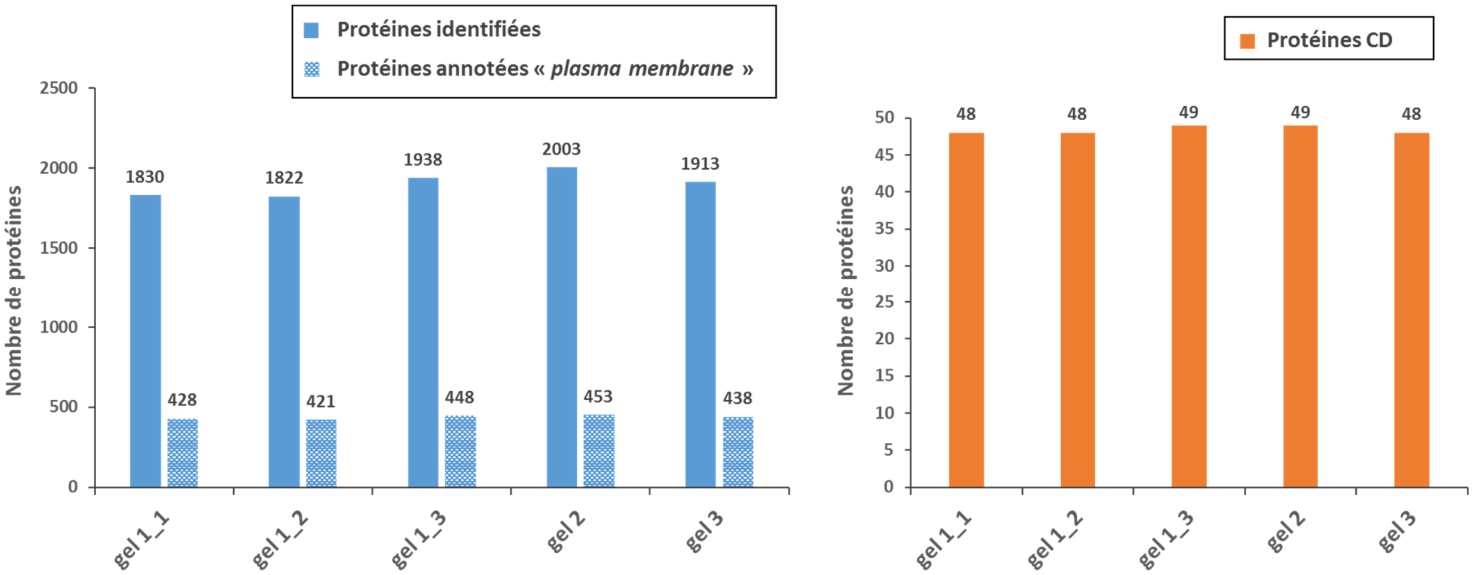


Figure I-7 - Nombre moyen de protéines identifiées sur les trois réplicats d'injection, ainsi que de protéines annotées « *plasma membrane* » et de protéines CD pour chaque réplicat technique de TG01

La répétabilité a également été évaluée en calculant les coefficients de variation (CV) inter et intra-gels, ainsi qu'entre les trois réplicats d'injection à partir des valeurs de quantification obtenues pour chaque peptide (somme des aires des trois premiers isotopes : P, P+1 et P+2). La répartition des CV calculés est représentée par les boîtes à moustache (Figure I-8). La boîte à moustache est une représentation graphique des données permettant de comparer dans notre cas la répartition des CV de plusieurs séries. L'amplitude des boîtes permet d'apprécier visuellement la dispersion des valeurs puisqu'elles contiennent 50 % de celles-ci. Les CV médians calculés, précisés dans les encadrés bleus, sont globalement inférieurs à 15 %, avec ceux calculés à partir des trois réplicats d'injection et intra-gel inférieurs à 10 % et inter-gels inférieur ou égal à 13 %. S'agissant d'une méthode de quantification relative employée dans une approche de découverte non ciblée, ces valeurs sont satisfaisantes et ce schéma analytique peut être considéré comme répétable. De plus, ces résultats sont cohérents, car chaque introduction d'une variabilité augmente les CV : CV médian triplicats d'injection < intra-gel < inter-gels.

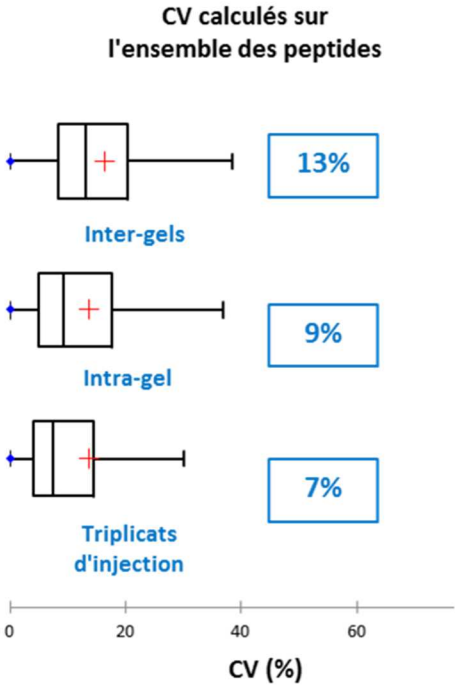


Figure I-8 - Boîtes à moustache représentant la répartition des CV inter-gels, intra-gel et entre les trois réplicats d'injection.
Les valeurs encadrées correspondent aux valeurs médianes

D'autre part, la stabilité du système chromatographique a été évaluée en calculant les CV des Tr à partir des données issues de Skyline pour les protéines CD, pour lesquelles une vérification manuelle a été effectuée. Le CV médian résultant est de 0,2 %, reflétant une très bonne stabilité chromatographique tout au long des analyses. De plus, la stabilité de l'ensemble du couplage au cours de cette étude de répétabilité a pu être évaluée à l'aide des peptides iRT (« *index Retention Time* »), qui ont été dopés en quantités égales dans chaque échantillon avant analyse nanoLC-MS/MS. Ces onze peptides ont été conçus de manière à être répartis sur l'ensemble du gradient chromatographique, et de manière à ce que leur séquence ne soit pas contenue dans la banque humaine. En ôtant le peptide le plus hydrophile et le plus hydrophobe du calcul du CV, car présentant une résolution médiocre du fait d'un signal réponse MS de mauvaise qualité, les médianes résultantes inférieures à 20 % reflètent qu'aucune dérive majeure de l'instrument n'est survenue au cours des analyses.

d. Apport de la vérification manuelle de l'intégration des pics par Skyline

Comme évoqué au *Chapitre I-III-2-b*, le logiciel Skyline commet souvent des erreurs d'intégration. Afin d'illustrer ce problème, qui peut être résolu par une vérification, voire une réintégration manuelle, le Tableau I-4 reporte les valeurs des CV médians inter et intra-gels calculés avant et après réintégration manuelle des peptides provenant des protéines CD (soit 640 peptides). Ces valeurs reflètent bien le

fait que la réintégration manuelle sur Skyline permet d'améliorer l'exactitude des données de quantification XIC, comme suggéré par Ramus *et collaborateurs*[108].

	Avant réintégration manuelle	**Après réintégration manuelle**
CV inter-gels *(médiane)*	12 %	10 %
CV intra-gel *(médiane)*	8 %	7 %

Tableau I-4 - CV médians inter et intra-gels obtenus avant et après réintégration manuelle des pics correspondants aux 640 peptides issus des protéines CD sur Skyline

Par ailleurs, des tests ont été effectués pour améliorer les résultats en appliquant des filtres automatiques, tels que : ôter les peptides les plus hydrophiles (Tr > 35min), ou ceux ayant été modifiés car ils présentent souvent des pics chromatographiques peu reproductibles, ou encore ceux ayant un « *isotope dot product* » (voir *Chapitre I-III-2-b*) strictement inférieur à 0,9. Cependant, l'application de ces filtres ne permet pas d'améliorer de manière significative les CV, d'autant qu'ils induisent la perte de beaucoup de peptides, soit d'informations. De ce fait, le meilleur moyen d'obtenir des résultats les plus exactes possible reste la vérification manuelle de l'intégration.

3- Conclusion

Ces différentes optimisations ont pu mettre en avant que le SG, qui emploie le SDS comme détergent, permet d'obtenir un nombre satisfaisant de protéines identifiées, notamment de protéines membranaires et plus particulièrement de protéines CD en un temps d'analyse raisonnable, ce qui permet d'envisager l'analyse d'un grand nombre d'échantillons au cours d'une même séquence d'analyses. Son utilisation pour l'étude de « *ghosts* » *membranaires* de CSC de glioblastomes avec une quantification XIC fournit des résultats répétables entraînant ainsi l'implémentation de ce schéma analytique dans le projet de recherche de biomarqueurs de glioblastomes.

II- Evaluation et optimisation d'une préparation d'échantillons en gel sans fractionnement : le « *Tube-Gel* »

Lorsque l'utilisation d'un détergent comme le SDS est requise lors d'une étude, il est souvent d'usage d'employer un gel 1D SDS-PAGE pour permettre l'élimination du détergent par des lavages avant digestion enzymatique et analyse LC-MS/MS du fait de son incompatibilité avec ces étapes. Dans le cas d'une étude impliquant une approche de quantification sans marquage, nécessitant une bonne répétabilité, l'utilisation d'un gel « *Stacking* » sans fractionnement est souvent préférée. Cependant, le SG peut s'avérer laborieux lorsqu'un grand nombre d'échantillons est à analyser, du fait qu'il implique :

1) Une étape de réduction des ponts disulfures avant dépôt sur gel,
2) De préparer et de faire migrer parallèlement plusieurs gels en même temps, puisqu'un gel permet de charger au maximum 10 échantillons,
3) De souvent déposer l'intégralité des échantillons sur les gels, en évitant le transfert d'un puits à un autre, de manière à ce que les gels soient distinguables les uns des autres afin de retrouver quelle bande correspond à quel échantillon,
4) Une étape de migration des protéines,
5) Une étape de fixation des protéines dans le gel,
6) Une étape de coloration pour révéler la présence des protéines,
7) De découpe du gel,
8) La réduction des ponts disulfures et l'alkylation des cystéines,
9) Des étapes de lavage du gel afin de retirer le SDS.

La préparation d'échantillons « *Tube-Gel* » (TG), qui consiste à faire polymériser un gel de polyacrylamide directement dans l'échantillon en solution, et ainsi à piéger les protéines dans une matrice de gel, permet de s'affranchir des étapes 1 à 4 ainsi que de l'étape 6. Elle semble de ce fait être une bonne alternative au SG lorsque l'utilisation du SDS et l'absence de fractionnement sont désirées, ainsi qu'une rapidité de mise en œuvre, de manière à augmenter le nombre de conditions à

analyser. Le TG a été décrit en 2005 pour l'analyse de protéines membranaires, car il supporte les hautes concentrations de détergents, comme le SDS[139].

Les travaux effectués au cours de cette thèse sur la préparation d'échantillons « *Tube-gel* » ont été menés en binôme avec le Dr Luc FORNECKER.

1- Evaluation d'une préparation d'échantillons innovante pour la protéomique quantitative sans marquage XIC : le « *Tube-Gel* »

Le TG a depuis sa première description été appliqué à l'étude de protéomes membranaires majoritairement[140-142] et a fait l'objet de quelques modifications[141, 142]. Parmi ces études, une a fait l'objet d'une quantification par marquage iTRAQ[141], et une seconde d'une quantification sans marquage par comptage de spectres MS/MS[143]. Bien que le TG ait été comparé à une préparation d'échantillons FASP et à une DL d'un point de vue qualitatif (notamment au niveau des caractéristiques physico-chimiques des protéines et du nombre de coupures manquées)[144], les performances et la répétabilité du TG pour la quantification XIC n'ont cependant jamais été évaluées. A cette fin, mais aussi pour savoir si le TG peut être applicable aux nombreux projets de quantification XIC menés au laboratoire, une analyse comparative du TG, du SG et de la DL a été effectuée. La conception d'un échantillon parfaitement calibré a permis de comparer leurs performances qualitatives mais aussi quantitatives. Cet échantillon, constitué d'un fond complexe et constant de protéines de levure dopé par des concentrations croissantes d'un mélange équimolaire de 48 protéines humaines (UPS1, Sigma), permet de mimer des variations au sein d'un protéome complexe lorsque les échantillons de différentes concentrations d'UPS1 sont comparés, de manière à se placer dans des conditions proches d'une étude différentielle de recherche de biomarqueurs. Cette étude est résumée schématiquement en Figure 1 du papier ci-après.

Les résultats de cette étude ont globalement montré que les performances du TG en termes de protéines identifiées sont supérieures à celles du SG et de la DL. Il est à noter cependant que la préparation d'échantillons TG génère davantage de modifications propionamides, du fait du contact direct des protéines avec des monomères d'acrylamide, et davantage d'oxydations des méthionines du fait des conditions oxydantes conséquentes d'une polymérisation chimique du TG. En termes de résultats quantitatifs, le TG est au moins aussi répétable que le SG, voire plus répétable que la DL, et permet de détecter des protéines variantes au sein d'un mélange complexe avec une haute sensibilité et peu de faux-positifs (voir Figure 5 du papier). Ainsi, le TG peut être employé pour des études de

protéomique quantitative sans marquage, notamment lorsqu'un grand nombre d'échantillons est à analyser.

L'ensemble de ces résultats a fait l'objet d'une publication en 2016 dans le journal *Proteomics* et se trouve ci-après.

RESEARCH ARTICLE

Benchmarking sample preparation/digestion protocols reveals tube-gel being a fast and repeatable method for quantitative proteomics

Leslie Muller[*], *Luc Fornecker*[*], *Alain Van Dorsselaer, Sarah Cianférani and Christine Carapito*

Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC UMR 7178, Strasbourg, France

Sample preparation, typically by *in-solution* or *in-gel* approaches, has a strong influence on the accuracy and robustness of quantitative proteomics workflows. The major benefit of *in-gel* procedures is their compatibility with detergents (such as SDS) for protein solubilization. However, SDS-PAGE is a time-consuming approach. Tube-gel (TG) preparation circumvents this drawback as it involves directly trapping the sample in a polyacrylamide gel matrix without electrophoresis. We report here the first global label-free quantitative comparison between TG, stacking gel (SG), and basic liquid digestion (LD). A series of UPS1 standard mixtures (at 0.5, 1, 2.5, 5, 10, and 25 fmol) were spiked in a complex yeast lysate background. TG preparation allowed more yeast proteins to be identified than did the SG and LD approaches, with mean numbers of 1979, 1788, and 1323 proteins identified, respectively. Furthermore, the TG method proved equivalent to SG and superior to LD in terms of the repeatability of the subsequent experiments, with mean CV for yeast protein label-free quantifications of 7, 9, and 10%. Finally, known variant UPS1 proteins were successfully detected in the TG-prepared sample within a complex background with high sensitivity. All the data from this study are accessible on ProteomeXchange (PXD003841).

Received: July 6, 2016
Revised: September 23, 2016
Accepted: October 12, 2016

Keywords:
In-solution digestion / Label-free quantitative proteomics / Sample preparation / Stacking gel / Technology / Tube-gel

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

Recent increases in MS resolution, mass accuracy, and acquisition speed have enabled the use of global label-free methods based on extracted ion chromatograms (XICs) for the relative quantification of complex protein extracts. The stability of the LC-MS system and the repeatability and robustness of the data are major considerations in this type of approach, with highly repeatable sample handling being particularly important. Samples for bottom-up label-free shotgun experiments are most commonly prepared by *in-solution* or *in-gel* methods, which have their respective advantages and drawbacks, but both rely on efficient protein solubilization and extraction. This is usually achieved with chaotropes or strong detergents such as SDS [1]. However, since these agents are typically incompatible at high concentration with further enzymatic digestion and can interfere in subsequent RP-LC and MS analysis, they need to be removed [2]. *In-gel* approaches are advantageous in this context as the high-concentration SDS

Correspondence: Dr. Christine Carapito, Laboratoire de Spectrométrie de Masse BioOrganique, Université de Strasbourg, CNRS, IPHC UMR 7178, 25 rue Becquerel, 67087 Strasbourg, France
E-mail: ccarapito@unistra.fr
Fax: +33 3 68 85 27 81

Abbreviations: **FA**, formic acid; **FDP**, false discovery proportion; **FP**, false positive; **LD**, liquid digestion; **SG**, stacking gel; **TG**, tube-gel; **TP**, true positive; **XICs**, extracted ion chromatograms

[*]These authors equally contributed to this work.
Colour Online: See the article online to view Figs. 1–5 in colour.

Significance of the study

Sample preparation is a key step for the repeatability and robustness of quantitative proteomics workflows. The two most common methods used are *in-gel* or *in-solution* approaches. In our study, we evaluated and demonstrated the benefits of an alternative gel-based sample preparation approach, named tube-gel (TG), based on the direct trapping of the sample in a polyacrylamide gel without electrophoresis.

The aim of our study was to compare the TG sample preparation method with a gel-based approach (stacking gel [SG]) and a classical *in-solution* digestion procedure. We have therefore used a reference sample consisting of 48 standard human proteins spiked in a complex yeast cell lysate background. This calibrated sample, with a series of six UPS1 concentration spikes, was allowed to rigorously compare these three sample preparation methods for both qualitative protein identification and label-free quantification. The results obtained for protein identification and quantification with TG sample preparation were better than those achieved with *in-solution* digestion, and equivalent or better than those obtained with SG samples. This study allows us to propose TG as a fast, effective, and repeatable sample preparation procedure, in particular if a large number of samples have to be analyzed.

employed can be removed by multiple and intensive washing steps [3]. This, alongside the fractionating of complex protein mixtures it allows, explains the wide use of acrylamide gel-based sample preparation in proteomics. However, extensive gel fractionation is seldom compatible with further quantification. The stacking gel (SG) approach is thereby an appropriate gel-based alternative: the proteins are concentrated in one band, allowing accurate quantification (without fractionation) and normalization of the data during computational and statistical analysis for shotgun label-free quantitative proteomics [4]. Gel-based sample preparation is, nonetheless, a multistep, time-consuming technique. The gels must first be poured and the samples have to be mixed with a loading buffer before the gel can be run until the proteins are stacked. After electrophoresis, the proteins are fixed and the gel has to be stained to reveal the stacking slice, which is then diced into small pieces, and subjected to reduction and alkylation before *in-gel* digestion. This technique is unwieldy and difficult to implement when a large number of samples have to be processed. Furthermore, the risk of contamination and sample loss accrues at each step of the preparation processes.

The tube-gel (TG) protocol was introduced in 2005 [5] for its compatibility with a wide variety of detergents. It has facilitated the characterization of membrane proteins [6–10] and lipid rafts proteins [11], and the charge derivatization of peptides [12]. In TG preparation, the sample is directly trapped in a polyacrylamide-gel matrix without electrophoresis. The resulting gel is then diced and subjected to reduction, alkylation, and *in-gel* digestion. For quantitative proteomics, TG has only been used for iTRAQ labeling [6] or label-free spectral counting [11]. However, some concerns remain about the amenability of TG-prepared samples to the accurate quantification of differentially expressed proteins and to protein identification in complex biological samples [8,13]. In this context, we have conducted the first direct and extensive comparison of the TG, SG, and basic liquid digestion (LD) protocols on a reference sample consisting of 48 standard human proteins spiked in a complex yeast cell lysate background.

We first evaluated the qualitative repeatability of the different sample preparation methods in terms of the numbers and overlaps of identified proteins. Then, we used the three methods for further label-free XIC quantification and investigated whether known variant proteins could be detected in complex protein mixtures prepared in these ways.

2 Materials and methods

2.1 Sample preparation

A yeast cell lysate was prepared in an 8 M urea/0.1 M ammonium bicarbonate buffer (NH_4HCO_3). Having determined the protein concentration (RC-DC™; Bio-Rad), the lysate was adjusted to 1 μg/μL and used to resuspend a UPS1 standard mixture (Sigma) and perform a serial dilution thereof. Each resulting sample, with six different spiked amounts of UPS1 (final amounts of 0.5, 1, 2.5, 5, 10, and 25 fmol UPS1 for 800 ng of yeast lysate), was aliquoted to allow the preparation of each sample in technical triplicate as follows: 3 × 20 μL in a 2 mL Eppendorf tube for LD, 3 × 20 μL in a 500 μL Eppendorf tube for TG, and 60 μL in a 1.5 mL Eppendorf tube for SG.

2.2 Liquid digestion

Each aliquot was reduced with 12 mM DTT for 30 min at 37°C with agitation and alkylated with 40 mM iodoacetamide for 1 h in the dark. The urea concentration was lowered to 1 M by dilution with fresh 0.1 M NH_4HCO_3 and the proteins were cleaved by adding modified porcine trypsin (Promega) at an enzyme:protein ratio of 1:80 w/w overnight at 37°C. Enzymatic digestion was stopped by adding 10 μL of formic acid (FA) and a C18 SPE (Sep-Pak®C18 50 mg; Waters, Milford, MA) was performed: two equilibrations with 500 μL methanol, two with 500 μL ACN, and three with 0.1% FA.

After sample loading, the phase was washed three times with 750 μL of 0.1% FA and the peptides were eluted with 200 μL of 60% ACN in 0.1% FA. The collected extract was vacuum dried and resolubilized with 50 μL of H_2O/ACN/FA (98/2/0.1, v/v/v).

2.3 Stacking gels

After denaturation at 100°C for 5 min in loading buffer (5% SDS, 5% β-mercaptoethanol, 1 mM EDTA, 10% glycerol, 10 mM Tris pH 6.8), each aliquot was split in three parts and the proteins were concentrated in one band with a 4% SDS-PAGE gel (three different gels prepared 1 day in advance). The gels were fixed with 50% ethanol/3% phosphoric acid and stained with colloidal Silver Blue. Each band was excised and cut in four pieces prior to *in-gel* digestion. The gel pieces were washed four times with 100 μL of 75% ACN and 25% NH_4HCO_3 at 25 mM and dehydrated with 100 μL of ACN. The cysteine residues were reduced by adding 10 mM DTT for 30 min at 60°C and 30 min at room temperature, and alkylated by adding 55 mM iodoacetamide for 20 min in the dark. The bands were then washed three times by adding 50 μL of 25 mM NH_4HCO_3 and 50 μL of ACN. After two dehydrations with 50 μL of ACN, the proteins were cleaved in an adequate volume to cover all the gel pieces with a modified porcine trypsin (Promega) solution at a 1:80 w/w enzyme–protein ratio. Digestion was performed overnight at 37°C. Tryptic peptides were extracted twice under agitation, first with 40 μL of 60% ACN in 0.1% FA for 1 h and then with 40 μL of 100% ACN for 1 h. The collected extracts were pooled, the excess ACN was vacuum dried, and the samples were resolubilized with 50 μL of H_2O/ACN/FA (98/2/0.1, v/v/v).

2.4 Tube-gels

In each aliquot, 2% SDS, 32% water, 7.5% acrylamide/bis-acrylamide, and 0.25% TEMED were added for a final volume of 100 μL. The solutions were then vortexed and centrifuged until all bubbles had disappeared. Ammonium persulfate (0.25%) was added to initiate polymerization and the tube was briefly vortexed and rapidly centrifuged to remove residual bubbles as these can interfere with the polymerization process. After fixation with 50% ethanol/3% phosphoric acid, TGs were cut in 2 mm high sections and each section in ~2 mm^2 pieces. The gel pieces were washed, the cysteine residues were reduced and alkylated, and the protein digestion and extraction were conducted, as described above for the SGs, in a 2-mL Eppendorf tube (with the exception that all volumes used were four times larger). After vacuum drying, samples were resolubilized with 50 μL of H_2O/ACN/FA (98/2/0.1, v/v/v).

2.5 Nano-LC-MS/MS analysis

The nano-LC-MS/MS analysis was performed on a nanoAcquity UPLC device (Waters) coupled to a Q-Exactive Plus mass spectrometer (Thermo Scientific, Bremen, Germany). Peptide separation was performed on an ACQUITY UPLC BEH130 C18 column (250 mm × 75 μm with 1.7 μm diameter particles) and a Symmetry C18 precolumn (20 mm × 180 μm with 5 μm diameter particles; Waters). The solvent system consisted of 0.1% FA in water (solvent A) and 0.1% FA in ACN (solvent B). The samples (2 μL) were loaded into the enrichment column over 3 min at 5 μL/min with 99% of solvent A and 1% of solvent B. The peptides were eluted at 450 μL/min with the following gradient of solvent B: from 1 to 8% over 2 min, 8 to 35% over 77 min, and 35 to 90% over 1 min.

The 54 samples (18 samples per preparation protocol) were injected in a random order. The MS capillary voltage was set to 1.8 kV at 250°C. The system was operated in a data-dependent acquisition mode with automatic switching between MS (mass range 300–1800 m/z with $R = 70\,000$, automatic gain control fixed at 3×10^6 ions, and a maximum injection time set at 50 ms) and MS/MS (mass range 200–2000 m/z with $R = 17\,500$, automatic gain control fixed at 1×10^5, and the maximal injection time set to 100 ms) modes. The ten most abundant peptides (intensity threshold 2×10^5) were selected on each MS spectrum for further isolation and higher energy collision dissociation fragmentation, excluding unassigned and monocharged ions. The dynamic exclusion time was set to 60 s.

2.6 Data analysis

The raw data obtained for each sample preparation protocol were processed separately using MaxQuant (version 1.5.3.30) [14]. Peaks were assigned with the Andromeda search engine with full trypsin specificity. The database used for the searches was concatenated in house with the *Saccharomyces cerevisiae* entries extracted from the UniProtKB-SwissProt database (16 April 2015, 7806 entries) [15] and those of the UPS1 proteins (48 entries). The minimum peptide length required was seven amino acids and a maximum of one missed cleavage was allowed. The precursor mass tolerance was set to 20 ppm for the first search and 4.5 ppm for the main search. The fragment ion mass tolerance was set to 20 ppm. Methionine oxidation was always set as a variable modification and peptides with modified methionines, as well as their unmodified counterparts, were excluded from protein quantification. Cysteine carbamidomethylation was set as a fixed modification for LD, as commonly applied. For TG and SG, carbamidomethylation was set as a variable modification to account for the potential propionamide modifications of cysteine residues. Cysteine propionamidation was thus also set as a variable modification for the TG and SG samples. For

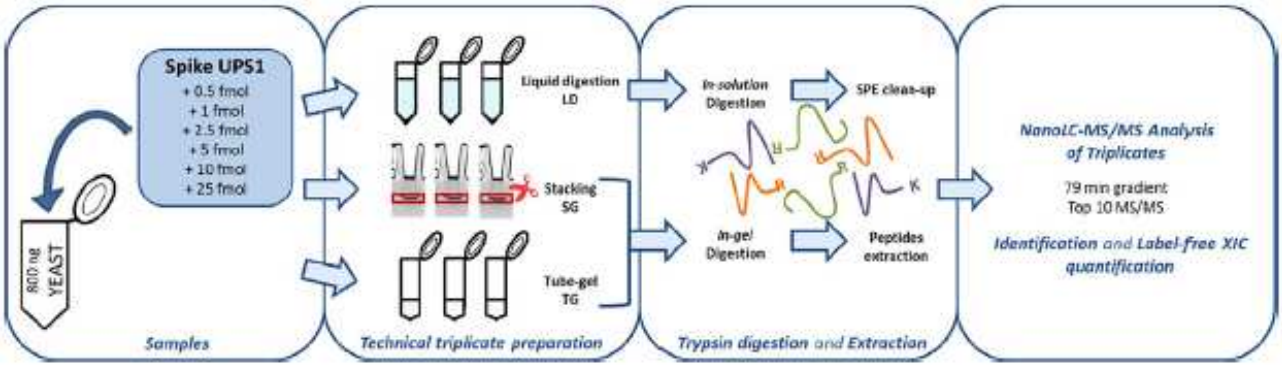


Figure 1. Experimental workflow used to compare the TG, SG, and LD sample preparation protocols in terms of their repeatability and ability to reveal known variant proteins in subsequent experiments. The samples were first prepared by spiking six known amounts of UPS1 (an equimolar mixture of 48 human proteins) in a yeast background. Three technical replicates were prepared with each protocol, digested, and extracted. Protein identification and label-free XIC quantification was performed with the software MaxQuant after nano-LC-MS/MS analysis.

protein quantification, the "match between runs" option was enabled. The maximum false discovery rate was 1% at peptide and protein levels with the use of a decoy strategy. The final protein lists were obtained after suppression of contaminants, reverse entries, and proteins only identified with modified peptides. We used the "proteingroups.txt" files with LFQ intensities (normalized intensities) for proteins quantification and "modificationspecificpeptides.txt" files. We considered only protein groups identified with at least one unique peptide. GO annotations were extracted for identified proteins using an in-house developed software suite (Mass Spectrometry Data Analysis, https://msda.unistra.fr/) [15]. Venn diagrams were drawn with eulerAPE (version 2.0.3) [16]. All data were deposited in the ProteomeXchange repository and are accessible through the ID PXD003841 [17].

2.7 Statistical data analysis

The Perseus package (v. 1.5.2.6; http://141.61.102.17/maxquant_doku) was used to perform statistical analysis on the pairwise comparisons of different UPS1 spiking amounts obtained for each sample protocol. We submitted the $\log_2$ transformed normalized intensities to Welch's *t*-tests. This analysis was only applied to the data from proteins quantified in at least two of the three technical replicates for each UPS1 spiking amount. Proteins were considered variants if the Welch's *t*-test *p*-value was less than 0.05 and if the difference calculated by Perseus between the means of the $\log_2$ transformed normalized intensities for the triplicates of each UPS1 spiking amount was ≤ -1 or ≥ 1 (corresponding to fold changes of 1/2 and 2, respectively).

3 Results and discussion

3.1 Design of a calibrated sample for analytical workflow benchmarking

This study was designed to rigorously compare LD-, SG-, and TG-prepared samples for qualitative protein identification and label-free quantification. First, a series of six UPS1 calibrated spikes (0.5, 1, 2.5, 5, 10, and 25 fmol) in a yeast background was prepared to mimic a real complex sample with controlled fold changes. This allowed the protein concentration to be varied from very low to high in a nonvariant complex background proteome. The samples spiked with different amounts of UPS1 were divided in order to perform three technical replicates of the three preparation protocols, resulting in a total of 54 samples that were independently processed thereafter. The TG- and SG-prepared samples were digested and extracted according to the *in-gel* procedure, whereas the LD samples were digested *in-solution* and cleaned up by SPE. The 54 resulting protein digests were then randomly injected onto the nano-LC-MS/MS system with a 79-min gradient and top-10 MS/MS method. The proteins were identified and quantified using label-free XIC (Fig. 1).

3.2 Comparison of the TG, SG, and LD protocols for protein identification

The total numbers of yeast and UPS1 proteins identified by MS/MS at each spike point are shown in Fig. 2 for the technical triplicates analyzed in a random order. On average 1323, 1788 and 1979 yeast proteins were identified in the LD-, SG-, and TG-prepared samples, respectively. More yeast proteins were always identified in the 18 TG-prepared samples (1886–2057 proteins) than in those prepared using the

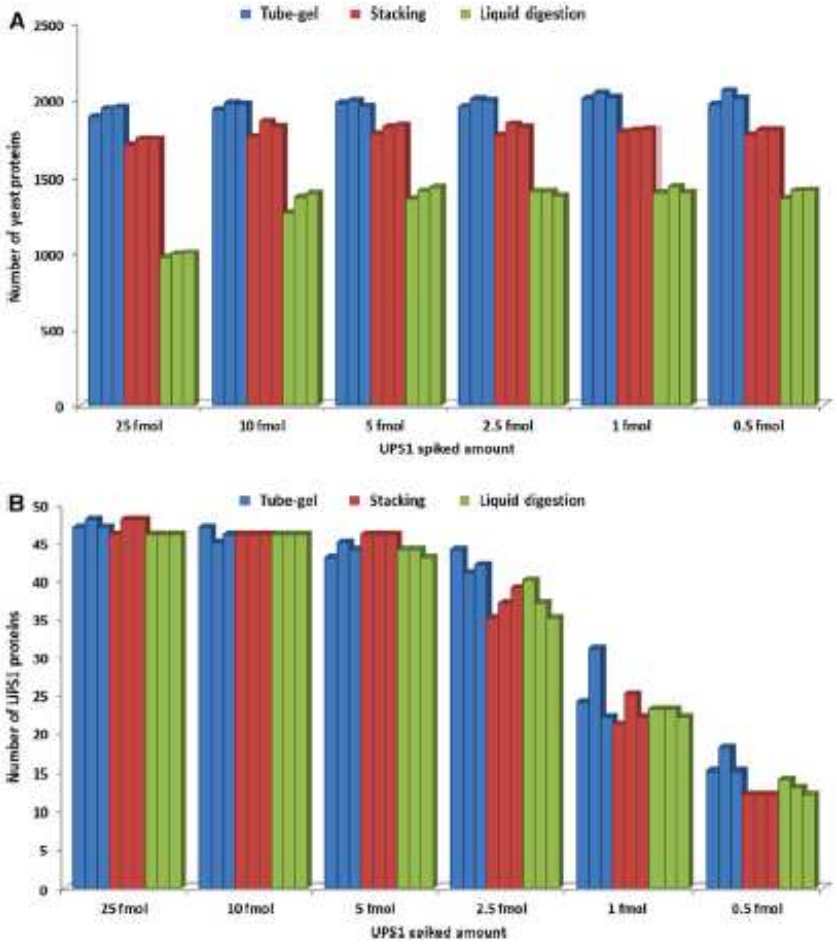


Figure 2. Number of (A) yeast and (B) UPS1 proteins identified by MS/MS in the three technical replicates for each UPS1 concentration prepared with the three preparation protocols, viz. tube-gel, stacking gel, and liquid digestion.

SG (1699–1854 proteins) or LD (970–1440 proteins) protocols. A similar and repeatable number of yeast proteins were identified in each sample prepared with the same protocol, in agreement with the constant yeast proteome background expected. Of note, the number of yeast proteins identified in the three 25 fmol LD replicates was lower than for the other spike points (Fig. 2A) due to a technical issue resulting in a 20% lower injected volume at a specific position in the autosampler for the three independently processed technical replicates. Nevertheless, this spike point was not excluded as we applied LFQ normalization.

Figure 2B shows that for the three highest spiking amounts, at least 43 of the 48 UPS1 proteins were identified in all the samples; however, all 48 proteins were only identified at the highest protein concentration and with the TG and SG protocols. The number of proteins identified then decreased gradually with the amount of UPS1 spiked, down to 12, 12, and 15 at 0.5 fmol in the LD-, SG-, and TG-prepared samples, respectively.

The higher number of proteins identified on average in the TG samples may be explained by the fewer preparation steps required for this method, which ensures an optimal incorporation of the sample in the gel and reduces the risk of protein loss.

The Venn diagram in Supporting Information Fig. 1 represents the distribution of all yeast proteins identified in the 18 analyses for the three preparation protocols. It shows that 55% of the yeast proteins identified in each type of sample are common to all three, with a common proportion of 76% between the TG and SG samples. The TG samples yielded the highest proportion of proteins uniquely identified throughout all protocols (10%). For SG and LD, this proportion was 5 and 4%, respectively. Compared with the gel-based samples, Supporting Information Fig. 2A shows that those prepared by LD yield a higher proportion of shorter proteins, while the TG protocol seems to favor longer proteins than does the SG method. When considering additional intrinsic physicochemical properties of identified proteins (isoelectric points and hydrophobicity indexes), the only significant difference reveals that gel-based protocols favor more hydrophobic proteins (Supporting Information Fig. 2B and C). Actually, general conclusions are difficult to draw from this experimental setup as a common protein pellet was used for the three protocols and no protocol-specific cell lysis procedures were applied.

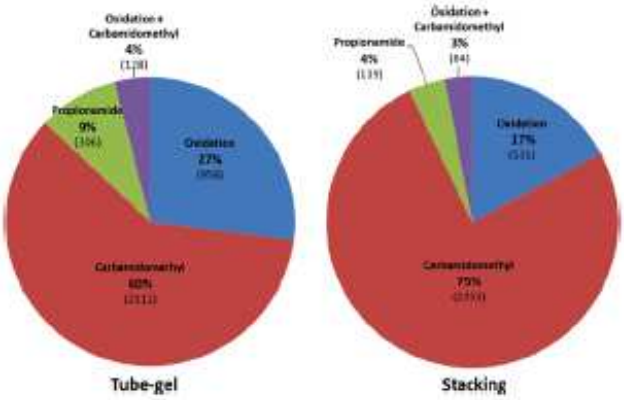


Figure 3. Percentages of different types of peptide modifications identified in the 18 analyses performed on tube-gel (left) and stacking gel samples (right). The proportions of mixed modifications (viz. carbamidomethyl + propionamide, oxidation + propionamide, and carbamidomethyl + propionamide + oxidation) are less than 1% in both cases.

Indeed, while strong chemical lysis could have been used prior to gel-based preparation, this is incompatible with LD and would therefore have biased subsequent comparisons.

The cellular localization distributions resulting from each sample preparation procedure were investigated by extracting GO annotations for the identified yeast proteins. As expected, these were mostly cytoplasmic and nuclear proteins. Similar localization distributions were obtained with the three protocols. Nevertheless, slightly more plasma membrane and endoplasmic reticulum proteins were identified in the TG samples, while more secreted proteins were detected in those prepared by LD (Supporting Information Fig. 3). These tendencies are more pronounced if only the fraction of proteins uniquely identified in each sample type is considered, ignoring nuclear proteins, which are overrepresented in the fraction of proteins identified only in the TG samples (Supporting Information Fig. 4). Again, more substantial variations would have been observed had the cell lysis protocols been adapted separately for each sample preparation protocol.

Supporting Information Fig. 5 shows the Venn diagram of sets of peptides identified in the 18 analyses performed for the three sample preparation protocols. The overlap between the sets (21%) is much lower than for the protein-level analysis. A significantly lower number of peptides (12 528) were identified in the LD samples than in the *in-gel* preparations (~20 000), for which 50% of the identified sequences were common to the two.

Figure 3 shows the proportions of different peptide modifications identified in the TG and SG samples. Note that since gel preparations can induce propionamide modifications due to the reaction of acrylamide monomers on cysteine residues [18], carbamidomethylation and propionamidation of cysteine residues were both defined as variable modifications during database searches for the TG and SG results. A total of 306 peptides (12% of all cysteine-containing peptides) were identified as partially propionamidated in the TG samples versus 119 (5% of all cysteine-containing peptides) in the SG ones. Polymerization in the TG protocol takes place directly in the presence of the sample, leading to direct contact between proteins and free acrylamide, and this may explain the higher number of propionamide modifications detected. Proteins are less exposed to acrylamide with the SG protocol as fewer monomers are present after polymerization [18] (particularly if the gel is prepared 1 day before the migration) and because the electrophoresis buffer contains glycine that can react with free acrylamide [19]. An alternative method would involve reducing the proteins before acrylamide/bis-acrylamide polymerization to induce systematic propionamidation [20]. As cysteine carbamidomethylation was set as a fixed modification and used for quantification in the analysis of the LD samples, the cysteine-containing peptides were retained in the quantification for the TG and SG samples by summing both variable modifications.

More oxidized methionines were detected in the TG than in the SG samples (958 vs. 531), probably because of direct contact (during TG preparation) between the proteins and ammonium persulfate, a strong oxidizing agent. Since methionine oxidation was set as a variable modification for all three protocols, all the methionine-containing peptides identified as oxidized, as well as their unmodified counterparts, were excluded from the subsequent quantification.

3.3 Comparison of the TG, SG, and LD protocols for protein quantification

The second objective of this study was to compare the three protocols in terms of quantification accuracy, repeatability, and detection of known UPS1 proteins as variants. For this purpose, we first set up three pairwise comparisons between different UPS1 concentrations. The overall quantification linearity was derived from the coefficients of determination (R^2) between the spiked amount and the peptides signal intensities (only peptides quantified in at least two replicates in at least three spiking amounts were considered). Then, for each pairwise comparison, we compared the expected fold changes with the experimentally measured fold changes for UPS1 proteins. Repeatability was assessed via the CV of the UPS1 proteins normalized intensities between technical replicates, and between all 18 analyses for the yeast proteins. Finally, the ability of the workflow to detect known variant proteins was investigated by constructing volcano plots ($-\log_{10}$(*p*-value) vs. difference) and via sensitivity and false discovery proportion (FDP) criteria.

The three comparisons set up to mimic investigations of real complex biological samples with different protein contents were as follows: (A) 25 fmol versus 2.5 fmol UPS1, mimicking a large magnitude fold change between variant proteins with the low point being close to the quantification limit for most proteins; (B) 25 fmol versus 5 fmol UPS1, mimicking an intermediate fold change; and (C) 10 fmol versus

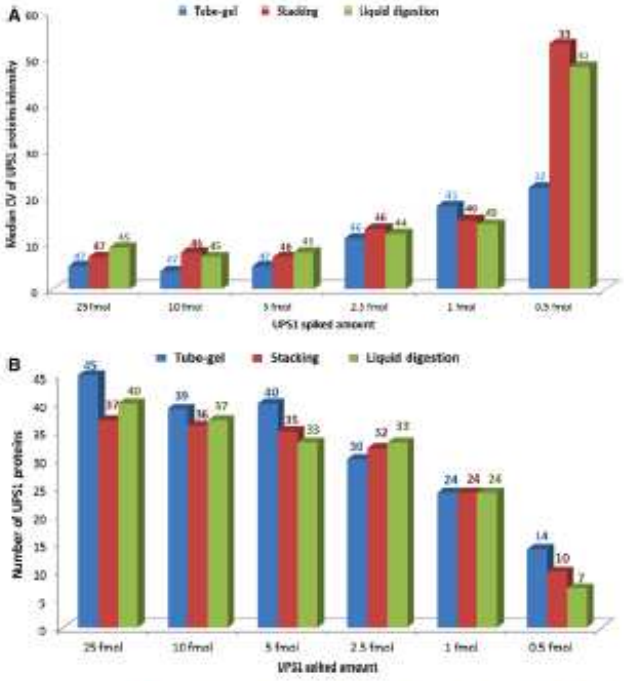


Figure 4. (A) Median CV calculated for different spiked amounts of UPS1 from the normalized intensities of UPS1 proteins between measurements performed on three technical replicates for each of the three preparation protocols. In each case, the number of UPS1 proteins taken into account for the calculation of the median CV is shown above the corresponding bar. (B) From the same dataset: the total number of UPS1 proteins identified with two or more unique peptides and a normalized intensities CV < 20% between the technical replicates is represented.

5 fmol, mimicking a small fold change. We did not consider the data obtained with UPS1 spiked at less than 2.5 fmol because the number of proteins identified was too low to make the comparison meaningful.

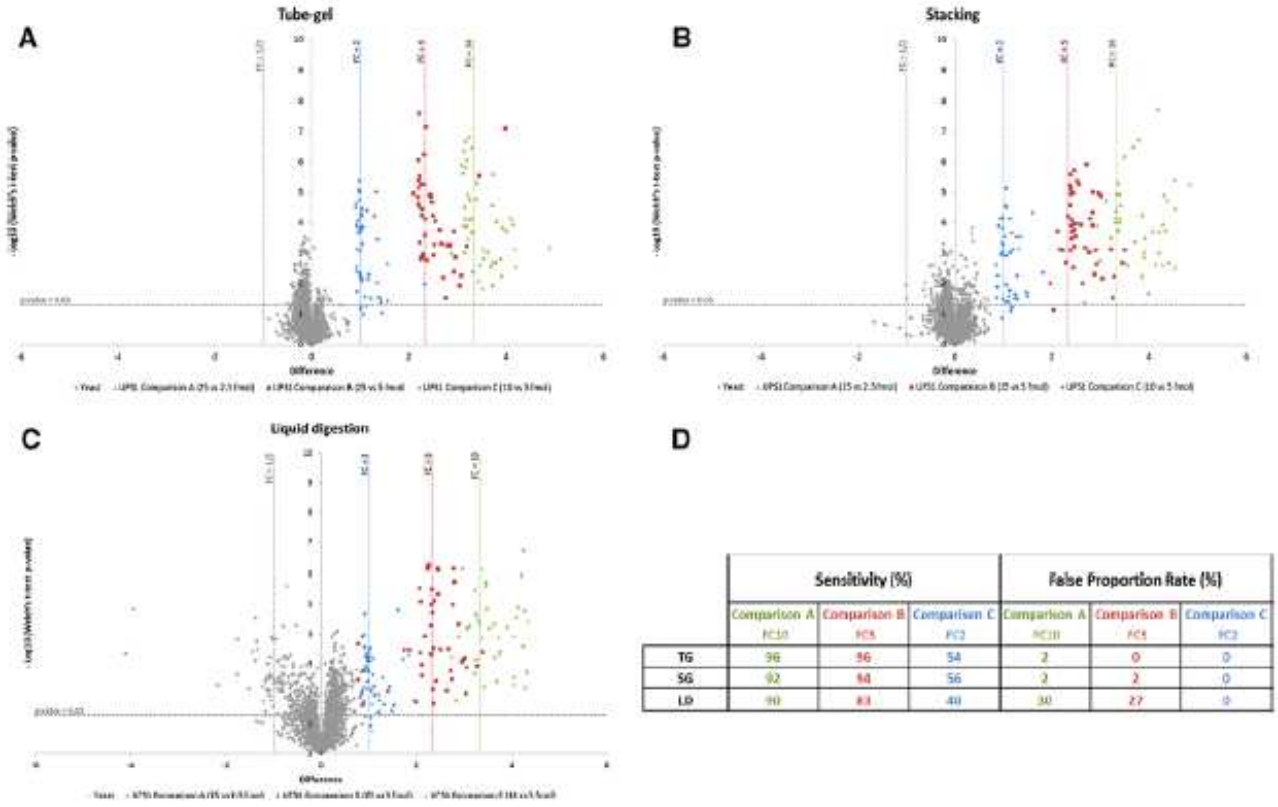
In terms of quantification linearity, the correlation between the intensity and the six spiked amounts of UPS1 was determined for all peptides quantified in at least two replicates with a minimum of three spiking amounts. The highest correlation was obtained for the TG samples, with the data for 79% of the peptides having $R^2 \geq 0.9$. In comparison, the same assessment of the data measured for the SG and LD samples yielded values of 69 and 64%, respectively (Supporting Information Fig. 6).

Then, measured fold changes were compared to theoretical fold changes at the protein level for the three comparisons (A–C). Fold changes were calculated as the ratio of the average normalized intensities measured in both conditions. Parts A–C of Supporting Information Fig. 7 show the distribution of fold changes obtained for samples prepared using the three protocols during comparisons A–C, respectively. The median fold changes for all UPS1 proteins identified with each protocol are presented in Supporting Information Fig. 7D. The UPS1 proteins that could not be quantified in the data for the 2.5 and 5 fmol spiking experiments were withdrawn for the fold change representations and calculations of the median fold changes. For comparison A, the fold changes best centered on the theoretical value (expected fold change of 10) were those derived from the TG data, while the difference between the theoretical and observed median fold changes was the highest for the SG data (13.9, compared with 11.2 and 10.4 for the LD and TG analysis, respectively). All three preparation protocols yielded fold changes well distributed around the expected value for comparisons B (expected fold change of 5) and C (expected fold change of 2), with a particularly small scatter for comparison C. These results highlight the strength of our global workflow even for low fold changes, provided that the data to compare are not too noisy and that the MS signals are not run close to their quantification limit.

The repeatability of each preparation protocol was evaluated by calculating the median CV of the normalized intensities measured between the 18 analyses for the yeast proteins (only considering those for which at least one normalized intensity was obtained for each spiking amount triplicate) and between technical replicate for the UPS1 proteins (only considering those for which at least one normalized intensities was obtained for each spiking amount). The median CV calculated for the yeast proteins were 7, 9, and 10%, for the TG, SG, and LD samples, based on the normalized intensities for 1780, 1612, and 1098 proteins, respectively. No significant difference was observed in the distributions of these CV between the three protocols (Supporting Information Fig. 8A). The median CV for the UPS1 proteins are plotted in Fig. 4A for each protein concentration, along with the corresponding number of proteins considered in the calculation. The median CV are less than 10% for the 25, 10, and 5 fmol UPS1 spiking experiments, with the lowest values obtained with TG-prepared samples. For the measurements performed at 2.5 and 1 fmol UPS1, the medians for all three datasets range from 10 to 20%. After spiking with 0.5 fmol UPS1, the normalized intensities varied much less in the TG data; in this case, the median CV were 22, 53, and 48% for the TG, SG, and LD samples, respectively. No significant difference was observed in the distributions of these CV between the three protocols, except for the distribution observed at 1 fmol with SG that was more dispersed than those obtained with TG and LD (Supporting Information Fig. 8B).

Figure 4B presents an additional analysis of these data: the total number of UPS1 proteins identified with two or more unique peptides and low normalized intensities CV (<20%) between the technical replicates are represented. With these criteria, more UPS1 proteins were identified in the TG than in the SG or LD data obtained after spiking from 25 to 5 fmol. At 2.5 fmol UPS1, however, slightly more proteins were identified in the LD data. Note finally that at the lowest protein concentration (0.5 fmol UPS1), the TG samples were the only ones to allow the repeatable identification of more than 10 UPS1 proteins with these stringent criteria, in line with upper presented results.

Parts A, B, and C of Fig. 5 show the volcano plots obtained for the pairwise comparisons conducted on the TG, SG, and



	Sensitivity (%)			False Proportion Rate (%)		
	Comparison A FC10	Comparison B FC5	Comparison C FC2	Comparison A FC10	Comparison B FC5	Comparison C FC2
TG	96	96	54	2	0	0
SG	92	94	56	2	2	0
LD	90	83	40	30	27	0

Figure 5. Volcano plots ($-\log_{10}$ (*p*-value) vs. difference calculated by Perseus) for the comparison experiments (A–C, see (D)) conducted on (A) tube-gel, (B) stacking gel, and (C) liquid digestion samples. The *p*-values are those of Welch's *t*-tests performed on the differences between the means (from triplicate experiments) of the $\log_2$ transformed normalized intensities measured under the two compared conditions. Proteins not quantified in at least two of the three technical replicates for the two compared conditions were not considered. Proteins were considered variant if the associated *p*-value was < 0.05 and the difference was ≤ -1 or ≥ 1. (D) The sensitivity and false discovery proportion (FDP) calculated for the three comparisons and three types of samples. The true positive (TP) and false positive (FP) were the UPS1 and yeast proteins identified as variant. The sensitivity was then calculated as the proportion of UPS1 proteins identified as variant (TP/48), and the FDP as FP/(TP + FP).

LD datasets, respectively. The UPS1 proteins are mostly well distinguished from the yeast background with all three workflows, although there is more overlap in the LD data (Fig. 5C). In line with the criteria usually applied in global discovery proteomics, proteins for which the fold change was significant (difference calculated by Perseus ≤ -1 or ≥ 1, equivalent to a fold change of 1/2 or 2) and the differential expression was significant (*p*-value < 0.05) were considered variant. UPS1 proteins identified as variant were considered true positives (TPs), and the same criteria were applied for the yeast background proteins to determine the false positives (FPs). The sensitivity was then quantified as the proportion of UPS1 proteins identified as variant (TP/48), and the FDP as FP/(TP + FP) [21]. These values are shown Fig. 5D as calculated for the three types of samples and three comparison experiments. In comparisons A and B, the UPS1 proteins were better discriminated from the yeast background in the TG and SG data than in the LD data, with the TG samples yielding the highest sensitivities with an equivalent or lower FDP than that calculated for the SG experiments. The FDPs calculated for the LD data are markedly higher. As expected given the stringent fold change criterion applied, the sensitivities obtained for comparison C were markedly lower (~50%), as proteins with an experimental fold change slightly less than 2 were excluded. As shown in Supporting Information Fig. 7C, the experimental fold changes were well distributed around the expected value for this comparison. No false positives were identified with any of the three protocols for comparison C. As for comparisons A and B, the sensitivities obtained with the TG and SG samples were better than those calculated for the LD data.

4 Concluding remarks

The results obtained in this study highlight the value of TG preparation for the identification and label-free XIC quantification of proteins present at low abundance in complex biological samples. The results obtained with TG samples were better than those achieved with LD samples (often considered as the gold standard), and equivalent or better than those obtained with SG samples. No attempt was made to

compare TG with the filter-aided sample preparation protocol, as Fischer and Kessler [20] have already shown that the two procedures are equivalent in terms of the number of proteins identified, while filter-aided sample preparation is more prone to sample loss.

By using a series of spiking experiments with known amounts of UPS1 proteins, this study demonstrates that the TG procedure is at least as effective and repeatable as the SG approach, and more effective than LD for label-free intensity-based protein quantification. There was no propensity among the proteins identified in the TG samples toward any specific subcellular localization. The TG protocol allowed differentially expressed proteins to be identified within complex biological samples, with less dispersed fold changes than with SG or LD samples. The time gain offered by the TG approach makes it very attractive, particularly if large sample series have to be analyzed.

The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium database with the identifier PXD003841 [17]. This work was supported financially by the "Agence Nationale de la Recherche" (ANR) and the French Proteomic Infrastructure (ProFI; ANR-10-INBS-08-03). LM was supported by a doctoral fellowship from the French Ministry of Research. We thank Dr. T. Rabilloud for his advices and valuable feedback on the manuscript.

The authors have declared no conflict of interest.

5 References

[1] Zhang, N., Chen, R., Young, N., Wishart, D. et al., Comparison of SDS- and methanol-assisted protein solubilization and digestion methods for *Escherichia coli* membrane proteome analysis by 2-D LC-MS/MS. *Proteomics* 2007, *7*, 484–493.

[2] Zhou, J. Y., Dann, G. P., Shi, T., Wang, L. et al., Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. *Anal. Chem.* 2012, *84*, 2862–2867.

[3] Rosenfeld, J., Capdevielle, J., Guillemot, J. C., Ferrara, P., In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. *Anal. Biochem.* 1992, *203*, 173–179.

[4] Gautier, V., Mouton-Barbosa, E., Bouyssie, D., Delcourt, N. et al., 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*, 527–539.

[5] Lu, X., Zhu, H., Tube-gel digestion: a novel proteomic approach for high throughput analysis of membrane proteins. *Mol. Cell. Proteomics* 2005, *4*, 1948–1958.

[6] Han, C. L., Chien, C. W., Chen, W. C., Chen, Y. R. et al., A multiplexed quantitative strategy for membrane proteomics: opportunities for mining therapeutic targets for autosomal dominant polycystic kidney disease. *Mol. Cell. Proteomics* 2008, *7*, 1983–1997.

[7] Smolders, K., Lombaert, N., Valkenborg, D., Baggerman, G., Arckens, L., An effective plasma membrane proteomics approach for small tissue samples. *Sci. Rep.* 2015, *5*, 10917.

[8] Zhou, J., Xiong, J., Li, J., Huang, S. et al., Gel absorption-based sample preparation for the analysis of membrane proteome by mass spectrometry. *Anal. Biochem.* 2010, *404*, 204–210.

[9] Cao, R., He, Q., Zhou, J., He, Q. et al., High-throughput analysis of rat liver plasma membrane proteome by a nonelectrophoretic in-gel tryptic digestion coupled with mass spectrometry identification. *J. Proteome Res.* 2008, *7*, 535–545.

[10] Cao, L., Clifton, J. G., Reutter, W., Josic, D., Mass spectrometry-based analysis of rat liver and hepatocellular carcinoma Morris hepatoma 7777 plasma membrane proteome. *Anal. Chem.* 2013, *85*, 8112–8120.

[11] Yu, H., Wakim, B., Li, M., Halligan, B. et al., Quantifying raft proteins in neonatal mouse brain by 'tube-gel' protein digestion label-free shotgun proteomics. *Proteome Sci.* 2007, *5*, 17.

[12] An, M., Dai, J., Wang, Q., Tong, Y., Ji, J., Efficient and clean charge derivatization of peptides for analysis by mass spectrometry. *Rapid Commun. Mass Spectrom.* 2010, *24*, 1869–1874.

[13] Lin, Y., Liu, H., Liu, Z., Liu, Y. et al., Development and evaluation of an entirely solution-based combinative sample preparation method for membrane proteomics. *Anal. Biochem.* 2013, *432*, 41–48.

[14] Cox, J., Hein, M. Y., Luber, C. A., Paron, I. et al., Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* 2014, *13*, 2513–2526.

[15] Carapito, C., Burel, A., Guterl, P., Walter, A. et al., MSDA, a proteomics software suite for in-depth mass spectrometry data analysis using grid computing. *Proteomics* 2014, *14*, 1014–1019.

[16] Micallef, L., Rodgers, P., eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. *PLoS One* 2014, *9*, e101717.

[17] Vizcaino, J. A., Deutsch, E. W., Wang, R., Csordas, A. et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 2014, *32*, 223–226.

[18] Rabilloud, T., Variations on a theme: changes to electrophoretic separations that can make a difference. *J. Proteomics* 2010, *73*, 1562–1572.

[19] Geisthardt, D., Kruppa, J., Polyacrylamide gel electrophoresis: reaction of acrylamide at alkaline pH with buffer components and proteins. *Anal. Biochem.* 1987, *160*, 184–191.

[20] Fischer, R., Kessler, B. M., Gel-aided sample preparation (GASP)—a simplified method for gel-assisted proteomic sample generation from protein extracts and intact cells. *Proteomics* 2015, *15*, 1224–1229.

[21] Ramus, C., Hovasse, A., Marcellin, M., Hesse, A. M. et al., Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *J Proteomics* 2016, *132*, 51–62.

2- Poursuite des optimisations du protocole « *Tube-Gel* » pour la protéomique quantitative sans marquage

Le premier travail du protéomiste consiste souvent à choisir un tampon de lyse et de solubilisation des protéines, de manière à ce que l'échantillon puisse répondre au mieux à la question biologique posée. En effet, nous avons pu voir au *Chapitre II-I* que les protéines membranaires requéraient l'utilisation de détergents comme le SDS. Cependant, comme évoqué au *Chapitre I-II-2*, le choix de la préparation d'échantillons va dépendre du tampon d'extraction utilisé. Le TG s'est avéré être une bonne alternative au SG lorsque les protéines sont solubilisées dans un tampon à base de SDS[145]. Il a par ailleurs été employé lors de l'utilisation de tampons à base d'un mélange SDS-urée[140, 146] ou encore de CHAPS et de Triton X-100[139]. Afin de simplifier l'étape consistant à choisir un couple tampon d'extraction-préparation d'échantillons, la polyvalence du TG vis-à-vis de plusieurs tampons de solubilisation a été évaluée. Le but des optimisations réalisées ici est de permettre au protéomiste de se focaliser uniquement sur le choix du tampon d'extraction selon le type de protéines à extraire ou les modifications induites, et de se diriger vers un TG, rapide à implémenter quel que soit le nombre de conditions à comparer, du fait de sa polyvalence.

Sachant qu'une polymérisation de l'acrylamide par une méthode chimique, c'est-à-dire par l'utilisation de TEMED comme catalyseur et de persulfate d'ammonium (APS) comme initiateur :

- N'est pas compatible avec la présence de détergents cationiques du fait de la précipitation de l'APS dans ces conditions de faible pH[147], menant à l'inhibition de la polymérisation[148],

- Est très oxydante de par la présence d'APS,

L'utilisation d'une photopolymérisation nécessitant l'utilisation de colorants (comme le bleu de méthylène, l'éosine ou la riboflavine (FMN)), d'un couple oxydo-réducteur et l'activation par la lumière a été envisagée[149]. Ce type de photopolymérisation présente plusieurs avantages par rapport à la polymérisation chimique, comme :

- Une tolérance plus large aux conditions expérimentales, c'est-à-dire qu'elle n'est pas inhibée à pH acides ou en présence de thiols. Dans le cas de l'utilisation du bleu de méthylène comme colorant, la gamme de pH permettant la polymérisation s'étend de 3 à 10[150].

- La rapidité de polymérisation, ainsi qu'un meilleur complètement du fait d'un meilleur taux de conversion de monomères d'acrylamide en polymères[151, 152], limitant ainsi les modifications

des protéines par l'acrylamide[153], comme c'est le cas pour la modification propionamide des cystéines.

- L'absence de pouvoir oxydant[149, 152, 154], permettant de générer moins d'oxydations des méthionines.

Ainsi, dans un premier temps, 32 combinaisons de tampon d'extraction – technique de polymérisation de l'acrylamide ont été effectuées en TG de 50 et 100 µl. Le TG, lors de son évaluation réalisée au *Chapitre II-II-1*, a été polymérisé dans des tubes Eppendorfs avec un volume final de 100 µl. Dans l'optique de faciliter la manipulation des TG, ceux-ci ont été ici polymérisés sur plaque 96 puits. Ainsi, en plus du volume « originel » de 100 µl, un volume plus adéquat à la manipulation sur plaque 96 puits de 50 µl a également été testé. Dans un second temps, les 10 meilleures combinaisons ont fait l'objet d'une évaluation pour la protéomique quantitative sans marquage à l'aide d'une approche par comptage de spectres MS/MS. La Figure II-1 résume les deux étapes permettant l'optimisation du protocole TG afin de le rendre polyvalent.

Ces travaux ont été menés en collaboration avec le Dr Thierry RABILLOUD, directeur de recherche au sein de l'Institut de Biosciences et Biotechnologies de Grenoble (CEA).

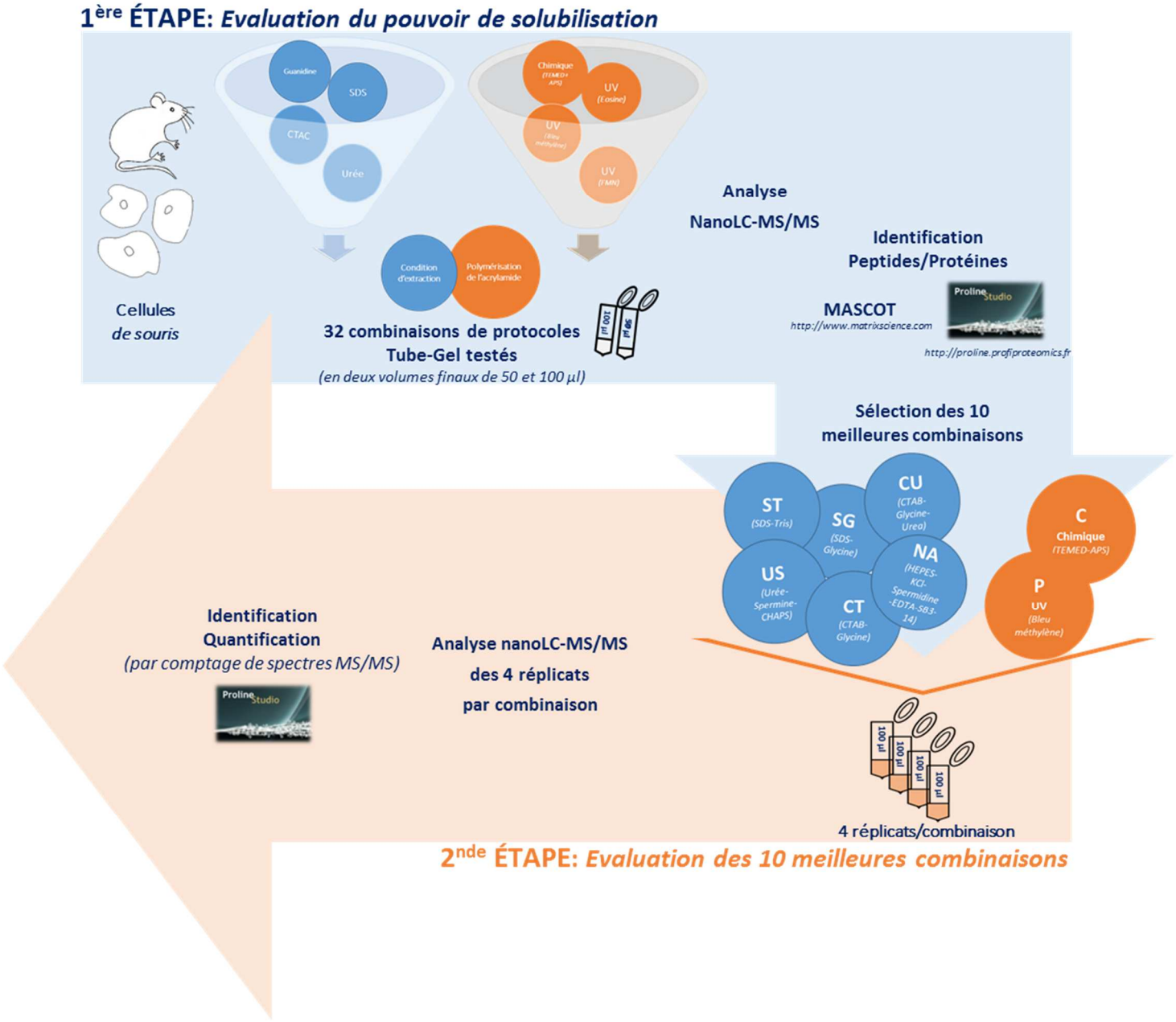


Figure II-1 - Résumé des étapes analytiques pour l'optimisation du protocole « *Tube-Gel* »

a. Exploration du pouvoir de solubilisation de 64 protocoles

Trente-deux combinaisons de tampons d'extraction-méthode de polymérisation de l'acrylamide ont été testées en volume de TG final de 50 et 100 µl, soit 64 conditions au total sur un lysat total de cellules de souris (J774, ECACC, Salisbury.) Malgré l'incompatibilité théorique de certaines combinaisons (comme détergent cationique CTAC (chlorure de cétrimonium) et polymérisation chimique), l'ensemble des conditions ont réellement été testées. Une polymérisation des TG n'a pas toujours été obtenue avec un volume de 50 µl, en raison d'une grande surface de contact avec l'oxygène inhibant la polymérisation, conséquente de l'utilisation de plaques 96 puits.

Le nombre de protéines identifiées pour chacune des conditions est résumé dans le Tableau II-1. Nous distinguons trois catégories :

- Les tampons à base de SDS qui donnent les résultats d'identification les plus faibles. Cela peut s'expliquer par une manipulation difficile de ce type de tampons de par la présence de SDS.

- Les tampons à base de CTAC, mais aussi à base d'urée-spermine et guanidine-spermine qui permettent d'obtenir les meilleurs résultats d'identification.

- L'extraction en conditions natives, qui malgré une faible dénaturation des protéines donne lieu à des résultats d'identification intermédiaires aux deux premières catégories.

Méthode de polymérisation \ Tampon de solubilisation		CTAC-Phosphate	CTAC-Phosphate-Urée	CTAC-Citrate-Chlorure de guanidine	SDS-Tris	SDS-Citrate	Natif	Urée-Spermine	Guanidine-Spermine
Photopolymérisation *Bleu de méthylène*	50 µl	1487	1647		785	1163	1164	1541	1404
	100 µl	1609	1881	1757	997	1038	1353	1712	1422
Photopolymérisation *Eosine*	50 µl	1629	1766		950	1108	1215	1598	1362
	100 µl	1882	1900	1697	852	1081	1263	1641	1510
Photopolymérisation *FMN*	50 µl				690	1011	996	1219	1521
	100 µl	1528	1943	1615	877	1008	1238	1679	1561
Polymérisation Chimique *TEMED/APS*	50 µl				912	1068	1235	1831	
	100 µl				897	883	1446	1798	

Tableau II-1 - Nombre de protéines identifiées avec au moins un peptide unique pour chaque combinaison testée.
Les cellules quadrillées représentent les conditions pour lesquelles la polymérisation du TG n'a pas fonctionné

Les résultats en termes de nombre de peptides identifiés suivent les mêmes tendances. Concernant les modifications induites par les différents protocoles, très peu de différences sont observées :

- Entre 15 et 20 % d'oxydations des méthionines sont observées sur l'ensemble des peptides dans l'ensemble des protocoles. L'absence de conditions oxydantes inhérente à l'utilisation de la photopolymérisation semble ne pas permettre la diminution d'induction de cette modification. Notons cependant que cette modification peut ne pas uniquement provenir de la technique de polymérisation de l'acrylamide, mais d'autres processus comme le temps de contact de l'échantillon avec l'air ambiant par exemple.

- Aux alentours de 1 % de modifications propionamides des cystéines sur l'ensemble des peptides pour tous les protocoles, sauf ceux impliquant l'utilisation d'urée-spermine et guanidine-spermine en combinaison avec une photopolymérisation pour lesquels la proportion de propionamides varie entre 5 et 10 %. Le taux de cette modification devrait être limité du fait de l'utilisation de CTAC, qui inactive les cystéines de par le pH acide, mais aussi par l'utilisation de photopolymérisation, qui présente un meilleur taux de conversion de monomères d'acrylamide et une plus grande rapidité que la polymérisation chimique. Or, les taux de propionamides induites étant déjà très faibles, ces optimisations n'ont pas apporté d'amélioration.

- Entre 5 et 10 % de carbamidométhylations des cystéines sur la totalité des peptides pour les protocoles présentant environ 1 % de propionamides. En effet les combinaisons générant davantage de propionamides engendrent en compensation moins de carbamidométhylations, la modification des cystéines par les monomères d'acrylamide intervenant avant leur alkylation au cours du protocole. Enfin, les combinaisons employant le tampon SDS-Tris ne présentent que 2 % de cystéines carbamidométhylées, reflétant que l'utilisation de ce tampon permet l'identification de deux fois moins de peptides à cystéines, probablement du fait de la difficulté de manipulation.

D'autre part, les annotations GO ont été extraites afin d'obtenir la distribution dans les différents compartiments cellulaires des protéines extraites pour chaque protocole. Aucune différence significative n'a été observée entre les protocoles, avec comme ordres de grandeur : 20 % de protéines annotées « *plasma membrane* », 40 % de protéines nucléaires et 55 % de protéines cytoplasmiques.

Pour finir, la polymérisation chimique en présence de CTAC n'a pas abouti, comme attendu par la théorie. Par ailleurs, le tampon guanidine-spermine semble ne pas être compatible avec une polymérisation chimique.

b. Evaluation des combinaisons les plus pertinentes

A partir des résultats de cette première étape d'évaluation du pouvoir de solubilisation, certaines combinaisons ont été retenues et parfois modifiées de manière à évaluer leur répétabilité par rapport au protocole de référence. Ainsi, les combinaisons employant des tampons à base de CTAC ont été conservées du fait des bons résultats d'identification, mais le phosphate a été remplacé par de la glycine, et la condition CTAC-Citrate/Chlorure de guanidine qui n'a pas toujours permis d'obtenir la polymérisation du TG a été ôtée. Les protocoles à base de tampon SDS ont malgré les difficultés de manipulation été retenus car ils font office de référence. Le citrate a cependant été remplacé par de la glycine. Le protocole natif générant des résultats plutôt satisfaisants malgré le faible pouvoir dénaturant du tampon a lui aussi était maintenu. Pour finir, le protocole à base d'urée et de spermine a été modifié par l'ajout de CHAPS. L'ensemble des protocoles testés sont résumés dans le Tableau 1 du papier prochainement soumis pour publication.

L'ensemble des combinaisons entre ces tampons d'extraction et une polymérisation chimique ainsi qu'une photopolymérisation à l'aide de bleu de méthylène comme colorant ont été préparées en quatre réplicats de TG de volume final de 100 µl, dans des tubes Eppendorfs, afin de limiter les problèmes liés aux plaques 96 puits. Ces réplicats ont permis d'évaluer la répétabilité des différents protocoles par une méthode de quantification sans marquage par comptage de spectres MS/MS.

Les résultats, disponibles dans le papier qui sera prochainement soumis pour publication dans *Scientific Reports*, ont démontré que :

- L'ensemble des conditions testées sont équivalentes en termes de nombres de protéines identifiées,

- La photopolymérisation n'affecte pas les résultats, aussi bien qualitatifs que quantitatifs, si ce n'est dans certains cas au niveau des modifications induites,

- Les protocoles sont globalement répétables avec des CV calculés à partir des valeurs de quantification sur les quatre réplicats de l'ordre de 20 %.

Finalement, l'ensemble de ces résultats a démontré la polyvalence du TG, et le choix d'un protocole pourra être dirigé en fonction des modifications souhaitées ou non, de la facilité de manipulation ou encore si l'étude porte sur les protéines cytosoliques davantage représentées dans les protocoles natifs.

En plus de ces optimisations, la possibilité de réduire au maximum le volume du TG sera prochainement étudiée, notamment dans le but de permettre l'analyse de quantités de protéines de plus en plus réduites. En effet les quantités requises pour des analyses protéomiques représentent souvent une difficulté pour les biologistes, en particulier dans le contexte d'échantillons préparés par microdissection laser. Ainsi, des TG polymérisés au sein de tubes en verre très fins offriraient de grandes surfaces de contact favorables pour les digestions enzymatiques, d'autant que ces optimisations permettraient de s'affranchir de l'étape de découpe des TG qui peut s'avérer chronophage lorsqu'un grand nombre d'échantillons est à analyser.

III-Optimisation d'une méthode d'extraction des protéines et de préparation d'échantillons pour l'étude de biopsies ganglionnaires

Les lymphomes B diffus à grandes cellules (LBDGC) représentent l'une des nombreuses entités des lymphomes non-Hodgkiniens. Ils proviennent de la prolifération de lymphocytes B anormaux qui donnent naissance par leur accumulation à des tumeurs, qui se développent généralement dans les ganglions et les tissus lymphatiques. Bien que les cellules B périphériques malignes peuvent être facilement obtenues en quantité suffisante dans les fluides biologiques comme le plasma, et que les fluides sont souvent préférés pour l'étude de ces maladies car facilement accessibles par des moyens peu invasifs, il est possible que leur protéome ne reflète pas toujours leur point d'origine[155-157]. De ce fait, les tissus lymphatiques, et notamment les ganglions lymphatiques, semblent être des matériaux de départ optimaux pour l'étude des lymphomes, d'autant que des prélèvements sont systématiquement effectués à l'hôpital pour le diagnostic de la maladie[156, 158, 159]. Cependant, l'utilisation de ganglions lymphatiques pour les analyses de protéomique souffre de quelques désavantages, comme[156] :

- Le nombre d'échantillons disponibles, puisque même si un prélèvement est réalisé de manière systématique, ces tissus ne sont pas toujours congelés.

- La présence de sang ou de plasma dans les prélèvements pouvant interférer avec l'analyse protéomique.

- La représentativité de la maladie au sein de la biopsie analysée. Notons cependant que ce problème peut survenir au même titre lors de l'établissement du diagnostic à l'hôpital.

Dans ce contexte, une optimisation de l'extraction de protéines à partir de ganglions lymphatiques pour l'analyse protéomique globale par spectrométrie de masse a été effectuée dans le cadre de l'étude de la recherche de biomarqueurs de résistance au traitement des LBDGC. Cette étude a été réalisée en binôme avec le Dr Luc FORNECKER.

1- Optimisation d'extraction de protéines à partir de tissus frais

a. Optimisation de l'extraction de protéines à l'aide d'un potter

Au laboratoire, des études ont été menées sur différents tissus, notamment sur de la peau humaine. L'extraction des protéines était alors effectuée à l'aide d'un tampon Laemmli et d'un stimuli mécanique exercé à l'aide d'un potter en verre[160].

Ainsi, à partir de trois prélèvements effectués sur un même ganglion lymphatique congelé, trois protocoles ont été évalués, impliquant : une extraction au potter simple, une extraction au potter suivie d'une étape de sonication, ainsi qu'une extraction au potter accompagnée d'une étape de précipitation à l'acétone. L'ensemble des protocoles testés est résumé par la Figure III-1.

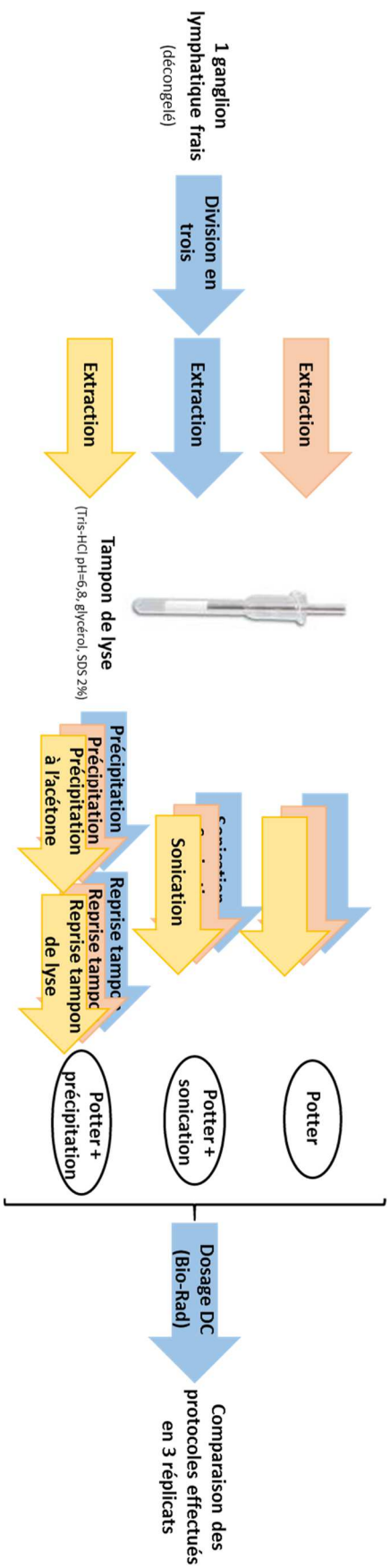


Figure III-1 - Résumé des trois protocoles d'extraction des protéines réalisée au potter à partir de ganglions lymphatiques testés

Les trois protocoles ont été évalués à l'aide de données issues d'un dosage DC, commercialisé par Bio-Rad, basé sur la méthode de Lowry[161]. Cette méthode permet de réduire les acides aminés aromatiques de manière à former un complexe bleu qui absorbe à une longueur d'onde de 750nm. Sachant que la fréquence des acides aminés aromatiques peut varier d'un échantillon à l'autre, et que l'échantillon de référence permettant d'établir la droite d'étalonnage est une protéine seule qui n'est pas représentative des échantillons complexes (l'albumine de sérum bovin, BSA), les résultats de ce type de dosage ne permettent d'obtenir qu'une estimation globale de la quantité de protéines présentes dans l'échantillon. Les résultats sont donnés par la Figure III-2.

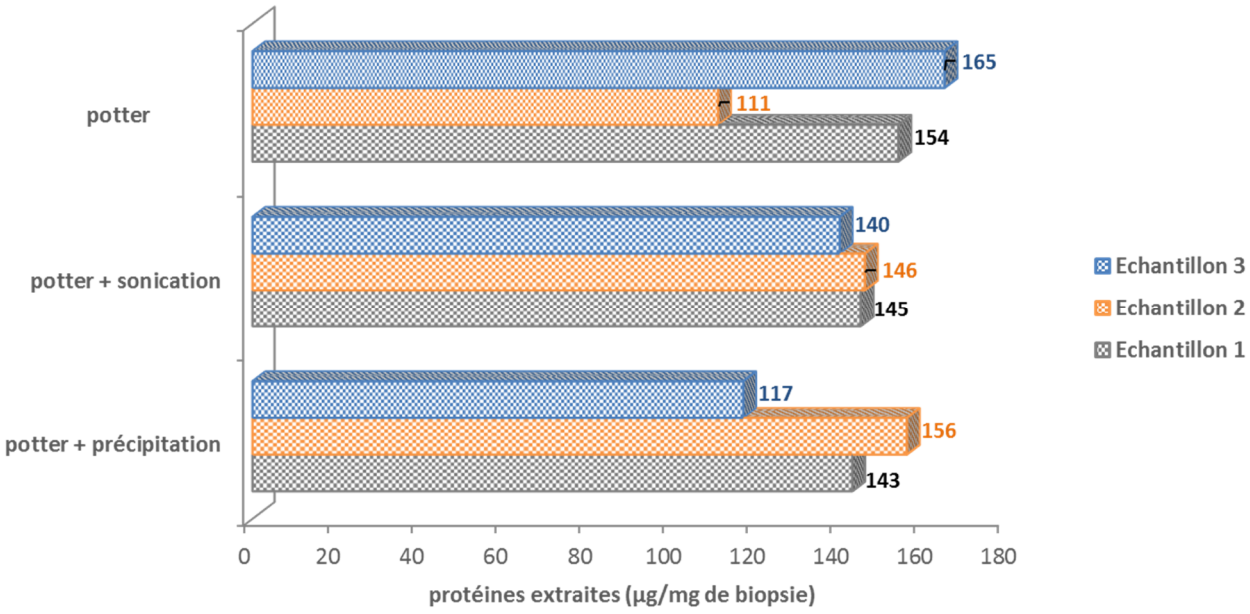


Figure III-2 - Quantité de protéines extraites par trois protocoles différents à partir de trois échantillons de ganglions lymphatiques (Dosage DC, Bio-Rad)

Le protocole qui semble fournir les résultats les plus répétables est la méthode d'extraction au potter suivie d'une étape de sonication, sous réserve que les résultats d'un dosage DC dans des mêmes conditions de manipulation soient répétables. Les méthodes d'extractions classique au potter et au potter suivie d'une étape de précipitation à l'acétone donnent des écarts de concentrations pouvant différer d'un-tiers d'un échantillon à l'autre, alors que la méthode d'extraction au potter accompagnée d'une étape de sonication diffère de plus ou moins 4 %. C'est pourquoi la méthode d'extraction au potter avec sonication sera préférentiellement utilisée pour la suite des optimisations.

b. Optimisation de la préparation d'échantillons

La méthode d'extraction des protéines à partir de tissus ganglionnaires impliquant l'utilisation de SDS, la préparation d'échantillons devra permettre de rendre l'échantillon compatible avec une analyse LC-MS/MS. De ce fait, un TG ainsi qu'un SG découpé en deux bandes ont été testés sur des extraits protéiques de trois ganglions différents. Ils ont fait l'objet d'une analyse nanoLC-MS/MS sur le Q-Exactive + de THERMO FISHER SCIENTIFIC (Figure III-3). A partir des résultats d'identification, chaque méthode a pu être comparée.

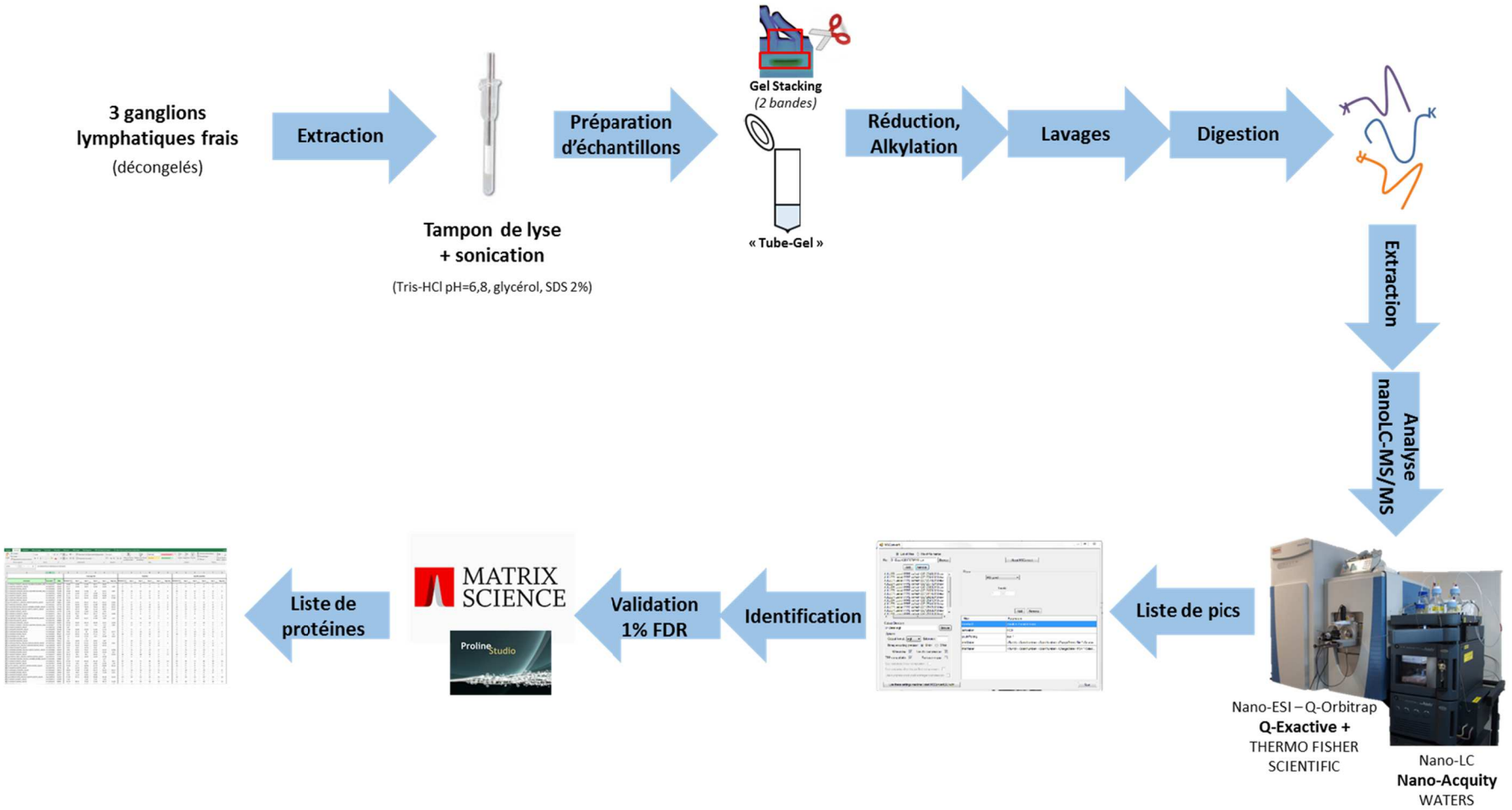


Figure III-3 - Schéma analytique de la comparaison des protocoles de préparation d'échantillons pour les extraits protéiques issus de tissus ganglionnaires

Les résultats d'identification sont donnés par la Figure III-4. Ceux-ci permettent de rendre compte que le SG découpé en deux bandes permet d'identifier davantage de protéines, notamment pour l'échantillon 1, cependant nous comparons ici un échantillon résultant de deux analyses (SG) à une préparation TG s'effectuant en une seule analyse. Etant donné que l'étude de la recherche de biomarqueurs de résistance au traitement des LBDGC impliquera une quantification sans marquage, une seule analyse par échantillon est préférée afin de diminuer le temps d'analyse pour un échantillon. C'est pour cette raison qu'une seconde comparaison, impliquant cette fois-ci un SG découpé en une seule bande (la bande d'intérêt) et un TG a été effectuée pour deux échantillons biologiques seulement, du fait d'un manque d'échantillons disponibles pour les optimisations. Les analyses ont à cette occasion été réalisées sur un couplage nanoLC-MS/MS avec le spectromètre de masse Impact-HD de BRUKER. Les résultats de cette nouvelle comparaison sont donnés par l'histogramme de la Figure III-5.

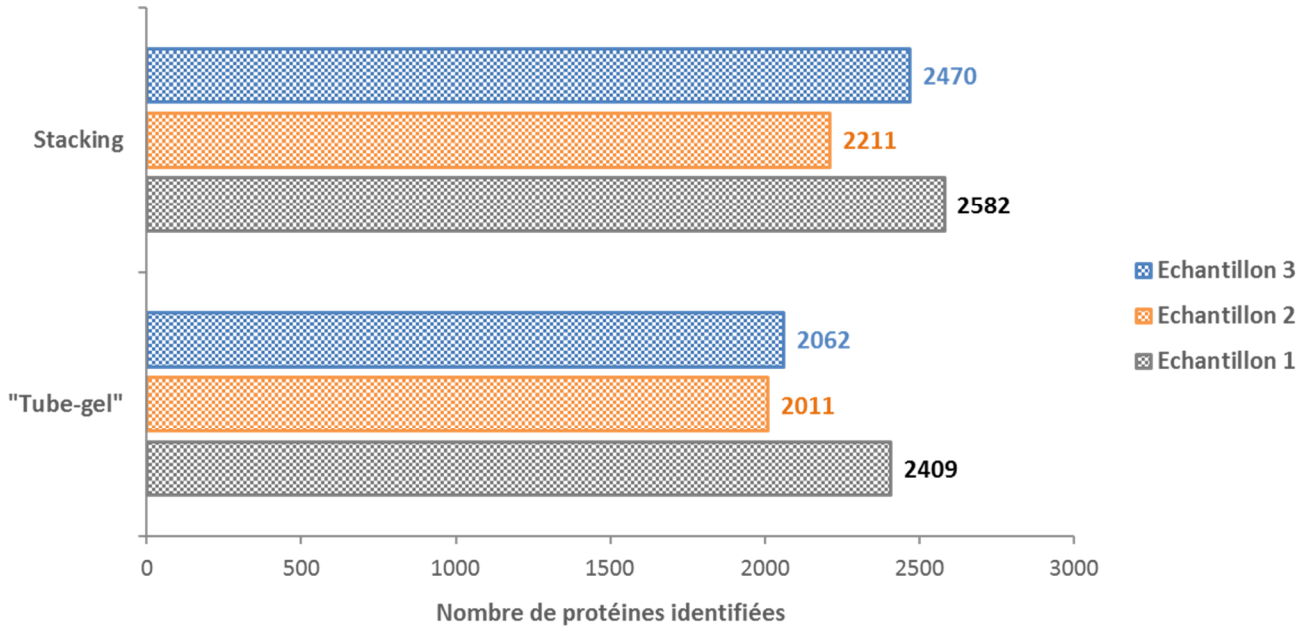


Figure III-4 - Résultats d'identification des protéines du protocole SG découpé en 2 bandes et du protocole TG pour trois extraits protéiques issus de tissus ganglionnaires

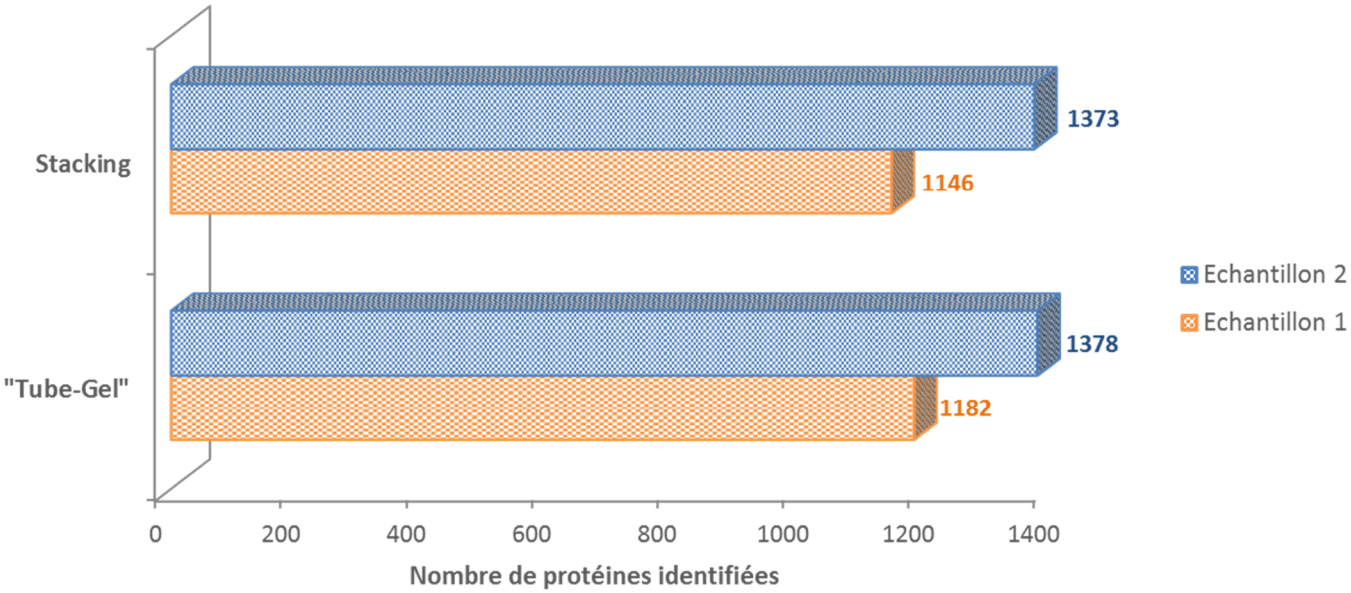


Figure III-5 - Résultats d'identification des protéines du protocole SG découpé en 1 bande et du protocole TG pour deux extraits protéiques issus de tissus ganglionnaires

Ces résultats montrent que le SG et le TG donnent des résultats similaires pour les deux échantillons de protéines extraites de tissus ganglionnaires. Concernant les modifications, tout comme dans l'étude du TG (*Chapitre II-II-1*), davantage de modifications propionamides des cystéines et oxydations des méthionines sont observées pour le TG par rapport au SG.

c. Conclusion

Les résultats de ces optimisations proposent d'extraire les protéines depuis les tissus ganglionnaires à l'aide d'un tampon de lyse à base de SDS et d'un stimuli mécanique généré par un potter, suivis d'une sonication de l'échantillon. Par ailleurs, le TG et le SG donnent des résultats similaires, mais pour des raisons de rapidité de préparation d'échantillons permettant de limiter la manipulation de l'échantillon ainsi que du temps de préparation réduit permettant d'augmenter le nombre de conditions, le TG sera préféré pour l'étude de la recherche de biomarqueurs de résistance des LBDGC. Bien que les peptides oxydés soient généralement exclus pour la quantification sans marquage de type XIC, et que le taux d'oxydations des méthionines est plus élevé dans le cas du TG, cela ne fausse pas les résultats comme le montre l'étude du *Chapitre II-II-1*. En ce qui concerne le nombre plus important de propionamides, ceci ne pose pas de problème particulier si ces modifications sont incluses pour la quantification, en n'oubliant pas de les combiner avec la modification carbamidométhylation des cystéines pour prendre en compte le fait qu'un même peptide dans deux conditions peut ne pas résulter en la même proportion de propionamides ou de carbamidométhylations induites sur les cystéines.

2- Evaluation de l'extraction à partir de tissus frais par rapport à des tissus inclus en paraffine

La disponibilité des échantillons frais congelés est souvent limitée du fait d'un coût de stockage élevé, rendant la collecte pour les analyses protéomiques compliquée[162, 163]. Cette faible disponibilité est également liée au fait que depuis le XIXème siècle, les tissus ne sont pas systématiquement congelés mais systématiquement fixés au formaldéhyde et inclus en paraffine de manière à les préserver et permettre leur stockage à température ambiante pendant de longues années[164-166]. Ces échantillons FFPE (pour « *Formalin-Fixed and Paraffin-Embedded* ») forment de ce fait une librairie de tissus, provenant de nombreuses pathologies, qui représente une manne pour la collecte rapide de larges cohortes d'échantillons, notamment pour des étapes de validation de biomarqueurs. Par ailleurs, si ces tissus FFPE conçus pour les études d'histopathologies permettaient également l'étude des protéines, ils constitueraient un échantillon idéal pour les études translationnelles[166]. La difficulté principale associée à l'utilisation d'échantillons FFPE pour l'analyse protéomique réside dans l'extraction des protéines. En effet, la fixation au formaldéhyde permet de former des ponts méthylènes stables entre les protéines, et l'inversion de cette réaction théoriquement irréversible accordant la possibilité d'extraire les protéines et de les identifier est un réel défi[162, 164-166]. Une première tentative a été réalisée en 1998[167]. Depuis, le nombre d'études de protéomique menées sur des tissus FFPE n'a cessé de croître comme l'illustre la Figure III-6.

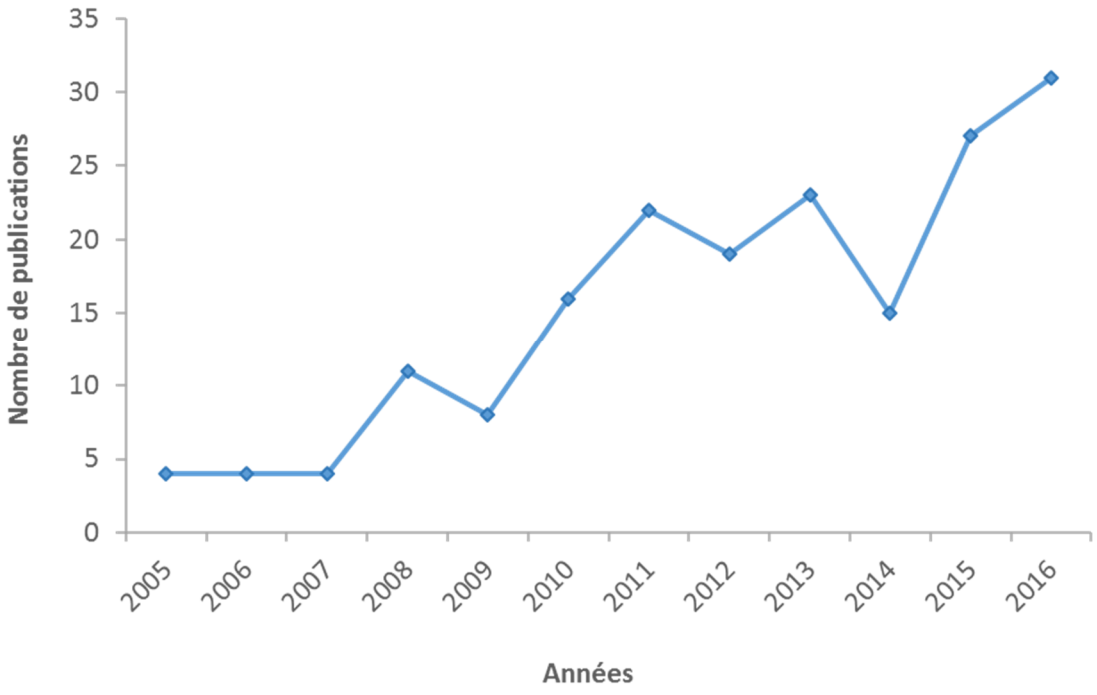


Figure III-6 - Evolution des publications de protéomique menées sur des tissus FFPE de 2005 à 2016
Données PubMed du 9 Août 2017 avec les mots clés « proteomics, FFPE »

Les étapes communes à ces papiers pour l'extraction des protéines à partir de tissus FFPE sont[164, 166] :

- Le **déparaffinage**, qui consiste à solubiliser la paraffine à l'aide de solvants apolaires, puis à réhydrater le tissu avec différentes concentrations d'alcools.

- L'**inversion de la réaction de pontage** au formaldéhyde qui s'effectue par des temps de chauffe de l'échantillon contenu dans un tampon de lyse. Il s'agit généralement d'un chauffage entre 90 et 100°C pendant 20 à 30 minutes, suivi de 60 à 80°C pendant 2 à 3 heures. La composition du tampon de lyse est supposée avoir moins d'influence sur cette étape que la température de chauffage[162].

L'inconvénient majeur de ces tissus FFPE est l'absence de procédure standardisée du processus de fixation[162, 166]. En effet, la température d'inclusion et le temps de fixation sont supposés avoir une influence délétère sur le tissu, mais des études se contredisent encore à ce sujet[164, 166, 168].

a. Optimisation de l'extraction de protéines à partir de tissus inclus en paraffine

Dans l'optique de former de grandes cohortes d'échantillons de ganglions lymphatiques pour une future étape de validation de candidats dans le cadre de l'étude de recherche de biomarqueurs de résistance des LBDGC, une optimisation de l'extraction des protéines à partir de tissus FFPE a été réalisée. De nombreux papiers ont démontré que l'utilisation d'un tampon de lyse à base de SDS permettait d'extraire la plus grande quantité de protéines, probablement du fait de son rôle dual de détergent et de dénaturant qui est supposé concourir au déroulement des protéines lors de l'hydrolyse des ponts méthylènes par les hautes températures[162, 164]. Cependant, la présence d'agents réducteurs dans le tampon d'extraction fait débat quant à son influence sur cette étape d'extraction des protéines à partir de tissus FFPE[164, 166]. Ainsi, nous avons voulu établir si l'ajout d'un agent réducteur tel que le DTT pouvait améliorer l'extraction des protéines de ganglions lymphatiques FFPE par inversement de la réaction de pontage au formaldéhyde à l'aide de trois échantillons différents. Après extraction avec et sans DTT, chaque échantillon a fait l'objet d'une préparation d'échantillons TG. Etant donné que la présence de thiols, comme c'est le cas en présence de DTT, peut inhiber la polymérisation d'acrylamide, l'échantillon 1 a également fait l'objet d'un SG. Le schéma analytique de l'optimisation de cette extraction est détaillé en Figure III-7.

3 ganglions lymphatiques FFPE

2 bains de xylène 5 minutes sous agitation

3 bains d'éthanol (100% - 90% - 70%) 5 minutes

Déparaffinage

Réhydratation

Extraction Avec DTT

Extraction Sans DTT

Tampon de lyse (Tris-HCl pH=6,8, glycérol, SDS 2%, DTT 100mM) 100°C pendant 20min 80°C pendant 2h Sonication

Tampon de lyse (Tris-HCl pH=6,8, glycérol, SDS 2%) 100°C pendant 20min 80°C pendant 2h Sonication

Préparation d'échantillons

Préparation d'échantillons

« Tube-Gel »

+

Gel Stacking *(2 bandes - mélangées)* Pour échantillon 1

« Tube-Gel »

Réduction, Alkylation

Lavages

Digestion

Analyse nanoLC-MS/MS

Nano-ESI – Q-Orbitrap **Q-Exactive +** THERMO FISHER SCIENTIFIC

Nano-LC **Nano-Acquity** WATERS

Traitement de données

Identification + Quantification XIC

Figure III-7 - Schéma analytique de l'optimisation de l'extraction à partir de tissus FFPE

Comparaison des TG

Dans un premier temps, les TG précédés d'une extraction avec et sans DTT des trois échantillons ont été comparés. Le nombre de protéines identifiées pour chaque échantillon est présenté en Figure III-8. Pour les échantillons 1 et 3, le nombre de protéines identifiées est supérieur avec une extraction sans DTT comparée à une extraction avec DTT. Cette tendance n'est cependant pas observée pour l'échantillon 2. Nous notons par ailleurs que le nombre de protéines identifiées en présence de réducteur est stable et répétable sur les trois échantillons, malgré un aspect visuel des TG inhabituel dans ces conditions. Pour un même échantillon, la comparaison de l'extraction avec DTT et sans DTT permet d'observer que 66 à 67 % des protéines identifiées sont communes aux deux types d'extraction. En descendant en granularité et en s'intéressant aux peptides, les mêmes tendances qu'au niveau protéique sont observées, soit davantage de peptides identifiés sans DTT par rapport à une extraction avec DTT, excepté pour l'échantillon 2. La proportion de peptides communs entre les deux extractions est faible (de 27 à 34 %) malgré un bon recouvrement au niveau protéique. En se penchant sur le nombre de coupures manquées au sein de chacune de ces conditions, nous observons davantage de coupures dans l'échantillon 2 extrait sans DTT (+20 %) par rapport au même échantillon extrait en présence de DTT. Etant donné que pour les deux autres échantillons, seulement 6 % de coupures manquées supplémentaires sont observées lorsque l'extraction est effectuée sans DTT, un problème de digestion sur l'échantillon 2 sans DTT peut être pointé, ce qui expliquerait qu'un nombre de protéines identifiées moins important est observé lors d'une extraction sans DTT par rapport en présence de DTT.

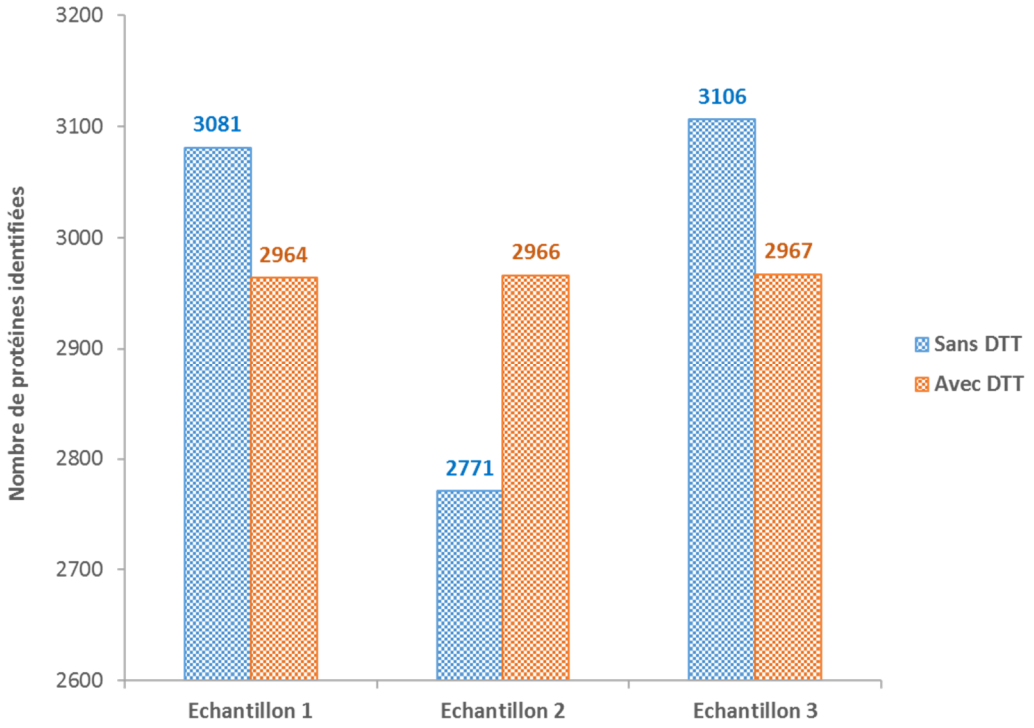


Figure III-8 - Nombre de protéines identifiées avec au moins un peptide unique (par MS/MS et « *matching* ») dans les trois échantillons extraits avec et sans DTT et préparés en TG

Concernant les modifications induites dans les deux types d'extraction, deux fois plus d'oxydations des méthionines sont observées dans les échantillons extraits sans DTT par rapport à ceux extraits avec DTT (Figure III-9). Ceci s'explique par le fait que la présence de DTT permet de réduire les méthionines sulfoxydes. Pour ce qui est des modifications induites sur les cystéines, davantage de modifications propionamides sont générées dans les échantillons extraits en présence de DTT (Figure III-10), ce qui s'explique par le fait qu'en présence de DTT, les ponts disulfures sont réduits et offrent davantage de sites pouvant réagir avec les monomères d'acrylamide que lors d'une extraction sans DTT. Cette réaction avec les monomères d'acrylamide n'est cependant pas complète malgré la réduction des ponts disulfures en présence de DTT, puisque certains peptides à cystéines présentent également des modifications carbamidométhylations, survenant après la polymérisation, soit après le contact avec les monomères d'acrylamide dans le schéma analytique.

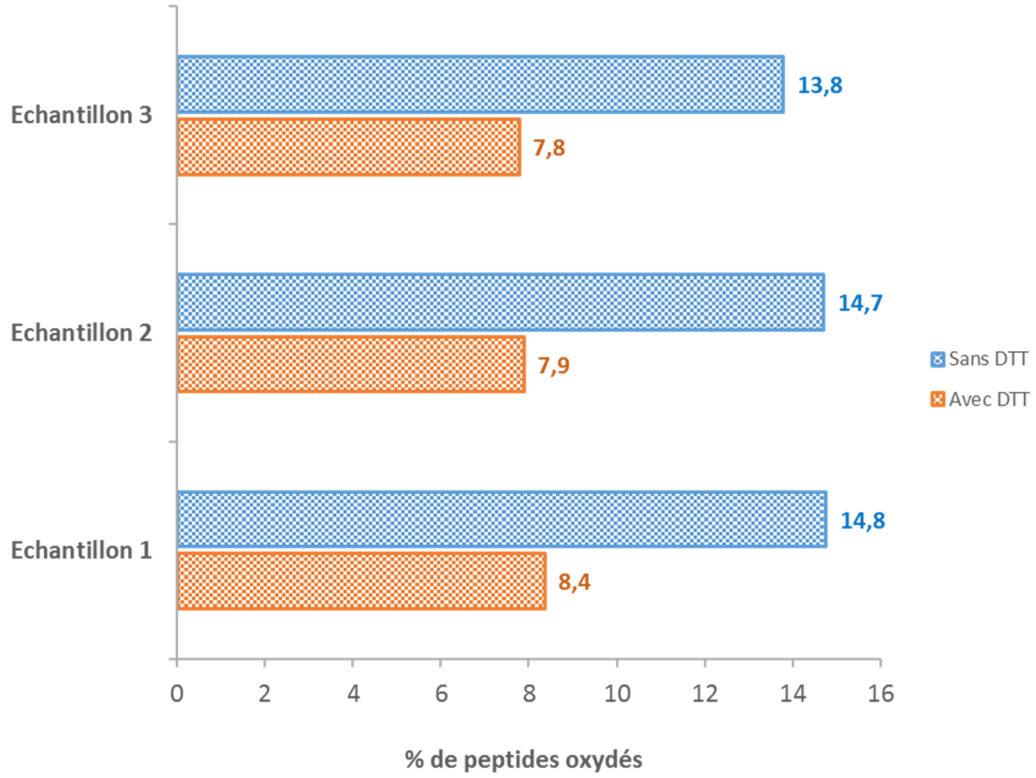


Figure III-9 - Pourcentage de peptides oxydés dans les trois échantillons extraits avec et sans DTT et préparés en TG

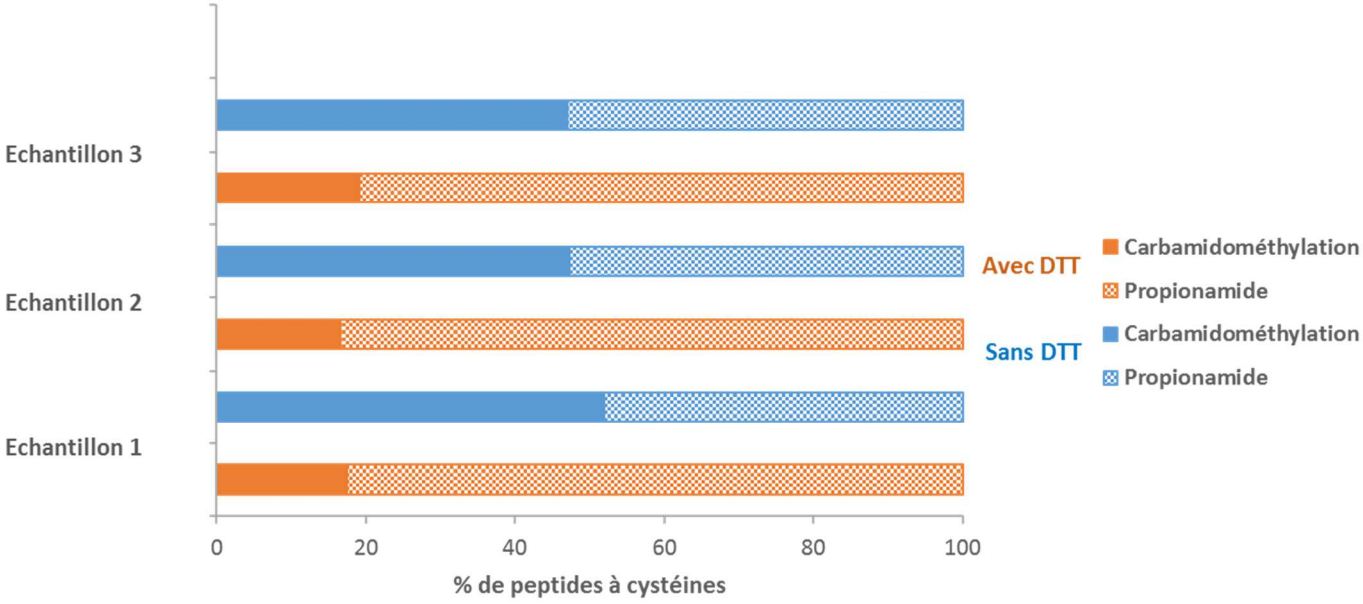


Figure III-10 - Distribution des modifications induites sur les cystéines pour les trois échantillons extraits avec et sans DTT et suivis d'une préparation TG

La distribution du nombre de peptides identifiés par protéine est équivalente entre les deux extractions, avec une médiane de 3 peptides par protéines. Aucune différence n'est observée au niveau de la distribution des protéines extraites en fonction de leur masse moléculaire, comme le

montre la Figure III-11. Enfin, le pourcentage de couverture de séquence est équivalent entre les deux extractions, avec une médiane de 12 %.

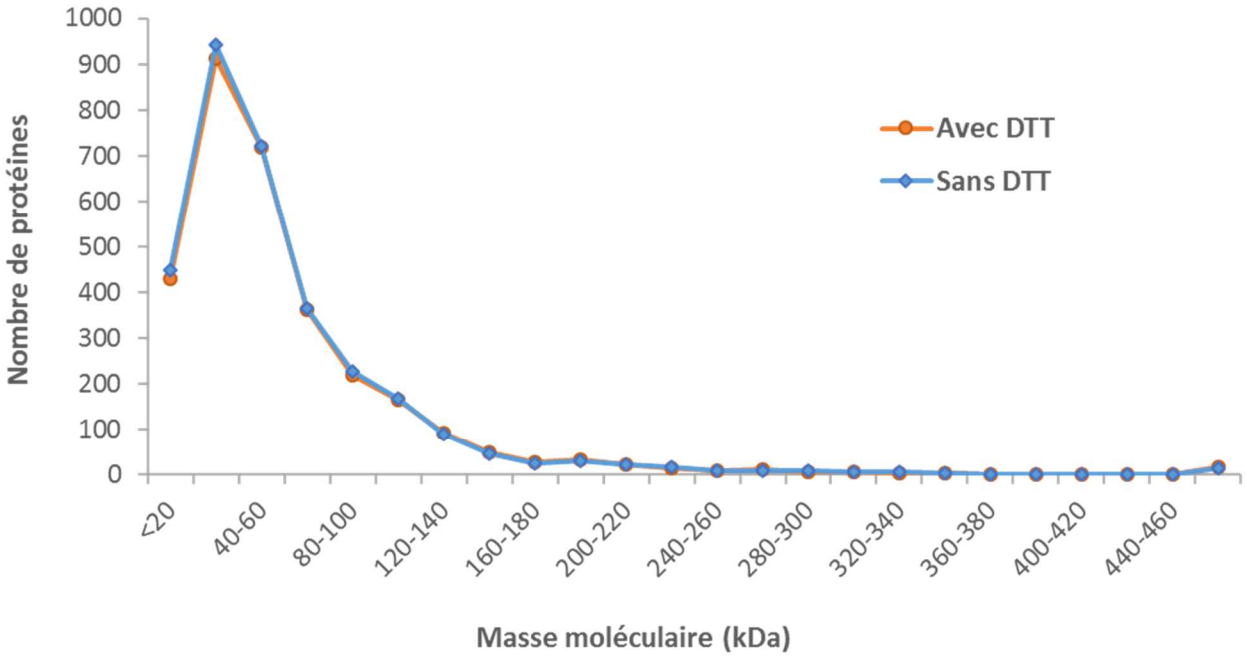


Figure III-11 - Distribution des masses moléculaires des protéines identifiées dans l'ensemble des trois échantillons extraits avec et sans DTT

A l'aide de données de quantification protéique issues de MaxQuant pour les trois réplicats, nous avons pu calculer les coefficients de variation pour chacune des extractions. Ainsi nous avons pu établir les boîtes à moustache de la Figure III-12. Celles-ci indiquent que les deux protocoles semblent stables et répétables avec des médianes de 12 et 13 % pour le TG avec une extraction en présence de DTT et en absence de DTT respectivement. De plus, les coefficients de corrélation entre réplicats sont de l'ordre de 0,98 pour l'extraction sans DTT et de 0,99 pour l'extraction avec DTT. Ainsi, ces deux protocoles d'extraction suivis d'une préparation d'échantillons TG permettent d'obtenir des résultats répétables.

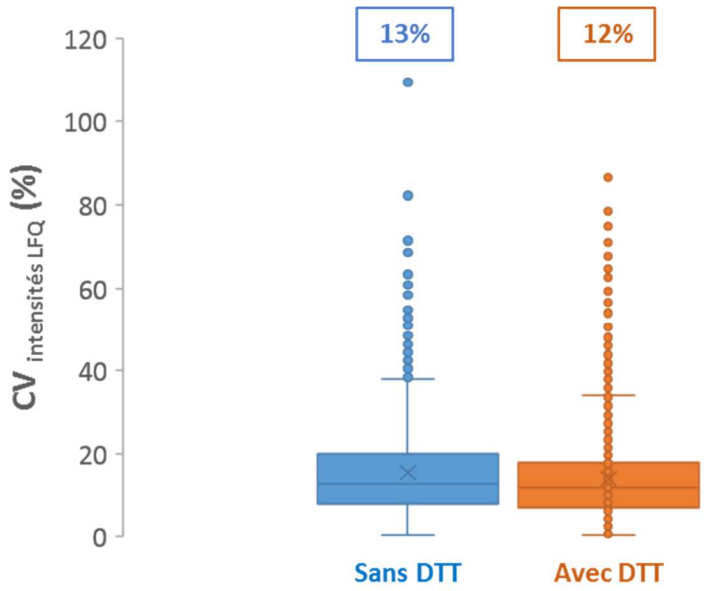


Figure III-12 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des valeurs d'intensités LFQ pour chacune des protéines identifiées dans chaque réplicat, pour l'extraction en présence de DTT et l'extraction en l'absence de DTT.
Les valeurs encadrées correspondent aux CV médians pour chaque extraction

Pour finir, les valeurs d'intensités protéiques ont permis d'établir des courbes de corrélation entre le protocole d'extraction en présence et en absence de DTT pour un même échantillon (Figure III-13). Les coefficients de corrélation supérieurs à 0,9 indiquent que malgré une extraction différente, les valeurs d'intensités corrèlent plutôt bien. Par ailleurs, ces valeurs moins bonnes que lors de la comparaison des réplicats peut s'expliquer par le fait que certaines protéines (toujours les mêmes) ont l'air d'être présentes en quantité plus importante dans les échantillons extraits en présence de DTT. Il s'agit de protéines abondantes dans ces échantillons, qui ne montrent pas de grand intérêt biologique pour la recherche de biomarqueurs de résistance des LBDGC, comme la myosine, le collagène et la laminine (en orange sur la Figure III-13). L'extraction en présence de DTT semble favoriser l'extraction de ces protéines, présentes en grande quantité dans ce type d'échantillon. Une hypothèse qui pourrait expliquer ce phénomène serait que ces protéines, de grande taille, présentent de nombreuses cystéines dans leur séquence. Ainsi, l'ajout de DTT dans le tampon permettrait de mieux les déplier en réduisant les ponts disulfures, et les dénaturer, ce qui faciliterait leur extraction. Ces protéines sont d'ailleurs identifiées avec davantage de peptides dans les échantillons extraits en présence de DTT qu'en absence de DTT (avec 27 et 3 peptides uniques respectivement pour la laminine LAMC-1 par exemple).

En conclusion, l'extraction sans DTT suivie d'un TG permet d'identifier le plus grand nombre de protéines, et ce, avec un nombre de coupures manquées et d'oxydations des méthionines plus élevé

qui n'impacte pas les résultats d'études de protéomique quantitative. Ainsi, la préparation d'échantillons TG précédée d'une extraction de protéines à partir de ganglions lymphatiques FFPE sans DTT est à privilégier pour ce type d'études.

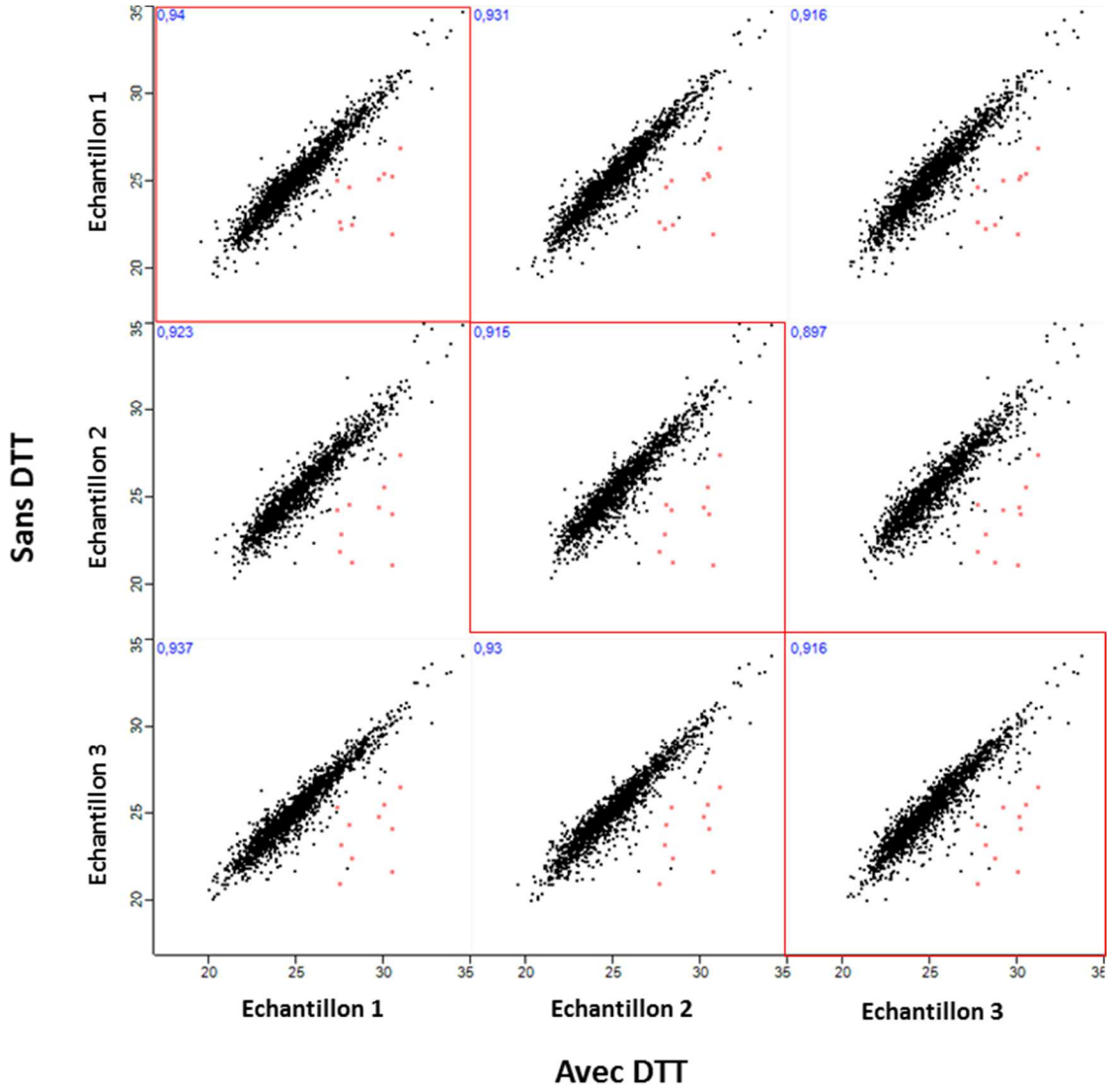


Figure III-13 - Courbes de corrélation établies à l'aide des valeurs d'intensités LFQ pour chaque protéine pour les trois échantillons extraits avec et sans DTT et suivis d'un TG.
Les encadrés rouges indiquent la corrélation pour un même échantillon entre l'extraction en présence de DTT et l'extraction en absence de DTT. Les points oranges correspondent aux mêmes protéines dans toutes les corrélations

Comparaison des TG suivant l'extraction avec et sans DTT avec le gel « Stacking » suivant l'extraction avec DTT

En parallèle, l'extraction sans DTT suivie d'un TG a été comparée à une extraction avec DTT suivie d'un TG et d'un SG. Concernant le nombre de protéines, une centaine de protéines supplémentaires sont

identifiées avec le TG précédé d'une extraction sans DTT par rapport au TG et SG suivant une extraction avec DTT (Figure III-14). Le nombre de protéines identifiées par ces deux dernières préparations d'échantillons suivant l'extraction au DTT est équivalent, ce qui semble témoigner que le DTT n'a pas vraiment influé sur la polymérisation du TG, ou tout du moins sur le piégeage des protéines dans le TG.

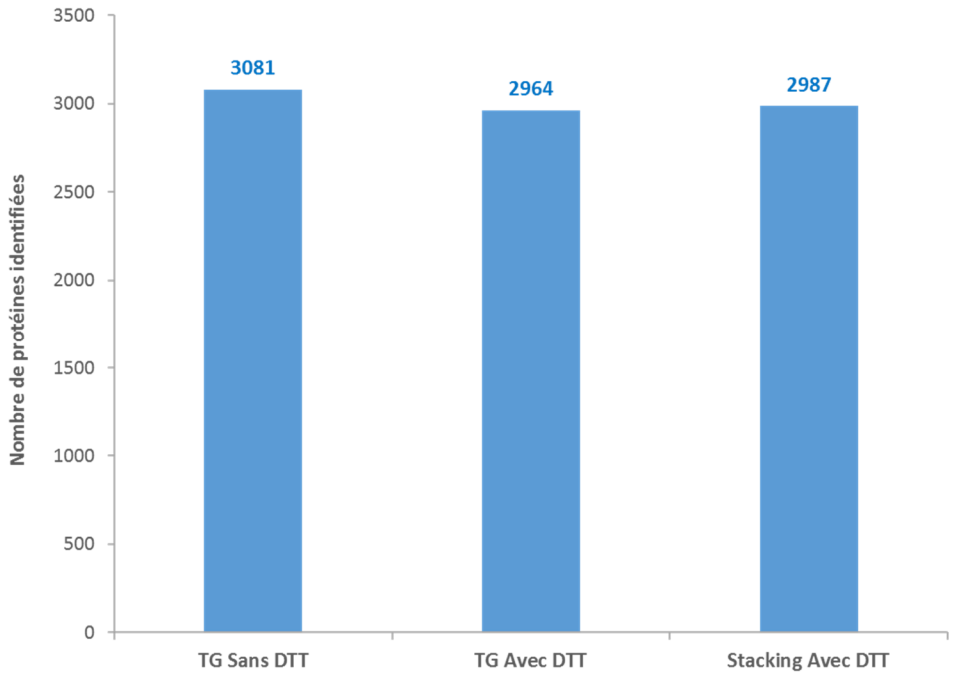


Figure III-14 - Nombre de protéines identifiées dans l'échantillon 1 extrait sans DTT et suivi d'un TG, ainsi que dans l'échantillon 1 extrait sans DTT suivi d'un TG et d'un SG

Lorsque ces trois conditions sont comparées deux à deux, l'extraction sans DTT suivie d'un TG montre 68 % de protéines communes avec le TG précédé d'une extraction en présence de DTT, et 72 % de protéines communes avec le gel « *Stacking* » précédé d'une extraction avec DTT. Par ailleurs, les deux préparations d'échantillons suivant l'extraction en présence de DTT présentent 68 % de protéines communes, reflétant qu'aucune différence majeure n'est observée entre le TG et le SG pour cette extraction.

Concernant les modifications, comme observé auparavant, l'extraction sans DTT génère davantage d'oxydations des méthionines que l'extraction avec DTT. Cependant, le TG et le SG d'une même extraction présentent un même taux d'oxydations (Figure III-15). Pour ce qui est des modifications induites sur les peptides à cystéines, la Figure III-16 permet d'observer que, comme attendu, l'extraction avec DTT suivie d'un TG génère le plus grand nombre de modifications propionamides pour les raisons évoquées précédemment. L'extraction avec DTT suivie d'un SG génère moins de cette

même modification que l'extraction sans DTT suivie d'un TG, et ceci s'explique facilement par le fait que le contact des protéines avec les monomères d'acrylamide est plus direct dans une préparation d'échantillons TG que dans un SG.

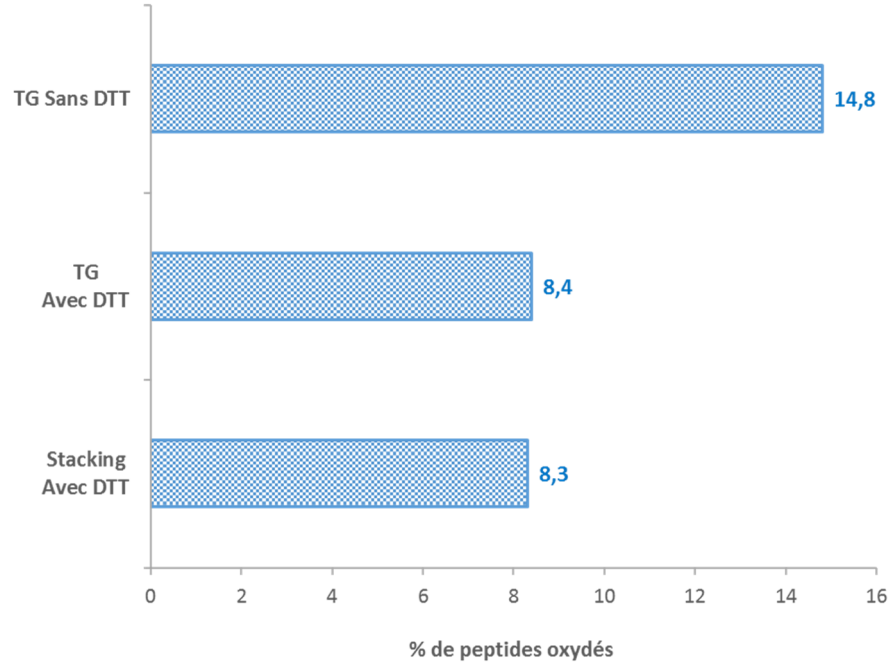


Figure III-15 - Pourcentage de peptides oxydés pour l'échantillon 1 extrait sans DTT et suivi d'un TG, ainsi que dans l'échantillon 1 extrait sans DTT suivi d'un TG et d'un SG

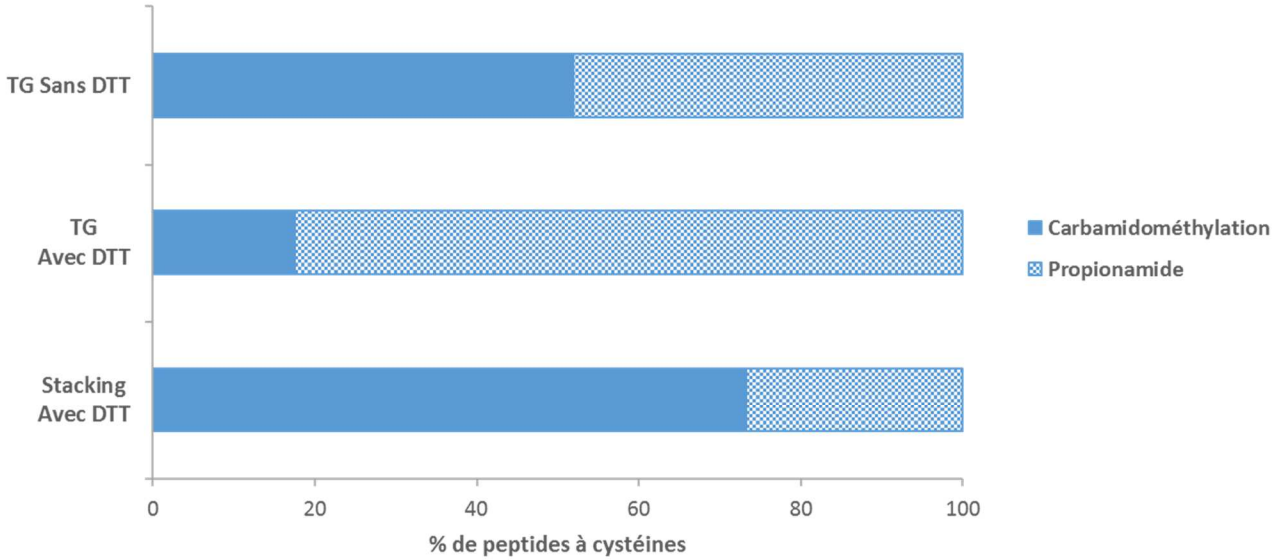
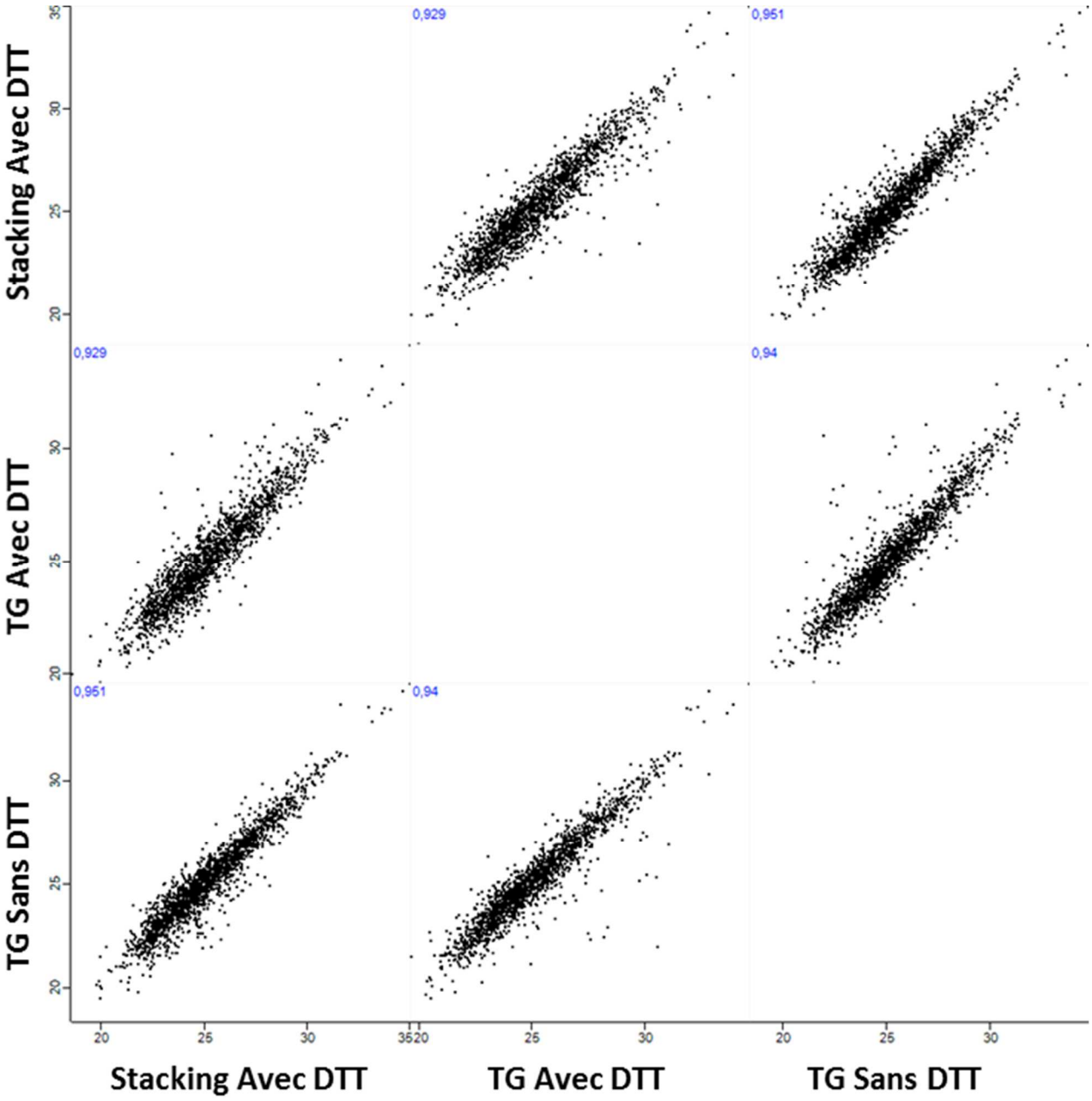


Figure III-16 - Distribution des modifications induites sur les cystéines pour l'échantillon 1 extrait sans DTT et suivi d'un TG, ainsi que dans l'échantillon 1 extrait sans DTT suivi d'un TG et d'un SG

Pour finir, à l'aide de données de quantification protéique, des courbes de corrélation ont été établies (Figure III-17). Ces données montrent que le SG avec une extraction en présence de DTT semble

davantage corréler avec la préparation d'échantillons TG précédée d'une extraction sans DTT. Ainsi, même si le nombre d'identifications est relativement équivalent entre le TG et le SG pour une extraction avec DTT, le SG devrait tout de même être préféré au TG puisqu'il corrèle davantage avec les TG précédés d'une extraction en absence de DTT (0,95) qu'avec les TG en présence de DTT (0,93), et que les TG avec et sans DTT entre eux (0,94). Ceci peut être le reflet discret de l'incompatibilité de



la polymérisation d'un TG en présence de DTT.

Figure III-17 - Courbes de corrélation établies à l'aide des valeurs d'intensités LFQ pour chaque protéine avec les coefficients de corrélation associés pour l'échantillon 1 extrait sans DTT et suivi d'un TG, ainsi que pour l'échantillon 1 extrait sans DTT suivi d'un TG et d'un SG

b. Evaluation de l'extraction de protéines à partir de tissus FFPE par rapport à l'extraction à partir de tissus frais congelés

Comme évoqué auparavant, les tissus frais ne sont pas systématiquement congelés et sont de ce fait pas toujours disponibles. C'est pour cette raison que les tissus FFPE apparaissent comme étant une alternative aux échantillons frais. Cependant, la question se pose de savoir si le protéome de tissus FFPE est proche, voire identique à celui de tissus frais. Ainsi, la comparaison de tissus FFPE et frais est primordiale pour assurer la fiabilité de l'usage de tissus FFPE pour l'analyse protéomique. Des études comparant ces deux types de tissus ont déjà été menées par différentes équipes, cependant l'analyse des tissus frais était souvent indépendante de celle des tissus FFPE[163]. Dans le cas des ganglions lymphatiques pour la recherche de biomarqueurs de résistance au LBDGC, nous avons pu disposer de 8 échantillons FFPE pour lesquels une partie du ganglion avait été congelée en raison d'une quantité de matériel suffisante. A l'aide de ces 16 échantillons, nous avons pu comparer l'extraction de protéines à partir de tissus frais à celle à partir de l'équivalent FFPE au sein d'une même étude, d'une même séquence d'analyses. Le schéma analytique utilisé est détaillé en Figure III-18.

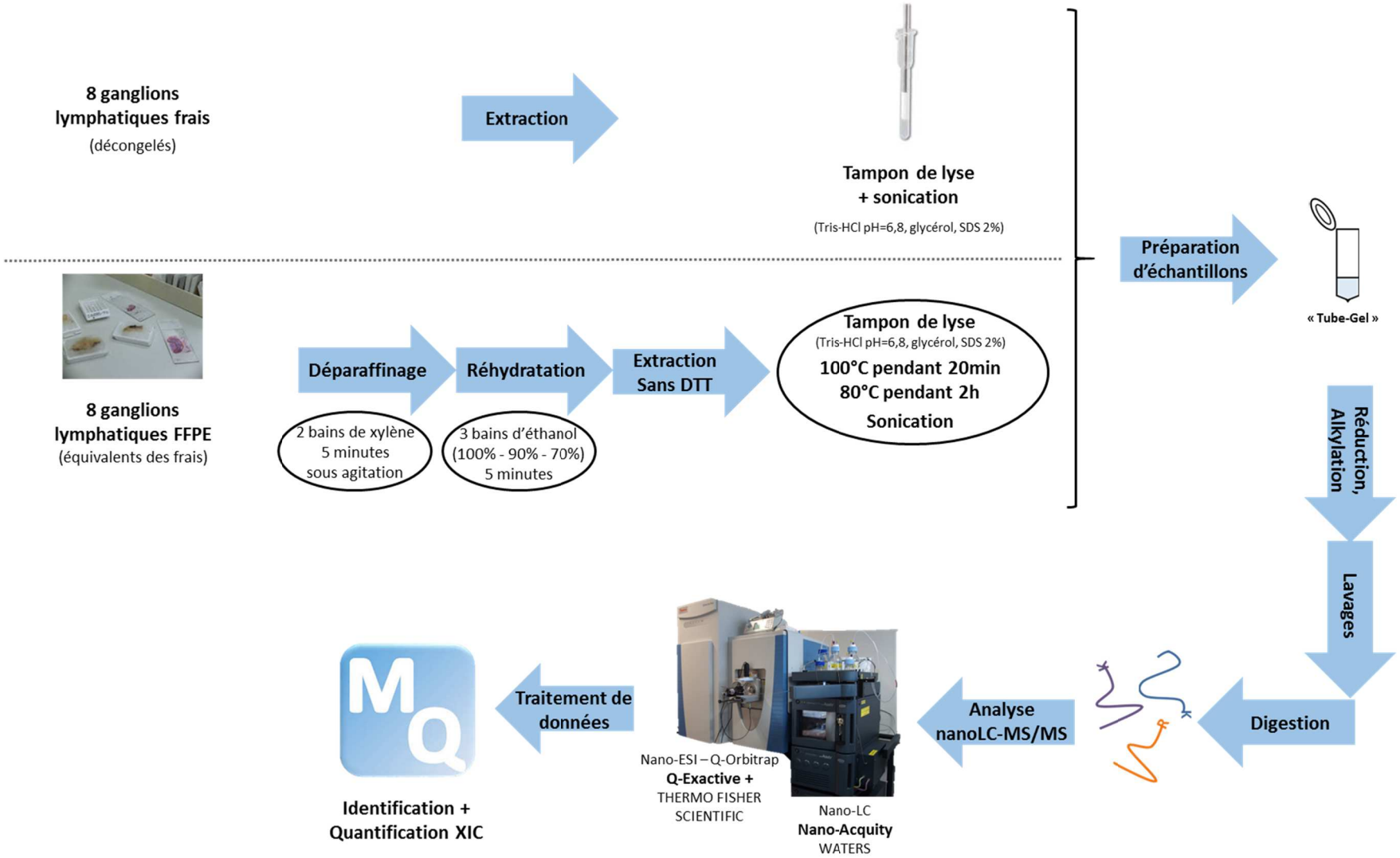


Figure III-18 - Schéma analytique de l'étude consistant à comparer l'extraction de protéines à partir de tissus frais à celle à partir de tissus FFPE correspondants

Le nombre de protéines identifiées dans les 8 échantillons de ganglions lymphatiques frais décongelés est toujours supérieur à celui des ganglions lymphatiques FFPE (en moyenne 10 % d'identifications en plus), comme l'illustre la Figure III-19. Les mêmes tendances sont observées au niveau peptidique.

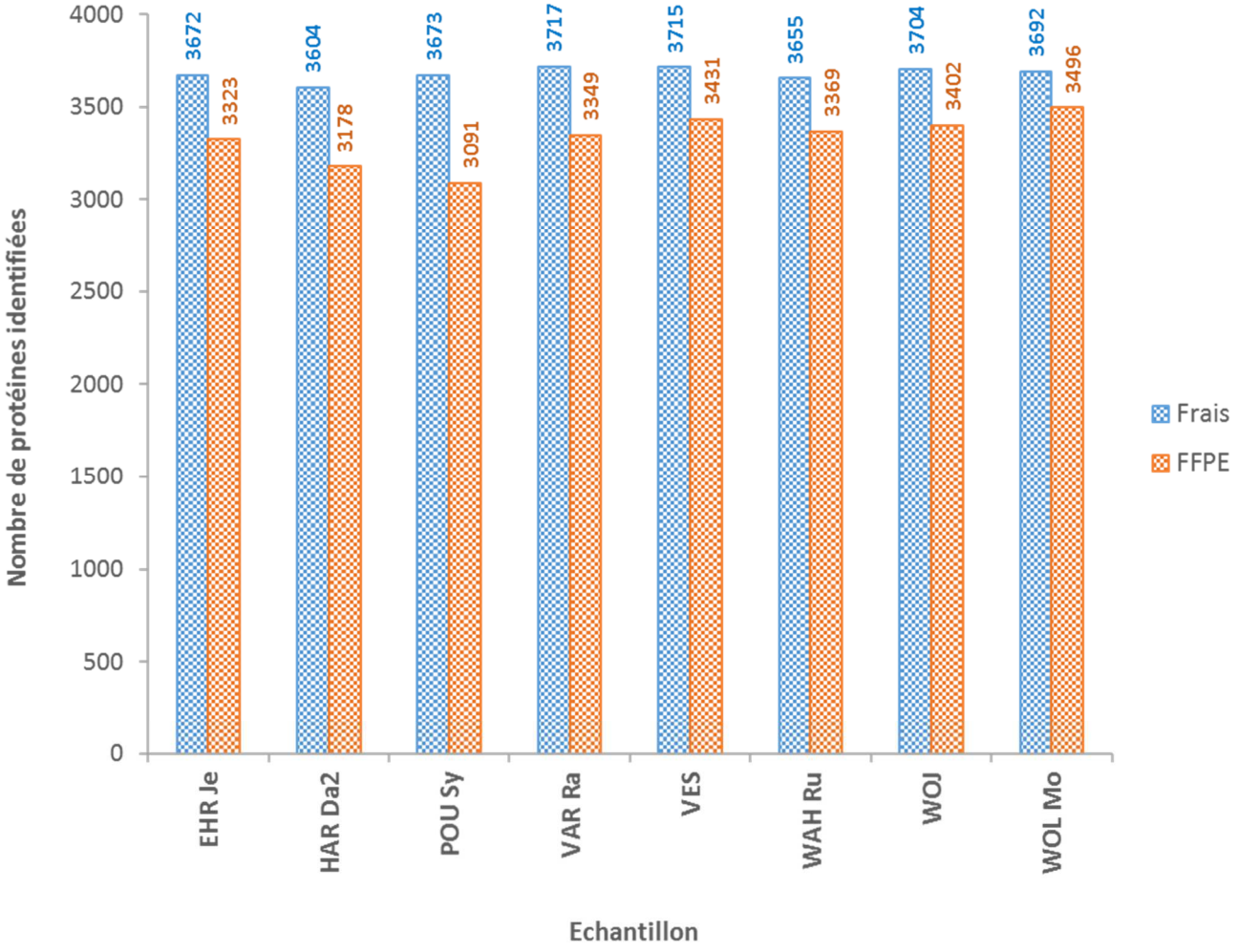


Figure III-19 - Nombre de protéines identifiées avec au moins un peptide unique dans les huit échantillons frais décongelés et leurs équivalents FFPE

La proportion de protéines communes entre échantillon frais et FFPE s'étend de 56 à 71 % pour les huit ganglions, alors que pour les peptides, le recouvrement est plus faible (de 31 à 41 %). Ces valeurs, du même ordre de grandeur qu'un réplicat d'analyse au niveau protéique, permettent de dire que les tissus frais et FFPE permettent d'obtenir des résultats d'identification de protéines relativement équivalents. En s'intéressant à la distribution du nombre de peptides ayant permis l'identification de chaque protéine, nous pouvons observer que pour l'extraction de tissus frais, les protéines sont identifiées avec une médiane de 4 peptides, alors que pour l'extraction de tissus FFPE, les protéines sont identifiées avec une médiane de 3 peptides impactant ainsi le pourcentage médian de couverture de séquence (13 et 11 % pour les tissus frais et les tissus FFPE respectivement). Cette légère différence peut être due au fait que dans les échantillons FFPE, les protéines sont pontées et que l'inversion de

cette réaction de pontage au formaldéhyde n'est pas totale. Ces résultats sont en adéquation avec ceux de la littérature[162, 163].

En ce qui concerne les modifications induites lors de l'extraction de tissus frais et FFPE suivie d'une préparation d'échantillons TG, aucune réelle tendance n'est observée quant aux oxydations des méthionines (Figure III-20). Cependant, au niveau des modifications induites sur les cystéines, l'extraction à partir des tissus frais a tendance à générer davantage de modifications propionamides quand couplée à une préparation d'échantillons TG (Figure III-21).

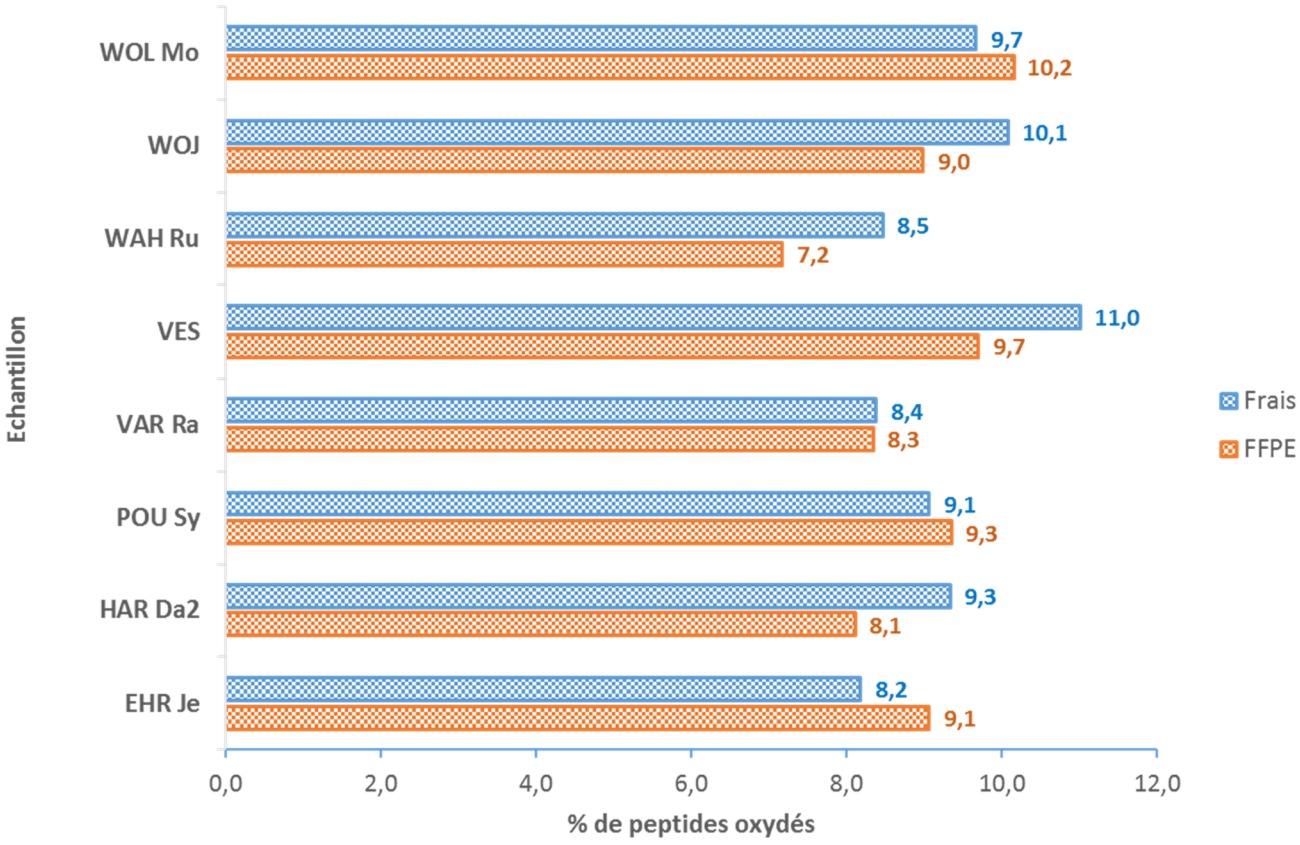


Figure III-20 – Pourcentage de peptides oxydés pour l'extraction des huit échantillons frais et leur équivalent FFPE suivie d'une préparation d'échantillons TG

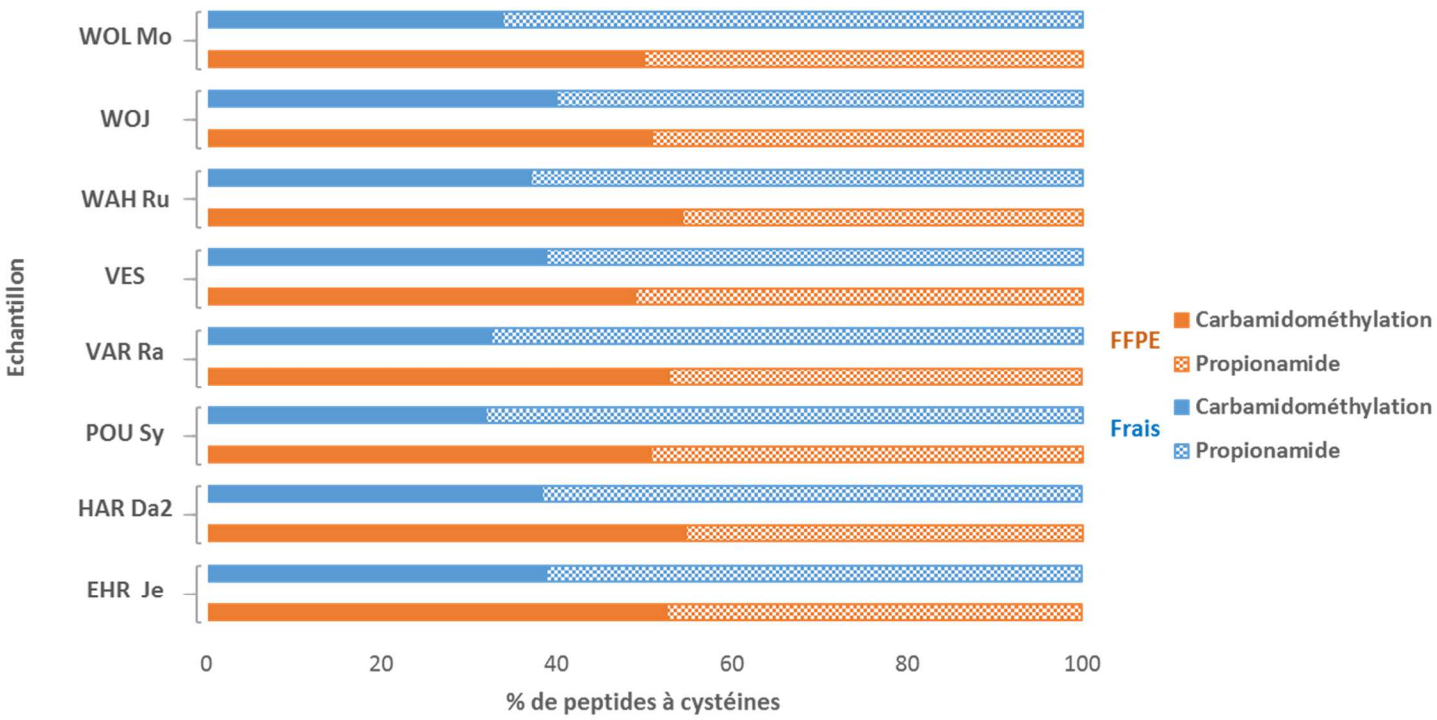


Figure III-21 - Distribution des modifications induites sur les cystéines pour l'extraction des huit échantillons frais et leur équivalent FFPE suivie d'une préparation d'échantillons TG

La distribution des masses moléculaires des protéines identifiées dans au moins un des huit échantillons pour les tissus frais et les tissus FFPE sont similaires, démontrant ainsi que l'inclusion en paraffine n'influence pas l'extraction des protéines en fonction de leur taille, contrairement à ce qui a été suggéré par Alessandro TANCA *et collaborateurs*[163] (Figure III-22). Cette même étude a montré que davantage de coupures manquées étaient observées dans les extraits protéiques de FFPE par rapport à ceux issus de tissus frais. Dans nos mains, 2,5 % de coupures ont été omises dans les échantillons FFPE par rapport aux frais. Par ailleurs, nous n'avons pas observé de différence quant à la distribution des acides aminés entre les deux types de tissus.

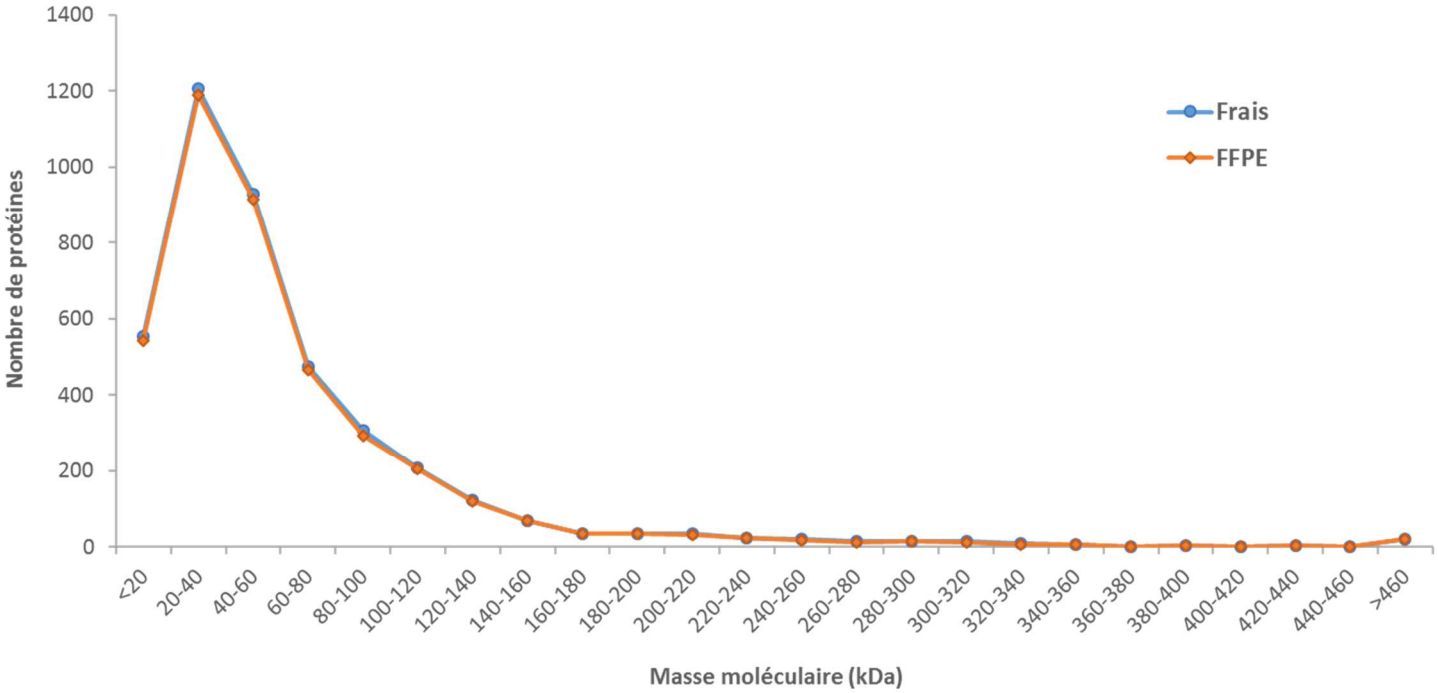


Figure III-22 - Distribution des masses moléculaires des protéines extraites

Pour finir, à partir des données de quantification protéique issues de MaxQuant, nous avons établi des courbes de corrélation (Figure III-23). La diagonale en surbrillance permet de distinguer la corrélation frais-FFPE pour un même échantillon. Les coefficients de corrélation s'étendent de 0,83 à 0,95, reflétant une corrélation acceptable pour des échantillons ayant subi des traitements pré-analytiques très différents (les FFPE ont été largement manipulés par rapport aux échantillons congelés).

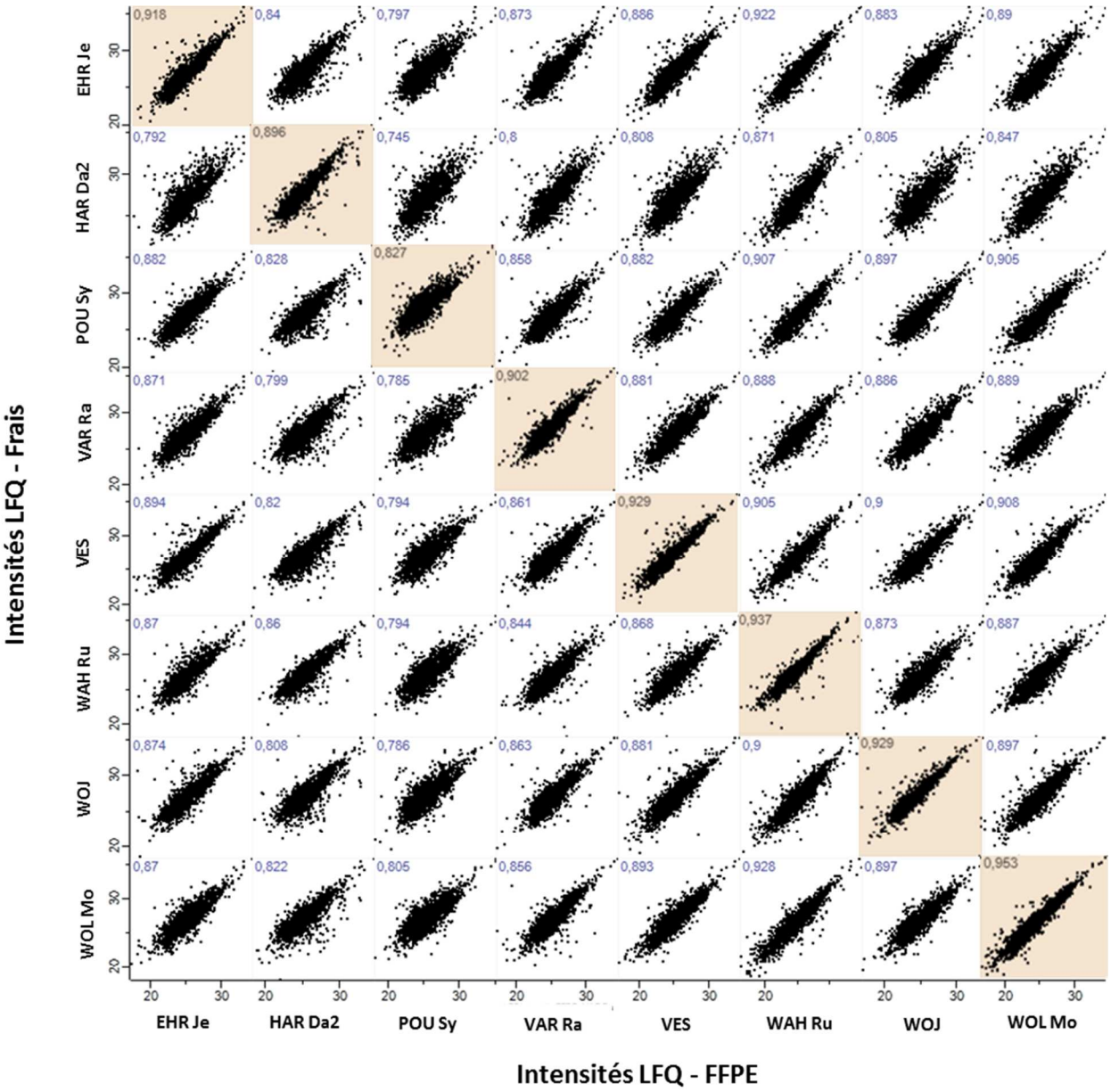


Figure III-23 - Courbes de corrélation établies à l'aide des valeurs d'intensités LFQ pour chaque protéine pour chacun des huit échantillons extraits à partir de tissus frais et de tissus FFPE avec les coefficients de corrélation associés.
Les courbes en surbrillance représentent la corrélation entre les valeurs obtenues après extraction du tissu frais et du tissu FFPE provenant d'un même ganglion lymphatique

c. Conclusion

Ces différentes optimisations d'extractions à partir de ganglions lymphatiques ont permis de rendre compte que l'extraction en présence et en l'absence de DTT donne des résultats d'identification plus ou moins similaires. Cependant, une extraction sans DTT sera préférée lorsqu'une préparation d'échantillons TG sera souhaitée, notamment lors de l'analyse de grandes séries d'échantillons, afin d'éviter tout problème de polymérisation du TG. Par ailleurs, l'extraction de protéines à partir d'échantillons FFPE (sans DTT) permet d'obtenir des résultats relativement proches de ceux obtenus par une extraction à partir d'échantillons frais, tenant compte du fait que les FFPE ont subi davantage

de traitements de pré-stockage que les tissus frais. Ainsi, les tissus FFPE peuvent être employés pour des études de validation de candidats biomarqueurs de résistance au traitement dans les LBDGC nécessitant de grandes cohortes d'échantillons.

IV- Optimisation d'une préparation d'échantillons pour l'étude d'un protéome urinaire

L'urine est un fluide biologique, facilement accessible, sécrété par les reins après filtration du plasma. Elle peut être obtenue en grande quantité par des techniques non invasives, et peut être aisément collectée sur plusieurs jours[169]. C'est pourquoi elle représente un matériel de départ attractif pour la recherche de biomarqueurs de diverses pathologies liées aux conditions rénales[169-172]. Cette matrice a cependant souvent été négligée pour les études de protéomique en raison d'une quantité de protéines supposée très insuffisante. C'est au début des années 2000 que des papiers ont été publiés prouvant la présence d'une quantité de protéines dans les urines compatible avec l'analyse protéomique, la rendant de fait attrayante pour ce type d'échantillons[169].

Diverses caractéristiques relatives à la complexité du milieu rendent l'analyse protéomique de l'urine compliquée, comme une large gamme dynamique, une haute variabilité autant inter-individus qu'intra-individus inhérente à des facteurs tels que le régime, les traitements médicamenteux ou le style de vie, mais aussi la présence en concentrations variables de sels, de contaminants ou encore une diversité de pH pouvant notamment réduire l'efficacité de digestion[169-171, 173, 174]. Ces propriétés induisent l'utilisation de méthodes d'extraction et de purification des protéines permettant de réduire le volume et de concentrer les protéines, et de retirer les interférents (sels et débris cellulaires). Il s'agit généralement de procédés multi-étapes pouvant introduire des biais non souhaités pour l'analyse protéomique quantitative sans marquage.

L'étude du protéome urinaire de patients greffés rénaux est apparue comme particulièrement intéressante pour suivre l'état des reins de tels patients, dès lors qu'il est actuellement réalisé à l'aide d'une biopsie, beaucoup plus invasive qu'un recueil d'urine. Ainsi, dans le cadre du projet de recherche de biomarqueurs de suivi du greffon de patients greffés rénaux, mené en binôme avec le Dr David Marx, l'optimisation d'une préparation d'échantillons permettant l'analyse de protéines urinaires par LC-MS/MS a été effectuée. Etant donné que l'objectif de ce projet réside dans l'application en routine d'un diagnostic de suivi, une méthode d'extraction de protéines et de préparation d'échantillons simple, rapide et robuste pour un usage au quotidien est désirée. Parmi les méthodes les plus utilisées dans le domaine de la protéomique urinaire, nous distinguons :

- Les **méthodes de précipitation**, qui permettent de se débarrasser des sels et autres molécules interférentes. Elles entraînent l'emploi de divers solvants organiques, même si la précipitation au TCA reste la plus utilisée dans le cas de l'urine[169, 170]. Ces méthodes impliquent néanmoins

une resolubilisation du culot souvent compliquée, et une grande manipulation de l'échantillon rendant ces stratégies difficilement applicables à plus grande échelle.

- L'**ultrafiltration**, comme le FASP, qui nécessite un volume d'échantillon faible. A l'aide d'une membrane de cellulose présentant un seuil de coupure de 10 ou de 30 kDa, cette technique permet de retenir les protéines au-dessus du filtre et de retirer les petites molécules telles que les sels, les réducteurs et alkylants, ainsi que les lipides et les acides nucléiques. Elle permet également de digérer les protéines directement sur le filtre, restreignant la manipulation de l'échantillon au niveau du filtre, ce qui limite l'introduction de variabilités[170, 171]. Des développements permettent aujourd'hui l'application de la méthode FASP sur des plaques 96 puits[175, 176], particulièrement attractives pour l'application en routine à grande échelle [177].

D'autres méthodes existent, mais sont peu employées comme l'**ultracentrifugation différentielle**, ou la **dialyse** qui nécessite souvent de grands volumes d'échantillon et qui est particulièrement chronophage, et donc pas envisageable pour une application à haut débit[169]. D'après la littérature, les méthodes de précipitation et le FASP permettent d'obtenir des couvertures du protéome différentes, mais complémentaires[169].

1- Evaluation d'un kit commercial et d'une méthode d'ultrafiltration

Dans l'optique d'une future application en routine d'un diagnostic de suivi des greffés rénaux, ainsi que pour l'étude d'une cohorte de 32 échantillons urinaires par protéomique quantitative sans marquage dans le cadre du projet de recherche de biomarqueurs de suivi de greffe rénale, deux méthodes d'extraction et de préparation d'échantillons urinaires rapides, et paraissant applicables à grande échelle ont été testées. Il s'agit :

- De la méthode **FASP** : développée originellement par MANZA *et collaborateurs*[30], la méthode d'ultrafiltration permettant de retirer les interférents tels que les sels et de digérer les protéines au sein d'un même dispositif a été nommée FASP par le groupe de Mathias MANN[31]. Malgré les avantages évoqués plus haut, la possible défaillance de certains filtres peut entraîner un risque non négligeable de perte totale de l'échantillon. Certaines équipes proposent de tester chaque filtre avant utilisation[178], tandis que d'autres préconisent d'utiliser des filtres à surface horizontale plane plutôt que verticale[179]. Par ailleurs, de nombreuses optimisations ont depuis été effectuées par plusieurs équipes, notamment par le groupe de Mathias MANN qui propose d'augmenter le seuil de coupure à 30 kDa au lieu de 10 kDa pour diminuer les temps de centrifugation, de limiter la quantité de protéines de départ à 100 µg

maximum, ou encore d'utiliser une combinaison d'enzymes comme LysC et trypsine (MED-FASP pour « MultiEnzyme Digestion FASP ») pour obtenir un meilleur rendement de conversion en peptides[180].

- Du **kit iST commercialisé par la société PreOmics**, qui constitue une chambre de réaction close dans laquelle les protéines sont lysées, réduites, alkylées et digérées, et qui permet l'extraction des peptides[33]. Ce kit a été pensé de manière à réaliser un minimum de manipulations de l'échantillon, et peut être opéré sur des plaques 96 puits. Ce type d'approche, impliquant un support avec une phase solide faisant office de réacteur dans lequel l'échantillon est purifié, réduit, alkylé et digéré, gagnent en popularité du fait qu'elles facilitent et accélèrent la préparation d'échantillons.

Ces deux méthodes ont été testées sur un même prélèvement urinaire provenant d'un sujet sain, exempt de débris cellulaires, divisé en six afin de réaliser trois réplicats techniques pour chaque protocole. Ces échantillons ont fait l'objet d'une analyse nanoLC-MS/MS comme le montre la Figure IV-1 résumant le schéma analytique d'optimisation de la préparation d'échantillons pour l'analyse protéomique d'urine.

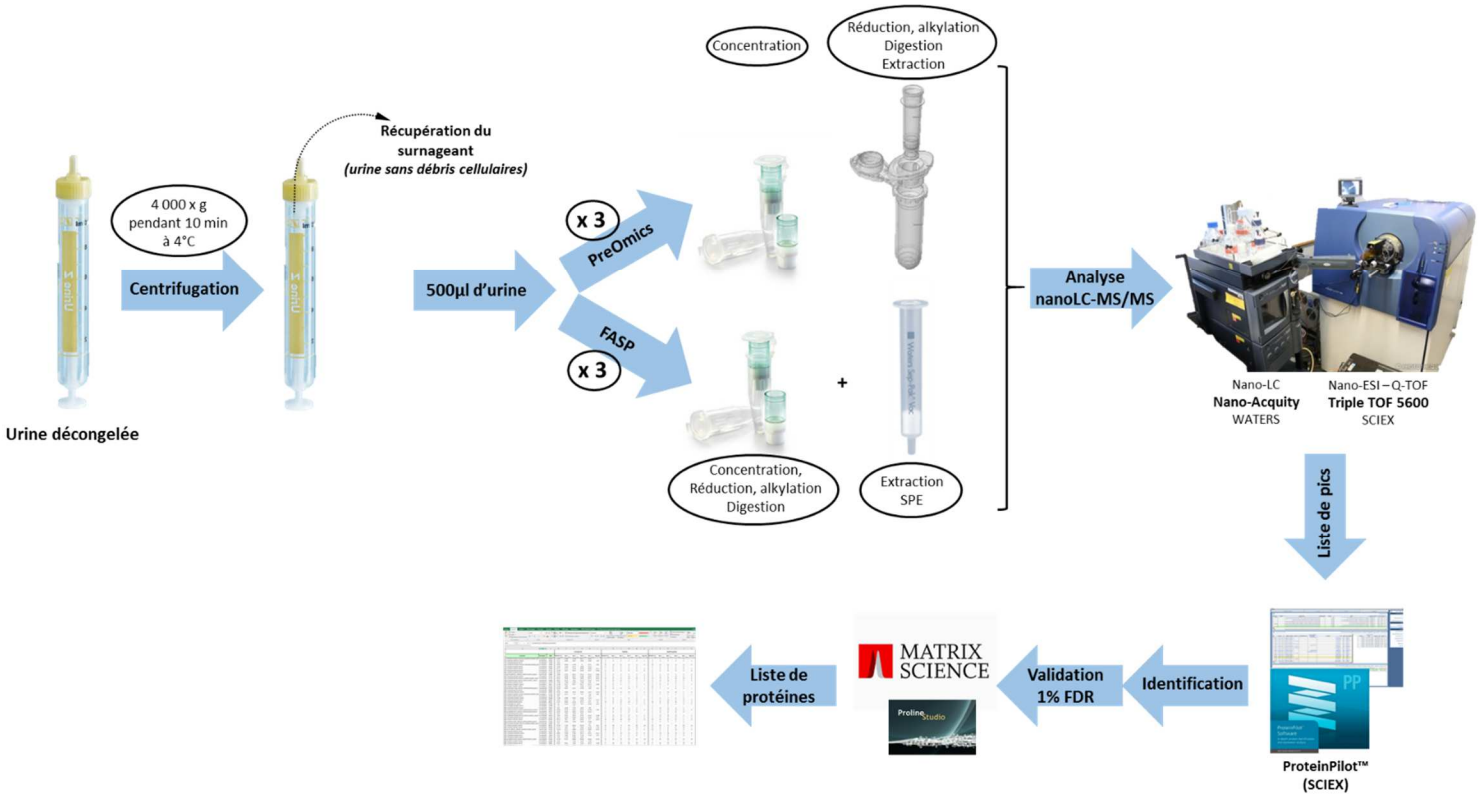


Figure IV-1 - Schéma analytique d'optimisation de la préparation d'échantillons pour l'analyse protéomique d'urine

Le protocole FASP a permis d'identifier davantage de protéines sur l'ensemble des trois réplicats (+ 20 %) que le kit PreOmics, avec des moyennes de protéines identifiées avec au moins un peptide unique de 773 et 624 respectivement. Cependant, le nombre de protéines communes aux deux protocoles reste relativement élevé pour des protocoles différents, soit 57 % comme l'illustre le diagramme de Venn de la Figure IV-2.

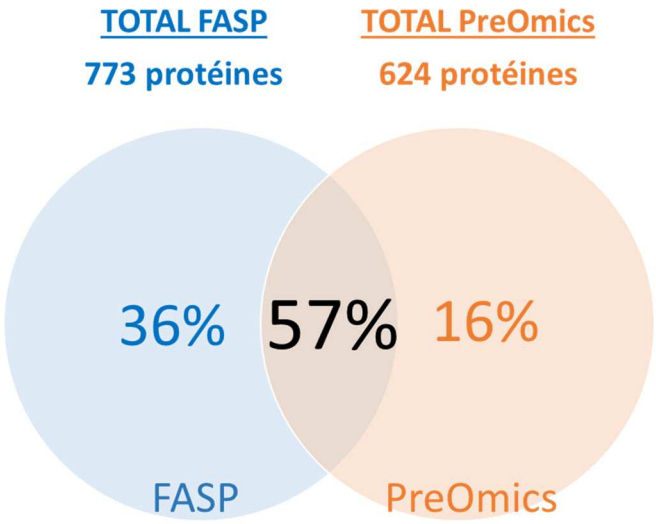


Figure IV-2 - Diagramme de Venn opposant les listes de protéines identifiées avec au moins un peptide unique dans l'ensemble des trois réplicats pour le protocole FASP et pour le protocole commercialisé par PreOmics

Aucun des deux protocoles ne semble favoriser une certaine gamme de taille des protéines comme le montre la Figure IV-3.

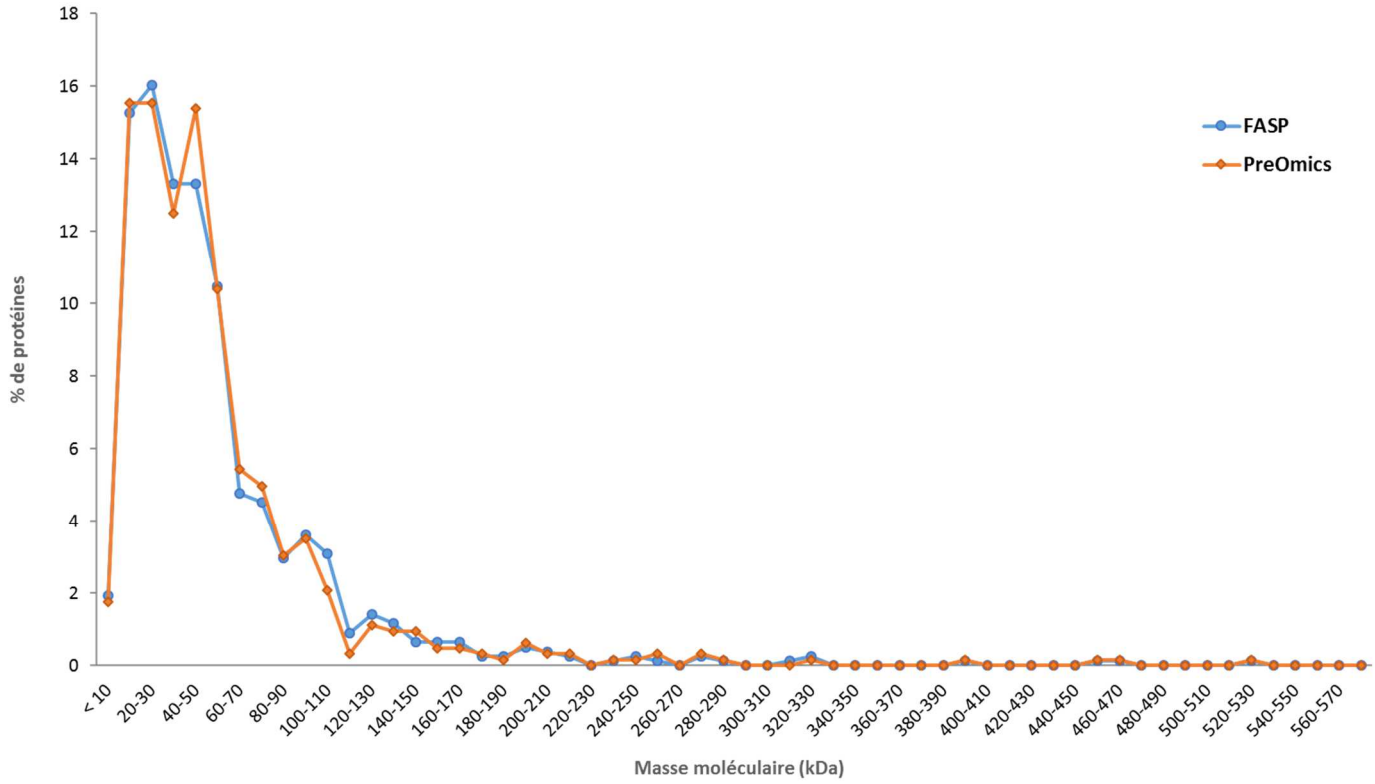


Figure IV-3 - Distribution des masses moléculaires des protéines extraites par les deux protocoles

En ce qui concerne les peptides, la tendance observée est plus marquée que celles des protéines, avec une soixantaine de pourcents de peptides identifiés en plus par le protocole FASP par rapport au protocole commercialisé par PreOmics, comme l'illustre la Figure IV-4. Notons cependant que le FASP génère davantage de coupures manquées (19 %) que le kit iST de PreOmics (7 %), ce qui s'explique par l'utilisation de la trypsine simple dans le cas du FASP, et d'un mélange LysC-trypsine pour le kit iST.

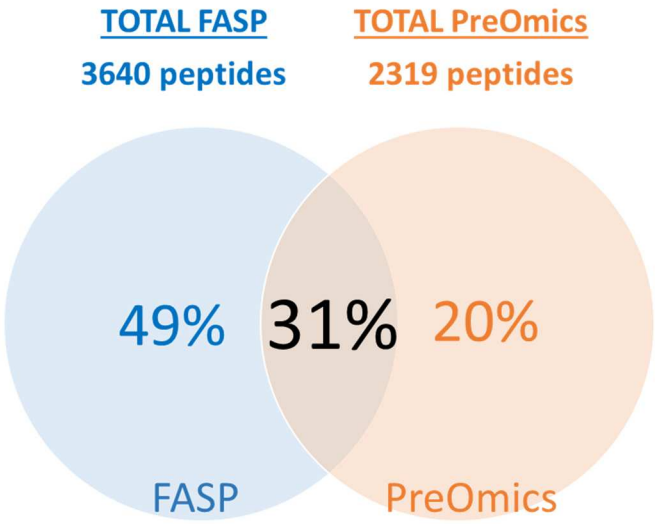


Figure IV-4 - Diagramme de Venn opposant les listes des séquences peptidiques identifiées dans l'ensemble des trois réplicats pour le protocole FASP et pour le protocole commercialisé par PreOmics

En moyenne, chaque protéine est identifiée avec 4,7 peptides pour le protocole FASP et avec 3,7 peptides pour le kit iST. Ce phénomène est renforcé par la Figure IV-5, qui montre que davantage de protéines présentent une couverture de séquence supérieure à 30 % avec le protocole FASP par rapport au protocole PreOmics, sachant que la distribution des protéines selon leur masse moléculaire est similaire pour les deux protocoles.

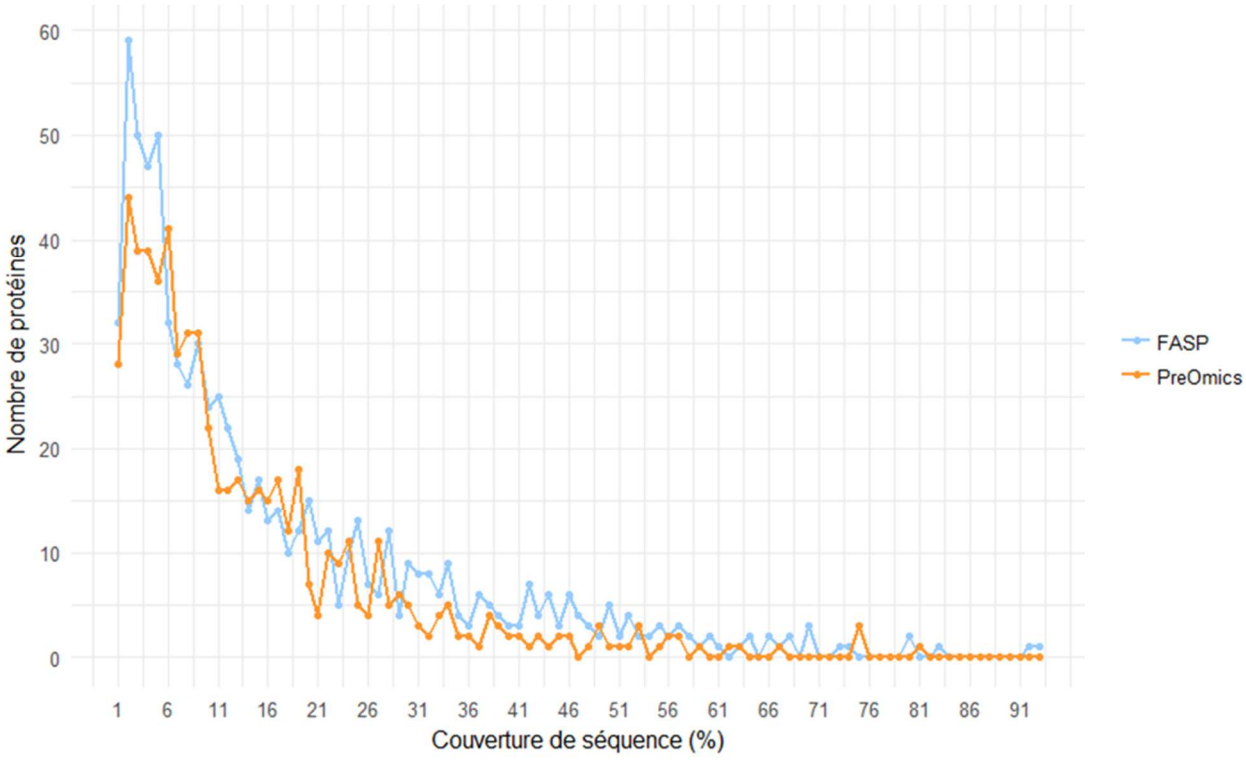


Figure IV-5 - Distribution de la couverture de séquence pour les protocoles FASP et PreOmics

Nous avons pu observer de manière surprenante que seuls 9 % des peptides ont été modifiés par l'utilisation du protocole PreOmics, alors que cette proportion est de 40 % pour le protocole FASP. En s'intéressant de plus près à ces modifications, nous avons pu nous rendre compte que très peu de carbamidométhylations des cystéines ont été induites par l'utilisation du protocole PreOmics. Ceci est en fait dû au fait que seuls 27 peptides contenant des cystéines (soit 1 % des peptides) ont pu être identifiés avec le kit iST, alors que ceux-ci sont au nombre de 1264 dans le protocole FASP (soit 35 % des peptides). Ce phénomène peut venir d'une possible inefficacité des réactions de réduction des ponts disulfures et d'alkylation des cystéines permettant de déstructurer les protéines et de faciliter l'accès de l'enzyme aux protéines, du fait peut-être de l'utilisation d'hydrochlorure de Tris(2-carboxyethyl)phosphine (TCEP) comme agent réducteur et de chloroacétamide, au lieu de DTT et d'IAA dans le cas du FASP. En effet, ces derniers pouvant réagir ensemble, ils ne peuvent être contenus dans un même tampon de lyse permettant la réduction et l'alkylation, comme c'est le cas dans le kit iST. Bien que la combinaison TCEP-chloroacétamide semble un peu moins efficace que celle couplant le DTT et l'IAA[181], ceci n'explique pas le taux si faible de peptides à cystéines, d'autant que des tests effectués en parallèle sur des échantillons autres que des urines n'ont pas abouti à ce phénomène. Pour ce qui est des oxydations, nous avons pu observer que le protocole PreOmics induit davantage d'oxydations (8 %) que le protocole FASP (5 %). S'agissant d'un dispositif commercial, la composition

exacte des tampons n'est pas connue, ce qui rend l'argumentaire de l'ensemble des observations difficile.

Enfin, la répétabilité a été grossièrement évaluée à l'aide de données de comptage de spectres issues du logiciel Proline, étant donné qu'il n'y a à l'heure actuelle pas d'outil permettant d'extraire de manière automatisée et correcte des données XIC à partir d'analyses générées par des spectromètres de masse TripleTOF commercialisés par SCIEX. Même si le logiciel MaxQuant est donné comme étant compatible avec les données de ce constructeur, nous avons pu observer au laboratoire qu'en réalité, très peu de protéines pouvaient être quantifiées avec cet outil. Ainsi, nous avons pu établir des boîtes à moustache représentant la répartition des CV calculés avec le nombre de spectres ayant permis l'identification de chaque protéine, sur les trois réplicats de chaque protocole (Figure IV-6).

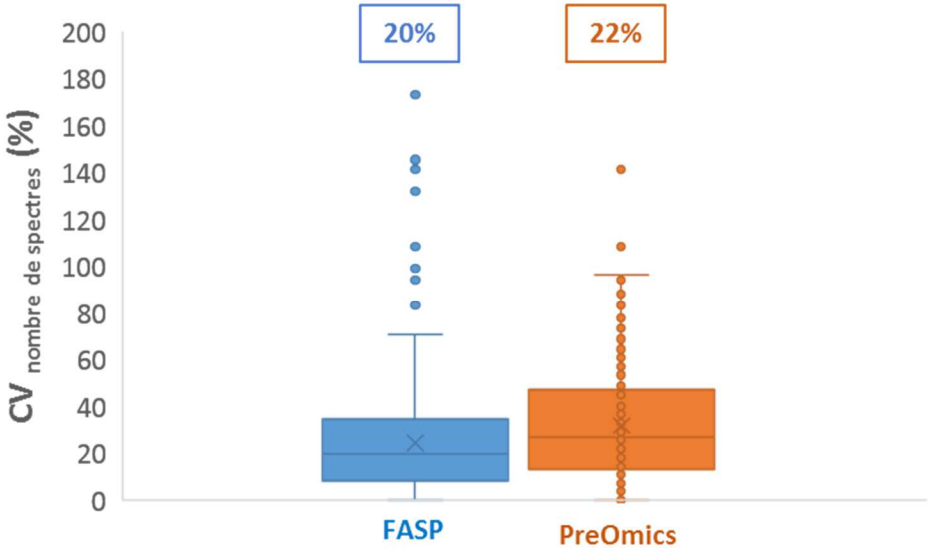


Figure IV-6 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide du nombre de spectres ayant permis l'identification de chaque protéine dans chaque réplicat pour le protocole FASP et le kit iST commercialisé par PreOmics.
Les valeurs encadrées correspondent aux CV médians pour protocole

Le protocole FASP semble être davantage répétable que le protocole PreOmics, avec un CV médian de 20 % et une répartition plus étroite des valeurs autour de la médiane. De plus, nous avons pu observer que 65 % des protéines identifiées dans les trois réplicats pour le protocole FASP sont communes, alors que seules 56 % des protéines sont communes aux trois réplicats pour le protocole PreOmics qui présente d'ailleurs un réplicat qui fait figure d'exception avec 16 % de protéines identifiées uniquement dans ce réplicat, contre 5 % pour les deux autres.

En conclusion, le protocole FASP permet d'identifier davantage de protéines, et ce, avec davantage de peptides que le protocole PreOmics, sans biais observé quant à la nature des peptides comme c'est le

cas pour le protocole PreOmics qui ne permet pas ou peu d'identifier des peptides à cystéines. Il génère cependant davantage de coupures manquées que le kit iST, mais ceci pourra encore être optimisé, notamment pour les étapes de validation des potentiels candidats par SRM, en implémentant une digestion à la LysC avant la digestion à la trypsine, ou encore en implémentant des solutions commerciales comme la trypsine rapide ou le mélange LysC-trypsine rapide commercialisés par Promega. Par ailleurs, ce protocole semble être suffisamment répétable pour être implémenté au moment de l'analyse protéomique quantitative sans marquage de la cohorte d'échantillons, dans le cadre du projet de recherche de biomarqueurs de suivi de greffons rénaux. De plus, cette technique peut être automatisée afin de traiter plusieurs échantillons en parallèle par l'implémentation de plaques 96 puits, pour une future application d'un test diagnostique en routine.

Des tests ont par ailleurs été effectués sur des plaques 96 puits contenant des membranes PVDF (pour « *PolyVinyliDene Fluoride* »). Ce protocole, publié sous le nom MStern[182], permet un traitement de l'échantillon beaucoup plus rapide que le FASP, puisqu'il met en jeu des pores cent fois plus larges que les membranes utilisées pour le FASP, et permet l'élution des peptides avec des solvants directement compatibles avec l'analyse LC-MS/MS. La rétention des protéines s'effectue ici par adsorption sur une surface hydrophobe et non pas par exclusion stérique. Malgré l'attrait de cette technique, celle-ci est limitée à une quantité de départ de 10 à 15 µg de protéines et génère davantage de coupures manquées que le FASP. Au cours de notre essai, nous n'avons pu identifier qu'une centaine de protéines, rendant cette technique non compétitive au FASP dans nos mains pour l'analyse protéomique d'échantillons urinaires.

CHAPITRE III

APPLICATION DES DEVELOPPEMENTS A DES PROJETS DE RECHERCHE DE BIOMARQUEURS PAR ANALYSE PROTEOMIQUE QUANTITATIVE SANS MARQUAGE

CHAPITRE III – APPLICATION DES DEVELOPPEMENTS A DES PROJETS DE RECHERCHE DE BIOMARQUEURS PAR ANALYSE PROTEOMIQUE QUANTITATIVE SANS MARQUAGE

Le protéome est une entité hautement dynamique, sensible aux variations externes et internes[183]. En effet, les désordres biologiques telles que les maladies ou la prise de traitements médicamenteux peuvent induire une variation rapide de l'expression des protéines[173, 184, 185]. Détecter ces changements au travers de signatures biologiques, nommées biomarqueurs, est un enjeu clé dans le domaine médical, puisqu'ils ont un rôle :

- **Diagnostic**, idéalement précoce, de maladie,
- **Pronostique**, en renseignant sur l'évolution de la maladie,
- **Prédictif**, en mettant en avant une forme de maladie résistante aux traitements par exemple,
- Dans la **stratification** de la maladie,
- Dans le **suivi** de la progression de la maladie,
- Dans le traitement (**cibles thérapeutiques**)[8, 10, 184-186].

L'analyse protéomique par spectrométrie de masse est un outil populaire dans le domaine de la recherche de biomarqueurs protéiques. Les approches quantitatives globales sans marquage permettent d'obtenir des informations permettant de comprendre les processus biologiques sous-jacents aux divers désordres étudiés de manière rapide et simple, et sont de ce fait couramment employées[3, 93, 186]. Il est largement admis que l'expression des protéines d'un individu sain à un autre est similaire, avec des variations statistiquement non pertinentes, contrairement à celles dans le cas d'une situation pathologique. Ainsi, les approches différentielles comparant les profils protéiques pathologiques à ceux d'individus sains permettent de mettre en avant ces différences d'expression, signatures de maladies[10, 184]. Ces modifications d'expression sont rarement restreintes à une protéine unique, mais à plusieurs, qui peuvent être sur- ou sous-exprimées, afin d'assurer davantage de spécificité[10, 185, 186]. L'attribution d'une signature formelle par l'intermédiaire d'un panel de protéines à une maladie est, malgré les outils statistiques souvent puissants, un défi majeur[10].

Les stratégies de protéomique quantitative différentielle pour la recherche de biomarqueurs ont suscité un réel engouement au début des années 2000, notamment suite à la publication d'un papier dans le journal *The Lancet*[187], clamant une avancée majeure dans le diagnostic précoce des cancers de

l'ovaire. Le test avait alors rapidement été amené vers l'étape clinique, mais des doutes quant à sa fiabilité ont grandi lorsque les données de cette étude ont été réanalysées par des groupes indépendants. Des irrégularités de pratiques expérimentales et de collectes d'échantillons ont ainsi été mises en évidence[188, 189]. Cette publication, ainsi que beaucoup d'autres mettant en avant des biomarqueurs d'intérêt limité et non spécifiques, tels que des marqueurs d'inflammations, ont hélas très largement contribué à ternir la réputation de l'analyse protéomique dans le domaine de la recherche de biomarqueurs[184, 188]. La communauté scientifique, consciente de ces problèmes a depuis activement œuvré pour fiabiliser la recherche de biomarqueurs par analyse protéomique, en sensibilisant les chercheurs sur différents points à respecter afin d'assurer la robustesse et la spécificité des biomarqueurs. Il s'agit :

- De la mise en place de **collaborations étroites** entre biologistes, massistes et statisticiens, dans le but de mettre en commun les compétences nécessaires à ce type d'études afin de définir clairement le projet clinique et de choisir les échantillons appropriés[3, 8, 10, 190-192],

- De la mise en place de **procédures standardisées** pour la collecte et le stockage des échantillons[192],

- Du **contrôle du processus analytique**[9, 10]**,**

- Et de la **séparation claire des phases de découverte et de validation**[184, 191]. Malgré l'apparente simplicité de ces approches différentielles, le chemin entre la découverte de biomarqueurs et leur application en clinique est souvent long et périlleux. Non seulement les approches de découverte permettent souvent d'établir des listes conséquentes de protéines différentiellement exprimées, d'où la nécessité d'impliquer de nombreuses compétences pour réduire la liste à un panel d'une dizaine de candidats biomarqueurs, mais surtout cette dernière étape de validation est longue et coûteuse. En effet, celle-ci doit être effectuée sur de larges cohortes de plusieurs centaines de patients, afin de vérifier que les candidats biomarqueurs sont bien représentatifs de la population et spécifiques de la maladie[184, 191]. Les critères évalués lors de cette étape ultime avant l'application en routine sont la spécificité, la sensibilité, l'utilité pour l'usage clinique ainsi que la robustesse[184, 188, 191]. Cette étape de validation a malheureusement souvent été omise et conséquente de nombreux échecs[8, 184, 185, 188].

Conscient de ces points de faiblesse, le protéomiste doit, dans le cadre de projets de recherche de

biomarqueurs, assurer le contrôle du processus analytique pour garantir qu'une différence d'abondance provienne d'une vraie différence biologique. Cela passe par l'utilisation de schémas analytiques simples et fiables, et notamment par des préparations d'échantillons qui maintiennent l'abondance réelle des protéines de par leur répétabilité, et qui garantissent que les deux groupes sont analysés dans des conditions comparables[10]. Ainsi, les différents développements évoqués au *Chapitre II* ont été appliqués au cours de ces travaux de thèse à différents projets de recherche de biomarqueurs cibles de thérapies, de résistance au traitement, mais également de suivi.

I- Projet de recherche de biomarqueurs de cellules souches cancéreuses de glioblastomes

Le projet de recherche de biomarqueurs de cellules souches cancéreuses de glioblastomes a été mené en collaboration avec l'équipe du laboratoire d'innovation thérapeutique de la faculté de pharmacie d'Illkirch, dirigée par le Dr Jacques HAIECH.

1- Contexte

Le glioblastome correspond au grade IV des gliomes[193]. Il s'agit d'une tumeur cérébrale qui se développe à partir des cellules constituant le tissu de soutien du système nerveux : les cellules gliales. Le glioblastome correspond à la tumeur cérébrale primitive maligne la plus agressive[194, 195]. C'est également la tumeur la plus fréquente, touchant les personnes de 45 à 70 ans avec une incidence de 5 à 7 personnes sur cent mille par an en Europe[195-197]. Le glioblastome est très hétérogène, et de par sa croissance invasive dans le tissu cérébral environnant, présente un caractère infiltrant rendant la résection chirurgicale complète impossible[195, 198, 199]. Un traitement adjuvant par radio- et/ou chimiothérapies après l'intervention chirugicale ne permet à l'heure actuelle toujours pas d'éviter la récidive inévitable de la tumeur, mais permet tout de même d'améliorer la survie des patients qui se situe à une médiane de 15 mois post-diagnostic[194, 195, 198, 200].

Une théorie largement controversée postule qu'une sous-population immature de cellules au sein même de la tumeur, nommée cellules souches cancéreuses (CSC), jouerait un rôle dans l'initiation et la récidive de la tumeur[195, 198, 201-204]. Ces cellules présentent des caractéristiques similaires aux cellules souches non tumorales, telles que :

- La capacité à **s'auto-renouveler**[195, 198, 201, 203, 205],
- La capacité à **se propager sans facteur de croissance**[195],
- La **résistance aux thérapies**, notamment de par leur protection dans des niches de cellules souches[195, 196, 198, 201, 203, 205],
- La capacité à **rester en état de quiescence** pendant des périodes prolongées, ce qui expliquerait que des récidives ou des métastases surviennent après une longue période, mais aussi le mécanisme de résistance aux chimiothérapies anti-prolifératives[201],

- La capacité à **métastaser**[201],

- Et la **difficulté à les isoler**[203].

Même si cette hypothèse reste très controversée, la présence de telles cellules a été démontrée pour les glioblastomes[203]. L'éradication de ces CSC qui sont à l'origine de la prolifération de cellules cancéreuses serait nécessaire et suffisante pour stopper l'expansion de la tumeur, et prévenir la récidive[205]. C'est pourquoi, l'idée de combiner les traitements actuels à des thérapies ciblées contre les CSC de glioblastomes a été émise, de manière à améliorer l'efficacité des traitements des glioblastomes[195, 198, 201, 205] (Figure III-1).

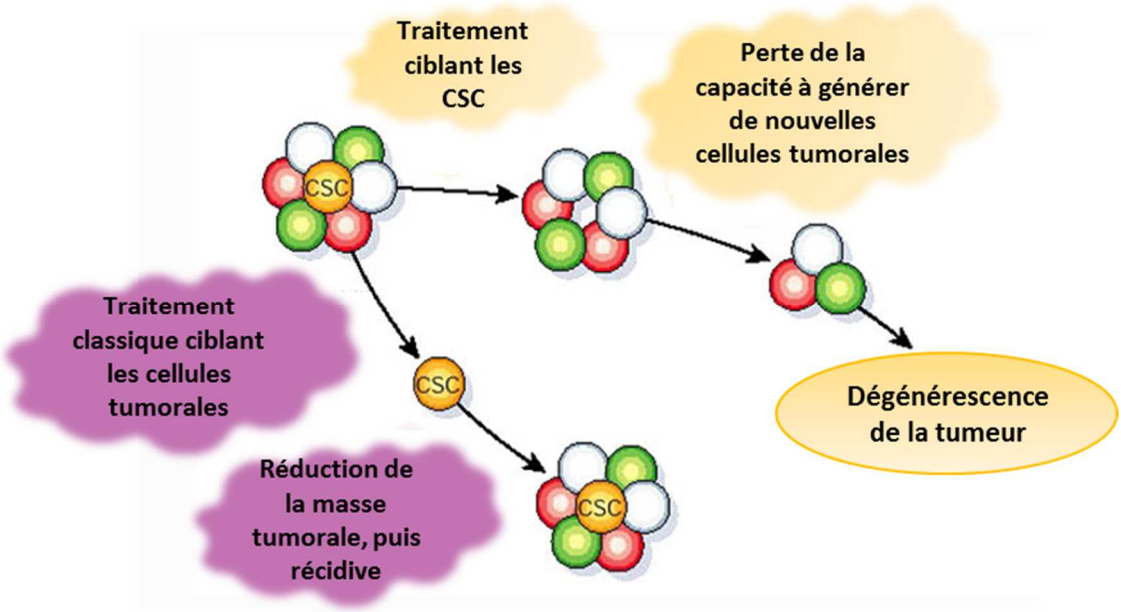


Figure I-1 - Schéma reflétant l'apport d'une thérapie ciblant les CSC dans la dégénération de la tumeur (adapté de [205])

Afin de développer de telles thérapies, il est nécessaire d'isoler les CSC et d'identifier de manière efficace des marqueurs à leur surface qui pourront être des cibles potentielles de thérapies. C'est dans l'optique de mettre en avant des marqueurs de surface spécifiques des CSC que nous avons mené une étude protéomique différentielle opposant des « *ghosts* » *membranaires* de CSC de glioblastomes à ceux d'astrocytes humains et/ou d'une lignée cellulaire de glioblastome (U87). Cette approche est innovante du fait qu'elle met en œuvre des préparations membranaires de CSC de glioblastomes. En effet, très peu d'études de recherche de biomarqueurs de glioblastomes humains, portant sur les protéines exprimées à la surface des CSC et enchâssées dans la membrane plasmique, qui représentent des cibles thérapeutiques de choix, ont été menées[197].

2- Stratégie analytique employée

Une approche de protéomique quantitative sans marquage XIC a été mise en œuvre sur des « *ghosts* » *membranaires*, préparés par l'équipe du Dr Jacques HAIECH, provenant :

- De deux lignées cellulaires de CSC issues de glioblastomes de patients décédés. Les tissus tumoraux ont été fournis par le Dr Hervé CHNEIWEISS qui dirige une équipe de recherche sur la plasticité gliale à Paris, et les préparations de CSC ont quant à elles été effectuées par l'équipe du Dr Jacques HAIECH comme décrit par Cristina PATRU *et collaborateurs*[206]. Chaque lignée a fait l'objet de deux préparations de CSC, ainsi les échantillons nommés TG01, OB1 (qui ont la particularité d'être non-mutés au niveau du gène p53, suppresseur de tumeur[207]), ainsi que TG10 et TG16 (p53 mutés) sont des réplicats biologiques,

- D'une culture d'astrocytes humains, nommée HA,

- Et d'une lignée cellulaire de glioblastome, U87. Il s'agissait initialement d'une culture issue d'un glioblastome d'une patiente décédée en 1968 (U87-MG). Depuis 2016, l'origine exacte de cette lignée cellulaire de glioblastome est discutée[208].

Le schéma analytique utilisé dans le cadre de ce projet est celui qui a été optimisé au *Chapitre II- I-2* et détaillé dans la Figure I-2. Brièvement, les échantillons TG01, OB1, TG10, TG16, HA et U87 ont été préparés en triplicats de gel « *Stacking* » découpés en deux bandes, et dont les extraits peptidiques ont été réunis avant analyse nanoLC-MS/MS sur un spectromètre de masse de type Q-TOF (Impact-HD, BRUKER). Des peptides iRT ont été dopés dans chaque échantillon afin de suivre les performances du couplage dans les conditions d'analyse. En supplément de ces peptides iRT, un contrôle externe, soit un digeste de BSA, a été analysé entre chaque échantillon pour suivre la stabilité du couplage pendant la semaine d'analyse. Au total, 18 échantillons ont été analysés dans un ordre aléatoire. Les données ont été traitées à l'aide du logiciel MaxQuant, et le traitement statistique a été effectué à partir des intensités protéiques normalisées LFQ avec le logiciel Perseus.

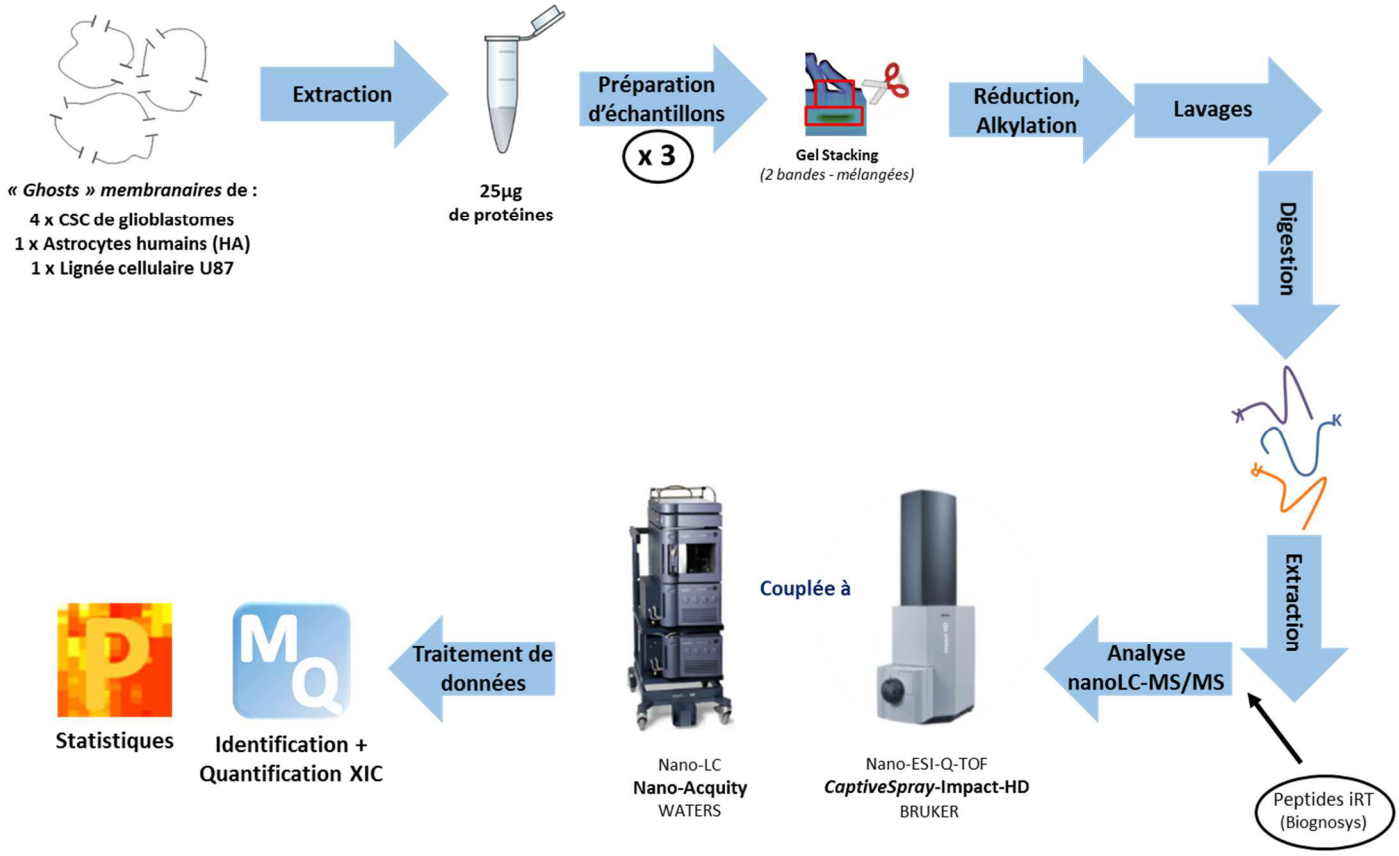


Figure I-2 - Schéma analytique utilisé pour la recherche de biomarqueurs de surface de CSC de glioblastomes

3- Résultats

a. Contrôles qualités

Contrôle qualité externe

Deux peptides du digeste de BSA analysé entre chaque échantillon ont été suivis pour vérifier les performances du couplage pendant la séquence d'analyses. Il s'agit :

- Du peptide YICDNQDTISSK, doublement chargé, avec un rapport masse sur charge de 722,32,
- Et du peptide YLYEIAR, également doublement chargé, avec un rapport masse sur charge de 464,25.

Ces deux peptides ont permis dans un premier temps de suivre les performances du système de chromatographie, à l'aide de leur temps de rétention ainsi que de leur largeur à mi-hauteur (Figures I-3 et I-4 respectivement). Etant donné que l'analyse d'un digeste de BSA constitue le test de performance de l'ensemble des couplages disponibles au laboratoire, des seuils critiques ont été définis pour chaque instrumentation.

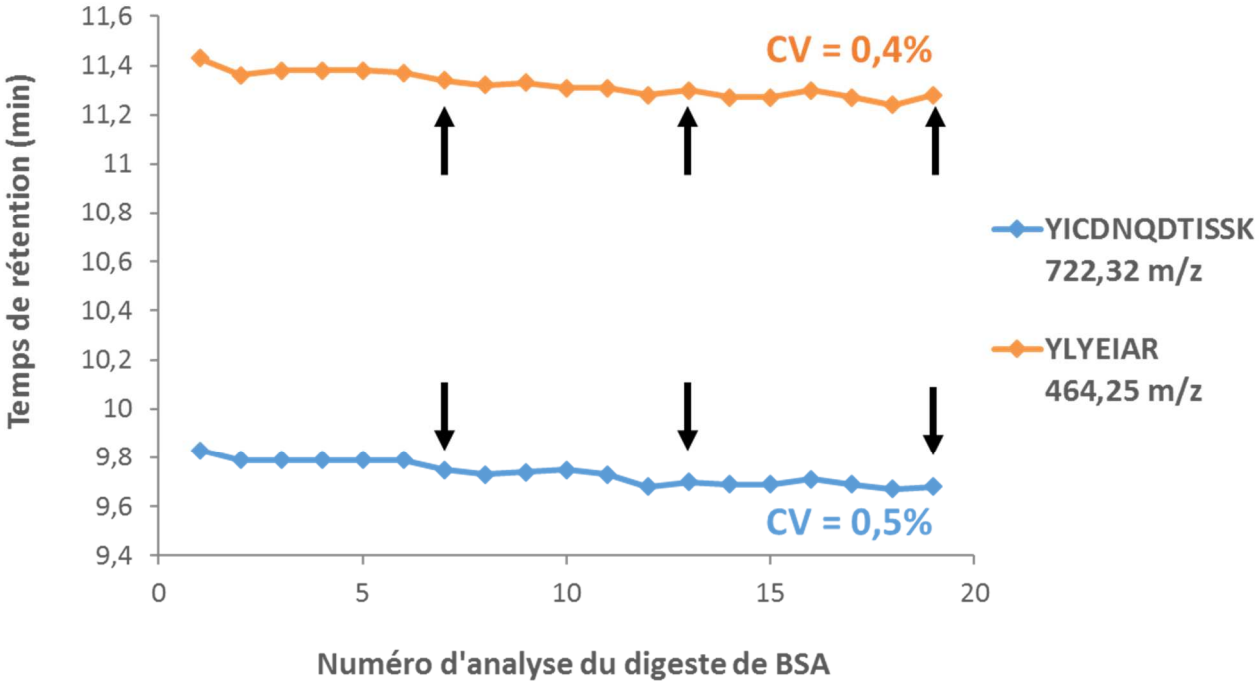


Figure I-3 - Suivi du temps de rétention des deux peptides extraits du digeste de BSA
Les flèches indiquent le moment auquel une nouvelle préparation du digeste de BSA a été réalisée

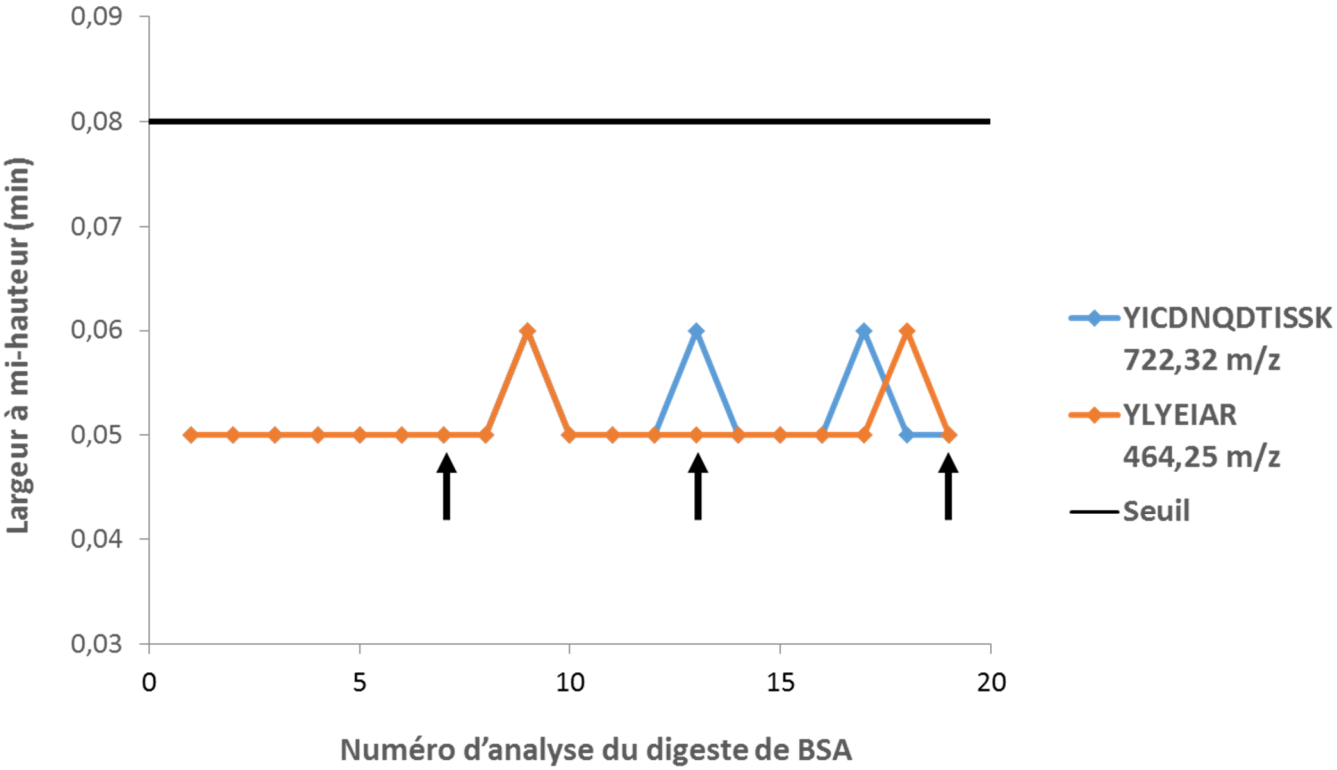


Figure I-4 - Suivi des largeurs à mi-hauteur des traces chromatographiques des deux peptides extraits du digeste de BSA
Les flèches indiquent le moment auquel une nouvelle préparation du digeste de BSA a été réalisée

Les coefficients de variation des Tr des deux peptides étant très faibles (0,4 et 0,5 %), tout comme la variation des largeurs à mi-hauteur qui se situent d'ailleurs toujours sous la valeur seuil, reflètent une bonne stabilité du système chromatographique.

Dans un deuxième temps, le suivi de leur intensité et de leur résolution en masse ont permis de suivre la stabilité et l'état d'encrassement du spectromètre de masse sur la semaine d'analyse. Ces données sont fournies par les Figures I-5 et I-6.

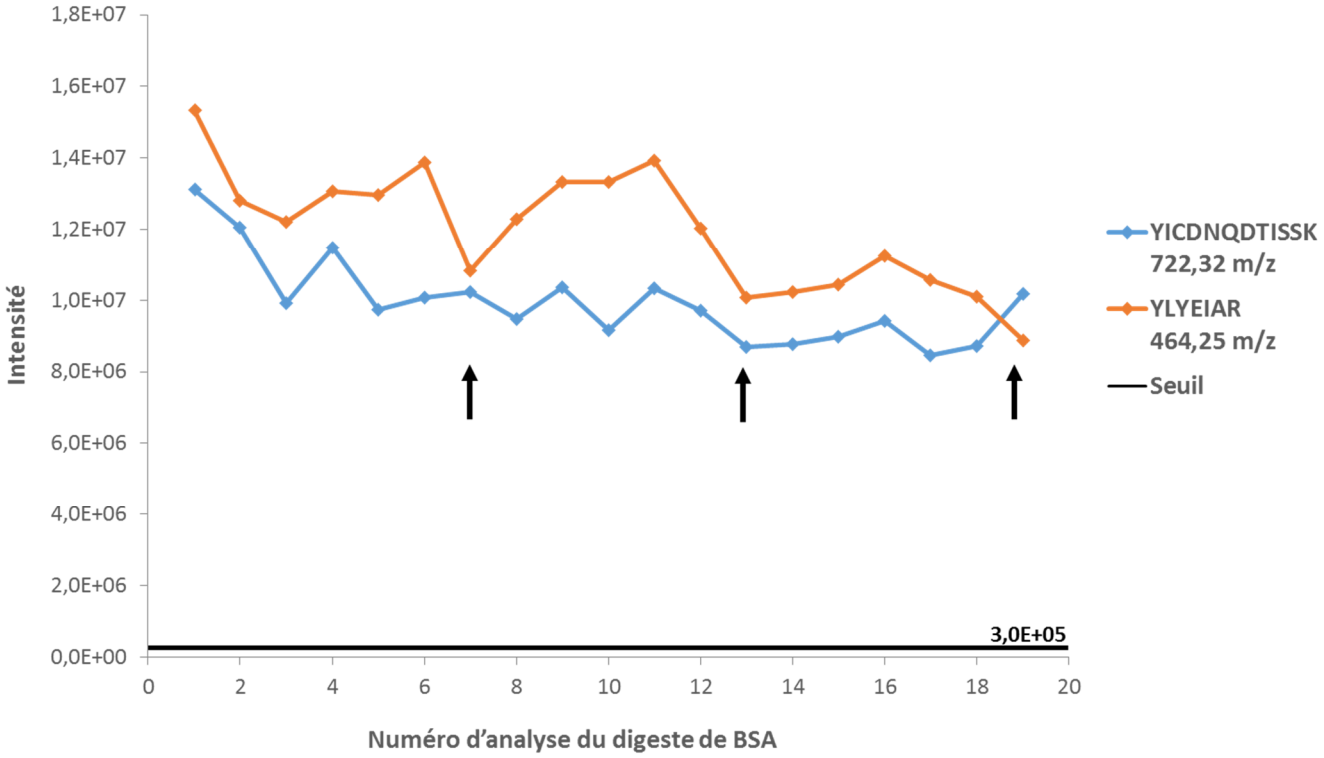


Figure I-5 – Suivi de l'intensité des deux peptides extraits du digeste de BSA.
Les flèches indiquent le moment auquel une nouvelle préparation du digeste de BSA a été réalisée

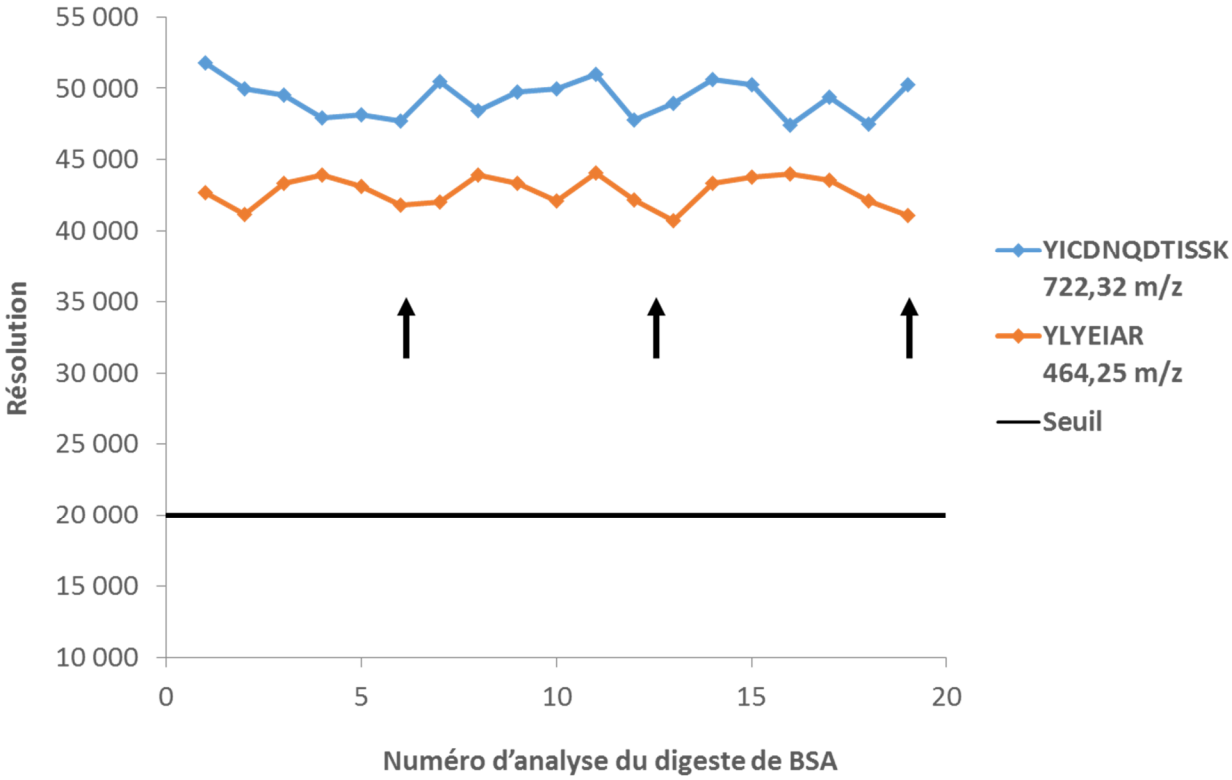


Figure I-6 - Suivi de la résolution en masse des deux peptides extraits du digeste de BSA.
Les flèches indiquent le moment auquel une nouvelle préparation du digeste de BSA a été réalisée

L'intensité et la résolution en masse des deux peptides peu variables, et au-dessus des valeurs seuil, reflètent que le spectromètre de masse n'a pas montré de signe d'encrassement durant la semaine d'analyse, et a été relativement stable.

Contrôle qualité interne

Les onze peptides iRT dopés dans chacun des échantillons ont permis de suivre, dans les conditions d'analyse, la stabilité du couplage. Dans un premier temps, la répartition des coefficients de variation calculés à partir des Tr sur l'ensemble des échantillons pour chaque peptide ont permis d'établir la boîte à moustache de la Figure I-7. La valeur médiane des CV inférieure à 1 %, nous permet de conclure que le système chromatographique, dans les conditions d'analyse, est resté stable tout au long de la séquence. Par ailleurs, la Figure I-8 permet d'apprécier la répartition des CV calculés à partir des aires obtenues pour chaque peptide iRT par le logiciel Skyline. La valeur médiane de 28 % reflète quant à elle les fortes variations observées sur certains peptides iRT présentant un faible rapport signal sur bruit. En ôtant ces peptides, la valeur médiane des CV des iRT ayant un facteur de réponse en masse correct s'élève à 13 %, reflétant ainsi une bonne répétabilité du système.

L'ensemble de ces contrôles qualité ont permis de rendre compte de la répétabilité et de la robustesse du couplage nanoLC-MS/MS pendant les analyses des échantillons du projet de recherche de biomarqueurs de surface de CSC de glioblastomes.

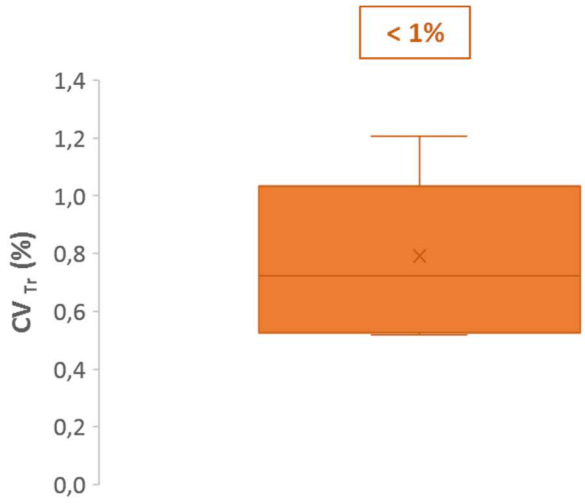


Figure I-7 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des Tr extraits de Skyline pour chaque peptide iRT dans l'ensemble des échantillons.
La valeur encadrée correspond à la valeur médiane

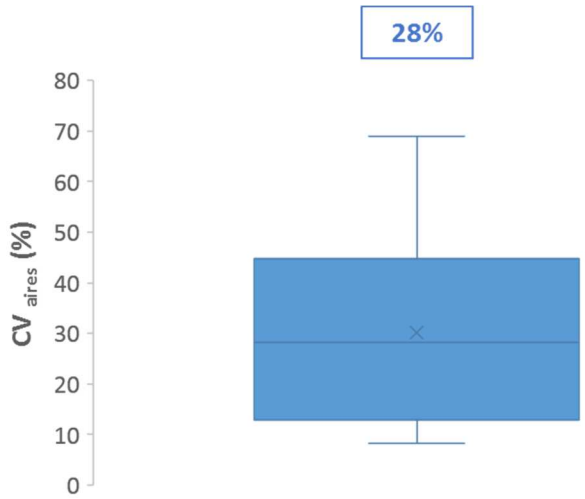


Figure I-8 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des valeurs d'aires extraites de Skyline pour chaque peptide iRT quantifié dans l'ensemble des échantillons.
La valeur encadrée correspond à la valeur médiane

b. Résultats d'identification et de quantification

En ce qui concerne les résultats relatifs à l'analyse des échantillons, 4096 protéines au total ont pu être identifiées, avec au minimum un peptide unique, sur l'ensemble des analyses. Parmi celles-ci, vingt-sept pourcents sont annotées « *plasma membrane* » par l'outil GO. A l'origine du projet, soit avant

mon arrivée au laboratoire, les biologistes étaient particulièrement intéressés par les récepteurs couplés aux protéines G (RCPG), car ils représentent les cibles de 30 à 50 % des médicaments actuellement sur le marché[209]. Cependant, malgré les nombreuses tentatives menées par l'équipe du laboratoire, peu de ces protéines ont pu être mises en avant par MS du fait de leur faible abondance, et de leur structure présentant beaucoup de domaines transmembranaires et des domaines accessibles (dans le milieu intra ou extracellulaire) de séquences courtes. Dans ce contexte, le projet s'est porté sur l'étude des CD, étant donné qu'ils sont davantage représentés dans les analyses de protéomique, et qu'ils peuvent être facilement ciblés et validés à l'aide d'anticorps disponibles commercialement dirigés contre ce type de protéines. Ainsi, parmi ces 27 % de protéines annotées « *plasma membrane* », 8 % sont des CD, soit 94 protéines (Figure I-9).

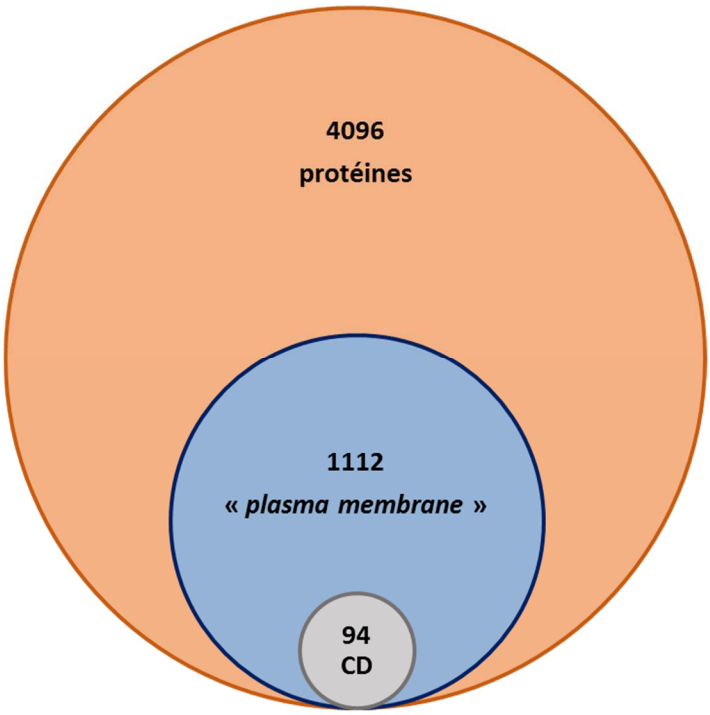


Figure I-9 - Nombre de protéines identifiées dans l'ensemble des échantillons avec au moins un peptide unique, ainsi que le nombre de protéines annotées « *plasma membrane* » et CD par GO

La Figure I-10 illustre le nombre de protéines identifiées en moyenne dans les trois réplicats pour chaque échantillon. Ainsi, les réplicats biologiques OB1, TG01, ainsi que TG10 et TG16 permettent d'identifier un nombre similaire de protéines. Les échantillons HA et U87 permettent d'identifier moins de protéines que les échantillons issus de CSC de tumeurs humaines. Pour l'ensemble de ces échantillons, en moyenne 29 à 33 % de protéines identifiées sont annotées « *plasma membrane* ».

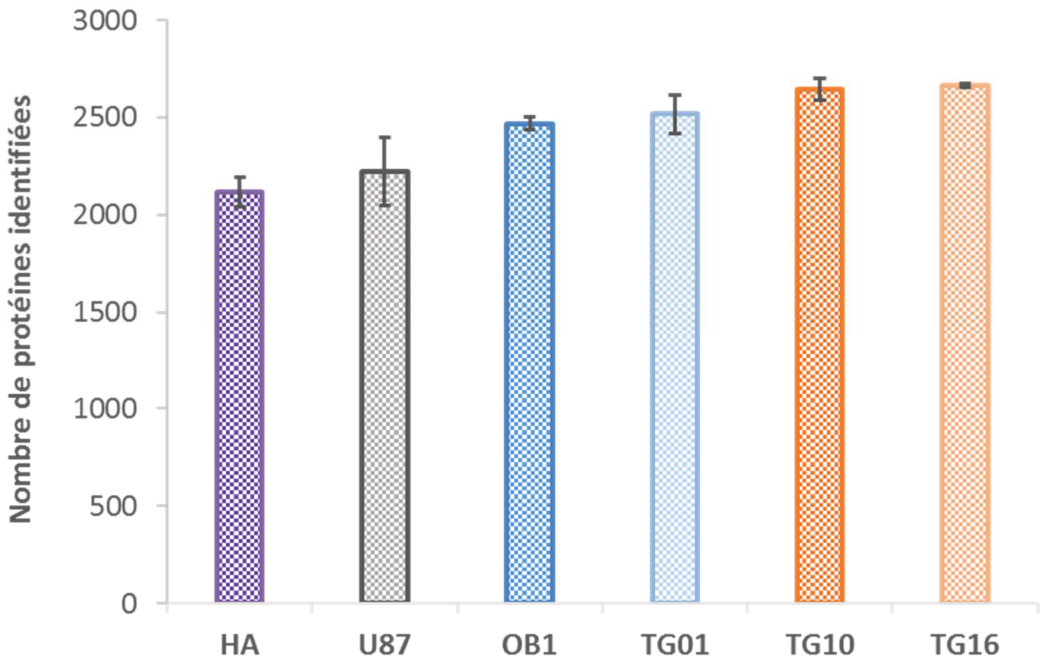


Figure I-10 - Nombre moyen de protéines identifiées avec au moins un peptide unique dans les trois réplicats de chaque échantillon

Par ailleurs, afin d'établir quelles protéines sont différentiellement exprimées, deux comparaisons ont été effectuées à l'aide d'un test statistique de Welch et du logiciel Perseus, à partir des données de quantification protéique normalisées LFQ :

- TG01, OB1, TG10, TG16 (groupe nommé CSG) contre HA pour différencier les protéines différentiellement exprimées dans les cellules saines par rapport aux CSC de glioblastomes,

- Et CSG contre U87 afin de différencier les protéines différentiellement exprimées dans les cellules cancéreuses différenciées de glioblastomes par rapport aux CSC de glioblastomes.

En ce qui concerne la comparaison CSG contre HA, 1147 protéines ont été identifiées comme étant différentiellement exprimées ($p < 0,05$), parmi lesquelles 614 sont surexprimées dans le groupe CSG (ratio CSG/HA > 0) (Figure I-11). Ce grand nombre de protéines différentiellement exprimées peut s'expliquer non seulement par la différence du nombre de protéines identifiées entre l'échantillon HA et les échantillons CSG, mais également par la structure des cohortes d'échantillons inadaptée aux outils statistiques (taille des cohortes et déséquilibre du nombre d'individus par groupe). Au total, huit protéines CD sont surexprimées dans les échantillons de CSC de glioblastomes, comme le montre le « *volcano plot* » de la Figure I-12.

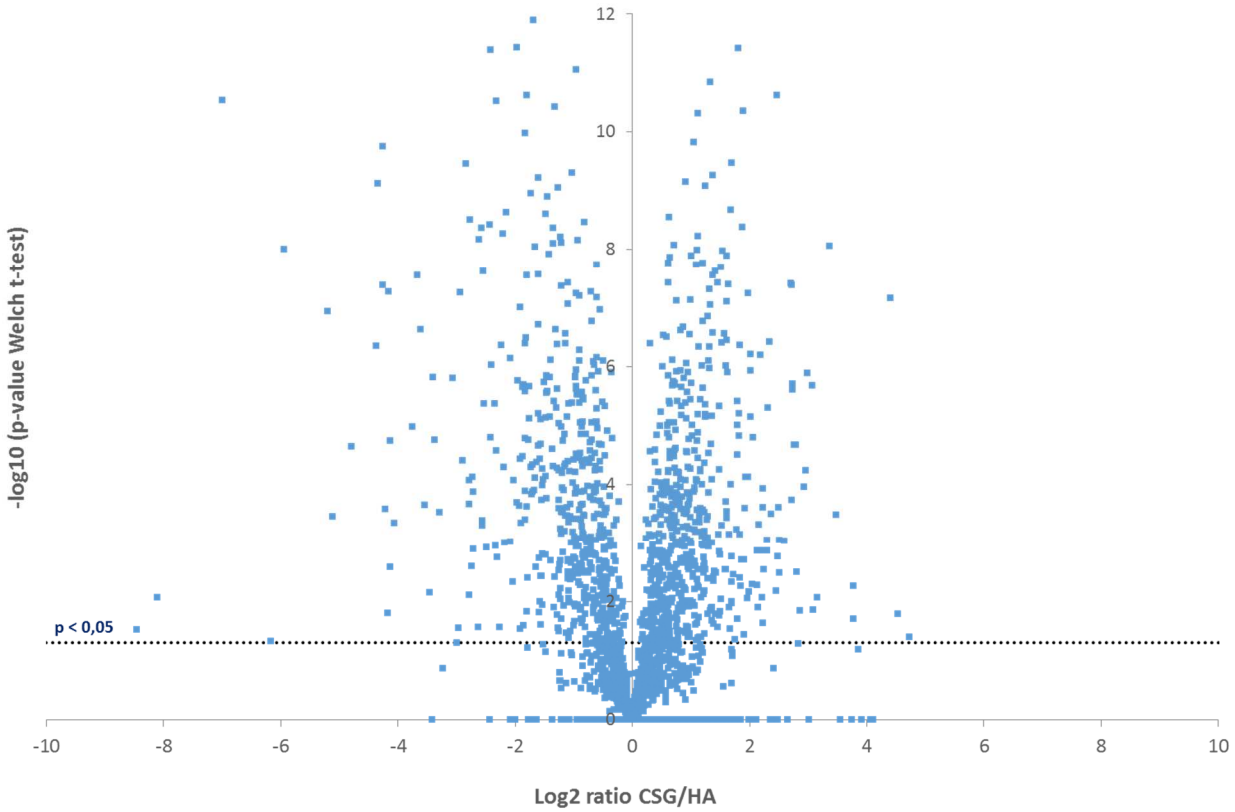


Figure I-11 – « *Volcano plot* » réalisé à l'aide des données issues du test statistique de Welch opposant les échantillons CSG à HA

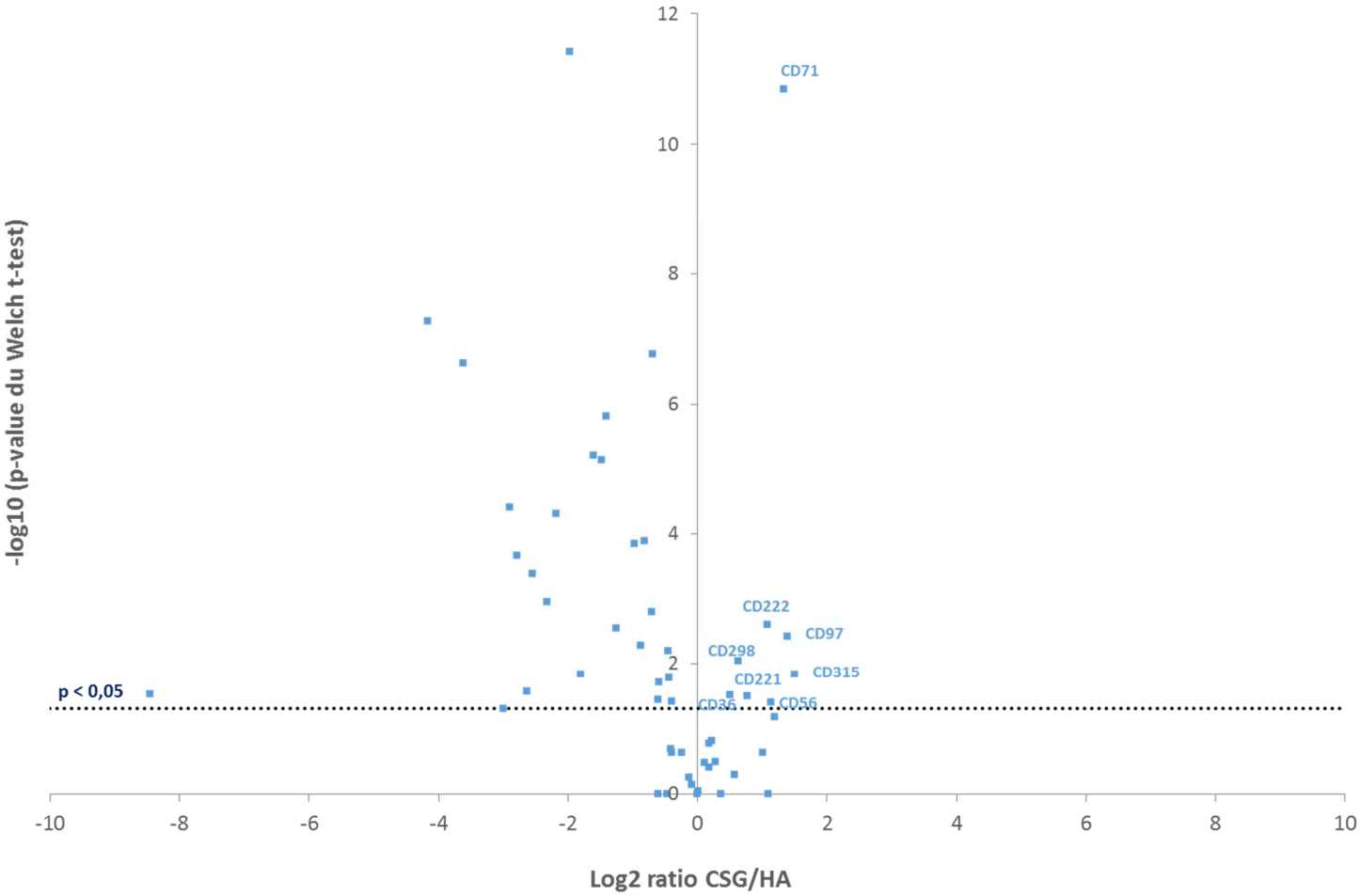


Figure I-12 – « *Volcano plot* » réalisé à l'aide des données pour les protéines CD issues du test statistique de Welch opposant les échantillons CSG à HA

En ce qui concerne la comparaison CSG contre U87, 1583 protéines ont été identifiées comme étant différentiellement exprimées ($p < 0,05$), et plus particulièrement, 775 sont surexprimées dans le groupe CSG (ratio CSG/U87 > 0) (Figure I-13). De même que pour la précédente comparaison, ce grand nombre de protéines différentiellement exprimées peut s'expliquer par le déséquilibre du nombre d'individus entre le groupe U87 et CSG, ainsi que par un nombre d'individus par groupe inadapté aux analyses statistiques. Au total, dix protéines CD sont surexprimées dans les échantillons de CSC de glioblastomes, comme l'illustre le « *volcano plot* » de la Figure I-14.

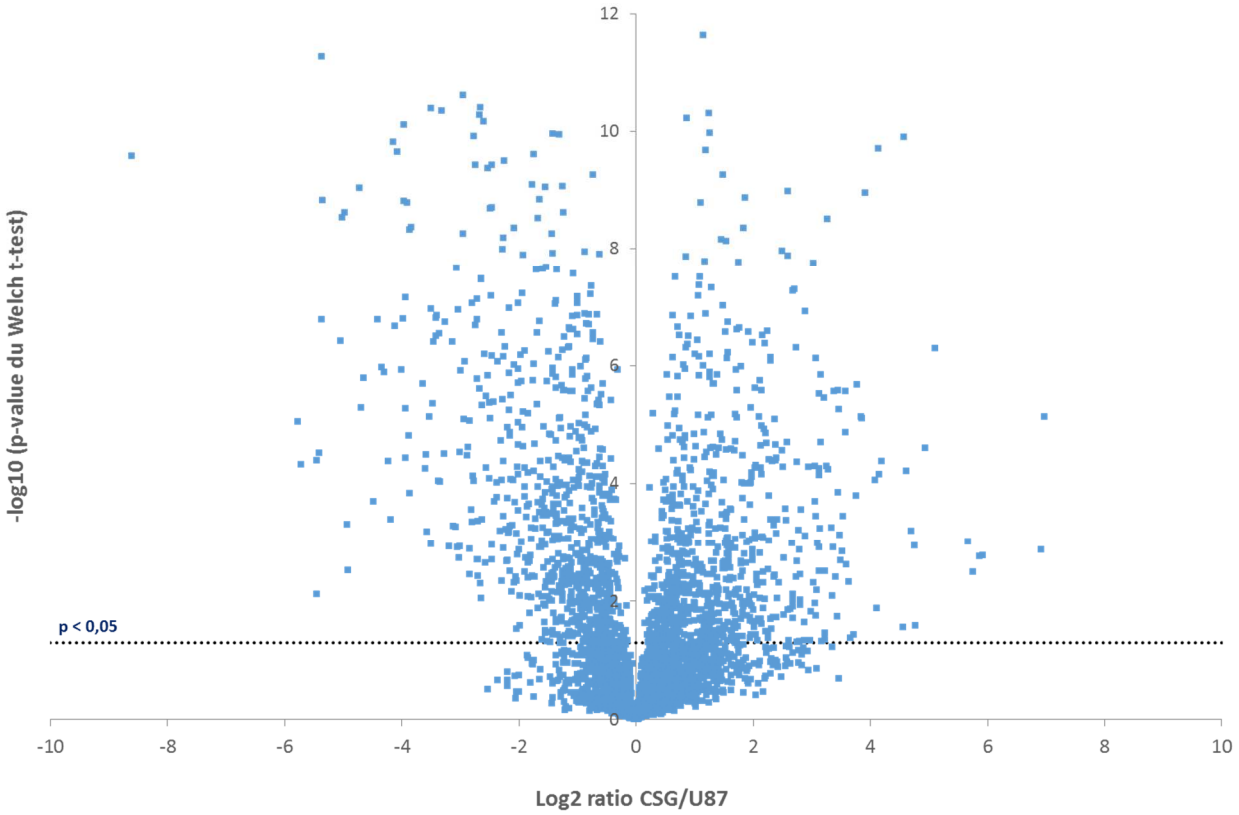


Figure I-13 – « *Volcano plot* » réalisé à l'aide des données issues du test statistique de Welch opposant les échantillons CSG à U87

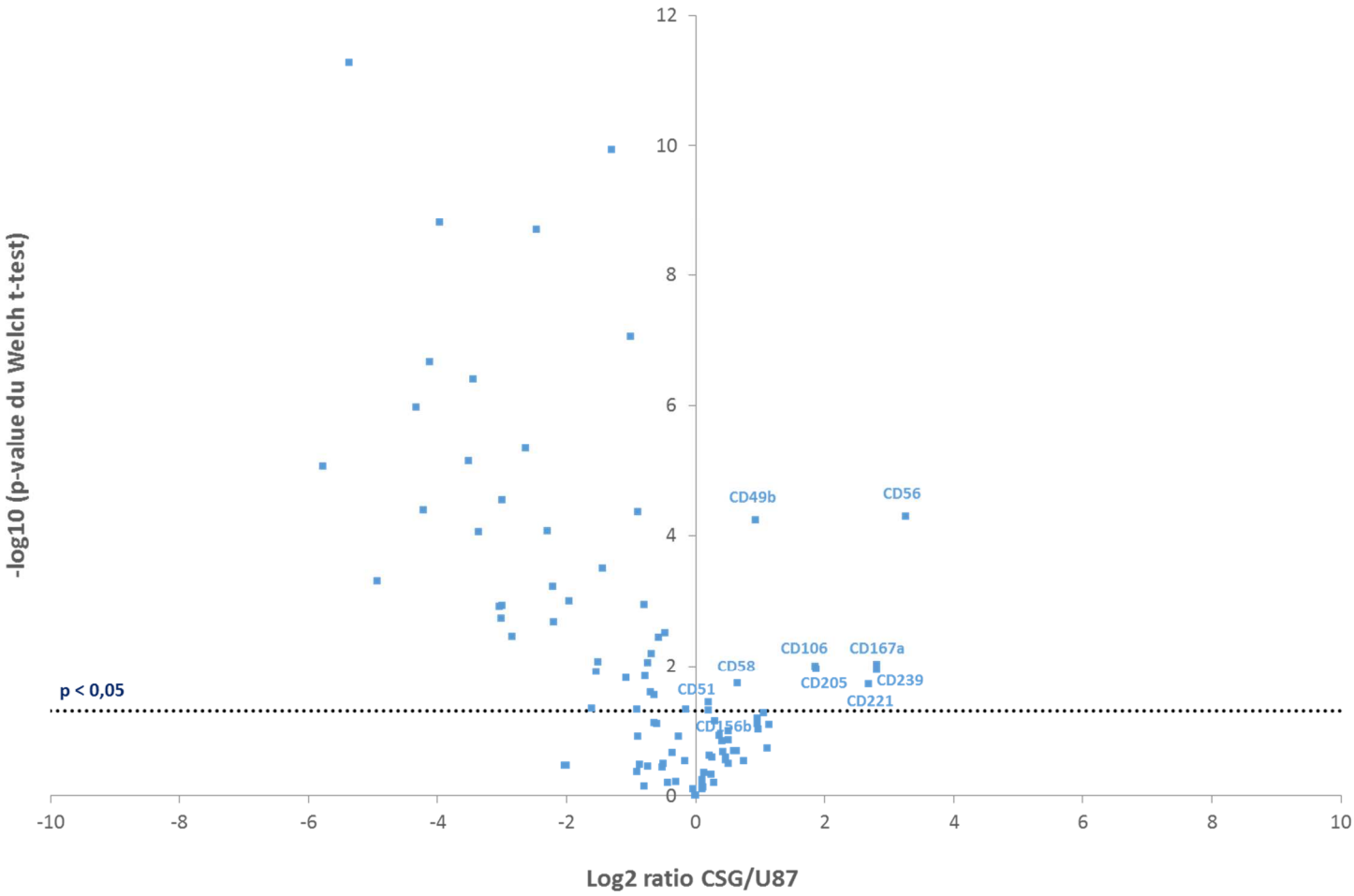


Figure I-14 – « *Volcano plot* » réalisé à l'aide des données pour les protéines CD issues du test statistique de Welch opposant les échantillons CSG à U87

Le détail de l'ensemble des protéines CD surexprimées dans chacune des comparaisons dans les CSC de glioblastomes est donné par le Tableau I-1. Celui-ci permet notamment de rendre compte que deux protéines sont retrouvées comme étant différentiellement exprimées dans les deux comparaisons : le CD221 et le CD56, identifiés respectivement avec 16 et 9 peptides uniques dans nos analyses. Ces protéines pourraient de ce fait être de potentiels marqueurs de CSC de glioblastomes. En effet, elles sont à la fois surexprimées dans les CSC par rapport à des cellules saines à partir desquelles les glioblastomes peuvent se développer (HA), et surexprimées par rapport aux cellules cancéreuses de glioblastomes différenciées. En ce qui concerne le CD56 (P13591, Neural-cell adhesion molecule 1), il a été testé par immunohistochimie sur des échantillons semblables, et a permis de distinguer les territoires différenciés des territoires indifférenciés au sein même de la tumeur. Il permettrait de ce fait de définir le niveau de malignité de la tumeur. Cette protéine, impliquée dans l'adhésion neurones-neurones, a notamment été décrite comme jouant un rôle dans la biologie des tumeurs, mais son expression n'est pas restreinte aux tissus cérébraux[210-213].

CD surexprimés dans CSG contre HA *(p<0,05 et ratio>0)*	CD surexprimés dans CSG contre U87 *(p<0,05 et ratio>0)*
	CD56
CD71	CD49b
CD222	CD51
CD97	CD58
CD298	CD156
CD315	CD205
CD36	CD167a
CD221	CD239
CD56	CD221
	CD156b

Tableau I-1 - Tableau résumant les protéines CD surexprimées dans les deux comparaisons. Les protéines en orange sont les protéines retrouvées comme différentiellement exprimées dans les deux comparaisons

4- Conclusion et perspectives

En conclusion, l'approche différentielle de protéomique quantitative sans marquage XIC a permis de mettre en évidence, non pas une éventuelle cible de thérapie, mais davantage un marqueur permettant potentiellement de distinguer les territoires différenciés des territoires indifférenciés au sein de la tumeur, le CD56 (ou P13591). Il est à noter cependant que cette étude présente différentes faiblesses que sont :

- L'implication d'un très faible nombre d'échantillons et de réplicats biologiques, notamment pour les échantillons contrôles (HA et U87) qui présentent non seulement un poids statistique très faible, mais ne permettent pas d'apprécier la représentativité de cette population. Cette étude s'inscrit néanmoins dans un contexte particulier avec une disponibilité très réduite d'échantillons.

- La possible modification du protéome d'origine des CSC de par l'utilisation de cultures cellulaires de plusieurs semaines permettant de les isoler.

Par ailleurs, la recherche de biomarqueurs cibles de thérapie des CSC soulève une difficulté liée au développement de thérapies ciblées, qu'est l'identification d'un ou plusieurs biomarqueurs spécifiques des CSC qui ne seraient pas partagés avec d'autres types cellulaires, dans le but d'éviter les effets secondaires.

Le CD56 doit maintenant faire l'objet d'une étude de validation à plus grande échelle, impliquant un grand nombre d'échantillons. Cette étude risque cependant de se heurter à la difficulté inhérente au projet de recherche de biomarqueurs de glioblastomes qu'est la disponibilité d'échantillons. L'intégration plus systématique de données de transcriptomique ou de génomique, complémentaires à celles de protéomique, pourraient jouer un rôle dans les années à venir dans la découverte de biomarqueurs notamment de glioblastomes, en permettant de corréler les différents niveaux d'expression[3].

II- Projet de recherche de biomarqueurs associés à une chimiorésistance primaire dans les lymphomes B diffus à grandes cellules

Ce projet de recherche de biomarqueurs associés à une chimiorésistance primaire dans les LBDGC a été mené en binôme avec le Dr Luc FORNECKER, hématologue praticien hospitaliers aux Hôpitaux Universitaires de Strasbourg (HUS) et docteur en chimie.

1- Contexte

Les lymphomes constituent un groupe hétérogène de malignités hématologiques se développant à partir des lymphocytes B ou T, et présentant une haute morbidité et mortalité. Sur la base de paramètres à la fois morphologiques et moléculaires, nous distinguons deux types de lymphomes :

- Les **lymphomes de Hodgkin**,
- Et les **lymphomes non-Hodgkiniens**[159].

Au sein de ce dernier type, les LBDGC représentent l'entité la plus fréquente avec un tiers des patients[214]. Il s'agit d'une tumeur se développant aux dépens des lymphocytes B, agressive, et rapidement prolifératives[157, 215]. Au sein des LBDGC, une classification moléculaire permet de distinguer trois sous-types : GCB (pour « *Germinal Center B-cell like* »), ABC (pour « *activated B-cell like* ») et PMBL (pour « *Primary mediastinal B-cell lymphoma* »). Le pronostic varie selon le sous-type dont le patient est atteint. Ainsi, un sous-type ABC est associé à une issue plus défavorable[216]. En effet, malgré les traitements de première ligne que sont la chimiothérapie CHOP (cyclophosphamide, doxorubicine, vincristine et prednisone) associée au Rituximab, qui a d'ailleurs été une avancée majeure dans le traitement de ces pathologies ces vingt dernières années, le rendement de guérison reste de 40-50%[215]. Ceci s'explique par le fait que la moitié des patients atteints de LBDGC développe une résistance au traitement menant au décès, malgré leur prise en charge thérapeutique[214, 217]. Cette résistance présente à l'heure actuelle un sérieux défi pour les hématologues.

Une meilleure compréhension de la résistance à ces traitements est primordiale afin d'améliorer la qualité et l'efficacité des thérapies. C'est dans ce contexte que nous avons mené une étude protéomique différentielle à partir de biopsies ganglionnaires congelées au moment du diagnostic, opposant 8 patients réfractaires à 12 patients sensibles au traitement de première ligne, et ce afin de

mieux comprendre la biologie sous-jacente à la chimiorésistance, voire de trouver des signatures permettant de déceler cette résistance au moment du diagnostic et d'adapter les traitements en conséquence.

2- Stratégie analytique employée

L'originalité de ce projet repose sur le matériel de départ utilisé, soit des biopsies ganglionnaires fraîches pour la recherche de biomarqueurs de lymphome, comme développé au *Chapitre II-III*, combiné à une approche de protéomique globale. De plus, de par l'implication du Dr Luc FORNECKER au laboratoire, la sélection des échantillons a pu être pleinement maîtrisée par notre équipe.

Les deux cohortes ont été réalisées à partir d'une liste de patients pour lesquels un prélèvement congelé était disponible au Centre de Ressources Biologiques des HUS. La sélection s'est ensuite effectuée selon plusieurs critères, tels que :

- Un diagnostic établi de LBDGC selon la classification de l'OMS 2008[218],
- L'application du même traitement de première ligne, soit la combinaison Rituximab-CHOP,
- Un prélèvement effectué au moment du diagnostic disponible au Centre de Ressources Biologique,
- La disponibilité des données cliniques,
- Et l'obtention du consentement du patient quant à l'utilisation des prélèvements pour la recherche.

Ainsi, la cohorte « réfractaire » contenait 8 prélèvements de patients réfractaires primaires aux traitements de première ligne, ou qui ont rechuté de manière précoce malgré une réponse complète initiale. La cohorte « sensible » était quant à elle constituée de 12 prélèvements de patients qui ont répondu de manière complète au traitement de première ligne et qui n'ont pas rechuté.

Le schéma analytique de l'étude protéomique différentielle menée résulte des différentes optimisations réalisées au cours du *Chapitre II-III-1*. Il est résumé par la Figure II-1.

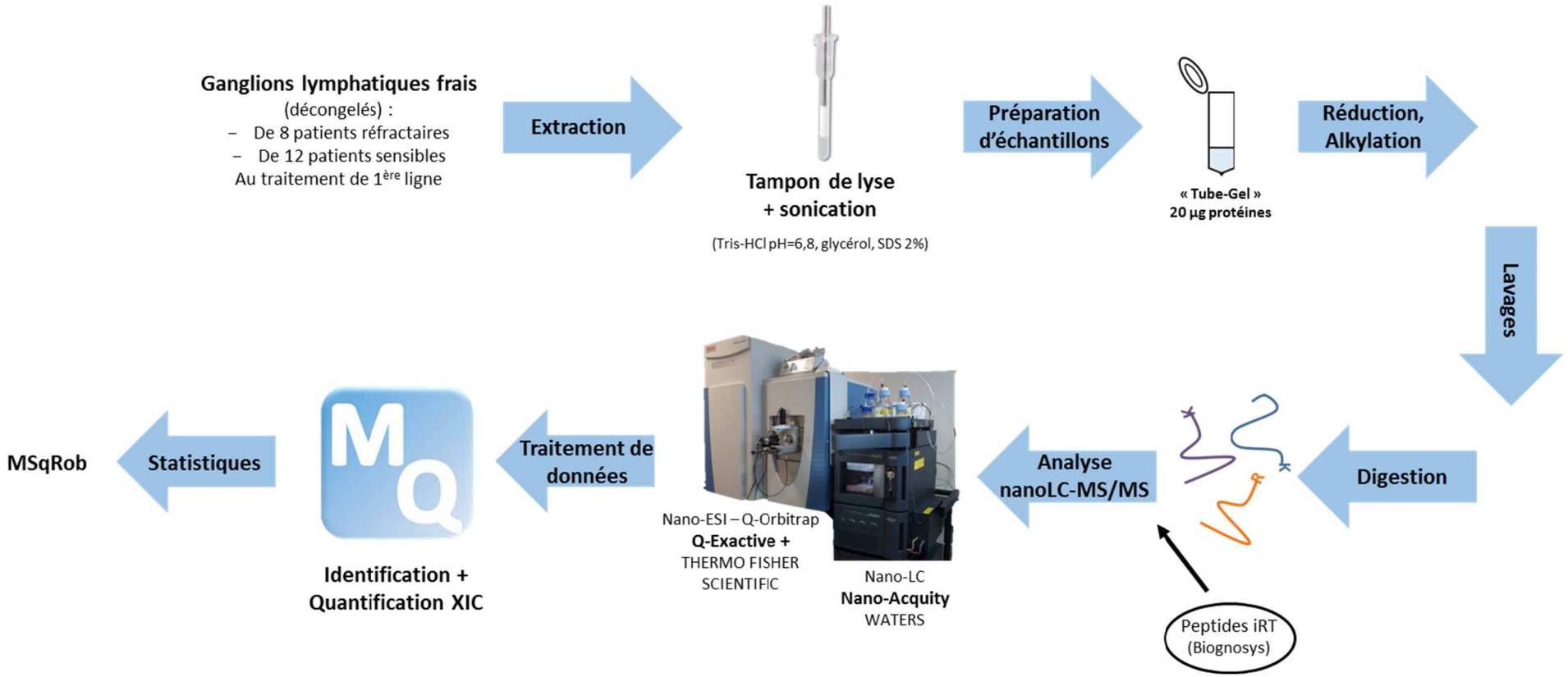


Figure II-1 - Schéma analytique employé afin d'établir une signature protéique permettant de déceler précocement une chimiorésistance

A partir de l'extraction des protéines de tissus frais, vingt microgrammes de protéines ont été préparés en TG et ont fait l'objet d'une analyse nanoLC-MS/MS sur un spectromètre de masse de type Q-Orbitrap (Q-Exactive +, THERMO FISHER SCIENTIFIC). Les données ont été traitées à l'aide du logiciel MaxQuant, et le traitement statistique des données de quantification sans marquage XIC a été effectué à partir des données peptidiques, à l'aide du « *package* » MSqRob[219] disponible sous R, par le Dr Frédéric BERTRAND, statisticien à l'Institut de Recherche Mathématique Avancée (IRMA) de Strasbourg. Cet outil statistique permet d'évaluer les ratios protéiques directement à partir des intensités peptidiques, soit directement à partir des mesures effectuées par le spectromètre de masse, et réduit les biais inhérents aux approches employant les données protéiques inférées en tenant compte de la variabilité de réponses entre peptides. Il permet notamment d'éviter la perte de certaines protéines. De plus, il surpasse les approches protéiques dans le cas de recherche de protéines différentiellement exprimées au sein de données clairsemées en raison de données manquantes. En résumé, l'utilisation de cet outil s'explique par une meilleure exactitude, sensibilité et spécificité par rapport aux approches employant les données protéiques inférées.

3- Résultats

a. Qualitatifs

D'un point de vue qualitatif, 4822 protéines au total ont été identifiées avec au moins un peptide unique sur l'ensemble des 20 échantillons. De manière plus détaillée, le nombre de protéines identifiées pour chaque échantillon est donné par la Figure II-2.

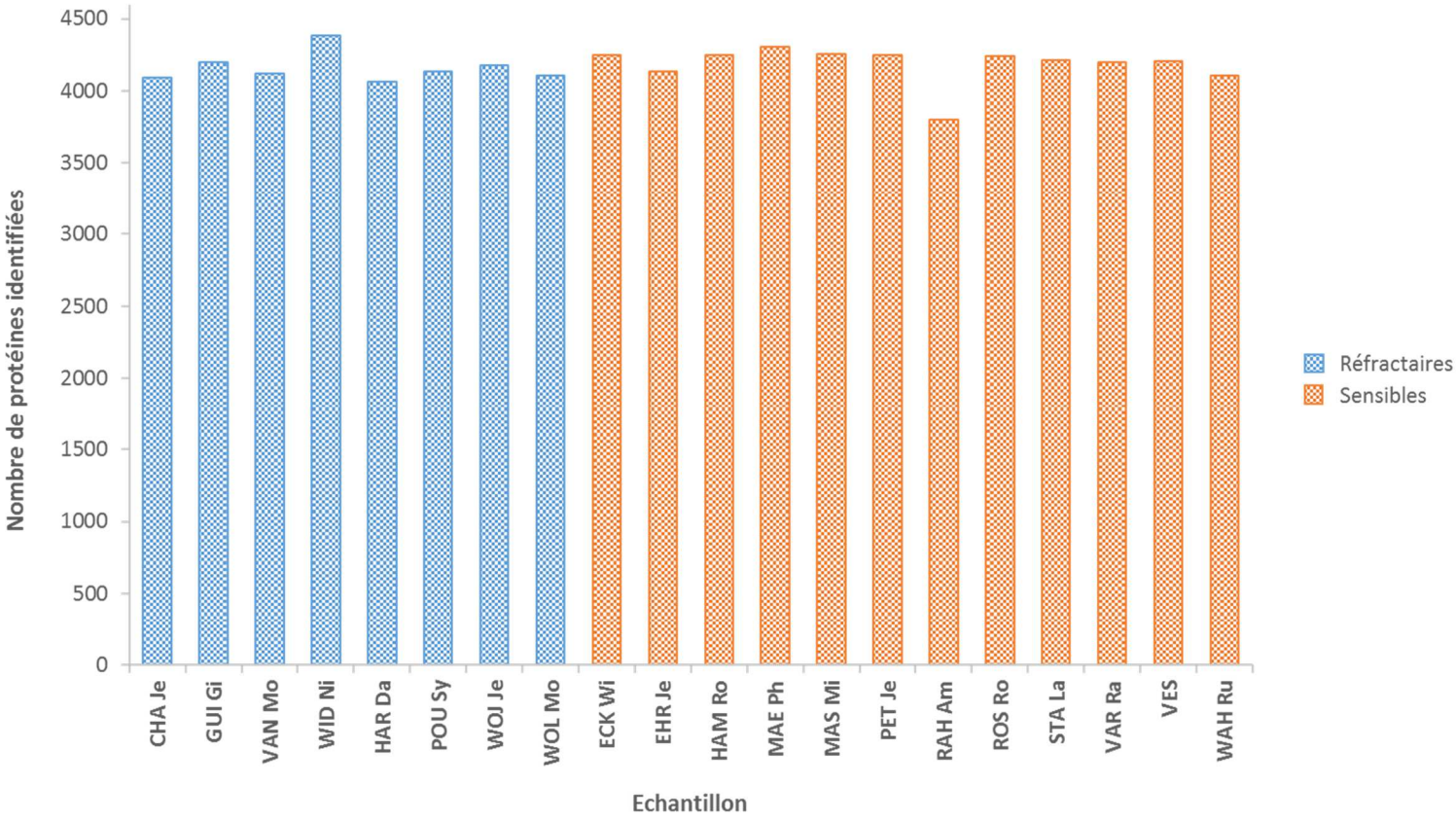


Figure II-2 - Nombre de protéines identifiées avec au moins un peptide unique dans chaque échantillon

Les identifications sont relativement similaires d'un échantillon à l'autre avec 4063 à 4385 protéines identifiées, hormis pour l'échantillon du groupe des sensibles (RAH Am) pour lequel un nombre plus faible, de 3802 protéines, a été identifié.

b. Quantitatifs

Contrôle qualité

Afin de s'assurer que l'ensemble des analyses a été effectué dans les mêmes conditions, la stabilité du couplage a été suivie à l'aide des onze peptides iRT dopés dans chacun des échantillons avant analyse nanoLC-MS/MS. Ainsi, dans un premier temps, les CV des Tr pour chacun des peptides, calculés sur l'ensemble des échantillons, a permis d'établir la boîte à moustache de la Figure II-3. Celle-ci permet de rendre compte, avec une valeur médiane inférieure à 1 %, de la stabilité du système chromatographique tout au long de la séquence d'analyses des échantillons.

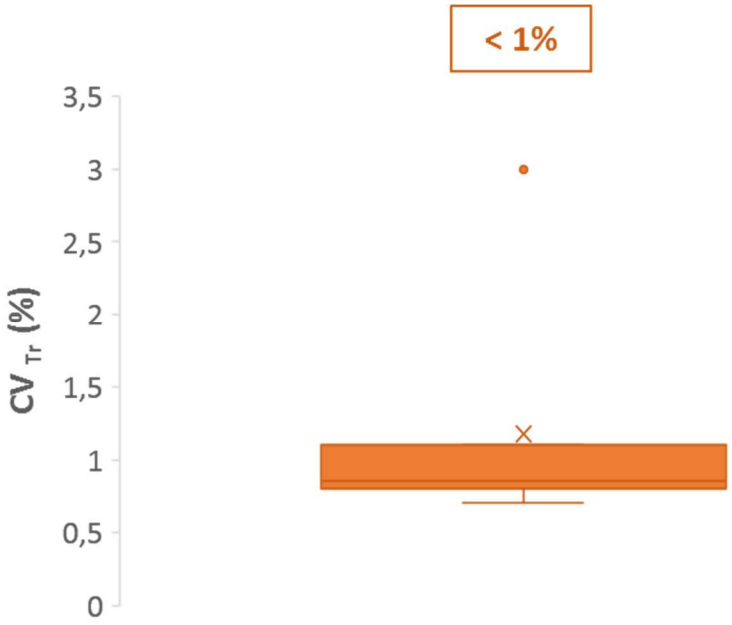


Figure II-3 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des Tr extraits de Skyline pour chaque peptide iRT considéré dans l'ensemble des échantillons.
La valeur encadrée correspond à la valeur médiane

Dans un deuxième temps, les aires extraites, pour chacun des peptides dont le rapport signal sur bruit était correct dans chacune des analyses (2 peptides ont été ôtés pour le calcul des CV), à l'aide du logiciel Skyline ont permis de calculer des CV permettant de suivre la stabilité du couplage. La répartition de ces CV est donnée par la Figure II-4. La valeur médiane de 16 % permet d'apprécier la bonne stabilité et la bonne répétabilité du couplage dans les conditions d'analyse.

Ce contrôle qualité a permis de rendre compte de la répétabilité, ainsi que de la robustesse du couplage nanoLC-MS/MS pendant les analyses des échantillons du projet de recherche de biomarqueurs associés à une chimiorésistance primaire dans les LBDGC.

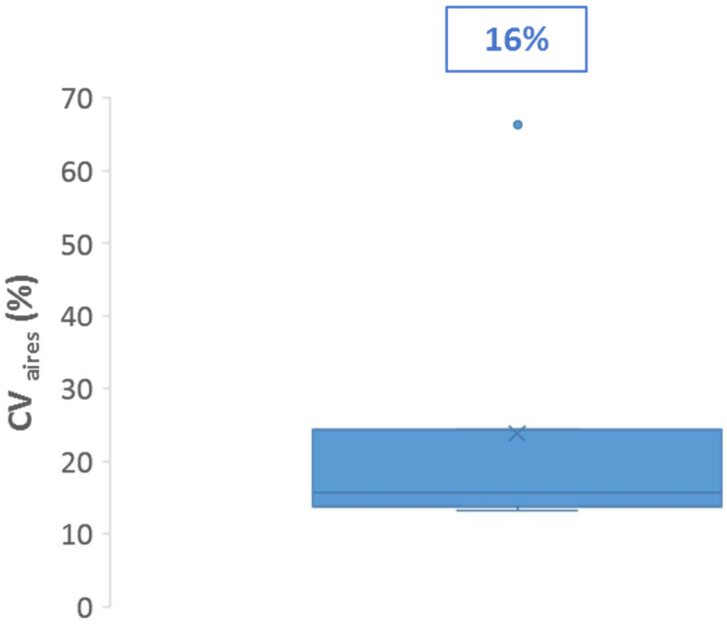


Figure II-4 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des valeurs d'aires extraites de Skyline pour chaque peptide iRT considéré, quantifié dans l'ensemble des échantillons.
La valeur encadrée correspond à la valeur médiane

Echantillons de ganglions lymphatiques de LBDGC

En ce qui concerne les échantillons des deux cohortes, les profils globaux de l'ensemble des intensités protéiques pour chaque échantillon ont été comparés à l'aide la Figure II-5.

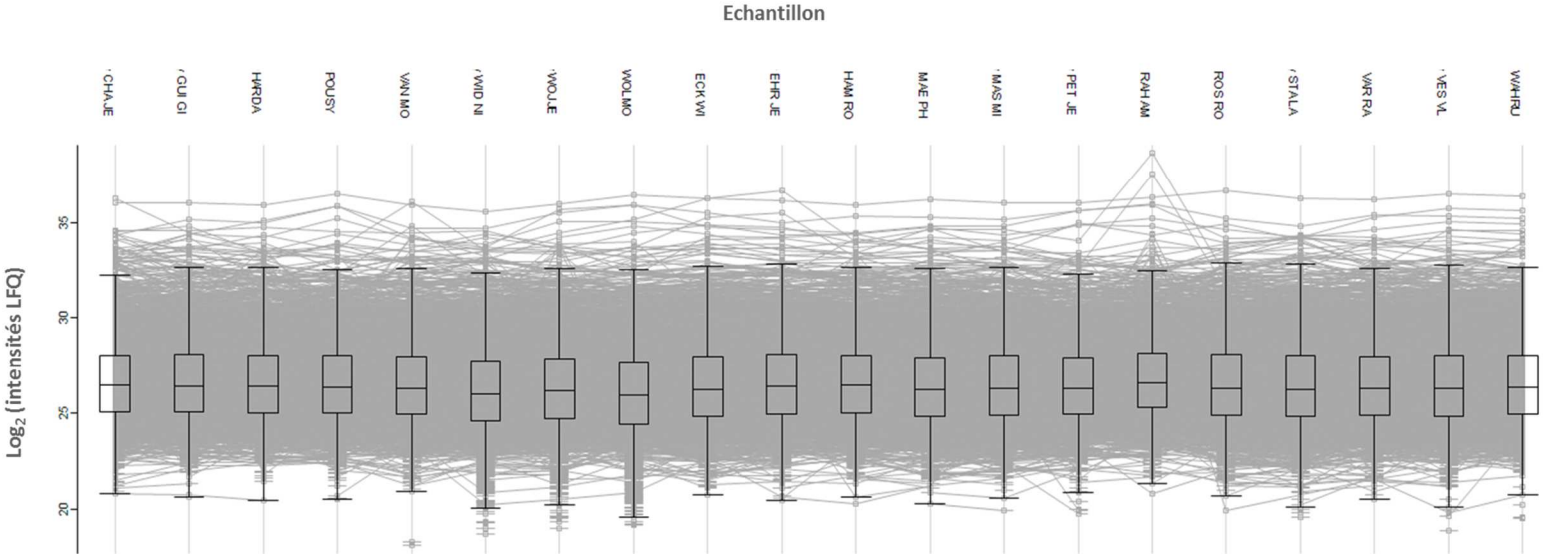


Figure II-5 - Profil de l'ensemble des intensités protéiques pour chacun des 20 échantillons analysés

Ces profils semblent, tout comme au niveau qualitatif, relativement similaires entre les différents échantillons. L'échantillon qui présentait un nombre plus faible de protéines identifiées (RAH Am) présente tout de même un profil d'intensités semblable aux autres échantillons. De surcroît, le

coefficient de corrélation moyen calculé à partir des coefficients de corrélation établis à partir des valeurs d'intensités normalisées de chaque échantillon entre eux est de 0,89.

Une fois la qualité des données attestée, un test statistique opposant le groupe des réfractaires au groupe des sensibles a été mené à partir des intensités peptidiques, à l'aide du modèle « *Peptide-level Robust Ridge Regression* » décrit par Ludger GOEMINNE *et collaborateurs*. Au total 586 protéines, couvrant trois ordres de magnitude (Figure II-6) ont été identifiées comme étant différentiellement exprimées entre les deux groupes avec une « *p-value* » et un taux de faux-positifs inférieurs à 5 %. La liste de ces protéines différentiellement exprimées est donnée en *Annexe 1*.

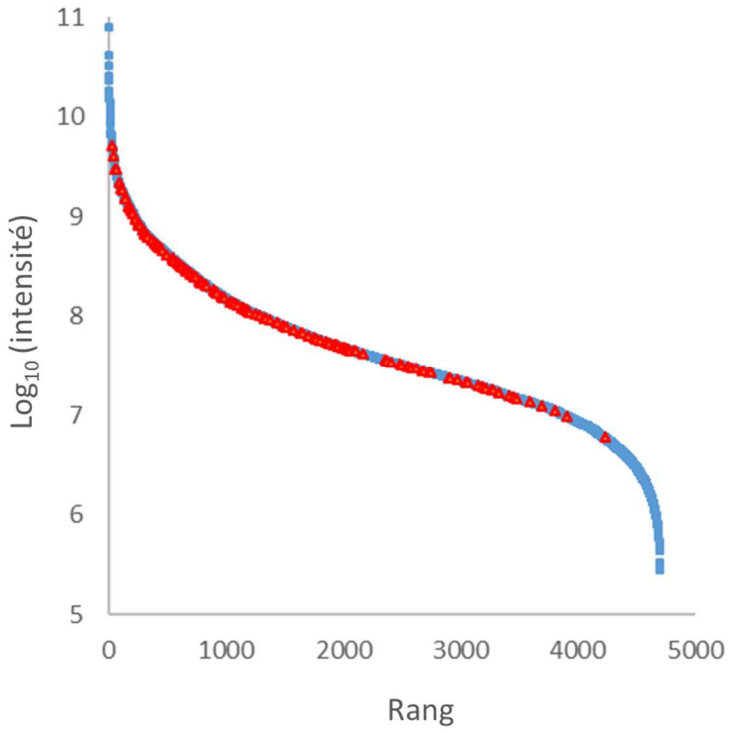


Figure II-6 - Gamme dynamique de l'ensemble des échantillons - en rouge sont représentées les protéines différentiellement exprimées (avec un FDR<5 %) après avoir effectué un test statistique opposant le groupe des réfractaires au groupe des sensibles

En parallèle des analyses protéomiques, des analyses transcriptomiques (RNA-séq) ont été menées sur 17 de ces échantillons, en collaboration avec les Drs Seiamak BAHRAM et Raphaël CARAPITO du laboratoire d'immuno-rhumatologie moléculaire de Strasbourg. En effet, pour un échantillon du groupe des réfractaires et deux échantillons du groupe des sensibles, la totalité du matériel de départ a été utilisé pour l'extraction des protéines, et n'était de ce fait plus disponible pour effectuer une extraction d'ARN. A partir des données de comptage normalisées par la taille des gènes de 17535 gènes, un test statistique de Wald opposant le groupe des réfractaires au groupe des sensibles a été effectué. En fixant une valeur seuil de « *p-value* » ajustée inférieure à 0,1, au total 244 gènes ont été

identifiés comme étant différentiellement exprimés entre les deux groupes. La liste des gènes différentiellement exprimés est donnée en *Annexe 2*.

L'intégration des deux types de données à l'aide du logiciel Perseus a permis d'exprimer 2936 ratios protéines/gènes. A partir de ces données, vingt-deux protéines ont été trouvées comme étant différentiellement exprimées à la fois en protéomique et en transcriptomique (Tableau II-1). Au sein de ces 22, seize protéines ont été trouvées comme étant surexprimées dans les deux types de données (Figure II-7). Parmi ces 16 protéines, nous trouvons IDO1 (P14902) et PREX1 (Q8TCU6).

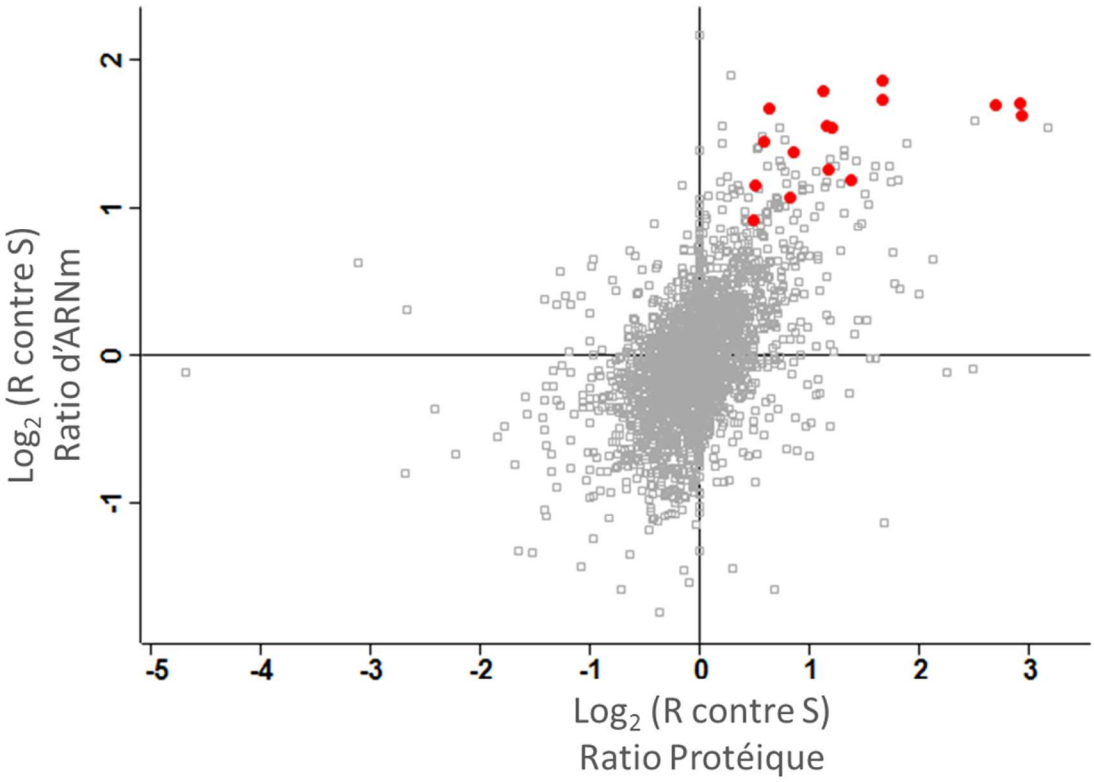


Figure II-7 – Visualisation conjointe des ratios protéiques et d'ARN messager (Résistant/Sensible)
En rouge sont représentées les 16 protéines surexprimées dans le groupe des réfractaires à la fois en transcriptomique et en protéomique

N° accession protéine	Ratio protéique	Ratio transcrits	Nom de gène
O43927	2,95	1,63	CXCL13
P14902	2,92	1,71	IDO1
P20718	2,70	1,70	GZMH
P52790	1,67	1,86	HK3
P05109	1,67	1,73	S100A8
O75558	1,39	1,20	STX11
P35237	1,20	1,55	SERPINB6
Q5TEJ8	1,19	1,27	THEMIS2
P04179	1,16	1,56	SOD2
P01009	1,14	1,80	SERPINA1
P26447	0,85	1,38	S100A4
Q8TCU6	0,83	1,07	PREX1
P00450	0,64	1,68	CP
P49863	0,58	1,46	GZMK
P48960	0,51	1,15	CD97
O75915	0,49	0,92	ARL6IP5
P40259	-1,65	-1,32	CD79B
P25398	-0,43	-0,88	RPS12
P22087	-0,39	-0,80	FBL
P62753	-0,31	-0,94	RPS6
P62269	-0,26	-1,07	RPS18
P01024	0,31	-1,44	C3

Tableau II-1 - Liste des protéines et gènes différentiellement exprimés dans les données de protéomique et transcriptomique avec les ratios (Résistant/Sensible) correspondants

- En ce qui concerne la protéine IDO1 (ou Indoleamine 2,3-dioxygenase 1), elle joue un rôle dans la dégradation du tryptophane, générant un microenvironnement immunosuppresseur qui provoque l'apoptose des cellules T ayant pour rôle de détruire les cellules tumorales, et induit l'augmentation des cellules T régulatrices qui régulent la quantité des cellules T. L'inhibition de cette protéine pourrait potentiellement améliorer l'efficacité des immunothérapies chez les patients réfractaires[220].

- Pour ce qui est de PREX1 ou (Phosphatidylinositol 3,4,5-trisphosphate-dependent Rac exchanger 1), elle est connue pour être surexprimée dans de nombreux cancers tels que le cancer du sein, de la prostate ou le mélanome. Avec PREX2 elle joue un rôle oncogénique dans les cancers humains en stimulant la migration et l'invasion des cellules tumorales[221].

A l'aide de l'outil d'enrichissement 2D du logiciel Perseus[222] intégrant les annotations GO et KEGG (pour « *Kyoto Encyclopedia of Genes and Genomes* »), il a été possible de mettre en évidence des voies de signalisation enrichies dans le groupe des réfractaires, comme la cascade du complément et de la coagulation. Des études suggèrent que la cascade du complément est impliquée dans le développement de tumeurs, du fait de la génération d'un microenvironnement immunosuppresseur[223].

4- Conclusion et perspectives

Les deux protéines, ainsi que les deux voies de signalisation mises en évidence par cette étude, doivent être évaluées à plus grande échelle sur une cohorte indépendante afin de valider qu'elles sont bien signatures de la résistance aux traitements de première ligne dans les LBDGC. Pour ce faire, des méthodes biologiques telles que la cytométrie en flux ou encore la spectrométrie de masse ciblée avec des stratégies de type SRM peuvent être employées. Les tissus inclus en paraffine représentent une manne pour la constitution de grandes cohortes d'échantillons pour cette étape d'évaluation/de validation de potentiels biomarqueurs. C'est pour cette raison que les développements méthodologiques du *Chapitre II-III-2* ont été menés et ont démontré la faisabilité et la fiabilité de l'extraction de protéines à partir de tissus FFPE permettant d'envisager une validation des potentiels biomarqueurs mis en évidence dans cette étude sur une grande cohorte d'échantillons FFPE.

En parallèle, des développements bioinformatiques vont être menés à partir des spectres MS/MS peptidiques n'ayant pas mené à une identification lors de cette étude de découverte. Les cancers étant souvent liés à des mutations, les résultantes protéiques (variants de séquences) ne sont pas présentes dans les banques de séquences protéiques alors qu'elles ont probablement générées des spectres de fragmentation. Ces protéines pourraient cependant être des signatures idéales de ce type de maladie. Ainsi, l'intégration de données de transcriptomique pourrait dans ce cas :

- Permettre de proposer des séquences protéiques propres à la maladie ou aux échantillons (spécialisées ou personnalisées), afin d'effectuer une nouvelle recherche protéique à l'aide de la nouvelle banque ainsi générée,

- Ou permettre, après avoir mis en évidence la surexpression d'un spectre MS/MS caractéristique de la fragmentation d'un peptide après un test statistique incluant à la fois les spectres MS/MS ayant mené à une identification ainsi que ceux n'ayant pas mené à une identification, d'identifier le peptide/la protéine correspondant(e) par séquençage *de novo*.

III-Projet de recherche de biomarqueurs de suivi de greffons rénaux

Ce projet de recherche de biomarqueurs a été mené en binôme avec le Dr David MARX, néphrologue, actuellement étudiant en thèse de recherche au laboratoire, et en collaboration avec le Pr Sophie CAILLARD des HUS.

1- Contexte du projet

La transplantation rénale représente une thérapie de choix pour les patients atteints de maladies rénales au stade terminal, qui améliore leur survie et leur qualité de vie[172, 224, 225]. Elle est actuellement réalisée sur 80 000 patients dans le monde par an. Malgré les améliorations des protocoles d'immunosuppression et de surveillance des patients, qui ont notamment permis d'augmenter la durée de vie à court terme du greffon, le rendement de survie des greffons reste à l'heure actuelle insuffisant avec 10 à 20 % des greffés qui font un rejet dans les soixante premiers jours post-transplantation[225, 226], en raison de phénomènes inflammatoires au niveau du greffon, ou de maladies glomérulaires, de fibroses, etc. De nos jours, une augmentation de la créatinine plasmatique permet de rendre compte d'une complication mais de manière tardive et non spécifique. En effet, le suivi de cette valeur ne permettant pas de différencier les différentes affections, une biopsie du greffon est souvent réalisée pour établir un diagnostic, souvent tardif[225-228]. Cet acte est cependant invasif, chronophage, et peut entrainer des complications. De ce fait, il n'est pas envisageable de le réaliser de manière fréquente sur le long terme. C'est pourquoi il y a un réel besoin d'une méthode de diagnostic non invasive, sensible, et spécifique des complications pouvant survenir sur le greffon rénal. Cette méthode nécessite de trouver des biomarqueurs fiables et précoces, qui permettraient d'évaluer de manière rapide ce qui cause les dommages au greffon, et ainsi de mettre en place un traitement précoce et préventif du rejet avant l'élévation de la créatininémie[172, 224-228]. L'urine apparaît comme une matrice de choix (voir *Chapitre II-IV*) pour partir à la recherche d'un jeu de biomarqueurs universel au panel de complications permettant le suivi du greffon, et d'éviter la biopsie fréquente et parfois inutile.

C'est dans ce contexte qu'une étape de découverte de biomarqueurs de suivi de l'état du greffon rénal dans l'urine a été menée par une approche de protéomique quantitative différentielle sans marquage XIC, opposant 16 échantillons de patients greffés stables à 16 échantillons de patients présentant des complications. Cette étude a fait l'objet d'un financement de pré-maturation par la SATT Conectus de Strasbourg. Elle s'inscrit dans un contexte particulier, celui de mettre au point une méthode

diagnostique applicable en routine en clinique, c'est-à-dire dont la préparation d'échantillons soit simple, rapide et robuste (c'est-à-dire applicable à n'importe quel échantillon d'urine).

2- Stratégie analytique employée

Cette étude a été menée sur 96 échantillons sélectionnés parmi 5400 prélèvements urinaires, collectés et congelés au département de Néphrologie et Transplantation du Nouvel Hôpital Civil (NHC) de Strasbourg, pour des patients ayant subi une transplantation rénale après mai 2014, pour lesquels une biopsie concomitante au prélèvement urinaire permettant de statuer sur l'état du greffon a été effectuée, et ayant signé un consentement pour participer à des études sur les facteurs influençant le pronostic des transplantations rénales. A partir de ces 96 échantillons, 32 mélanges (ou « *pools* ») de 3 échantillons ont été réalisés. Le choix des « *pools* » a été fait de manière à homogénéiser les échantillons pour ainsi diminuer la variabilité inter-individus, étant donné que l'urine est un échantillon très variable d'un individu à l'autre, et renforcer le signal propre au statut représenté par le « *pool* ». Cette technique est d'ailleurs souvent employée pour les études de protéomique portant sur l'urine[172, 173]. Ainsi, deux cohortes ont été réalisées :

- La première cohorte est constituée de 16 « *pools* » « contrôles ». Il s'agit de 16 mélanges de 3 prélèvements urinaires de patients transplantés qui ne montrent pas de complication aiguë à la biopsie, et dont les fonctions rénales étaient stables dans les semaines précédant et suivant la biopsie.

- La seconde cohorte est constituée de 16 « *pools* » de patients transplantés, incluant 9 types de complications, non divulguées afin de respecter la confidentialité du projet.

De par l'implication du Dr David MARX au laboratoire, la sélection des échantillons a pu être rapide et pleinement maîtrisée par notre équipe.

Le schéma analytique employé est celui optimisé au *Chapitre II-IV* et détaillé dans la Figure III-1. Ainsi, après avoir décongelé les 96 prélèvements urinaires, ceux-ci ont été centrifugés afin de retirer les débris cellulaires. Le mélange de trois prélèvements provenant de patients différents a ensuite été effectué de manière à générer les deux cohortes de 16 « *pools* » qui ont été aliquotés en six exemplaires de 1,2 ml. Un aliquot a été envoyé au laboratoire de biochimie du NHC de Strasbourg, afin de déterminer la protéinurie et la créatininurie de chaque « *pool* », tandis que 500 µl d'un deuxième « *pool* » a été utilisé pour notre étude de découverte de biomarqueurs. Le choix de cette méthode de dosage de la protéinurie s'explique ici par l'objectif premier d'être dans des conditions similaires à

celles d'un test appliqué en routine à l'hôpital. Il est a noté que les aliquots restants ont été congelés et stockés à -80°C de manière préventive. Pour les 32 « *pools* » de notre étude, les protéines ont été concentrées, réduites, et alkylées à l'aide d'un filtre d'ultracentrifugation avec un seuil de coupure de 10 kDa suivant le protocole FASP. Au vu des dosages protéiques, deux groupes de digestion ont été générés de manière à garder des ratios trypsine/protéine acceptables :

- Un groupe d'échantillons comprenant au total 1 à 100 µg de protéines (les quantités réelles s'étalant de 13,5 à 95,5 µg de protéines),

- Et un groupe d'échantillons comprenant plus de 100 µg de protéines au total (avec des quantités réelles allant de 102 à 224 µg de protéines).

Ce choix s'explique ici aussi par la volonté de simplifier au maximum le protocole, dans le but de le rendre applicable à l'hôpital. Après extraction des peptides par SPE, chaque échantillon a été divisé en deux afin de mener deux séquences d'analyses distinctes :

- Une première série nommée « $Q_{éq}$ » pour laquelle la quantité protéique a été ajustée en fonction de la concentration protéique urinaire pour chaque échantillon, de manière à analyser une même quantité protéique pour chaque échantillon,

- Et une seconde série nommée « $V_{éq}$ » pour laquelle un même volume d'échantillon urinaire de départ a été analysé quelle que soit sa concentration protéique, comme c'est souvent le cas pour l'étude de fluides biologiques.

Des peptides iRT ont été dopés dans chacun des échantillons avant analyse nanoLC-MS/MS, effectuée sur un spectromètre de masse de type Q-Orbitrap. Les données brutes ont été traitées avec MaxQuant, et le traitement statistique a été effectué à partir des données peptidiques, issues de la quantification sans marquage XIC, par le Dr Frédéric BERTRAND à l'aide du « *package* » disponible sous R : MSqRob.

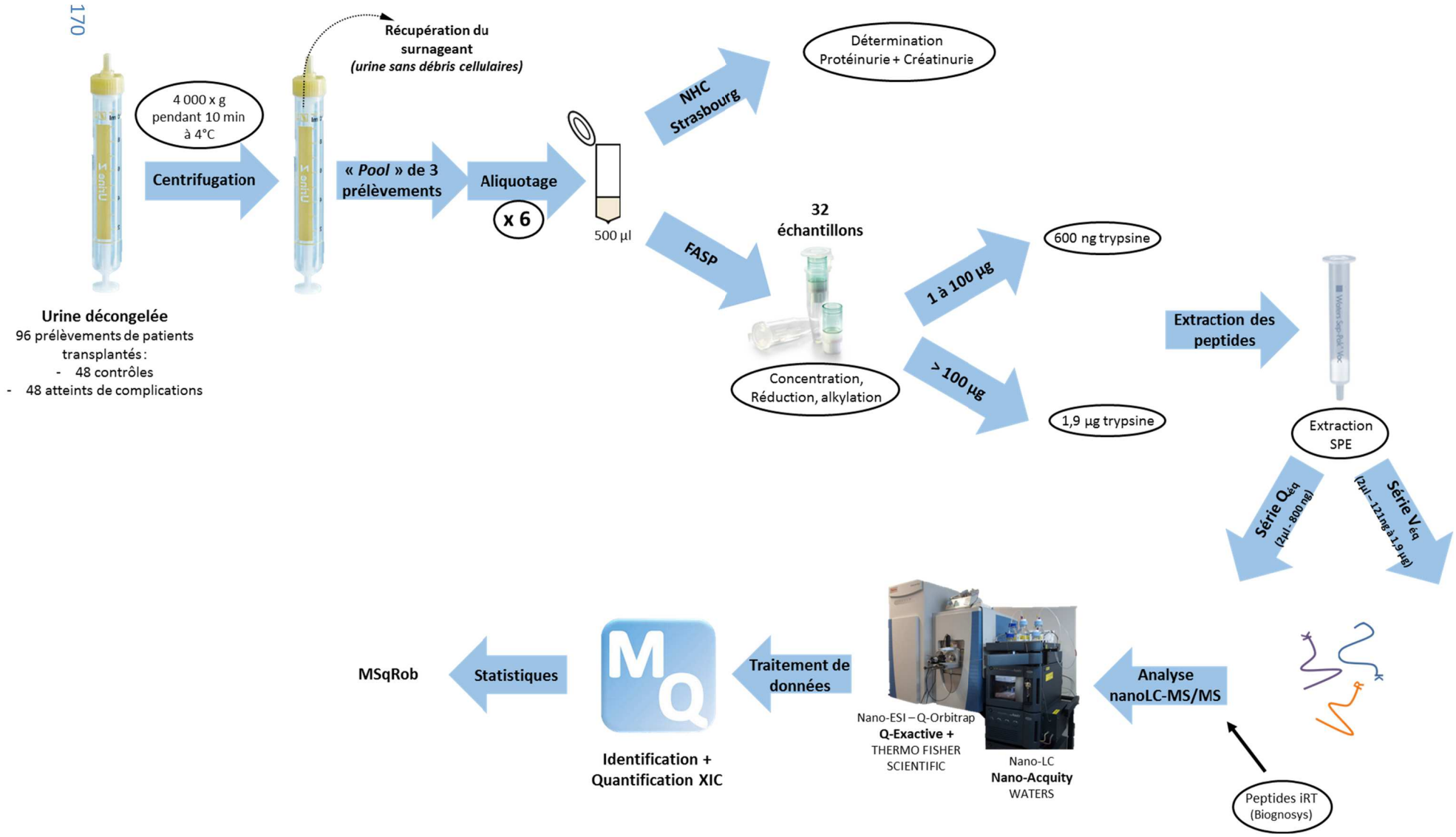


Figure III-1 - Schéma analytique employé pour la recherche de biomarqueurs urinaires permettant le suivi des patients ayant subi une transplantation rénale

3- Résultats

a. Evaluation de l'apport de l'option « *Match between runs* » de MaxQuant sur des séries différentes

Au-delà du projet à proprement parler, l'apport de l'option MaxQuant permettant le transfert des identifications d'un échantillon à l'autre, nommée « *Match between runs* », a pu être évaluée grâce aux deux séries distinctes ($V_{éq}$ et $Q_{éq}$). En effet, l'ensemble des données générées pour ces deux séries ont été traitées au cours d'une même analyse MaxQuant ($V_{éq}$+$Q_{éq}$), en activant cette option dans le but d'augmenter le nombre d'identifications dans chacun des groupes, tout en effectuant des normalisations séparées pour les deux conditions d'analyses ($V_{éq}$ et $Q_{éq}$). Afin d'en tirer des conclusions, les données des deux séries ont également été traitées de manière séparées ($V_{éq}$ et $Q_{éq}$ individuellement), avec le même logiciel, en activant ici aussi l'option « *Match between runs* ».

Dans un premier temps, le nombre total de protéines identifiées dans le traitement MaxQuant ($V_{éq}$+$Q_{éq}$), et les traitements individuels $V_{éq}$ et $Q_{éq}$ est donné par le Tableau III-1.

	Nombre de protéines identifiées Avec FDR < 1%	
	Série $Q_{éq}$	**Série $V_{éq}$**
Analyse individuelle ($Q_{éq}$ ou $V_{éq}$)	1315	1225
Analyse simultanée ($V_{éq}$+$Q_{éq}$)	1326 + 1%	1319 + 8%

Tableau III-1 - Nombre de protéines identifiées selon les différents traitements MaxQuant

Ces valeurs permettent de rendre compte que l'ajout des 32 analyses d'une série à un traitement MaxQuant d'une autre série avec l'option « *Match between runs* » activée, permet d'augmenter de 1 et 8 % le nombre de protéines identifiées par rapport au traitement individuel MaxQuant des 32 analyses de la série $V_{éq}$ et $Q_{éq}$. De manière plus détaillée, les Figures III-2 et III-3 illustrent la différence des pourcentages de protéines identifiées par MS/MS et par « *matching* » (du fait de l'option « *Match between runs* ») entre le traitement MaxQuant ($V_{éq}$+ $Q_{éq}$) par rapport au traitement individuel de la série $Q_{éq}$ et $V_{éq}$ respectivement.

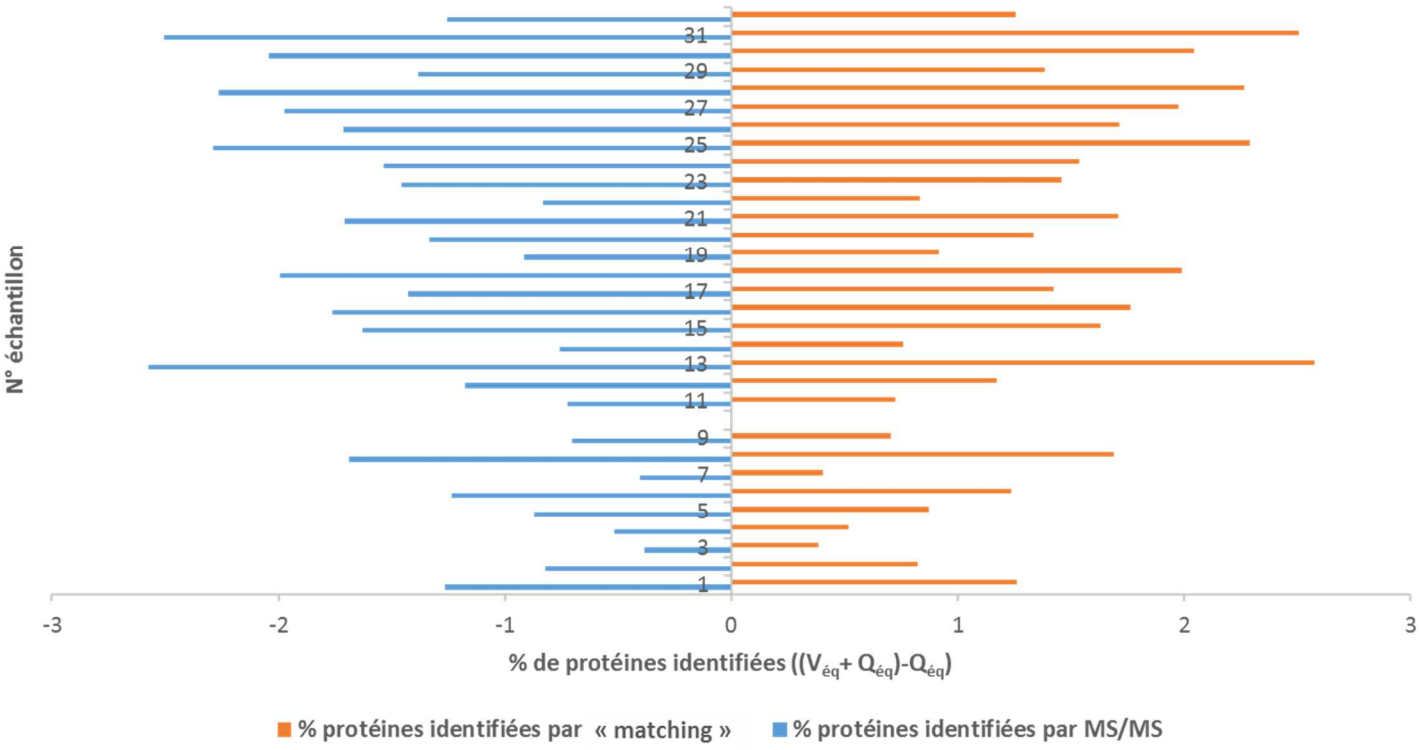


Figure III-2 - Pourcentage de protéines identifiées (par MS/MS et « *matching* ») résultant de la différence de pourcentages de protéines identifiées (par MS/MS et par « *matching* ») dans les échantillons $Q_{éq}$ entre le traitement ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $Q_{éq}$

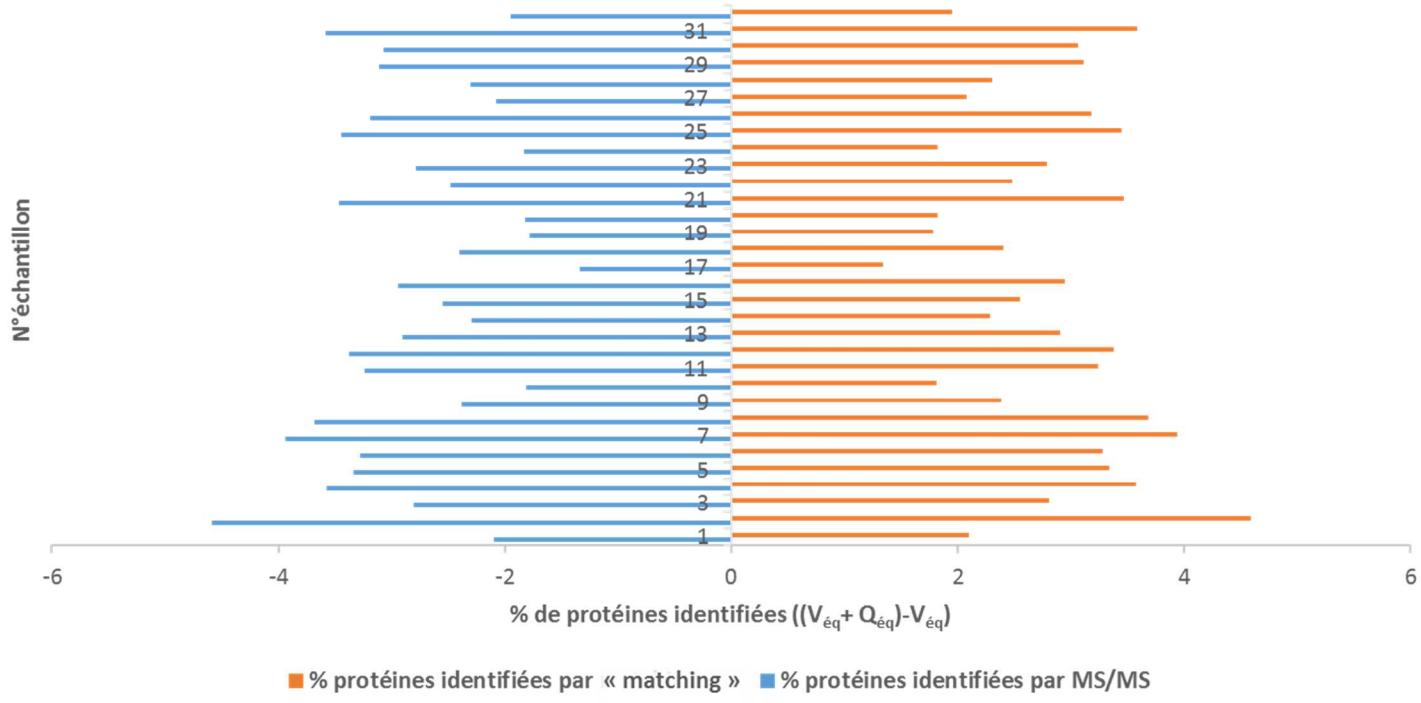


Figure III-3 - Pourcentage de protéines identifiées (par MS/MS et « *matching* ») résultant de la différence de pourcentages de protéines identifiées (par MS/MS et par « *matching* ») dans les échantillons $V_{éq}$ entre le traitement ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $V_{éq}$

Etonnement, l'apport des analyses de la série $V_{éq}$ et de l'option « *Match between runs* » au traitement des données de la série $Q_{éq}$ permet d'identifier moins de protéines par MS/MS que lors du traitement des analyses de la série $Q_{éq}$ seule (alors que le nombre de spectres MS/MS acquis dans la série $Q_{éq}$ et donnant lieu à une identification ne devrait pas varier). En réalité, cette différence s'explique par l'étape de validation à 1 % de FDR qui s'effectue sur des jeux de données de taille différente dans le traitement simultané et dans le traitement individuel. En contrepartie, le traitement MaxQuant ($V_{éq}$+$Q_{éq}$) permet d'augmenter le nombre de protéines identifiées par « *matching* », avec 1 à 2 % de protéines supplémentaires identifiées avec l'ajout d'analyses de la série $V_{éq}$ à la série $Q_{éq}$, et l'activation de l'option « *Match between runs* ». Cette même tendance, avec cependant des valeurs légèrement plus élevées (+2 à 4 % d'identifications par « *matching* » et -2 à 4 % par MS/MS), est observée pour la comparaison traitement MaxQuant ($V_{éq}$+$Q_{éq}$) – MaxQuant $V_{éq}$. Les mêmes observations ont été effectuées au niveau des identifications peptidiques.

Par la suite, les mêmes comparaisons de traitements MaxQuant ont été effectuées, mais en s'intéressant cette fois-ci aux pourcentages de protéines quantifiées. Les Figures III-4 et III-5 illustrent la différence de pourcentages de protéines quantifiées (intensités brutes et intensités normalisées LFQ) entre le traitement MaxQuant ($V_{éq}$+$Q_{éq}$) par rapport au traitement individuel de la série $Q_{éq}$ et $V_{éq}$ respectivement.

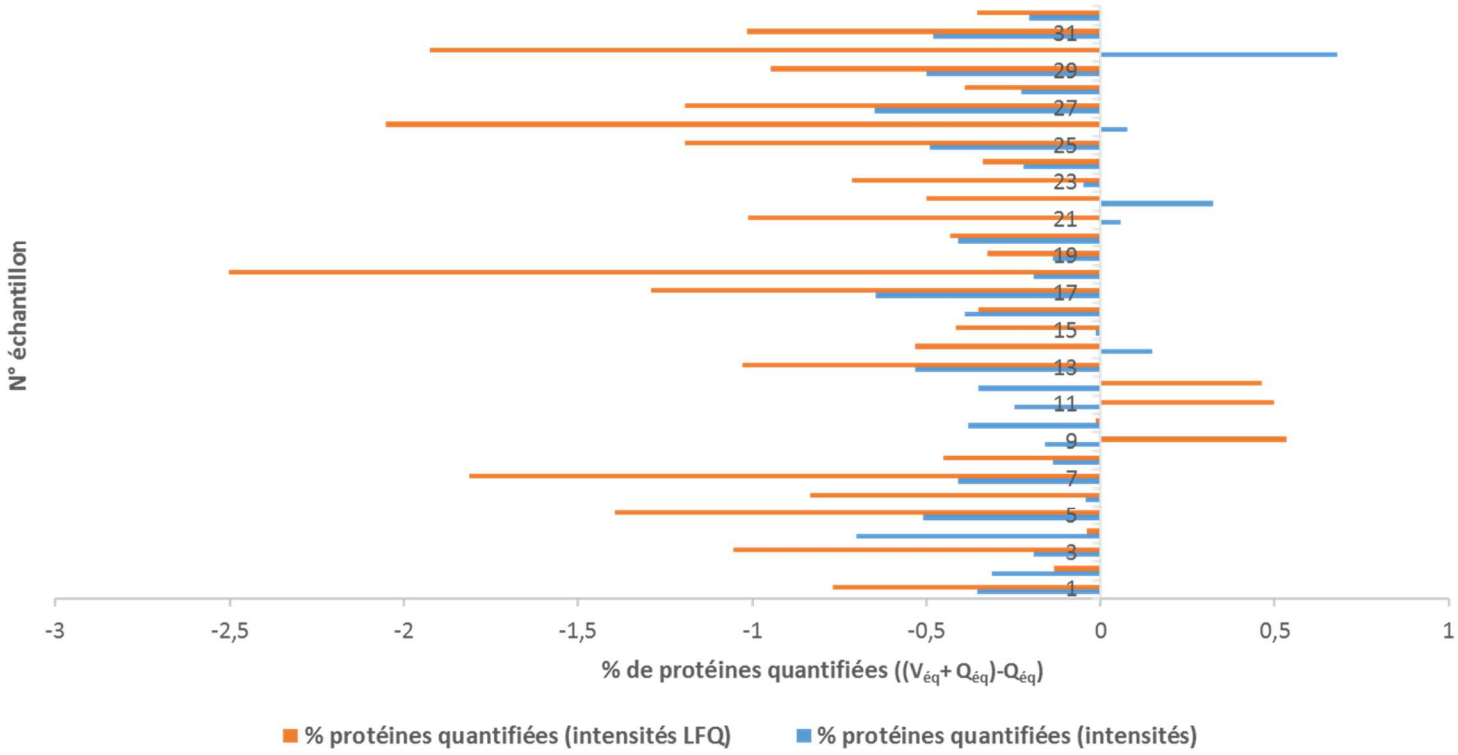


Figure III-4 - Pourcentage de protéines quantifiées (intensités brutes et intensités normalisées LFQ) résultant de la différence de pourcentages de protéines quantifiées dans les échantillons $Q_{éq}$ entre le traitement MaxQuant ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $Q_{éq}$

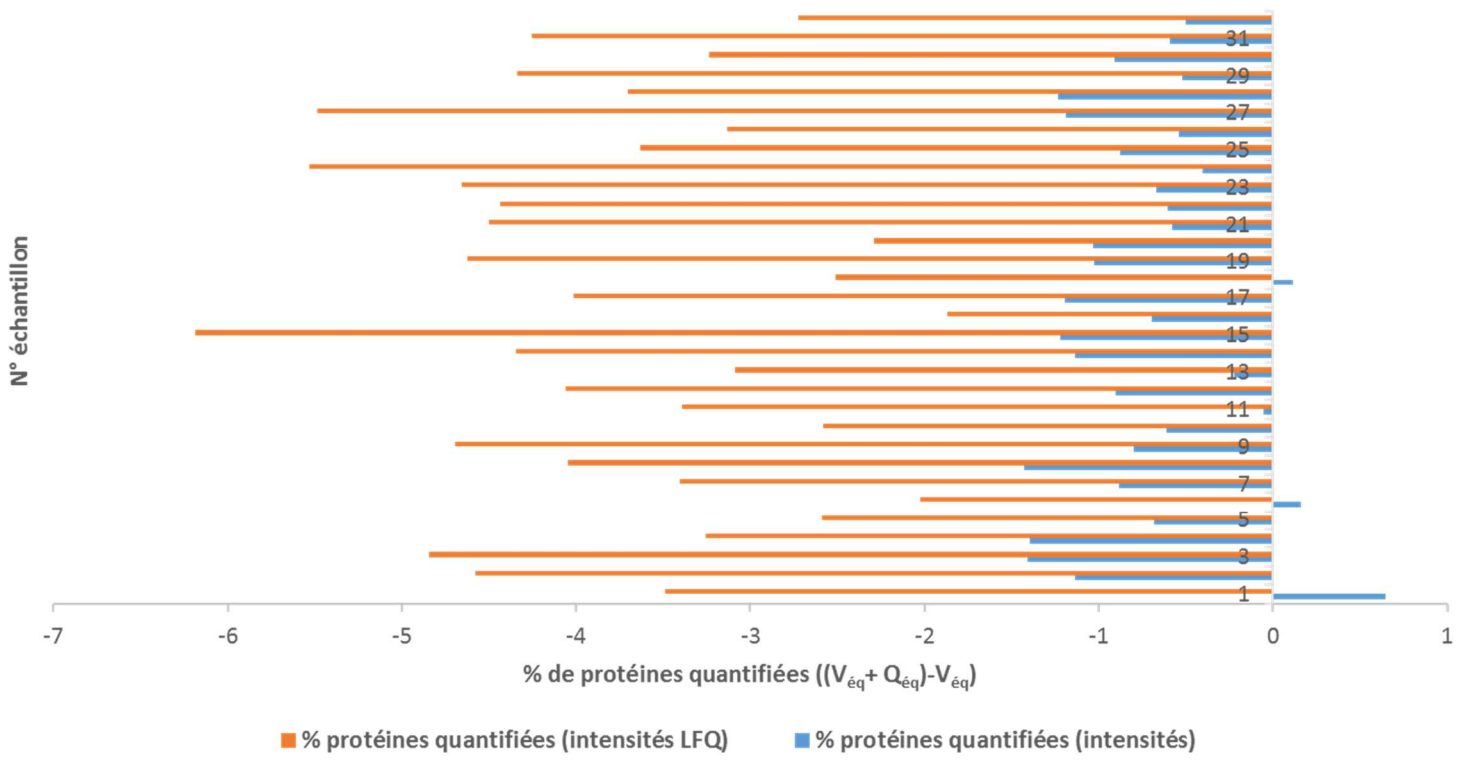


Figure III-5 - Pourcentage de protéines quantifiées (intensités brutes et intensités normalisées LFQ) résultant de la différence de pourcentages de protéines quantifiées dans les échantillons $V_{éq}$ entre le traitement MaxQuant ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $V_{éq}$

Pour la comparaison du traitement MaxQuant ($V_{éq}$+$Q_{éq}$) par rapport au traitement MaxQuant $Q_{éq}$ individuel, le pourcentage de protéines quantifiées est, pour la majorité des échantillons, plus faible pour le traitement simultané des séries par rapport au traitement individuel (Figure III-4). Le nombre de protéines identifiées étant plus important dans l'analyse simultanée des deux séries, la différence de pourcentages de protéines quantifiées est en défaveur du traitement simultané, alors que la tendance inverse est observée en nombre de protéines quantifiées, notamment au niveau des intensités brutes des protéines (Figure III-6). Pour les intensités normalisées LFQ, le nombre de protéines quantifiées est pour quelques échantillons tout de même en défaveur d'un traitement simultané des deux séries.

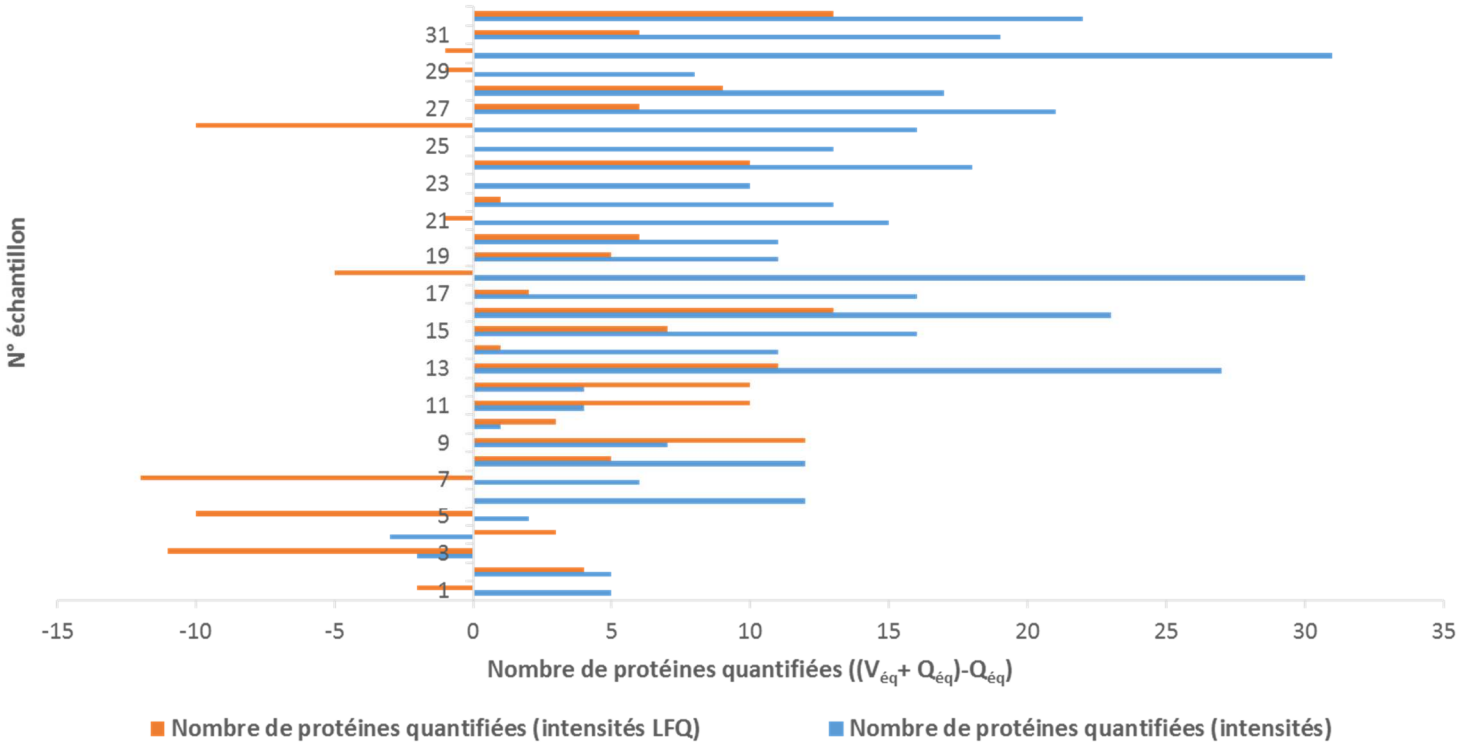


Figure III-6 - Nombre de protéines quantifiées (intensités brutes et intensités normalisées LFQ) résultant de la différence du nombre de protéines quantifiées dans les échantillons $Q_{éq}$ entre le traitement MaxQuant ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $Q_{éq}$

Pour la comparaison du traitement MaxQuant ($V_{éq}$ + $Q_{éq}$) par rapport au traitement MaxQuant $V_{éq}$, il est clair que malgré l'apport de protéines identifiées, le pourcentage de protéines quantifiées est moins important dans le traitement simultané des deux séries par rapport au traitement individuel (Figure III-5). Cette différence est d'autant plus marquée lorsque les intensités sont normalisées. En nombre, le traitement simultané est clairement favorable lorsqu'uniquement les intensités brutes sont considérées, alors qu'il est défavorable lorsque l'emploi des intensités normalisées est envisagé (Figure III-7).

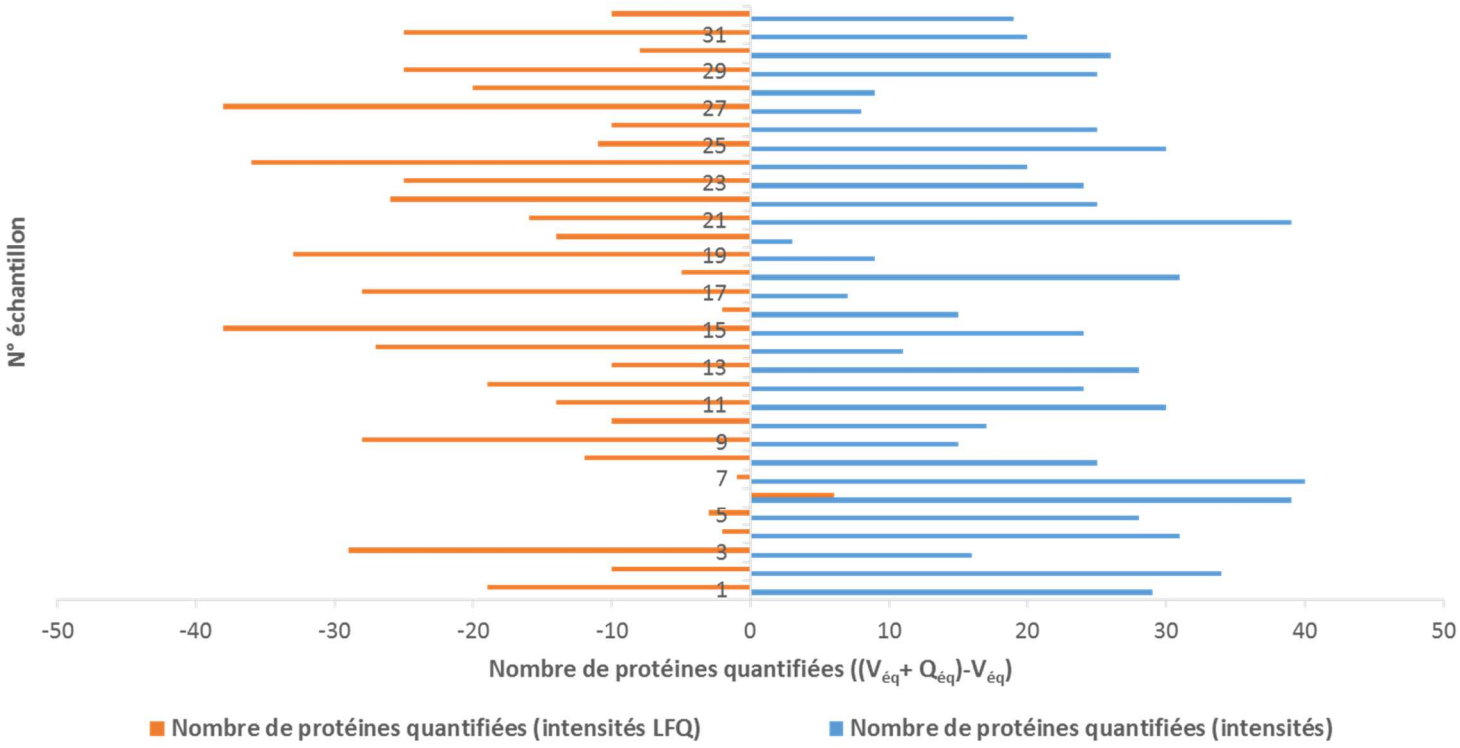


Figure III-7 - Nombre de protéines quantifiées (intensités brutes et intensités normalisées LFQ) résultant de la différence du nombre de protéines quantifiées dans les échantillons $V_{éq}$ entre le traitement MaxQuant ($V_{éq}$+ $Q_{éq}$) et le traitement individuel de $V_{éq}$

Au niveau des peptides quantifiés, aucune différence n'a été notée entre les deux types de traitement (simultané et individuel) lors de l'établissement de la différence de pourcentages de peptides quantifiés, alors qu'en nombre, une moyenne de 230 et 82 peptides supplémentaires (sur un nombre de 5100 et 4800 respectivement) sont quantifiés lors d'un traitement simultané des deux séries par rapport au traitement individuel de $Q_{éq}$ et $V_{éq}$ respectivement.

En conclusion, l'option « *Match between runs* » lors du traitement simultané des deux séries ($V_{éq}$ et $Q_{éq}$), permet d'augmenter le nombre de protéines totales identifiées, notamment pour la série $V_{éq}$. En ce qui concerne la quantification, le traitement simultané permet de quantifier moins de protéines en proportion, mais davantage en nombre. Cependant, si l'étude est menée avec les intensités normalisées LFQ, l'analyse simultanée des deux séries ne permet pas d'augmenter le nombre de protéines quantifiées, au contraire. De plus, la normalisation a globalement pour effet de diminuer le nombre de protéines quantifiées.

Au final, vu le faible apport du traitement simultané, et dans le but de simplifier le traitement à l'aide du « *package* » MSqRob en réduisant le fichier en ôtant les échantillons de la série $V_{éq}$, il a été décidé d'utiliser le traitement MaxQuant individuel de la série $Q_{éq}$ pour la recherche de biomarqueurs. Il n'a pas été jugé judicieux d'utiliser la série $V_{éq}$ pour la recherche de biomarqueurs, puisque des quantités

analysées différentes peuvent fausser les conclusions d'expression différentielle entre les deux groupes. De plus, l'analyse de volumes équivalents a mené, pour certains échantillons, à l'analyse d'une très faible quantité de protéines, diminuant ainsi la gamme dynamique couverte par l'analyse de ces échantillons.

b. Résultats qualitatifs

Comme évoqué précédemment, 1315 protéines au total ont été identifiées avec au moins un peptide unique sur l'ensemble des 32 échantillons « *pools* ». De manière plus détaillée, le nombre de protéines identifiées pour chaque échantillon est illustré par la Figure III-8.

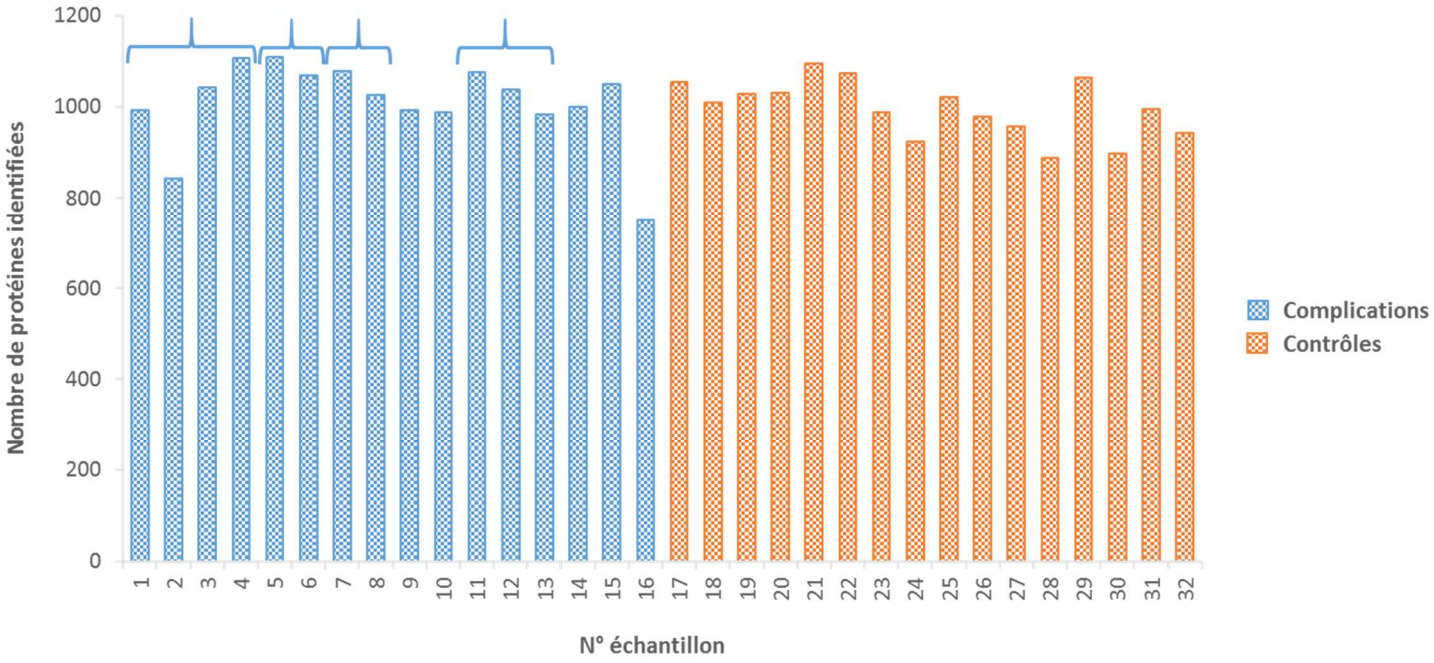


Figure III-8 - Nombre de protéines identifiées avec au moins un peptide unique dans chaque échantillon.
Les accolades signalent des réplicats biologiques

Les identifications sont relativement similaires d'un échantillon à l'autre dans le groupe contrôle avec 898 à 1065 protéines identifiées, tandis que pour le groupe des complications, les identifications vont de 750 à 1109 protéines identifiées, avec des nombres d'identifications plus faibles observés pour les échantillons 2 et 16. Il est à noter que l'échantillon 16 correspond à un échantillon pour lequel aucun autre réplicat biologique n'était disponible, et il est difficile de dire si ce nombre de protéines identifiées plus faible est propre à la complication qu'il représente ou si il est dû à un problème de l'échantillon ou de l'analyse de l'échantillon en lui-même. L'échantillon 2 est quant à lui un des quatre réplicats biologiques de la complication représentée par les échantillons 1, 2, 3 et 4 qui sont tous relativement différents les uns des autres en termes d'identification, ce qui laisse à penser que cette complication présente des variabilités d'un échantillon à l'autre.

Par ailleurs, il est à noter que les échantillons 18 et 28 n'ont été constitués que de deux prélèvements urinaires de patients différents en raison d'un troisième prélèvement contaminé, du fait d'une fissure du pot de recueil lors du stockage. Cela ne semble cependant pas affecter les résultats d'identification de ces échantillons.

c. Résultats quantitatifs

Contrôle qualité

Les peptides iRT, dopés dans chacun des échantillons, ont permis de suivre la stabilité du couplage dans les conditions d'analyse, tout comme pour les précédents projets de recherche de biomarqueurs. Parmi ces onze peptides, deux peptides ainsi qu'un des deux états de charge d'un troisième peptide ont dû être ôtés car ils présentaient des signaux instables sur l'ensemble des analyses. Ainsi, dans un premier temps, la répartition des coefficients de variation calculés à partir des Tr sur l'ensemble des échantillons pour chaque peptide ont permis d'établir la boîte à moustache de la Figure III-9. La valeur médiane des CV inférieure à 1 %, nous permet de conclure que le système chromatographique, dans les conditions d'analyse, est resté stable tout au long de la séquence. Par ailleurs, la Figure III-10 permet d'apprécier la répartition des CV calculés à partir des aires obtenues par le logiciel Skyline. La valeur médiane de 20 % reflète la bonne stabilité et la bonne répétabilité du couplage dans les conditions d'analyse.

Ce contrôle qualité a permis de rendre compte de la répétabilité, ainsi que de la robustesse du couplage nanoLC-MS/MS pendant les analyses de l'ensemble des échantillons de la série $Q_{éq}$ du projet de recherche de biomarqueurs urinaires pour le suivi de greffons rénaux.

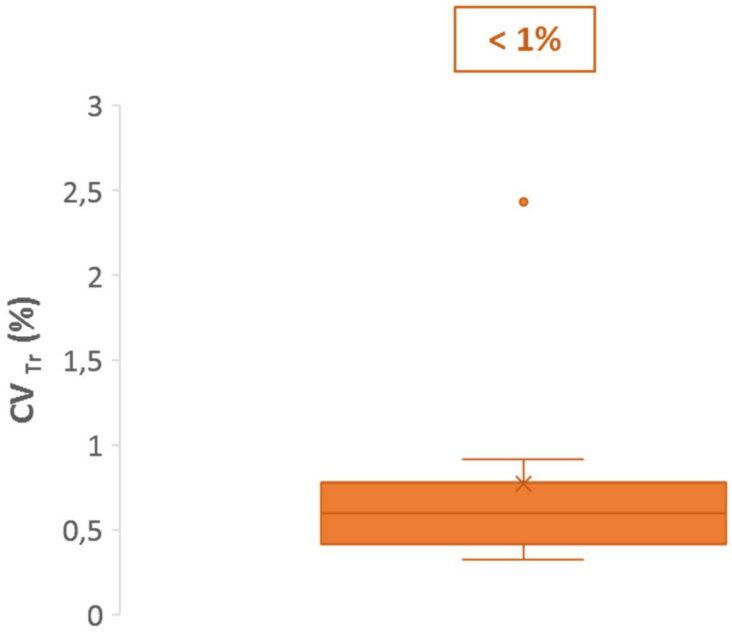


Figure III-9 - Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des Tr extraits de Skyline pour les neuf peptides iRT considérés dans l'ensemble des échantillons de la série $Q_{éq}$.
La valeur encadrée correspond à la valeur médiane

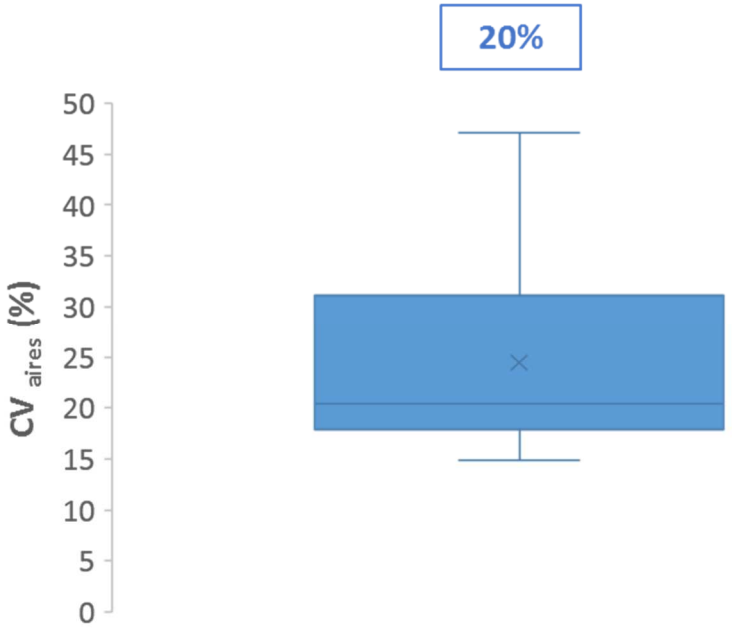


Figure III-10 -Boîte à moustache représentant la distribution des coefficients de variation calculés à l'aide des valeurs d'aires extraites de Skyline pour les neufs peptides iRT considérés, quantifiés dans l'ensemble des échantillons de la série $Q_{éq}$.
La valeur encadrée correspond à la valeur médiane

Echantillons de la série $Q_{éq}$

En ce qui concerne les 32 échantillons de la série $Q_{éq}$, leur profil d'intensités protéiques individuel ont été comparés à l'aide de la Figure III-11.

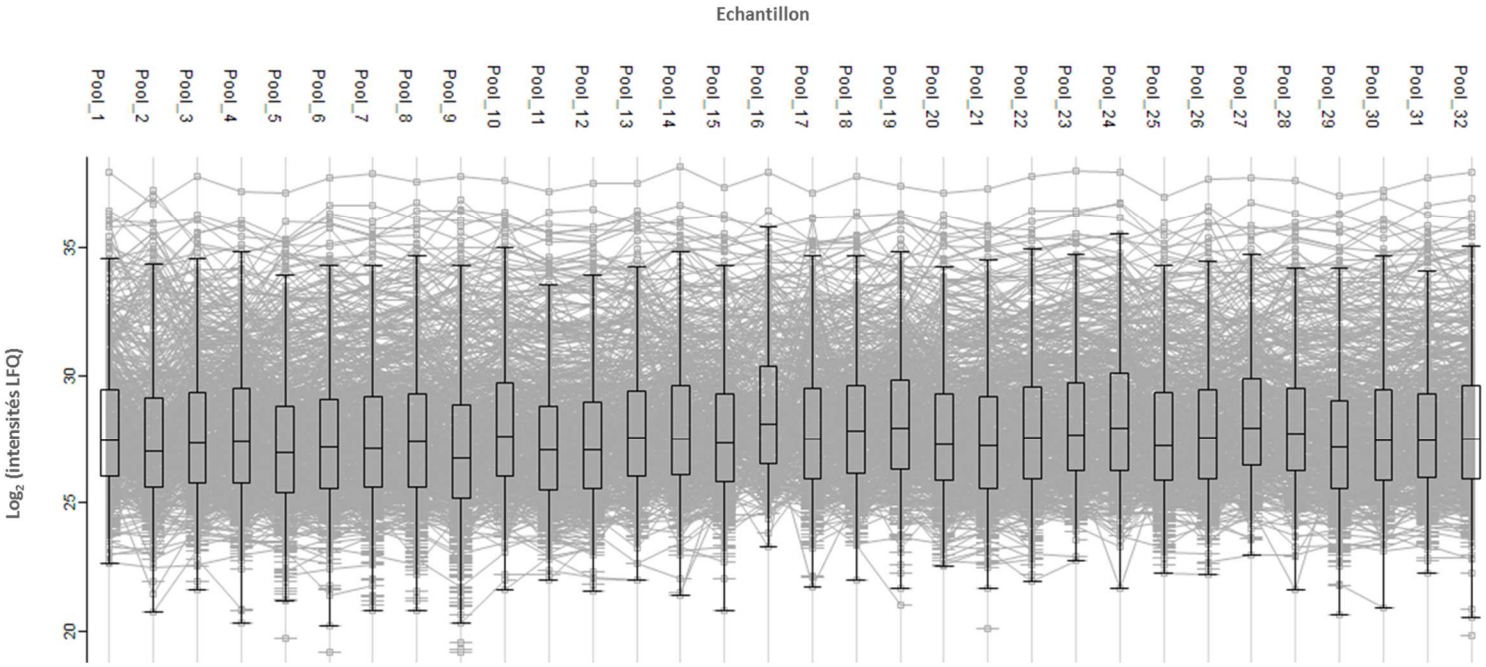


Figure III-11 - Profil de l'ensemble des intensités protéiques pour chacun des 20 échantillons analysés

Les profils globaux d'intensités de l'ensemble des échantillons semblent relativement semblables les uns aux autres. Tout comme au niveau qualitatif, l'échantillon 16 semble légèrement différent, cependant cela n'est pas observé pour l'échantillon 2. Malgré un nombre de protéines identifiées inférieur, les intensités de l'échantillon 16 sont globalement décalées vers des valeurs plus élevées par rapport aux autres échantillons. Ceci peut éventuellement s'expliquer par le fait que si cet échantillon contenait effectivement moins de protéines en nombre comparé aux autres échantillons, et était donc moins complexe que les autres, l'analyse d'une même quantité de ce mélange moins complexe a en réalité engendré l'analyse d'une plus grande quantité de protéines individuelles de l'échantillon 16 par rapport aux autres échantillons. La normalisation LFQ aurait cependant dû permettre de palier à ce problème. Comparé aux échantillons ganglionnaires, les valeurs d'intensités pour les échantillons urinaires sont dispersées sur une gamme dynamique un peu plus large.

Par la suite, un test statistique opposant le groupe des prélèvements urinaires « contrôles » au groupe des prélèvements urinaires de différentes complications a été mené à partir des intensités peptidiques, à l'aide du modèle « *Peptide-level Robust Ridge Regression* » décrit par Ludger GOEMINNE *et collaborateurs*. Au total, 58 protéines ont été identifiées comme étant différentiellement exprimées

entre les deux groupes avec une « *p-value* » et un taux de faux-positifs inférieur à 5 %. Ce projet ayant fait l'objet d'un financement par la SATT Conectus de Strasbourg dans le cadre d'un projet de pré-maturation, l'identité de ces protéines ne peut être communiquée.

4- Conclusion et perspectives

Cette étape de découverte de biomarqueurs urinaires pour le suivi de greffons rénaux a permis de mettre en évidence une cinquantaine de protéines différentiellement exprimées entre les deux groupes. Parmi ces protéines, une trentaine ont été sélectionnées dans le but d'évaluer leur efficacité par SRM sur une cohorte de 200 échantillons urinaires. La poursuite de ce projet est actée avec un projet de maturation financé par la SATT Conectus de Strasbourg.

Par ailleurs, la préparation d'échantillons employée dans ce projet, d'une durée inférieure à deux jours, permet d'envisager son transfert en milieu hospitalier pour le diagnostic en routine de suivi de greffons rénaux. Cette durée de préparation peut cependant encore être réduite en employant un mélange d'enzymes LysC-trypsine rapide commercialisé par Promega qui permettrait, non seulement de réduire le taux de coupures manquées qui était de l'ordre de 28 % pour cette étape de découverte, mais aussi de réduire le temps de digestion et de préparation en omettant les étapes de réduction et d'alkylation.

TRAVAUX COMPLEMENTAIRES

TRAVAUX COMPLEMENTAIRES

Les différentes compétences acquises au cours de ces travaux de thèse ont été mises à contribution d'autres projets de collaboration qui n'ont pas été développé dans ce manuscrit. Ceux-ci, qui ont fait ou vont prochainement faire l'objet d'une publication, sont résumés ci-après.

I- Recherche de partenaires et caractérisation d'une protéine issue de l'archée *Haloferax volcanii*

Ce projet a été mené en collaboration avec le Dr Marc GRAILLE et son étudiant en thèse, Trãn Vãn Nhân, du laboratoire de biochimie de l'école polytechnique de Palaiseau.

Chez les eucaryotes, et plus spécifiquement chez la levure *Saccharomyces cerevisiae*, la protéine Trm112 interagit avec et active des méthyltransférases telles que Mtq2, Trm9, Trm11 et Bud23, ce qui a pour conséquence de modifier les acteurs de la traduction des ARNm en protéines par le ribosome[229]. Les dysfonctionnements des complexes formés par Trm112 et ses partenaires ont d'ailleurs été associés à des maladies neuro-dégénératives et des cancers[230, 231]. L'archée *Haloferax volcanii* constitue un modèle d'études, notamment du fait de similarités entre les mécanismes de traduction des eucaryotes et des archées, pour lequel des outils génétiques ont été développés.

1- Recherche de partenaires de Trm112

Haloferax volcanii permet d'étudier et de comprendre d'un point de vue évolutif le(s) rôle(s) des orthologues de Trm112, dans le but de mieux appréhender les mécanismes de traduction chez les eucaryotes. De ce fait, l'objectif de ce projet consistait en la caractérisation du réseau d'interaction de la protéine Trm112 chez *Haloferax volcanii* par l'identification de ses partenaires au travers d'une approche de spectrométrie de masse.

Ainsi, les échantillons suivants ont été préparés par nos collaborateurs :

- quatre co-immunoprécipitations de la protéine Trm112-Flag tagguée, réalisées après une réaction de pontage chimique *in vivo* au formaldéhyde, de manière à stabiliser le complexe[232, 233]. Il est à noter que la réaction de pontage au formaldéhyde a été inversée par chauffage après la purification.

- une co-immunoprécipitation de Trm112-Flag tag sans pontage préalable,

- ainsi que trois co-immunoprécipitations du flag-tag seul. Cet échantillon permet d'identifier les interactions non spécifiques.

L'ensemble de ces échantillons ont fait l'objet d'une préparation d'échantillons par gel « *Stacking* », ont été digérés à la trypsine et analysés par microLC-MS/MS sur un couplage microLC (NanoLC 400 Eksigent-SCIEX) - spectromètre de masse de type Q-TOF (TripleTOF 6600 – SCIEX). Les échantillons ont ensuite été soumis à une recherche dans une banque de données contenant les séquences protéiques issues d'UniProt pour l'espèce *Haloferax volcanii*. Une comparaison des listes de protéines identifiées, et plus particulièrement une comparaison du nombre de spectres ayant permis ces identifications entre les échantillons pontés et les échantillons de flag-tag seul, a mené à l'identification de 25 méthyltransférases enrichies dans les échantillons pontés. Parmi ces méthyltransférases, des orthologues de Mtq2, Trm9 et Trm11 de *Saccharomyces cerevisiae* ont été retrouvés. Des tests enzymatiques ont d'ailleurs pu mettre en évidence des activités similaires de Mtq2-Trm112 et Trm9-Trm112 d'*H. volcanii* à celles de leurs orthologues eucaryotes.

Ainsi, l'apport de la spectrométrie de masse pour l'identification de partenaires a permis d'approfondir l'état des connaissances du réseau d'interaction de la protéine Trm112 dans le mécanisme de la traduction d'ARNm.

2- Recherche de méthylation du motif GGQ de la protéine aRF1

Le complexe formé par Trm112 et Mtq2 méthyle le facteur de terminaison eRF1 impliqué dans la libération des protéines nouvellement synthétisées sur la chaîne latérale d'une glutamine chez *S. cerevisiae*[234, 235]. L'objectif de cette étude était ici de caractériser la protéine aRF1, orthologue de eRF1 chez *H. volcanii*, et plus spécifiquement, d'étudier son état de méthylation dans les co-immunoprécipitations d'aRF1 provenant :

- De la souche sauvage,
- D'une souche Mtq2 délétée,
- D'une souche Trm112 délétée ,
- Et d'un essai de méthylation *in vitro* qui permet de maîtriser le rendement de méthylation.

L'objectif était de déterminer si la protéine Trm112 est nécessaire à la méthylation d'aRF1.

Ainsi, des gels « *Stacking* » ont été préparés pour chaque échantillon de manière à retirer la grande quantité de sels présente dans les échantillons, du fait de la précipitation de la protéine en absence de sels. Dans un premier temps, trois enzymes ont été testées sur la protéine aRF1 recombinante exprimée chez *Escherichia Coli*, soit la trypsine, la pepsine et la thermolysine. L'objectif était de trouver une enzyme qui permette de mesurer le motif GGQ de la protéine par spectrométrie de masse, soit plus précisément de générer un peptide de 7 à 25 acides aminés contenant le motif GGQ. Ainsi, les digestes peptidiques de ces trois enzymes pour la protéine recombinante ont été analysés par nanoLC-MS/MS sur un couplage nanoLC (Nano-Acquity, WATERS) – spectromètre de masse de type Q-Orbitrap (Q-Exactive +, THERMO FISHER SCIENTIFIC). Par la suite, des recherches ont été effectuées dans une banque de donnée contenant la séquence de cette protéine, ainsi que celle de la protéine aRF3 (souvent présente dans les échantillons du fait de sa forte interaction avec aRF1), des enzymes utilisées et des kératines humaines. Pour la trypsine, au maximum une coupure manquée était autorisée. Pour la thermolysine, cinq coupures manquées ont été autorisées, alors que pour la pepsine, l'enzyme n'a pas été spécifiée. Deux recherches ont été effectuées pour chaque enzyme : une première avec les modifications variables de méthylation et de diméthylation de la glutamine, et une seconde avec la modification variable de méthylation sur l'ensemble des acides aminés susceptibles de portés une méthylation, soit T, S, E, D, L, I, R, Q, N, K, H et C.

- De par la proximité de nombreuses lysines et arginines autour du motif GGQ (Figure I-1), la trypsine n'a pas permis de générer des peptides suffisamment grands pour être détectés en MS.

- La thermolysine (qui clive théoriquement en partie N-terminale des acides-aminés L, F, V, I, A et M) semblait quant à elle permettre de détecter le motif GGQ par MS. Cela a bien été le cas, cependant les spectres de fragmentation de ces peptides présentaient un faible rapport signal sur bruit et étaient peu informatifs malgré des optimisations d'énergies de collision ou encore de méthodes de liste d'inclusion.

- La pepsine (qui clive théoriquement en partie C-terminale des acides-aminés F, L, Y et W mais est relativement aspécifique) a permis de mesurer le peptide par MS. Les spectres des peptides permettant l'identification du motif GGQ (méthylé, diméthylé ou non) ont été vérifiés manuellement afin de retirer d'éventuelles ambiguïtés. Par la suite, pour l'ensemble des peptides dont les spectres MS/MS étaient validés, les courants d'ions ont été extraits par le logiciel Skyline dans l'ensemble des échantillons afin de vérifier leur absence ou leur présence

dans chacun de ces échantillons, dans le but de palier au phénomène de sous-échantillonnage inhérent au mode d'acquisition DDA.

MSSDAEEASEDRRKYEFKKVIEDLREYEGSGTQLVTIYIPEDKQISDVVEHVITEHS
EASNIKSKQTRTNVQDALTSIKDRLRYYGNFPPDNGIVMFSGAVDAGGGQTTM
VTKVLESPPEPIQSFRYHCDSNFLTGPLEDMLMDKGLFGLIVLDRREANVGWLK
GKRVEPVKSASSLVPGKQRK**GGQ**SAQRFARLRLEAIDNFYQEVAGMANDLFVP
KRHDMDGILVGGPSPTKDEFLDGDYLHHELQDLVVGKFDVAYTDESGLYDLVDA
GQDALADQEVIKDKKQMEEFFEKLHRGNESTYGFEATRKNLVMGSVDRLLISED
LRKDIAVYDCGGQEEYELVDHRHDTPTHECDDGSEAELKDREDVIEWLMDLAD
QRGTETKFISTDFEKGEQLYDAFGGIAGILRYSTGVHHHHHH

Figure I-1 – Séquence de la protéine aRF1 d'*Haloferax volcanii*
Les peptides théoriques résultant de la digestion sont soulignés en bleu pour la trypsine, en vert pour la thermolysine et en orange pour la pepsine

Au final, la pepsine a permis de conclure sur la présence de l'espèce non méthylée, méthylée et diméthylée dans la souche sauvage ainsi que dans l'échantillon Trm112 délété. Aucun des peptides ayant permis de détecter le motif GGQ n'a été mesuré dans les échantillons Mtq2 délété et *in vitro*, malgré de nombreux essais. L'échantillon *in vitro* va faire l'objet d'une nouvelle préparation, étant donné que les conditions expérimentales ont depuis été optimisées, et l'échantillon Mtq2 délété va être purifié une seconde fois afin d'effectuer une nouvelle analyse par MS.

II- Recherche de zones d'interactions par une technique de pontage chimique dans le cadre de différents complexes

La technique de pontage chimique est une méthode permettant de déterminer l'architecture globale de complexes protéiques en localisant des surfaces d'interactions entre les différents partenaires[236]. Il s'agit d'une approche de protéomique structurale qui emploie une stratégie de protéomique « *Bottom-up* ». Cette technique consiste à ponter chimiquement un complexe protéique dans son état natif, ou pseudo-natif, à l'aide d'un agent pontant[237]. Cet agent, généralement bi-fonctionnel et non-clivable, a pour particularité :

- De se lier aux chaînes latérales d'acides aminés spécifiques,

- D'imposer une restriction de distance de par la taille de son bras espaceur. En effet, deux résidus seront liés pas l'agent pontant, uniquement si la distance spatiale entre ces deux résidus est inférieure à la taille maximum que peut prendre le bras espaceur[238].

Ainsi, l'identification de ces pontages chimiques permet de définir quels résidus sont liés, fournissant de cette manière des informations de proximité[239]. Ces informations restent cependant complémentaires à d'autres techniques structurelles comme l'échange hydrogène/deutérium, la cristallographie à rayons X, etc[240].

La digestion enzymatique de tels échantillons génère des mélanges très complexes, au sein desquels les informations les plus pertinentes, soit les ponts inter-protéiques, sont très peu abondants et difficiles à détecter[241]. Par ailleurs, ces peptides inter-protéiques pontés génèrent des spectres de fragmentation très complexes, puisqu'ils résultent de la fragmentation simultanée de deux peptides et de l'agent pontant. Les possibilités d'interprétation sont de ce fait nombreuses et augmentent l'espace de recherche de n^2, rendant l'utilisation de moteurs de recherche de protéomique « classique » comme Mascot inutilisables[236]. Il n'y a à l'heure actuelle pas de consensus sur les outils de traitement de données issues de pontage chimique, et de nombreuses équipes ont développé leur propre logiciel dédié à ce type de données (tels que xQuest/xProphet[242, 243], pLink[236], etc.). En plus de ces difficultés, le défi actuel de cette technique réside dans la validation automatisée des résultats[237]. En effet, une validation manuelle très chronophage des données, qui requiert de l'expérimentation et qui est sujette aux erreurs, est encore nécessaire de nos jours.

Au cours de deux projets de collaboration, deux mélanges d'agents pontants lourd et léger, se liant aux chaînes latérales des lysines ont été employés : le DSS d_0/d_{12} (disuccinimidyl suberate) et le BS^3 d_0/d_4 (Sulfo-DSS) avec des bras espaceurs de 11,4 Ångström. L'utilisation d'un mélange lourd/léger s'explique par la génération d'un doublet caractéristique détectable sur les spectres MS, qui permet de faciliter la vérification manuelle des résultats. Après digestion à la trypsine, les échantillons résultant du pontage chimique ont été analysés par nanoLC-MS/MS. Les données collectées ont été soumis à une recherche avec le logiciel pLink, et seuls les peptides pontés intra- et inter-protéiques pour lesquels l'agent pontant à la fois lourd et léger a été identifié ont été vérifiés manuellement.

1- Etude du complexe de la myomégaline

Afin de comprendre les mécanismes moléculaires impliqués dans le contrôle et la régulation de la dynamique des microtubules, nos collaborateurs du centre de recherche en cancérologie de Marseille, le Dr Ali BADACHE et son étudiant Habib BOUGUENINA, ont étudié l'interactome de la myomégaline[244, 188]

[245]. Pour ce faire, une stratégie de pontage chimique a notamment été appliquée à l'étude du complexe formé par la myomégaline et ses huit autres partenaires.

Cette approche a permis d'identifier de nombreux pontages intra-protéiques ainsi que 9 pontages inter-protéiques, parmi lesquels :

- 5 entre la myomégaline et PRKAC
- 1 entre la myomégaline et CDK5RAP2
- 2 entre CDK5RAP2 et AKAP9
- 1 entre AKAP 9 et EB1.

Une visualisation des données par l'outil disponible gratuitement sur internet, XiNet[246], a permis d'établir la Figure II-1 et d'apprécier la proximité des protéines PRKAC et myomégaline.

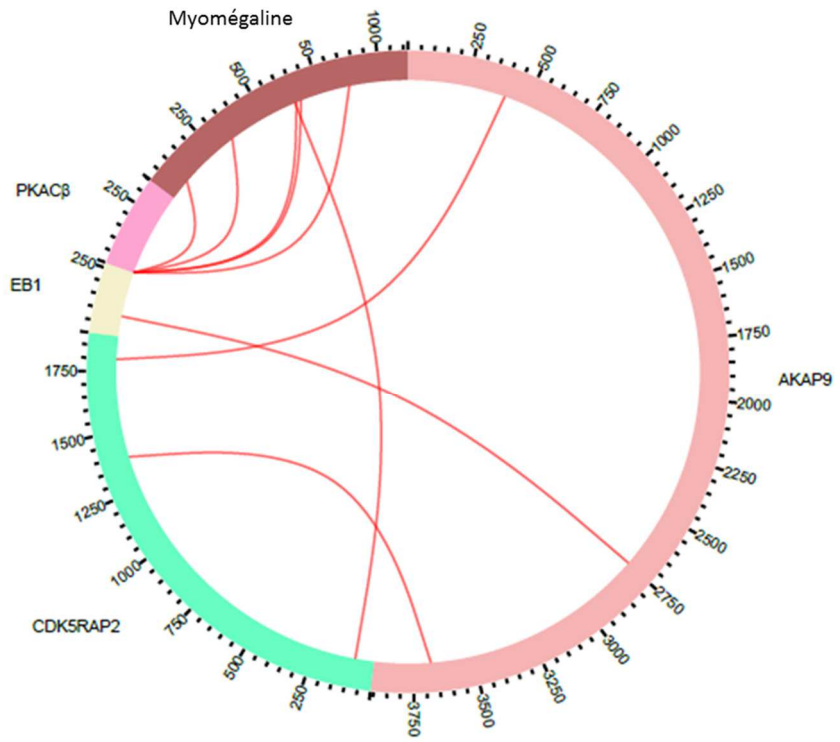


Figure II-1 - Visualisation XiNet des pontages inter-protéiques entre la myomégaline et certains de ses partenaires

Les collaborateurs ont pu, en plus de la cartographie moléculaire, établir la Figure II-2, fournissant des informations complémentaires dans la connaissance de l'organisation du complexe de la

myomégaline, qui s'avère relativement alambiquée. Ces travaux ont fait l'objet d'une publication dans le journal *PNAS* en 2017.

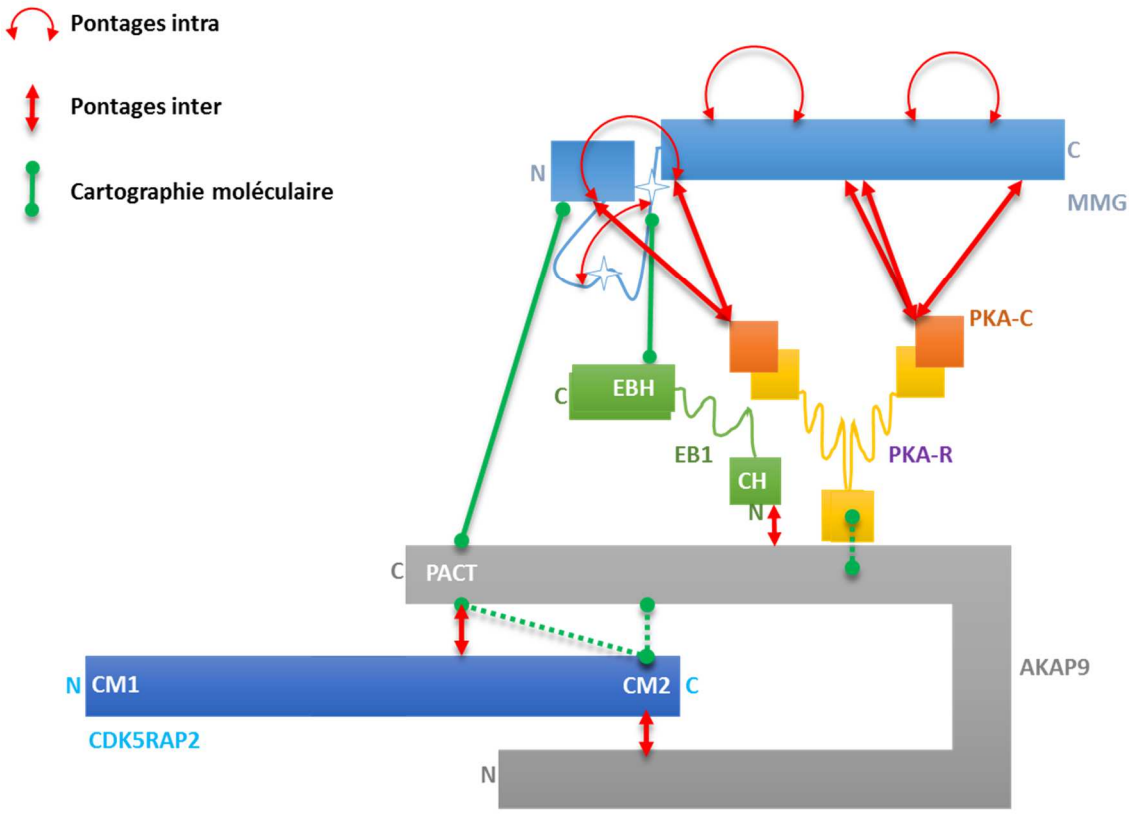


Figure II-2 - Association des données de pontage chimique et de cartographie moléculaire pour l'étude du complexe de la myomégaline (abrégée MMG sur cette Figure)

2- Etude du complexe TFIIH au travers de différentes sous-unités

L'étude du complexe TFIIH au travers de différentes sous-unités a été mené en collaboration avec le Dr Arnaud POTERSZMAN, de l'équipe de biologie structurale intégrative de l'Institut de Génétique et de Biologie Moléculaire et Cellulaire d'Illkirch.

TFIIH est un facteur de transcription général, essentiel pour la transcription de l'ARN polymérase II, qui est impliqué dans le processus de réparation de l'ADN. Il s'agit d'un complexe multi-protéique composé de dix sous-unités réparties en deux sous-complexes : le cœur-TFIIH et le complexe à activité kinase CAK[247]. La mutation de TFIIH peut résulter en des maladies humaines telles que des cancers ou des maladies récessives autosomiques[248]. Son réseau d'interaction est complexe et actuellement peu compris au niveau moléculaire, c'est pourquoi il a fait l'objet de plusieurs études structurelles au laboratoire. En effet, en plus d'études d'échange hydrogène/deutérium, différents édifices de complexité croissante que sont p34-p44 (sous-édifice centrale de l'intégrité structurelle du cœur à 6),

CAK/XPD et le cœur à 6 ont été étudié par une approche de pontage chimique afin de définir des zones d'interactions.

L'étude de deux de ces édifices (cœur à 6 ainsi que de CAK/XPD) a été menée en binôme avec le Dr Julien MARCOUX alors qu'il était post-doctorant au laboratoire. Ainsi, à l'aide de cette approche de pontage chimique, nous avons pu établir :

- 35 pontages inter-protéiques et 46 intra-protéiques pour le cœur à 6, ainsi que 4 pontages inter-protéiques et 60 intra-protéiques pour CAK ayant permis de générer la Figure II-3,
- Et 21 pontages inter-protéiques entre p34-p44 représentés sur la Figure II-4 ainsi que 9 pontages intra-protéiques p44 et pour p34.

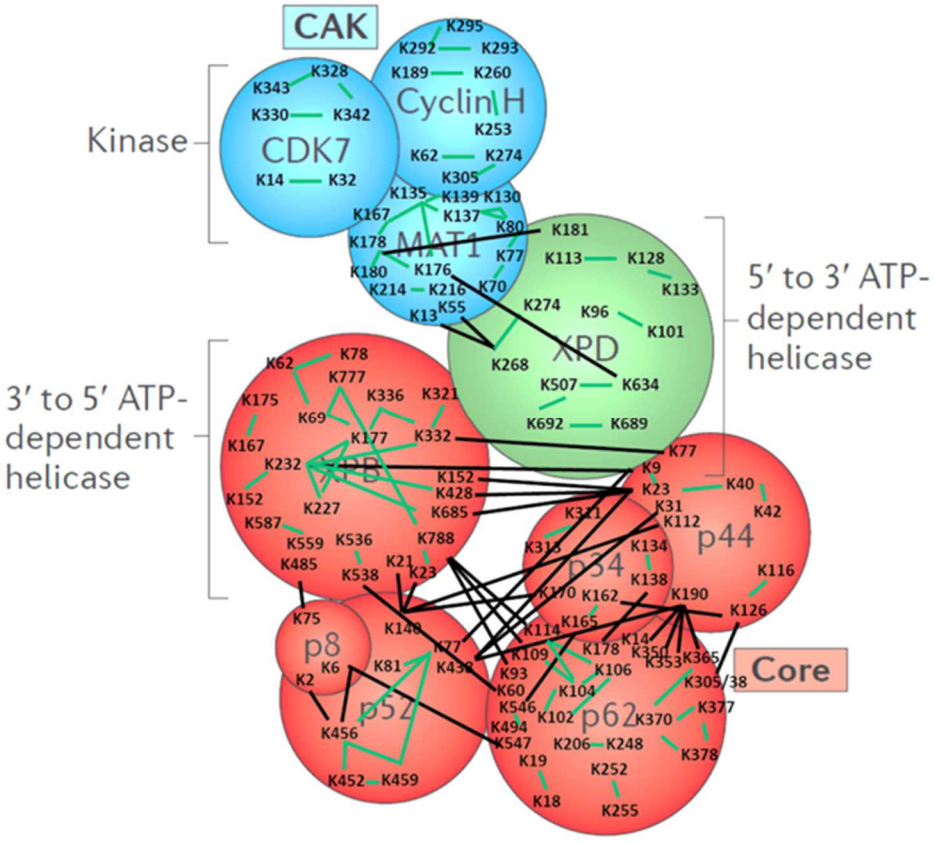


Figure II-3 - Visualisation des pontages inter et intra-protéiques identifiés avec l'analyse des peptides pontés des édifices CAK/XPD et cœur à 6

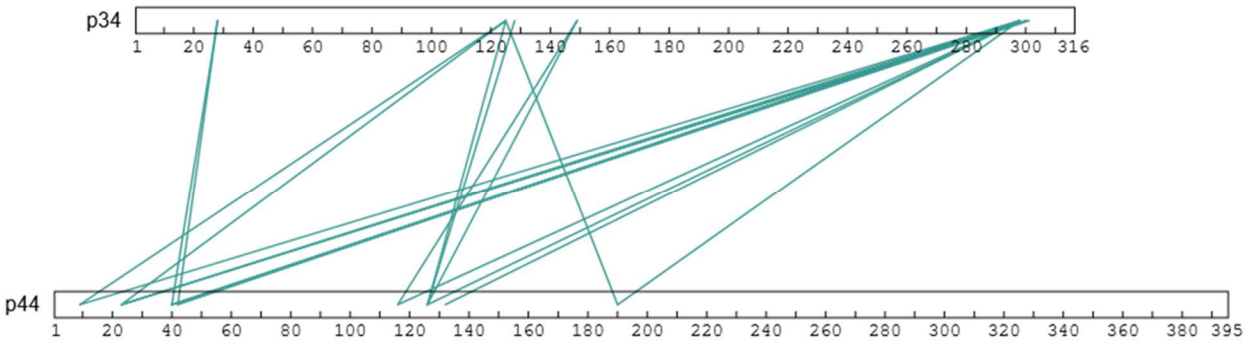


Figure II-4 - Visualisation des pontages inter-protéiques identifiés dans l'édifice p34-p44

L'ensemble de ces résultats ont été fournis à nos collaborateurs et vont prochainement faire l'objet d'une publication.

CONCLUSION GENERALE

Les objectifs de ces travaux de thèse étaient de développer et optimiser des méthodes de préparation d'échantillons pouvant répondre aux exigences des stratégies de protéomique quantitative sans marquage (XIC ou comptage de spectres), à savoir une simplicité et une rapidité de mise en œuvre permettant l'analyse d'un grand nombre d'échantillons, ainsi qu'une haute répétabilité.

C'est dans cet objectif que j'ai développé et évalué la préparation d'échantillons « *Tube-Gel* ». J'ai pu démontrer son efficacité et sa répétabilité pour la protéomique quantitative sans marquage à haut-débit vis-à-vis de deux autres techniques couramment employées, et ainsi implémenter cette préparation d'échantillons innovante à de nombreux projets au laboratoire. Par ailleurs, j'ai pu poursuivre les développements de cette préparation d'échantillons et démontrer l'universalité du protocole « *Tube-Gel* » à l'aide de différentes conditions d'extraction et de polymérisation.

En parallèle, j'ai procédé à des optimisations de préparations d'échantillons permettant de répondre à diverses questions biologiques dans le domaine de la recherche de biomarqueurs de pathologies par analyse protéomique. Ainsi, j'ai pu mettre au point un protocole permettant d'optimiser l'analyse des protéines membranaires dans le cadre de la recherche de biomarqueurs de glioblastomes. Ces protéines sont à l'heure actuelle souvent des cibles thérapeutiques préférentielles et sont de ce fait souvent au centre des recherches de biomarqueurs protéiques. Cependant, leur extraction et leur analyse restent aujourd'hui encore un défi. La mise au point de ce protocole a permis de mettre en évidence un candidat biomarqueur permettant de distinguer les domaines différenciés et non différenciés au sein des tumeurs que sont les glioblastomes.

J'ai également optimisé des méthodes d'extraction de protéines à partir de tissus ganglionnaires à la fois frais et FFPE pour l'étude de la résistance au traitement des LBDGC. Ces optimisations ont permis d'étudier les LBDGC à partir de tissus ganglionnaires frais, et de mettre en évidence des protéines intéressantes qui pourraient permettre de mieux comprendre et appréhender le mécanisme de chimiorésistance. Ces candidats pourront être évalués sur des tissus FFPE dont l'extraction a d'ores et déjà été optimisée.

Par ailleurs, j'ai développé une méthode de préparation d'échantillons permettant l'analyse du protéome urinaire dans le but de trouver des biomarqueurs de suivi de greffons rénaux. Cette matrice est extrêmement variable d'un individu à l'autre, et l'objectif final de ce projet est de mettre au point une méthode diagnostique applicable en routine à l'hôpital. Ainsi, une méthode de préparation d'échantillons simple, rapide et robuste vis-à-vis des difficultés inhérentes à ce type d'échantillons était

requise. Grâce aux développements effectués, le projet se poursuit avec la vérification et la validation des candidats biomarqueurs proposés par SRM sur une cohorte de 200 échantillons.

Enfin, les différentes compétences acquises au cours de ces travaux ont été mises à contribution d'autres projets de collaboration de protéomique structurale.

En conclusion, ces travaux de thèse reflètent que la préparation d'échantillons est une étape primordiale du schéma analytique. Celle-ci doit être adaptée à la fois à la question biologique posée et au milieu étudié. Bien entendu, une préparation d'échantillons répétable permet d'éviter l'introduction de variabilités de manière précoce dans le schéma analytique d'études de protéomique quantitative sans marquage, cependant les étapes suivantes doivent au même titre être maîtrisées et suffisamment répétables afin de générer des données fiables. Ceci est particulièrement important dans la recherche de biomarqueurs de pathologies humaines, de manière à ne pas faire passer des candidats biomarqueurs erronés à des étapes de vérification et de validation très coûteuses.

Plus personnellement, au cours de ces travaux de thèse j'ai pu acquérir de nombreuses connaissances notamment sur la préparation d'échantillons, l'analyse par nanoLC-MS/MS, l'entretien de spectromètres de masse et le traitement bio-informatique de données. J'ai également pu développer mon esprit critique et pris conscience de l'enjeu de chaque manipulation de l'échantillon au cours du schéma analytique pouvant être source de variabilités. Par ailleurs, j'ai appris à communiquer mes résultats au travers de communications par affiche ou orale, mais aussi au travers de réunions avec les collaborateurs.

PERSPECTIVES

Au vu des performances en constante évolution des spectromètres de masse, qui permettent d'analyser des mélanges complexes en des temps de plus en plus réduits, et de la maturité des stratégies de protéomique « *Bottom-up* », l'enjeu des années à venir consistera à analyser des cohortes d'échantillons toujours plus importantes, ce qui permettra notamment de se placer dans des conditions plus favorables aux traitements statistiques. De ce fait, des préparations d'échantillons simples, rapides, sans fractionnement et répétables deviendront d'autant plus indispensables.

Par ailleurs, malgré la relative maturité de la protéomique « *Bottom-up* », de nombreux efforts doivent encore être réalisés de manière à tirer beaucoup plus d'informations des données générées. En effet, la quantité de données et par extension la quantité de spectres MS/MS générées ont très largement augmenté ces dernières années. Cependant, une large proportion de ces informations n'est

actuellement pas exploitée avec en moyenne 75 % de spectres non identifiés[61]. Ces améliorations peuvent notamment passer par la protéogénomique et l'intégration de données multi- « omiques » permettant par exemple de générer des banques personnalisées[80]. Ceci peut être particulièrement intéressant dans le domaine médical et la médecine personnalisée[8]. Le développement de stratégies multi-« omiques » reste cependant relativement récent et l'intégration de ces données reste compliquée de par le manque d'expertise et d'outils dédiés. Ainsi, des progrès restent à faire dans ce domaine afin de rendre ces approches plus attractives.

De par les différentes collaborations menées au cours de ces travaux, il me semble important d'intégrer de manière plus systématique des statisticiens aux projets de recherche de biomarqueurs, étant donné qu'ils ont l'expertise et la connaissance permettant d'employer les outils les plus adaptés aux données générées et aux questionnements biologiques. A ce même titre, l'analyse protéomique par MS nécessite une expertise dans le domaine de l'analytique et de la spectrométrie de masse, de manière à être au fait de l'importance de maîtriser l'ensemble du schéma analytique. Ce type d'expertise pourrait notamment être implémentées dans le futur dans les hôpitaux pour le diagnostic, voire la médecine personnalisée en routine par MS[249].

BIBLIOGRAPHIE

1. Neverova, I. & Van Eyk, J.E. Role of chromatographic techniques in proteomic analysis. *J Chromatogr B Analyt Technol Biomed Life Sci* **815**, 51-63 (2005).

2. Aebersold, R. & Goodlett, D.R. Mass spectrometry in proteomics. *Chem Rev* **101**, 269-295 (2001).

3. Matthews, H., Hanison, J. & Nirmalan, N. "Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes* **4** (2016).

4. Mann, M., Hendrickson, R.C. & Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-473 (2001).

5. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198-207 (2003).

6. Yates, J.R., 3rd Mass spectrometry. From genomics to proteomics. *Trends Genet* **16**, 5-8 (2000).

7. Nesvizhskii, A.I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics* **4**, 1419-1440 (2005).

8. Tanase, C., Albulescu, R. & Neagu, M. Proteomic Approaches for Biomarker Panels in Cancer. *J Immunoassay Immunochem* **37**, 1-15 (2016).

9. Dias, M.H., Kitano, E.S., Zelanis, A. & Iwai, L.K. Proteomics and drug discovery in cancer. *Drug Discov Today* **21**, 264-277 (2016).

10. Boschetti, E., Chung, M.C. & Righetti, P.G. "The quest for biomarkers": are we on the right technical track? *Proteomics Clin Appl* **6**, 22-41 (2012).

11. Bantscheff, M., Lemeer, S., Savitski, M.M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal Bioanal Chem* **404**, 939-965 (2012).

12. Pasing, Y., Colnoe, S. & Hansen, T. Proteomics of hydrophobic samples: Fast, robust and low-cost workflows for clinical approaches. *Proteomics* **17** (2017).

13. Anderson, N.L. & Anderson, N.G. Proteome and proteomics: new technologies, new concepts, and new words. *Electrophoresis* **19**, 1853-1861 (1998).

14. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301 (1988).

15. Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F. & Whitehouse, C.M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71 (1989).

16. Nilsson, T. et al. Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat Methods* **7**, 681-685 (2010).

17. Zhang, Y., Fonslow, B.R., Shan, B., Baek, M.C. & Yates, J.R., 3rd Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* **113**, 2343-2394 (2013).

18. Gregorich, Z.R., Chang, Y.H. & Ge, Y. Proteomics in heart failure: top-down or bottom-up? *Pflugers Arch* **466**, 1199-1209 (2014).

19. Garcia, B.A. What does the future hold for Top Down mass spectrometry? *J Am Soc Mass Spectrom* **21**, 193-202 (2010).

20. Toby, T.K. et al. Proteoforms in Peripheral Blood Mononuclear Cells as Novel Rejection Biomarkers in Liver Transplant Recipients. *Am J Transplant* **17**, 2458-2467 (2017).

21. Fornelli, L. et al. Top-down proteomics: Where we are, where we are going? *J Proteomics* (2017).

22. Moradian, A., Kalli, A., Sweredoski, M.J. & Hess, S. The top-down, middle-down, and bottom-up mass spectrometry approaches for characterization of histone variants and their post-translational modifications. *Proteomics* **14**, 489-497 (2014).

23. Wu, C. et al. A protease for 'middle-down' proteomics. *Nat Methods* **9**, 822-824 (2012).

24. Feist, P. & Hummon, A.B. Proteomic challenges: sample preparation techniques for microgram-quantity protein analysis from biological samples. *Int J Mol Sci* **16**, 3537-3563 (2015).

25. Tubaon, R.M., Haddad, P.R. & Quirino, J.P. Sample clean-up strategies for ESI mass spectrometry applications in bottom-up proteomics: Trends from 2012 to 2016. *Proteomics* (2017).

26. Speers, A.E. & Wu, C.C. Proteomics of integral membrane proteins--theory and application. *Chem Rev* **107**, 3687-3714 (2007).

27. Chang, Y.H. et al. New mass-spectrometry-compatible degradable surfactant for tissue proteomics. *J Proteome Res* **14**, 1587-1599 (2015).

28. Norris, J.L., Porter, N.A. & Caprioli, R.M. Mass spectrometry of intracellular and membrane proteins using cleavable detergents. *Anal Chem* **75**, 6642-6647 (2003).

29. Yasui, K., Uegaki, M., Shiraki, K. & Ishimizu, T. Enhanced solubilization of membrane proteins by alkylamines and polyamines. *Protein Sci* **19**, 486-493 (2010).

30. Manza, L.L., Stamer, S.L., Ham, A.J., Codreanu, S.G. & Liebler, D.C. Sample preparation and digestion for proteomic analyses using spin filters. *Proteomics* **5**, 1742-1745 (2005).

31. Wisniewski, J.R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359-362 (2009).

32. Ilavenil, S. et al. Removal of SDS from biological protein digests for proteomic analysis by mass spectrometry. *Proteome Sci* **14**, 11 (2016).

33. Kulak, N.A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Methods* **11**, 319-324 (2014).

34. Zougman, A., Selby, P.J. & Banks, R.E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006-1000 (2014).

35. Laemmli, U.K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680-685 (1970).

36. Rabilloud, T. et al. Power and limitations of electrophoretic separations in proteomics strategies. *Mass Spectrom Rev* **28**, 816-843 (2009).

37. Neuhoff, V., Stamm, R. & Eibl, H. Clear Background and Highly Sensitive Protein Staining with Coomassie Blue Dyes in Polyacrylamide Gels - a Systematic Analysis. *Electrophoresis* **6**, 427-448 (1985).

38. Candiano, G. et al. Blue silver: A very sensitive colloidal Coomassie G-250 staining for proteome analysis. *Electrophoresis* **25**, 1327-1333 (2004).

39. Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* **1**, 845-867 (2002).

40. Rabilloud, T. & Lelong, C. Two-dimensional gel electrophoresis in proteomics: a tutorial. *J Proteomics* **74**, 1829-1841 (2011).

41. Righetti, P.G. The Monkey King: a personal view of the long journey towards a proteomic Nirvana. *J Proteomics* **107**, 39-49 (2014).

42. Tsiatsiani, L. & Heck, A.J. Proteomics beyond trypsin. *FEBS J* **282**, 2612-2626 (2015).

43. Xie, F., Smith, R.D. & Shen, Y. Advanced proteomic liquid chromatography. *J Chromatogr A* **1261**, 78-90 (2012).

44. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699-711 (2004).

45. Michalski, A., Cox, J. & Mann, M. 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* **10**, 1785-1793 (2011).

46. Sleno, L. & Volmer, D.A. Ion activation methods for tandem mass spectrometry. *J Mass Spectrom* **39**, 1091-1112 (2004).

47. Biemann, K. Mass spectrometry of peptides and proteins. *Annu Rev Biochem* **61**, 977-1010 (1992).

48. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed Mass Spectrom* **11**, 601 (1984).

49. Wysocki, V.H., Tsaprailis, G., Smith, L.L. & Breci, L.A. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom* **35**, 1399-1406 (2000).

50. Dongre, A.R., Jones, J.L., Somogyi, A. & Wysocki, V.H. 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* **118**, 8365-8374 (1996).

51. Olsen, J.V. et al. Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4**, 709-712 (2007).

52. Syka, J.E., Coon, J.J., Schroeder, M.J., Shabanowitz, J. & Hunt, D.F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc Natl Acad Sci U S A* **101**, 9528-9533 (2004).

53. Zubarev, R.A. et al. Electron capture dissociation for structural characterization of multiply charged protein cations. *Anal Chem* **72**, 563-573 (2000).

54. Frese, C.K. et al. Unambiguous phosphosite localization using electron-transfer/higher-energy collision dissociation (EThcD). *J Proteome Res* **12**, 1520-1525 (2013).

55. Blueggel, M., Chamrad, D. & Meyer, H.E. Bioinformatics in proteomics. *Curr Pharm Biotechnol* **5**, 79-88 (2004).

56. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567 (1999).

57. Geer, L.Y. et al. Open mass spectrometry search algorithm. *J Proteome Res* **3**, 958-964 (2004).

58. Craig, R., Cortens, J.P. & Beavis, R.C. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* **3**, 1234-1242 (2004).

59. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).

60. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**, 1794-1805 (2011).

61. Griss, J. et al. Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat Methods* **13**, 651-656 (2016).

62. Skinner, O.S. & Kelleher, N.L. Illuminating the dark matter of shotgun proteomics. *Nat Biotechnol* **33**, 717-718 (2015).

63. Chick, J.M. et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **33**, 743-749 (2015).

64. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. & Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520 (2017).

65. McEntyre, J. Linking up with Entrez. *Trends Genet* **14**, 39-40 (1998).

66. Benson, D.A. et al. GenBank. *Nucleic Acids Res* **43**, D30-35 (2015).

67. O'Leary, N.A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).

68. Wu, C.H. et al. The Protein Information Resource. *Nucleic Acids Res* **31**, 345-347 (2003).

69. Berman, H.M. et al. The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242 (2000).

70. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204-212 (2015).

71. Reinders, J.S., A. Proteomics Methods and Protocols, Edn. 1. (Humana Press, 2009).

72. Deutsch, E.W. et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* **45**, D1100-D1106 (2017).

73. Vizcaino, J.A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**, 223-226 (2014).

74. Jones, P. et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res* **34**, D659-663 (2006).

75. Desiere, F. et al. The PeptideAtlas project. *Nucleic Acids Res* **34**, D655-658 (2006).

76. Lane, L. et al. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res* **40**, D76-83 (2012).

77. Gaudet, P. et al. The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res* **45**, D177-D182 (2017).

78. Schaeffer, M. et al. The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinformatics* (2017).

79. Jaffe, J.D., Berg, H.C. & Church, G.M. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**, 59-77 (2004).

80. Nesvizhskii, A.I. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**, 1114-1125 (2014).

81. Zhang, B. et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387 (2014).

82. Vaudel, M., Barsnes, H., Raeder, H. & Berven, F.S. Using Proteomics Bioinformatics Tools and Resources in Proteogenomic Studies. *Adv Exp Med Biol* **926**, 65-75 (2016).

83. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-214 (2007).

84. Huttlin, E.L., Hegeman, A.D., Harms, A.C. & Sussman, M.R. Prediction of error associated with false-positive rate determination for peptide identification in large-scale proteomics experiments using a combined reverse and forward peptide sequence database strategy. *J Proteome Res* **6**, 392-398 (2007).

85. Navarro, P. & Vazquez, J. A refined method to calculate false discovery rates for peptide identification using decoy databases. *J Proteome Res* **8**, 1792-1796 (2009).

86. Bantscheff, M. & Kuster, B. Quantitative mass spectrometry in proteomics. *Anal Bioanal Chem* **404**, 937-938 (2012).

87. Ong, S.E. et al. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics* **1**, 376-386 (2002).

88. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J.R. & Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* **7**, 383-385 (2010).

89. Gygi, S.P. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* **17**, 994-999 (1999).

90. Ross, P.L. et al. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* **3**, 1154-1169 (2004).

91. Thompson, A. et al. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* **75**, 1895-1904 (2003).

92. Cappadona, S., Baker, P.R., Cutillas, P.R., Heck, A.J. & van Breukelen, B. Current challenges in software solutions for mass spectrometry-based quantitative proteomics. *Amino Acids* **43**, 1087-1108 (2012).

93. Ong, S.E. & Mann, M. Mass spectrometry-based proteomics turns quantitative. *Nat Chem Biol* **1**, 252-262 (2005).

94. Rose, K. et al. A new mass-spectrometric C-terminal sequencing technique finds a similarity between gamma-interferon and alpha 2-interferon and identifies a proteolytically clipped gamma-interferon that retains full antiviral activity. *Biochem J* **215**, 273-277 (1983).

95. Mirgorodskaya, O.A. et al. Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards. *Rapid Commun Mass Spectrom* **14**, 1226-1232 (2000).

96. Yao, X., Freas, A., Ramirez, J., Demirev, P.A. & Fenselau, C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* **73**, 2836-2842 (2001).

97. Liu, H., Sadygov, R.G. & Yates, J.R., 3rd A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **76**, 4193-4201 (2004).

98. Blondeau, F. et al. Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc Natl Acad Sci U S A* **101**, 3833-3838 (2004).

99. Powell, D.W. et al. Cluster analysis of mass spectrometry data reveals a novel component of SAGA. *Mol Cell Biol* **24**, 7249-7259 (2004).

100. Rappsilber, J., Ryder, U., Lamond, A.I. & Mann, M. Large-scale proteomic analysis of the human spliceosome. *Genome Res* **12**, 1231-1245 (2002).

101. Ishihama, Y. et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* **4**, 1265-1272 (2005).

102. Braisted, J.C. et al. The APEX Quantitative Proteomics Tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics* **9**, 529 (2008).

103. Lu, P., Vogel, C., Wang, R., Yao, X. & Marcotte, E.M. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**, 117-124 (2007).

104. Bondarenko, P.V., Chelius, D. & Shaler, T.A. Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* **74**, 4741-4749 (2002).

105. Chelius, D. & Bondarenko, P.V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* **1**, 317-323 (2002).

106. Schwanhausser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).

107. Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P. & Geromanos, S.J. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. *Mol Cell Proteomics* **5**, 144-156 (2006).

108. Ramus, C. et al. Benchmarking quantitative label-free LC-MS data processing workflows using a complex spiked proteomic standard dataset. *J Proteomics* **132**, 51-62 (2016).

109. Vandenbrouck, Y. et al. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *J Proteome Res* **15**, 3998-4019 (2016).

110. Sturm, M. et al. OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics* **9**, 163 (2008).

111. Monroe, M.E. et al. VIPER: an advanced software package to support high-throughput LC-MS peptide identification. *Bioinformatics* **23**, 2021-2023 (2007).

112. Bouyssie, D. et al. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol Cell Proteomics* **6**, 1621-1637 (2007).

113. Schilling, B. et al. 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* **11**, 202-214 (2012).

114. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).

115. Cox, J. et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526 (2014).

116. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301-2319 (2016).

117. Vidova, V. & Spacil, Z. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal Chim Acta* **964**, 7-23 (2017).

118. Gillet, L.C. et al. 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* **11**, O111 016717 (2012).

119. Comai, L.K., J.E. ; Mallick, P. Proteomics Methods and Protocols, Edn. 2017. (Humana Press, Methods in molecular biology; 2017).

120. Silva, J.C. et al. Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem* **77**, 2187-2200 (2005).

121. Geromanos, S.J., Hughes, C., Ciavarini, S., Vissers, J.P. & Langridge, J.I. Using ion purity scores for enhancing quantitative accuracy and precision in complex proteomics samples. *Anal Bioanal Chem* **404**, 1127-1139 (2012).

122. Egertson, J.D. et al. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* **10**, 744-746 (2013).

123. Purvine, S., Eppel, J.T., Yi, E.C. & Goodlett, D.R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847-850 (2003).

124. Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A. & Yates, J.R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* **1**, 39-45 (2004).

125. Tsou, C.C. et al. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* **12**, 258-264, 257 p following 264 (2015).

126. Li, Y. et al. Group-DIA: analyzing multiple data-independent acquisition mass spectrometry data files. *Nat Methods* **12**, 1105-1106 (2015).

127. Domon, B. & Aebersold, R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* **28**, 710-721 (2010).

128. Wallin, E. & von Heijne, G. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7**, 1029-1038 (1998).

129. Alfonso-Garrido, J., Garcia-Calvo, E. & Luque-Garcia, J.L. Sample preparation strategies for improving the identification of membrane proteins by mass spectrometry. *Anal Bioanal Chem* **407**, 4893-4905 (2015).

130. Vit, O. & Petrak, J. Integral membrane proteins in proteomics. How to break open the black box? *J Proteomics* **153**, 8-20 (2017).

131. Moore, S.M., Hess, S.M. & Jorgenson, J.W. Extraction, Enrichment, Solubilization, and Digestion Techniques for Membrane Proteomics. *J Proteome Res* **15**, 1243-1252 (2016).

132. Wang, X. & Liang, S. Sample preparation for the analysis of membrane proteomes by mass spectrometry. *Protein Cell* **3**, 661-668 (2012).

133. Miguet, L. et al. Proteomic analysis of malignant B-cell derived microparticles reveals CD148 as a potentially useful antigenic biomarker for mantle cell lymphoma diagnosis. *J Proteome Res* **8**, 3346-3354 (2009).

134. Miguet, L. et al. Proteomic analysis of malignant lymphocyte membrane microparticles using double ionization coverage optimization. *Proteomics* **6**, 153-171 (2006).

135. Mause, S.F. & Weber, C. Microparticles: protagonists of a novel communication network for intercellular information exchange. *Circ Res* **107**, 1047-1057 (2010).

136. Audran, E. (2012).

137. Leon, I.R., Schwammle, V., Jensen, O.N. & Sprenger, R.R. Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol Cell Proteomics* **12**, 2992-3005 (2013).

138. Zhou, J. et al. Evaluation of the application of sodium deoxycholate to proteomic analysis of rat hippocampal plasma membrane. *J Proteome Res* **5**, 2547-2553 (2006).

139. Lu, X. & Zhu, H. Tube-gel digestion: a novel proteomic approach for high throughput analysis of membrane proteins. *Mol Cell Proteomics* **4**, 1948-1958 (2005).

140. Cao, R. et al. High-throughput analysis of rat liver plasma membrane proteome by a nonelectrophoretic in-gel tryptic digestion coupled with mass spectrometry identification. *J Proteome Res* **7**, 535-545 (2008).

141. Han, C.L. et al. A multiplexed quantitative strategy for membrane proteomics: opportunities for mining therapeutic targets for autosomal dominant polycystic kidney disease. *Mol Cell Proteomics* **7**, 1983-1997 (2008).

142. Zhou, J. et al. Gel absorption-based sample preparation for the analysis of membrane proteome by mass spectrometry. *Anal Biochem* **404**, 204-210 (2010).

143. Yu, H. et al. Quantifying raft proteins in neonatal mouse brain by 'tube-gel' protein digestion label-free shotgun proteomics. *Proteome Sci* **5**, 17 (2007).

144. Fischer, R. & Kessler, B.M. Gel-aided sample preparation (GASP)--a simplified method for gel-assisted proteomic sample generation from protein extracts and intact cells. *Proteomics* **15**, 1224-1229 (2015).

145. Muller, L., Fornecker, L., Van Dorsselaer, A., Cianferani, S. & Carapito, C. Benchmarking sample preparation/digestion protocols reveals tube-gel being a fast and repeatable method for quantitative proteomics. *Proteomics* **16**, 2953-2961 (2016).

146. Cao, L., Clifton, J.G., Reutter, W. & Josic, D. Mass spectrometry-based analysis of rat liver and hepatocellular carcinoma Morris hepatoma 7777 plasma membrane proteome. *Anal Chem* **85**, 8112-8120 (2013).

147. Eley, M.H., Burns, P.C., Kannapell, C.C. & Campbell, P.S. Cetyltrimethylammonium bromide polyacrylamide gel electrophoresis: estimation of protein subunit molecular weights using cationic detergents. *Anal Biochem* **92**, 411-419 (1979).

148. Caglio, S., Chiari, M. & Righetti, P.G. Gel polymerization in detergents: conversion efficiency of methylene blue vs. persulfate catalysis, as investigated by capillary zone electrophoresis. *Electrophoresis* **15**, 209-214 (1994).

149. Lyubimova, T., Caglio, S., Gelfi, C., Righetti, P.G. & Rabilloud, T. Photopolymerization of polyacrylamide gels with methylene blue. *Electrophoresis* **14**, 40-50 (1993).

150. Righetti, P.G. & Gelfi, C. Electrophoresis gel media: the state of the art. *J Chromatogr B Biomed Sci Appl* **699**, 63-75 (1997).

151. Rabilloud, T., Vincon, M. & Garin, J. Micropreparative one- and two-dimensional electrophoresis: improvement with new photopolymerization systems. *Electrophoresis* **16**, 1414-1422 (1995).

152. Lyubimova, T. & Righetti, P.G. On the kinetics of photopolymerization: a theoretical study. *Electrophoresis* **14**, 191-201 (1993).

153. Bonaventura, C., Bonaventura, J., Stevens, R. & Millington, D. Acrylamide in Polyacrylamide Gels Can Modify Proteins during Electrophoresis. *Analytical Biochemistry* **222**, 44-48 (1994).

154. Righetti, P.G., Gelfi, C. & Bosisio, A.B. Polymerization Kinetics of Polyacrylamide Gels .3. Effect of Catalysts. *Electrophoresis* **2**, 291-295 (1981).

155. Bortner, J.D., Jr. et al. Proteomic profiling of human plasma by iTRAQ reveals down-regulation of ITI-HC3 and VDBP by cigarette smoking. *J Proteome Res* **10**, 1151-1159 (2011).

156. Zamo, A. & Cecconi, D. Proteomic analysis of lymphoid and haematopoietic neoplasms: there's more than biomarker discovery. *J Proteomics* **73**, 508-520 (2010).

157. Boyd, R.S., Dyer, M.J. & Cain, K. Proteomic analysis of B-cell malignancies. *J Proteomics* **73**, 1804-1822 (2010).

158. Jansen, C. et al. Protein profiling in pathology: analysis and evaluation of 239 frozen tissue biopsies for diagnosis of B-cell lymphomas. *Proteomics Clin Appl* **4**, 519-527 (2010).

159. Hu, Q.Y., Su, J., Jiang, H., Wang, L.L. & Jia, Y.Q. Potential role of proteomics in the diagnosis of lymphoma: a meta-analysis. *Int J Lab Hematol* **35**, 367-378 (2013).

160. Schnell, G. et al. Discovery and targeted proteomics on cutaneous biopsies infected by borrelia to investigate lyme disease. *Mol Cell Proteomics* **14**, 1254-1264 (2015).

161. Lowry, O.H., Rosebrough, N.J., Farr, A.L. & Randall, R.J. Protein measurement with the Folin phenol reagent. *J Biol Chem* **193**, 265-275 (1951).

162. Broeckx, V. et al. Comparison of multiple protein extraction buffers for GeLC-MS/MS proteomic analysis of liver and colon formalin-fixed, paraffin-embedded tissues. *Mol Biosyst* **12**, 553-565 (2016).

163. Tanca, A. et al. Comparability of differential proteomics data generated from paired archival fresh-frozen and formalin-fixed samples by GeLC-MS/MS and spectral counting. *J Proteomics* **77**, 561-576 (2012).

164. O'Rourke, M.B. & Padula, M.P. Analysis of formalin-fixed, paraffin-embedded (FFPE) tissue via proteomic techniques and misconceptions of antigen retrieval. *Biotechniques* **60**, 229-238 (2016).

165. Broeckx, V. et al. Formalin-fixed paraffin-embedded tissue: The holy grail of clinical proteomics. *Proteomics Clin Appl* **8**, 735-736 (2014).

166. Giusti, L. & Lucacchini, A. Proteomic studies of formalin-fixed paraffin-embedded tissues. *Expert Rev Proteomics* **10**, 165-177 (2013).

167. Ikeda, K. et al. Extraction and analysis of diagnostically useful proteins from formalin-fixed, paraffin-embedded tissue sections. *J Histochem Cytochem* **46**, 397-403 (1998).

168. Craven, R.A. et al. Proteomic analysis of formalin-fixed paraffin-embedded renal tissue samples by label-free MS: assessment of overall technical variability and the impact of block age. *Proteomics Clin Appl* **7**, 273-282 (2013).

169. Olszowy, P. & Buszewski, B. Urine sample preparation for proteomic analysis. *J Sep Sci* **37**, 2920-2928 (2014).

170. Court, M. et al. Toward a standardized urine proteome analysis methodology. *Proteomics* **11**, 1160-1171 (2011).

171. Kalantari, S., Jafari, A., Moradpoor, R., Ghasemi, E. & Khalkhal, E. Human Urine Proteomics: Analytical Techniques and Clinical Applications in Renal Diseases. *Int J Proteomics* **2015**, 782798 (2015).

172. Sigdel, T.K. et al. Mining the human urine proteome for monitoring renal transplant injury. *Kidney Int* **89**, 1244-1252 (2016).

173. Apweiler, R. et al. Approaching clinical proteomics: current state and future fields of application in fluid proteomics. *Clin Chem Lab Med* **47**, 724-744 (2009).

174. Vaezzadeh, A.R., Briscoe, A.C., Steen, H. & Lee, R.S. One-step sample concentration, purification, and albumin depletion method for urinary proteomics. *J Proteome Res* **9**, 6082-6089 (2010).

175. Yu, Y. et al. Urine sample preparation in 96-well filter plates for quantitative clinical proteomics. *Anal Chem* **86**, 5470-5477 (2014).

176. Yu, Y., Bekele, S. & Pieper, R. Quick 96FASP for high throughput quantitative proteome analysis. *J Proteomics* **166**, 1-7 (2017).

177. Sun, Z. et al. Toward Biomarker Development in Large Clinical Cohorts: An Integrated High-Throughput 96-Well-Plate-Based Sample Preparation Workflow for Versatile Downstream Proteomic Analyses. *Anal Chem* **88**, 8518-8525 (2016).

178. Hernandez-Valladares, M. et al. Reliable FASP-based procedures for optimal quantitative proteomic and phosphoproteomic analysis on samples from acute myeloid leukemia patients. *Biol Proced Online* **18**, 13 (2016).

179. Lipecka, J. et al. Sensitivity of mass spectrometry analysis depends on the shape of the filtration unit used for filter aided sample preparation (FASP). *Proteomics* **16**, 1852-1857 (2016).

180. Wisniewski, J.R. Quantitative Evaluation of Filter Aided Sample Preparation (FASP) and Multienzyme Digestion FASP Protocols. *Anal Chem* **88**, 5438-5443 (2016).

181. Muller, T. & Winter, D. Systematic Evaluation of Protein Reduction and Alkylation Reveals Massive Unspecific Side Effects by Iodine-containing Reagents. *Mol Cell Proteomics* **16**, 1173-1187 (2017).

182. Berger, S.T. et al. MStern Blotting-High Throughput Polyvinylidene Fluoride (PVDF) Membrane-Based Proteomic Sample Preparation for 96-Well Plates. *Mol Cell Proteomics* **14**, 2814-2823 (2015).

183. Rocken, C., Ebert, M.P. & Roessner, A. Proteomics in pathology, research and practice. *Pathol Res Pract* **200**, 69-82 (2004).

184. Rifai, N., Gillette, M.A. & Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotechnol* **24**, 971-983 (2006).

185. Sajic, T., Liu, Y. & Aebersold, R. Using data-independent, high-resolution mass spectrometry in protein biomarker research: perspectives and clinical applications. *Proteomics Clin Appl* **9**, 307-321 (2015).

186. Liang, S. et al. Quantitative proteomics for cancer biomarker discovery. *Comb Chem High Throughput Screen* **15**, 221-231 (2012).

187. Petricoin, E.F. et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359**, 572-577 (2002).

188. Buchen, L. Cancer: Missing the mark. *Nature* **471**, 428-432 (2011).

189. Check, E. Proteomics and cancer: running before we can walk? *Nature* **429**, 496-497 (2004).

190. Ioannidis, J.P. Biomarker failures. *Clin Chem* **59**, 202-204 (2013).

191. Lescuyer, P., Hochstrasser, D. & Rabilloud, T. How shall we use the proteomics toolbox for biomarker discovery? *J Proteome Res* **6**, 3371-3376 (2007).

192. Poste, G. Bring on the biomarkers. *Nature* **469**, 156-157 (2011).

193. Louis, D.N. et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol* **131**, 803-820 (2016).

194. Nicolaidis, S. Biomarkers of glioblastoma multiforme. *Metabolism* **64**, S22-27 (2015).

195. Pointer, K.B., Clark, P.A., Zorniak, M., Alrfaei, B.M. & Kuo, J.S. Glioblastoma cancer stem cells: Biomarker and therapeutic advances. *Neurochem Int* **71**, 1-7 (2014).

196. Jovcevska, I. et al. TRIM28 and beta-actin identified via nanobody-based reverse proteomics approach as possible human glioblastoma biomarkers. *PLoS One* **9**, e113688 (2014).

197. Autelitano, F. et al. Identification of novel tumor-associated cell surface sialoglycoproteins in human glioblastoma tumors using quantitative proteomics. *PLoS One* **9**, e110316 (2014).

198. Rapp, C. et al. Identification of T cell target antigens in glioblastoma stem-like cells using an integrated proteomics-based approach in patient specimens. *Acta Neuropathol* **134**, 297-316 (2017).

199. Kalinina, J., Peng, J., Ritchie, J.C. & Van Meir, E.G. Proteomics of gliomas: initial biomarker discovery and evolution of technology. *Neuro Oncol* **13**, 926-942 (2011).

200. Ghosh, D. et al. A Cell-Surface Membrane Protein Signature for Glioblastoma. *Cell Syst* **4**, 516-529 e517 (2017).

201. Clevers, H. The cancer stem cell: premises, promises and challenges. *Nat Med* **17**, 313-319 (2011).

202. Baker, M. Cancer stem cells tracked. *Nature* **488**, 13-14 (2012).

203. Sampetrean, O. & Saya, H. Characteristics of glioma stem cells. *Brain Tumor Pathol* **30**, 209-214 (2013).

204. He, J., Liu, Y. & Lubman, D.M. Targeting glioblastoma stem cells: cell surface markers. *Curr Med Chem* **19**, 6050-6055 (2012).

205. Reya, T., Morrison, S.J., Clarke, M.F. & Weissman, I.L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105-111 (2001).

206. Patru, C. et al. CD133, CD15/SSEA-1, CD34 or side populations do not resume tumor-initiating properties of long-term cultured cancer stem cells from human malignant glio-neuronal tumors. *BMC Cancer* **10**, 66 (2010).

207. Lacroix, M., Linares, L.K. & Le Cam, L. [Role of the p53 tumor suppressor in metabolism]. *Med Sci (Paris)* **29**, 1125-1130 (2013).

208. Allen, M., Bjerke, M., Edlund, H., Nelander, S. & Westermark, B. Origin of the U87MG glioma cell line: Good news and bad news. *Sci Transl Med* **8**, 354re353 (2016).

209. Garland, S.L. Are GPCRs Still a Source of New Targets? *J Biomol Screen* **18**, 947-966 (2013).

210. Zeromski, J., Nyczak, E. & Dyszkiewicz, W. Significance of cell adhesion molecules, CD56/NCAM in particular, in human tumor growth and spreading. *Folia Histochem Cytobiol* **39 Suppl 2**, 36-37 (2001).

211. Ravandi, F. et al. CD56 expression predicts occurrence of CNS disease in acute lymphoblastic leukemia. *Leuk Res* **26**, 643-649 (2002).

212. Kurokawa, M. et al. CD56: a useful marker for diagnosing Merkel cell carcinoma. *J Dermatol Sci* **31**, 219-224 (2003).

213. Waziri, A. et al. Preferential in situ CD4+CD56+ T cell activation and expansion within human glioblastoma. *J Immunol* **180**, 7673-7680 (2008).

214. Fu, K. et al. Addition of rituximab to standard chemotherapy improves the survival of both the germinal center B-cell-like and non-germinal center B-cell-like subtypes of diffuse large B-cell lymphoma. *J Clin Oncol* **26**, 4587-4594 (2008).

215. Maxwell, S.A., Cherry, E.M. & Bayless, K.J. Akt, 14-3-3zeta, and vimentin mediate a drug-resistant invasive phenotype in diffuse large B-cell lymphoma. *Leuk Lymphoma* **52**, 849-864 (2011).

216. Thieblemont, C. et al. The germinal center/activated B-cell subclassification has a prognostic impact for response to salvage therapy in relapsed/refractory diffuse large B-cell lymphoma: a bio-CORAL study. *J Clin Oncol* **29**, 4079-4087 (2011).

217. Liu, Y. et al. Identification of differentially expressed proteins in chemotherapy-sensitive and chemotherapy-resistant diffuse large B cell lymphoma by proteomic methods. *Med Oncol* **30**, 528 (2013).

218. Sabattini, E., Bacci, F., Sagramoso, C. & Pileri, S.A. WHO classification of tumours of haematopoietic and lymphoid tissues in 2008: an overview. *Pathologica* **102**, 83-87 (2010).

219. Goeminne, L.J., Gevaert, K. & Clement, L. Peptide-level Robust Ridge Regression Improves Estimation, Sensitivity, and Specificity in Data-dependent Quantitative Label-free Shotgun Proteomics. *Mol Cell Proteomics* **15**, 657-668 (2016).

220. Hennequart, M. et al. Constitutive IDO1 Expression in Human Tumors Is Driven by Cyclooxygenase-2 and Mediates Intrinsic Immune Resistance. *Cancer Immunol Res* (2017).

221. Srijakotre, N. et al. P-Rex1 and P-Rex2 RacGEFs and cancer. *Biochem Soc Trans* **45**, 963-977 (2017).

222. Cox, J. & Mann, M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* **13 Suppl 16**, S12 (2012).

223. Pio, R., Ajona, D. & Lambris, J.D. Complement inhibition in cancer therapy. *Semin Immunol* **25**, 54-64 (2013).

224. Ho, J., Rush, D.N. & Nickerson, P.W. Urinary biomarkers of renal transplant outcome. *Curr Opin Organ Transplant* **20**, 476-481 (2015).

225. Salvadori, M. & Tsalouchos, A. Biomarkers in renal transplantation: An updated review. *World J Transplant* **7**, 161-178 (2017).

226. Gwinner, W., Metzger, J., Husi, H. & Marx, D. Proteomics for rejection diagnosis in renal transplant patients: Where are we now? *World J Transplant* **6**, 28-41 (2016).

227. Loftheim, H. et al. Urinary proteomic shotgun approach for identification of potential acute rejection biomarkers in renal transplant recipients. *Transplant Res* **1**, 9 (2012).

228. Kim, S.C., Page, E.K. & Knechtle, S.J. Urine proteomics in kidney transplantation. *Transplant Rev (Orlando)* **28**, 15-20 (2014).

229. Bourgeois, G., Marcoux, J., Saliou, J.M., Cianferani, S. & Graille, M. Activation mode of the eukaryotic m2G10 tRNA methyltransferase Trm11 by its partner protein Trm112. *Nucleic Acids Res* **45**, 1971-1982 (2017).

230. Shimada, K. et al. A novel human AlkB homologue, ALKBH8, contributes to human bladder cancer progression. *Cancer Res* **69**, 3157-3164 (2009).

231. Swinehart, W.E. & Jackman, J.E. Diversity in mechanism and function of tRNA methyltransferases. *RNA Biol* **12**, 398-411 (2015).

232. Guerrero, C., Milenkovic, T., Przulj, N., Kaiser, P. & Huang, L. Characterization of the proteasome interaction network using a QTAX-based tag-team strategy and protein interaction network analysis. *Proc Natl Acad Sci U S A* **105**, 13333-13338 (2008).

233. Sutherland, B.W., Toews, J. & Kast, J. Utility of formaldehyde cross-linking and mass spectrometry in the study of protein-protein interactions. *J Mass Spectrom* **43**, 699-715 (2008).

234. Heurgue-Hamard, V. et al. The glutamine residue of the conserved GGQ motif in Saccharomyces cerevisiae release factor eRF1 is methylated by the product of the YDR140w gene. *J Biol Chem* **280**, 2439-2445 (2005).

235. Liger, D. et al. Mechanism of activation of methyltransferases involved in translation by the Trm112 'hub' protein. *Nucleic Acids Res* **39**, 6249-6259 (2011).

236. Yang, B. et al. Identification of cross-linked peptides from complex samples. *Nat Methods* **9**, 904-906 (2012).

237. Walzthoeni, T. et al. False discovery rate estimation for cross-linked peptides identified by mass spectrometry. *Nat Methods* **9**, 901-903 (2012).

238. Leitner, A., Faini, M., Stengel, F. & Aebersold, R. Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends Biochem Sci* **41**, 20-32 (2016).

239. Nguyen-Huynh, N.T. et al. Chemical cross-linking and mass spectrometry to determine the subunit interaction network in a recombinant human SAGA HAT subcomplex. *Protein Sci* **24**, 1232-1246 (2015).

240. Serpa, J.J. et al. Mass spectrometry-based structural proteomics. *Eur J Mass Spectrom (Chichester)* **18**, 251-267 (2012).

241. Artigues, A. et al. Protein Structural Analysis via Mass Spectrometry-Based Proteomics. *Adv Exp Med Biol* **919**, 397-431 (2016).

242. Rinner, O. et al. Identification of cross-linked peptides from large sequence databases. *Nat Methods* **5**, 315-318 (2008).

243. Leitner, A., Walzthoeni, T. & Aebersold, R. Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline. *Nat Protoc* **9**, 120-137 (2014).

244. Roubin, R. et al. Myomegalin is necessary for the formation of centrosomal and Golgi-derived microtubules. *Biol Open* **2**, 238-250 (2013).

245. Wu, J.C. et al. Molecular Pathway of Microtubule Organization at the Golgi Apparatus. *Developmental Cell* **39**, 44-60 (2016).

246. Combe, C.W., Fischer, L. & Rappsilber, J. xiNET: cross-link network maps with residue resolution. *Mol Cell Proteomics* **14**, 1137-1147 (2015).

247. Radu, L. et al. The intricate network between the p34 and p44 subunits is central to the activity of the transcription/DNA repair factor TFIIH. *Nucleic Acids Res* (2017).

248. Luo, J. et al. Architecture of the Human and Yeast General Transcription and DNA Repair Factor TFIIH. *Mol Cell* **59**, 794-806 (2015).

249. Duarte, T.T. & Spencer, C.T. Personalized Proteomics: The Future of Precision Medicine. *Proteomes* **4** (2016).

PARTIE EXPERIMENTALE

I- Préparations d'échantillons

1- Recherche de biomarqueurs de glioblastomes

Gel « *Stacking* » : Les protéines ont été dénaturées à 100 °C pendant 10 minutes dans un tampon contenant 5 % de SDS, 5 % de 2-mercaptoéthanol, 1 mM d'EDTA, 10 % de glycérol et 10 mM de Tris-HCl à pH 6,8. Au total, 25 µg de protéines par échantillon ont été migrées dans la partie haute (gel de concentration à 4 % d'acrylamide) d'un gel 1D SDS-PAGE. Les gels ont été fixés à l'aide d'une solution composée de 50 % d'éthanol et 3 % d'acide phosphorique, et colorés au bleu colloïdal « *Silver Blue* ». La bande de « *Stacking* » a été excisée, ainsi que la partie supérieure à cette bande, découpée en deux bandes.

Réduction/Alkylation en gel : Les bandes de gel ont été décolorées à l'aide de 100 µl d'un mélange ACN/bicarbonate d'ammonium à 25 mM ($^3/_4/^1/_4$). Cette opération a été répétée 4 fois. Celles-ci ont été déshydratées avec 50 µl d'ACN avant réduction des cystéines avec 50 µl de DTT à 10 mM pendant 30 minutes à 60 °C, puis 30 minutes à température ambiante. L'alkylation a ensuite été réalisée à l'aide d'IAA à 55 mM pendant 20 minutes dans le noir. Pour finir, les bandes de gel ont été lavées trois fois à l'aide de 50 µl de bicarbonate d'ammonium à 25 mM, puis déshydratées deux fois avec 50 µl d'ACN. Les bouts de gel ont été stockés à -20 °C jusqu'à la digestion enzymatique.

Digestion : Les protéines de la bande « *Stacking* » ont été digérées à l'aide de 12,5 µl d'une solution de trypsine porcine modifiée (Promega) à 0,04 µg/µl dans 25 mM de bicarbonate d'ammonium. Pour les deux bandes correspondant à la partie haute du gel, les protéines ont été digérées avec 5 µl de solution de trypsine modifiée à 12,5 ng/µl dans 25 mM de bicarbonate d'ammonium. Du tampon bicarbonate d'ammonium a été ajouté à chaque échantillon de manière à atteindre un volume final de 25 µl, afin que la totalité des bouts de gel soient immergés. La digestion a été réalisée sur la nuit à 37 °C.

Extraction des peptides du gel : Deux extractions successives ont été réalisées afin d'extraire les peptides du gel. La première a été réalisée pendant une heure avec 40 µl d'une solution composée de 60 % d'ACN et 0,1 % de FA. La seconde a été effectuée avec 40 µl d'ACN à 100 % pendant une heure. Les extraits peptidiques ont été réunis dans un puits, et l'excès d'ACN a été évaporé sous vide à l'aide d'un SpeedVac™ (Thermo Scientific, Waltham, USA) après dopage des échantillons aux peptides iRT.

Les peptides ont été resolubilisés dans 50 µl d'un mélange H_20/ACN/FA (98/2/0,1), et ont été soniqués dans la glace pendant 20 minutes avant analyse nanoLC-MS/MS.

2- Recherche de biomarqueurs de LBDGC

Extraction de protéines à partir de tissus frais : A partir d'une dizaine de milligrammes de biopsies, les protéines ont été extraites en deux fois à l'aide d'un potter en verre. Une première fois à l'aide de 100 µl de tampon Laemmli sans bleu de bromophénol et sans DTT (soit 62,5 mM de Tris-HCl à pH 6,8, 10 % de glycérol et 2 % de SDS), puis une seconde fois à l'aide de 85 µl de ce même tampon. Les deux extraits, ainsi que les morceaux de biopsies restants, ont été réunis dans un Eppendorf et ont été soumis à une sonication de 5 minutes dans un bain à ultrasons. La quantité de protéines pour chaque échantillon a été dosée à l'aide d'un dosage DC (Bio-Rad). Au total, 20 µg ont été préparés en « *Tube-Gel* ».

Extraction de protéines à partir de tissus FFPE : Les extractions ont été réalisées sur des copeaux de tissus FFPE de 20 µm. Dans un premier temps, les copeaux ont été déparaffinés à l'aide de deux bains de 5 minutes avec 1 ml de xylène sous agitation. Chaque bain a été suivi d'une centrifugation à 12000 x g pendant 3 minutes de manière à retirer le surnageant. Les tissus ont par la suite été réhydratés à l'aide de trois bains successifs de 5 minutes avec 1 ml d'éthanol à 100 %, 90 % et 70 % respectivement. Du tampon Laemmli (soit 62,5 mM de Tris-HCl à pH 6,8, 10 % de glycérol et 2 % de SDS) a ensuite été ajouté aux tissus FFPE. Les échantillons ont été chauffés dans ce tampon à 100 °C pendant 20 minutes, et à 80 °C pendant 2 heures de manière à inverser la réaction de pontage au formaldéhyde et extraire les protéines. Pour finir, les échantillons ont été soniqués pendant 20 secondes dans un bain à ultrasons avant d'effectuer le dosage protéique (dosage DC, Bio-Rad) et de préparer les TG.

« *Tube-Gel* » : Dans un tube Eppendorf de 0,5 ml, différents produits ont été ajoutés à l'extrait protéique de manière à obtenir les concentrations finales suivantes : 2 % de SDS, 32 % d'H_2O, 7,5 % d'acrylamide/Bis-acrylamide et 0,25 % de TEMED. Les solutions ainsi générées ont été vortexées et centrifugées de manière à retirer l'ensemble des bulles, étant donné que l'oxygène inhibe la polymérisation. Les TG résultants ont été fixés avec une solution contenant 45 % de méthanol et 5 % d'acide acétique avant d'être découpés en morceaux d'environ 2 mm^2.

Réduction/alkylation : Ce protocole est le même que celui détaillé pour la recherche de biomarqueurs de glioblastomes. Cependant, les volumes de solution utilisés ont été quadruplés.

Digestion : Les protéines ont été digérées à l'aide de 40 µl d'une solution de trypsine porcine modifiée (Promega) à 0,01 µg/µl dans 25 mM de bicarbonate d'ammonium, à laquelle 110 µl de tampon bicarbonate d'ammonium à 25 mM ont été ajoutés de manière à recouvrir l'ensemble des bouts de gel. La digestion a été réalisée à 37 °C sur la nuit.

Extraction des peptides du gel : Ce protocole est le même que celui détaillé pour la recherche de biomarqueurs de glioblastomes. Cependant, les volumes de solution utilisés ont été quadruplés. Les échantillons ont été évaporés sous vide avant d'être resolubilisés dans 50 µl d'un mélange H_20/CAN/FA (98/2/0,1) contenant les peptides iRT, et soniqués dans la glace pendant 20 minutes avant analyse nanoLC-MS/MS.

3- Recherche de biomarqueurs protéiques urinaires

Protocole FASP : Chaque échantillon d'urine a été centrifugé à 4 °C pendant 10 minutes à 4000 x g afin de se débarrasser des débris cellulaires. Six aliquots de 1,2 ml de surnageant ont été préparés et stockés à -80 °C. A l'aide d'un aliquot, 500 µl d'urine ont été déposés sur des filtres microcon avec un seuil de coupure de 10 kDa. Les échantillons ont été centrifugés à 14000 x g pendant 40 minutes à température ambiante de manière à concentrer les protéines sur le filtre. Par la suite, deux-cent microlitres d'un tampon UA (urée 8 M dans 0,1 M de Tris-HCl à pH 8,5) ont été ajoutés et centrifugés pendant 40 minutes à 14000 x g à température ambiante. Cette étape a été répétée deux fois. La réduction des cystéines a été effectuée à 37 °C pendant 30 minutes à l'aide de 100 µl de DTT à 10 mM dans UA. L'alkylation des cystéines a ensuite été réalisée avec 100 µl d'IAA à 0,05 M dans UA pendant 20 minutes à l'abri de la lumière. Les filtres ont été centrifugés à 14000 x g pendant 40 minutes à température ambiante avant ajout de 100 µl d'UA aux filtres et centrifugation à 14000 x g pendant 40 minutes. Un échange de tampon a finalement été réalisé par ajout de 100 µl de bicarbonate d'ammonium à 0,05 M suivi d'une centrifugation à 14000 x g pendant 40 minutes à température ambiante. Cette étape a été réalisée à deux reprises.

Digestion : Pour finir, les échantillons ont été digérés à l'aide de 40 µl de trypsine à 0,016 µg/µl pour les échantillons contenant 1 à 100 µg de protéines, et de trypsine à 0,047 µg/µl pour les échantillons contenant une quantité supérieure à 100 µg. La digestion a été réalisée sur la nuit à 37 °C. Après digestion, les échantillons ont été centrifugés à 14000 x g pendant 40 minutes, puis 50 µl de chlorure de sodium ont été ajoutés avant d'être centrifugés à 14000 x g pendant 20 minutes. Cette étape a été réalisée deux fois avant d'ajouter 5 µl d'acide trifluoroacétique à chacun des échantillons de manière à stopper la digestion.

Extraction des peptides : Les peptides ont été extraits et déssalés par SPE. Ainsi, les cartouches Sep-Pack® C18 (WATERS) ont dans un premier temps été mouillées deux fois avec 1 ml de méthanol, et trois fois avec 1 ml d'ACN, avant d'être conditionnées 3 fois à l'aide d'eau acidifiée avec 0,1 % de FA. Par la suite, les échantillons ont été chargés sur les cartouches SPE. Celles-ci ont ensuite été lavées deux fois par 1 ml d'eau acidifiée avec 0,1 % de FA, et l'élution a été effectuée avec 600 µl d'un mélange contenant 60 % d'ACN et 0,1 % de FA.

Les échantillons ont été divisés en deux avant d'être évaporés sous vide. Ainsi,

- les premières moitiés ont constitué la série $Q_{éq}$, pour laquelle les volumes de solution de reprise (H_20/CAN/FA (98/2/0,1) contenant les peptides iRT) ont été adaptés à chaque échantillon de manière à obtenir des solutions à 400 ng/µl de protéines de départ théorique.

- et les deuxièmes ont constitué la série $V_{éq}$, pour laquelle 112 µl de solution de reprise (H_20/CAN/FA (98/2/0,1) contenant les peptides iRT) ont été ajoutés aux culots peptidiques.

L'ensemble des échantillons a été soniqué dans la glace avant analyse nanoLC-MS/MS.

II- Conditions chromatographiques

Les séparations chromatographiques ont été, pour l'ensemble des analyses, effectuées à l'aide de systèmes Nano-Acquity UPLC (WATERS) sur des colonnes ACQUITY UPLC BEH 130 C18 de 250 mm x 75 µm et un diamètre de particules de 1,7 µm. Avant séparation, les peptides ont été concentrés sur une précolonne Symmetry C18 (WATERS) de 20 mm x 180 µm avec un diamètre de particules de 5 µm. Le solvant A consistait en de l'eau acidifiée avec 0,1 % de FA, et le solvant B en de l'ACN acidifié avec 0,1 % de FA. Les gradients utilisés étaient différents selon les projets (voir Tableau Partie expérimentale – 1), cependant la concentration des peptides sur la précolonne a toujours été effectuée pendant 3 minutes à 5 µl/min avec 99 % de solvant A et 1 % de solvant B.

Projet	Spectromètre de masse	Débit chromatographique	Gradient chromatographique utilisé
Biomarqueurs Glioblastome	Impact-HD	450 nl.min^{-1}	2 à 35% de B pendant 180 min, Puis 35 à 80% de B en 1 min
Biomarqueurs LBDGC	Q-Exactive +	450 nl.min^{-1}	1 à 35% de B en 120 min, Puis 35 à 80 % en 1 min
Biomarqueurs Protéome urinaire	Q-Exactive +	450 nl.min^{-1}	1 à 35% de B en 90 min, Puis 35 à 90 % en 1 min

Tableau Partie expérimentale – 1- Détail des gradients et débits chromatographiques employés dans les différents projets, ainsi que les spectromètres de masse

III – Paramètres d'acquisition par spectrométrie de masse

1- Impact-HD (BRUKER)

Cet instrument a été utilisé pour l'ensemble des analyses liées au projet de recherche de biomarqueurs de glioblastomes.

Pour ces analyses, le système fonctionnait avec un changement automatique entre le mode MS (500 ms/spectre sur une gamme de masse de 150 à 2200 m/z) et le mode MS/MS. Les ions les plus intenses sur le spectre MS au court d'un temps de cycle fixe (MS + MS/MS) de 3,5 secondes, avec une préférence pour les ions multi-chargés et une stricte exclusion des ions monochargés, ont été sélectionnés pour être isolés et fragmentés par CID. L'exclusion dynamique a été fixée à 60 secondes, et les précurseurs ont été reconsidérés pour être à nouveau isolés et fragmentés si leur intensité était trois fois plus élevée que lors de la première sélection. La tension du capillaire a été fixée à 1300 Volts, et sa température à 150 °C.

2- Q-Exactive + (THERMO FISHER SCIENTIFIC)

Cet instrument a été employé pour le projet de recherche de biomarqueurs protéiques urinaires et de résistance aux traitements des LBDGC. Les paramètres étaient alors les suivants.

La tension du capillaire était fixée à 1,8 kVolts et sa température à 250 °C. Le système fonctionnait en mode DDA avec un changement automatique entre les modes MS (gamme de masse de 300 à 1800 m/z, une résolution de 70000, un « *Automatic Gain Control* » de $3x10^6$ ions et un temps d'injection maximum de 50 ms) et MS/MS (gamme de masse de 200 à 2000 m/z avec une résolution de 17500,

un « *Automatic Gain Control* » de $1x10^5$ et un temps d'injection maximum de 100ms). Les dix ions les plus intenses avec une intensité supérieure à $2x10^5$ sur le spectre MS ont été sélectionnés afin d'être isolés et fragmentés par HCD. Les ions monochargés, et dont la charge n'a pas été assignée, ont été exclus de cette sélection. Le temps d'exclusion dynamique a quant à lui été fixé à 60 secondes.

ANNEXE 1

LISTE DES PROTEINES DIFFERENTIELLEMENT EXPRIMEES DANS LE PROJET DE RECHERCHE DE BIOMARQUEURS ASSOCIES A UNE CHIMIORESISTANCE PRIMAIRE DANS LES LYMPHOMES B DIFFUS A GRANDES CELLULES

Le ratio protéique correspond au ratio (Résistants/Sensibles)

N° accession protéine	Ratio protéique
P01023	0,28
Q9BTE6	-0,43
Q9NRN7	0,60
Q7Z7G0	-1,08
P09110	0,74
Q15027	0,42
P21399	0,56
Q9H2P0	-0,44
P54922	0,92
Q09666	0,15
O95433	-1,00
P27144	0,65
P14550	0,32
P15121	0,35
P54886	-0,26
P30837	-0,46
P09972	0,30
Q8IZ07	-0,52
Q5VYY1	2,52
P04083	0,34
P07355	0,33
P08758	0,58
P08133	1,09
Q16853	-0,79
P13798	0,20
P06727	-0,49
P04114	0,48
Q9HC16	1,06
P02649	0,36
Q9BQE5	0,63
O95236	1,28
P07741	0,99
Q07960	0,26
P98171	0,32
O75915	0,49
Q12797	0,74
Q6DD88	0,38
P05023	-0,27
O75947	-0,55
P25311	0,30
Q9NYF8	-0,37
P21810	-1,06
Q9UBW5	0,52
P51451	1,08

N° accession protéine	Ratio protéique
P53004	0,45
Q14137	-0,49
P07738	-0,98
P17213	-3,12
O95861	-0,83
Q10589	-0,90
O14981	-0,42
Q13895	-0,54
P02745	1,06
Q07021	-0,47
P01024	0,31
P0C0L4	1,16
P04003	0,98
P01031	0,64
P02748	0,68
Q9BUH6	0,44
O76075	-0,17
Q05682	-0,93
P49913	-1,57
P07384	0,22
Q14444	-0,39
P04040	-0,34
P16152	0,59
Q8IX12	-0,32
Q96CT7	-0,46
Q4VC31	-0,61
P08571	0,65
P28907	0,56
P40259	-1,65
P01732	1,44
P21926	-0,90
P48960	0,51
Q99459	-0,30
Q00534	-0,91
O75420	0,67
P02795	1,32
P00751	0,62
P08603	0,50
Q14839	-0,38
Q8IWX8	-0,41
O75339	-1,53
Q07065	-0,28
O00299	0,74
Q9UDT6	-1,79

N° accession protéine	Ratio protéique
O76031	-0,42
P10909	1,25
O75153	-0,61
Q99715	-1,69
Q05707	-0,73
P02462	1,45
P12109	-0,42
Q14019	0,35
P00450	0,64
Q99829	0,31
Q9BRF8	0,74
Q8N684	-0,37
Q9H3G5	0,82
P17927	0,77
O75718	-0,58
O75534	-0,18
P01040	1,36
P33240	-0,41
P07858	0,69
Q9UBR2	0,45
Q13616	-0,32
Q93034	-0,33
O43927	2,95
P27707	-0,48
P07585	-0,73
Q9H773	-0,57
Q7Z4W1	0,77
O95865	0,65
Q9NVP1	-0,91
Q9BUQ8	-0,36
Q9GZR7	-0,48
Q96GQ7	-0,51
Q86XP3	-0,35
Q9H0S4	-0,60
P26196	-0,26
O43583	-0,57
Q13574	0,44
Q7L2E3	-0,30
Q08211	-0,23
O60610	0,31
P10515	-0,25
O60884	-0,55
O75165	0,47
Q99615	-0,33

N° accession protéine	Ratio protéique
Q9ULA0	0,65
Q92608	0,22
Q8NF50	-0,33
O60496	1,44
Q9C005	-0,64
Q12882	0,60
Q16555	0,28
P15924	1,68
P51452	0,61
Q14204	0,23
O75923	0,95
Q14213	1,75
Q99848	-0,44
P42126	1,17
Q5VYK3	0,32
Q15075	-0,30
Q96C19	0,54
Q9H4M9	0,33
P20042	-0,22
P55884	-0,18
O15371	-0,21
P60228	-0,33
O75821	-0,35
Q96JJ3	0,48
Q92979	-0,68
Q9Y6C2	0,45
P07814	-0,21
Q96HE7	0,41
P30040	0,41
O95571	0,52
P15311	-0,57
P00734	0,49
Q9H098	-0,60
Q9BZQ8	0,36
Q5R3K3	1,81
Q92520	-0,90
Q8NCA5	-1,07
P49327	-0,15
P22087	-0,39
P23142	-0,75
Q9UBX5	-0,99
P35555	1,02
P09467	0,64
P30273	1,76

N° accession protéine	Ratio protéique
P12314;Q92637	1,52
P02671	1,23
P02675	1,55
P02679	1,80
Q02790	-0,29
P21333	-0,29
O95466	0,26
Q96RU3	-0,57
Q16658	-0,36
P02794	0,63
P02792	0,70
Q8IY81	-0,71
Q96AE4	-0,19
O15117	0,54
P10253	0,41
P51570	0,36
Q96C23	0,87
P32455	0,93
P32456	1,15
P28676	0,56
P57678	-0,53
Q9UJY4	-0,63
Q8WWP7	0,59
Q9NUV9	0,94
Q96F15	0,76
Q8NHV1	1,31
Q9H4G4	0,64
P17900	0,60
Q96IJ6	0,51
Q9Y5P6	0,35
P04899	0,69
Q9BVP2	-0,74
P13224	-3,23
P07203	0,74
O75791	1,17
P48637	0,46
P78347	-0,45
A4D1E9	-0,80
Q9BZE4	-0,74
P10144	1,29
P20718	2,70
P49863	0,58
O75367	-0,20
O95479	0,65

N° accession protéine	Ratio protéique
Q5VWC8	-1,08
P40939	0,34
P08631	0,57
Q9H583	-0,40
P52790	1,67
P06340	-1,06
P04440	-1,01
P01920	-0,83
P28845	1,54
Q3SXM5	-0,58
P14625	0,21
Q0VDF9	-0,32
P10809	-0,35
P61604	-0,53
P98160	-0,38
O43719	-0,56
Q7Z6Z7	-0,14
P41252	-0,29
O75874	0,46
P14902	2,92
P80217	1,52
P11717	0,55
P12268	-0,37
Q92835	-0,43
Q8N201	-0,45
P46940	0,39
Q13576	0,42
Q6DN90	-0,64
P05161	-0,82
O14498	-2,23
P20702	-0,63
P07476	2,49
Q96CX2	0,47
O60341	-0,39
Q92945	-0,20
Q9NQT8	0,77
Q9P2G3	2,13
P52292	-0,31
Q14974	-0,21
P83111	0,83
Q53H82	-0,93
P28838	0,44
Q14847	-0,55
P18428	1,07

N° accession protéine	Ratio protéique
Q14739	-0,41
Q13094	0,88
P49257	-0,29
Q12907	0,34
Q8NF37	-0,40
P02750	0,87
Q12912	-0,85
Q07954	0,56
P42704	-0,39
Q8N386	1,19
Q9H9A6	-0,56
Q5S007	1,23
O15116	0,48
P09960	0,74
Q14767	-0,97
P51884	-0,91
Q9BS40	-1,35
P61626	0,59
Q9UDY8	-0,81
Q92918	-0,32
Q9UBB5	-0,70
Q96RQ3	-0,69
P49736	-0,28
P33991	-0,35
P33992	-0,33
Q14566	-0,39
P33993	-0,42
P48163	1,09
P23368	-0,44
Q9H8H3	0,44
O14880	0,63
P41218	0,74
P25325	0,53
Q9UBG0	-0,73
Q9BYD6	-0,44
P52815	-0,62
Q9P015	-0,58
P49406	-0,41
Q9NWU5	-0,44
Q8N983	-0,46
Q13405	-0,60
P82932	-0,97
P00403	0,38
P11586	-0,22

N° accession protéine	Ratio protéique
Q14764	0,37
P20592	-1,17
Q9NR99	0,66
Q9BQG0	-0,39
O75592	-0,55
P35749	-0,97
O94832	-0,68
O00160	0,80
B0I1T2	0,44
Q9NZM1	1,61
Q9NZM1	1,74
Q9UJ70	0,65
P43490	0,27
Q15021	-0,44
Q9BPX3	-0,46
P19338	-0,38
Q00653	0,44
Q9ULX3	-0,76
Q14978	-0,70
O00567	-0,25
Q9BXD5	0,74
Q9H0P0	-0,64
Q9BSD7	-0,49
Q02818	0,37
O43809	-0,25
Q8WUM0	-0,20
O75694	-0,23
Q92621	-0,29
Q8NFH4	-0,46
P20774	-1,17
Q6UX06	-3,70
Q6UWY5	-1,18
O60313	-0,26
P02763	1,31
P22059	-0,27
Q92882	0,41
Q9UNF0	0,41
Q99497	0,26
O95453	-0,71
P09874	-0,49
Q460N5	0,57
Q9UKK3	0,48
Q8IXQ6	0,53
Q9BVG4	-0,47

N° accession protéine	Ratio protéique
Q8WW12	-0,77
Q13442	-0,46
Q9BUL8	-0,40
Q14690	-0,53
Q9NTI5	-0,29
Q8IZL8	-0,39
O00541	-0,63
O15067	-0,50
O60925	-0,65
Q9UHV9	-0,76
Q99471	-0,53
O15212	-0,58
P08237	-0,65
O95336	0,35
Q96G03	0,40
Q99623	-0,67
Q7RTV0	-0,69
O43175	-0,49
P47712	1,78
P16885	-0,29
P08567	0,50
P00747	0,63
O60568	0,57
O15162	0,62
P29590	0,87
O75439	-0,30
Q8TCS8	-0,48
Q7Z3K3	-0,34
O15160	-0,37
P16435	-0,43
Q15181	0,84
Q9NQ55	-0,73
Q06203	-0,48
Q96SB3	-0,44
Q15257	0,47
Q8TCU6	0,83
P10644	0,83
P05771	0,86
Q99873	-0,63
Q9UMS4	-0,26
O43395	-0,45
Q6P2Q9	-0,16
Q14558	0,37
P25789	0,50

N° accession protéine	Ratio protéique
P20618	0,33
P40306	0,72
P49720	0,47
P28072	-1,30
P28065	0,70
O75832	-0,63
O43242	-0,32
Q9UL46	0,55
P61289	-0,48
Q8WXF1	-0,39
Q96EY7	-0,46
Q96BW5	1,07
P18031	0,51
P23469	0,68
Q15397	-0,46
Q00577	0,31
Q15269	-0,58
Q92626	-0,66
Q9ULZ3	0,72
P11216	0,57
P06737	1,08
Q9UL25	-0,41
Q9H2M9	-0,45
P54727	-0,90
P78406	-0,44
P62826	-0,29
P43487	-0,41
P49792	-0,28
P46060	-0,19
P98175	-0,27
Q96PK6	-0,22
Q9NW13	-0,50
P02753	-0,58
P18754	-0,64
Q04864	-0,73
P51606	0,63
Q14699	-0,80
Q96DB5	-0,47
P12724	-1,33
Q63HN8	0,78
O60930	0,59
O43148	-0,62
Q9H4A4	0,49
P62750	-0,39

N° accession protéine	Ratio protéique
P61353	-0,31
P63173	-0,52
P36578	-0,35
Q02878	-0,32
P18124	-0,24
P62424	-0,32
Q96P16	-1,36
P25398	-0,43
P62249	-0,29
P62269	-0,26
P39019	-0,28
P15880	-0,24
P60866	-0,35
P62851	-0,39
P61247	-0,33
P62753	-0,31
P46781	-0,29
P10301	0,63
P23921	-0,67
P05386	-0,55
Q9Y3B9	-0,51
O43818	-0,48
O76021	-0,46
P31949	0,55
P26447	0,85
P05109	1,67
P06702	1,89
Q9NSI8	-2,68
Q9UHR5	-0,67
O43290	-0,22
O14595	-0,34
O95810	-1,34
P01009	1,14
P01011	0,72
P30740	0,93
P35237	1,20
P50453	0,95
P05546	0,47
P36955	-0,56
P05155	1,00
Q01105	-0,33
Q15459	-0,41
O75533	-0,33
Q13435	-0,30

N° accession protéine	Ratio protéique
Q9BX95	-0,55
Q9BZZ2	0,95
Q13291	-2,42
O75746	-0,53
Q9UJS0	-0,44
Q99808	-0,56
Q9UGQ3	1,19
Q9GZT3	-0,51
O60264	-0,36
Q14683	-0,44
Q9UQE7	-0,33
A6NHR9	-0,42
Q7KZF4	-0,26
O75643	-0,17
P09661	-0,31
Q15036	0,36
O60749	-0,25
P35610	0,91
P00441	-0,37
P04179	1,16
O75398	0,71
P02549	-1,19
Q01082	-0,56
Q9Y6N5	0,73
Q9Y5M8	-0,35
Q13243	-0,44
Q13242	-0,44
Q08945	-0,23
P51692	0,58
O95793	-0,35
O75558	1,39
O15400	-0,26
Q15833	0,35
P53597	-0,53
Q15022	-0,53
Q9UH65	-0,58
O43760	-0,76
Q01995	-0,80
Q03518	0,86
Q03519	0,87
O15533	0,86
Q9BW92	-0,58
Q8IV04	0,34
Q66K14	0,41

N° accession protéine	Ratio protéique
Q15369	-0,44
Q13488	0,61
P17987	-0,20
Q15554	-0,68
P02786	-0,41
Q5TEJ8	1,19
Q9Y2W1	-0,38
Q3ZCQ8	-0,74
Q9Y490	0,10
P42166	-0,55
Q9H3N1	-0,29
P24821	-0,72
Q03169	1,10
Q68CZ2	0,96
Q9NS69	-1,00
O94826	-0,37
P11387	-0,64
P67936	-0,42
P51580	0,47
O14773	0,44
P29144	-0,20
P12270	-0,28
Q13077	-0,80
Q2NL82	-0,56
P02766	-0,81
P49411	-0,35
O43396	-0,39
O15042	-0,51
P41226	0,78
Q9Y385	-0,65
P17480	-0,46
P09936	-1,40
Q92900	-0,29
P22695	-0,22
O60287	-0,42
Q9Y4E8	0,32
Q9Y5J1	-0,49
O75691	-1,00
Q9NYH9	-0,43
P15498	0,26
P19320	-0,50
P13611	-1,00
P18206	-0,63
Q9P253	0,37

N° accession protéine	Ratio protéique
Q96AX1	0,22
Q96QK1	0,26
Q99986	-0,50
P04004	1,00
P04275	0,51
Q9GZL7	-0,47
Q9BV38	-0,42
Q9UNX4	-0,39
Q8NI36	-0,46
P54577	-0,19
Q04917	-0,42
O75844	-0,52
Q86UK7	-0,99
Q15942	-0,56

ANNEXE 2

LISTE DES GENES DIFFERENTIELLEMENT EXPRIMES DANS LE PROJET DE RECHERCHE DE BIOMARQUEURS ASSOCIES A UNE CHIMIORESISTANCE PRIMAIRE DANS LES LYMPHOMES B DIFFUS A GRANDES CELLULES

Le ratio transcrits correspond au ratio (Résistants/Sensibles)

Nom de gène	Ratio transcrits
ABCA2	1,11
ABCD2	1,71
ALOX15B	1,89
ANGPTL4	1,56
APBB2	-1,31
AQP9	2,11
ARHGEF3	1,30
ARL6IP5	0,92
BCL11B	1,54
BHLHE40	0,94
BMP7	-1,79
BMP8B	1,86
BNC2	-1,29
BTBD9	0,80
BTD	0,80
C10orf35	1,62
C10orf47	1,45
C11orf21	1,57
C3	-1,44
CCL11	-1,85
CCL4	1,43
CCL5	1,31
CCNB1IP1	-1,01
CCR2	1,95
CD247	1,65
CD3E	1,49
CD3G	1,37
CD5	1,41
CD6	1,56
CD79B	-1,32
CD97	1,15
CDC42SE1	0,72
CDH23	1,46
CENPV	-1,99
CHI3L1	1,61
CHST1	-1,38
CIDEB	1,19
CLDN7	2,29
CLPTM1L	-0,72
COLEC12	-1,18
CP	1,68
CPM	1,54
CRAT	1,06
CRHBP	-1,53
CRTC3	1,13

Nom de gène	Ratio transcrits
CXCL13	1,63
CXCL5	1,66
DCBLD2	-1,09
EEF1G	-0,87
EMP3	1,02
EOMES	1,41
ERAP2	-1,18
FAM171B	-2,04
FAM186B	-1,77
FAM208B	-1,18
FAM78A	0,90
FBL	-0,80
FCGR2C	1,96
FGR	1,40
FKBP5	1,15
FMN1	1,35
FOXO3B	-1,25
FPR2	1,62
FURIN	1,06
G0S2	1,53
GBP5	1,67
GCSAM	-1,43
GIPR	1,44
GLB1L2	1,63
GLUL	1,46
GPBAR1	2,02
GRIP1	-1,80
GZMA	1,43
GZMH	1,70
GZMK	1,46
GZMM	1,48
HK3	1,86
HLA-C	1,45
HSPA12B	-1,54
HTRA1	-1,55
IDO1	1,71
IFNG	1,88
IL18RAP	1,89
IL21	1,66
IL2RB	1,50
IL32	1,39
IL4R	-1,09
IL6R	1,11
INPP4A	1,04
IRF2BPL	0,98

Nom de gène	Ratio transcrits
ITGA11	-1,68
ITGBL1	-1,61
ITK	1,68
ITPK1	0,83
KANK1	-2,08
KAT2B	0,98
KCNJ2	2,02
LAG3	1,55
LAMC3	1,81
LIMA1	-1,59
LMTK3	1,75
LOC100129269	-1,57
LOC100216546	-1,52
LOC100506888	-1,53
LOC100616530	-1,56
LTK	2,14
MARCH1	1,06
MARCH2	0,93
MARCO	2,17
MET	1,54
MGAT3	-1,16
MGC12916	1,50
MMP11	-1,27
MPND	1,03
MSX2	1,86
MT1F	1,81
MT1G	2,18
MT1H	1,60
MT1L	1,65
MT1M	2,45
MT1X	2,51
MT2A	2,36
MUC1	1,74
MYADML	-1,65
NACC2	1,25
NDRG1	1,38
NKG7	1,85
NLE1	-0,86
NRXN2	1,65
NSUN5	-0,79
NTM	-1,85
NUP85	-0,81
OBSCN	1,66
PALD1	-1,74
PCDHGB8P	-1,36

Nom de gène	Ratio transcrits
PDCD1	1,37
PDGFC	-1,41
PI15	1,89
PIAS3	0,69
PILRA	1,50
PLA2G16	1,11
PLCB2	1,12
PLCH2	1,62
PLTP	1,53
POTEH	1,62
PPM1L	-1,43
PPM1N	1,74
PREX1	1,07
PRND	-1,91
PTK6	1,50
QSOX2	-1,13
RAB27A	1,07
RCAN3	0,69
RFFL	0,90
RGN	-1,68
RIN3	1,21
RINT1	-1,05
RPL12	-0,96
RPL13	-1,08
RPL13A	-0,99
RPL18A	-0,89
RPL24	-0,77
RPL32	-0,92
RPL35A	-1,16
RPL36A	-1,04
RPL5	-0,97
RPS12	-0,88
RPS15A	-1,08
RPS18	-1,07
RPS6	-0,94
RPS8	-0,93
S100A4	1,38
S100A6	1,06
S100A8	1,73
SCARA5	1,59
SCARNA16	-1,27
SEMA4B	1,14
SERPINA1	1,80
SERPINB6	1,55
SH2B2	-1,45

Nom de gène	Ratio transcrits
SH2D1A	1,43
SIGIRR	1,50
SIGLEC14	2,19
SIRPG	1,65
SLC11A1	1,96
SLC25A35	1,00
SLC2A4RG	1,06
SLC43A1	-1,09
SLED1	1,51
SNHG4	-1,06
SNHG8	-1,09
SNORA1	-0,85
SNORA10	-1,32
SNORA14B	-1,05
SNORA21	-1,34
SNORA24	-1,60
SNORA27	-1,73
SNORA33	-1,54
SNORA53	-0,96
SNORA55	-1,27
SNORA56	-1,12
SNORA5C	-1,43
SNORA6	-1,29
SNORA62	-1,60
SNORA64	-1,45
SNORA65	-1,72
SNORA66	-1,43
SNORA67	-1,40
SNORA71D	-1,73
SNORA74A	-1,54
SNORA80B	-1,85
SNORA84	-1,01
SNORD15A	-1,89
SNORD17	-1,85
SNX10	1,53
SNX29P2	-1,59
SOBP	-1,34
SOD2	1,56
SORL1	-1,54
ST6GALNAC3	1,66
STEAP3	1,52
STOM	1,41
STX11	1,20
TBC1D9	1,16
TBX21	1,82

Nom de gène	Ratio transcrits
TCEB3C	-1,62
TCEB3CL	-1,53
TERT	-1,62
THEMIS2	1,27
THSD7A	-1,52
TMEM155	1,70
TMEM241	-1,53
TMEM98	-1,54
TMTC4	-1,31
TSPAN32	1,50
TSPAN6	-1,35
VSIG4	1,86
ZBED2	2,34
ZBTB16	1,71
ZC2HC1B	-2,26
ZC3HAV1L	-1,33
ZNF581	-1,10
ZNF608	-1,36
ZNF711	-2,03

Leslie MULLER

Développements de méthodes de préparation d'échantillons pour l'analyse protéomique quantitative : application à la recherche de biomarqueurs de pathologies

Résumé

Les stratégies de protéomique quantitative sans marquage sont très attractives dans le domaine de la recherche de biomarqueurs de pathologies. Cependant, elles requièrent une pleine maîtrise du schéma analytique et de sa répétabilité. Plus particulièrement, la préparation d'échantillons nécessite d'être suffisamment répétable pour ne pas impacter la qualité et la fiabilité des résultats.

Les objectifs de cette thèse étaient de développer et d'optimiser des méthodes analytiques pour la protéomique quantitative, en particulier pour l'étape de préparation d'échantillons. Ainsi, un protocole innovant, simple, rapide et permettant l'analyse quantitative sans marquage d'un grand nombre d'échantillons avec une haute répétabilité a été développé et optimisé : le « *Tube-Gel* ». Par ailleurs, des préparations d'échantillons adaptées à différentes matrices biologiques pour la recherche de biomarqueurs ont été élaborées. Les méthodes mises au point et leur application ont permis de proposer des candidats biomarqueurs pour plusieurs pathologies : le glioblastome, les lymphomes B diffus à grandes cellules, et les complications survenant sur les greffons rénaux.

Mots-clés : Analyse protéomique, Spectrométrie de masse, Quantification sans marquage, Préparation d'échantillons

Résumé en anglais

Label-free quantitative proteomics strategies are very attractive for diseases biomarkers researches. These approaches require the full control and the repeatability of the analytical workflow. In particular, the sample preparation has to be repeatable enough to ensure the quality and reliability of the results.

Objectives of this work were to optimize and develop analytical methods for quantitative proteomics, with a special focus on the sample preparation step. Thus, an innovative, easy and fast protocol allowing the analysis of high sample numbers with high repeatability was developed and further optimized: the "Tube-Gel" protocol. Besides, sample preparations adapted to a variety of biological matrices were developed for the search of biomarkers. The developed methods and their application allowed the identification of potential biomarkers for a variety of diseases: glioblastoma, diffuse large B-cell lymphomas and renal transplants failures.

Keywords: Proteomic analysis, Mass spectrometry, Label-free quantification, Sample preparation