Université de Lille Nord de France
Ecole doctorale Biologie Santé
-
Université de Sherbrooke
Programme de Biochimie

Développement de stratégies protéomiques pour la découverte de nouvelles protéines codées dans des séquences codantes non canoniques chez les eucaryotes

Par
Vivian Delcourt
Programmes de Biologie-Santé (EDBSL) et Biochimie (UdeS)

Thèse présentée en vue de l'obtention du grade de philosophiae doctor (Ph.D.)
en Biologie-Santé (EDBSL) et Biochimie (UdeS)

14 Décembre, 2017

Membres du jury d'évaluation
Pr. Michelle Scott, évaluatrice interne au programme de biochimie et présidente de jury
Pr. François-Michel Boisvert, évaluateur externe au programme de biochimie
Dr. Julia Chamot-Rooke, rapporteur
Dr. Jean Armengaud, rapporteur

Pr. Isabelle Fournier, co-directrice de thèse
Pr. Xavier Roucou, co-directeur de thèse
Dr. Julien Franck, co-encadrant de thèse

Nous ne sommes savants que de la science présente.

Michel Eyquem de Montaigne

REMERCIEMENTS

Au terme de ces trois années pendant lesquelles j'ai beaucoup appris, il m'apparaît essentiel de remercier les personnes sans qui ce travail n'aurait été possible :

Ma directrice de thèse à l'Université Lille 1, Pr. Isabelle Fournier, pour les connaissances qu'elle m'a transmises en chimie analytique et spectrométrie de masse, sa confiance, son encadrement, ses encouragements et la sympathie qu'elle a exprimée à mon égard.

Mon directeur de thèse à l'Université de Sherbrooke, Pr. Xavier Roucou, pour son accueil dans son laboratoire, pour les connaissances qu'il m'a transmises en biochimie et biologie, pour s'être assuré que mon arrivée à Sherbrooke se passe le mieux possible, sa confiance, son encadrement et son soutien.

Le Pr. Michel Salzet, directeur du laboratoire de Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM) à l'Université Lille 1 pour son accueil au sein de son laboratoire et pour avoir tout mis en œuvre pour que je puisse réaliser mon doctorat.

Je vous exprime, à tous les trois, toute ma gratitude pour votre confiance et j'espère avoir été à la hauteur de vos attentes.

Je remercie aussi particulièrement Dr. Julien Franck, qui m'a d'abord encadré en Master-2 puis co-encadrant de mon doctorat à l'Université Lille 1 et grâce à qui j'ai beaucoup appris en spectrométrie de masse et chimie analytique. Son encadrement, son aide, son soutien, ses encouragements, sa disponibilité, sa bienveillance et son amitié m'ont été précieux.

Je tiens également à remercier les membres du jury, Dr. Julia Chamot-Rooke et Dr. Jean Armengaud de me faire l'honneur d'être rapporteurs de mes travaux.

J'exprime également ma reconnaissance envers Pr. Michelle Scott et Pr. François-Michel Boisvert qui ont accepté d'évaluer mes travaux de doctorat.

Je remercie le Dr. Maxence Wisztorski, pour son aide et la sympathie qu'il a toujours témoignée à mon égard.

Je remercie également Jean-François Jacques et Mylène Brunelle, tous deux professionnels de recherche au laboratoire de Xavier Roucou pour leur aide, leur soutien, leurs précieux conseils et le temps qu'ils ont pu m'accorder.

Mes remerciements vont également aux membres permanents du laboratoire PRISM avec qui il aura été un plaisir de travailler dans une ambiance toujours conviviale.

Je remercie également les étudiants et ex-étudiants que j'ai eu la chance de rencontrer au laboratoire PRISM et dans le laboratoire de Xavier Roucou : Marie B., Max B., Max G., Sondos, Annie, Marc-André, Hélène, Marie D., Stéphanie, Dounia, Jusal, Benoit,

Antonella, Adriana, Tanina, Tony, Philipe, Tristan et Flore avec qui j'ai eu la chance de travailler dans une atmosphère agréable. J'exprime également mes remerciements à Jean-François Lucier, Maxime Lévesque et Jules Gagnon, pour leur aide précieuse lors de ma découverte des "joies" de la bioinformatique, des codes et scripts sous linux/R/etc. Dans un même registre, je remercie Julien Clerk-Lamalice pour son aide précieuse avec LaTeXet dans l'édition du code de son *template*.

D'un point de vue personnel, je tiens à remercier ma mère pour avoir assuré notre éducation et s'être assurée que nous ne manquions de rien malgré le décès de notre père, parti beaucoup trop vite... Je remercie aussi particulièrement ma sœur, mon frère et plus largement l'ensemble de ma famille pour leur soutien inébranlable et leur affection à toute épreuve.

Je remercie également ma belle famille pour leurs encouragements et l'attention qu'ils m'ont toujours témoignée.

Je remercie aussi mes amis qui ont toujours été présents malgré la distance. Aussi je remercie particulièrement Clem D, Clem L, Julien, Romain et Aurélien, amis de (très) longue date pour leur amitié inconditionnelle.

Enfin, j'exprime mon immense gratitude envers Laure (*alias* Doudou) qui partage ma vie depuis maintenant plus de cinq ans, d'avoir accepté de quitter famille et amis pour vivre à l'étranger avec moi, pour ta présence dans les bons moments mais aussi et surtout lors des périodes plus difficiles, pour ton soutien et tes encouragements constants, pour tes attentions quotidiennes, ton souci de mon bien-être parfois au détriment du tien, pour ta compréhension et ton affection indéfectible.

Du fond du cœur, merci.

RÉALISATIONS SCIENTIFIQUES

Publications

Publiées :

Mouilleron, H.*, **Delcourt, V.***, & Roucou, X. (2015). Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic acids research*, 44(1), 14-23.

Delcourt, V.*, Franck, J.*, Leblanc, E., Narducci, F., Robin, Y. M., Gimeno, J. P., Quanico J., Wisztorski M., Kobeissy F., Jacques J. F., Roucou X., Salzet M. & Fournier I. (2017). Combined Mass Spectrometry Imaging and Top-down Microproteomics reveals evidence of a hidden proteome in ovarian cancer. *EBioMedicine*.

Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., & Roucou, X. (2017). Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. *Proteomics*.

En révision :

Delcourt V.*, Franck J.*, Quanico J, Gimeno J. P., Wisztorski M., Raffo-Romero A., Kobeissy F., Roucou X., Salzet M. & Fournier I. Top-down microproteomics bridged to MALDI MS imaging reveals the molecular physiome of brain regions. *Molecular and cellular proteomics*

Samandi S.*, Roy A. V.*, **Delcourt V.**, Lucier J. F., Gagnon J., Beaudoin M. C., Vanderperre B., Breton M. A., Motard J., Jacques J. F., Brunelle M., Gagnon-Arsenault I., Fournier I., Ouangraoua A., Hunting D. A., Cohen A. A., Landry C. R., Scott M. S. & Roucou X. Deep transcriptome annotation suggests that small and large proteins encoded in the same genes often cooperate. *eLife*

En préparation :

Delcourt V., Brunelle M., Roy A. V., Jacques J. F., Salzet M., Fournier I. & Roucou X. A protein translated from a short ORF originally excluded from gene annotations is the main translation product of *MIEF1*.
A soumettre

Non canonical ORF encoded peptides are source of novel cancer markers.
A rédiger

* : co-premier auteur

Communications orales

Parallel reaction monitoring deciphers the stoichiometry of two distinct proteins coded by the human dual-coding gene *MIEF1*. *Canadian National Proteomics Network 2017*. Toronto, Ontario, Canada

Expression endogène et stœchiométrie de deux protéines humaines différentes codées dans un même gène. *Symposium annuel de biochimie de l'Université de Sherbrooke*. Sherbrooke, Québec, Canada

Recent developments for on tissue proteomics analysis. *Rencontres du Club Jeunes de la Société Française de Spectrométrie de Masse (RCJSM) 2015*. Montélimar, Drôme, France

Posters

Combination of MALDI-MS Imaging and top-down microproteomics allows identification of specific proteoforms, potential biomarkers and uncovers hidden proteins. Journée André Verbert 2017. Lille, France.

A novel proteomic approach with a database of non-canonical protein-coding sequences reveals the contribution of the short proteome. *Keystone, Omics strategies to study the proteome 2017*. Breckenridge, Colorado, Etats-Unis d'Amérique

A novel proteomic approach with a database of non-canonical protein-coding sequences reveals the contribution of the short proteome. *Symposium PROTEOMEUS 2017*. Sherbrooke, Québec, Canada

Method for clinical application of MALDI Imaging driven top-down microproteomics for discovery of tumor biomarkers. International Mass Spectrometry Conference (IMSC-2016). Toronto, Ontario, Canada

Deep functional proteomic identification of alternative proteins gives insight into their functions. *Symposium annuel PROTEO (PROTEO-2016)*. Québec, Québec, Canada

Deep functional proteomic identification of alternative proteins gives insight into their functions. *Symposium PROTEOMEUS 2016*. Sherbrooke, Québec, Canada

Top-Down Protein Identification from Tissue Sections by Microproteomics Approach. *Amercian Society for Mass Spectrometry (ASMS-63) 2015*. St Louis, Missouri, Etats-Unis d'Amérique

RÉSUMÉ

Développement de stratégies protéomiques pour la découverte de nouvelles protéines codées dans des séquences codantes non canoniques chez les eucaryotes

Par
Vivian Delcourt
Programmes de Biologie-Santé (EDBSL) et Biochimie (UdeS)

Thèse présentée en vue de l'obtention du diplôme de philosophiae doctor (Ph.D.) en Biologie-Santé (EDBSL) et Biochimie (UdeS)

La vision traditionnelle de la synthèse protéique chez les eucaryotes comprend un ARN messager (ARNm) qui porte un seul cadre de lecture ouvert (ORF), aussi appelé séquence codante (CDS). Chaque gène codant eucaryote produit généralement une protéine canonique et éventuellement une ou plusieurs isoformes. L'ensemble de ces protéines constitue le protéome.

Cependant, de nombreuses évidences expérimentales récentes démontrent que le protéome des eucaryotes a été sous-estimé, et que les cellules sont capables de synthétiser des protéines qui n'étaient jusqu'alors pas prédites. Ces protéines nommées « alternatives » (altProts) peuvent être issues de la traduction d'ORFs non annotés ou alternatifs contenus sur des ARNms, ou des ARNs annotés comme non codants (ARNnc). Les altProts ne sont donc pas des isoformes des protéines canoniques mais des protéines avec des séquences complètement nouvelles.

Ces découvertes ont notamment été possibles grâce aux progrès techniques réalisés en biochimie analytique et particulièrement dans les approches d'analyses protéomiques basées sur la spectrométrie de masse qui permettent l'identification à grande échelle des protéines. Dans le cadre de ces analyses, deux approches sont privilégiées. La première, ou *bottom-up* se base sur les produits peptidiques issus d'une digestion enzymatique des protéines intactes quand la seconde ou *top-down*, développée plus récemment, est basée sur la mesure des protéines entières par spectrométrie de masse. Les protéines ou produits peptidiques sont alors fréquemment séparés par une étape de chromatographie liquide couplée à un spectromètre de masse. Le spectromètre de masse génère alors des spectres « MS » et « MS/MS » de ces analytes. Sur la base des informations récoltées dans les spectres MS et MS/MS, il est alors possible d'identifier la protéine en comparant ces spectres aux valeurs théoriques contenues dans une base de données de séquences protéiques. La découverte récente des altProts résulte précisément de ce dernier point. Jusqu'à présent, seulement les protéines prédites par l'annotation des génomes étaient présentes dans ces bases de don-

nées. Ce n'est que récemment que les séquences des altProts ont été prédites, permettant ainsi leur identification.

Les travaux réalisés dans cette thèse s'articulent autour du développement de stratégies aidant à la découverte et la caractérisation des altProts par approches protéomiques bottom-up et top-down. Ces aspects sont décrits dans plusieurs publications scientifiques qui seront présentées dans ce manuscrit. Elles comprennent une revue de bibliographie, deux publications relatives à l'application de l'approche top-down par micro-extractions de tissus de cerveau de rat et de biopsie tumorale ovarienne et une publication relative à la détermination de la stœchiométrie de deux protéines, l'une alternative et l'autre canonique toutes deux issues du même gène.

Mots-clés: Protéomique, Protéines alternatives, Spectrométrie de masse

SUMMARY

Development of proteomics strategies for the discovery of novel proteins encoded within non-canonical open reading frames in eukaryotic species

By
Vivian Delcourt
Program: Biology & Health (EDBSL) and Biochemistry (UdeS)

Thesis presented for the obtention of Doctor degree diploma in Biology & Health (EDBSL) and Biochemistry (UdeS)

The traditional view of protein synthesis in eukaryotic species involves one messenger RNA (mRNA) bearing a single open reading frame (ORF), also termed coding sequence (CDS). Thus, each eukaryotic coding gene may produce one canonical protein and possibly one or more of its isoforms. The whole pool of these proteins is termed the proteome.

However, numerous experimental evidence report that eukaryotic proteomes may have been under-estimated and that cells are capable of synthetizing proteins which had not been predicted thus far. These proteins, termed "alternative proteins" (altProts) may be translated from non-canonical or alternative ORFs localized in mRNAs or from RNAs annotated as non-coding (ncRNA). These altProts are not isoforms of canonical proteins but entirely new protein sequences.

These discoveries were made possible thanks to technical progresses in analytical chemistry and particularly in mass spectrometry-based proteomics analyses, which allow large-scale identification of proteins. These analyses are based on two main strategies; the "bottom-up" approach is based on the peptidic products of enzymatic digestion of native proteins whereas the second and more recent approach, termed "top-down", is based on the analysis of intact protein by mass spectrometry. Proteins or peptidic products are often separated by a liquid chromatography coupled to a mass spectrometer. Mass spectrometer records "MS" and "MS/MS" spectra of these analytes. Based on the information found in MS and MS/MS spectra, it is then possible to identify the protein from which the analyte is derived by comparing these spectra to theoretical spectra derived from a protein sequence database. Until recently, only proteins included in genome annotation were reported in these databases. It is only recently that altProts have been predicted, allowing their identification.

The work described in this thesis is focused on the development of experimental strategies helping the discovery and characterization of altProts using bottom-up and top-down approaches. The findings are described in scientific publications which are included in the

thesis. These publications include a review, two publications on the application of the top-down approach using micro-extractions on rat brain tissue and ovarian tumor biopsy and one publication related to the stoichiometry elucidation of a canonical and an alternative protein both encoded within the same gene.

Keywords: Proteomics, Alternative proteins, Mass spectrometry

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

Abréviation	Description
aa (ou AA)	Acide aminé (ou *Amino acid*)
ADN (ou DNA)	Acide désoxyribonucléique (ou *Desoxyribonucleic acid*)
ADNc (ou cDNA)	ADN complémentaire (ou *Complementary DNA*)
AGC	Contrôle automatique du gain (ou *Automatic gain control*)
altORF	Cadre de lecture ouvert alternatif (ou *Alternative open reading frame*)
altProt	Protéine alternative (ou *Alternative protein*)
AQUA	Quantification absolue par peptides synthétiques marqués par des isotopes stables (ou *Stable isotope labelled synthetic peptides for absolute quantification*)
ARN (ou RNA)	Acide ribonucléique (ou *Ribonucleic acid*)
ARNm (ou mRNA)	ARN messager (ou *messenger RNA*)
ARNnc (ou ncRNA)	ARN non-codant (ou *non-coding RNA*)
CDS (ou refORF)	Séquence d'ADN codante canonique (ou *Canonical coding DNA sequence*)
CRISPR	Courtes répétitions palindromiques groupées et régulièrement espacées (ou *Clustered Regularly Interspaced Short Palindromic Repeats*)
CV	Coefficient de variation (ou *Coefficient of variation*)
CyPrP	Protéine prion cytoplasmique (ou *Cytoplasmic prion protein*)
Da	Dalton (unité de mesure)
DDA	Acquisition data-dépendante (ou *Data dependant acquisition*)
DIA (ou SWATH)	Acquisition data-indépendante (ou *Data independant acquisition*)
ESI	Ionisation par électronébuliseur (ou *Electrospray ionization*)
EST	Séquence motif d'expression (ou *Expressed sequence tag*)
FASP	Préparation d'échantillon assistée par filtre (ou *Filter-aided sample preparation*)
FDR	Taux de faux positifs (ou *False discovery rate*)
GFP	Protéine fluorescente verte (ou *Green fluorescent protein*)

GRCh	Consortium de référence du génome humain (ou *Genome Reference Consortium Human*)
HCV	Virus de l'hépatite C (ou *Hepatitis C virus*)
IB (ou WB)	Immunobuvardage (ou *Western blot*)
IF	Immunofluorescence
ISH	Hybridation *In situ* (ou *In situ hybridization*)
IRES	Site interne d'entrée du ribosome (ou *Internal ribosome entry site*)
KO	Inactivation génique (ou *Knock-out (gene inactivation)*)
LAP	Etiquette de localisation et de purification par affinité (ou *Localization and affinity purification tag*)
LMJ	Micro jonction liquide (ou *Liquid micro junction*)
lncRNA	Long ARN non-codant (ou *Long non-coding RNA*)
MALDI	Désorption-ionisation laser assistée par matrice (ou *Matrix assisted laser desorption-ionization*)
miRNA	Micro ARN
MRM	Suivi de réactions multiples (ou *Multiple reaction monitoring*)
MS	Spectrométrie de masse (ou *Mass spectrometry*)
MS/MS (ou MS^2)	Spectrométrie de masse en tandem (ou *Tandem mass spectrometry*)
MSI	Imagerie par spectrométrie de masse (ou *Mass spectrometry imaging*)
m/z	Rapport masse/charge (ou *Mass/charge ratio*)
nanoLC	Nano chromatographie liquide (ou *Nano liquid chromatography*)
NCE	Energie de collision normalisée (ou *Normalized collision energy*)
ORF	Cadre de lecture ouvert (ou *Open reading frame*)
PAM	Microdissection manuelle assistée par parafilm (ou *Parafilm assisted manual microdissection*)
PCR	Réactions en chaîne par polymerase (ou *Polymerase chain reaction*)
PRM	Suivi des réactions en parallèle (ou *Parallel reaction monitoring*)
PrP	Protéine prion (ou *Pion protein*)
PSM	Correspondance peptide-spectre (ou *Peptide-spectrum match*)
PTM	Modification post-traductionnelle (ou *Post-translational modification*)
refORF	Cadre de lecture de référence (ou canonique) (ou *Reference open reading frame*)

RNApolII	ARN polymérase II (ou *RNA polymerase II*)
RNA-seq	Séquençage des ARN à haut débit (ou *High throughput RNA sequencing*)
ROI	Région d'intérêt (ou *Region of interest*)
SEP	Peptide encodé à partir d'un court cadre de lecture ouvert (ou *Short open reading frame encoded peptide*)
shRNA	Petits ARN en épingle à cheveux (ou *Short hairpin RNA*)
siRNA	Petit ARN interférent (ou *Small interfering RNA*)
SILAC	Culture cellulaire avec incorporation d'acides aminés marqués par des isotopes stables (ou *Stable isotope labeling by amino acids in cell culture*)
smORF	Petit cadre de lecture ouvert (ou *Small open reading frame*)
sORF	Court cadre de lecture ouvert (ou *Short open reading frame*)
SDS	Sodium dodecyl sulfate
SDS-PAGE	Electrophorèse en gel de polyacrylamide avec SDS (ou *Sodium dodecyl sulfate polyacrylamide gel electrophoresis*)
SRM	Suivi de réaction sélectionnée (ou *Selected reaction monitoring*)
TOF	Temps de vol (ou *Time of flight*)
uORF	Cadre de lecture ouvert en amont du CDS (ou *Upstream open reading frame*)
UTR	Région non traduite (ou *Untranslated region*)
WT	Sauvage ; retrouvé naturellement (ou *Wild type*)

1 INTRODUCTION GÉNÉRALE

L'annotation systématique des génomes d'eucaryotes après séquençage permet de définir et d'annoter les gènes et les régions intergéniques. Cette annotation est une étape essentielle dans la biologie moderne car elle permet de structurer et de faciliter l'accès à l'information génétique. Les gènes appelés "codants" sont *in fine* traduits en protéines qui remplissent un vaste pan des activités biologiques nécessaires à la vie de la cellule. La vision canonique des gènes codants chez les eucaryotes, une fois transcrits en ARN messagers (ARNm), suppose qu'ils ne portent qu'un seul cadre de lecture ouvert (ORF) aussi appelé séquence codante (CDS). Ils ne coderaient donc que pour une protéine ou des isoformes de cette protéine par le processus d'épissage alternatif qui produit des isoformes d'ARNm. Par souci de clarté, cette protéine canonique sera appelée protéine de référence dans la suite de la thèse.

L'annotation des gènes est réalisée selon les règles préétablies suivantes :

1. Les gènes sont généralement conservés à travers l'évolution.

2. Ils contiennent une séquence d'une taille minimale de 100 codons qui codera pour une protéine.

3. Les gènes qui sont transcrits mais qui présentent un CDS d'une taille inférieure à 100 codons sont automatiquement annotés comme non-codants, et sont donc considérés comme ne produisant pas de protéines.

4. Un gène code pour une unique protéine de référence ou ses isoformes, Ces isoformes sont produites par épissage alternatif, l'utilisation de promoteurs alternatifs, ou bien par édition de l'ARN.

5. Dans un même transcrit considéré comme codant, un seul ORF est effectivement traduit en protéine par le ribosome.

De façon générale, ces règles ont pour conséquence que la séquence codante unique annotée correspond à l'ORF le plus long. Cependant, de nombreuses évidences expérimentales rapportent l'expression de protéines codées dans des ORFs alternatifs, appelées protéines alternatives. En effet, chaque transcrit, qu'il soit annoté comme codant ou non, peut contenir plusieurs ORFs. Ces ORFs peuvent être localisés dans les régions annotées comme non traduites (5' et 3' UTR), dans un cadre de lecture décalé au sein de la séquence codante

de la protéine canonique, et dans des transcrits annotés comme non-codants. L'expression des ORFs alternatifs a été mise en évidence grâce au développement des techniques de profilage ribosomal qui séquencent les portions de transcrits protégés par les ribosomes, et des techniques de protéomique par spectrométrie de masse qui identifient les protéines par détection de leurs fragments peptidiques (approche *bottom-up*) ou de leurs séquences intactes (approche *top-down*).

Les stratégies de protéomique présentent l'avantage majeur de détecter les protéines synthétisées, donc suffisamment stables pour avoir une activité biologique spécifique. L'approche *bottom-up*, qui est la plus souvent employée, identifie les protéines sur la base de leur fragments peptidiques issus d'une digestion protéolytique. Ceci peut constituer un frein à la détection des protéines alternatives qui sont dans leur grande majorité des petites protéines de taille inférieure à 100 acides aminés. Elles sont donc moins susceptibles de générer des peptides suite à une digestion enzymatique. L'approche *top-down*, quant à elle, semble particulièrement performante pour identifier les protéines de petite taille puisqu'elle permet l'identification de protéines intactes. De plus, l'approche *top-down* permet d'obtenir des informations quant aux modifications post traductionnelles que présentent les protéines sans nécessiter de préparation d'échantillon spécifique requises en *bottom-up*.

Hypothèses

Les approches protéomiques par spectrométrie de masse peuvent aider la détection et la caractérisation des protéines alternatives. Toutefois, les approches actuelles en protéomique ne semblent pas optimales pour leur étude. Est-il possible de développer des approches protéomiques plus adaptées à l'étude des protéines alternatives afin d'en évaluer l'expression ?

Objectif 1

L'étude des protéines alternatives est une discipline nouvelle en pleine expansion. De nombreuses études font état de leur expression et décrivent parfois leur fonction, que ce soit à travers des études centrées sur une protéine alternative ou à large échelle. Parallèlement à ce nouveau domaine, l'étude de petites protéines de taille inférieure à 100 acides aminés principalement codées dans des ARN annotés comme non-codant s'est développé (Matsumoto *et al.*, 2017; D'Lima *et al.*, 2017). Les protéines alternatives incluent ces petites protéines mais également toutes les autres protéines non annotées qui sont codées par des altORFs localisés dans les ARNms. Il paraît donc important d'établir un état des

lieux des avancées réalisées pour la caractérisation des petites protéines alternatives issues d'ARNms ou ARNnc, des mécanismes expliquant leur expression, des changements que leur découverte implique dans l'annotation des gènes et enfin des défis à relever pour leur étude.

Objectif 2

La caractérisation de petites protéines représente un défi important en protéomique dite « *bottom-up* », par leur faible capacité à générer des peptides protéolytiques qui sont requis pour leur identification. La protéomique dite « *top-down* » permet de détecter des protéines de faible poids moléculaire ou des produits de clivage de larges protéines, ce qui étend le champ du protéome accessible par les approches de protéomique par spectrométrie de masse. De plus, les stratégies d'analyse de régions de tissus par microprotéomique offrent une étude du protéome plus fidèle que les modèles de lignées cellulaires habituellement employés, notamment grâce à l'analyse du microenvironnement cellulaire. Il apparaît important de tester cette approche pour l'identification des protéines alternatives et de petites protéines de référence au sein de tissus biologiques.

Objectif 3

Les biomarqueurs sont des molécules dont la mesure est associée à un état physiologique ou pathologique et sont notamment employés pour le dépistage, le diagnostic et le pronostic de diverses affections biologique. Les protéines et leurs modifications peuvent constituer une signature spécifique d'un trouble de santé ou de la réponse à un traitement. Toutefois, la découverte de biomarqueurs protéiques est principalement associée à l'approche *bottom-up*. L'emploi de la protéomique par *top-down* pour l'évaluation et la découverte de biomarqueurs reste rare malgré ses performances pour la caractérisation de protéines intactes et ses capacités d'observations de modifications post traductionnelles. Dans ce contexte, l'association de stratégies spécifiques pour l'étude de tissus biologiques telles que l'imagerie par spectrométrie de masse, la microprotéomique et l'approche *top-down* pourrait s'avérer être efficace pour la découverte de biomarqueurs protéiques de référence et alternatifs. Dans cet objectif, nous emploierons la méthode développée lors de l'objectif 2 pour identifier des protéines de référence et alternatives au sein de régions saines, tumorales et nécrotiques d'un tissu. Par la détection spécifique de protéines au sein des différentes régions, nous pourrons établir la preuve de concept que les protéines alternatives, tout comme les protéines de référence, pourront être source de nouveaux biomarqueurs.

Objectif 4

Par la remise en cause du dogme « un gène – une protéine », l'annotation fonctionnelle des gènes « multicodants » doit être modifiée en conséquence. En effet, certaines informations peuvent se révéler pertinentes dans l'étude de phénotypes associés à ces gènes comme l'identité et la localisation des séquences codantes du gène, mais également les niveaux d'expression des protéines du gène et leur stœchiométrie. Pour cet objectif, les expériences sont focalisées sur le gène *MIEF1*. Ce gène modèle code pour la protéine canonique MiD51 et la protéine alternative altMiD51, identifiée dans Vanderperre *et al.* (2013), détectée au sein du laboratoire et également été observée après analyse d'empreintes ribosomales. L'objectif sera de déterminer les quantités absolues des protéines de ce gène au sein de divers échantillons biologiques, ce qui mènera à la détermination de leur stœchiométrie.

2 BIBLIOGRAPHIE

2.1 Du génome à la protéine

Le génome représente l'ensemble des séquences d'acides désoxyribonucléiques (ADN) nucléaires d'un organisme eucaryote. Il contient l'intégralité des informations nécessaires à la vie, dès les premières étapes de l'embryogenèse jusqu'à la mort de l'organisme. Il est composé de gènes, regroupant introns et exons, et de régions intergéniques (Figure 1).

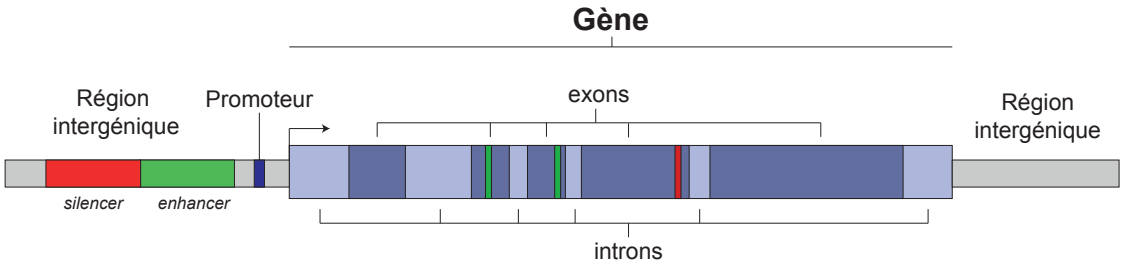


FIGURE 2.1 – **Structure générale d'un gène**
Les séquences correspondant à des codons d'initiation et de terminaison de la traduction canonique du gène sont respectivement représentées en vert et rouge.

2.1.1 Séquençage et annotation du génome

Avec l'avènement des méthodes de séquençage à haut débit, notamment grâce à la découverte de la *Thermophilus aquaticus* polymérase (Chien *et al.*, 1976; Saiki *et al.*, 1988), une enzyme capable de copier de façon fidèle de l'ADN à une vitesse de l'ordre du millier de base par minute, plus de 80% du génome humain ont pu être séquencés pour la première fois en 2001 (Venter *et al.*, 2001; Lander *et al.*, 2001). Ces études ont permis d'estimer que 75% du génome correspondent à des régions intergéniques et 25% à des gènes, eux-mêmes divisés en exons (1.1%) et introns (24%). Depuis, de nombreuses versions du génome ont été confirmées et affinées.

Une fois séquencé et assemblé, le génome doit être annoté. Pour ce faire, il est nécessaire de comparer les séquences observées expérimentalement avec ce qui est déjà connu. Plusieurs outils sont utilisés pour l'annotation. En général, les séquences obtenues sont comparées à d'autres séquences obtenues comme des ADN complémentaires (ADNc ou cDNA) - copies complémentaires des acides ribonucléiques messagers (ARNm) obtenues *via* transcription inverse -, des données de séquençage d'ARN (RNA-seq), des protéines

déjà caractérisées ou encore de séquences motifs d'expression (EST) de l'organisme à annoter (Yandell et Ence, 2012). Les connaissances déjà acquises pour d'autres organismes sont également informatives. En effet, étant donné qu'un grand nombre de gènes sont conservés à travers de nombreuses espèces, les annotations référencées pour un organisme peuvent aider à l'annotation du génome d'autres organismes au sein du même groupe ou règne phylogénétique, voire même pour des règnes éloignés. Du fait de l'évolution, la reconnaissance de séquences au sein d'organismes différents est un outil puissant qui comporte toutefois certaines limites.

Il est également possible de prédire les gènes *ab initio*, en absence d'évidence expérimentale, mais cela nécessite des procédures particulières telles que l'entraînement d'outils statistiques sur le génome à annoter pour affiner la prédiction. Ces outils sont en général moins précis, notamment en ce qui concerne les prédictions de jonctions introns-exons (Yandell et Ence, 2012).

Lors de ces processus d'annotation, certaines étapes peuvent aussi limiter la découverte de nouveaux gènes. Il est fréquent que certains gènes soient automatiquement classés comme "non-codants" du fait de l'absence de séquence codante d'une longueur arbitraire minimale de 300 nucléotides (ou 100 codons) (Goffeau *et al.*, 1996; Basrai *et al.*, 1997). Ainsi, si ladite séquence n'est pas reconnue pour ses similitudes avec d'autres gènes/transcrits ou protéines, elle est automatiquement rejetée des séquences potentiellement codantes du génome. De plus, comme plusieurs étapes de l'annotation reposent sur les connaissances déjà référencées dans diverses bases de données génomiques, transcriptomiques et protéiques (Pruitt et Maglott, 2001; Kent *et al.*, 2002; Flicek *et al.*, 2011; UniProt-Consortium *et al.*, 2008), il est fréquent que de nombreuses séquences ne soient pas reconnues, limitant ainsi la découverte de nouvelles séquences codantes.

Les gènes, malgré leur importance, sont passifs dans la vie cellulaire. Pour remplir leur fonction, ils doivent être copiés en transcrits ARN qui sont les intermédiaires cellulaires entre un gène et le rôle biologique qui lui est associé.

2.1.2 Transcription de l'ADN en ARN

La transcription est le mécanisme cellulaire finement régulé permettant la synthèse d'ARN à partir d'ADN. L'ensemble des ARNs ainsi formés sont des éléments-clés de l'expression des gènes et constituent le transcriptome. Le transcriptome est composé de plusieurs classes d'ARNs qui sont transcrits par trois ARN polymérases. Les ARN messagers contiennent des séquences codantes qui seront traduites en protéines, et sont transcrits par l'ARN polymérase II (ARNpolII). Les ARN non codants, également transcrits par l'ARN-

polII, regroupent les longs ARNs non codants (lncRNA) qui interviennent dans la régulation de l'expression de gènes et qui génèrent parfois de nouvelles protéines (Mercer *et al.*, 2009), les micro ARN (miRNA) impliqués dans la dégradation des ARN et la régulation traductionnelle (He et Hannon, 2004), et les petits ARN nucléolaires nécessaires à la biogénèse du ribosome (Dupuis-Sandoval *et al.*, 2015). Les ARN de transfert (ARNt), transcrits par l'ARN polymérase 3, interviennent quant à eux dans le transport d'acides aminés et sont nécessaires à la synthèse protéique. Les ARN ribosomiques, transcrits par l'ARN polymérase 1, sont des composants structurels des ribosomes essentiels pour la synthèse protéique.

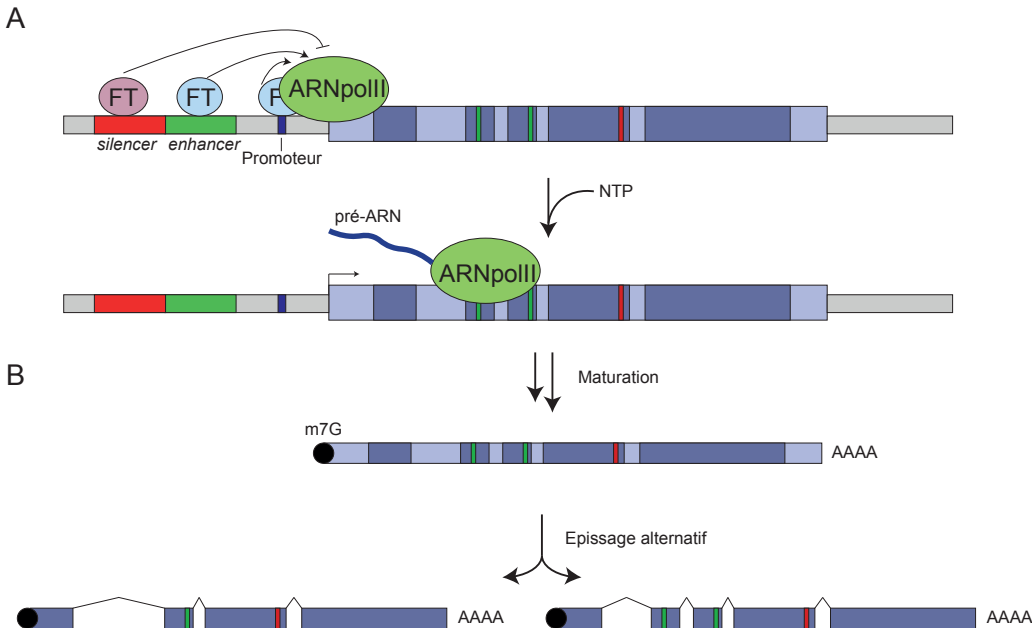


FIGURE 2.2 – **Mécanismes de transcription et maturation-épissage**
A. Transcription de l'ADN en ARN par l'ARNpolII. **B.** Maturation et épissage alternatif du pré-ARN en ARN mature.
FT : facteur de transcription, m7G : 7-méthylguanosine, NTP : nucléotides tri-phosphate

Sous l'action d'un *stimulus* ou de manière constitutive, l'ADN qui est empaqueté sous forme de chromatine se décondense au niveau des régions proches du gène à transcrire. En amont du gène se trouvent des séquences d'ADN régulatrices appelées *silencer*, *enhancers* et promoteurs qui régulent ou activent respectivement la transcription des gènes. Ces séquences sont spécifiquement reconnues par des facteurs de transcription dont le rôle est notamment de réprimer ou d'activer la transcription en agissant sur l'ARNpolII. Dans la situation d'activation de la transcription, le facteur de transcription recrute les différentes protéines nécessaires à l'initiation de la transcription au site promoteur pour former avec l'ARNpolII un complexe multimérique ADN-protéines. Une fois ce complexe assemblé,

l'ARNpolII lit le brin d'ADN du gène pour en former une copie (Nikolov et Burley, 1997; Vaquerizas *et al.*, 2009) (Figure 2 A).

Un pré-ARN est produit une fois la transcription achevée. Ce transcrit n'est toutefois pas fonctionnel. Tout d'abord, une coiffe est ajoutée à l'extrémité 5' du pré-ARN et remplit des fonctions de protection et d'intermédiaire dans l'initiation de la traduction. Le pré-ARN est également polyadénylé à son extrémité 3' ce qui lui confère davantage de stabilité. Cette queue polyA intervient également lors de la traduction, notamment pour circulariser le transcrit. Enfin, les introns et éventuellement certains exons sont éliminés lors de l'épissage alternatif par l'intermédiaire du spliceosome (Figure 2 B). Bien qu'ils soient fréquemment représentés de manière séquentielle, ces mécanismes peuvent se dérouler co-transcriptionellement *in-vivo* (Bentley, 2014).

2.1.3 Traduction de l'ARN en protéine

La traduction est l'étape finale de l'expression des gènes codants. Lors de ce processus, l'intégralité d'une séquence codante est traduite en protéine selon le code génétique. La traduction comprend trois étapes-clés : l'initiation, l'élongation et la terminaison.

D'un point de vue canonique, la traduction concerne les transcrits ARNm, provenant des gènes dits "codants". Ces transcrits sont divisés en trois régions différentes : la région non traduite en 5' (5'UTR), la séquence codante (CDS) et la région non traduite en 3' (3'UTR) (Figure 3).

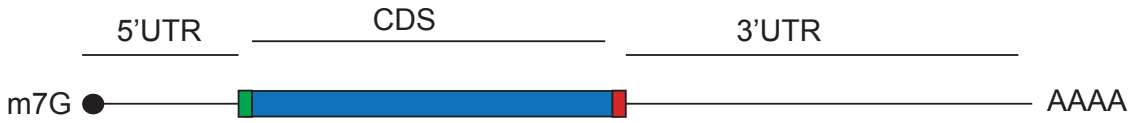


FIGURE 2.3 – **Représentation canonique d'un transcrit ARNm**
m7G : 7-méthylguanosine

2.1.3.1 La traduction coiffe-dépendante et modèle de "scanning" du ribosome

Initiation de la traduction

Lors de l'initiation de la traduction, le facteur d'initiation eIF2 capte un ARN de transfert de méthionine, premier acide aminé incorporé lors de la traduction, et se fixe sur la sous unité 40S du ribosome pour former avec d'autres eIFs le complexe de préinitiation 43S (Figure 4 A). Parallèlement, d'autres facteurs d'initiation de la famille eIF4 se fixent à la coiffe en 5' de l'ARNm et le circularise par la queue polyA (Figure 4 A et B). Le complexe 43S est alors recruté en région 5' par les eIF4. La petite sous-unité lit l'ARNm dans le sens

5' vers 3' jusqu'à rencontrer un codon d'initiation "AUG" (Figure 4 C). Il est possible que le ribosome initie à d'autres codons d'initiation (Ingolia *et al.*, 2011), mais cette initiation non canonique s'effectue en proportion moins importante. Aussi, des séquences voisines influent sur l'efficacité de l'initiation de la traduction et la séquence considérée comme optimale pour l'initiation est GCCACC**AUG**G (Kozak, 2002). Les facteurs d'initiation sont alors libérés pour être remplacés par la sous-unité 60S du ribosome et ainsi former un ribosome complet 80S capable de traduire l'ARNm en protéine (Jackson *et al.*, 2010) (Figure 4 D).

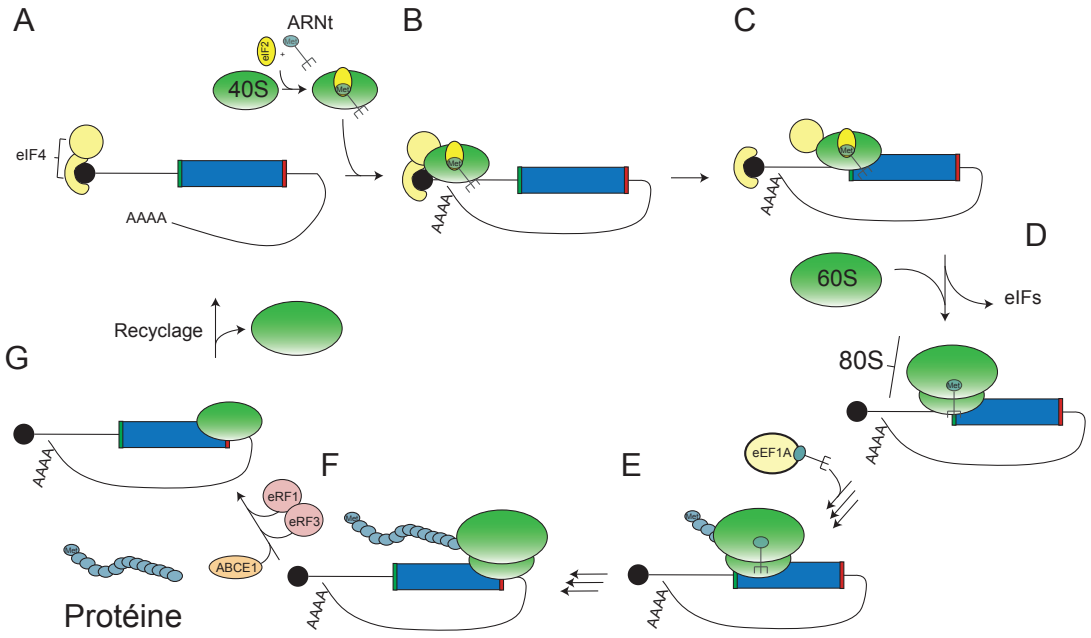


FIGURE 2.4 – **Traduction d'un ARNm en protéine**
Inspiré de (Pisarev *et al.*, 2010; Villa et Fraser, 2014; Wu *et al.*, 2016)

Elongation

Lors de l'élongation, le ribosome lit l'intégralité de la séquence de l'ARNm par groupe de 3 nucléotides, appelés codons. Le ribosome lie les acides aminés à la séquence protéique naissante grâce à la protéine eEF1A dont le rôle est de recruter les ARNs de transfert. L'élongation est un processus relativement rapide, de l'ordre de 4,7 acides aminés par seconde (Wu *et al.*, 2016), capable de synthétiser des peptides de quelques acides aminés à plusieurs milliers d'acides aminés (Figure 4 E).

Terminaison

La traduction s'achève par la lecture d'un codon STOP "UAA/UGA/UAG" par le ribosome (Figure 4 F). La reconnaissance du codon STOP par la protéine eRF1 entraîne le recrutement de la protéine eRF3 qui libère la protéine nouvellement synthétisée. Enfin, la protéine ABCE1 dissocie le ribosome en deux sous-unités 40S et 60S (Pisarev *et al.*, 2010; Villa et Fraser, 2014) (Figure 4 F). Le ribosome peut alors être recyclé et réassemblé sur un autre transcrit ou initier de nouveau la traduction au sein du même transcrit (Figure 4 G).

2.1.3.2 La traduction coiffe-indépendante

Le mécanisme de traduction peut se produire indépendamment de la coiffe. L'initiation de la traduction repose alors essentiellement sur la présence de sites internes d'entrée du ribosome (IRES), de séquences polyU, de séquences complémentaires du brin 18S de l'ARN ribosomal dans les régions 5' et 3' UTR (Weingarten-Gabbay *et al.*, 2016) ou d'adénosines méthylées en position 6 (m6A) en 5'UTR (Meyer *et al.*, 2015) (Figure 5). Ces séquences sont capables de mobiliser les protéines nécessaires au recrutement de la petite sous-unité du ribosome qui débutera la lecture du transcrit et initiera la traduction comme décrit précédemment.

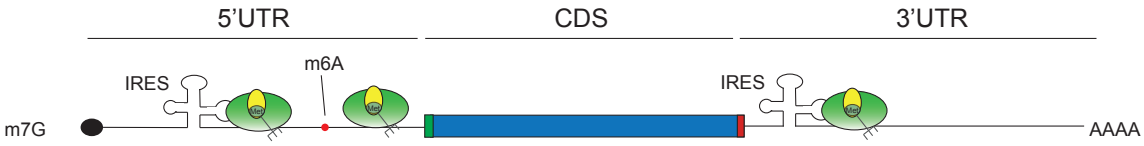


FIGURE 2.5 – **Traduction coiffe-indépendante**
IRES : site d'entrée interne du ribosome, m7G : 7-méthylguanosine, m6A : 6-méthyladénosine

Une fois synthétisées, l'ensemble des protéines, appelé protéome, est responsable de la majorité des fonctions biologiques et est étudié par des approches appelées protéomiques. Dans ces approches, on s'intéresse à l'étude de leur expression au sein de tissus ou cellules, à leur environnement et partenaire(s) d'interaction, leur(s) modification(s) post-traductionnelle(s) (PTMs), leur potentiel de biomarqueur ou encore leur structure tridimensionnelle. L'étude des protéines a longtemps consisté en l'application de techniques biochimiques ou biophysiques encore employées. Toutefois, l'émergence de la spectrométrie de masse (MS) pour l'analyse des protéines a récemment permis des avancées considérables par sa polyvalence et sa capacité à appréhender des systèmes et processus biologiques complexes.

2.2 L'étude des protéines par spectrométrie de masse

2.2.1 Généralités sur la spectrométrie de masse

La spectrométrie de masse (MS) est une technique analytique qui consiste à séparer les molécules chargées (ou ions) en fonction de leur rapport masse/charge. Le spectromètre de masse est composé de trois parties essentielles : la source d'ions qui génère les ions moléculaires, l'analyseur qui les sépare en fonction de leur rapport masse/charge (m/z) et le détecteur qui convertit l'impact ou l'oscillation des ions en un signal proportionnel à la quantité de l'analyte dans le mélange étudié.

Lors de la première moitié du XX^{ème} siècle, la MS fut utilisée pour des applications physicochimiques telles que la détermination d'isotopes et l'étude des atomes. Ce n'est que qu'à partir des années 1950 que les travaux de McLafferty, Biemann et Djerassi ont permis de mettre en évidence les applications de la MS pour l'étude de la structure de composés organiques de faible poids moléculaire (Griffiths, 2008).

Par la suite, la mise au point des sources de désorption-ionisation assistée par matrice (MALDI Karas et Hillenkamp (1988)) et *electrospray* (ESI, Whitehouse *et al.* (1989)) ont ouvert la MS à l'étude des biomolécules de plus haut poids moléculaire dont les protéines. La source MALDI se distingue par sa tolérance aux potentiels contaminants d'un échantillon tandis que la source ESI est plus facilement couplée à des chromatographies liquides (Griffiths, 2008). Ces deux sources complémentaires sont encore à l'heure actuelle les deux techniques les plus utilisées dans les analyses de biomolécules, particulièrement en protéomique en association avec divers types d'analyseurs.

Lors de l'analyse de protéines en MS, le spectromètre de masse enregistre des spectres *full scan* des analytes présents au sein du mélange. Ce spectre renseigne alors sur la masse moléculaire du ou des ions au sein de l'échantillon (Figure 6 A). Afin d'identifier un analyte par détermination de sa séquence en acides aminés, l'expérimentateur ou le programme de contrôle du spectromètre de masse peut, quand l'appareil le permet, déclencher la fragmentation de l'analyte. On obtient alors un spectre de fragmentation (ou MS en tandem, MS/MS) qui permet de déterminer la composition en acides aminés de la protéine ou du peptide présent dans le mélange (Figure 6 B).

2.2.2 Modes d'acquisition des spectromètres de masse pour l'étude des protéines

La majorité des approches protéomiques emploient des spectromètres de masse équipés d'une source nano-ESI, utilisés en couplage avec une nano-chromatographie liquide. Il

existe cependant divers modes d'acquisition en fonction des applications désirées.

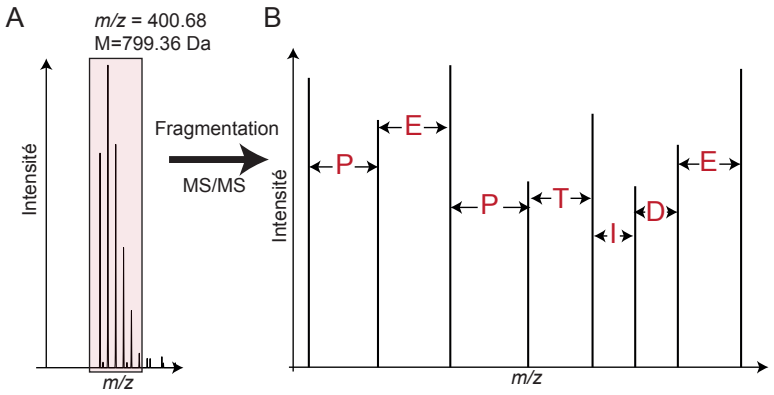


FIGURE 2.6 – **Spectres *full scan* et MS/MS**
Spectres *full scan* d'un peptide et mesure de sa masse précise **(A.)** et fragmentation de ce peptide permettant d'en déduire sa séquence en acides aminés PEPTIDE **(B.)**

2.2.2.1 Acquisition des spectres dépendante de l'intensité

Le spectromètre de masse qui opère en fonction de l'intensité des ions (ou en mode *data-dependant*, DDA) enregistre des spectres de manière cyclique. D'abord, un spectre *full-scan* recense les "N" espèces présentes au sein de l'échantillon et les classe selon leurs intensités relatives dans le mélange. Le spectromètre de masse isole ensuite successivement les N espèces et enregistre N spectres MS/MS séparément. Une fois le cycle terminé et les N spectres enregistrés, un nouveau cycle débute. On appelle plus communément ce mode d'acquisition "Top-N". Enfin, si l'exclusion dynamique est activée, le spectromètre de masse s'efforcera d'isoler les ions qu'il n'a pas fragmenté depuis un temps donné, évitant ainsi la fragmentation successive des mêmes précurseurs.

A l'heure actuelle, l'acquisition DDA est la méthode privilégiée pour la majorité des études protéomiques, notamment pour ses aptitudes de couverture du protéome (Aebersold et Mann, 2016).

2.2.2.2 Spectrométrie de masse ciblée

La MS ciblée, à l'inverse du mode d'acquisition DDA, ne tente pas de fragmenter les molécules en fonction de leur abondance et de leur apparition dans un spectre full scan. Ce mode d'acquisition se focalise uniquement sur les m/z que l'expérimentateur lui aura indiqué. Ceci est particulièrement intéressant pour détecter ou quantifier une molécule ou un groupe de molécules d'intérêt. La technique est plus sensible et précise que l'approche DDA mais ne permet de détecter qu'un nombre restreint d'analytes. Pour augmenter ses

capacités de couverture, il est alors nécessaire de programmer les acquisitions, notamment en fonction de leur temps de rétention chromatographique. Elle regroupe les stratégies *single reaction monitoring* (SRM) ou *multiple reaction monitoring* (MRM) (Picotti et Aebersold, 2012) et, plus récemment, *parallel reaction monitoring* (PRM) (Peterson *et al.*, 2012).

2.2.2.3 Acquisition par fragmentation indépendante de l'intensité

Récemment, une méthode d'acquisition s'est développée et connaît des applications en protéomique par MS. Ce mode réalise la fragmentation séquentielle successive de tous les ions compris dans une certaine gamme de *m/z* par tranches (appellé DIA ou SWATH). Ainsi, si le spectromètre de masse est paramétré dans ce mode d'acquisition entre *m/z* 400 et *m/z* 1200, il déclenche la fragmentation de tous les analytes compris dans la tranche *m/z* 400-425, puis entre *m/z* 424-450 et ainsi jusque *m/z* 1 174-1200 (Rosenberger *et al.*, 2014). Ce cycle sera alors répété tout au long de l'acquisition. Les spectres sont ensuite comparés à une banque de spectres obtenue au préalable par approche DDA ou issue d'une base de données spectrales ce qui permet d'identifier les analytes présents dans l'échantillon.

Ce mode d'acquisition a l'avantage d'être beaucoup plus précis que les approches de DDA pour la quantification mais permet une moindre couverture du protéome (Bruderer *et al.*, 2015).

2.2.3 Stratégies d'étude des protéines

L'étude des protéines par MS s'appuie essentiellement sur deux approches : l'approche *bottom-up* qui consiste à analyser les protéines après digestion enzymatique, et l'approche *top-down* qui permet d'analyser les protéines intactes.

2.2.3.1 Approche bottom-up

L'approche *bottom-up* repose sur l'utilisation d'enzymes protéolytiques qui permettent de cliver les différentes protéines contenues dans un mélange complexe en peptides.

Afin de décomplexifier un échantillon biologique complexe, il est possible de séparer préalablement les protéines par des techniques telles que l'électrophorèse sur gel de polyacrylamide (SDS-PAGE) ou encore par chromatographie d'exclusion stérique. Les protéines sont ensuite digérées par une enzyme protéolytique. Différentes enzymes peuvent être utilisées en fonction de leur spécificité de sites de clivage mais la majorité des études protéomiques emploient la trypsine, compte tenu de son efficacité, son coût et sa spécificité (Tsiatsiani et Heck, 2015).

Les peptides issus de la digestion peuvent éventuellement être séparés à leur tour, soit pour décomplexifier l'échantillon, soit pour enrichir l'échantillon en peptides d'intérêt comme ceux présentant une modification post-traductionnelle particulière (Xu *et al.*, 2010a; Pan *et al.*, 2011; Sharma *et al.*, 2014). Les peptides sont par la suite analysés en MS. Ceux dont la séquence ne peut provenir que d'une seule protéine (appelés peptides uniques) permettent alors l'identification de protéines (Figure 7 A-D).

L'approche *bottom-up* dite *shotgun* consiste à identifier et quantifier le plus grand nombre de protéines à partir d'un échantillon complexe. Elle reste, jusqu'à présent, la technique offrant la plus grande couverture du protéome. Par cette approche, il est possible d'identifier, pour des échantillons très fractionnés ou pour des gradients de chromatographie très longs, un nombre de protéine de l'ordre de 10 000 protéines par lignée cellulaire ou tissu biologique (Nagaraj *et al.*, 2011; Beck *et al.*, 2011; Geiger *et al.*, 2012; Coscia *et al.*, 2016). Toutefois, cette technique reste problématique pour l'étude de petites protéines qui génèrent un plus faible nombre de peptides suite au traitement enzymatique. Aussi, certaines protéines ne génèrent pas le moindre peptide unique. Dans ce cas, il est possible d'employer différentes enzymes capables de générer des peptides différents.

Bien que l'approche *bottom-up* soit la méthode de préparation d'échantillon la plus employée dans les études protéomiques, les progrès récents permettent désormais d'identifier des protéines intactes sans traitement protéolytique.

2.2.3.2 Approche top-down

Les récentes avancées des analyseurs à haute résolution, notamment des analyseurs de type *Orbitrap* (Makarov, 2000), mais également des techniques de fragmentation, ont permis d'identifier un grand nombre de protéines à partir de mélanges complexes (Tran *et al.*, 2011). Cette approche est pour l'instant majoritairement appliquée avec le mode d'acquisition DDA.

L'avantage majeur de l'approche *top-down* est qu'elle donne accès aux masses intactes des protéines de l'échantillon, ce que ne permet pas l'approche *bottom-up*. Les modifications post-traductionnelles sont également directement accessibles alors qu'elles nécessitent en général des étapes d'enrichissement par *bottom-up* (Catherman *et al.*, 2014) (Figure 7 E-H). Les analyses par approche *top-down* offrent ainsi la possibilité de décrire le protéome d'un échantillon à un niveau de complexité jusqu'alors difficilement accessible par protéomique *bottom-up* par l'observation de "protéoformes" (Smith *et al.*, 2013; Toby *et al.*, 2016). En effet, les protéines étaient jusqu'alors décrites selon le terme « isoforme », ce qui prend en considération l'épissage alternatif. Grâce à l'observation de la protéine

dans son état intact, il est alors possible de déterminer avec précision d'une part l'épissage, mais également les diverses combinaisons de modifications post traductionnelles que le précurseur protéique peut présenter. Ainsi, si l'on considère un gène produisant deux transcrits épissés, qui tous deux, une fois traduits, présentent un site de modification post-traductionnelle, la protéomique par approche *top-down* permettra l'identification éventuelle des quatre « protéoformes » (deux isoformes modifiées ou non) de manière simultanée. La protéomique par approche *bottom-up* quant à elle nécessitera d'une part, l'identification de fragments protéolytiques spécifiques de la région épissée et d'autre part, l'identification de peptides sous leurs formes intactes et modifiées pour offrir des performances comparables en termes de d'identification protéique.

Toutefois, l'analyse protéomique en *top-down* reste encore moins efficace que les approches *bottom-up*/*shotgun* en termes de couverture de protéome avec un nombre absolu d'identification de protéines uniques de l'ordre de centaines, voire un millier lorsque les protéines sont préfractionnées (Catherman *et al.*, 2014). Ceci s'explique d'un point de vue expérimental par le long temps d'acquisition des spectres qui augmente avec la masse moléculaire des ions à observer (Durbin *et al.*, 2016). De ce fait, l'approche *top-down* semble plus performante pour la caractérisation de petites protéines ce qui la rend complémentaire à l'approche *bottom-up*. Enfin, l'analyse de protéines par *top-down* nécessite en général une étape de précipitation de protéines afin d'éliminer les détergents et sels, incompatibles avec la MS en source ESI. Ceci occasionne des difficultés supplémentaires, notamment pour resolubiliser les protéines avant séparation par chromatographie liquide (Kachuk et Doucette, 2017). Elle reste néanmoins une approche prometteuse et ses capacités sont constamment améliorées.

2.2.4 Identification des protéines en spectrométrie de masse

L'identification par MS des protéines ou des peptides qui en sont issus repose sur la comparaison entre un spectre expérimental mesuré et un spectre théorique, déduit à partir de la séquence protéique/peptidique contenue dans une base de données (Eng *et al.*, 1994; Liu *et al.*, 2012).

2.2.4.1 Corrélation entre spectres expérimentaux et théoriques

En DDA, chaque spectre est comparé aux peptides ou protéines présentes dans une base de données de séquences protéiques. Les protéines (approche *top-down*) ou peptides (approche *bottom-up*) dont la masse correspond à celle des ions observés lors d'un spectre *full-scan* sont sélectionnés pour être identifiés. Les ions observés lors de la fragmenta-

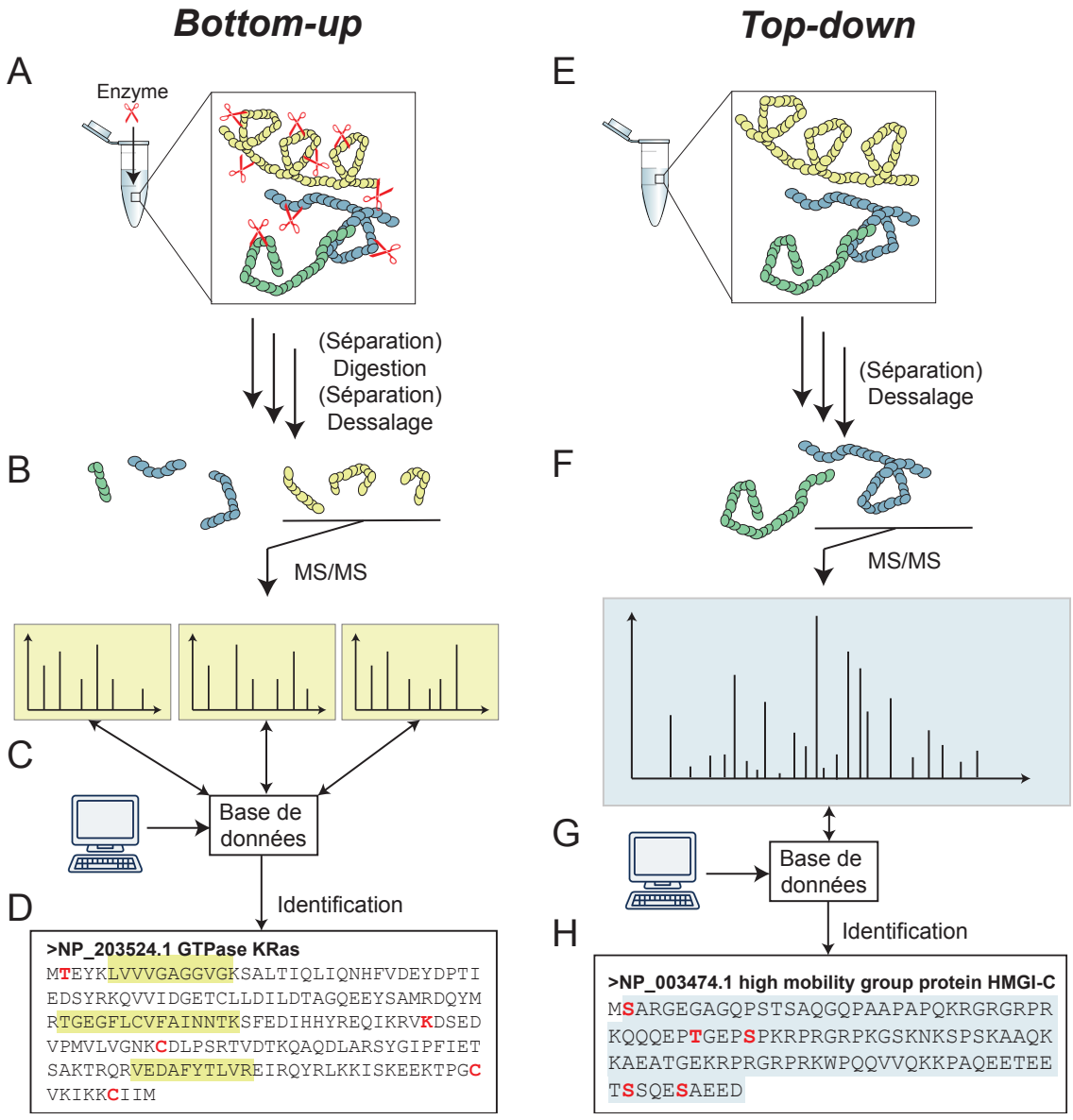


FIGURE 2.7 – **Représentation schématique des stratégies protéomiques *bottom-up* (gauche, A-D) et *top-down* (droite, E-H)**
Préparation des échantillons **(A et E)** et analyse en MS **(B et F)**. Les spectres obtenus sont analysés par approche bioinformatique à l'aide d'une base de données de protéines **(C et G)**. Cette étape permet l'identification de protéine(s) **(D et H)**. Les modifications post-traductionnelles sont marquées en rouge.

tion sont ensuite comparés aux masses attendues des fragments théoriques. L'algorithme d'identification crée alors un score basé sur la corrélation entre les fragments théoriques et expérimentaux pour déduire la séquence de l'analyte fragmenté. L'algorithme déduit alors la protéine identifiée sur la base de la séquence observée la plus probable (Eng *et al.*, 1994; Liu *et al.*, 2012).

En protéomique ciblée ou en DIA/SWATH, l'identification repose souvent sur la comparaison des spectres mesurés et des spectres expérimentaux référencés dans une banque de spectres et sur leur temps de rétention chromatographique. Les intensités relatives des différents fragments sont rapportées à celles de la banque de spectres et la similarité spectrale est calculée sous la forme du calcul d'un angle de contraste (Wan *et al.*, 2002) ou produit vectoriel (Stein et Scott, 1994). L'identification de la protéine requiert une similarité spectrale élevée.

2.2.4.2 Validation statistique des identifications protéiques

Une fois que l'algorithme de recherche a assigné une séquence à chaque spectre, il est nécessaire de filtrer les résultats pour éliminer les potentiels faux-positifs et obtenir une estimation du taux de faux-positifs (ou *false discovery rate*, FDR). Pour ce faire, les bases de données de séquences protéiques (appelées cibles, ou *target*) sont mélangées avec des séquences de protéines fictives (appelées leurres, ou *decoy*). Si l'algorithme d'identification corrèle un spectre avec une séquence *decoy* et un score associé bas, il est très probable que ce spectre soit de qualité insuffisante pour donner une identification fiable. En revanche s'il assigne une séquence avec un haut score à une protéine *target*, alors il est très probable que le spectre observé provienne bel et bien de cette protéine (Elias et Gygi, 2007).

En approche *bottom-up*, les séquences des protéines *target* sont associées le plus souvent à des séquences de protéines inversées *decoy*. Tous les spectres d'une expérience de MS sont classés en fonction de leur score, allant du plus élevé au plus faible. Grâce au nombre d'identifications *target* et *decoy* (Figure 8 A), on peut estimer un nombre de vrais positifs, faux positifs et ainsi en déduire un taux de faux positifs (ou *false discovery rate*, FDR) (Elias et Gygi, 2007). Pour cela, on définit un score limite minimal en deçà duquel le FDR désiré est atteint (en général 1%) (Elias et Gygi, 2007) (Figure 8 B). Chaque paire spectre-peptide est appelée correspondance spectre-peptide (ou *peptide-spectrum match*, PSM). Il est possible d'affiner à nouveau les résultats en estimant un FDR au niveau peptidique, après association des PSMs, et au niveau protéique.

Tandis que l'approche *bottom-up* requiert des séquences *decoy* inversées, l'estimation de faux positifs en approche *top-down* s'appuie davantage sur la correction des scores selon Bonferoni (Catherman *et al.*, 2013). Il est toutefois possible d'appliquer une approche *target-decoy* avec des séquences de protéines mélangées (Kellie *et al.*, 2011). Puisque l'analyse *top-down* concerne les protéines intactes, l'estimation des faux positifs se limite au niveau spectral.

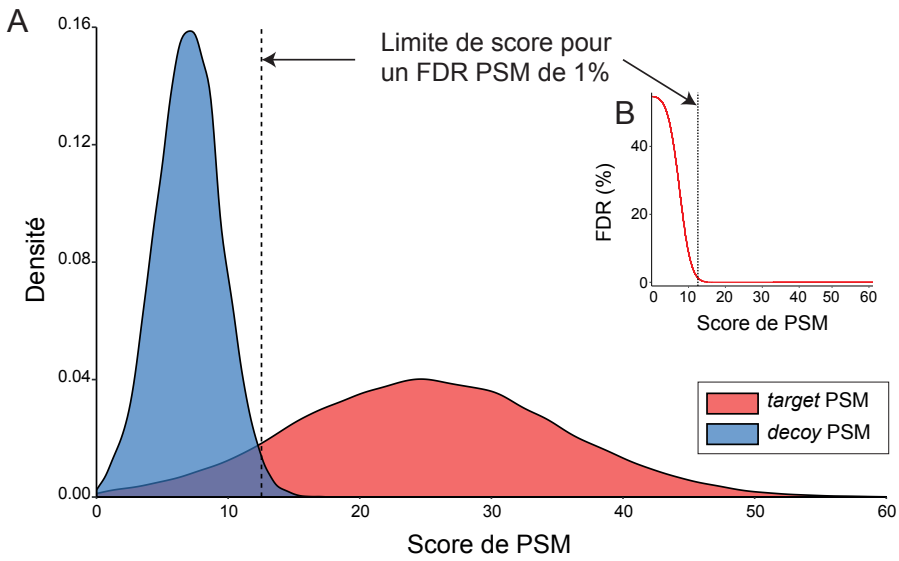


FIGURE 2.8 – **Stratégie de validation statistique des identifications**
Chaque spectre est assigné à une séquence peptidique normale (*target*) ou leurre (*decoy*) qui sont classées selon leur score **(A.)**. On calcule alors un score limite pour un taux de faux positifs de 1% **(B.)**. Les PSMs ayant un score supérieur au score limite (lignes pointillées) seront conservés pour la suite de l'analyse alors que ceux ayant un score inférieur seront écartés.

En dépit de ces outils, il est souvent délicat d'expliquer des phénomènes biologiques en s'appuyant uniquement sur des données d'identification. C'est pourquoi il est crucial de pouvoir quantifier les protéines au sein d'échantillons biologiques pour associer des variations d'expression à des observations phénotypiques.

2.2.5 La quantification des protéines en spectrométrie de masse

2.2.5.1 La quantification sans marquage

La quantification des protéines sans marquage est encore, à l'heure actuelle, la méthode de quantification des protéines la plus employée, notamment pour son faible coût et sa facilité de mise en œuvre. Elle s'appuie sur deux méthodes principales par DDA : le comptage de spectres et la quantification basée sur l'intensité du signal du précurseur. Plus récemment, les approches DIA/SWATH sont utilisées comme autres méthodes de quantification.

Comptage spectral
Le nombre de spectres MS/MS qui donnent lieu à une identification pour une protéine par DDA est proportionnel à la quantité de cette dernière. Ainsi, les protéines les plus abon-

dantes génèrent un grand nombre de spectres et inversement pour les moins abondantes. Cette approche est toutefois biaisée. Les grandes protéines sont susceptibles de générer un plus grand nombre de peptides uniques et donc artificiellement un plus grand nombre de spectres. Il est alors nécessaire de normaliser les valeurs mesurées en rapportant le comptage à la longueur de la protéine, ou à son nombre de peptides uniques théoriques. Ce biais associé au mode d'acquisition DDA explique en partie pourquoi la quantification par comptage spectral est l'approche de quantification par MS la moins précise (Liu *et al.*, 2016; Choi *et al.*, 2017).

Quantification par intensité du précurseur

Par DDA, il est possible d'obtenir des informations quantitatives plus précises par les approches de quantification en MS1 (intensité du précurseur). Dans cette méthode, on enregistre systématiquement la trace chromatographique de l'ion précurseur qui mène à une identification par MS/MS. Il est possible d'utiliser les paires précurseur-spectre MS/MS pour identifier ces mêmes précurseurs dans des échantillons similaires, à l'aide de la mesure de masse précise et de son temps de rétention chromatographique. Ceci permet notamment de réduire le nombre de valeurs manquantes lors d'analyses statistiques ultérieures (Cox *et al.*, 2014). Enfin, cette méthode permet également d'estimer les quantités absolues des protéines. En effet, si on considère que la quantité d'histones, protéines nécessaires à l'empaquetage de l'ADN dans les cellules, est stable pour un type cellulaire donné, il est alors possible de rapporter le signal global des protéines à celui des histones et d'en déduire leur quantité (Wiśniewski *et al.*, 2014). Cette approche est toutefois moins précise que les approches de quantification absolue par dilution isotopique car elle considère que la réponse du signal est identique pour chaque peptide.

Fragmentations indépendantes de l'intensité

Le mode d'acquisition DIA/SWATH offre la plus grande précision pour la quantification de protéines dans un échantillon biologique sans marquage. Par cette méthode, tous les ions sont fragmentés successivement sans sélection préalable ce qui élimine les problèmes de valeurs manquantes en DDA (Aebersold et Mann, 2016). Les protéines sont alors quantifiées grâce à l'intensité des ions fragments de leurs peptides.

2.2.5.2 Méthodes de quantification par marquage

Les méthodes de quantification par marquage emploient des isotopes stables qui permettent de multiplexer l'analyse de plusieurs échantillons en une expérience de MS. Elles

sont généralement plus précises que les approches sans marquage, mais sont également plus coûteuses, du fait de l'emploi d'isotopes stables.

Marquage métabolique

Le marquage métabolique par des isotopes stables d'acides aminés (SILAC) implique l'utilisation de milieux de cultures contenant des acides aminés synthétiques enrichis en atomes stables comme le ^{13}C ou ^{14}N. Ces atomes ne se sont pas des radionucléides et ne sont présents qu'à l'état de traces dans la nature. Les acides aminés marqués sont incorporés aux cellules en culture (ou parfois aux animaux, Geiger *et al.* (2013)). Les protéines de cellules sans milieu SILAC et celles avec milieu SILAC sont mélangées dans des proportions équivalentes. Les peptides marqués et non marqués sont détectés de façon quasi-systématique lors de l'élution chromatographique et leur intensité relative sont comparées. Cette approche est limitée aux cellules en culture et donc incompatible avec l'étude de tissus biologiques. De plus, pour chaque marquage, le spectromètre de masse pourra potentiellement fragmenter l'ion léger et l'ion lourd, ce qui diminue automatiquement ses capacités de couverture du protéome. Cette méthode limite l'analyse à un maximum de 3 échantillons par expérience de MS.

Tableau 2.1 – **Evaluation des méthodes de quantification en MS**
D'après Liu *et al.* (2016)

Méthode de quantification	**Marquage**	**Précision**	**Reproductibilité**
Comptage spectral	Sans	+	++
Intensité du précurseur	Sans	+++	++++
DIA/SWATH	Sans	++++	++++
SILAC	Avec	++++	++++
Marquage chimique MS1	Avec	+++	+++
Marquage chimique MS2	Avec	+++	+++
AQUA	Avec	++++	++++

Marquage chimique

Le marquage chimique consiste à induire une réaction entre les peptides de digestion enzymatique et un réactif marqué. Pour chaque échantillon, un type de réactif est employé. Tous les échantillons sont ensuite mélangés dans des proportions égales. Il existe deux types de marquages chimiques : le marquage avec observation de précurseurs légers et lourds (MS1, similaire au SILAC) et le marquage avec rapporteur de fragmentation (MS2). L'utilisation de rapporteurs MS2 permet de multiplexer jusqu'à 10 échantillons simultanément sans diminuer les performances du spectromètre de masse. En effet, toutes les espèces mélangées ont la même masse et ne sont différenciées que par des ions rapporteurs visibles dans les spectres MS/MS (Aebersold et Mann, 2016). Cette stratégie est applicable aux tissus et est notamment employée dans la caractérisation de tissus cancéreux (Zhang *et al.*, 2016a; Mertins *et al.*, 2016).

Quantification absolue par étalon interne synthétique

La quantification absolue par étalon interne synthétique (AQUA) consiste à incorporer un peptide marqué par des isotopes stables au sein de l'échantillon, préférentiellement au moment de l'ajout de l'enzyme (Gerber *et al.*, 2003; Shuford *et al.*, 2012). Avec l'ajout d'une quantité connue d'un peptide synthétique, il est possible de déterminer de manière absolue la quantité endogène du peptide issu de la protéine d'intérêt. On peut, par cette approche, déterminer le nombre de copies d'une protéine par cellule ou pour une quantité de tissu donnée (Gerber *et al.*, 2003). Elle est néanmoins limitée à un faible nombre de protéines, est principalement employée dans des expériences de protéomique ciblées et n'offre pas la possibilité de multiplexer différents échantillons.

2.2.6 Applications de la spectrométrie de masse spécifiques aux tissus

2.2.6.1 L'imagerie par spectrométrie de masse

L'imagerie par MS est une technique qui permet de cartographier des molécules exogènes ou endogènes comme les métabolites, peptides et protéines au sein de tissus biologiques. Elle repose en grande partie sur l'utilisation de spectromètres de masse équipés d'une source MALDI.

L'analyte est co-crystalisé avec une matrice. L'excitation de ces cristaux par un rayon laser désorbe et ionise les molécules contenues dans les cristaux. Ces ions sont séparés au sein de l'analyseur et leur masse est déterminée. Si l'on applique la matrice à la surface d'une section de tissu biologique, comme un organe ou une biopsie, on peut alors évaluer la

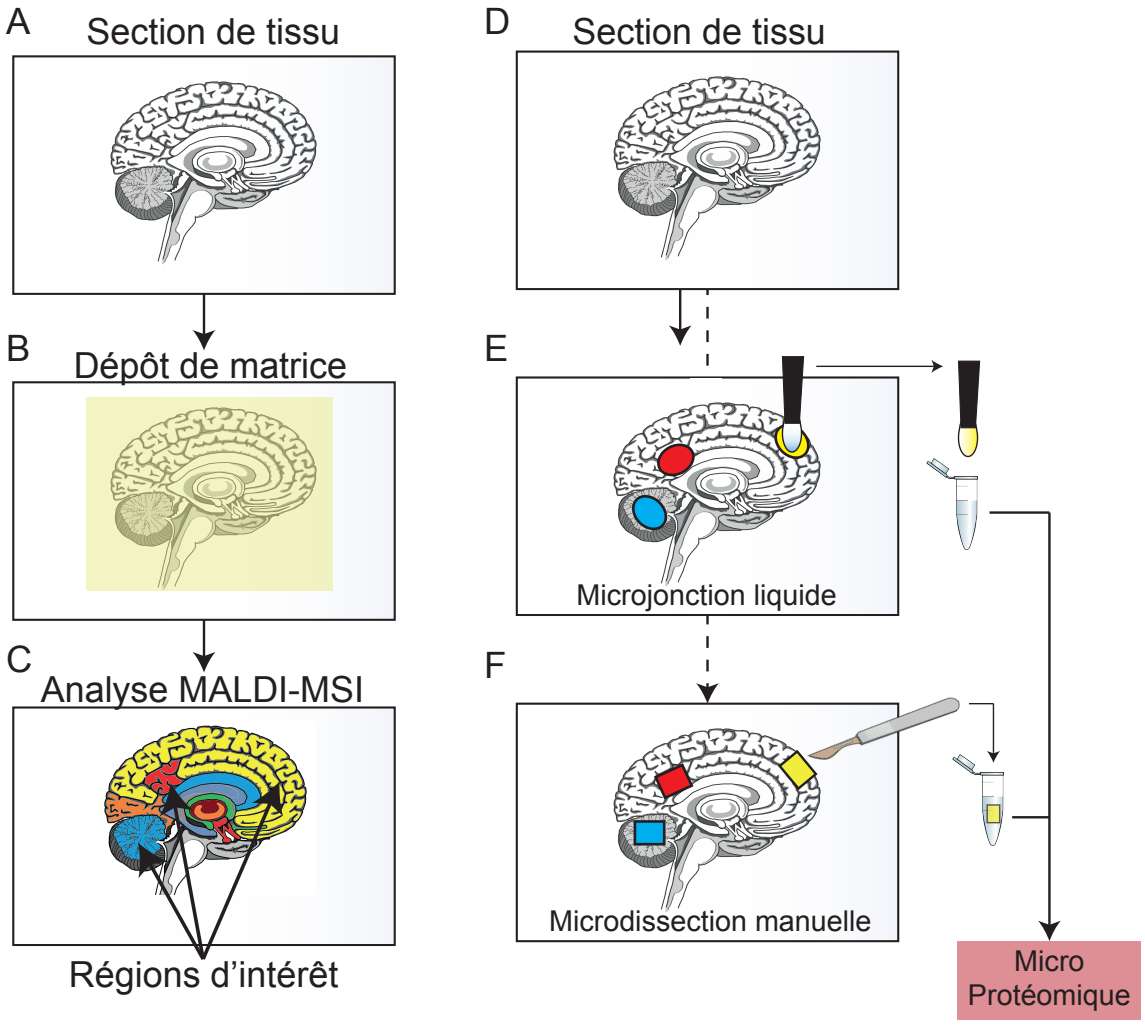


FIGURE 2.9 – **Stratégies de MALDI-MSI et micro-protéomique**
A. et D. Préparation des sections de tissus. **B.** Dépôt de matrice nécessaire au processus de désorption-ionisation MALDI. **C.** Expérience d'imagerie MALDI et définition de régions d'intérêt. **E.** Micro-extractions par jonction liquide. **F.**. Micro-extractions manuelles.

composition en ions de ce tissu. L'association des spectres et de leurs coordonnées permet de reconstruire une image moléculaire du tissu (Caprioli *et al.*, 1997).

Cette méthode connaît un grand succès pour classifier des tissus par histologie moléculaire, notamment pour délimiter une région saine d'une région tumorale. Elle est par exemple employée pour la découverte de biomarqueurs (Lemaire *et al.*, 2007a,b; Franck *et al.*, 2009a).

Certaines stratégies permettent également d'identifier des peptides issus de digestion *in situ* par imagerie MALDI (Franck *et al.*, 2009b) mais elles sont en général limitées aux protéines les plus abondantes.

2.2.6.2 Analyse microprotéomique

Compte tenu du manque de méthodes d'identification robustes associées à l'imagerie MALDI, cette dernière est parfois employée pour définir des régions d'intérêt desquelles les protéines sont extraites et analysées par approches des protéomiques qui couvrent plus largement le protéome.

Dans un premier temps, un tissu biologique ou biopsie est préparé et analysé par imagerie MALDI (Figure 9 A et B) et les régions d'intérêt sont déterminées par classification moléculaire du tissu (Figure 9 C). En parallèle, des sections adjacentes du tissu sont préparées et déposées sur une lame (Figure 9 D). Les protéines des régions d'intérêt sont ensuite extraites par application d'un solvant par micro-jonction liquide (Quanico *et al.*, 2013) (Figure 9 E) ou par microdissection manuelle assistée par parafilm (Franck *et al.*, 2013) (Figure 9 F). Il est également possible de réaliser des microdissections laser (Dilillo *et al.*, 2017).

Les approches protéomiques décrites précédemment permettent d'identifier et de quantifier des protéines à l'aide de bases de données. Celles-ci sont issues des annotations des génomes et transcriptomes ainsi que de la caractérisation de protéines selon le dogme : 1 gène - 1 protéine. Toutefois, de nombreuses évidences récentes rapportent l'expression de nouvelles protéines issues de cadres de lecture alternatifs de gènes codants ou annotés comme non codants.

2.3 Les cadres de lecture alternatifs et protéines alternatives

2.3.1 Définition de protéine alternative

Un cadre de lecture ouvert (ORF) est défini par la région comprise entre deux codons STOP qui peut contenir un codon d'initiation de la traduction. Généralement, cette définition est simplifiée en considérant qu'un ORF représente une séquence potentiellement codante, comprise entre un codon initiateur (ATG) et un codon STOP. Les cadres de lecture alternatifs (altORFs) représentent toutes les séquences pouvant coder pour des protéines différentes des protéines canoniques (ou protéines de référence) recensées dans les diverses bases de données. Ces protéines sont appelées protéines alternatives (Vanderperre *et al.*, 2013; Mouilleron *et al.*, 2015).

Lors de la traduction, le ribosome lit un transcrit par groupe de trois nucléotides appelés codons et le traduit en protéine. Il existe donc trois cadres de lecture au sein d'un transcrit. Un décalage d'une base dans le sens de lecture entraîne un changement de cadre de lecture

et donc un codon différent (Figure 10 A). Une fois traduites, les protéines issues des trois cadres de lecture ouverts différents ont des séquences différentes (Mouilleron *et al.*, 2015) (Figure 10 B).

Les séquences codantes issues d'altORFs peuvent se situer dans des régions dites "non traduites" (UTRs), dans des cadres de lecture chevauchant le cadre de lecture canonique d'un ARNm (Figure 10 C) ou encore au sein des ARN non-codants (Figure 10 D). Les altORFs sont généralement classés en différents groupes en fonction de la localisation du codon initiateur de la traduction "AUG" sur le transcrit (Tableau 2).

Tableau 2.2 – **Classification des altORFs en fonction de la localisation du codon d'initiation**
D'après Mouilleron *et al.* (2015)

altORF$^{5'}$	altORFs dont le codon d'initiation est localisé dans la région 5'UTR
altORF^{CDS}	altORFs dont le codon d'initiation est localisé dans un cadre de lecture chevauchant le CDS canonique
altORF$^{3'}$	altORFs dont le codon d'initiation est localisé dans la région 3'UTR
altORF^{nc}	altORFs dont le codon d'initiation est localisé sur un ARN non-codant putatif

De manière générale, les séquences protéiques issues de la traduction d'altORFs sont plus courtes que celles des protéines de référence. En effet, dans la très grande majorité des cas, lors de l'annotation du génome, le plus grand ORF pourvu d'une séquence codante est automatiquement annoté comme CDS (ou refORF) dès lors que sa longueur est supérieure à 100 codons (Basrai *et al.*, 1997). Il en résulte que les ORFs restants, ou altORFs, sont en général plus petits (Vanderperre *et al.*, 2013; Mouilleron *et al.*, 2015).

Il est important de noter que les protéines alternatives se distinguent des isoformes de protéines canoniques. En effet, on appelle isoforme une protéine produite à partir d'un transcrit épissé. L'absence ou l'ajout d'un exon modifie une partie de la séquence de la protéine. Elle garde toutefois une homologie de séquence très forte avec la protéine canonique. Afin de limiter la prédiction de faux positifs alternatifs, c'est à dire d'isoformes non prédites dans les bases de données protéiques, la séquence de la protéine de référence est comparée à chacune des séquences des protéines alternatives prédites issues du même gène. Dans le cas où une protéine alternative présente une homologie de séquence importante avec la protéine de référence du gène sur une part significative de sa longueur totale, elle est étiquetée comme nouvelle isoforme potentielle (Mouilleron *et al.*, 2015). Cette

vérification d'homologie de séquence n'est pas nécessaire pour les protéines alternatives prédites à partir de transcrits non-codants car ils ne sont pas associés à une protéine de référence.

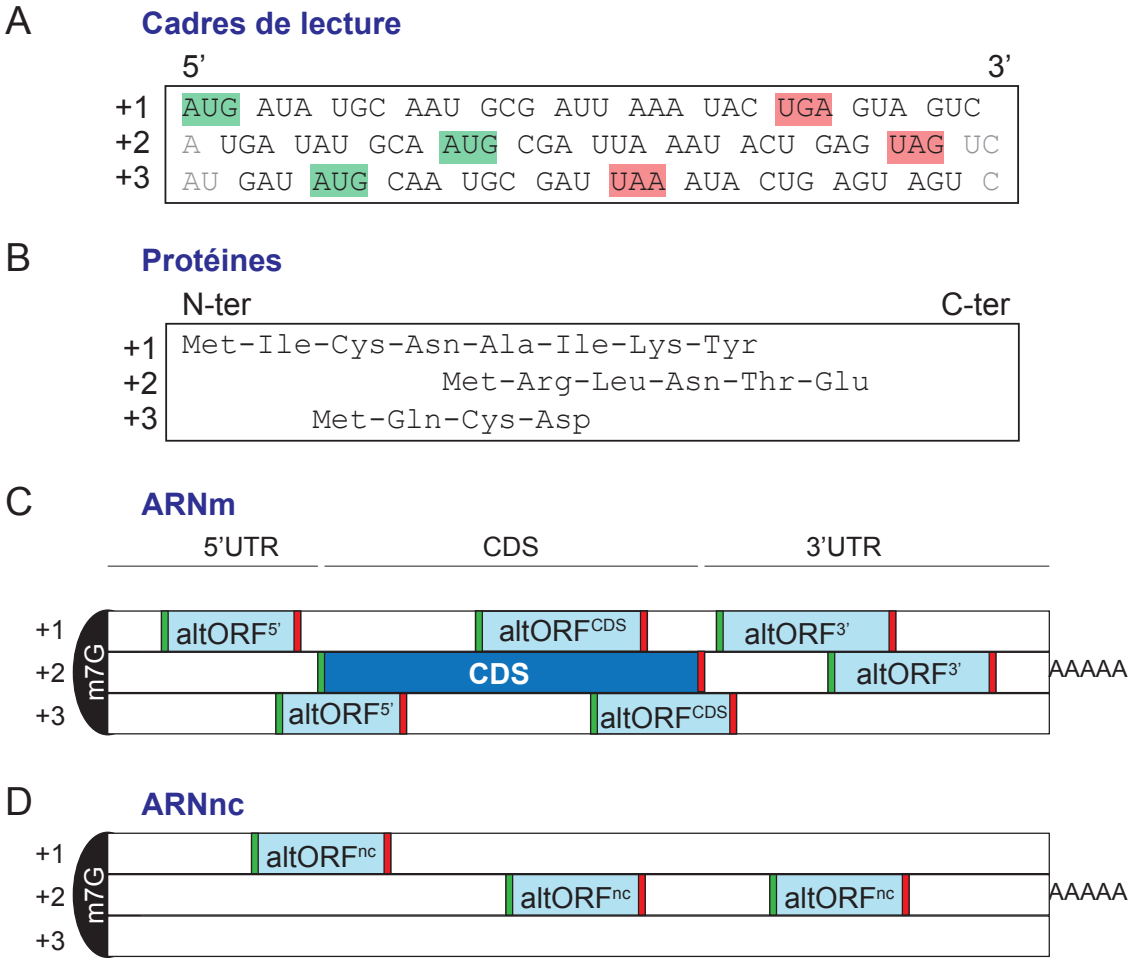


FIGURE 2.10 – **Schéma général des protéines alternatives**
A. Séquence en nucléotides d'un transcrit fictif présentant des séquences codantes potentielles dans ses trois cadres de lecture ouverts. **B.** Traduction *in silico* du transcrit A. **C. et D.** Schéma représentatif des cadres de lecture ouverts observables au sein d'un ARN messager (C.) ou au sein d'un ARN non-codant (D.). La séquence codante pour la protéine canonique est annotée "CDS".

2.3.2 Stratégies de prédiction des protéines alternatives

Il est possible de prédire les protéines alternatives à partir de plusieurs types de données.

A partir de données transcriptomiques

Les techniques de séquençage à haut débit des transcrits (RNASeq) offrent une profondeur

de couverture du transcriptome de plus en plus grande. Ces données peuvent être utilisées pour la prédiction d'ORFs canoniques et alternatifs. Deux approches sont possibles. L'une d'entre elles utilise des données générées de façon interne pour un type de tissu ou de cellule spécifique, en fonction des échantillons à étudier. Cette approche a l'avantage de n'incorporer dans les prédictions que les transcrits effectivement observés dans le/les échantillon(s) d'intérêt. Ceci limite notamment l'étendue des prédictions (Ma *et al.*, 2016). La deuxième approche s'appuie sur la prédiction des altORFs à l'échelle du transcriptome complet référencé dans les différentes bases de données de séquences de transcrits (Pruitt et Maglott, 2001; Flicek *et al.*, 2011). Cette approche présente l'avantage majeur d'être plus complète, car tous les transcrits référencés seront considérés. La prédiction ne sera alors pas conditionnée par l'observation de certains transcrits, critère limitant pour les transcrits de très faible abondance. Elle peut néanmoins être source de faux positifs. En effet, dans le cas où un transcrit est exprimé dans un type de tissu ou lignée cellulaire très spécifique, la prédiction des altORFs de ce transcrit pour l'étude d'un système biologique différent peut engendrer des erreurs.

A partir de données génomiques

Un organisme pour lequel on dispose uniquement du séquençage du génome peut être relativement difficile à étudier. Avec l'aide des outils bioinformatiques et des prédictions d'exons, il est possible de déduire la séquence des gènes avec une relative précision. Après avoir annoté les protéines de référence putatives par homologie de séquence avec celles répertoriées pour d'autres espèces, les altORFs peuvent être prédits.

La prédiction peut être basée sur la traduction *in silico* des six cadres de lecture d'un génome complet (trois pour le brin sens, et trois pour le brin antisens). Cette technique est particulièrement utilisée dans les approches de protéogénomique mais augmente considérablement la taille de l'espace de recherche. Elle limite également la prédiction des séquences dans les jonctions exons-exons (Nesvizhskii, 2014).

Enfin, seuls les altORFs d'une taille minimale de 30 codons sont généralement considérés (Vanderperre *et al.*, 2013; Samandi *et al.*, 2017). Bien qu'il n'y ait pas de critère de taille minimale pour la fonctionnalité d'une protéine ou d'un peptide, prédire à l'échelle du génome les ORFs sans restriction de taille résulterait en un nombre extrêmement grand de séquences. En effet, certaines ORFs non fonctionnels peuvent apparaître au cours de l'évolution. De plus, il est délicat de prédire une fonction ou d'étudier la conservation des ORFs courtes. L'étude des protéines issues d'ORFs plus petits que 30 codons reste possible en privilégiant une approche ciblant un/des gène(s) d'intérêt.

2.3.3 Découverte de protéines encodées à partir de cadres de lecture alternatifs

La première observation de protéine encodée à partir d'un cadre de lecture alternatif a été décrite pour le gène *CDKN2A/INK4a*. Ce gène était alors connu pour encoder une protéine de 168 acides aminés (156 chez l'humain) "p16INK4a", inhibitrice des *Cyclin dependant kinase* et capable d'arrêter la progression du cycle cellulaire. A partir d'un transcrit issu de l'épissage alternatif et qui débute par un exon différent, le gène encode une protéine de 169 acides aminés (132 chez l'humain) "p19-ARF". Cette protéine issue d'un altORF décalé est également capable d'arrêter le cycle cellulaire (Ouelle *et al.*, 1995). Ce gène code donc pour deux protéines qui ne sont pas des isoformes, toutes deux impliquées dans la régulation du cycle cellulaire (Figure 11).

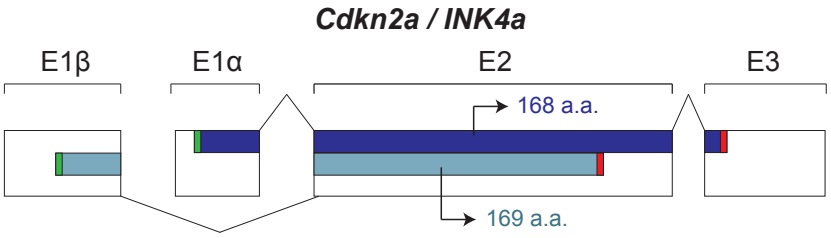


FIGURE 2.11 – **Schéma représentatif des transcrits codant pour les protéines du gène *CDKN2A/INK4a* chez la souris**
Les ORFs de la protéine p16-INK4a et de p19-ARF sont respectivement marqués en bleu foncé et bleu clair.

Depuis, d'autres exemples de protéines alternatives encodées à partir de gènes codants ont été décrits. Parmi ces exemples, le transcrit qui code pour Histone H4 code également pour un peptide lié à la croissance œstrogène-dépendante (Bab *et al.*, 1999; Smith *et al.*, 2005). Le gène XLαs/Gαs code la protéine alternative Alex qui interagit avec sa protéine de référence (Klemke *et al.*, 2001; Abramowitz *et al.*, 2004). Le gène *RPP14* codant pour une sous unité du complexe *ribonuclease P* code également pour la thioester-dehydratase "HsHTD" à partir d'un altORF localisé dans la région 3'UTR du transcrit ; cette nouvelle protéine mitochondriale est associée à la synthèse des acides gras (Autio *et al.*, 2008) (Figure 12 A). Le gène de la protéine Prion code pour altPrP à partir d'un cadre de lecture décalé de son CDS (Vanderperre *et al.*, 2011) (Figure 12 B). Le gène ATXN1 code pour une protéine alternative, qui interagit directement avec la protéine canonique (Bergeron *et al.*, 2013). Le gène *MKKS* code pour deux protéines alternatives uMKKS1 & 2 traduites à partir de la région 5'UTR (Akimoto *et al.*, 2013). Le récepteur à l'angiotensine II de type 1 est co-traduit avec un peptide de 7 acides aminés intervenant dans sa signalisation cellulaire (Liu *et al.*, 2014; Yosten *et al.*, 2016) (Figure 12 C). Le récepteur à

adénosine 2-A est co-traduit avec une protéine alternative présente dans la région 5'UTR de son transcrit. L'expression de sa protéine alternative est régulée positivement lors de la stimulation du récepteur de façon post-transcriptionnelle (Lee *et al.*, 2014). Plus récemment, le gène *MUC1* a été décrit pour exprimer une protéine alternative dont l'expression est corrélée avec celle de sa protéine de référence (Chalick *et al.*, 2016) à partir d'un cadre de lecture décalé de son CDS. Enfin, le gène qui code pour l'insuline code également pour une protéine alternative immunogénique associée au diabète de type 1 à partir d'un cadre de lecture décalé de son CDS (Kracht *et al.*, 2017).

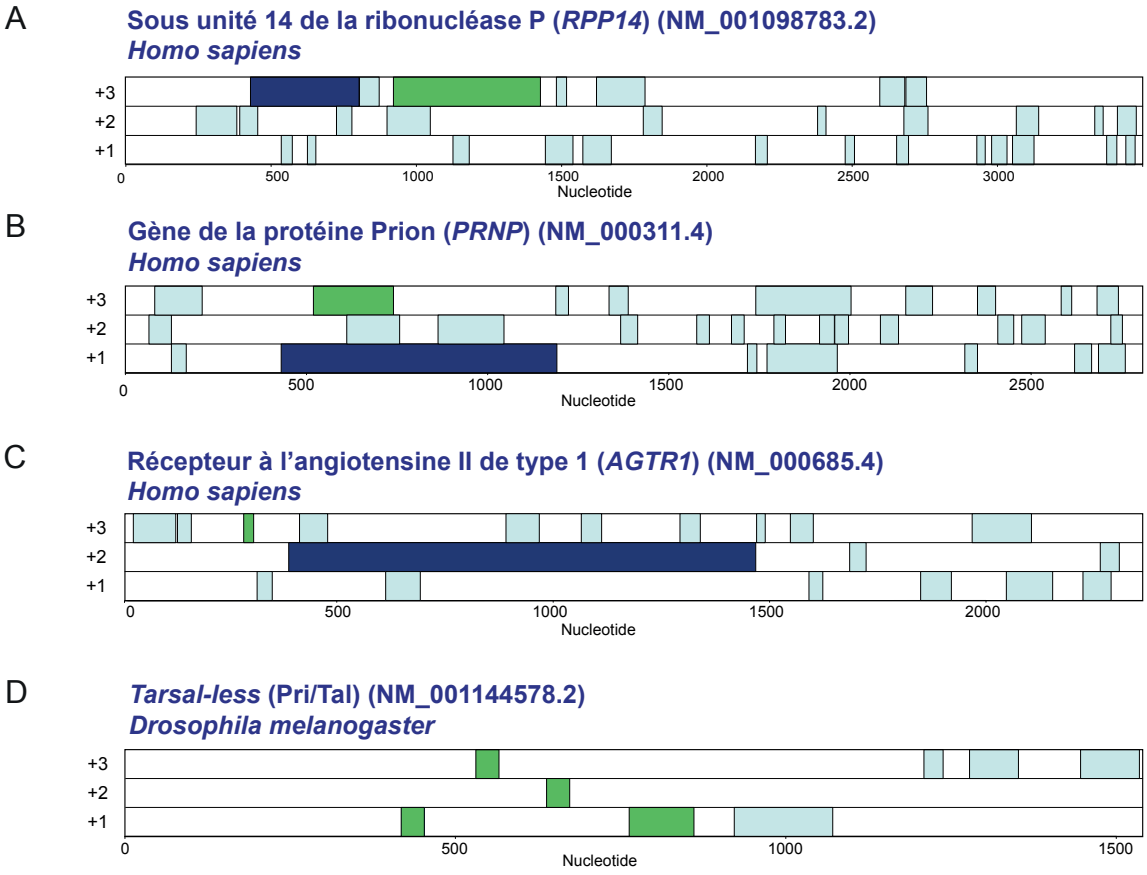


FIGURE 2.12 – **Exemple d'ARNs avec évidences d'expression d'une ou plusieurs protéines alternatives**
Représentation schématique des transcrits canoniques des gènes *RPP14* (**A.** Autio *et al.* (2008)), *PRNP* (**B.**, Vanderperre *et al.* (2011)), *AGTR1* (**C.**, Liu *et al.* (2014)) et Tal (**D.**, Galindo *et al.* (2007)). Les altORFs prédits sont colorés en bleu clair, les séquences codantes des protéines de référence en bleu foncé et les altORFs codant pour une protéine alternative en vert.

Aussi, de nombreuses études ont mis en évidence l'expression de protéines alternatives à partir de transcrits annotés comme non-codants. Un exemple remarquable concerne le

gène *Tarsal-less* chez la drosophile avec quatre ORFs successifs codants (Galindo *et al.*, 2007) ; les quatre protéines correspondantes sont impliquées dans le développement embryonnaire et conservées à travers de nombreuses espèces eucaryotes (Galindo *et al.*, 2007; Kondo *et al.*, 2010; Zanet *et al.*, 2015) (Figure 12 D). D'autres études rapportent la découverte de protéines encodées à partir de transcrits initialement annotés comme non-codants, et leur annotation a depuis été modifiée. Le gène *LOC550643/LINC01420* renommé *NBDY* encode une protéine de 68 acides aminés impliquée dans le clivage de la coiffe des transcrits (D'Lima *et al.*, 2017). Enfin, un autre gène initialement classé comme non-codant par l'annotation automatique du génome est effectivement traduit en une protéine de 90 acides aminés qui régule la régénération musculaire (Matsumoto *et al.*, 2017).

Ces découvertes soulèvent la question de l'efficacité de l'annotation automatique des génomes et transcriptomes. Ces cadres de lecture ouverts encodent des protéines initialement non prédites. Les manquements du processus d'annotation automatique ont considérablement retardé la découverte et caractérisation de ces protéines. Toutefois, compte tenu du nombre d'ORFs prédits au sein d'un transcrit, l'étude de leur expression s'avère être un processus complexe. De plus, les mécanismes de traduction des altORFs sont connus mais restent relativement peu étudiés.

2.3.4 Mécanismes de traduction des protéines alternatives

Hormis le modèle général de *scanning* du ribosome, divers mécanismes peuvent expliquer la traduction de différents ORFs au sein d'un transcrit.

Le premier mécanisme qui explique la traduction de plusieurs ORFs au sein d'un transcrit est la réinitiation ribosomale. Lors de ce processus, le ribosome traduit un cadre de lecture dans son intégralité. Une fois la traduction achevée, la grande sous-unité se détache alors que la petite sous-unité continue de scanner le transcrit. Si elle rencontre un codon d'initiation de la traduction, la grande sous unité du ribosome est recrutée à nouveau, et le ribosome reconstitué traduit une séquence codante au sein du transcrit (Kozak, 1987, 2001) (Figure 13 A). Ce mécanisme est toutefois modulé en fonction de la distance entre les deux ORFs, la présence de codon d'initiation "AUG" ainsi que la présence de structures en "épingle à cheveux" (Kozak, 1987, 2001).

Le deuxième mécanisme de traduction de différents cadres de lecture au sein d'un transcrit est appelé *leaky scanning*. Il survient lorsqu'un ribosome, qui lit un transcrit selon le modèle de *scanning*, n'initie pas au premier codon d'initiation qu'il rencontre. Il est alors possible qu'il initie la traduction à un codon d'initiation en aval et donc éventuellement un altORF (Kozak, 1995, 2002) (Figure 13 B).

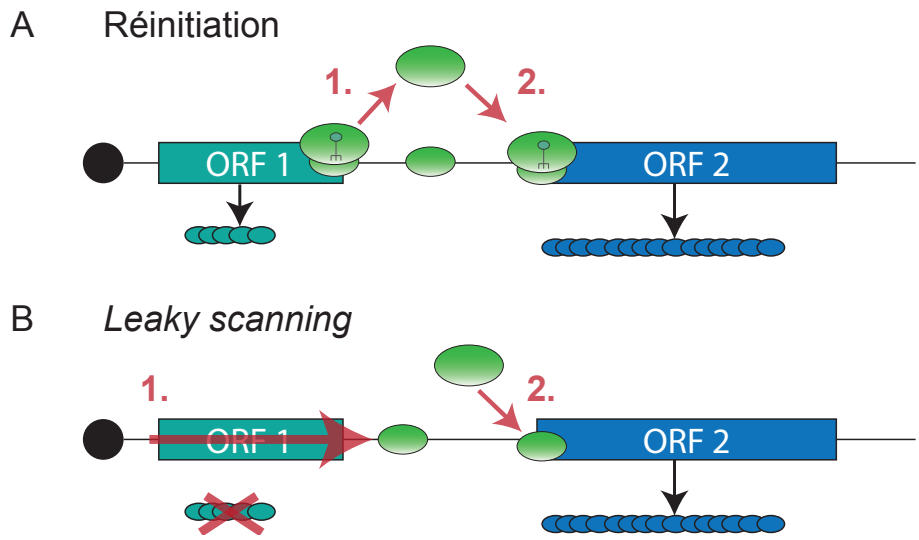


FIGURE 2.13 – **Mécanismes de traduction des altORFs**
A. Mécanisme de réinitiation ribosomale avec synthèse d'une première protéine issue de l'ORF1 suivi du détachement (1.) de la grande sous-unité et réinitiation (2.). **B.** Mécanisme de *leaky scanning* dans lequel le ribosome ne traduit pas le premier ORF (1.) et initie au début du deuxième ORF (2.).

Enfin, le troisième mécanisme qui explique l'expression de différents ORFs au sein d'un transcrit est la traduction coiffe indépendante lorsque le transcrit contient des séquences capables de recruter les complexes d'initiation de la traduction (Figure 5).

2.3.5 Etude fonctionnelle des protéines alternatives

Déterminer la fonction de protéines alternatives équivaut à caractériser de nouvelles protéines. Des outils de prédiction bioinformatiques de fonctions sont couramment utilisés pour reconnaître des séquences en acides aminés associées à des activités biologiques tels qu'InterPro (Mitchell *et al.*, 2014) ou Pfam (Finn *et al.*, 2016). Toutefois, les protéines alternatives ont relativement peu d'homologie de séquences avec les protéines de référence et présentent peu de domaines protéiques connus.

Une étape importante dans la caractérisation d'une nouvelle protéine est de valider l'expression dans un contexte transcriptionnel proche du contexte originel, c'est-à-dire de reproduire une copie d'un transcrit d'intérêt. Une étiquette moléculaire permettant la détection par les approches de biochimie classiques tels que l'immunobuvardage (ou *western-blot*, WB) ou l'immunofluorescence (IF) est habituellement ajoutée. Le WB valide le poids moléculaire de la protéine alternative alors que l'IF indique sa localisation cellulaire et permet d'émettre des hypothèses sur sa fonction biologique.

La technique d'interférence de transcrits par petits ARNs en épingle à cheveux (shRNA) ou interférants (siRNA) est une technique couramment utilisée pour évaluer l'impact d'une diminution d'expression d'une protéine. Elle est toutefois difficile à mettre en œuvre pour l'étude des protéines alternatives et particulièrement pour celles issues d'un transcrit qui code déjà pour une protéine de référence. En effet, leur utilisation diminuerait à la fois l'expression de la protéine de référence et de la protéine alternative (Delcourt *et al.*, 2017). Il faudra alors privilégier des approches d'édition du génome telles que le CRISPR-Cas9 pour inhiber l'expression d'une protéine sans affecter celle des autres (Ran *et al.*, 2013).

Réciproquement, les observations réalisées suite à l'utilisation de si- ou shRNA pour des protéines de référence dont le gène code également pour une protéine alternative ne sont pas exclusivement imputables à la protéine de référence. Le phénotype observé peut en effet résulter de la diminution des deux protéines.

Les observations réalisées suite à la surexpression hétérologue d'une protéine de référence soulèvent également des interrogations. Dans le cas où la séquence codante surexprimée contient un altORF dans un cadre de lecture décalé, certaines observations imputées à la protéine de référence pourraient en fait résulter de l'activité d'une protéine alternative issue de cet altORF, si cette dernière s'exprime effectivement (Delcourt *et al.*, 2017).

Les approches protéomiques telles que la purification par affinité aident également à déterminer l'activité d'une protéine alternative en établissant son interactome. Si les partenaires d'interaction de la protéine alternative sont impliqués dans une voie biologique particulière, on peut émettre l'hypothèse que la protéine alternative y participe également.

2.3.6 Détection à large échelle des protéines alternatives

Les protéines alternatives sont détectables à large échelle essentiellement par deux techniques : le profilage ribosomal et la protéogénomique par MS.

2.3.6.1 Le profilage ribosomal

Le profilage ribosomal consiste à séquencer les fragments de transcrits protégés par les ribosomes en phase d'initiation ou d'élongation de la traduction. Cette technique peut être employée sur des cellules en culture traitées avec des inhibiteurs de la traduction ou après extraction des ribosomes sans traitement préalable (Ingolia *et al.*, 2011; Brar et Weissman, 2015; Ingolia, 2016).

Profilage ribosomal pour la détermination des régions initiatrices de la traduction
Certaines substances sont connues pour inhiber la transition du ribosome entre les phases d'initiation et d'élongation. La harringtonine (Ingolia *et al.*, 2011), la lactimidomycine (Lee *et al.*, 2012) et, dans des conditions particulières, la puromycine (Fritsch *et al.*, 2012), sont capables de bloquer les ribosomes au site d'initiation avant la phase d'élongation de la traduction. L'utilisation de ces substances permet, dans la plupart des cas, de déterminer le site d'initiation avec une précision de l'ordre du nucléotide (Lee *et al.*, 2012). Cette approche distingue les ORFs qui déclenchent effectivement l'initiation de la traduction des autres ORFs. Elle indique également avec précision le codon d'initiation des protéines synthétisées, crucial pour déterminer les codons non canoniques (Ingolia *et al.*, 2011) (Figure 14, gauche).

Profilage ribosomal en phase d'élongation de la traduction
Cette technique détermine les séquences des transcrits en cours de traduction pendant le processus d'élongation sans nécessiter de traitement préalable (Ingolia, 2016) (Figure 14, droite).

Détermination des séquences protégées par les ribosomes
Après capture et purification des ribosomes, les échantillons sont traités avec une enzyme de type *nucléase*, capable de dégrader les transcrits. Après inactivation de la *nucléase*, le ribosome est dissocié et le fragment de transcrit protégé par le ribosome est isolé. Ces fragments sont ensuite amplifiés par transcription inverse et séquencés. Les séquences obtenues sont ensuite alignées sur le génome.

Pour l'étude des protéines alternatives, l'utilisation du profilage ribosomal en phase d'initiation est privilégiée. En effet, il peut être délicat de déterminer quel ORF est effectivement traduit lors de l'élongation pour les ORFs qui se chevauchent. Il est alors nécessaire d'utiliser des algorithmes particuliers qui étudient la périodicité des lectures de séquençage (Michel *et al.*, 2012; Calviello *et al.*, 2016) (Figure 14 C).

Ces approches se révèlent très efficaces pour déterminer les régions traduites du transcriptome. Les altORFs détectés dans les ARNm sont majoritairement localisés dans les régions 5'UTR et CDS (ORF canonique ou dans un cadre de lecture décalé) et relativement peu dans les régions 3'UTR. Toutefois, une approche récente a mis en évidence la présence de ribosomes dans les régions 3'UTR (Miettinen et Björklund, 2014).

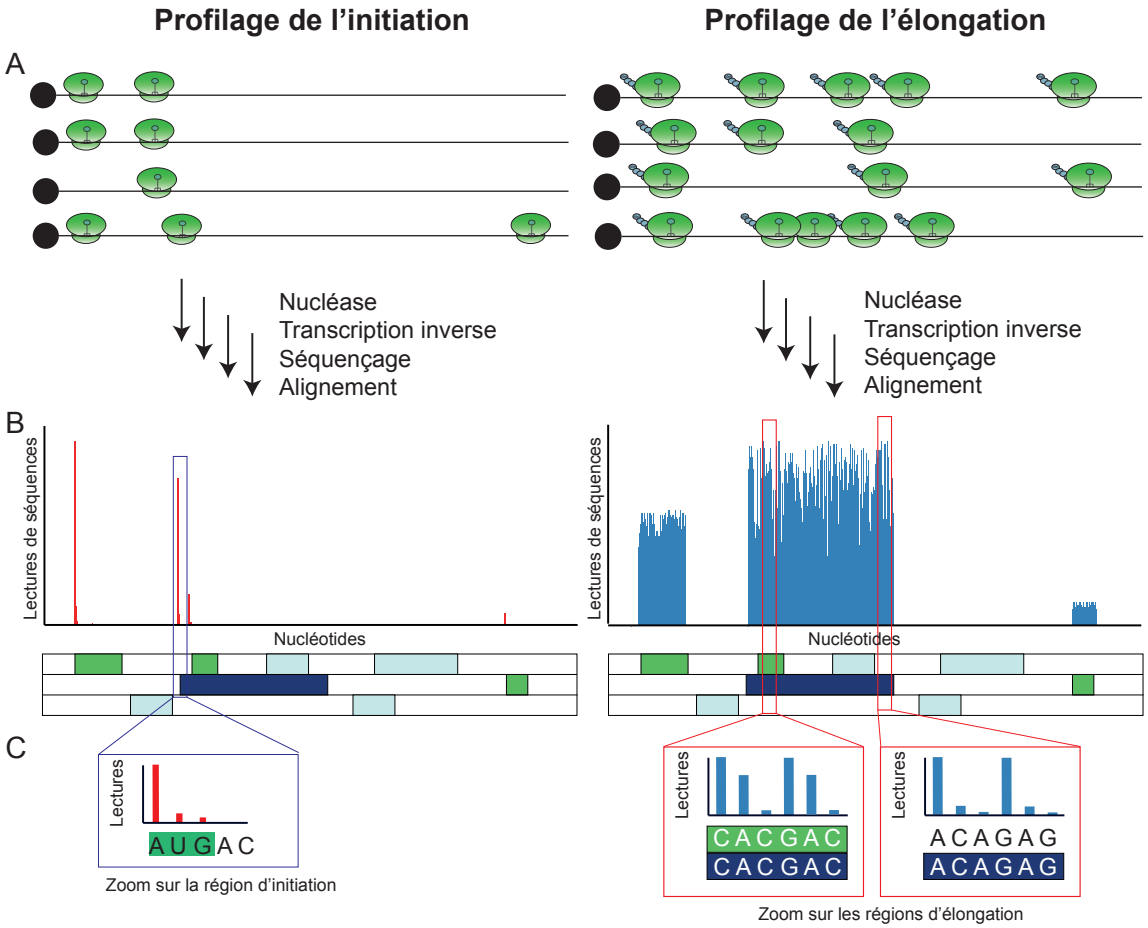


FIGURE 2.14 – **Stratégies de profilage ribosomal**
Représentation schématique des stratégies de profilage ribosomal d'initiation (gauche) ou d'élongation (droite). **A.** Collecte des ribosomes sur les transcrits en élongation ou en initiation suite à un traitement de quelques minutes à la harringtonine ou lactimidomycine. **B.** Histogrammes des lectures des différentes séquences obtenues après amplification et séquençage des fragments d'ARN protégés par les ribosomes. Les altORFs effectivement traduits sont marqués en vert et le CDS canonique en bleu. **C.** Analyse de la périodicité des codons pour déterminer l'ORF traduit, particulièrement utile en profilage d'élongation.

2.3.6.2 L'identification de protéines alternatives par spectrométrie de masse

La détection de protéines alternatives par MS est une technique développée récemment qui repose sur l'identification de leur séquence en acides aminés observée en MS. Bien que le profilage ribosomal ait mis en évidence la traduction de nombreux altORFs, leur expression n'avait pas encore été démontrée à large échelle par MS. C'est avec le développement des approches de protéogénomique que les protéine alternatives ont été observées.

La protéogénomique est la discipline qui définit les régions codantes du génome à partir

d'analyses protéomiques par spectrométrie de masse sans nécessairement prendre considérer les bases de données de protéines de référence. Elle permet ainsi d'étudier le protéome d'organismes pour lesquels les connaissances protéiques et/ou transcriptomiques sont faibles, notamment par l'emploi de bases de données issues de la traduction *in-silico* des six cadres de lecture de séquençage génomique partiel ou complet ou des trois cadres de lecture de données transcriptomiques (Armengaud *et al.*, 2014). La protéogénomique offre également la possibilité de caractériser des mutations, polymorphismes génétiques et ainsi défini les modifications du génome responsables du développement de pathologies cancéreuses (Zhang *et al.*, 2014; Mertins *et al.*, 2016; Zhang *et al.*, 2016a). Enfin, la protéogénomique permet de détecter de nouvelles régions codantes par la considération des régions présumées « non-codantes » du génome ou transcriptome lors de l'identification des spectres (Vanderperre *et al.*, 2013). Elle peut alors être associée au profilage ribosomal pour affiner les identifications ou offrir un mode de détection complémentaire. Il est toutefois nécessaire d'utiliser des approches statistiques adaptées à la protéogénomique afin de limiter le nombre de faux positifs (Nesvizhskii, 2014).

C'est en 2013 avec les publications de Slavoff *et al.* (2013) et Vanderperre *et al.* (2013) que la détection des protéines alternatives par MS a été confirmée. Depuis, d'autres publications ont corroboré ces observations (Ma *et al.*, 2014; Kim *et al.*, 2014; Samandi *et al.*, 2017).

L'identification des protéines alternatives nécessite toutefois des précautions particulières. En effet, l'utilisation d'une base de données contenant les prédictions de protéines alternatives augmente considérablement l'étendue des séquences protéiques. Lors de la validation statistique des spectres, l'identification d'un grand nombre de PSMs de protéines de référence induit un taux de faux positifs de PSMs alternatifs artificiellement bas. Afin de limiter les faux positifs, il est nécessaire d'utiliser des approches de FDR séparés (alternatif - référence) plutôt que combinés (Nesvizhskii, 2014; Menschaert et Fenyö, 2015).

Cette analyse a été réalisée lors de la préparation de la publication Samandi *et al.* (2017). Dans cette publication, nous avons réanalysé les données de protéomique de quatre études déjà publiées qui n'avaient été interrogées que pour les protéines de référence. Une proportion non négligeable de spectres non identifiés pour les protéines de référence restaient alors non assignés et pouvaient permettre l'identification de protéines alternatives. Dans un premier temps, nous avons appliqué une approche de FDR combinée puis évalué la différence de distribution de scores pour les protéines alternatives et les protéines de références. Une différence importante entre les groupes a été constatée. Nous avons alors appliqué une approche de FDR séparée afin de limiter le nombre de faux positifs potentiels. A la suite de

cette analyse, nous avons constaté qu'il existait toujours une différence importante entre les distributions des scores des PSMs des protéines de référence et des protéines alternatives (Annexe 1). Les protéines de référence sont en règle générale beaucoup plus grandes que les protéines alternatives. Elles sont donc plus susceptibles de générer des peptides et d'obtenir un score d'identification plus élevé.

En effet, si on compare les distributions des scores des PSMs de référence en fonction du nombre de peptides uniques, on constate que les protéines de référence identifiées avec un faible nombre de peptides uniques ont des distributions de scores significativement plus faibles (Annexe 2). De plus, les protéines de référence identifiées sur la base d'un seul peptide unique ont une taille significativement inférieure à celles identifiées avec au moins deux peptides uniques (Annexe 2).

Ce résultat montre que les protéines alternatives identifiées par approche de FDR séparée présentent des scores généralement plus faibles que les protéines de référence, mais cela est essentiellement dû à la longueur des protéines alternatives, et leur faible probabilité de générer plusieurs peptides uniques. Cet aspect démontre que l'identification de protéines alternative est complexe, même en employant des approches statistiques adaptées.

3 ARTICLE 1

Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA

Auteurs de l'article: Vivian Delcourt, Antanas Staskevicius, Michel Salzet, Isabelle Fournier, Xavier Roucou

Statut de l'article: Publié

Avant-propos: La complexité d'étude des petites protéines et la récente caractérisation de certaines d'entre elles a mené à la rédaction d'un article de revue. Ma contribution à cet article comprend une étude bibliographique des récentes découvertes relatives aux petites protéines, la création de figures ainsi que la participation à la rédaction de l'article sous la supervision de mes encadrants.

Résumé: Les peptides et petites protéines encodés à partir de courts ORFs chez les eucaryotes sont longtemps demeurés dans l'ombre des grandes protéines. Récemment, les progrès des techniques de protéomique par MS et de profilage ribosomal ont permis la détection de nombreuses petites protéines. La variété de leurs fonctions émerge également. Outre le défi technique que représente la détection des petites protéines, elles ont été largement ignorées par les annotations. N'étant pas recherchées, elles n'étaient pas détectées. Dans cette revue, nous identifions les conventions qui nécessitent une réévaluation, y compris la supposition qu'un ARNm ne contient qu'une séquence codante. La découverte à grande échelle des petites protéines et de leur fonction impliquera la remise en cause de certains paradigmes et la mise à jour de leur annotation, encore largement perçues comme information codante non pertinentes en opposition aux séquences codantes déjà annotées.

Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA

Vivian Delcourt, Antanas Staskevicius, Michel Salzet, Isabelle Fournier, Xavier Roucou

Journal : Proteomics
Editeur : Wiley

3.1 Manuscrit

3.1.1 Abstract

Short ORF-encoded peptides and small proteins in eukaryotes have been hiding in the shadow of large proteins for a long time. Recently, improved identifications in MS-based proteomics and ribosome profiling resulted in the detection of large numbers of small proteins. The variety of functions of small proteins is also emerging. It seems to be the right time to reflect on why small proteins remained invisible. In addition to the obvious technical challenge of detecting small proteins, they were mostly forgotten from annotations and they escaped detection because they were not sought. In this review, we identify conventions that need to be revisited, including the assumption that mature mRNAs carry only one coding sequence. The large-scale discovery of small proteins and of their functions will require changing some paradigms and undertaking the annotation of ORFs that are still largely perceived as irrelevant coding information compared to already annotated coding sequences.

3.1.2 Introduction

There is definitively some buzz around the largely unexplored territory of short ORFs, generally defined as ORFs below 100 codons, and their translation products in prokaryotes and in eukaryotes (Ramamurthi et Storz, 2014; Storz *et al.*, 2014; Su *et al.*, 2013; Chu *et al.*, 2015; Saghatelian et Couso, 2015; Andrews et Rothnagel, 2014; Landry *et al.*, 2015; Pueyo *et al.*, 2016a; Yang *et al.*, 2011; Staudt et Wenkel, 2011; Feller, 2012; Ericson *et al.*, 2014; Hellens *et al.*, 2016). Here, we will focus our discussion on eukaryotes, mainly mammals. Based on the weight of the size criterion in the way annotations are performed, annotated coding ORFs or protein-coding sequences (CDSs) are virtually always the longest ORFs within a coding gene or transcript (Goffeau *et al.*, 1996; Carninci *et al.*, 2005). Thus, unannotated ORFs within coding genes and transcripts are obligatory shor-

ter than CDSs, and although most of these ORFs are shorter than the conventional 100 codons cut-off used to qualify as short ORFs, a fraction of them are longer (Vanderperre *et al.*, 2013; Chalick *et al.*, 2016; Vanderperre *et al.*, 2011; Bergeron *et al.*, 2013; Raj *et al.*, 2016). We previously used the terms alternative ORFs or altORFs for all unannotated ORFs because they require alternative initiation compared to canonical translation initiation sites mapped to CDSs, a large fraction overlap annotated CDSs in an alternative reading frame, and they may generate alternative translation products different from annotated protein (Vanderperre *et al.*, 2013). In this review, we extend our discussion to unannotated ORFs and their translation products in general.

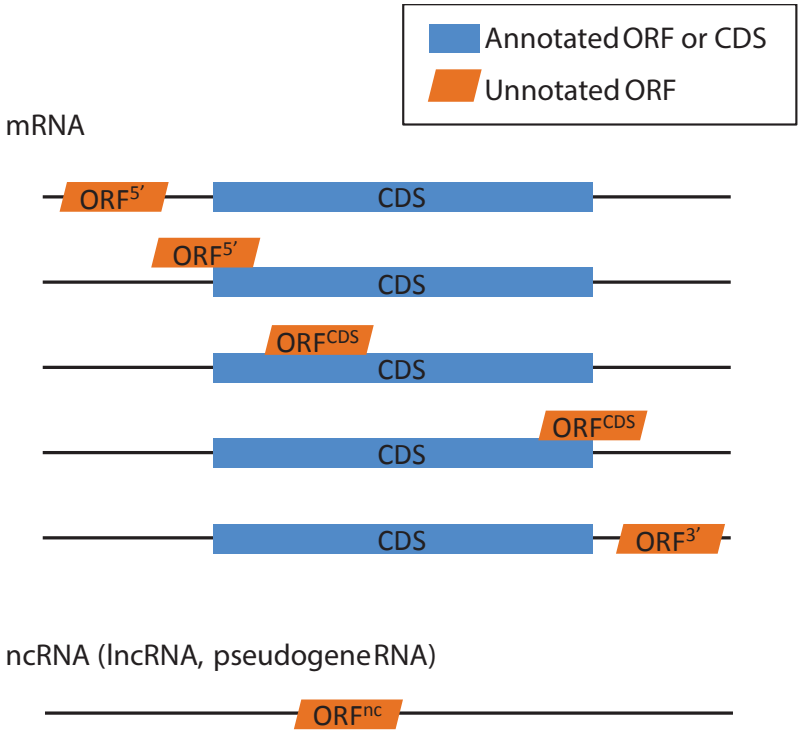


FIGURE 3.1 – **Unannotated ORFs may be found in mRNAs or non-coding RNAs.** Molecules shown here represent mature or processed RNAs. Unannotated ORFs are found in untranslated regions of the transcriptome. Within mRNAs, these regions include 5'UTRs, 3'UTRs, and alternative reading frames overlapping annotated CDSs. An $ORF^{5'}$ (also termed upstream ORF or uORF) may be found outside the CDS in the three reading frames, or partially overlapping a CDS in one of the two alternative reading frames. ORF^{CDS} may be found completely nested inside the CDS, or partially overlapping the 3'UTR in one of the two alternative reading frames . An $ORF^{3'}$ may be in one of the 3 reading frames. Non-coding RNAs do not have a tripartite structure similar to mRNAs, and an ORF^{nc} may be found anywhere.

Unannotated ORFs are generally classified in different groups (Figure 1). These ORFs,

particularly short ORFs, are disturbing. They violate some basic concepts in genome and transcriptome annotations, including the conventional 100 codons cut-off used to identify non-coding ORFs (Goffeau *et al.*, 1996; Carninci *et al.*, 2005; Dinger *et al.*, 2008; Ulitsky et Bartel, 2013) and the one mature mRNA - one protein paradigm (Mouilleron *et al.*, 2015). They have been hiding and remained undetected until large scale mapping of ribosome occupancy using ribosome profiling exposed unexpected large numbers of translated ORFs (Ingolia *et al.*, 2011; Lee *et al.*, 2012; Ingolia *et al.*, 2014; Fields *et al.*, 2015; Bazzini *et al.*, 2014; Crappé *et al.*, 2014; Calviello *et al.*, 2016; Brar et Weissman, 2015). Many small proteins translated from ORFs have now been detected by MS-based proteomics (Crappé *et al.*, 2014; Menschaert *et al.*, 2013; Koch *et al.*, 2014; Ma *et al.*, 2014). Although the physiological function of the majority of these novel small proteins is not known, some have important functions in muscle contraction (Magny *et al.*, 2013; Anderson *et al.*, 2015; Nelson *et al.*, 2016), development (Kondo *et al.*, 2007, 2010) and signaling (Chng *et al.*, 2013; Pauli *et al.*, 2014). Finally, the field is recent and it is difficult to predict the implications of the discovery of this new genetic information that has been hiding in genomes.

Here, we mainly discuss some elements of the modern view of protein translation that are definitively not compatible with the recent discoveries of small proteins, and we briefly present some strategies that are used to improve the exploration of this new territory of the proteomic world.

3.1.3 Basic concepts and corollaries of the modern view of the protein coding information

Annotations of protein-coding genes have greatly contributed to the development and the application of omics-based technologies in modern experimental biology and medicine. Each annotated gene, CDS, transcript and protein has a specific entry in databases that are routinely used for research, and these databases are indispensable in the emerging approach of precision medicine. For example, protein databases are central to the success of MS-based protein identification (Aebersold et Mann, 2003). Unfortunately, databases have a dark side ; they limit the scope of discoveries that can be made since any unannotated information cannot be detected in a biological sample. Here, we highlight some general concepts in annotations, some of which are clearly hampering the discovery of small proteins.

3.1.3.1 Coding and non-coding genes and transcripts

Several strategies are used for the annotation of coding sequences (CDSs), including CDS identification based on known protein sequences and ab initio predictions of the most likely CDS (Yandell et Ence, 2012; Mudge et Harrow, 2016). And typically, genes are annotated as coding or non-coding. For example, the Ensembl human gene annotation (Genome Reference Consortium Human Build 38) contains 20,310 coding genes and 37,118 non-coding genes (including 14,589 pseudogenes).

In eukaryotes, alternative splicing, alternative transcription and alternative polyadenylation are mechanisms that produce different RNA isoforms with possible variations in the CDS (de Klerk et 't Hoen, 2015). Therefore, there are more CDSs than protein-coding genes, and most genes encode multiples protein isoforms. CDSs are annotated in databases such as RefSeq (O'Leary *et al.*, 2015), Ensembl (Cunningham *et al.*, 2014) and the Consensus CDS (Farrell *et al.*, 2013). As of April 2017, the consensus CDS database contains 18,889 coding genes and 32,524 CDSs for *Homo sapiens*.

If a transcript generated from a protein-coding gene does not contain the annotated CDS or a variation of this CDS, or if it contains a retained intron, it is labeled as a transcript without ORF or non-coding. Thus, according to current annotations, a coding gene can generate both coding and non-coding RNAs. An example is illustrated in Figure 2 for the centromere protein A, CENPA. Transcripts 1 and 2 code for the canonical 140 amino acid CENPA protein and a shorter 114 amino acid isoform, respectively. Annotations indicate that the third transcript does not contain any protein-coding ORF. Thus, CENPA is annotated as a coding-gene expressing coding and non-coding transcripts. One could imagine that transcripts 1 and 2 are not expressed in a specific tissue; in that tissue, this gene would functionally be a non-coding gene. One could also imagine that in the same tissue, CENPA could be protein-coding in conditions where transcripts 1 and 2 are expressed and non-coding in conditions where transcript 3 only is expressed. This example illustrates how annotations simplify the information and may not reflect the biology.

The presence of unannotated ORFs adds another level of complexity. An overlapping ORF^{CDS} in the +2 reading frame is not annotated (Figure 2). Yet, ribosome profiling data detect initiating ribosomes with a number of reads similar to reads on the annotated CDS (Michel *et al.*, 2013). Thus, transcript 3 may be a coding transcript after all. Examples of transcripts for which annotation changed from non-coding to coding are shown in table 1.

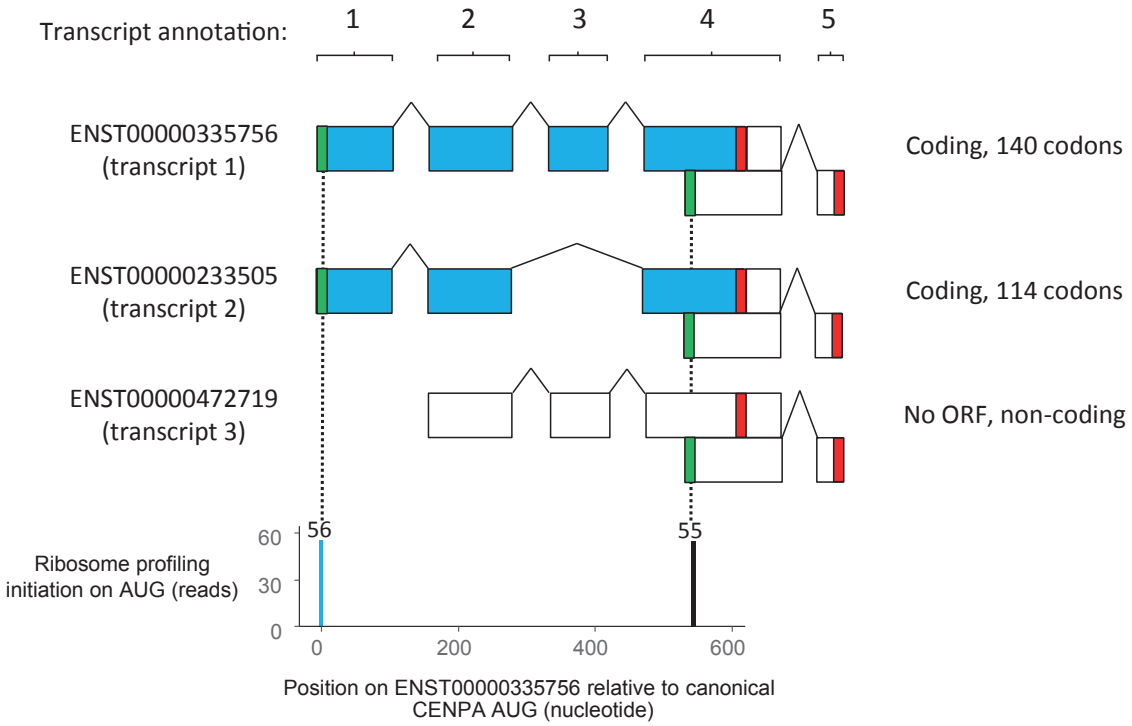


FIGURE 3.2 – **Typical Ensembl transcripts annotation for a human gene**
Three transcripts from the CENPA gene are illustrated. Sequences annotated as coding sequences are shown as blue boxes, and sequences annotated as untranslated regions are shown as open boxes. Green boxes represent AUG translation initiation sites. Red boxes represent stop codons. ENST00000335756 carries the canonical 420 bps CDS with 4 coding exons. The end of exon 4 after the stop codon is non-coding. ENST00000233505 is an isoform lacking exon 3 and carries a shorter version of the canonical CDS. ENST00000472719 is an isoform lacking exon 1. In the absence of the translation initiation site, the Ensembl annotation indicates the absence of a canonical CDS, and this transcript is believed to be non-coding. An unannotated overlapping long ORF of 53 codons in the +2 reading frame starts with an AUG codon within exon 4 and ends with a stop codon in exon 5. Translation initiation ribosome profiling data clearly show initiating ribosomes on both the annotated CDS and the unannotated ORF with similar number of reads (graph below). Thus, transcripts 1 and 2 might be bicistronic, and transcript 3 is likely a coding transcript for a novel small protein.

Tableau 3.1 – **The discovery of small proteins leads to changes in annotations**

Annotation[a]	Gene	Gene type	Ensembl Transcript	Transcript type	Protein (aa)
GRCh37	APELA / ELABELA (ENSG00000248329)	Non coding	ENST00000507152	lincRNA[b]	-
GRCh38		Coding		Coding	54
GRCh37	MRLN (ENSG00000227877)	Non-coding	ENST00000414264	lincRNA[b]	-
GRCh38		Coding		Coding	46
GRCh37	SLC35A4 (ENSG00000176087)	Coding	ENST00000323146[c]	Coding	324
GRCh38		Coding	ENST00000623481	Novel coding	103[d]

[a]Ensembl human genome assembly.

[b]lincRNA refers to long intergenic non-coding RNAs in Ensembl annotations

[c]Although this canonical transcript contains both the CDS and a short ORF encoding a novel 103 amino acids protein, and was proposed to be a bicistronic mRNA (Andreev *et al.*, 2015a), it is annotated as coding the 324 aa SLC35A4 protein only.

[d]This small protein is not an isoform, and SLC35A4 is a dual-coding gene.

3.1.3.2 One mRNA, one protein

Protein synthesis is one of the most complex and energetically demanding cellular processes (Rolfe et Brown, 1997). The eukaryotic ribosome is a molecular machine with four RNA molecules and at least 80 proteins. Ribosome biogenesis and assembly demands major cellular resources, including 200 assembly protein factors and 80 small nucleolar RNAs (Woolford et Baserga, 2013; Thomson *et al.*, 2013). Every stage of translation, initiation, elongation, termination and recycling is also regulated by a large number of specific factors. Overall, if ribosome biosynthesis is not taken into account, more than 100 protein-coding genes and 60 non-protein-coding genes are involved in translation.

The aim of this elaborate translation machinery is to interact with mRNAs and decode CDSs. In the current view of protein synthesis, mRNAs harbour a tripartite structure with a unique CDS and two untranslated regions (5'- and 3'-UTRs). Thus, although the life

of an mRNA is complex (Moore, 2005), its function would be to carry a single coding message, the annotated CDS.

3.1.3.3 The 100 codons cut-off

ORFs are expected to occur by chance in a long nucleotide sequence, and establishing a 100 codon cut-off has been one of the key criteria for the annotation of CDSs, and for the separation of mRNAs from non-coding RNAs (Dinger *et al.*, 2008; Ulitsky et Bartel, 2013). Although additional criteria are used to identify likely CDSs, including conservation and known protein domains, the 100 codons cut-off is the main bottleneck that has prevented short ORFs from being included in annotation databases.

3.1.3.4 Implications

These concepts have an important corollary : the complete set of the coding information is restricted to annotated CDSs, and the proteome is primarily determined according to this assumption. Thus, short ORFs are excluded from the transcriptome coding potential, and the predicted corresponding proteins are omitted from the reference proteome. Consequently, short ORFs and small proteins are not well represented in current databases. The size distribution of the current set of human consensus CDSs indicates that median and average lengths are 434 and 569 codons, respectively (Figure 3), confirming previous results obtained with the analysis of a collection of cDNAs (Frith *et al.*, 2006), GENCODE annotated CDSs (Raj *et al.*, 2016), and a protein database (Landry *et al.*, 2015).

The fraction of CDSs with a maximum length of 100 codons is only 2.5 %. Interestingly, this observation was used as evidence supporting the 100 codons cut-off to discriminate coding from non-coding RNAs (Dinger *et al.*, 2008; Ulitsky et Bartel, 2013). Our discussion below suggests that it is rather because the contribution of small proteins has been largely overlooked that they are few in databases.

Since the fraction of small proteins in databases is so low, there is a generally unrecognized perception that small proteins have less important functions than large proteins. This issue will not be addressed here but it is interesting to note small proteins of 100 amino acids or less are structural subunits or key regulators of two vital molecular machines, the FO-F1 ATP synthase and the sarcoplasmic reticulum Ca^{2+}-ATPase, respectively (table 2).

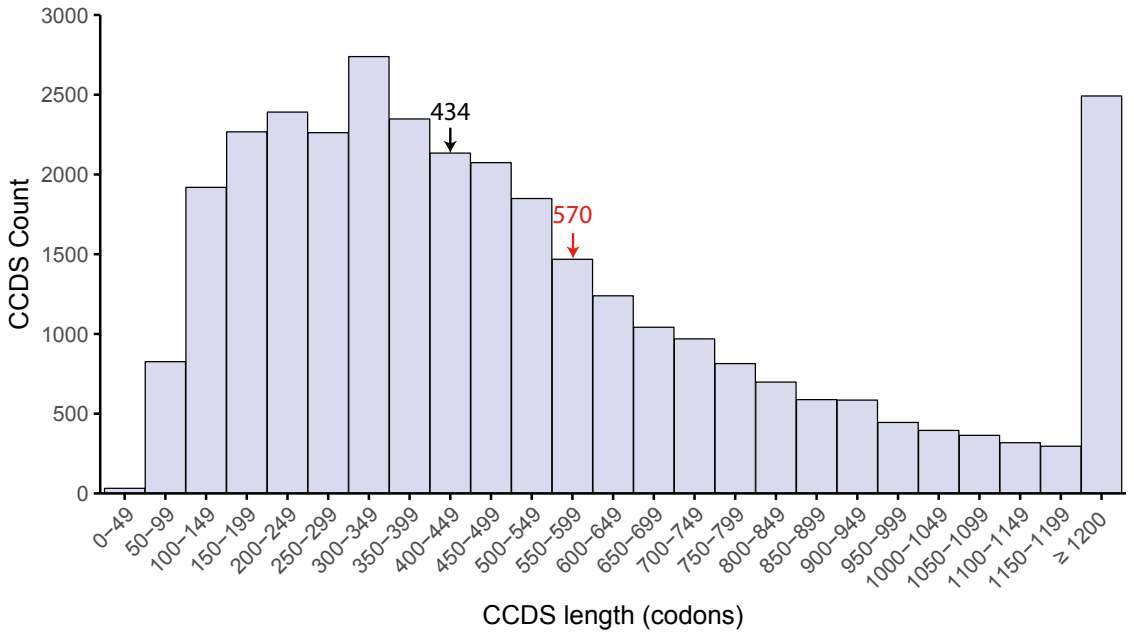


FIGURE 3.3 – **Distribution of the number and the size of human consensus CDSs** Calculations were performed with consensus CDS release 20. Median, 434 codons ; average, 570 codons.

Tableau 3.2 – **Small proteins and molecular machines**

Protein	Size (aa)	Description
ATP synthase subunit f	94	ATPsynthase, H^+ transporting, mitochondrial Fo complex subunit F2
ATP synthase subunit epsilon	51	ATP synthase, H^+ transporting, mitochondrial F1 complex, epsilon subunit
ATP synthase subunit e	69	ATP synthase, H^+ transporting, mitochondrial Fo complex, subunit E
ATP synthase subunit g 2	100	ATP synthase, H^+ transporting, mitochondrial Fo complex, subunit G2
Sarcolipin	31	Sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA), negative regulator
Phospholamban	52	Sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA), negative regulator

3.1.4 Translation outside of annotated CDSs : delinquent translation machinery or outdated concepts ?

Do ribosomes interact with transcripts annotated as mRNAs but not those annotated as non-coding ? Do ribosomes decode annotated CDSs only when they interact with mRNAs ? Do only annotated CDSs code for functional proteins ? These questions, which might have seemed senseless a few years ago, are now fundamental questions and the answers are no. A large number of studies clearly show that cells can decode both annotated CDSs and unannotated ORFs, and some databases now include information about ORFs with evidence of expression (Michel *et al.*, 2013; Wan et Qian, 2013; Crappé *et al.*, 2014; Olexiouk *et al.*, 2015; Hao *et al.*, 2017). Two databases are specific for short ORFs and proteins shorter than 100 amino acids in eukaryotic species (Olexiouk *et al.*, 2015; Hao *et al.*, 2017).

3.1.4.1 The role of serendipity in the discovery of overlapping ORFs

Since ORFs are not annotated and the modern view of an mRNA is still a transcript carrying a single annotated CDS, the discovery of small proteins translated from unconventional ORFs is often associated with serendipity (Ramamurthi et Storz, 2014; Storz *et al.*, 2014; Saghatelian et Couso, 2015; Mouilleron *et al.*, 2015). In my laboratory, we stumbled upon a first small protein following a combination of circumstances using a green fluorescent protein (GFP) tagged cytosolic form of the prion protein (CyPrP), CyPrP^{GFP} (Grenier *et al.*, 2006; Beaudoin *et al.*, 2009). In this construct, GFP is inserted into a natural restriction site within the unstructured N-terminal domain of CyPrP (Figure 4A). Cells transfected with CyPrP^{GFP} display RNA aggregates in the cytoplasm (Goggin *et al.*, 2008), and the expression of several RNA helicases, including DDX3 was tested by western blot using a commercial anti-DDX3 antibody, Ab37160 (Figure 4B). DDX3 was detected at the expected size in mock-transfected cells, but a second band labeled X appeared in cells expressing CyPrP^{GFP}. The hypothesis of a novel DDX3 isoform was unlikely based on the observation that three other antibodies directed against different regions of DDX3, Ab50703, NB-200-194 and NB200-195 did not detect protein X (Figure 4B). This experiment was later repeated with CyPrP^{GFP} C-terminal deletion mutants (Figure 4C). Strikingly, the electrophoretic mobility of protein X increased with the deletions. Eventually, we discovered that PrP CDS contains an ORF^{CDS} in the +3 reading frame (Vanderperre *et al.*, 2011) (Figure 4A, C). In our experiments, the GFP CDS was inserted inside the ORF^{CDS}. We realized that there are no stop codons in the +3 reading frame of GFP CDS, and that a chimeric small protein containing frameshifted GFP might be expressed. In a control experiment, transfected cells expressing frameshifted GFP displayed an epitope that was

detected with antibody Ab37160 (Figure 4D). Thus, protein X was a chimeric protein containing the Ab37160 epitope translated from an ORF^{CDS} in the prion protein CDS. If a commercial anti- DDX3 antibody different from Ab37160 had been used in the first place, the expression of this novel small protein would have gone unnoticed, and researchers expressing the prion protein in their experiments would not have known that both a large and a small protein are co-expressed in their experiments. These observations are proof of principle that ribosomes are able to translate two overlapping messages from the same transcript. As discussed below, many studies have confirmed this feature.

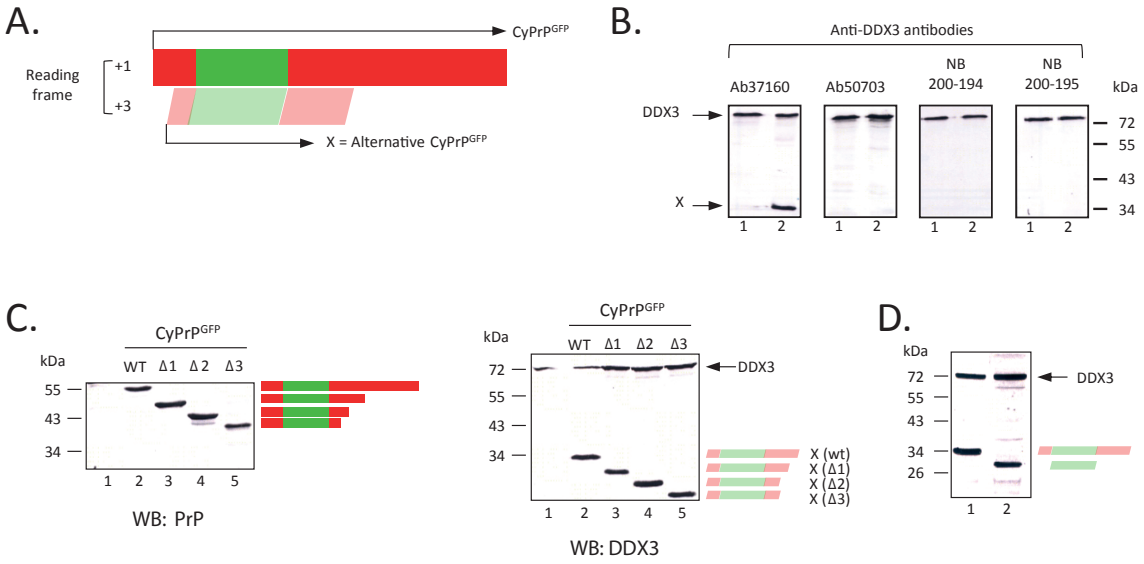


FIGURE 3.4 – **Translation of an overlapping ORF in the prion protein (PrP) CDS**
A. Diagram of CyPrP^{GFP} CDS (+1 reading frame) and frameshifted alternative CyPrP^{GFP} (+3 reading frame). GFP CDS is shown in green, CyPrP CDS is shown in red. **B.** Western blot analysis of mock-transfected HEK293 cells (lane 1) and CyPrP^{GFP}-transfected cells (lane 2) with four antibodies directed against DDX3, as indicated. DDX3 is indicated by an arrow above the 72 kDa marker. An unknown protein labeled X is detected in CyPrP^{GFP}-expressing cells with antibodies Ab37160. **C.** Western blot analysis of different CyPrP^{GFP} C-terminal deletion mutants with anti-PrP (left blot) or anti-DDX3 (right blot, Ab37160) antibodies. Lane 1 : Mock-transfected cells ; lane 2 : cells transfected with wild-type CyPrP^{GFP} ; lanes 3-5 : deletion mutants. A diagram of each construct is indicated on the right side of each blot. C-terminal deletions within CyPrP^{GFP} introduce C-terminal deletions within overlapping alternative CyPrP^{GFP}. **D.** Western blot analysis of cells transfected with CyPrP^{GFP} (lane 1) or frameshifted GFP (lane 2) with anti-DDX3 antibodies (Ab37160).

3.1.4.2 Overlapping viral ORFs decoded by the human translation machinery

Some of the earliest evidence that mammalian ribosomes may translate ORFs^{CDS} comes from viruses which use unusual strategies to optimize their coding capacity in small genomes. For space constraints, we provide below a few examples for human viruses only. The hepatitis C virus (HCV) combines the polyprotein and the overlapping strategies. The 10 canonical proteins of the hepatitis C virus are coded in a long CDS that is translated into a large polyprotein. Further proteolytic cleavage results in the production of 7 non-structural and 3 structural proteins. In addition to these proteins, additional smaller proteins encoded by CDSs overlapping the main CDS in a different reading frame are also produced. Such alternative reading frame proteins can be detected in patients developing specific antibodies (Morice *et al.*, 2009). One of these proteins, the alternate F protein seems to modulate the immune system (Park *et al.*, 2016; Samrat *et al.*, 2014; Xu *et al.*, 2014). West Nile Virus also generates alternative reading frame proteins (Faggioni *et al.*, 2012). Many other RNA viruses express multiple proteins translated from ORFs^{CDS} within their mRNAs (Firth et Brierley, 2012). Oncogenic viruses such as Kaposi's sarcoma-associated herpesvirus and Epstein-Barr virus also produce alternative reading frame proteins encoded in alternative CDSs overlapping the latency-associated nuclear antigen and Epstein-Barr nuclear antigen 1 CDSs, respectively (Kwun *et al.*, 2014). Whether these novel proteins have a function in the oncogenic activity of these viruses remains to be determined. Importantly, these observations indicate that small ORFs^{CDS} hiding in large annotated viral CDSs are translated in infected human cells. Ribosome profiling confirmed the translation of a large number of viral annotated CDSs and short ORFs^{CDS} in human cells infected with cytomegalovirus and Kaposi's sarcoma-associated herpesvirus (Stern-Ginossar et Ingolia, 2015). In this study, a fraction of the small proteins were also confirmed by MS-based proteomics.

3.1.4.3 Overlapping ORFs (ORFs^{CDS}) in mammals

There is accumulating evidence of functional ORFs^{CDS} in mammalian mRNAs (Vanderperre *et al.*, 2011, 2013; Bergeron *et al.*, 2013; Nekrutenko *et al.*, 2005; Klemke *et al.*, 2001; Lee *et al.*, 2014; Michel *et al.*, 2012), and computational approaches predict a large number of ORFs^{CDS} in human and murine cDNAs and transcripts (Chung *et al.*, 2007; Ribrioux *et al.*, 2008; Xu *et al.*, 2010b). Several cDNAs containing the annotated CDS and an ORFs^{CDS} ligated into expression vectors generate a small protein in addition to the larger annotated proteins in cultured cells (Vanderperre *et al.*, 2011, 2013; Bergeron *et al.*, 2013; Nekrutenko *et al.*, 2005; Klemke *et al.*, 2001; Lee *et al.*, 2014) and in vivo (Li *et al.*, 2009;

Kracht *et al.*, 2017). These remarkable observations confirm a counterintuitive feature of mammalian ribosomes which are able to decode ORFs^{CDS} in addition to annotated CDSs in the same transcripts. The translation of both the CDS and the ORFs^{CDS}s can be easily tested. Generally, an epitope tag fused in-frame with a CDS facilitates the detection of a protein of interest by western blotting or immunofluorescence (Figure 5). However, this approach does not allow the detection of a small protein co-expressed from an ORF^{CDS}. Addition of a second epitope in-frame with the ORF^{CDS} makes the small protein detectable (Figure 5). In other words, what you see is what you've tagged ; you won't see what you haven't tagged.

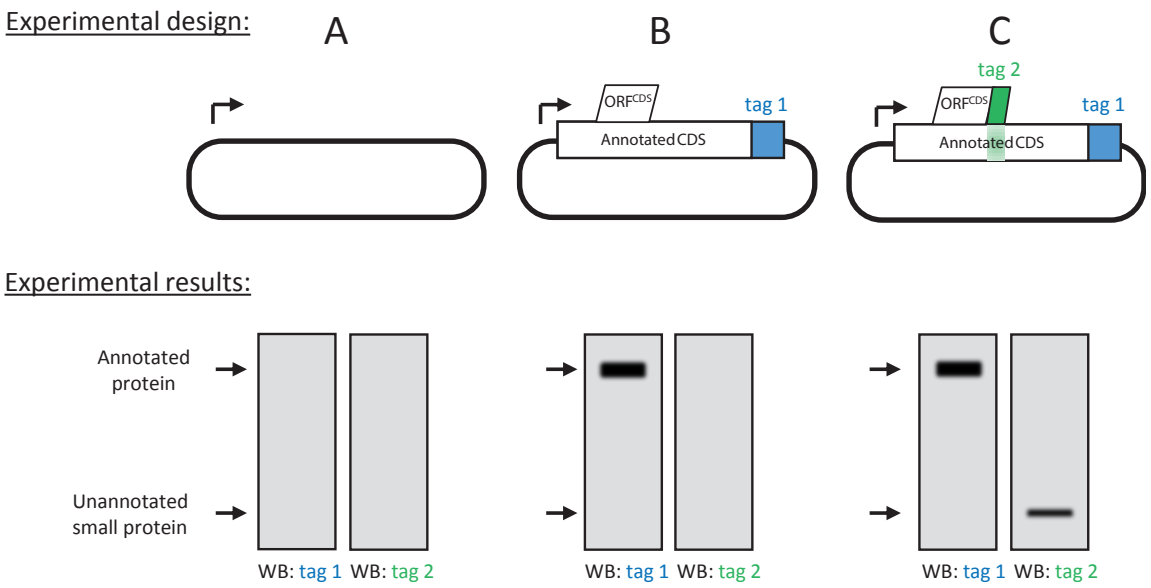


FIGURE 3.5 – **A simple double epitope tagging strategy for the detection of ORFs^{CDS}** Upper panel : experimental design. (A) Empty vector. (B) Expression plasmid containing an epitope tagged CDS (tag1). (C) Expression plasmid containing both an epitope tagged CDS (tag 1) and an epitope tagged ORF^{CDS} (tag 2). Tag 2 is in-frame with the ORF^{CDS}, but out-of-frame with the CDS. Tag 2 should not introduce a stop codon in the CDS frame. Bottom panel : after transfection and expression, cell lysates are analyzed by western blot with anti-tag 1 and anti-tag 2 antibodies. (A) No signals are detected in mock-transfected cells. (B) Anti-tag 1 antibodies detect the expression of the protein of interest. A second protein expressed from the ORF^{CDS} remain invisible. (C) Anti-tag 2 antibodies detect the expression of the unannotated small protein.

3.1.4.4 ORFs in regions annotated as untranslated : mRNA UTRs, long non-coding RNAs (lncRNAs), and pseudogenes RNAs

A combination of computational and experimental approaches provide strong evidence for the translation of ORFs present in regions classified as untranslated, from yeast to human (Vanderperre *et al.*, 2013; Ingolia *et al.*, 2014; Bazzini *et al.*, 2014; Andreev *et al.*,

2015a; Prabakaran *et al.*, 2014; Smith *et al.*, 2014; Ji *et al.*, 2015; Mackowiak *et al.*, 2015; Ruiz-Orera *et al.*, 2014; Aspden *et al.*, 2014; Popa *et al.*, 2016). These unannotated ORFs include ORFs$^{5'}$ upstream of the CDS, also termed upstream ORFs (Wethmar *et al.*, 2013), ORFs$^{3'}$ downstream of the CDS, and ORFs^{nc} in long non-coding RNA and pseudogenes transcripts (Figure 1).

A number of studies demonstrate function in development (Kondo *et al.*, 2010, 2007; Galindo *et al.*, 2007), DNA repair (Slavoff *et al.*, 2014), muscle contraction (Anderson *et al.*, 2015; Magny *et al.*, 2013; Nelson *et al.*, 2016), cell signaling (Pauli *et al.*, 2014; Chng *et al.*, 2013; Matsumoto *et al.*, 2017), mRNA decapping (D'Lima *et al.*, 2017) and phagocytosis (Pueyo *et al.*, 2016b). As a consequence of these discoveries, an update is performed in databases and the annotation of some genes and their transcripts switches from non-coding to coding (table 1). Once in the databases, the small proteins are officially in the reference proteome and can be detected in routine MS-based proteomics experiments.

3.1.4.5 Mechanisms for the translation of ORFs

Similar to mRNAs, many lncRNAs are transcribed by RNA polymerase II and are modified with a 5'cap structure and a 3' poly A tail (Quinn et Chang, 2016). From the ribosome perspective, there is no specific mechanism requirement for the translation of ORFs^{nc}. Possible mechanisms for the translation of several ORFs within the same transcript have already been reviewed (Mouilleron *et al.*, 2015) and will not be detailed here. Briefly, these mechanisms can be separated into two classes. In the first class, leaky scanning and translation reinitiation are compatible with the widely accepted ribosome scanning model for translation initiation (Kozak, 2002; Hinnebusch *et al.*, 2016). The second class includes the recruitment of ribosomes on translation initiation sites by tethering, cap-assisted internal initiation and RNA looping (Chappell *et al.*, 2006; Martin *et al.*, 2011; Paek *et al.*, 2015).

The regulation of the translation of ORFs is still unknown but, some insights into the translation of ORFs$^{5'}$ have recently been published. In contrast to annotated CDSs, translation of ORFs$^{5'}$ is resistant to oxidative stress (Andreev *et al.*, 2015a), and glucose and oxygen deprivation (Andreev *et al.*, 2015b). Unconventional translation of ORFs$^{5'}$ mediated by the alternative initiation factor eiF2A seems to be crucial in tumor initiation (Sendoel *et al.*, 2017).

3.1.5 Taking on the challenges of unifying the nomenclature, annotating, detecting and deciphering the function of currently unannotated ORFs and proteins

3.1.5.1 Challenge 1 : the nomenclature

The heterogeneity in the nomenclature of short ORFs and their translation products in the literature is a good indication that this rapidly expanding research domain is relatively new. ORFs below 100 codons are termed short ORFs (sORFs) (Slavoff *et al.*, 2013), small ORFs (smORFs) (Basrai *et al.*, 1997; Saghatelian et Couso, 2015; Bazzini *et al.*, 2014; Olexiouk *et al.*, 2015; Smith *et al.*, 2014; Mackowiak *et al.*, 2015; Aspden *et al.*, 2014; Zanet *et al.*, 2015), or upstream ORFs (uORFs) when they are located in 5'UTRs. The cutoff may reach 150 codons in some studies (Slavoff *et al.*, 2013). Their translation products are labeled peptides (Pueyo *et al.*, 2016a; Magny *et al.*, 2013; Kondo *et al.*, 2010, 2007; Ruiz-Orera *et al.*, 2014; Galindo *et al.*, 2007), small proteins (Ramamurthi et Storz, 2014; Storz *et al.*, 2014; Yang *et al.*, 2011; Cabrera-Quio *et al.*, 2016), sORF-encoded peptides or SEP (Ma *et al.*, 2014; Slavoff *et al.*, 2014, 2013), micropeptides (Crappé *et al.*, 2013), and microproteins (D'Lima *et al.*, 2017). We have termed unannotated ORFs and proteins, alternative ORFs and alternative proteins, respectively to highlight the fact that they represent different ORFs and translation products compared to current annotations (Vanderperre *et al.*, 2013, 2011; Bergeron *et al.*, 2013; Mouilleron *et al.*, 2015).

In the short term, it will be useful to introduce a unified nomenclature to avoid confusion and to better organize the field of small proteins.

3.1.5.2 Challenge 2 : the detection

There is no straightforward protocol to detect the short proteome, and it is not as simple as "just mass spec your protein gel dye fronts to death" (Feller, 2012). Recent computational and biochemical advances, and reduction in the size cutoff for potential coding-ORFs are helping the discovery of small proteins. The different approaches include ribosome profiling (Raj *et al.*, 2016; Lee *et al.*, 2012; Ingolia *et al.*, 2014; Bazzini *et al.*, 2014; Calviello *et al.*, 2016; Michel *et al.*, 2013, 2012; Wan et Qian, 2013; Olexiouk *et al.*, 2015; Ji *et al.*, 2015; Crappé *et al.*, 2013; Michel et Baranov, 2013) and proteogenomics (Vanderperre *et al.*, 2013; Crappé *et al.*, 2014; Menschaert *et al.*, 2013; Koch *et al.*, 2014; Ma *et al.*, 2014; Menschaert *et al.*, 2013; Olexiouk et Menschaert, 2016; Ma *et al.*, 2016). Proteogenomics involve the extraction of ORFs from genomic or transcriptomic data and the generation of custom-made protein databases for MS-based proteomics identification (Nesvizhskii, 2014).

3.1.5.3 Challenge 3 : the function

There have been recent successes in the discovery of the function of small proteins (Anderson *et al.*, 2015; Magny *et al.*, 2013; Nelson *et al.*, 2016; Kondo *et al.*, 2010; Pauli *et al.*, 2014; Chng *et al.*, 2013; Slavoff *et al.*, 2014; Matsumoto *et al.*, 2017; D'Lima *et al.*, 2017; Pueyo *et al.*, 2016b). Yet, undertaking the functional analyses of small proteins is a multilevel challenge. First, the selection of an unannotated ORF for functional investigation is tricky. ORFs are more likely to occur by chance in the genome (Fickett, 1995; Basrai *et al.*, 1997). Computational tools to find homologs were developed with large proteins but may not perform as well on short ORFs (Cheng *et al.*, 2011). Protein domains such as those annotated by the InterPro database were determined with annotated and mostly large proteins (Mitchell *et al.*, 2014), and the majority of small proteins are unlikely to contain these conventional domains. Second, it is technically challenging to work with small proteins. For example, the use of tags with small proteins is always a concern as it can modify some biochemical features and interfere with normal localization (Viallet et Vo-Dinh, 2003). Third, knockdown experiments to validate function are not possible with standard small interfering or short hairpin RNAs (siRNAs or shRNAs respectively) approaches when the small protein under investigation is encoded in an already annotated mRNA. siRNAs or shRNAs would reduce the expression of both the small and the annotated large proteins at the same time. Here, gene editing methods are required to remove an initiation site or introduce a stop codon to specifically stop the expression of the short ORF. Gene editing may be particularly tricky for ORFs^{CDS} because it would be preferable not to modify the sequence of the annotated protein.

3.1.6 Conclusion and perspective

Only long CDSs and large proteins made it through into current annotation databases routinely used to detect the coding genome and the proteome. Yet, there is strong evidence from ribosome profiling and proteogenomics that there are within our cells new genetic information and novel proteins the vast majority of which remains invisible, similar to a genomic and a proteomic dark matter. The exploration of this dark matter requires questioning the assumptions on which the annotations were based, including (1) that mRNAs carry a single CDS because only one CDS is annotated ; and (2) that RNAs annotated as non-coding are necessarily non-coding. Just as annotations are not rigid and change according to new experimental evidence, assumptions should not be interpreted as absolute dogmas either.

The extent of the proteomic dark matter made of small proteins is difficult to assess. Floo-

ding current databases used for MS-based proteomics with all possible ORFs is not an option ; the majority of short ORFs may represent expected biological noise (Landry *et al.*, 2015), and inserting great numbers of irrelevant ORFs would result in large databases that cause challenges for peptide identification (Nesvizhskii, 2014). The trend in recent years has been to create customized databases based on a proteogenomic approach. Parallel to the identification of small proteins with this approach, it will be important to facilitate access to and update a database containing the sequences of all detected small proteins. This will enable the MS-based proteomics community to validate and detect the expression of small proteins in routine experiments, an important advance towards the democratization and functional characterization of the small proteome.

3.1.7 Acknowledgements

This research was supported by CIHR grants MOP-137056 and MOP-136962 to X.R, and a Canada Research Chair in Functional Proteomics and Discovery of New Proteins to X.R. X.R is member of the Fonds de Recherche du Québec Santé-supported Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke. We thank the staff from the Centre for Computational Science at the Université de Sherbrooke, Compute Canada and Compute Québec for access to the Mammouth supercomputer.

3.2 Conclusion et perspectives

Cette revue bibliographique est publiée dans l'édition spéciale *Special Issue on short ORF-encoded peptides* du journal *Proteomics* (Wiley).

Les peptides et protéines encodées par les petits ORFs sont depuis peu décrits pour leur expression et parfois leur fonction. Cette part du protéome, jusqu'alors isolée et inexplorée, fut révélée par la technique du profilage ribosomal et l'observation de ribosomes en cours de traduction au sein d'ORFs non-codantes. Plus récemment, les techniques de protéomique par spectrométrie de masse ont également confirmé ce constat. Ces techniques démontrent que le protéome des eucaryotes a été sous-évalué et qu'une part significative de protéines et donc de fonctions biologiques restent à découvrir. Du fait de sa description récente, le protéome issu des ORFs reste largement méconnu de la communauté scientifique.

Dans cette revue, nous référençons les récentes découvertes concernant les protéines encodées à partir de petits ORFs et manqués par les processus d'annotation du génome. Après un rappel des règles qui régissent l'annotation des génomes et leur assimilation dans les diverses bases de données, nous énonçons les questions qui sont soulevées suite aux découvertes d'ORFs codant au sein d'ARNncs et de séquences codantes différentes des CDS au sein d'ARNms. En effet, il est nécessaire de s'interroger sur l'efficacité du processus d'annotation et de ses limites. Ce processus peut constituer un frein à la découverte de protéines d'importance capitale et de fonctions biologiques nouvelles. De plus, la découverte de ces petites protéines permet d'affiner les connaissances des processus biologiques, aspect particulièrement important avec l'avènement de la médecine de précision et l'émergence de la médecine personnalisée.

Ces découvertes entrainent des modifications dans les annotations du génome au sein des différentes bases de données, soit par l'annotation de régions non-codantes en régions codantes, soit par le référencement d'un nouveau gène codant ou par l'ajout d'une nouvelle région codante dans un gène. L'annotation des génomes est donc dynamique et est corrigée suite à la description de ces nouvelles régions codantes. Ces changements sont possible grâce à la remise en question des dogmes communément admis lors de l'annotation, à savoir qu'un ARNm ne contient qu'un ORF codant, et que cet ORF doit être d'une taille supérieure à 100 codons. Nous rapportons également qu'il est nécessaire de s'interroger sur le potentiel codant de transcrits épissés d'ARNms ne contenant pas de CDS canoniques.

Aussi, nous exposons les expériences qui ont entrainé la découverte d'altPrP, l'une des

premières protéines alternatives décrite. C'est en effet grâce à des expériences conduites dans une optique complètement différente et l'emploi d'un anticorps commercial ciblant DDX3 qui détecte un épitope a un poids moléculaire inattendu lorsque CyPrP^{GFP} est surexprimé que cette découverte fut possible. Après avoir écarté l'hypothèse d'une isoforme courte de DDX3 et réalisé diverses expériences de biologie moléculaire que cette protéine fut identifiée. En effet, la protéine fluorescente GFP possède, dans un cadre de lecture décalé de son CDS, l'épitope reconnu par l'anticorps anti-DDX3. L'emploi d'autres anticorps commerciaux dirigés contre des régions différentes de la protéine DDX3 n'aurait pas révélé l'expression d'une protéine alternative. Ainsi d'autres découvertes n'auraient sans doute pas pu être réalisées.

Après avoir décrit les différents mécanismes qui peuvent expliquer l'expression de petites protéines, nous avons démontré, à l'aide de divers exemples recensés à travers de nombreuses études, que les dogmes régissant l'annotation des génomes doivent être réévalués. En effet, la machinerie ribosomale, la traduction de protéines et l'étendue du protéome s'avèrent être sous-évalués. Le ribosome ne sélectionne pas strictement une séquence codante à traduire sur la base de sa conservation, sa longueur ou la présence d'un large domaine protéique. De petites protéines sont effectivement traduites à partir de petits ORFs au sein d'ARNm ou ARNnc. Aussi, la taille d'une protéine n'est pas associée avec l'absence ou la moindre importance de sa fonction. En effet, de nombreuses protéines occupent des fonctions capitales voire essentielles.

Il est également nécessaire d'établir une nomenclature pour ces protéines. En effet, du fait de son expansion récente et des multiples subdivisions qui existe au sein de cette discipline, beaucoup de termes différents sont employés pour décrire des choses semblables. L'information scientifique s'en retrouve diluée, ce qui peut constituer un frein à son expansion et sa diffusion. Aussi, l'étude de petites protéines au sens large implique des défis techniques. En effet, le vaste pan des techniques de biochimie a été développé et optimisé sur l'hypothèse que le protéome fonctionnel était plus grand que cent acides aminés. Des techniques de biochimie courantes comme le *western-blot* ou l'immunofluorescence sont généralement délicates à mettre en œuvre pour l'étude de petites protéines et nécessitent parfois l'emploi de protocoles adaptés. Enfin nous rappelons que la taille d'une protéine n'est pas un critère limitant pour sa fonctionnalité. Toutefois, la détermination de la fonction d'une protéine constitue un défi important.

Dans cette revue nous mentionnons que ces petites protéines sont notamment découvertes grâce aux progrès récents réalisés en protéomique par MS. En effet, grâce aux avancées techniques et méthodologiques récentes en instrumentation et bioinformatique, il est dé-

sormais possible d'identifier plusieurs milliers de protéines à partir de divers échantillons biologiques. Cependant, les bases de données employées pour l'identification de protéines ne tiennent généralement pas compte des protéines encodées dans les altORFs. Les approches de protéogénomique et l'emploi de bases de données modifiées contenant les traductions *in silico* des trois cadres de lecture des transcrits (ou les six cadres du génome) ont permis de détecter un grand nombre de protéines alternatives et donc de nouvelles régions codantes. En effet, dans chaque expérience de protéomique par MS, un grand nombre de spectres de haute qualité spectrale restent sans assignation. Ceux-ci peuvent provenir de peptides modifiés chimiquement ou post-traductionnellement, de séquences protéiques comprenant des polymorphismes ou des mutations génétiques, mais aussi de nouvelles protéines encodées à partir d'altORFs.

Ces stratégies d'identification sont très fréquemment associées à l'approche de protéomique par MS dite "*bottom-up*". Cependant, cette approche connait des faiblesses et particulièrement pour la caractérisation de petites protéines. En effet l'emploi d'enzymes protéolytiques pour la génération de peptides est un frein important pour la détection de petites protéines et de protéines alternatives. Celles-ci contiennent moins de sites de coupure, sont donc moins susceptibles de générer des peptides uniques et donc plus difficile à identifier. De plus, l'emploi d'une enzyme protéolytique entraîne la génération de plusieurs dizaines voire centaine de milliers de peptides issus de protéines de référence. De ce fait, les probabilités d'identifier une protéine alternative diminuent drastiquement tant par leurs faibles prédispositions à générer des peptides que par leur dilution parmi un grand nombre de peptides de protéines de référence.

L'emploi de l'approche *top-down* émergente pourrait constituer une solution appropriée pour l'identification de petites protéines et de protéines alternatives. En effet, cette approche permet l'identification de protéines intactes sans employer d'enzyme protéolytique. De plus, les analyseurs de masse sont plus efficaces pour déterminer la masse d'un composé de faible poids moléculaire. Ils sont donc prédisposés à détecter des petites protéines. Enfin, l'application de l'approche *top-down* n'a pas encore été évaluée pour la détection de protéines alternatives. Par sa complémentarité, il serait donc intéressant de tester ses capacités pour l'identification de petites protéines en général et plus particulièrement de protéines alternatives.

4 ARTICLE 2

Top-down microproteomics bridged to MALDI MS imaging reveals the molecular physiome of brain regions

Auteurs de l'article: Vivian Delcourt, Julien Franck, Jusal Quanico, Jean-Pascal Gimeno, Maxence Wisztorski, Firas Kobeissy, Xavier Roucou, Michel Salzet, Isabelle Fournier

Statut de l'article: En révisions

Avant-propos: Dans cet article, nous nous sommes appliqués à développer une approche de microprotéomique par *top-down* MS à partir de régions localisées de cerveau de rat, servant de modèle de référence. Dans cet article, j'ai réalisé les expériences et analyses d'imagerie de métabolites permettant de définir les trois régions d'intérêt à analyser. J'ai par la suite, réalisé les extractions localisées de ces trois régions par deux approches de microprotéomique, l'extraction par microjonction liquide et la microdissection manuelle assistée par parafilm. Après avoir paramétré le spectromètre de masse pour l'acquisition de spectres de protéines intactes, j'ai analysé les données d'identification par différents programme bio-informatiques. Certaines expériences, notamment les expériences d'imagerie de protéines par MS pour la corrélation ainsi que les analyses de signalisation cellulaire, ont été menées par des collaborateurs et encadrants. J'ai également participé à la rédaction de l'article et à la construction des figures et tableaux et rédigé la réponse aux reviewers avec l'aide des collaborateurs et de mes encadrants.

Résumé: L'analyse microprotéomique de tissus par approche *top-down* à été réalisée sur trois régions de cerveau de rat, permettant l'identification de 123 protéines de référence. De plus, 8 nouvelles protéines alternatives issues de cadres de lecture alternatifs ont été identifiées. Certaines protéines exposent des modifications post-traductionnelles ou des troncations liées aux régions du cerveau où elles ont été identifiées et à leur fonction. Les protéines identifiées dans les trois régions ont été corrélées avec des expériences d'imagerie de protéines par MALDI. Par exemple, le fragment C-terminal de la protéine α-synucléine (95-140) clivé par la métalopeptidase de matrice 3 a été identifié avec une localisation spécifique dans le gyrus denté de l'hippocampe. Dans l'ensemble, nous avons mis en évidence une partie du protéome physiologique des trois régions de cerveau par la

caractérisation du protéome de référence et du protéome caché.

Top-down microproteomics bridged to MALDI MS imaging reveals the molecular physiome of brain regions

Vivian Delcourt, Julien Franck, Jusal Quanico, Jean-Pascal Gimeno, Maxence Wisztorski, Firas Kobeissy, Xavier Roucou, Michel Salzet, Isabelle Fournier

Journal : Molecular and Cellular Proteomics
Editeur : American Society for Biochemistry and Molecular Biology

4.1 Manuscrit

4.1.1 Abstract

Tissue top-down microproteomics was performed on 3 brain regions, leading to the characterization of 123 reference proteins. Moreover, 8 alternative proteins from alternative open reading frames (AltORF) were identified. Some proteins display specific post-translational modification profiles or truncation linked to the brain regions and their functions. Systems biology analysis performed on the microproteome identified in each region allowed to associate sub-networks with the functional physiology of each brain region. Back correlation of the identified proteins from the microproteome with tissue localization was then performed by MALDI mass spectrometry imaging. As an example, mapping of the distribution of the matrix metallopeptidase 3-cleaved C-terminal fragment of α-synuclein (aa 95-140) identified its specific distribution along the hippocampal dentate gyrus. Taken together, we established the molecular physiome of 3 rat brain regions through reference and hidden proteome characterization.

4.1.2 Introduction

On-tissue microproteomics provides a direct means to examine proteomic fluctuations at the cellular level in response to changes in the tissue microenvironment (Quanico *et al.*, 2013). Its importance is evident in physiopathological diseases such as cancer, where proteomic analysis of the complete tissue does not take into account tumor heterogeneity and thus the cellular cross-talks occurring in different regions of the tumor (Viale *et al.*, 2016; Massard *et al.*, 2016; Krönig *et al.*, 2015; Johann Jr *et al.*, 2009; Sugihara *et al.*, 2013; Celis *et al.*, 2002). Combined with MALDI mass spectrometry imaging (MSI) which can map the distribution of molecules (Bonnel *et al.*, 2011; Bruand *et al.*, 2011), on-tissue microproteomics can provide details of the molecular events occurring at cellular level in such discrete regions. In this context, our team made an ongoing effort to develop microscale techniques that can achieve reliable identification by shot-gun

proteomics and quantification of proteins within an area of the most limited size, and correlate these expression changes with alterations in cell phenotypes and/or biological state (Quanico *et al.*, 2013; Wisztorski *et al.*, 2016, 2013).

Liquid microjunction (LMJ) microextraction was the first technique developed for this purpose (Wisztorski *et al.*, 2016; Walworth *et al.*, 2010, 2011; Van Berkel et Kertesz, 2013, 2009; Franck *et al.*, 2013; Kertesz *et al.*, 2015; Kertesz et Van Berkel, 2014, 2013, 2010; Emory *et al.*, 2010; ElNaggar *et al.*, 2011). LMJ is the application of a droplet (1-2 μL) of solvent on top of a locally digested area, in order to extract peptides after on-tissue trypsin digestion. About 1500 protein groups from a tissue area of about 650 μm in diameter corresponding to less than 1900 cells can be identified (Quanico *et al.*, 2013). A method providing automatic microextraction and injection into the nanoLC-MS instrument from a tissue surface for shotgun microproteomics was also implemented. Thus an online LMJ coupling to on-tissue digestion using automatic microspotting of the digestion enzyme allows the analysis of a very limited area of the tissue section down to 250 μm spot size (corresponding to an equivalent average number of 300 cells) (Quanico *et al.*, 2016a). Application to ovarian cancer resulted in the identification of 1148 protein groups (Wisztorski *et al.*, 2013).

Parafilm-Assisted Microdissection (PAM) consists of mounting the tissue on a glass slide covered with a stretched layer of Parafilm M™ (Franck *et al.*, 2013; Zimmerman *et al.*, 2011; Quanico *et al.*, 2015). Regions of interest previously highlighted by MALDI-MSI are then manually microdissected. The microdissected areas are then submitted to in-solution digestion and nanoLC-MS/MS, allowing the identification and relative quantification of a large number of proteins (Franck *et al.*, 2013). Application to prostate cancer biomarker discovery led to the identification of 1251 proteins, 485 of which fit the Fisher's test criterion. 135 were upregulated and 73 downregulated in 8 prostate cancer biopsies (Quanico *et al.*, 2015).

All these strategies based on bottom-up proteomics remain limited as it is difficult to determine whether the protein is in its native or truncated form. Also, there is no direct information about post-translational modifications (PTMs) which often require specific enrichment steps. The top-down proteomics approach gives a unique solution for intact protein characterization with applications to monoclonal antibody characterization; de-novo sequencing and PTM elucidation without any conventional PTM-specific enrichment usually applied for bottom-up strategies and has already proven disease-monitoring capabilities for various pathologies (Nicolardi *et al.*, 2015; Kou *et al.*, 2016; Fellers *et al.*,

2015; Birner-Gruenberger et Breinbauer, 2015; Tran *et al.*, 2011; Liu *et al.*, 2013). However, this approach usually needs large amounts of protein samples and extensive fractionation techniques to be competitive with conventional bottom-up strategies in terms of unique protein IDs, mostly due to the need for accumulation of more microscans required for intact protein MS and MS/MS to generate spectra suitable for analysis. The molecular weight distribution tends to be restricted to lower molecular weight products as it remains challenging for the mass analyzer to measure the exact mass of high molecular weight compounds. Currently, top-down proteomics gives great opportunities for the better understanding of biological mechanisms and has been used complementary to bottom-up proteomics to gain information about PTMs, intact molecular weight and truncated forms of proteins, all of which can be critical for biomarker hunting. However, its association with tissue MALDI imaging and clinical investigations remains rare but promising (Ye *et al.*, 2014; Laouirem *et al.*, 2014). Notably, one study involving on-tissue extraction and direct infusion of protein extracts permitted the detection of a specific proteoform in non-alcoholic steatohepatitis patient tissues which could not be reliably identified by the bottom-up approach, showing great promises for disease characterization (Ye *et al.*, 2014; Laouirem *et al.*, 2014).

Recently, it has been shown that the proteome of higher mammals might has been under evaluated. We recently demonstrated the presence of several proteins issued from a mature mRNA which is normally assumed to contain a single Coding DNA Sequence (CDS). These proteins, so-called alternative proteins, are issued from alternative open reading frames (altORFs) and correspond to the hidden proteome (Vanderperre *et al.*, 2013). AltORFs are defined as potential protein-coding ORFs exterior to, or in different reading frames from, annotated CDSs in mRNAs and ncRNAs. Indeed, proteins translated from non-annotated altORFs were detected in several studies by MS (Vanderperre *et al.*, 2013; Mouilleron *et al.*, 2015). AltORFs are present in untranslated mRNA regions (UTRs) or overlap canonical or reference ORFs (refORFs) in a different reading frame. Thus, alternative proteins are not identical to reference proteins (Vanderperre *et al.*, 2013; Mouilleron *et al.*, 2015). For example, AltMRVI1, an alternative protein of the MRVI1 gene present in the 3'UTR region of the MRVI1 mRNA, has been shown to interact with BRCA1 (Vanderperre *et al.*, 2013). Translation of altORFs in human mRNAs in addition to refORFs provides access to a large set of novel proteins whose functions have not been characterized, and which cannot be detected using conventional protein databases. Moreover, conventional bottom-up proteomics is not well suited for their analysis because these proteins are relatively small (between 2 and 20 kDa) and more often do not contain

enzyme-cleavable sites. Thus, the number of enzymatically cleaved peptides generated is too small compared to those of reference proteins. Consequently, the probability of peptide and protein identification is poor, in the absence of low-mass protein enrichment steps. In this context, top-down proteomics offers better capabilities to detect alternative proteins, considering that no enzymatic digestion steps are used and this strategy is well suited to low-mass proteins.

In this article, further investigation of the hidden proteome on biological tissues was done. For this purpose, we developed a novel strategy based on MALDI MSI coupled to on-tissue top-down microproteomics to identify low-mass proteins and to localize them. We performed our analyses on rat brain to compare the reference proteome and the hidden proteome in different regions. Differential distributions of unique and common biological and functional pathways among the three different regions were then determined. A direct link can be drawn between the classes of proteins identified and the biological functions associated with each specific brain region. Interestingly, we identified different large peptide fragments from either neuropeptide precursors or from constitutive synapse proteins. These large peptides are different in each brain region and are in line with the presence of specific endocrine processing enzymes like prohormone convertases (Zheng *et al.*, 1997), neutral endopeptidases (Walther *et al.*, 2009) or angiotensin converting enzymes (Saavedra *et al.*, 1982; Harmer *et al.*, 2002).

We also showed the presence of specific PTMs associated to each brain region and in relation with their local function. Moreover, we demonstrated the presence of novel proteins issued from alternative ORFs and specific for each brain region. Finally, we performed back correlation between the identified and quantified microproteome and cellular localization with MALDI MSI. Taken together, we could depict a molecular proteomic pattern in three different rat brain regions in relation with the biological and physiological functions of each specific brain area.

4.1.3 Experimental Procedures

4.1.3.1 Experimental Design and Statistical Rationale

We first acquired MS images of lipids. These images were subjected to spatial segmentation to identify regions of interest (ROIs) that can be subjected to LMJ or PAM microproteomics. For this purpose, several tissue sections were obtained from rat brain. LMJ and PAM were followed by top-down proteomics for protein identification from 3 different

brain regions. Back correlation by MALDI MSI was then performed (n=3). Reference and alternative proteins were thus identified and localized in the 3 rat brain regions.

4.1.3.2 Chemicals

MS grade water (H_2O), acetonitrile (ACN), methanol (MeOH), ethanol (EtOH) and chloroform were purchased from Biosolve (Dieuze, France). The cleavable detergent ProteaseMAX was purchased from Promega (Charbonnieres, France). Parafilm M, 2,5- dihydroxybenzoic acid (DHB), sinapinic acid (SA), α-cyano-4-hydroxycinnamic acid (HCCA), aniline, sodium dodecyl sulfate (SDS), DL-dithiothreitol (DTT), trifluoroacetic acid (TFA) and formic acid (FA) were purchased from Sigma (Saint-Quentin Fallavier, France).

4.1.3.3 Tissues

Male Wistar rats of adult age were sacrificed by CO_2 asphyxiation and dissected. Brain tissues were frozen in isopentane at -50°C and stored at -80°C until use.

4.1.3.4 Tissue section preparation

For MALDI-MSI experiments, tissues were cut in 10 μm slices using a cryostat (Leica Microsystems, Nanterre, France) and were mounted on Indium Tin Oxide (ITO) coated glass slides (LaserBio Labs, Sophia-Antipolis, France) by finger-thawing. For LMJ and PAM, MSI-adjacent tissue slices were cut at 30 μm thickness. For LMJ, the tissues were mounted on polylysine glass slides (Thermo Fisher Scientific, Courtaboeuf, France) whereas for PAM, the tissues were mounted on Parafilm M-covered polylysine glass slides (Franck *et al.*, 2013). After tissue section preparation, the slides were immediately dehydrated under vacuum at room temperature for 20 minutes. The slides were then scanned and stored at - 80 °C until use.

4.1.3.5 MALDI-MSI

DHB matrix (50mg/mL in 6 :4 v/v MeOH/0.1% TFA in water) was then manually sprayed using a syringe pump connected to an electrospray nebulizer at a flow rate of 300 μL/h under nitrogen gas flow. The nebulizer was moved uniformly across the entire tissue until crystallization was sufficient to ensure optimal lipid detection. The tissue was then analyzed using an UltraFlex II MALDI-TOF/TOF mass spectrometer equipped with a Smartbeam Nd-YAG 355nm laser and controlled by FlexControl software (Bruker Daltonics, Bremen, Germany). Acquisition was performed in positive reflector mode with an m/z range of 50 to 900 and a spatial resolution of 300 μm. Each image pixel was obtained by

averaging 300 laser shots at a rate of 200 Hz. External calibration was performed using the Peptide calibration standard mix 6 (LaserBio Labs). Lipid ion distributions were generated using FlexImaging software version 3.0 (Bruker Daltonics).

For intact protein imaging, SA and HCCA liquid ionic matrices were used. These were prepared by dissolving the matrices in 7 :3 v/v ACN/0.1% TFA in water containing 7.2 μL aniline at a concentration of 15 and 10 mg/mL, respectively. The matrices were deposited on the tissue sections using ImagePrep (Bruker Daltonics). Images were acquired using the UltraFlex II instrument in positive linear mode with an m/z range of 3000-25000 and 2000-25000, respectively, at 50 μm resolution with the laser size set using "Medium" setting. Each image pixel was obtained by accumulating 500 laser shots at a rate of 200 Hz. External calibration was performed using the Protein Calibration standard I (Bruker Daltonics).

Peak detection and spatial segmentation analysis were then performed using SCiLS Software (SCiLS GmbH, Bremen, Germany) by applying the Bisecting k-Means with Correlation Distance approach. Segmentation was made with median normalization and medium denoizing. After analysis, the ROIs were determined by selecting regions where the correlation distances were significantly distant from one another.

4.1.3.6 Tissue Immunofluorescence

Immunofluorescence was performed on 10μm sagittal rat brain sections. The sections were immersed in blocking buffer (PBS 1x containing 1% bovine serum albumin, 1% ovalbumin, 2% Triton, 1% NDS, and 0.1M Glycine) for 1 hour. The primary antibody monoclonal mouse Anti-GFAP (1 : 500, Millipore, Molsheim, France) was diluted with the blocking buffer and applied to the sections except for the negative control where only the blocking buffer was applied. The sections were then incubated overnight at 4°C. The following day, the sections were washed three times with PBS 1x, and incubated for 1h at 37°C with the secondary antibody Alexa fluor donkey anti-mouse (1 :1000, Life Technologies, ThermoFisher Scientific, Courtaboeuf, France) diluted in blocking buffer without 0.1 M glycine. Afterwards, the sections were further washed with several changes of PBS 1x, stained with Sudan black 0.3% for 10 min in order to decrease the background generated by lipids, and were eventually counterstained with Hoechst solution (1 : 10,000). The slides were then washed with PBS 1x, and Dako fluorescent mounting medium was applied on the sections before putting cover slips. Confocal images were obtained using a confocal microscope (Leica Biosystems, Nussloch, Germany).

4.1.3.7 Intact protein extraction buffer

To ensure little-to-no protein hydrolysis by endogenous proteases, every step from buffer preparation to nanoLC-MS/MS analysis were carried out within the same day with on-ice conservation in between sample processing steps. A 1% solution of temperature- and acid-cleavable commercial detergent (ProteaseMAX) was prepared in 50 μM DTT and was aliquoted and immediately stored at -20 °C until use according to manufacturer's recommendations. The aliquots were processed the same day of sample extraction to ensure minimal degradation of the detergent over time. An aliquot was further diluted in ice-cold 50 μM DTT to obtain a final detergent concentration of 0.1% and stored on ice until use. Each aliquot was used within the day without conservation of the remaining solution.

4.1.3.8 LMJ experiments

To ensure optimal protein extraction, lipids were depleted from the tissue section by immersing the glass slides in consecutive solvent baths consisting of 70% and 95 % EtOH (1 min each time) and chloroform (30 s) with complete solvent evaporation under reduced pressure at room temperature between each washing step. The slides were then re-scanned to obtain better optical images with better contrast as the washing steps improve the visibility of the structures on the tissue section. The tissue slide for LMJ extraction was placed on a TriVersa NanoMate instrument (Advion, Ithaca, NY, USA). Proteins were then extracted from every ROI by completing six cycles of extraction consisting of pipetting up 1.5 μL of detergent solution, dispensing 0.8 μL of extraction buffer on the surface of the selected ROI with 10 iterations of up-and-down pipetting, aspiration of 2.5 μL by the device and expulsion of 4 μL from the pipette tip into a clean tube to ensure complete retrieval of the initial 1.5 μL volume for each cycle. Per ROI, the final collected volume was 9 μL ; the extracts were immediately placed on ice until further processing.

4.1.3.9 PAM experiments

10 μL of extraction buffer was transferred into a tube. Selected ROIs were manually dissected using a clean scalpel blade and transferred into the protein extraction buffer. Excision of the ROIs was performed with the aid of a microscope. The samples were placed on ice until further processing.

4.1.3.10 nanoLC-MS/MS

The extracts obtained using either the LMJ or PAM approaches were sonicated for 5 minutes and incubated at 55 ˚C for 15 minutes to ensure reduction of disulfide bonds.

These were then quickly centrifuged to rally condensation droplets at the bottom of the tube. The parafilm pieces were then carefully removed from the tubes using a pipette tip and the tubes were then heated at 95˚C for 10 minutes to ensure complete detergent dissociation. The tubes were then quickly centrifuged and placed on ice. 11 μL of 10 % ACN in 0.4 % FA in water were added to each tube to obtain a final ACN concentration similar to initial LC gradient conditions and the samples were stored at 4 ˚C until nanoLC-MS/MS analysis on the same day. 5 μL of each sample was loaded onto a 2 cm$\times$150 μm internal diameter (i.d.) PLRP-S (Varian, Palo Alto, California, USA) IntegraFrit sample trap-column (New Objective, Woburn, Massachussetts, USA) at a maximum pressure of 280 bar using a Proxeon EASY nLC-II chromatographic system (Proxeon, Thermo Scientific, Bremen, Germany). Proteins were separated on a 15 cm$\times$100 μm diameter i.d. PLRP-S column with a linear gradient of ACN from 5 to 100% and a flow rate of 300 nL/min. 10 μL of the samples were also injected and separated using a 3-h gradient.

Data were acquired on a Q-Exactive mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoESI source (Proxeon, Thermo Fisher Scientific, Bremen, Germany). 1.6 kV was applied on the PicoTip nanospray emitter (New Objective, Woburn, Massachussetts, USA) and the spectra were acquired in data-dependent mode using a top 3 strategy. Full scans were acquired by averaging 4 microscans at 70,000 resolution (at m/z 400) within a m/z range of 800 – 2000 with an AGC target of 1×10^6 and a maximum accumulation time of 200 ms. The three most abundant ions with charge states superior than +3 or unassigned were selected for fragmentation. Precursors were selected within an m/z selection window of 15 by the quadrupole and fragmented by averaging two microscans at a resolution of 70,000 with a Normalized Collision Energy (NCE) of 25. The AGC target was set to 1×10^6 with a maximum accumulation time of 500 ms. Dynamic exclusion was set to 20 s.

4.1.3.11 Data analysis

RAW files were processed with ProSight PC 3.0 or 4.0 (Thermo Fisher Scientifique, Bremen Germany). Spectral data were deisotoped using the cRAWler algorithm and searched against the complex Rattus norvegicus ProSightPC database version 2014_07. Using a similar approach, a second search was performed to detect alternative protein products, by interrogating RAW files with a concatenated custom database containing every reference proteins and their isoforms. These were generated from an in-silico transcriptome-wide translated database that contains every possible reference and alternative protein products from the Ensembl Rnor 6.0 transcripts sequence database with at least 30 amino acids

(Vanderperre *et al.*, 2013). For alternative protein identification, it was verified that the ID was coming from a specific precursor that was not identified during the reference protein search. Files were searched using a two-step search tree containing a 1-kDa precursor tolerant search ("Absolute") and a "Biomarker" search and MS/MS spectra were matched with sequences within a 15 ppm mass tolerance. Proteins were considered identified when one of the two steps gave expected values (E-value) inferior to 1×10^{-4}.

Likewise, data from PMID 27512083 (Quanico *et al.*, 2016b) were interrogated using the same search strategy with the concatenated database to identify alternative proteins which were not interrogated in the original publication. As ProsightPC's "Absolute" search mode adds multiple identifications for a single spectrum, output files were filtered using a custom R script. For each identified spectrum, 1) the one with the best E-Value and (2) identification that had the closest experimental mass compared to ProsightPC database was selected, which were concatenated in a single table. In this table, the ProsightPC PTMs were considered true if this PTM matches both its theoretical and experimental masses. On the other hand, mass shifts that matched known shifts were annotated accordingly (e.g. +80 for phosphorylation, +42 for acetylation) while undescribed shifts were automatically marked as unmodified (Supplementary data 1). Finally, a non-redundant identification file was generated (Supplementary data 2) containing information about identifications, methods, ROIs, found modifications, E-values, best P-score and spectral-count.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno *et al.*, 2015) partner repository with the dataset identifier PXD005424.

4.1.3.12 Subnetwork Enrichment Pathway Analyses and Statistical Testing

Elsevier's Pathway Studio version 10.0 (Ariadne Genomics/Elsevier) was used to deduce relationships among differentially expressed proteomics protein candidates using the Ariadne ResNet database (Bonnet *et al.*, 2009; Yuryev *et al.*, 2009). "Subnetwork Enrichment Analysis" (SNEA) algorithm was selected to extract statistically significant altered biological and functional pathways pertaining to each identified set of protein hits among the different groups. SNEA utilizes Fisher's statistical test set to determine if there are non-random associations between two categorical variables organized by specific relationships. Integrated Venn diagram analysis was performed using "the InteractiVenn" : a web-based tool for the analysis of complex data sets (Heberle *et al.*, 2015). See Supplementary Data

3 & 4 for the listed differential pathways.

4.1.4 Results

4.1.4.1 Top-down microproteomics and MALDI-MSI

Different types of molecules can be used in MALDI MSI to determine ROIs from biological tissues such as lipids, endogenous or tryptic peptides and proteins. However, lipid MALDI-MSI is the most convenient to our approach as it gives good spatial resolution and does not need extensive sample preparation steps. Our first developments were performed on rat brain tissue sections (Figure 1).

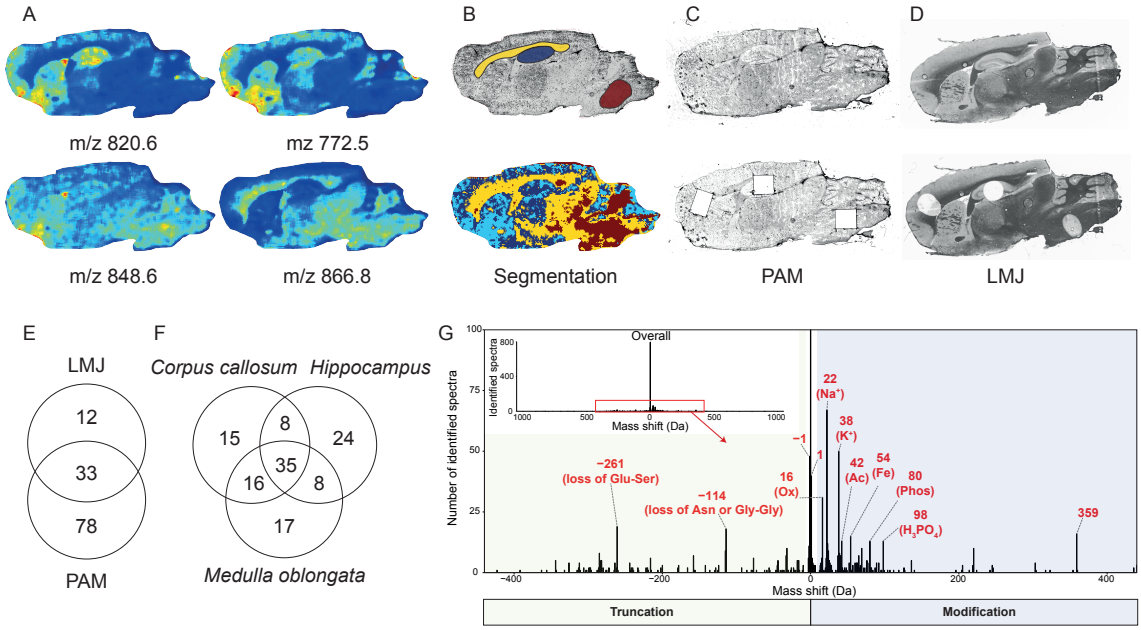


FIGURE 4.1 – **Top-down proteomics bridged to MALDI-MSI**
(A) Molecular images after median normalization of spectra followed by medium denoizing and automatic hotspot removal. (B) Optical image with highlighted regions of interest *Corpus callosum* (yellow), *Hippocampus* (blue) and *Medulla oblongata* (brown) (top) and spatial segmentation analysis using the Bisecting k-Means approach using Correlation as the distance metric (bottom). (C-D) Optical images of PAM and LMJ tissue sections with the top and bottom panels showing the tissue sections before and after ROI processing, respectively. (E) Venn Diagram of the extracted proteins per technique (LMJ or PAM) and (F) global unique identifications using both strategies. (G) Overall mass shifts of observed proteins precursors versus their theoretical masses (G, inset) and most abundant observed mass shifts within a ± 400 Da tolerance window (G) with annotation of known mass shifts. -114 Da corresponds to loss of “Asn” at N-term of ATP synthase-coupling factor 6, mitochondrial or loss of “Gly-Gly” at C-term of Ubiquitin monomer and -261 corresponds to loss of Glu-Ser at C-term of Thymosin beta-4

Different ROIs can be retrieved after lipid MALDI-MSI (Figure 1A) followed by non-supervised spatial segmentation analysis (Figure 1B, bottom) compared to the optical image (Figure 1B, top). Three ROIs in the *Hippocampus*, *Corpus callosum* and *Medulla oblongata* (Bregma Index lateral 1.90 mm) were selected for further processing as their segmentation profiles were sufficiently distinct.

Based on these selected ROIs, the two main strategies in order to perform micro-proteomics studies were then realized i.e., PAM (Figure 1C) or LMJ (Figure 1D). Based on the identified proteins, our approach mostly enables identification of low molecular weight (from 1.6 to 21.9 kDa) and most abundant proteins. These two strategies allowed the identification of proteins that were common within the three regions as well as specific ones. Analyses of the three ROIs gave a total of 123 proteins identified (Figure 1E & F, Supplementary data 1 & 2). 111 proteins have been identified in PAM and 45 in LMJ. The number of specific proteins identified is higher with PAM than with LMJ which might be related to tissue washing steps prior to protein extraction and smaller area of extraction. By combining the two approaches, 15 specific non-redundant proteins were identified from the *Corpus callosum*, 17 from *Medulla oblongata*, and 24 from *Hippocampus* (Figure 1E & F, Supplementary data 1 and 2 1 and 2). 35 are common to the 3 brain regions ; 16 are shared between *Corpus callosum* and *Medulla oblongata*, 8 between *Corpus callosum* and *Hippocampus*, and 8 between *Medulla oblongata* and *Hippocampus*. The majority of identified spectra exhibited a mass shift close to 0 Da (Figure 1G, inset). The mass tolerant identification approach allowed characterization of modified forms of proteins which can either be truncated compared to database prediction or modified (Figure 1G) in a similar fashion to what is described by Chick *et al.* (2015).

4.1.4.2 Systems biology analyses of the identified proteins

Functional enrichment analysis using Search Tool for Recurring Instances of Neighbouring Genes (STRING, Szklarczyk *et al.* (2014)) identified 4 GO terms associated with Molecular function : Hydrogen ions transmembrane transport (GO 0015078), Cytochrome-c oxidase activity (GO : 0004129), Ion transmembrane transporter activity (GO : 0015075), and Oxidoreductase activity (GO : 0016491). Systems biology analysis was then performed on the over-expressed proteins of each group for LMJ (Figure 2A) and for PAM (Figure 2C). Differential distributions of unique and common statistically significant biological and functional pathways among the three different regions are depicted in Figure 2A for LMJ and 2C for PAM, including 39 vs 18 pathways for *Corpus callosum*, 91 vs 34 pathways for *Medulla oblongata* and 31 vs 82 pathways for *Hippocampus* (Please re-

fer to Supplementary data 3 for the identity of each of the unique pathways). Combined differential pathways were analyzed across the three regions. Three pathways in LMJ vs 2 in PAM were shared between *Corpus callosum* and *Medulla oblongata*, 6 vs 15 pathways

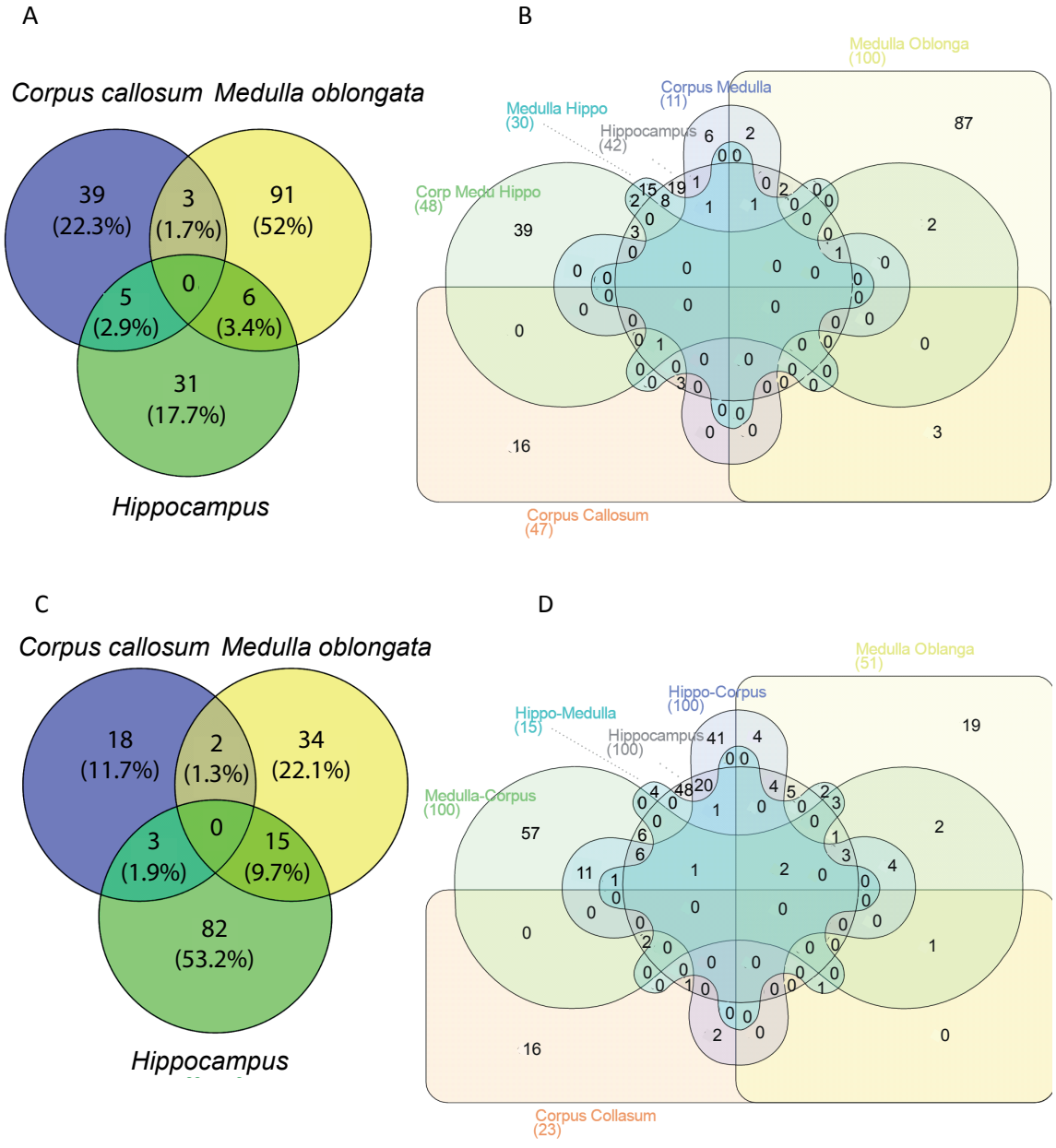


FIGURE 4.2 – **Biological and functional pathways among the brain regions**
Differential distribution of unique and common/intersected biological and functional pathways among the three brain regions (*Corpus callosum*, *Hippocampus* and *Medulla oblongata*) obtained with LMJ (A) or PAM (C) extraction methods. Each brain region was analyzed across the three regions using a comprehensive Venn analysis representation extracted from Subnetwork Enrichment Analysis (B with LMJ and D with PAM).

between *Hippocampus* & *Medulla oblongata*, and 5 vs 3 pathways between *Corpus callosum* and *Hippocampus*. Integrated Venn diagram analysis was performed using "the InteractiVenn" : a web-based tool for the analysis of complex data sets (Figures 3A-B) (Heberle *et al.*, 2015). See Supplementary data 3 for the listed differential pathways. Overexpressed proteins common to *Medulla oblongata* and *Hippocampus* (Figure 3A) are involved in learning, epilepsy, neuronal activity and plasticity, neurotransmission and ischemia. For *Hippocampus* and *Corpus callosum* (Figure 3A), the identified proteins are mainly involved in neurogenesis, cell proliferation and oxidative stress. For *Medulla oblongata* and *Corpus callosum* (Figure 3A), the pattern is more related to cell damage and life span. The same analysis for unique pathways in *Hippocampus* clearly showed protein patterns involved in neurogenesis, synaptogenesis, neurite outgrowth, neuroprotection, and axogenesis (Figure 3B, Supplementary data 4). For *Medulla oblongata* the proteins are mainly involved in pathways related to memory consolidation, epilepsy, cognition disorders, oligodendrocytes differentiation, amyotrophic lateral sclerosis, and spinocerebral ataxia type 1 (Figure 3B). For *Corpus callosum*, the proteins are mainly implicated in beta thalassemia, anemia and related hemoglobinopathies (Figure 3B). All the results are in line with biological and physiological functions of these 3 brain regions.

4.1.4.3 PTM analysis of identified proteins

PTM analysis of proteins from the 3 regions revealed the presence of 91 proteins which were identified with PTMs, of which, 29 were detected in the *Hippocampus*, 40 in the *Corpus callosum* and 37 in the *Medulla oblongata* (Supplementary data 2). Interestingly, some proteins show region-specific PTMs (Table 1, Supplementary data 2). As an illustration, the most abundant PTM of stathmin in the *Corpus callosum* (identified) and the *Hippocampus* (detected but not identified) was the Nter-Acetyl + 1 Phosphorylation, whereas in the *Medulla oblongata* (identified) it was the Nter-Acetylation (Figure 4). Similarly, neurogranin was specifically phosphorylated in the *Hippocampus*. Another example is the Astrocytic phosphoprotein (PEA-15), which was observed with a phosphorylated residue in the *Corpus callosum* but not in the *Medulla oblongata* (Table 1 & Supplementary data 2). Similarly, Parathymosin was identified with a mass shift of +79.94 Da in *Hippocampus* by two spectra and with 5.89 and 5.38 ppm mass errors compared to theoretical mass plus a phosphorylation, thus implying a phosphorylated residue (Table 1, Figure 1G & Supplementary data 1 & 2). These data clearly revealed that the PTM state of proteins is linked to the brain regions where they are localized, and consequently with the biological function of the protein in relation to the physiological function of the considered brain region.

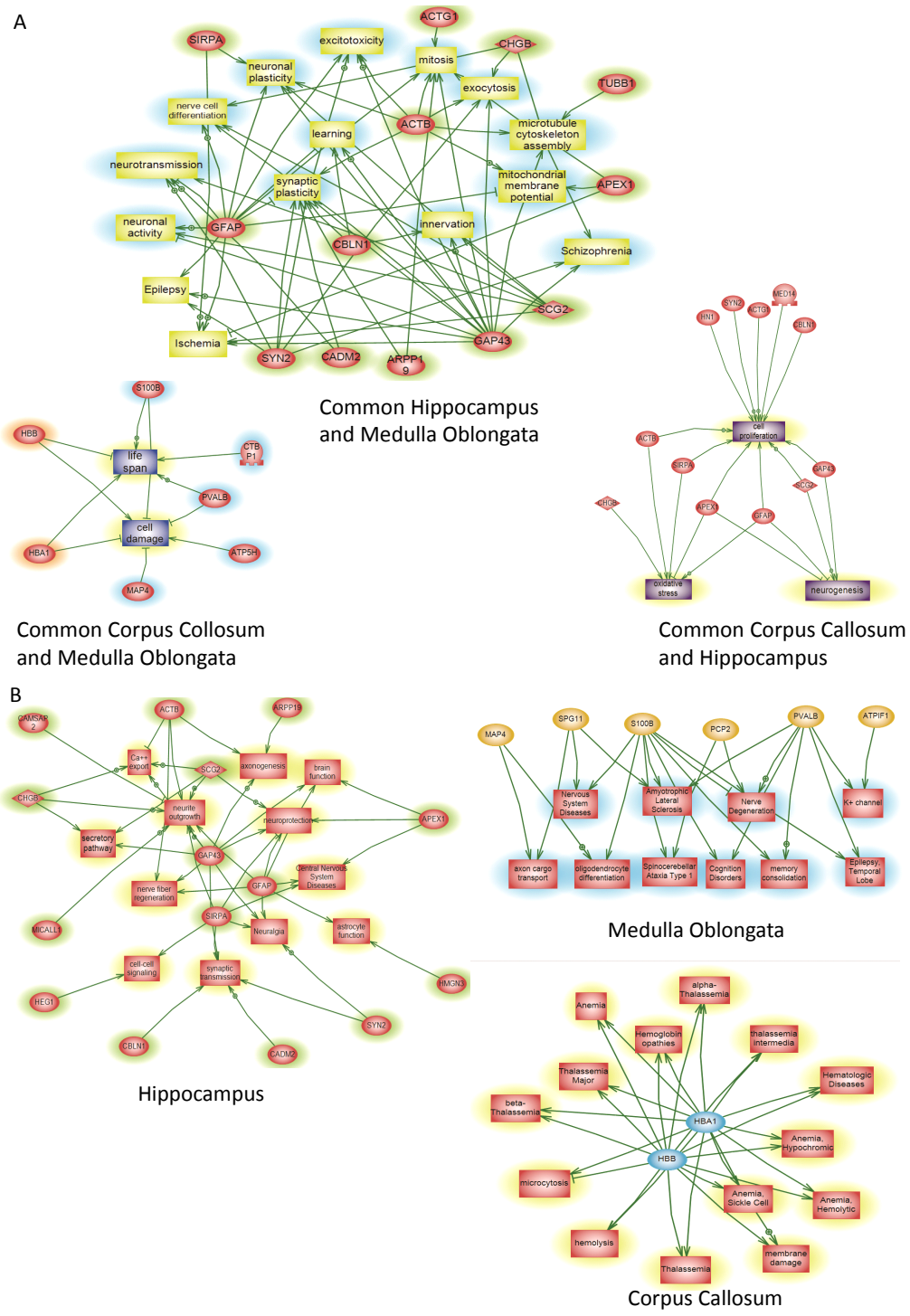


FIGURE 4.3 – **Global pathway analysis**
(A) Global pathway analyses of the over-expressed proteins common to two different regions i.e., *Hippocampus* and *Medulla oblongata*, *Hippocampus* and *Corpus callosum* and *Medulla oblongata* and *Corpus callosum*. (B) Over-expressed proteins in the *Hippocampus*, *Medulla oblongata* or *Corpus callosum* were involved in globally altered molecular pathways.

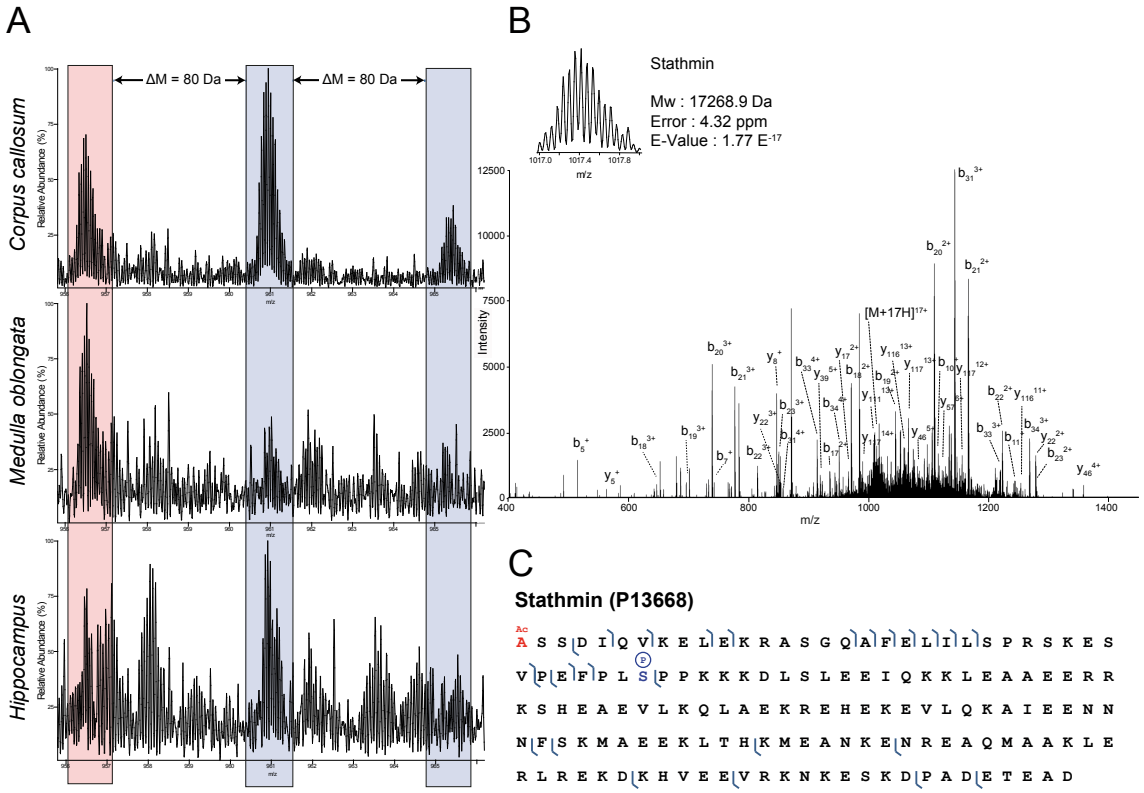


FIGURE 4.4 – **Region-specific PTM profile of Stathmin**
ROI-specific PTM profile of Stathmin (A) which was identified in *Corpus callosum* and *Medulla oblongata* but not in *Hippocampus*. N-terminal acetylation is marked red, while the phosphorylations are marked blue. Precursor (z = 17) and assigned fragmentation spectrum of Nter-Acetylated and phosphorylated (Ser-38) Stathmin (B) and fragmentation map (C).

4.1.4.4 Protein fragments linked to brain region localization

Data analyses revealed the presence of protein fragments in the three brain regions (Table 2 & Supplementary data 8). These fragments are derived from large proteins such as neuropeptide precursors (somatostatin, proenkephalin, secretogranin 1 & 2), Synuclein (alpha, beta and gamma), Synaptosomal associated protein 25, DNA-(apurinic or apyrimidinic) protein (APEX), Hematological and neurological expressed 1 protein (HN1), Myelin basic protein (MBP) and Thymosin beta 4. The generated fragments are linked to the presence of processing enzymes e.g. pro-protein convertases, neutral endopeptidases, angiotensin-converting enzymes and aminopeptidases, which are differentially expressed in the brain region (Zheng *et al.*, 1997; Walther *et al.*, 2009; Harmer *et al.*, 2002; Salzet *et al.*, 2000; Day et Salzet, 2002). Neuropeptide fragment precursors, neuromodulin and secretogranin 1 are principally detected in *Hippocampus* whereas fragments of MBP and somatostatin are detected in majority in *Medulla oblongata*. HN1 fragments are detected in *Hippocampus*, whereas Secretogranin 2 is present in both *Hippocampus* and *Medulla oblongata*.

4.1.4.5 Alternative protein identification

Three alternative proteins were detected in top-down microproteomics experiments. AltCd3e and AltMyo1f were detected in *Hippocampus* using LMJ and PAM, respectively, and AltGrb10 was detected in the *Medulla oblongata* using PAM (Table 3). These results suggest that the strategy was suitable for tissue top-down microproteomics studies of the reference and hidden proteomes. We then enlarged this study by re-analyzing previous data obtained using whole rat brain sections (PMID :27512083) (Quanico *et al.*, 2016b). Reanalysis of this dataset allowed the identification of 5 more alternative proteins (Table 3, Supplementary Data 6). These alternative proteins are translated from sequences located in mRNAs 3'UTR (AltSstr3, AltKcnq5, AltLdlr), 5'UTR regions (AltZbtb8a) of mRNAs and from a putative non-coding RNA (AltRn50_X_0580.1).

4.1.4.6 Back correlation to localization by MALDI MSI

Intact protein MSI experiments were performed to show ion distributions of the proteins identified by top-down MS. To this end, two images were acquired ; the first section was prepared with HCCA/aniline matrix and the second one with SA/aniline. The images were acquired only on the three ROIs specified in the previous imaging experiment. Peaks obtained from these images were then matched with the observed monoisotopic masses from the top-down MS analysis performed on the entire rat brain tissue section. Thirty eight protein IDs obtained from the reference proteome were assigned to peaks obtained from

both images with a delta mass cutoff of 10 Da (Supplementary Data 7). This includes five proteins previously matched also with top-down MS data, namely PEP-19 (Pcp4), ubiquitin (Ubc), thymosin β-4 (Tmsb4x), thymosin β-10 (Tmsb10), and calmodulin (Calm1) [34]. Figure 5A shows the ion image of m/z 4966 assigned as the intact form (as hematopoietic system regulatory peptide) of thymosin β-4 (theoretical mass = 4960.49). The specific localization of m/z 4966 in the *Hippocampus* can be clearly observed. Topdown data indicate that this isoform, detected as the $[M+5H]^{5+}$ charge state, is the N-acetylated isoform after methionine excision (Figure 5B). Its distribution in the *Hippocampus* in MSI correlates well with the topdown data where this form was detected using PAM. Furthermore, its detection by MSI and assignment of N-acetylation by topdown is in accord with the MSI database reported by Maier et al (Maier *et al.*, 2013). Figure 5C shows the mapping of m/z 5180 assigned as the C-terminal fragment of α-synuclein (observed as the $[M+4H]^{4+}$ charge state in topdown, Figure 5C), showing its particular intense distribution along the hippocampal dentate gyrus. Its distribution in the cerebral cortex observed in the ROI that includes the *Corpus callosum* was also detected in both MSI and topdown. To verify the specific formation of this fragment, the putative protease cleavage sites found in the full amino acid sequence of α-synuclein was mapped using the PROtease Specificity Prediction servER (PROSPER), where it can be observed that cleavage by matrix metallopeptidase 3 (MMP3) can induce the generation of the C-terminal fragment (Figure 5D). In situ hybridization of the genes that code for α-synuclein (Snca) and MMP3 in mouse brain obtained from the Allen Mouse Brain Atlas (http ://mouse.brain-map.org/) (Lein *et al.*, 2007) confirms the distribution of α-synuclein (strong) and MMP3 (weak) along the mouse hippocampal dentate gyrus (Figure 5E).

4.1.5 Discussion

We developed a novel strategy combining MALDI imaging with top-down microproteomics to determine localized proteoforms, including truncated forms, fragments, and possibly altprots. First, molecular histology was performed using MALDI-MSI and spatial segmentation in order to distinguish ROIs within a tissue. These ROIs were then subjected to protein microextraction with ProteaseMAX rather than SDS or organic solvents. Protein microextraction efficiency was confirmed by nanoLC high resolution MS/MS analysis of rat brain tissue since we identified a large number of proteins (123) compared to the 36 previously identified from a whole tissue proteomics study which performed extraction using acidified MeOH (Ye *et al.*, 2014). Only 19 proteins were in common with those identified from this study. The 17 proteoforms absent in our study are small peptides less than 4500 Da and are more related to the neuropeptide family, e.g. chromogranin-A, cholecystoki-

nin, proneuropeptide Y, secretogranin-2, proSAAS, cocaine- and amphetamine-regulated transcript protein, and oxysterol-binding protein, consistent with the brain regions selected in our study. Nevertheless, the common proteoforms identified are exactly the same with the same PTMs.

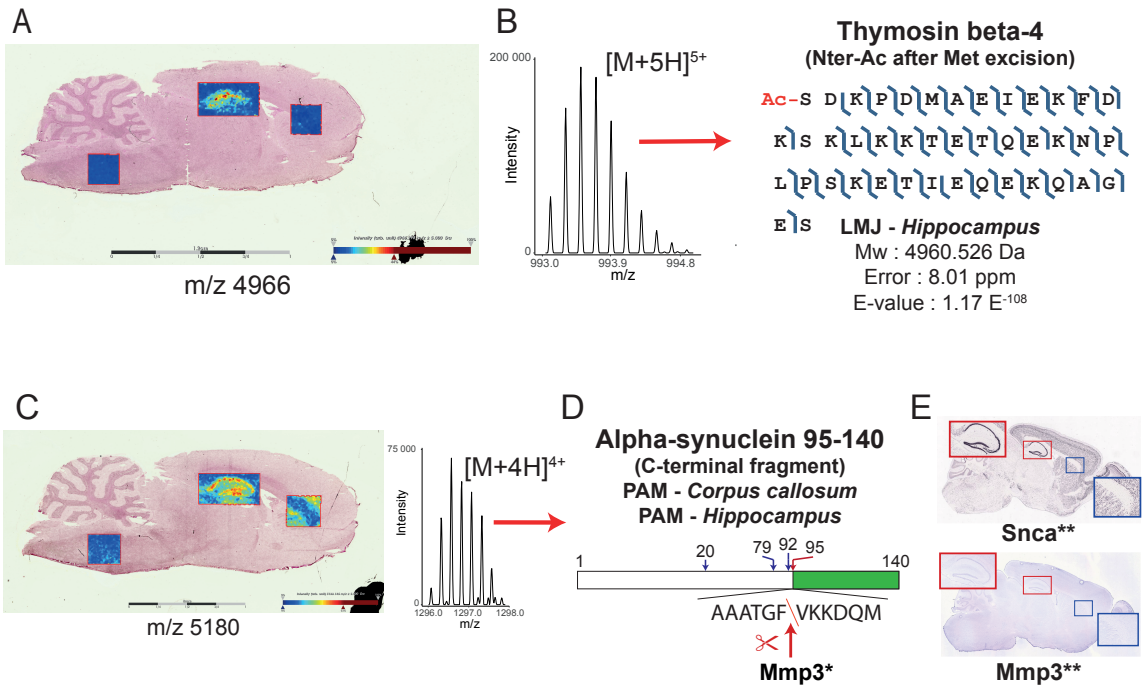


FIGURE 4.5 – **Back-correlation of proteins in MALDI-MSI and top-down microproteomics**
(A) MALDI-MSI of m/z 4966 attributed to intact N-terminally acetylated form of Thymosin beta-4 with corresponding identification by top-down microproteomics (B). (C) MALDI-MSI of m/z 5180 attributed to C-terminal fragment of alpha-synuclein identified in *Hippocampus* and *Corpus callosum*. (D) Schematic representation of protein fragment with predicted Stromelysin (Mmp3) cleavage sites (arrows and amino acid numbers) by PROSPER* (Song *et al.*, 2012) and identified form (red arrows). (E) In-situ- hybridization of alpha-synuclein (top) and Stromelysin (bottom). **Image credit : Allen Institute (Lein *et al.*, 2007).

It is interesting to note that LMJ and PAM do not identify the same proteins and are thus complementary, giving a total of 123 protein IDs overall. For example, somatostatin and peptide 143-185 of proenkephalin-A were specifically identified in LMJ samples whereas alpha-synuclein and neuromodulin where specifically identified in PAM experiments. Considering that the average size of brain cells is 15 μm and that we have microextracted 0.8 mm^2 with LMJ and 1 mm^2 with PAM, we estimate that we identified proteins from 4444 cells for LMJ and 5662 cells for PAM. By combining the two approaches, 15 specific and non-redundant proteins were identified from the *Corpus callosum*, 17 from

the *Medulla oblongata* and 24 from the *Hippocampus* (Tables 1 and 2). 35 are common to the 3 brain regions, 16 between *Corpus callosum* and *Medulla oblongata*, 8 between *Corpus callosum* and *Hippocampus* and 8 between *Medulla oblongata* and *Hippocampus*. Proteins identified with PAM are mainly present in the cytoplasm (62%), mitochondrial membrane (9.3%) or organelles and plasma membranes (28.7%). With LMJ, the proteins identified are from organelles (51.5%) and the cytoplasm (47.7%).

These studies performed by top-down proteomics are in line and complementary to our previous studies based on bottom-up proteomics (Quanico *et al.*, 2013; Franck *et al.*, 2013; Quanico *et al.*, 2015) as it gives information about the precursor mass and PTMs detectable by measuring the ΔM(s) between the intact precursor within a close retention time window. Indeed, our approach successfully discriminate stathmin PTMs between different regions of rat brain tissue (Figure 4). We showed that stathmin is more abundant in *Corpus callosum* and *Medulla oblongata* and its PTM pattern is specific for each of these two regions. The ratio phospho-stathmin/Nter-Ac was significantly higher in the *Corpus callosum*, suggesting a different biological activity in these two regions of the brain (Figure 4). Similarly, out of the 41 unique proteins that were identified with PTMs (Supplementary data 1 & 2), 22 had region specific PTMs (Table 1). The most prevalent PTMs are the N-acetylation of proteins and phosphorylation. For example, we found that α-synuclein presents one PTM, i.e., N-acetyl-L-methionine in *Medulla oblongata*, *Hippocampus* and *Corpus callosum*. In literature it has been shown that α-synuclein acetylation at Met in position 1 seems to be important for its proper folding (Sarafian *et al.*, 2013; Trexler et Rhoades, 2012). Similarly, the Astrocytic phosphoprotein (PEA-15) possesses N-acetyl-L-alanine in *Medulla oblongata* and N-acetyl-L-alanine plus O-phospho-L-serine in *Corpus callosum*. None of them have been previously identified (Lundby *et al.*, 2012).

In the same way, we identified protein fragments from proteins with particular distribution and presented a specific cleavage form across each brain region. Majority of the identified fragments are large neuropeptides like synenkephalin and secretogranins 1 & 2. These fragments are produced by enzymatic cleavage of the pro-protein convertase family like PC1/3, PC2 or PC5, PACE4 (Zheng *et al.*, 1997). We previously demonstrated the role of these enzymes in proenkephalin maturation (Day et Salzet, 2002; Salzet, 2001) and found some of these neuropeptide fragments in temporal lobe epilepsy (Mériaux *et al.*, 2014) and Alzheimer's disease (Kim *et al.*, 2015), such as secretogranins for example. Synenkephalin is implicated in circadian rhythm in the *Hippocampus* (Asai *et al.*, 2007),

Snap25 is implicated in synaptogenesis and memory consolidation (Hou *et al.*, 2004; Hong-jun *et al.*, 2002; Aigner *et al.*, 1995). As previously demonstrated, we confirmed that the somatostatin is present in *Medulla oblongata* (Johansson *et al.*, 1984) whereas we showed for the first time the presence of the hematological and neurological expressed 1 protein in the *Hippocampus* (fragment) and *Corpus callosum* (full length after methionine excision).

Besides these novel protein fragments, another small family of proteins has been identified from the hidden proteome. In fact, more and more evidence suggests that mRNAs contain more than one coding sequence and could be translated into an annotated or reference protein and at least one alternative protein (Vanderperre *et al.*, 2013; Mouilleron *et al.*, 2015). We tested if our strategy was able to detect intact alternative proteins. We identified 3 alternative proteins (Table 3) by the top-down microproteomics approach which share no sequence similarity with annotated rattus norvegicus proteins. Of the 5 novel altprots identified by reanalysis of the study on whole tissue sections (Alt-Kcnq5, Alt-Zbtb8a, Alt-Sstr3, Alt-Ldlr and a non-coding RNA Alt-Rn50_X_0580.1), 3 of them are receptors as reference proteins i.e. somatostatin 3 receptor, potassium voltage-gated channel subfamily Q member 5, and low density lipoprotein receptor. It is interesting to note that these 3 receptors are known to be expressed in *Hippocampus* specifically (Dournaud *et al.*, 1996; Yus-Najera *et al.*, 2003; Poirier *et al.*, 1993).

Back correlation of top-down microproteomics protein IDs with MALDI MS images allowed to localize 38 identified proteins (Supplementary Data 7). The correlation included proteins with PTM modifications or enzymatic cleavage whose distribution varies differently in the 3 regions in line with identified biological processes taking place in each individual region. As an example, the truncated, N-acetylated form of thymosin β-4 was mapped in MSI and its distribution was compared with the result of the topdown data, showing good correlation of the results from the two approaches. The C-terminal fragment of α-synuclein likewise showed very good correlation of results. More importantly, the distribution of this fragment in the hippocampal dentate gyrus in MSI can be correlated with the abundance of α-synuclein and MMP3 in the same region in ISH experiments on mouse brain. MMP3 can cleave α-synuclein at F94, yielding the natively unstructured C-terminal fragment aa 95-140 (5.74 kDa). Other MMP3-produced C-terminally truncated peptides of α-synuclein (aa 1-78, 1-91 and 1-93) have been reported under stress conditions, with aa 1-93 being implicated in dopamine neuronal loss in substantia nigra, suggesting that overexpression of the fragments could have a

significant impact in Parkinson's disease (Choi *et al.*, 2011). What role aa 95-140 has in this regard thus needs to be further investigated.

Taken together, our results show that top-down microproteomics linked to MALDI MSI can be used to search for biomarkers, PTM detection and to identify novel proteins expressed from altORFs.

4.1.6 Acknowledgements

Supported by grants from Région Nord Pas-de-Calais and PROTEO (V. Delcourt), University Lille 1 (BQR to Dr. Julien Franck), Canadian Institutes for Health Research (MOP-136962) and Canada Research Chairs in Functional Proteomics and Discovery of New Protein (Prof. X. Roucou), PRISM (Prof. M. Salzet), Ministère de l'Enseignement Supérieur et de la Recherche via Institut Universitaire de France (Prof. I. Fournier), SIRIC ONCOLille (Prof. I. Founier), and Grant INCa-DGOS-Inserm 6041.

Author contributions
IF, JF and MS conceived the study. VD, JF, JQ, JPG, MW, and FK performed the experiments. IF, JF and MS supervised the project, and participated in experimental design, data analyses and writing of the manuscript with contributions coming from all co-authors. MS, IF, XR and JF also obtained funds for the project.

Additional Information The authors declare no competing financial interests in this work.

Tableau 4.1 – **Region specific post-translationally modified proteins**
PTMs from ProsightPC were concatenated with imputed PTMs from mass shifts (i.e. Acetylation (+42) ; Phosphorylation (+80))

Region	Accession number	PTM(s)	Protein name	Theo. mass (Da)	Obs. mass (Da)	Shift (Da)
Corpus callosum	P13668	N-acetyl-L-alanine, O-phospho-L-serine	Stathmin	17268.9	17269	0.094
	P31399	N-acetyl-L-alanine	ATP synthase subunit d, mitochondrial	18662.6	18662.6	0.082
	G3V9C0	N-acetyl-L-serine	Histone H2A	14037.9	14038	0.05
	Q5U318	N-acetyl-L-alanine, O-phospho-L-serine	Astrocytic phosphoprotein PEA-15	15021.7	15021.8	0.05
	D3ZHW9	N-acetyl-L-serine	Protein Shfm1	8183.53	8183.6	0.044
	B2RZ27	N-acetyl-L-serine	Protein Sh3bgrl3	10381.2	10381.3	0.033
	D3ZZW2	N-acetyl-L-serine	Protein LOC100910678	6972.9	6972.9	-0.008
	D3ZTB5	N-acetyl-L-alanine	Protein S100a13	11101.9	11101	-0.909
Hippocampus	Q04940	O-phospho-L-serine + Phosphorylation (+80)	Neurogranin	7440.43	7520.5	80.041
	M0R5I3	Phosphorylation (+80)	High mobility group nucleosomal binding domain 3, isoform CRA_a	10236.4	10316.4	79.973
	Q5U1W8	Phosphorylation (+80)	High-mobility group nucleosome binding domain 1	9987.3	10067.3	79.964
	P04550	N-acetyl-L-serine + Phosphorylation (+80)	Parathymosin	11463.2	11543.1	79.942
	P62329	N-acetyl-L-serine + Acetylation (+42)	Thymosin beta-4	4960.49	5002.5	42.029
Medulla oblongata	P06302	N-acetyl-L-serine + Acetylation (+42)	Prothymosin alpha	12286.1	12328.1	41.993
	P04631	N-acetyl-L-serine	Protein S100-B	10648	10648.1	0.05
	B2RYS2	N-acetyl-L-alanine	Cytochrome b-c1 complex subunit 7	13460.9	13460.9	0.047
	P63041	N-acetyl-L-methionine	Complexin-1	15154.5	15154.5	0.047
	P0CC09	N-acetyl-L-serine	Histone H2A type 2-A	13997.9	13997.9	0.042
	P02625	N-acetyl-L-serine	Parvalbumin alpha	11829	11829	0.032
	P11951	N-acetyl-L-serine	Cytochrome c oxidase subunit 6C-2	8360.42	8360.4	0.022
	Q5U318	N-acetyl-L-alanine	Astrocytic phosphoprotein PEA-15	14941.8	14940.8	-0.918
	P31044	N-acetyl-L-alanine	Phosphatidylethanolamine-binding protein 1	20699.4	20698.4	-0.935

Tableau 4.2 – **Most detected truncated protein**
M.O : medulla oblongata ; C.C : corpus callosum ; Hi : hippocampus ; AA : amino acids

Accession number	Protein description	M.O	C.C	Hi	Detected length (AA)	Full length (AA)	Fragment (AA)	Fragment position
P21571	Chain [33-108] in ATP synthase-coupling factor 6, mitochondrial	✓	✓	✓	76	108	33-108	C-terminal fragment
P10818	Chain [27-111] in Cytochrome c oxidase subunit 6A1, mitochondrial	✓	✓	✓	85	111	27-111	C-terminal fragment
P11240	Chain [38-146] in Cytochrome c oxidase subunit 5A, mitochondrial	✓	✓	✓	109	146	38-146	C-terminal fragment
Q71UE8	Chain [1-76] in NEDD8	✓	✓	✓	76	81	1-76	N-terminal fragment
P35171	Chain [24-83] in Cytochrome c oxidase subunit 7A2, mitochondrial	✓	✓	✓	60	83	24-83	C-terminal fragment
P21571	ATP synthase-coupling factor 6, mitochondrial	✓	✓	✓	53	108	56-108	C-terminal fragment
P47942	Dihydropyrimidinase-related protein 2	✓	✓	✓	55	572	518-572	C-terminal fragment
Q63429	Polyubiquitin-C	✓	✓		74	810	1-74	N-terminal fragment
P28073	Proteasome subunit beta type-6	✓	✓		17	238	78-94	Internal fragment
P80432	Chain [17-63] in Cytochrome c oxidase subunit 7C, mitochondrial	✓	✓	✓	47	63	17-63	C-terminal fragment
D4A5W9	Synaptosomal-associated protein	✓	✓	✓	45	206	162-206	C-terminal fragment
P13668	Stathmin		✓		112	149	38-149	C-terminal fragment
P21571	ATP synthase-coupling factor 6, mitochondrial	✓	✓	✓	75	108	34-108	C-terminal fragment
F1LQ96	Gamma-synuclein	✓		✓	30	122	93-122	C-terminal fragment
F1LUV9	Neural cell adhesion molecule 1		✓		61	833	773-833	C-terminal fragment
P37377	Alpha-synuclein		✓		73	140	68-140	C-terminal fragment
O35314	Secretogranin-1			✓	30	675	292-321	Internal fragment
P19527	Neurofilament light polypeptide	✓			74	542	469-542	C-terminal fragment
Q5M7W5	Microtubule-associated protein 4	✓			16	1057	31-46	Internal fragment
P26772	10 kDa heat shock protein, mitochondrial	✓	✓		37	102	66-102	C-terminal fragment
Q8R1R5	CD99 antigen-like protein 2	✓	✓	✓	24	246	223-246	C-terminal fragment
Q6PCU8	Chain [36-108] in NADH dehydrogenase [ubiquinone] flavoprotein 3, mitochondrial	✓	✓	✓	73	108	36-108	C-terminal fragment
F1LQ96	Gamma-synuclein	✓		✓	48	122	75-122	C-terminal fragment

Tableau 4.3 – **Alternative protein products identified by tissue top-down proteomics**

Region	E-Value (P-score)	Observed Mass (Da)	Protein	Gene	AltORF localization on RNA	Transcript
Hippocampus	2.14 E-05 (3.80E-11)	4642.28	AltCd3e	Cd3e	3'UTR	ENSRNOT00000047291
	7.70E-05 (1.37E-10)	8154.94	AltMyo1f	Myo1f	CDS	ENSRNOT00000011513
Medulla Oblangata	1.06E-05 (1.89E-11)	15025.79	AltGrb10	Grb10	CDS	ENSRNOT00000085175
Whole brain section PMID : 27512083	3.15E-05 (2.24E-10)	4760.46	AltRn50_X_0580.1	Rn50_X_0580.1	ncRNA	ENSRNOT00000066392
	3.42E-05 (2.43E-10)	5000.62	AltSstr3	Sstr3	3'UTR	ENSRNOT00000009612
	4.23E-07 (7.52E-13)	3344.66	AltZbtb8a	Zbtb8a	5'UTR	ENSRNOT00000010983
	2.48E-09 (1.77E-14)	2825.44	AltKcnq5	Kcnq5	3'UTR	ENSRNOT00000040034
	1.36E-05 (9.68E-11)	4440.29	AltLdlr	Ldlr	3'UTR	ENSRNOT00000013496

4.2 Conclusion et perspectives

Les stratégies de microprotéomique, à l'inverse des modèles employant des lignées cellulaires ou la lyse complète d'un tissu, offrent une grande précision de mesure du protéome dans des régions localisées. Les techniques de micro-jonction liquide (LMJ) par dépôt d'un solvant à la surface d'une région d'intérêt d'un tissu (Quanico *et al.*, 2013) et de microdissection manuelle assistée sur parafilm (PAM) où la région d'intérêt est isolée du tissu à l'aide d'un scalpel (Franck *et al.*, 2013) se sont montrées efficaces pour décrire le protéome localisé de tissus biologiques. Ces techniques présentent l'avantage majeur d'analyser une population de cellules variées au sein de tissus et donc de capter protéome du microenvironnement cellulaire.

Ces techniques, alors réservées à l'application de l'approche de protéomique *bottom-up*, soit par la digestion localisée en micro-goutte (LMJ) ou en milieu liquide après extraction (LMJ et PAM), n'avaient pas encore été adaptées à l'analyse de protéine intactes par *top-down*. Cette technique se distingue par sa complémentarité vis-à-vis de l'approche *bottom-up*. En effet, elle détermine avec précision la masse du précurseur protéique tel qu'il est présent au sein de l'échantillon et donc à son état intact, tronqué ou en cours de dégradation. Elle offre aussi la possibilité de déterminer avec plus d'aisance les modifications post-traductionnelles des protéines, soit par leur observation directe ou par mesure de l'écart de masse entre la masse prédite de la protéine et sa mesure expérimentale. Enfin, par l'absence de traitement protéolytique, elle est particulièrement performante pour caractériser le petit protéome, moins facilement identifiable par approche *bottom-up*.

Son application nécessite toutefois des traitements d'échantillons spécifiques. En effet, les détergents sont requis pour l'extraction efficace de protéines à partir de matrices biologiques complexes mais sont par ailleurs incompatibles avec la spectrométrie de masse et notamment les couplages de chromatographie liquide et ESI-MS. Le détergent le plus employé pour l'extraction de protéines est le sodium-dodecyl sulfate, chargé négativement, qui est connu pour interagir avec les systèmes de chromatographie et masquer les charges positives des analytes lors de l'acquision de spectres en mode positif. Ce dernier est la plupart du temps éliminé par des étapes de précipitations et dessalage employant des solvants apolaires. Or, l'utilisation de ce type de solvant entraine l'agrégation des protéines, des difficultés accrues pour resolubiliser les protéines avant l'analyse en spectrométrie de masse et donc l'utilisation de quantités initiales d'extraits cellulaires importante (Kachuk et Doucette, 2017). Ce critère est particulièrement limitant en microprotéomique car l'utilisation de plus grandes quantités de matériel est associée à une région plus importante de

tissu extraite et donc moins spécifique.

Dans ce contexte, nous avons développé une méthode de préparation d'échantillon qui associe l'extraction localisée de tissu biologique et analyse de protéines intactes. L'emploi de détergents compatibles avec la spectrométrie de masse s'est avéré être efficace pour extraire et analyser les protéines intactes à partir de régions d'intérêt. Ces détergents ont la particularité de présenter des groupements réactifs en milieu acide et/ou sensibles à la température. Leur efficacité est comparable aux détergents communément employés mais représentent un coût associé plus grand. Enfin, leur atout majeur est l'absence d'étape de précipitation ce qui permet l'analyse d'une quantité minime de protéines.

Afin de définir des régions d'intérêt, des expériences d'imagerie par MALDI-MS ont été réalisées. Par ces expériences, il est possible de reconstruire les images moléculaires d'un tissu afin d'en étudier la composition. Les spectres sont alors classés par classification hiérarchique afin de distinguer les régions du tissu qui présentent des profils moléculaires différents. Suite à cette classification, il est alors possible de définir des régions d'intérêt pour en étudier le protéome.

Après extraction des différentes régions et analyse par spectrométrie de masse, nous avons pu identifier des protéines communes ou spécifiques de chacune des régions extraites. Le nombre de protéines identifiées, même s'il est moindre que ceux obtenus par application de l'approche bottom-up, est important compte tenu de la quantité estimée de cellules extraites par les deux techniques. Ces protéines ont des poids moléculaires compris entre 3 et 20 kDa, confirmant l'aptitude de l'approche *top-down* à analyser de petites protéines. La stratégie d'extraction PAM s'est révélée plus performante en termes d'identifications uniques de protéines par rapport à la LMJ. Cette différence peut être expliquée par les étapes de lavages aux solvants organiques réalisés avant d'extraire les protéines pour la LMJ. Ces solvants pourraient avoir éliminé une partie des protéines avant l'étape d'extraction. Cependant, certaines des protéines ont été spécifiquement identifiées par l'une ou l'autre des deux techniques, ce qui démontre qu'elles sont complémentaires. Par l'analyse des protéines associées à leurs localisations, nous avons pu déterminer les voies de signalisation cellulaire qui leur sont associées. Aussi, certaines protéines ont été identifiées avec des modifications post-traductionnelles (PTMs) qui sont parfois spécifiques des régions analysées ou dont la distribution est différente comme exposé pour la stathmine. Ce résultat confirme les capacités de l'approche *top-down* pour la détermination de PTMs.

Certaines protéines tronquées ont également été identifiées. Ces fragments de protéines peuvent être issus de la maturation ou de la dégradation des protéines. Les protéines tron-

quées ont un intérêt biologique particulier car elles sont associées à des activités biologiques différentes de leurs précurseurs de pleine longueur, notamment par la perte de portions contenant des résidus ou domaines protéiques nécessaires à leur fonction. Ce résultat est par ailleurs confirmé par la détection d'une forme tronquée de la protéine α-synucléine produite suite à un clivage par la stromélysine d'après le programme PROSPER (Song *et al.*, 2012). En réalisant des expériences complémentaires d'imagerie MALDI de protéines, nous avons également pu déterminer que ce fragment était précisément localisé au sein du gyrus denté de l'hippocampe et dans les régions voisines du corps calleux. Les protéines tronquées peuvent également constituer une source de biomarqueurs pathologiques comme démontré lors d'études antérieures (Lemaire *et al.*, 2007a; Longuespée *et al.*, 2012).

De plus, les expériences de microprotéomique ont également permis la détection de trois protéines alternatives localisées et de 5 protéines alternatives détectées dans une étude déjà publiée de microprotéomique (Quanico *et al.*, 2016b). La détection de protéines alternatives par approche *top-down* confirme l'hypothèse initiale que cette approche est adaptée pour la détection de ces protéines généralement de petite taille. Toutefois, le faible nombre de protéines alternatives identifiées et de leurs spectres associés suggère qu'une part de ces protéines seraient faiblement abondantes. Aussi, l'approche *top-down* est aussi associée à la sélection successive des différents états de charge des mêmes précurseurs protéiques. En effet, on observe souvent de multiples états de charge lorsqu'une protéine intacte est analysée par ESI-MS. En approche DDA, la sélection de l'ion pour la fragmentation est conditionnée par son intensité et, si l'exclusion dynamique est activée, par sa sélection antérieure. Or, la détection du même précurseur à des états de charge différents limite la fonction d'exclusion dynamique et donc la sélection d'ions moins abondants, ce qui peut constituer un frein à l'identification de protéines alternatives.

Enfin, le développement de cette technique sur un tissu modèle de cerveau de rat a prouvé la faisabilité d'associer imagerie par MALDI-MS et microprotéomique par *top-down*. Cette approche s'avère particulièrement efficace pour identifier le protéome de faible poids moléculaire, les éventuelles PTMs de ces protéines et certaines protéines alternatives à partir de régions d'intérêt d'un tissu biologique. La nouvelle méthode développée pourrait s'avérer être efficace pour la caractérisation de tissus pathologiques cancéreux pour la recherche de biomarqueurs protéiques. En effet, certaines protéines ou leurs fragments peuvent constituer des signatures spécifiques de pathologies diverses. Il est donc pertinent de l'employer pour évaluer ses capacités de détection de biomarqueurs issus de protéines de référence ou de nouveaux biomarqueurs potentiels issus de protéines alternatives.

5 ARTICLE 3

Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer

Auteurs de l'article: Vivian Delcourt, Julien Franck, Eric Leblanc, Fabrice Narducci, Yves-Marie Robin, Jean-Pascal Gimeno, Jusal Quanico, Maxence Wisztorski, Firas Kobeissy, Jean-François Jacques, Xavier Roucou, Michel Salzet, Isabelle Fournier

Statut de l'article: Publié

Avant-propos:

Comme démontré dans l'article précédent, il est désormais possible d'étudier les protéines intactes par approche *top-down* à partir de régions localisées d'un tissu biologique. Il serait intéressant d'appliquer cette méthode à un tissu cancéreux afin d'identifier de potentiels biomarqueurs protéiques issus de séquences canoniques ou alternatives. Dans cet article sous la supervision de mes encadrants, j'ai effectué les analyses d'imagerie par spectrométrie de masse afin de déterminer les régions d'intérêt. J'ai ensuite procédé aux microextractions protéiques localisées par deux approches de microprotéomique. J'ai par la suite identifié par des logiciels bio-informatiques les protéines de référence et alternatives au sein des divers échantillons analysés. Des analyses de signalisation ont ensuite été menées par nos collaborateurs. Après évaluation des différentes protéines alternatives identifiées, mes superviseurs ont souhaité poursuivre l'étude par la validation *in cellulo* de la protéine alternative altGNL1, identifiée à partir de la région tumorale du tissu. Nous avons réalisé le clonage de la protéine de référence, contenant la séquence codante pour la protéine alternative dans un cadre de lecture décalé de son CDS canonique, toutes deux étiquetées avec des rapporteurs différents. J'ai enfin procédé aux validations par *western-blots* et immunofluorescence après avoir surexprimé les différentes lignées cellulaires.

J'ai participé à la rédaction de l'article, à la construction des figures et tableaux, et aux réponses aux examinateurs après révision de l'article avec l'aide de nos collaborateurs et sous la direction de mes encadrants. Enfin, nous avons vérifié la version finale de l'article avant validation auprès de l'éditeur.

Résumé:

Contexte : Il a été récemment démontré que des protéines pouvaient être traduites à partir de cadres de lecture alternatifs (altORFs), augmentant l'étendue du protéome. L'approche *top-down* de protéomique par spectrométrie de masse permet l'identification de protéines intactes contenant des modifications post-traductionnelles ainsi que de formes tronquées provenant d'ORFs de référence et altORFs.

Méthodes : L'approche de microprotéomique par *top-down* a été appliquée sur des régions bénignes, tumorales et nécrotiques-fibrotiques de biopsie ovarienne, ce qui a permis d'identifier des protéines et leurs modifications spécifiques des régions analysées. Les régions d'intérêt ont été déterminées par imagerie par spectrométrie de masse MALDI et segmentation spatiale.

Découvertes : L'analyse à l'aide d'une base de données contenant les protéines canoniques et alternatives prédites a permis l'identification de 15 protéines alternatives. Parmi celles-ci, la protéine alternative G nucléolaire (altGNL1) identifiée dans la région tumorale est encodée à partir d'un ORF localisé dans un cadre de lecture décalé de la protéine canonique GNL1. La co-expression de GNL1 et altGNL1 a été validée par transfection dans des cellules HEK293 et HeLa à l'aide d'un plasmide d'expression contenant la séquence codante GNL1 étiquetée FLAG(V5). Les expériences de western-blot et immunofluorescence ont confirmé l'expression constitutive d'altGNL1-V5 et GNL1-FLAG.

Conclusions : Notre approche permet l'étude de l'expression protéique dans le contexte du cancer de l'ovaire de haut grade séreux, offrant la possibilité de détecter de potentiels marqueurs encore non évalués.

Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer

Vivian Delcourt, Julien Franck, Eric Leblanc, Fabrice Narducci, Yves-Marie Robin, Jean-Pascal Gimeno, Jusal Quanico, Maxence Wisztorski, Firas Kobeissy, Jean-François Jacques, Xavier Roucou, Michel Salzet, Isabelle Fournier

Journal : EBioMedicine
Editeur : Elsevier

5.1 Manuscrit

5.1.1 Abstract

Background : Recently, it was demonstrated that proteins can be translated from alternative open reading frames (altORFs), increasing the size of the actual proteome. Top-down mass spectrometry-based proteomics allows the identification of intact proteins containing post-translational modifications (PTMs) as well as truncated forms translated from reference ORFs or altORFs.
Methods : Top-down tissue microproteomics was applied on benign, tumor and necrotic-fibrotic regions of serous ovarian cancer biopsies, identifying proteins exhibiting region-specific cellular localization and PTMs. The regions of interest (ROIs) were determined by MALDI mass spectrometry imaging and spatial segmentation.
Findings : Analysis with a customized protein sequence database containing reference and alternative proteins (altprots) identified 15 altprots, including alternative G protein nucleolar 1 (AltGNL1) found in the tumor, and translated from an altORF nested within the GNL1 canonical coding sequence. Co-expression of GNL1 and altGNL1 was validated by transfection in HEK293 and HeLa cells with an expression plasmid containing a GNL1-FLAG(V5) construct. Western blot and immunofluorescence experiments confirmed constitutive co-expression of altGNL1-V5 with GNL1-FLAG.
Conclusions : Taken together, our approach provides means to evaluate protein changes in the case of serous ovarian cancer, allowing the detection of potential markers that have never been considered.

5.1.2 Introduction

With the recent advances in mass spectrometry (MS) based-proteomics, the application of top-down MS-based proteomic strategies now allows the analysis of complex protein mixtures in their intact state without the need for enzymatic digestion (Tran *et al.*, 2011). In a study by Ye *et al.* (2014), top-down MS-based proteomics coupled to MatrixAssisted Laser Desorption Ionization (MALDI) MS imaging (MALDI-MSI) of a rat brain post-treated with the NMDA receptor antagonist MK801 revealed 34 proteins with their specific post-translational modifications (PTMs) (Ye *et al.*, 2014). Recently, we performed MALDI-MSI coupled to top-down tissue microproteomics on 3 rat brain regions and demonstrated the possibility to identify specific proteoforms linked to the physiology of the tissue region ; several unique markers were identified showing different proteoforms of brain-specific proteins(data not shown). In this work, we investigated the pathological heterogeneity in ovarian serous cancer tumor microenvironment utilizing a top-down microproteomics approach. Specifically, we investigated proteome microenvironment alterations aiming to delineate and characterize specific protein profiles in benign, tumor and necrotic/fibrotic tumor regions by taking into account their PTMs and assessing their cleaved forms.

Our assessment also takes into account the identification of alternative proteins (AltProts) (Vanderperre *et al.*, 2013). AltProts are translated from alternative open reading frames (AltORFs). AltORFs can have different localizations : they can overlap annotated protein-coding sequences in a different reading frame, or can be present within untranslated regions (UTRs) of mature mRNAs (Mouilleron *et al.*, 2015; Vanderperre *et al.*, 2013). Thus, alternative proteins are completely different from annotated or reference proteins (Mouilleron *et al.*, 2015; Vanderperre *et al.*, 2013). AltORFs may also be present in transcripts annotated as non-coding RNAs (ncRNAs). Indeed, proteins translated from non-annotated AltORFs were detected in our previous studies by MS. Some of these alternative translation products have also been validated biologically and assessed for their biological activity. For example, we have shown that AltMRVI1 is translated from an AltORF overlapping the MRVI1 coding sequence in a different reading frame and interacts with BRCA1 (Vanderperre *et al.*, 2013). Translation of AltORFs in addition to annotated coding sequences opens the door to proteins that cannot be detected using conventional protein databases. Thus, due to their intriguing role, we aimed at investigating the profiles of the "hidden proteome" and assess their contribution in serous ovarian cancer. Additionally, these AltProts are mainly small proteins and the

top-down proteomics strategy seems to be a better alternative rather than the shotgun proteomics for their detection. This is so because, even if the shotgun approach remains the most efficient strategy for high throughput proteomics, the identification of small proteins in this approach can be hampered due to the low amount of generated tryptic peptides and the generally fewer presence of enzyme cleavage sites. Therefore, top-down proteomics offers a good alternative to identify small proteins or truncated forms as well as some PTMs from the reference or the hidden proteome. Overall, our aim is to identify and characterize reference and altprots as potential markers for serous ovarian cancer pathology.

5.1.3 Experimental Procedures

5.1.3.1 Tissue Collection

The ovarian biopsies were obtained from patients of the Centre Oscar Lambret (Lille, France) and from the CHRU de Lille Pathology Department. All experiments were approved by the local Ethics Committee (CPP Nord Ouest IV 12/10) in accordance with the French and European legislation on this topic. Methods of collection for human ovaries were performed in accordance with procedures that were approved by the Ethics Committee of the CHRU Lille. The study adhered to the principles of the Declaration of Helsinki and the Guidelines for Good Clinical Practice. All patients gave written informed consent before enrollment. The flash-frozen biopsies were stored at -80 °C until use.

5.1.3.2 Experimental Design and Statistical Rationale

We first performed MS imaging of lipids in order to perform spatial segmentation analysis to identify regions of interest (ROIs),which were then subjected to liquid micro-junction (LMJ) (Quanico *et al.*, 2013; Wisztorski *et al.*, 2016) or parafilm-assisted manual microdissection (PAM) (Franck *et al.*, 2013; Quanico *et al.*, 2015) methods of microextraction. LMJ and PAM were followed by top-down proteomics for protein identification from necrotic/fibrotic tumor, tumor and benign (B) regions (technical triplicate). Reference and alternative proteins were then identified and localized in the 3 tissue regions of the ovarian serous cancer biopsies.

5.1.3.3 Chemicals

MS grade water (H_2O), acetonitrile (ACN), methanol (MeOH), ethanol (EtOH), and chloroform were purchased from Biosolve (Valkensvaard, Netherlands). The cleavable detergent ProteaseMAX was purchased from Promega (Charbonnieres, France). Parafilm M

was purchased from Pechiney Plastic Packaging (Chicago, Illinois). 2,5- dihydroxybenzoic acid (DHB), sodium dodecyl sulfate (SDS), DL-dithiothreitol (DTT) trifluoroacetic acid (TFA) and formic acid (FA) were purchased from Sigma (Saint-Quentin-Fallavier, France).

5.1.3.4 Tissue Section Preparation

For MALDI-MSI experiments, tissues were cut in 10- µm slices using a cryostat (Leica Microsystems, Nanterre, France) and mounted on Indium Tin Oxide (ITO)-coated glass slides (LaserBio Labs, Sophia-Antipolis, France) by finger-thawing. For LMJ and PAM, consecutive tissue slices were also obtained but with a 30-µm thickness. For LMJ, the tissues were mounted on a polylysine glass slide. For PAM, on the other hand, the tissue sections were mounted on a parafilm M-covered glass slide (Franck *et al.*, 2013; Quanico *et al.*, 2015). After tissue section preparation, the sections were immediately dehydrated under vacuum at room temperature for 20 min. The slides were then scanned using a Nikon scanner and stored at -80 °C until use.

5.1.3.5 MALDI-MSI

DHB matrix (50 mg/mL) dissolved in 6 :4 (v/v) MeOH/0.1 % TFA in water was manually sprayed at a flow rate of 300 µL/h using a syringe pump connected to an electrospray nebulizer. The nebulizer was connected to a nitrogen line operated at 1 bar. The nebulizer was moved uniformly throughout the tissue until crystallization was sufficient to ensure optimal lipid detection. The tissue was then analyzed using an UltraFlex II MALDI-TOF/TOF (Time Of Flight) mass spectrometer equipped with a Smartbeam Nd-YAG laser (355 nm) and controlled by FlexControl software (Bruker Daltonics, Bremen, Germany). Lipid image acquisition was performed in positive reflector mode within an m/z range of 50 to 900 at a 300 µm resolution, and the obtained spectra were averaged from 300 laser shots per pixel. Peak detection and spatial segmentation analysis were then performed on the acquired images using SCiLS Software 2015b (SCiLS Lab GmbH, Bremen, Germany). For spatial segmentation, the Bisecting k-Means approach with Correlation as the distance metric was used. Spectra were subjected to median normalization and medium denoising prior to peak picking. After analysis, the ROIs were determined by selecting segments where the correlation distance metric is significantly distant from the other.

5.1.3.6 Intact Protein Extraction Buffer

To ensure minimal protein hydrolysis by endogenous proteases, every step from buffer preparation to nanoflow Liquid Chromatography (nanoLC)-MS/MS analysis was carried

out within the same day with on ice conservation in between sample processing steps. A 0.1 % (v/v) aliquot of temperature- and acid-cleavable commercial detergent (ProteaseMAX) was prepared from a 1 % (v/v) stock solution prepared in 50 µM DTT and immediately stored at -20 °C until use according to manufacturer's recommendations. Aliquots were processed within the day of sample extraction to ensure minimal degradation of the detergent over time, and remaining solutions were discarded.

5.1.3.7 LMJ Experiments

To ensure optimal protein extraction, lipids were removed from the tissue section by immersing the glass slide consecutively for 1 min each in 70% EtOH and in 95% EtOH then 30 s in chloroform with complete solvent evaporation under reduced pressure at room temperature between each washing step. The slide was then scanned again as washing steps improve structure visibility. The slide used for LMJ microextraction was placed inside a TriVersa NanoMate (Advion, Ithaca, NY, USA) instrument. Proteins were then extracted from every ROI by completing six cycles of extraction composed of the following steps : 1) aspirate 1.5 µl of detergent solution, 2) dispense 0.8 µl of the extraction buffer on the surface of the selected ROI with 10 iterations of up and- down pipetting, 3) aspirate 2.5 µl volume, and 4) expel 4 µl from the pipette tip into a clean tube to ensure complete retrieval of the initial 1.5 µl volume. Per ROI, the final collected volume was 9 µl. Each extract was immediately stored on-ice until further processing.

5.1.3.8 PAM Experiments

ROIs generated from spatial segmentation of MS images were cut using a scalpel. The pieces of parafilm M containing the tissue were then placed in a tube containing 10 µl of the extraction buffer and stored on-ice until further processing.

5.1.3.9 nanoLC-MS/MS

The extracts obtained with the LMJ or PAM strategy were sonicated for 5 min and the proteins were denatured at 55 °C for 15 min. The tubes were then quickly centrifuged to collect the extracts at the bottom of the tube. For extracts obtained using the PAM strategy, the parafilm pieces were carefully removed from the tubes using a pipette tip and the extracts were incubated at 95 °C for 10 min to ensure complete detergent dissociation. The tubes were then quickly centrifuged and stored on ice. 11 µl of 10% 0.4% FA in water were added to each tube so that the final ACN concentration is equal to the concentration of ACN at the beginning of the LC gradient. Samples were subjected to nanoLC-MS/MS analysis on the same day of sample preparation and were kept in the autosampler with the

thermostat set at 4 °C.

5 µl of the sample was loaded onto a 2 cm × 150 µm internal diameter IntegraFrit sample trap-column (New Objective, Woburn, Massachusetts, USA) at a maximum pressure 280 bar using a Proxeon EASY nLC-II (Proxeon, Thermo Scientific, Bremen, Germany). Proteins were separated on a 15 cm × 100 µm internal diameter PLRP-S column (Varian, Palo Alto, California, USA) with a linear gradient of ACN from 5 to 55% for 110 min and 55% to 90% for 25 min and a flow rate of 300 nL/min. Data were acquired on a Q-Exactive mass spectrometer (ThermoScientific, Bremen, Germany) equipped with a nanoESI (Electrospray Ionization) source (Proxeon, Thermo Scientific, Bremen, Germany) and a PicoTip nanospray emitter (New Objective, Woburn, Massachusetts, USA).

Data were acquired in data-dependent mode using a top 3 strategy. Full scans were acquired by averaging 4 microscans at 70,000 resolution within a m/z mass range of 800–2000 with an AGC target of 1×10^6 and a maximum accumulation time of 200 ms. The three most abundant ions with charge superior than 3 or unassigned were selected for fragmentation. Precursors were selected within a 15 m/z selection window by the quadrupole and fragmented with a Normalized Collision Energy (NCE) of 25; the (Automatic Gain Control) AGC target was set to 1×10^6 with a maximum accumulation time of 500ms. For each MS/MS spectrum, two microscans at a resolution of 70,000 at m/z 400 were acquired and averaged. Dynamic exclusion was set to 20 s.

5.1.3.10 Data Analysis

RAW files were processed with ProSight PC 3.0 (Thermo Fisher Scientific, Bremen, Germany). Spectral data were deisotoped using the cRAWler algorithm and searched against the complex *Homo sapiens* ProSightPC database version 2014_07 containing every canonical protein and its known PTMs. Files were searched with the "absolute mass" then "biomarker" search modes (Kellie *et al.*, 2010) in ProsightPC considering every PTM available in the complex database. A second search was performed to detect altORF products with a concatenated database composed of the *H. sapiens* UniProt Reference proteome (canonical and isoforms) of 01.16.2015 and an in-silico translated database of the *H. sapiens* of the transcripts from GenBank containing every ORF with potential protein product that had at least 29 amino acids with the same search strategy. Identification was considered positive when one of the two strategies gave an expected score (E-value) that was lower than 10^{-4}.

Raw files were also processed with ProSightPC 3.0 (Thermo Scientific) and Proteome Discoverer 2.1 (Thermo Scientific) utilizing the ProSightPD 1.0 node. Spectra were then searched using a three-tiered search tree. The first search was an Absolute Mass search with MS^1 tolerance of 100 Da and MS^2 tolerance of 10 ppm, against the complex Homo sapiens ProSightPC database version 2014_07 containing every canonical protein and its known PTMs. The second search was a ProSight Biomarker search with MS^1 tolerance of 10 ppm, MS^2 tolerance of 10 ppm, against the same database. Lastly, a second AbsoluteMass search was performed with MS^1 tolerance of 1000 Da, MS^2 tolerance of 10 ppm, using Delta M mode, against the same database.

False discovery rates (FDR) were estimated as described previously (Kellie *et al.*, 2011). Briefly, data were searched using scrambled protein sequences as decoys using identical strategies as above (absolute and biomarker modes). Logarithmic P-score distributions of decoy protein hits were analyzed for each search mode (absolute and biomarker) separately. Area under score distributions were calculated to reach 5% of total distribution starting from the best score (highest –log P), thus giving P-score cutoffs at 5% FDR for each search strategy. Proteins that had greater P-scores were removed from identification files.

UniProt accession numbers from each ovarian tissue technical replicate were combined and exported to UniProt "Retrieve/ID mapping" tool to recover files with accession numbers, Gene names and protein names (Supplementary Data 1). Venn diagrams were then generated by entering the UniProt combined accession number of each region into the University of Gent Venn diagram Web tool. The mass spectrometry top-down proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaíno *et al.*, 2015) partner repository with the dataset identifier PXD005420.

5.1.3.11 Subnetwork Enrichment Pathway Analyses and Statistical Testing

The gene names of identified proteins were used as input to retrieve a network from STRING (Szklarczyk *et al.*, 2014). The Elsevier's Pathway Studio version 10.0 (Ariadne Genomics, Elsevier) was used to deduce relationships among differentially expressed proteomics protein candidates using the Ariadne ResNet database (Bonnet *et al.*, 2009; Yuryev *et al.*, 2009). "Subnetwork Enrichment Analysis" (SNEA) algorithm was selected to extract statistically significant altered biological and functional pathways pertaining to each identified set of protein hits among the different groups. SNEA utilizes Fisher's statistical test set to determine if there are nonrandom associations between two categorical

variables organized by specific relationship. Integrated Venn diagram analysis was performed using “the InteractiVenn” : a web-based tool for the analysis of complex data sets (Heberle *et al.*, 2015).

5.1.3.12 Alternative Protein Validation

To validate the altprot product identified by top-down proteomics, one of the identified altprots was selected and cloned in the context of its reference protein. The plasmid contained the canonical G protein nucleolar 1 (GNL1) coding sequence with a C-terminal FLAG tag and the AltGNL1 coding sequence nested within the GNL1 coding sequence, but in a frameshifted ORF with a C-terminal V5 tag.

The DNA sequence was built using Gblocks which were assembled using the Gibson assembly (Gibson *et al.*, 2009) protocol using the NEBuilder HiFi DNA Assembly Cloning Kit (New England BioLabs, Ipswich, Massachusetts, USA) according to manufacturer’s recommendation into a pcDNA 3.1-expression vector (Invitrogen, Carlsbad, California, USA).

GNL1 and AltGNL1 co-expression was validated by western blot in HEK 293 cells transfected with polyethylenimine (PEI, Sigma) as transfection reagent (Hsu and Uludağ, 2012). Briefly, cells were grown in complete Dulbecco’s Modified Eagle’s Medium (DMEM, Wisent, St-Bruno, Québec, Canada) into a 6-well plate until 70-80% confluent. 1.6 µg plasmidic DNA was mixed into 80 µl of serum-free DMEM and 8 µl of 0.1% (w/v) PEI. The mixture was then gently mixed and let stand at room temperature for 10 min. 480 µl of complete media was then added to the mixture immediately and dropwise onto cells without media renewal. After 24 h of transfection, the cells were washed twice with phosphate-buffered saline (PBS) and lysed using 4% SDS. The lysate was sonicated and centrifuged and the protein content was estimated using Bicinchoninic Acid (BCA) assay (Thermo Scientific). 100 µg of protein was denatured and loaded onto a 15% SDS-PAGE gel. After migration, the proteins were transferred onto a polyvinyldiene difluoride membrane. The membrane was then blocked using 2.5%milk–supplemented, Tris-buffered saline with tween 20 (TBST, 1/1000 v/v). The membrane was rinsed with PBS and probed with anti-FLAG M2 mouse antibody (F1804, Sigma) and anti-V5mouse antibody (V8012, Sigma) overnight. The membrane was then washed three times with TBST and rinsed with PBS and probed with anti-mouse horseradish peroxidase antibody (7076S, Cell signaling technology) and visualized.

GNL1 and AltGNL1 co-expression was also validated at the cellular level by immunofluorescence. HeLa cells were transfected with GeneCellIn transfection reagent (BioCellChallenge, Toulon, France). Briefly, 25,000 HeLa cells per well were seeded into a 24-well plate and let to grow in complete DMEM media for 24 h. The cells were then transfected by adding 250 ng of plasmidic DNA into 100 µl of serum-free DMEM media and 1.5 µl of GeneCellIn transfection reagent. The mixture was let at room temperature for 15 min and then added dropwise to cells without media renewal. After 24 h of transfection, the cells were rinsed twice with PBS and fixed with 4% paraformaldehyde for 20 min and rinsed twice again. The cell membranes were permeabilized using 0.15% Triton X-100 for 5 min, then rinsed twice with PBS and then twice with Normal Goat Serum (NGS) blocking buffer for 20 min. Anti-FLAG rabbit antibody (F7425, Sigma) and anti-V5 mouse antibody were then added and cells were incubated overnight at 4 ° C. The cells were then rinsed twice with NGS and probed with anti-Mouse 488 antibody (A11017, Invitrogen) and anti-rabbit 568 antibody (A21069, Invitrogen) for an hour. The cells were rinsed twice with PBS and incubated with 4',6'-diamidino-2-phenylindole (DAPI) for 30 min. The cells were rinsed twice with PBS, mounted on a microscope slide with SlowFade (Thermo Scientific) and sealed. The slides were stored in the dark at 4 ° C until observation via confocal microscopy.

5.1.4 Results

5.1.4.1 Tumor Proteome Microenvironment Investigation

MALDI-MSI was performed on ovarian high grade serous carcinoma sections in order to perform non-supervised spatial segmentation analysis and identification of ROIs (Fig. 1a). The tissue was thus successfully classified by MALDI-MSI with three main clusters (Fig. 1b). One was associated with the benign region (blue-cyan), whereas the two others matched the tumor and necrotic/fibrotic tumor regions (brown-red and orange-yellow, respectively). The presence of the three regions was confirmed by a pathologist (Fig. 1b). A total of 18 samples from the three clustered regions were extracted and analyzed in triplicate employing the two extraction strategies (Fig. 1c). This resulted in the identification of 150 proteins in LMJ and 149 in PAM(Fig. 1c) at an estimated FDR of 5% for reference and "reference plus AltProts" concatenated protein databases (Supplementary Fig. 1a and b, respectively). The distribution using LMJ or PAM is as follows : 41 vs. 47 specific proteins in tumor regions, 24 vs. 27 in necrotic/fibrotic tumor regions, and 37 vs. 19 in benign regions (Fig. 1c). Overall, 61 proteins are specific to the tumor region, 44 to the necrotic/fibrotic tumor region and 48 to the benign region.

Thus, 237 proteins were identified by combining the data (see Supplementary Data 1) which is, to our knowledge, the highest number of identified proteins for tissue top-down microproteomics. From the list of identified proteins, some were already known to be involved in ovarian cancer (Table 1) but the ones we identified are mainly fragments of proteins e.g. a fragment of 58 amino acid residues derived from KRT8 (Fig. 1d). Among the identified proteins some are particularly interesting due to the fact that they are found in both tumor and necrotic-fibrotic tumor regions e.g. gamma-synuclein, Lupus la protein (SSB), Nucleophosmin (NPM1), Nuclease sensitive element-binding protein 1 (YBX1), Probable ATP-dependent RNA helicase DDX17 (DDX17), and Hematological and neurological expressed 1-like protein (JPT2). Others are found specifically in benign and necrotic/fibrotic tumor regions, such as salivary acidic proline-rich phosphoprotein 1/2 (PRH1). G antigen 7 (GAGE7), High mobility group protein B1 (HMGB1), Glycogen synthase (GYS1), G antigen 2B/2C (GAGE2B), and Cilia- and flagella-associated protein 44 (CFAP44) are only found in the necrotic/fibrotic tumor region.

Tableau 5.1 – **List of proteins and potential biomarkers identified within the necrotic/fibrotic tumor and tumor regions with referenced pathological involvement.** Here, the identification of the intact or fragmented form is emphasized (see also Supplementary Data 1).

Uniprot accession number	Protein name	Region	Ref.
P05787	Cytokeratin 8 (fragment)	Tumor, necrotic/fibrotic tumor	(Wang *et al.*, 2012)
P08729	Cytokeratin 7 (fragment)	Tumor	(Waldemarson *et al.*, 2012)
P14174	Macrophage migration inhibition factor	Tumor, necrotic/fibrotic tumor	(Hagemann *et al.*, 2005)
P23528	Cofilin-1 (fragment)	Tumor	(Li *et al.*, 2013)
P31949	Protein S100-A11	Tumor, necrotic/fibrotic tumor	(Liu *et al.*, 2015)
P46939	Utrophin (fragment)	Tumor	(Lomnytska *et al.*, 2006)
P53985	Monocarboxylate transporter 1 (fragment)	Necrotic/fibrotic tumor	(Chen *et al.*, 2010)
Q12906	Interleukin enhancer-binding factor 3 (fragment)	Necrotic/fibrotic tumor	(Guo *et al.*, 2012)
Q14247	Src substrate contactin (fragment)	Tumor, necrotic/fibrotic tumor	(Bourguignon *et al.*, 2001)
Q86Z02	Homeodomain-interacting protein kinase 1	Necrotic/fibrotic tumor	(Kondo *et al.*, 2003)
Q8NC51	Plasminogen activator inhibitor 1 RNA-binding protein (fragment)	Tumor, necrotic/fibrotic tumor	(Koensgen *et al.*, 2007)

STRING protein analysis of the tumor region associated with GO term analyses led to the identification of two major pathways (RNA binding (GO : 0003727), and poly(A) RNA

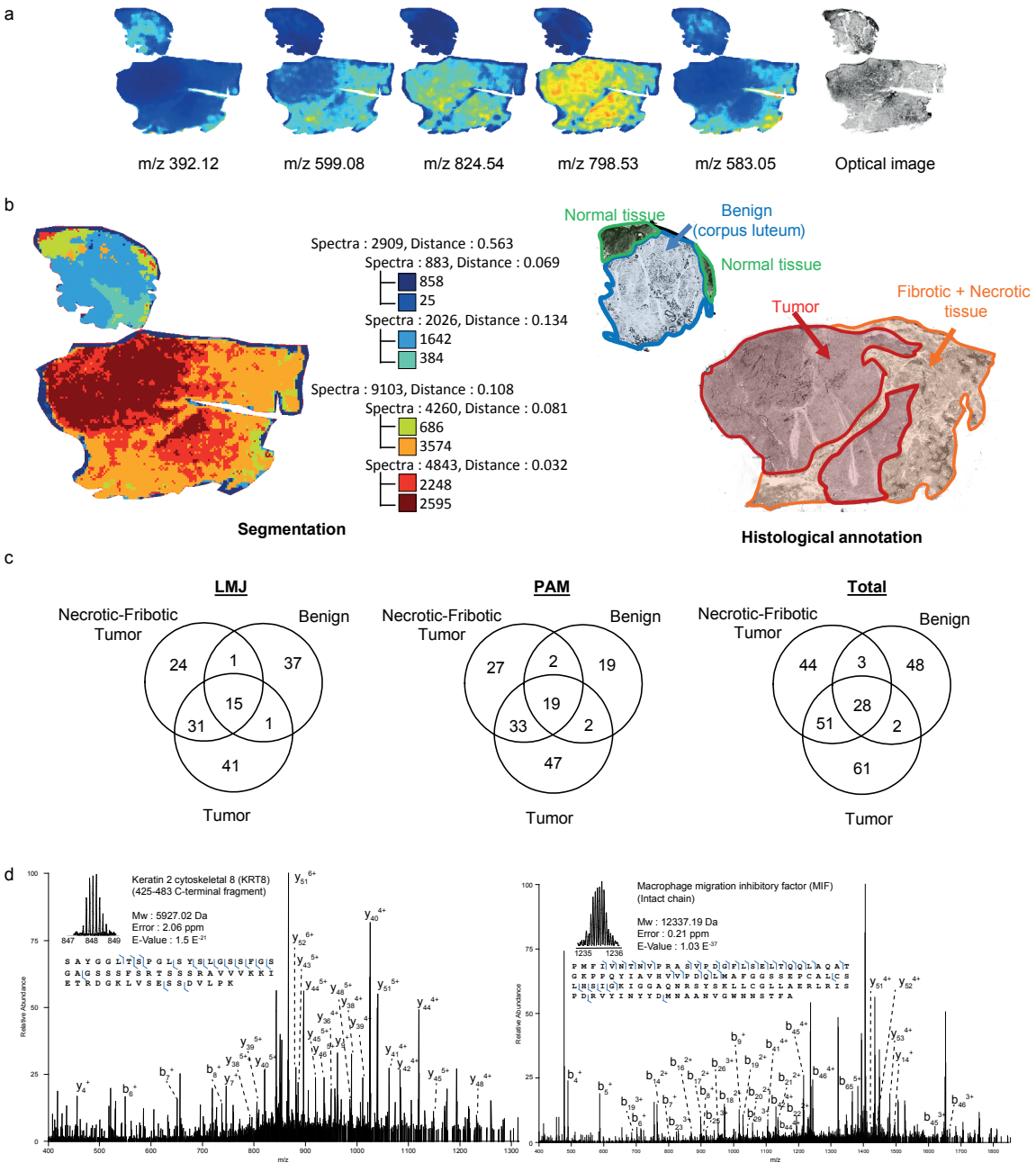


FIGURE 5.1 – **Association of MALDI-MSI and top-down microproteomics.**
(a) MALDI-MSI of lipids and optical image, (b) histological annotation and segmentation analysis using the Bisecting k- Means and Correlation Distance approach (left). (c) Venn diagram of the top-down gene reference products identified in the ovarian cancer tissue by the LMJ approach, the PAM approach and total, (d) Precursor and HCD fragmentation scan of Keratin 2 cytoskeletal 8 fragment 425-483 and Macrophage migration inhibitory factor (MIF).

binding (GO : 0044822). Cellular component GO overrepresentation analysis revealed that the proteins identified in the tumor are mainly found in exosomes - extracellular ve-

sicles (42.3%) and in the nucleus (57.6%). In the necrotic/ fibrotic tumor region, 50% of the proteins are found in the extracellular exosomes and in vesicles, 16% are in the nucleus and 34% in various organelles. In benign regions, 50% of the proteins are involved in cellular traffic, 12.3% are cytoplasmic and 27.7% are membrane-bound proteins. Global subnetwork analyses in both tumor, necrotic/fibrotic tumor and benign regions clearly showed differences in protein pathway involvement (Fig. 2). In the benign region, protein pathways are mainly implicated in cell survival, growth, motion, adhesion, differentiation and vascularization (Fig. 2a), whereas in the necrotic/fibrotic tumor region the proteins are mainly implicated in apoptosis, inflammation, neoplasm, acute phase reaction and oxidative stress (Fig. 2b). Tumor subnetwork global analysis showed pathways in neoplasm, autophagy, apoptosis, cell proliferation and tumor immunity (Fig. 2c).

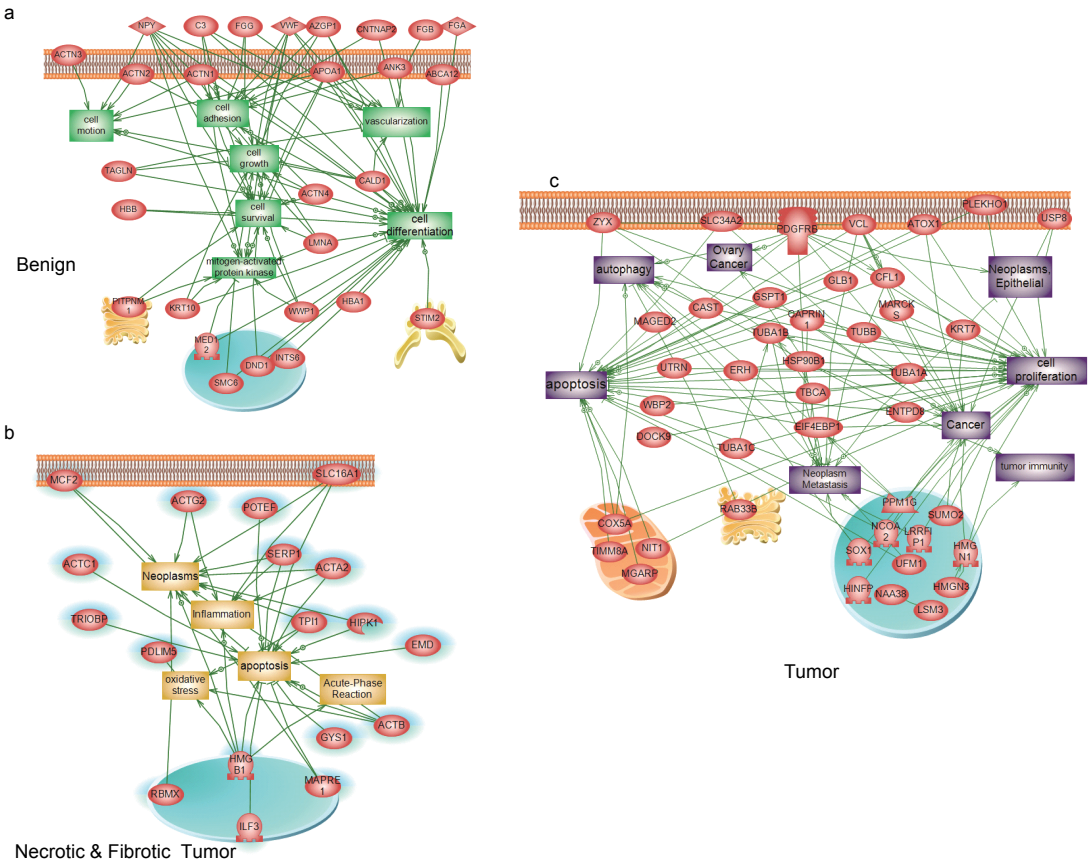


FIGURE 5.2 – **Systems biology analysis**
Global network identification of the proteins present in benign (a), necrotic-fibrotic tumor (b) and tumor regions (c).

Subnetwork enrichment analysis confirmed the global analysis (Fig. 3). In the benign region, the subnetworks revealed implication in muscle contraction and cell differentiation

(Fig. 3a). For necrotic/fibrotic tumor, the subnetworks are involved in Rho Kinase, RNA processing and Rhabdomyocyte disease pathways (Fig. 3b). Tumor subnetworks revealed implication in necrosis, interferon regulatory factor signaling pathway, myonecrosis, and cancer and T cell hypo responsiveness (Fig. 3c). Global network analysis between benign and necrotic/fibrotic tumor regions (Fig. 3d) revealed proteins involved in apoptosis, contraction and actin organization pathways. The same analysis between tumor and necrotic/fibrotic tumor revealed proteins involved in neoplasm and Smooth Muscle Cell (SMC) proliferation pathways. Comparison of proteins from tumor and benign regions showed proteins involved in cell death, cell growth, keloid and muscle cell differentiation.

5.1.4.2 Hidden Proteome : Alternative Proteins

We previously identified 6 AltProts using the shot-gun proteomic approach i.e. AltADCY1, AltCCDC152, AltKART34, AltMOBKL2B, AltPALLD, AltSMCHD1 (Vanderperre *et al.*, 2013). With the top-down microproteomics approach, 15 unknown proteinswere identified in patient biopsies including : AltApol6, AltCMBL, AltTLR5, AltPKHD1L1, AltLARS2-AS1, AtltSERPINE1, AltCSNK1A1L, AltGPC5, AltLTB4R, AltTMP1, AltGRAMD4, AltMTHFR, AltAGAP1, AltGNL1 and AltRP11- 576E20.1 (Table 2). Six altprots were identified in the benign region (AltTLR5, AltPKHD1L1, AtltSERPINE1, AltGPC5, AltGRAMD4, AltAGAP1), 5 in the necrotic/fibrotic tumor region (AltApol6, AltLARS2-AS1, AltLTB4R, AltTMP1, AltMTHFR) and 4 in the tumor (AltCMBL, AltGNL1, AltRP11-576E20.1, AltCSNK1A1L). The function of these proteins remains unknown. AltGNL1 was selected for further analysis (Fig. 4) based on immunofluorescence data provided by the Human Protein Atlas confirming the presence of its reference protein GNL1 in ovarian cancer tissue.

In order to validate the co-expression of GNL1 and nonannotated AltGNL1 proteins from the same gene, we transfected cells with an expression plasmid containing a GNL1-FLAG$^{(V5)}$ construct in HEK 293 cells (Fig. 5). In this construct, the Flag and V5 tags are in frame with GNL1 and AltGNL1, respectively. Both GNL1^{FLAG} and AltGNL1^{V5} are expressed and detected with anti-FLAG and anti-V5 antibodies, respectively (Fig. 5a & b). Co-expression at single cell level was confirmed by immunofluorescence (Fig. 5c). AltGNL1 displays a nuclear localization whereas GNL1 is present in the cytosol.

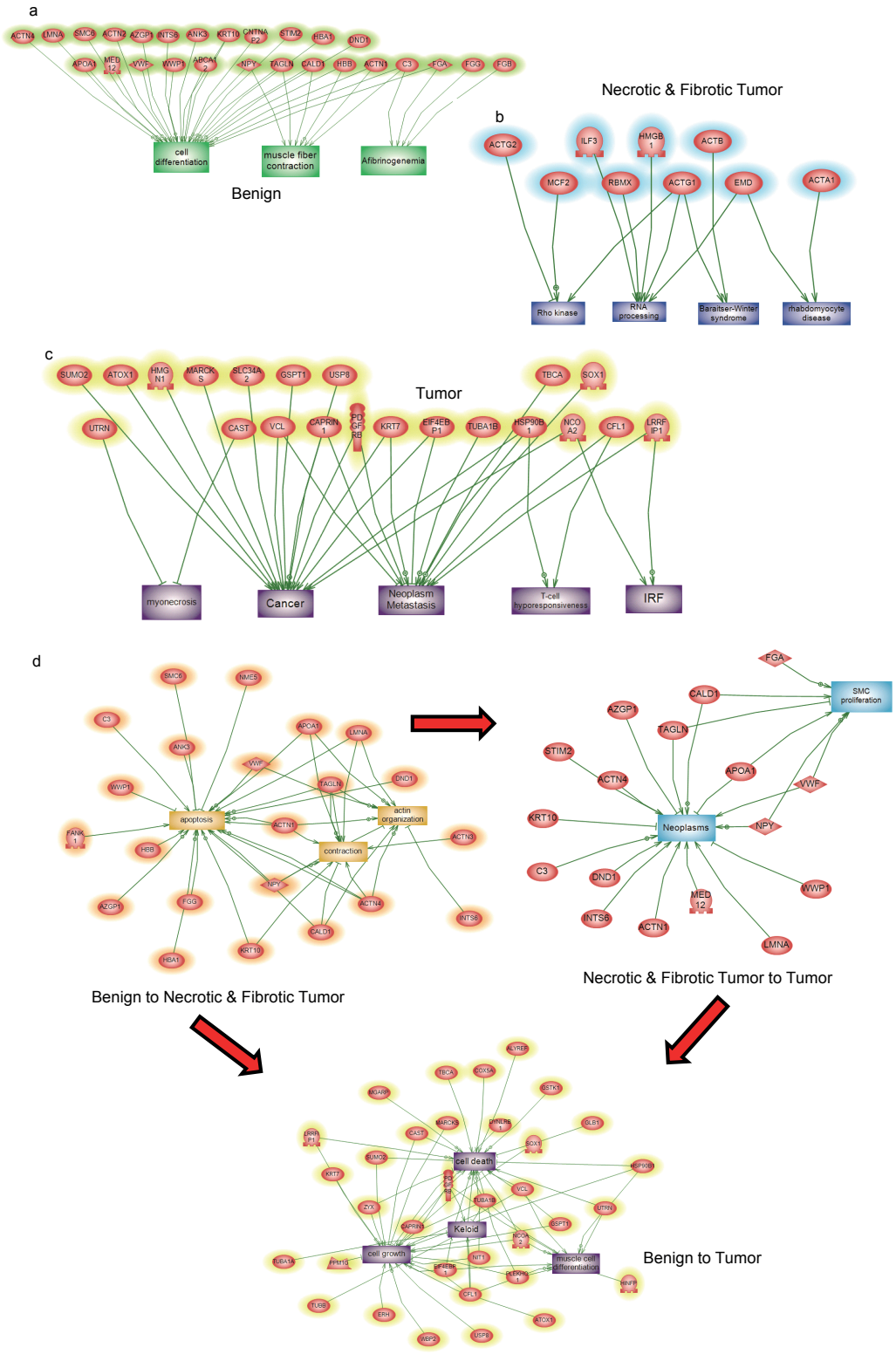


FIGURE 5.3 – **GO Enrichment analysis**
GO Enrichment analysis of benign (a), necrotic/fibrotic tumor (b) and tumor region (c). (d) Global network analysis between benign and necrotic-fibrotic tumor region, between tumor and necrotic-fibrotic tumor, and between benign and tumor showed proteins involved in cell death, cell growth, keloid and muscle cell differentiation.

Tableau 5.2 – **Alternative protein products from ovarian cancer biopsies identified by tissue top-down microproteomics.**
AltORF localization are either "CDS" for AltORFs nested within the canonical CDS but in a frameshifted ORF, "3'" for AltORF located in the 3'UTR region, CDS-3' for AltORF overlapped in the CDS and 3'UTR region or "ncRNA" for AltORFs found in putative non-coding transcripts.

E-value (P-score)	Theoretical mass	Gene	Transcript	Tissue	AltORF localization
9.34E-06 (-3.50E-11)	3864.85	APOL6	NM_030641.3	LMJ-necrotic-fibrotic tumor	3'
5.88E-05 (-2.20E-10)	3802.9	CMBL	NM_138809.3	PAM-tumor	3'
2.15E-05 (-8.04E-11)	4332.03	TLR5	NM_003268.5	LMJ-benign	CDS
3.91E-06 (-1.46E-11)	2814.3	PKHD1L1	NM_177531.4	PAM-benign	3'
8.36E-06 (-3.13E-11)	2509.14	GNL1	NM_005275.3	PAM-tumor	CDS
6.67E-07 (-1.25E-15)	4977.49	LARS2-AS1	NR_048543.1	LMJ-necrotic-fibrotic tumor	ncRNA
7.63E-06 (-1.43E-11)	3564.83	RP11-576E20.1	XR_241690.1	PAM-tumor	ncRNA
4.58E-06 (-8.59E-12)	3943.89	SERPINE1	NM_000602.4	LMJ-benign	CDS
1.75E-06 (-3.28E-12)	5429.66	CSNK1A1L	NM_145203.5	LMJ-tumor	3'
6.23E-06 (-1.17E-11)	4832.21	GPC5	NM_00446.5.1	LMJ-benign	CDS
6.12E-07 (-1.15E-12)	4873.55	LTB4R	NM_001143919.2.1	LMJ-necrotic-fibrotic tumor	3'
2.95E-06 (-5.54E-12)	3875.9	TMP1	XM_005254648.1.1	LMJ-necrotic-fibrotic tumor	3'
1.76E-06 (-3.30E-12)	4911.42	GRAMD4	XM_005261398.1.1	LMJ-benign	3'
3.39E-06 (-6.36E-12)	4852.45	MTHFR	NM_005957.4	LMJ-necrotic-fibrotic tumor	CDS-3'
8.57E-06 (-1.61E-11)	4908.42	AGAP1	XM_006712240.1	LMJ-benign	3'

5.1.5 Discussion

This work involves the use of tissue microproteomics to characterize the local proteome in three regions (necrotic/fibrotic tumor, tumor and benign region) of human ovarian cancer. These regions were analyzed by MALDI-MSI and discerned by spatial segmentation analysis (Alexandrov *et al.*, 2011; Bonnel *et al.*, 2011; Bruand *et al.*, 2011), and the proteins were microextracted utilizing LMJ and PAM approaches (Franck *et al.*, 2013; Quanico *et al.*, 2015; Wisztorski *et al.*, 2016). A total of 237 gene products within the three region swere identified. 61 proteins were specific to the tumor region, 44 to the necrotic/ fibrotic tumor region, and 48 to the benign region. The extracted protein profiles from the 3 regions are clearly different and subnetwork analysis revealed a possible progression in the nature of the protein pathways involved in the 3 regions. These results suggest a mechanism in cancer progression from benign to tumor and necrotic/fibrotic tumor regions by a progressive switch in the cell phenotype because we detected proteins common to these regions e.g. SSB, NPM1, YBX1, DDX17, HN1L or PHR1, HMGB1, GYS1, GAGE2B, CFAP44. Utilizing a systems biology approach, pathways implicated in muscle proliferation, cell differentiation, actin, cytoskeleton disorganization, apoptosis, neoplasia, and necrosis with Rho kinase activation are enriched and are likely to be involved in the switch in cell phenotype. In addition, T cell response is observed to be inhibited, leading a tolerant immune response towards the tumor. These results are consistent with spatial segmentation analysis showing that the tumor and necrotic-fibrotic tumor regions had a close histological molecular profile distinct from that of benign regions (see cluster tree, Fig. 1a).

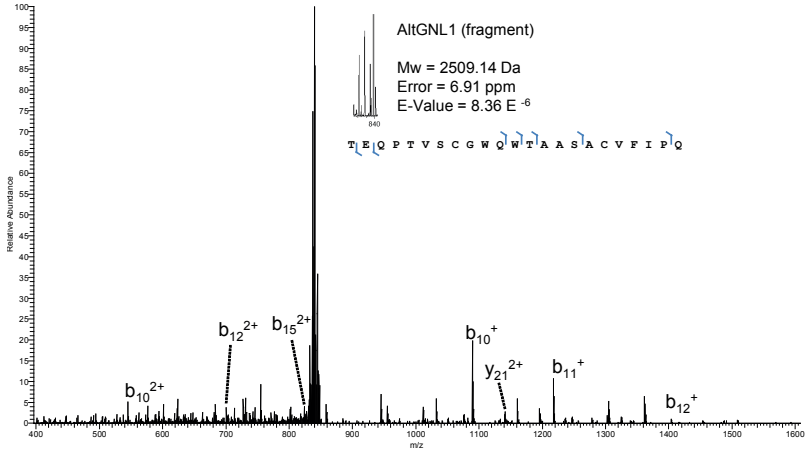


FIGURE 5.4 – **Precursor and HCD fragmentation scan of Alternative Guanine Nucleotide-binding Protein-like 1 (AltGNL1)**

Tissue top-down microproteomics gives insight on the tumor microenvironment with the identification of proteins involved in cancer processes, diagnosis and/or progression (Table 1). For example, the C terminal fragment (aa425–483) of Cytokeratin-8 (KRT8) has been detected in our experiments in the necrotic/fibrotic tumor and tumor regions (Fig. 1d). KRT8 was previously referenced as a potential biomarker for ovarian cancer (Wang *et al.*, 2012). We demonstrate here that in cancer regions, it is not the complete protein that is present but a C-terminal fragment of 58 amino acid residues.We previously obtained similar results for the C-terminal fragment of the immunoproteasome 11S, PA28 or Reg alpha, a marker for Grade III-IV serous ovarian cancer (Lemaire *et al.*, 2007a), as well as for Grade I and tumor relapse (Longuespée *et al.*, 2012). Similarly, a fragment (aa55–72) of Cytokeratin-7 (KRT7) was detected in the tumor region. KRT7 is already a marker for ovarian adenocarcinoma (Chu *et al.*, 2000; Waldemarson *et al.*, 2012), but here we demonstrate that, in fact, the fragment composed of 17 amino acid residues is potentially the actual marker in ovarian tumor. KRT8 and 7 were also reported to be highly expressed in ovarian cancer cell lines (Chu *et al.*, 2000). Protein S100- A11 was detected in the tumor and necrotic/fibrotic tumor regions and was also observed as being particularly highly expressed in ovarian cancer (Liu *et al.*, 2015). The pro-inflammatory cytokine Macrophage migration inhibitory factor (MIF) was detected in the necrotic/fibrotic tumor and tumor regions (Fig. 3d). MIF is already a potential biomarker for ovarian cancer and is associated with tumor growth, metastasis and poor prognosis (Simpson *et al.*, 2012). This protein is also a serum biomarker that distinguishes benign from malignant ovarian tumors in combination with other biomarkers (Agarwal *et al.*, 2007; Krockenberger *et al.*, 2008), and is associated with loss of p53 suppressor activity (Hudson *et al.*, 1999), inhibiting apoptosis and DNA damage repair. Several other proteins already linked to cancer were also identified, including nitrilase-1(Nit1), melanoma antigen family D 2 (MAGED2), Zyxin (ZYX), and ATX1 antioxidant protein 1 homolog (ATOX1). Nit1 is a negative regulator in primary T cells and is classified as a tumor suppressor in association with the fragile histidine-triad protein Fhit (Semba *et al.*, 2006) over-produced in non-small cell lung cancer (NSCLC) and may be a therapeutic target in ovarian cancer (Croce *et al.*, 1999). MAGED2 is also over-expressed in NSCLC (Sienel *et al.*, 2004). Zyxin, a Smad3-mediated TGF-β1 signaling target, regulates cancer cell motility and epithelial-mesenchymal transition during lung cancer development and progression (Beaino *et al.*, 2014; Mise *et al.*, 2012). Interestingly, some proteins identified in the present work have not yet been identified by the Cancer Network Galaxy (TCNG) e.g. Dolichyl-diphosphooligosaccharide-protein glycosyltransferase subunit 4 (OST4), Signal recognition particle receptor subunit alpha (SRPRA), and U6 snRNA-associated Sm-like protein LSm8 (LSM8).

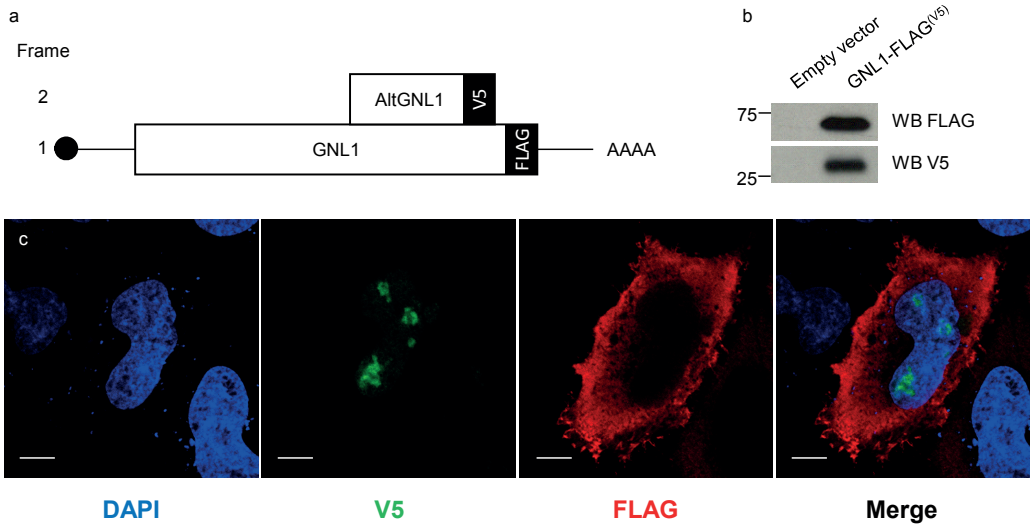


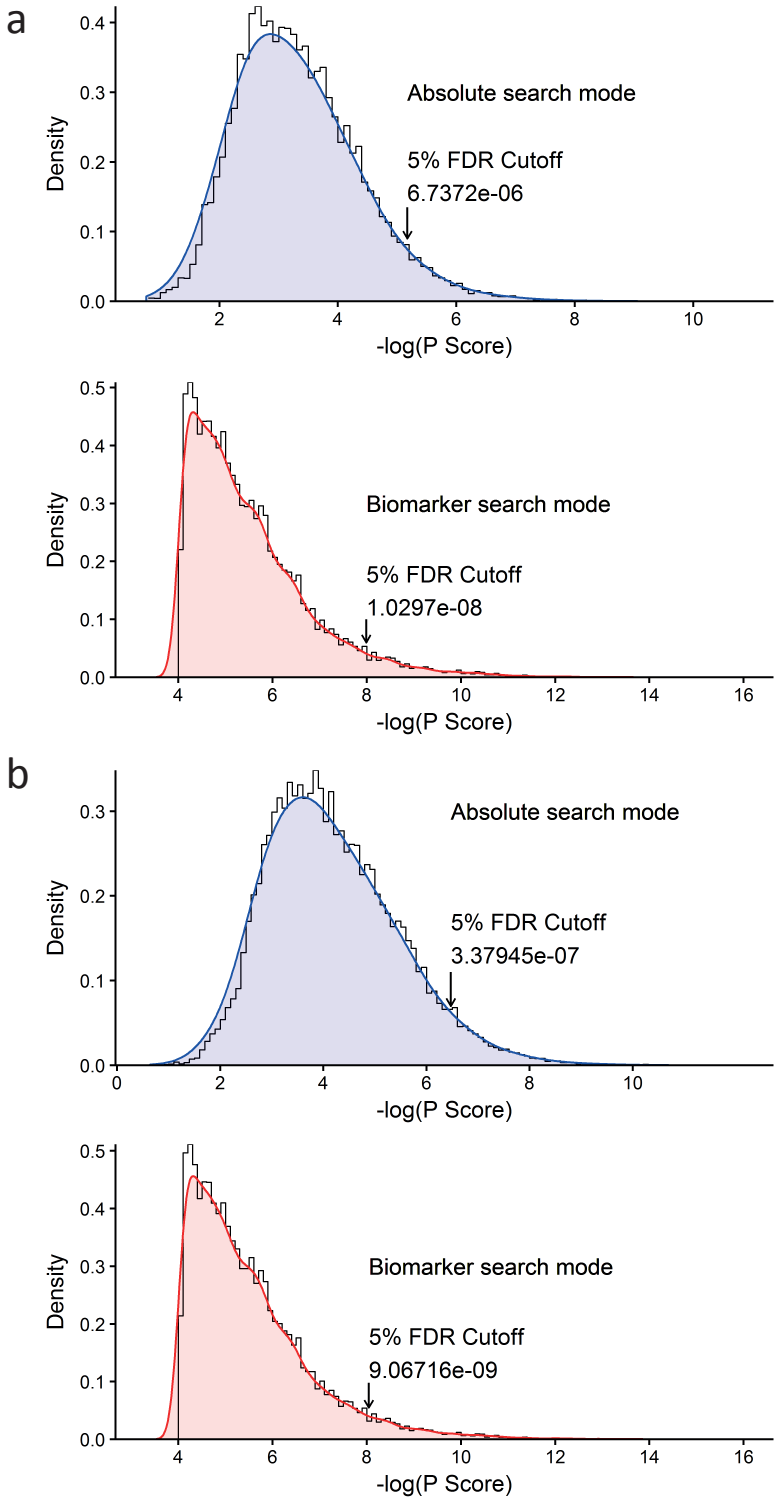
FIGURE 5.5 – **Validation of co-expression of reference protein GNL1 and its alternative protein AltGNL1**
(a) Schematic representation of themRNA product fromGNL1 (AltGNL1) plasmid used for validation. (b)Western blot showing co-expression of GNL1 (FLAG tagged) and AltGNL1 (V5 tagged) and (c) immunofluorescence assay showing co-expression at cell level with nucleus staining (DAPI, blue) AltGNL1-V5 (green), GNL1 (red) and merge panel. White bars represent 10 µm.

In addition to reference proteins,we identified altprots by top-down microproteomics. 6 altprots were detected in the benign region, 5 in the necrotic/fibrotic tumor region, and 4 in the tumor region. None of these 15 altprots were previously identified. Genes coding for these altprots are annotated as genes coding for receptors (TLR5, LTB4R, AGAP1R), enzymes (CMBL, SERPINE1, MTHF, CSNK1A1L, TMP1), or cytoplasmic or nuclear proteins (Apol6, GRAMD4, GNL1, PKHD1L1). AltLARS2-AS1 and AltRP11-576E20.1 are expressed from genes annotated as non-coding genes, and thus should be re-annotated.We focused our interest on AltProts detected in the cancer region, specifically AltGNL1. Indeed, the reference GNL1 protein was previously detected in ovarian cancer according to the Protein Atlas. None of the other 13 reference proteins were previously identified in proteomic or genomic large-scale studies on ovarian cancer. We validated the co-expression of the reference GNL1 protein with its AltProt AltGNL1 (Fig. 5b–c). Our results clearly demonstrate that both proteins are co-expressed from a single mRNA expressed from a cDNA construct. Immunofluorescence experiments showed that AltGNL1 displays nuclear localization whereas GNL1 is present in the cytosol (Fig. 5c). Our results confirm the presence of a hidden proteome which can constitute a reservoir of potential biomarkers and therapeutic targets.

Taken together, our results show that top-down microproteomics coupled with MALDI MSI can be used to detect proteins expressed from altORFs. These proteins can be used as putative diagnostic biomarkers that may have been missed in conventional proteomics approaches utilizing reference protein databases only. Our approach will be useful to determine the function of altprots in health and disease. Supplementary data to this article can be found online at http ://dx. doi.org/10.1016/j.ebiom.2017.06.001.

5.1.6 Funding sources

Supported by grants from Région Nord Pas-de-Calais (CM/YB N°2015.2097/12) and PROTEO (FRQNT-RS-188158) (V. Delcourt), University Lille 1 (BQR to Dr. Julien Franck, 2015), Canadian Institutes for Health Research (MOP-136962) and Canada Research Chairs in Functional Proteomics and Discovery of New Protein (Prof. X. Roucou), PRISM (Prof. M. Salzet), Ministère de l'Enseignement Supérieur et de la Recherche via Institut Universitaire de France (ESRS 0900500E) (Prof. I. Fournier), SIRIC ONCOLille (Prof. I. Founier), and Grant INCa-DGOS-Inserm 6041.



Supplementary Figure 1 : FDR estimation via interrogation of scrambled databases. Area of densities of highest negative logarithm of P-score distributions gave P-score cutoffs for absolute and biomarker search modes at 5% FDR. Results were generated using “reference” (a) and concatenated “reference and AltORFs” protein databases (b). P-score cutoffs were reported on density plots.

5.2 Conclusion et perspectives

Les stratégies de microprotéomique donnent accès à des informations relative au microenvironnement cellulaire, aspect particulièrement important pour l'analyse de tissus pathologiques. En effet, le microenvironnement cellulaire des tissus cancéreux contient les cellules malignes qui participent au développement de l'affection, mais également les cellules environnantes et la matrice extracellulaire. L'analyse de ces régions de tissus permet d'observer les interactions des cellules malignes *in situ*, d'évaluer l'expression de protéines qui favorisent l'évolution du cancer ou témoignant des mécanismes de défense de l'organisme. La stratégie d'extraction LMJ a déjà été appliquée à l'analyse de cancer de l'ovaire par approche *bottom-up*. Son application à l'étude de tissus fixés de cancer de l'ovaire par la stratégie LMJ s'était avérée efficace pour détecter de nombreuses protéines impliquées dans le développement tumoral et certains biomarqueurs protéiques (Wisztorski *et al.*, 2013). De façon similaire, la stratégie PAM associée à l'imagerie MALDI s'est quant à elle distinguée lors de l'analyse de tissus cancéreux de prostate. Son application a notamment révélé que l'expression de certaines protéines était modulée en fonction de leur localisation, entrainant l'activation ou la diminution de certaines voies biologiques dans les tissus tumoraux (Quanico *et al.*, 2015).

Par sa complémentarité vis-à-vis de l'approche *bottom-up*, l'application de la stratégie développée lors de l'article 2 pour l'étude de tissus cancéreux pourrait se révéler puissante pour détecter des biomarqueurs protéiques de petite taille. De plus, quelques études ont par ailleurs démontré les performances de l'approche *top-down* pour le diagnostic de pathologies telles que le diabète (Mao et Wang, 2014) ou la stéatohépatite non acloolique (Sarsby *et al.*, 2014). Dans ce contexte, nous avons appliqué la stratégie d'étude de tissu de l'article 2 par la définition de régions d'intérêt d'une biopsie d'ovaire cancéreux de haut grade séreux et par l'extraction de protéines intactes au sein de ces régions.

Les expériences d'imagerie MALDI de métabolites ont permis de distinguer trois régions d'intérêt principales. Après examen pathologique de sections de tissus adjacentes, un pathologiste a d'une part, confirmé avec les délimitations des régions d'intérêt du tissu avec une correspondance parfaite avec l'imagerie MALDI-MS, mais aussi défini le contexte pathologique de chacune des trois régions. L'imagerie par MALDI-MS s'est donc démontrée particulièrement efficace pour délimiter les régions tumorales, nécrotiques et bénignes de la biopsie d'ovaire, confirmant ses aptitudes en histologie moléculaire pour la caractérisation de tissus pathologiques.

Les extractions localisées selon les deux techniques de microprotéomique (LMJ et PAM)

réalisées pour chacune des régions d'intérêt en triplicats techniques ont permis d'identifier 237 protéines différentes. Certaines de ces protéines sont spécifiquement identifiées dans des régions du tissu et d'autres sont partagées par plusieurs régions. Il est important de souligner de nombreuses protéines sont communes aux régions tumorales et nécrotiques, ce qui suggère que ces régions de tissu présentent des similarités moléculaires. Ces similarités sont par ailleurs également suggérées par la classification hiérarchique d'imagerie MALDI.

Parmi les protéines référence identifiées, certaines ont déjà été décrites pour leur expression élevée dans des lignées cellulaires ou des tissus de cancer de l'ovaire ou leur implication dans le développement de la pathologie. En effet, la protéine Cofilin-1, associée à la résistance au Taxol, a été identifiée dans la région tumorale (Li *et al.*, 2013) ; les protéines S100-A11 et cortactine, détectées dans les régions tumorales et nécrotiques, sont associées à la migration et l'invasion des cellules tumorales (Liu *et al.*, 2015; Bourguignon *et al.*, 2001). La protéine inhibitrice de la migration des macrophages (MIF) a été détectée dans les régions tumorales et nécrotiques. MIF est une cytokine pro-inflammatoire associée à la croissance tumorale (Hagemann *et al.*, 2007), le développement de métastases (Simpson *et al.*, 2012) et un pronostic défavorable. Enfin, cette protéine est un biomarqueur potentiel du cancer de l'ovaire en association avec d'autres biomarqueurs protéiques (He *et al.*, 2012). Ces résultats démontrent que la méthode de préparation d'échantillon compatible avec l'approche *top-down* combinée à l'imagerie MALDI est une stratégie efficace pour la détection et éventuellement la découverte de biomarqueurs protéiques.

De façon similaire, l'application de cette méthodologie a également permis d'identifier de nouvelles protéines alternatives encodées à partir de cadres de lecture ouverts d'ARNms ou d'ARNnc à partir des trois régions étudiées et des deux modes d'extraction LMJ et PAM. Parmi celles-ci, une protéine alternative a été sélectionnée pour la validation de la coexpression avec sa protéine de référence *in cellulo*. AltGNL1 est une protéine alternative identifiée dans la région tumorale du tissu d'ovaire encodée à partir d'un cadre de lecture décalé du CDS de GNL1. La protéine de référence, GNL1, est référencée comme protéine extracellulaire ou nucléaire (UniProt, 2017), ou cytoplasmique et nucléolaire (Boddapati *et al.*, 2012). Un plasmide d'expression hétérologue a été construit pour la validation. Il comporte le CDS complet de GNL1 étiqueté par un FLAG à son extrémité C-terminale et, dans un cadre de lecture décalé, l'ORF d'AltGNL1 étiqueté par un V5. L'insertion de l'étiquette V5 au sein du CDS de GNL1 entraine une modification de sa séquence, cette construction ne pourrait donc pas être employée pour l'étude de la pro-

téine de référence. Cependant, l'insertion du V5 ne fait pas apparaître de codon STOP dans le cadre de lecture de GNL1, condition requise pour étudier la coexpression des deux protéines. Les expériences de WB après surexpression de ce plasmide dans des cellules HEK 293 ont confirmé la coexpression des deux protéines. Cependant le poids moléculaire reporté de la protéine alternative AltGNL1 est supérieur à son poids moléculaire prédit ce qui suggère que cette protéine alternative serait encodée grâce un codon d'initiation non canonique en amont de l'AUG prédit. Les expériences d'IF ont quant à elles permis de déterminer leurs localisations cellulaires dans des cellules HeLa. La protéine de référence est détectée principalement au sein du cytoplasme des cellules tandis que la protéine alternative semble être présente au sein des nucléoles. La localisation de la protéine de référence pourrait être influencée par la présence d'une étiquette V5 dans un cadre de lecture décalé. Toutefois la présence de la protéine alternative au sein des nucléoles est remarquable car cette localisation est partagée avec sa protéine de référence selon la littérature (Boddapati *et al.*, 2012). La localisation de protéines alternatives et protéines de référence encodées à partir du même gène est déjà rapportée dans la littérature (Abramowitz *et al.*, 2004; Bergeron *et al.*, 2013). Cette association est aussi compatible avec la fréquente observation de relation fonctionnelle entre protéine alternative et protéine de référence (Mouilleron *et al.*, 2015). De plus, la protéine de référence n'est pas détectée dans les tissus normaux d'ovaires tandis qu'elle est parfois détectée dans des tissus cancéreux d'ovaires alors que nous avons détecté la protéine alternative dans la région tumorale du tissu (Uhlén *et al.*, 2015). L'expression élevée de ces deux protéines pourrait constituer une signature de la pathologie, ce qui reste toutefois à valider.

Un autre aspect important qui découle de l'éventuelle coexpression de deux protéines à partir d'un unique gène et leur éventuelle relation fonctionnelle est l'évaluation des niveaux d'expression de l'une par rapport à l'autre et donc leur stœchiométrie. L'approche *top-down*, même si elle est particulièrement efficace pour la détection de petites protéines et identifie de nouvelles protéines alternatives, n'est pas encore associée à un mode de quantification absolue robuste. Or, la détermination de la quantité précise de protéines dans des échantillons biologiques requiert souvent des techniques de protéomique ciblées avec des étalons internes par approche *bottom-up*. De plus, la détermination de la relation stoechiométrique qui existe entre protéine de référence et alternative d'un même gène n'a pas encore été réalisée. Il apparaît primordial de développer une approche employant la quantification absolue par protéomique ciblée pour déterminer cette relation à partir d'un gène modèle connu pour exprimer deux protéines.

6 ARTICLE 4

A protein translated from a short ORF originally excluded from gene annotations is the main translation product of *MIEF1*

Auteurs de l'article: Vivian Delcourt, Mylène Brunelle, Annie Roy, Jean-François Jacques, Michel Salzet, Isabelle Fournier, Xavier Roucou

Statut de l'article: En rédaction

Avant-propos: Dans cet article, nous établissons pour la première fois la relation de stœchiométrie entre une protéine de référence et une protéine alternative encodées par le même gène, MiD51 et altMiD51. Pour ce faire, nous employons des techniques de quantification absolue par étalon interne (AQUA) et nous mesurons les quantités absolues des deux protéines dans divers lignées cellulaires et tissus.
Dans cette étude, sous la supervision de mes encadrants, j'ai mené le développement analytique de la méthode de quantification par spectrométrie de masse ciblée *PRM*, réalisé le paramétrage du spectromètre de masse et les diverses préparations d'échantillons, analyses bioinformatiques et de quantification. Pour la caractérisation des lignées éditées génétiquement par CRISPR-Cas9, les lignées cellulaires clonales ont été générées par Mylène Brunelle. J'ai validé l'édition des deux protéines par *western-blot* et MS ciblée. J'ai également généré, avec l'aide de Jean-François Jacques, le séquençage des allèles édités des cellules HeLa-KO-MiD51. J'ai enfin bénéficié de l'aide de nos collaborateurs pour certaines préparations d'échantillons.
Sous la direction de mes encadrants, j'ai participé à la rédaction de l'article, à la création de figures et tableaux. Cet article est actuellement en cours de rédaction.

Résumé: Les approches de protéogénomique et de profilage ribosomal démontrent de manière concomitante que les gènes peuvent coder pour une grande et une ou plusieurs petites protéines encodées à partir de séquences codantes annotées (CDS) ou alternatives non-annotées (altORFs). Toutefois, la stoechiométrie entre une grande et petite protéine traduite à partir du même gène reste inconnue. *MIEF1* est un gène codant qui a été récemment décrit pour encoder deux protéines, comporte un CDS et un altORF récemment annoté localisé dans le 5'UTR. Le profilage ribosomal et la protéomique par MS ont

démontré que cet altORF est effectivement traduit. De ce fait, *MIEF1* est un gène modèle codant à la fois pour une grande et une petite protéine. Dans cette étude, nous utilisons la quantification absolue à l'aide de peptides synthétiques comportant des isotopes stables et le suivi des réactions en parallèle (PRM) pour quantifier les deux protéines dans deux lignées cellulaires et un tissu de colon. Nous démontrons que le principal produit de traduction de *MIEF1* n'est pas la protéine canonique MiD51 de 463 acides aminés mais la petite protéine altMiD51 de 70 acides aminés. Nos résultats démontrent que le concept du CDS unique est inadéquat et donne un argument fort en faveur de la modernisation de l'annotation fonctionnelle des gènes.

A protein translated from a short ORF originally excluded from gene annotations is the main translation product of *MIEF1*

Vivian Delcourt, Mylène Brunelle, Annie Roy, Jean-François Jacques, Michel Salzet, Isabelle Fournier, Xavier Roucou

En rédaction

6.1 Manuscrit

6.1.1 Abstract

Proteogenomics and ribosome profiling concurrently show that genes may code for both a large and one or more small proteins translated from annotated coding sequences (CDSs) and unannotated alternative open reading frames (altORFs), respectively, but the stoichiometry between large and small proteins translated from the same gene is unknown. *MIEF1*, a protein-coding gene recently identified as a dual-coding gene, harbours a CDS and a newly annotated altORF located in the 5'UTR actively translated as shown by ribosome profiling, and MS-based proteomics. Thus, *MIEF1* is a prototypical gene coding for both a large and a small protein. Here, we use absolute quantification with stable isotope-labeled peptides and parallel reaction monitoring to determine levels of both proteins in two human cells lines and in human colon. We report that the main *MIEF1* translational product is not the canonical 463 amino acid MiD51 protein but the small 70 amino acid alternative MiD51 protein (altMiD51). These results demonstrate the inadequacy of the single CDS concept and provide a strong argument for modernizing functional annotations of genes.

6.1.2 Introduction

According to the traditional view of protein synthesis, each protein-coding gene harbours a single annotated ORF or coding sequence (CDS) encoding a canonical protein. However, genes contain more than one ORF, and the longest ORF is generally designated as the canonical CDS in genome annotations (Dinger *et al.*, 2008). In eucaryotes, alternative splicing results in the production of several mRNAs and the translation of different isoforms, in addition to the canonical protein. Hence, the translational output of a protein-coding gene is currently consealed to a canonical protein and one or several isoforms.

This concept was recently disproved by two modern approaches for the accurate measurement of translation, ribosome profiling and proteogenomics. Ribosome profiling maps

the regions of the transcriptome which are actively translated with nucleotide resolution (Brar et Weissman, 2015). Proteogenomics approaches use customized protein databases and mass spectrometry (MS)-based proteomics to detect translated proteins (Vanderperre *et al.*, 2013; Menschaert *et al.*, 2013; Ma *et al.*, 2014; Koch *et al.*, 2014; Bazzini *et al.*, 2014; Nesvizhskii, 2014; Ma *et al.*, 2016; Olexiouk et Menschaert, 2016). Both methods have revealed prevalent translation of ORFs outside of annotated CDSs, and of out-of-frame ORFs (Ingolia *et al.*, 2009; Brar et Weissman, 2015). For clarity, we term currently non-annotated ORFs, alternative ORFs or altORFs. These findings call into question the concept of the single CDS in eukaryotic mRNAs (Mouilleron *et al.*, 2015). In addition, they also highlight the need to redefine translated sequences (Ingolia, 2016), modernize functional genome annotations with shorter ORFs (Delcourt *et al.*, 2017), and reassess the translation output of protein coding-genes by considering smaller proteins in addition to larger canonical proteins. In particular, the cellular stoichiometry of a canonical protein versus a small protein encoded in the same gene and their respective concentrations is unknown. Yet, proteins are the primary effectors of biological processes and deciphering the function of a gene in health and disease requires accurate characterization of its products.

The mitochondrial elongation factor 1 gene or *MIEF1* also termed *SMCR7L/MiD51*, localized at the Chr22-q13.1 locus, codes for a mitochondrial receptor of Drp1, a GTPase which functions in mitochondrial fission (Palmer *et al.*, 2011; Zhang *et al.*, 2016b; Osellame *et al.*, 2016).

Ribosome profiling and proteogeomics studies recently demonstrated the translation of a stable 70 amino acid protein product encoded in a altORF localized in the 5'UTR (Figure 1 a & b) (Lee *et al.*, 2012; Vanderperre *et al.*, 2013; Kim *et al.*, 2014; Crappé *et al.*, 2014; Andreev *et al.*, 2015a; Sidrauski *et al.*, 2015; Calviello *et al.*, 2016; Samandi *et al.*, 2017). Thus, *MIEF1* is a prototypical gene coding for both a large and a small protein. For simplicity, we termed this novel protein alternative MiD51 or altMiD51. Remarkably, both proteins are localized at the mitochondria. MiD51 is an outer mitochondrial membrane protein whereas altMiD51 is located at the mitochondrial matrix and both are involved in mitochondrial fission (Osellame *et al.*, 2016; Samandi *et al.*, 2017).

Here, we employ a targeted proteomics approach based on AQUA peptides to reliably quantify the absolute amount of MiD51 and altMiD51 in two human cell lines and one human tissue, and thus we establish an improved map of the translational output of *MIEF1 / SMCR7L / MiD51*.

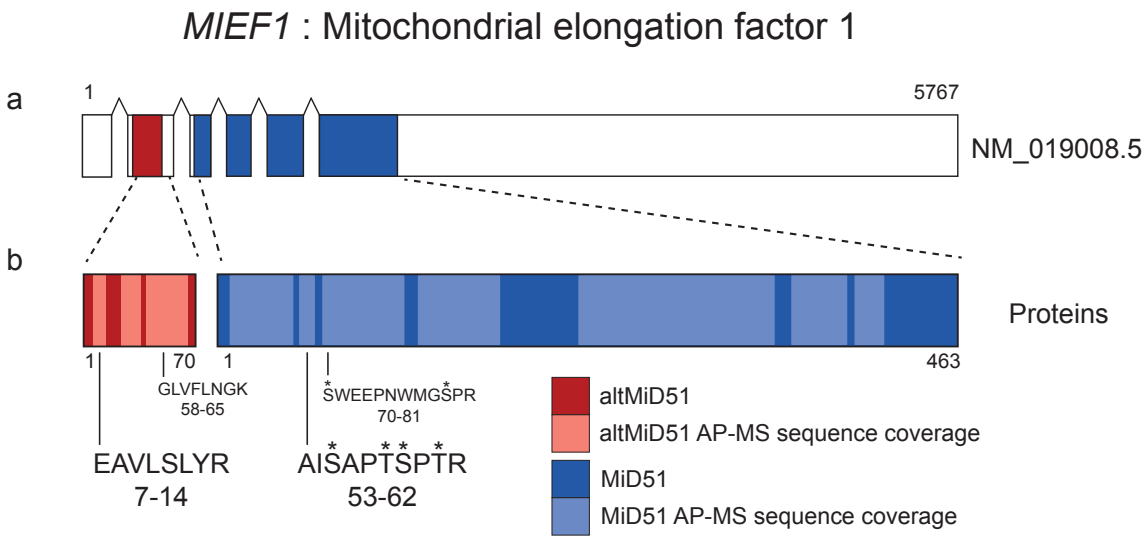


FIGURE 6.1 – **Schematic representation of human *MIEF1* RefSeq variant 1 mRNA and altMiD51 and MiD51 proteins.**
(a) Human *MIEF1* includes 11 exons (RefSeq, GRCh38.p7). The mRNA variant 1 (NM_019008.5) shown here contains 6 exons. The CDS (blue) is shared between exons 3 to 6. AltMiD51 ORF is localized within exon 2, annotated as a non-coding exon. (b) Sequence coverage in AP-MS experiments is represented in light colors. Proteotypic peptides sequence and positions (a.a.) are shown. EAVLSLYR and AISAPTSPTR peptides were selected for absolute quantification.
* : known phosphorylated residue (phosphosite.org, Hornbeck *et al.* (2014))

6.1.3 Results

6.1.3.1 Determination MiD51 and altMiD51 proteotypic peptides

In addition to the canonical CDS (Consensus CDS CCDS13995 ; RefSeq NM_019008.5 ; Ensembl ENST00000325301) and protein (RefSeq NP_06188, UniProt Q9NQG6 ; Ensembl ENSP00000327124), human *MIEF1* contains a functional and recently annotated altORF (GenBank HF548110) and small protein (Uniprot L0R8F8 ; GenBank CCO13821.1 ; Ensembl ENSP00000490747). Thus, *MIEF1* is clearly a prototypical dual-coding gene for which the absolute quantification of the large and small protein products is unknown. We evaluated the ability of MiD51 and altMiD51 to generate proteotypic peptides after trypsin digestion. Proteotypic peptides are specific for each protein and they must be consistently detected with excellent quality precursor and fragment mass transitions (Kuster *et al.*, 2005; Mallick *et al.*, 2007; Wilhelm *et al.*, 2014; Zolg *et al.*, 2017). In order to facilitate the detection of specific tryptic peptides for both proteins, we used affinity purification coupled with mass spectrometry. MiD51^{GFP} and altMiD51^{GFP} were

independently overexpressed in HeLa cells. Both proteins were affinity purified and analyzed via data-dependant (DDA) nano capillary liquid chromatography mass spectrometry (nanoLC-MS/MS). Several proteotypic peptides were detected for a total sequence coverage of 68.6 % and 71.9 % for altMiD51 and MiD51, respectively (Figure 1 b & supplementary data 1). After manual evaluation, best quality proteotypic peptides were selected for parallel reaction monitoring (PRM) optimization (Gallien *et al.*, 2014; Bourmaud *et al.*, 2016).

6.1.3.2 PRM optimization for MiD51 and altMiD51 proteotypic peptides

Selected proteotypic peptides were then validated in low-sensitivity PRM experiments in 3 kDa FASP processed samples (Wisniewski *et al.*, 2009). AltMiD51 is a small protein of 70 amino acids and a low M.W. cut-off is necessary to ensure protein retention during sample preparation. Since both MiD51 and altMiD51 are mitochondrial proteins (Samandi *et al.*, 2017), mitochondria were isolated from mock-transfected cells and from cells transfected with a cDNA containing both the CDS coding for MiD51 and the native 5'UTR containing the altORF coding altMiD51 (RefSeq transcript NM_019008.4). Two proteotypic peptides for each protein were detected in these mitochondrial extracts (Figure 1-Figure supplement 1). Signal intensity for endogenous mitochondrial HSP60 peptide shows that the protein concentration of mock and transfected mitochondrial extracts were similar, and that the intensity difference for altMiD51 and MiD51 peptides between mock-transfected and altMiD51/MiD51-transfected samples did not result from differences in mitochondria preparation.

As MiD51's most intensely detected peptide (AISAPTSPTR) bore known phosphosites (Figure 1 b), a second PRM method including a dephosphorylation step using calf intestinal phosphatase (CIP) was implemented (Wu *et al.*, 2011). A fraction of MiD51 was indeed phosphorylated since CIP treatment resulted in a 20 % increase in intensity for AISAPTSPTR (Figure 1-Figure supplement 2). The efficiency of CIP treatment was validated with two known HSP60 tryptic phosphorylated peptides, VGGTSDVEVNEK and VTDALNATR (phosphosite.org) with an increase in intensity of 103 % and 133 %, respectively. The intensity of a non-phosphorylated HSP60 peptide did not change significantly. AltMiD51 peptides are clearly non phosphorylated since CIP treatment did not change significantly their intensity (Figure 1-Figure supplement 2). Based on these results, we selected peptides EAVLSLYR and AISAPTSPTR for absolute quantitation of altMiD51 and MiD51, respectively.

Finally, the precision of the most sensitive PRM method across different samples was

estimated with the measure of the coefficient of variation (CV) on mitochondrial and whole cell extracts. Indeed, a CV below 20 % is required for absolute quantification (Gallien *et al.*, 2015). The CVs were systematically below 20 %, indicating that both mitochondrial and whole cell extracts were suitable for quantification (Figure 1-Figure supplement 3). Even though peptide intensities are higher in mitochondrial extracts, we decided to use whole cell lysates for absolute quantification of altMiD51 and MiD51 as their preparation does not involve cell fractionation, with the risk of variable mitochondrial recovery.

6.1.3.3 AltMiD51 and MiD51 protein abundances

Two synthetic stable isotope-labeled peptides for absolute quantification (AQUA) (Gerber *et al.*, 2003), EAVLSLYR and AISAPTSPTR, were spiked into the protein sample after trypsin digestion from HeLa cells and analyzed *via* PRM (Figure 1 c). A total of 5 y ion series transitions starting from the most N-terminal amino acid were measured and both peptides displayed at least one quantifiable transition within a range of 40 amol - 250 fmol and a CV < 20 % (Figure 2- Figure supplement 1).

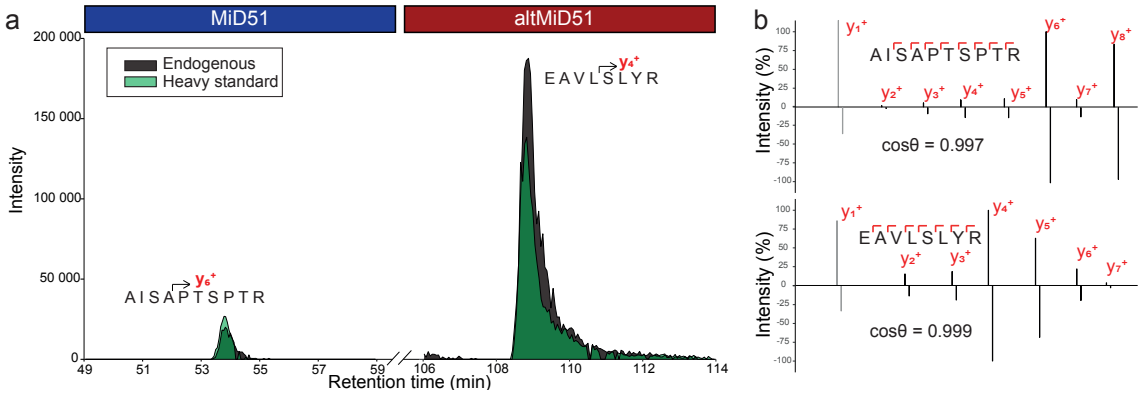


FIGURE 6.2 – **Extracted ion transition chromatograms of MiD51 and altMiD51 peptides in HeLa cells and spectral contrast angle analysis.**
(a) Extracted framgent-ion transition chromatograms of MiD51 ($[AISAPTSPTR+2H]^{2+} \rightarrow y_6^+$) and altMiD51 ($[EAVLSLYR+2H]^{2+} \rightarrow y_4^+$) peptides in HeLa cells. (b) Spectral contrast angle analysis of endogenous peptides (top) and stable isotope labelled synthetic peptides (bottom) extracted from Figure 2-Figure supplement 8.

Absolute quantification PRM experiments were performed by spiking AQUA peptides with trypsin into the digestion mixture as described by Gerber *et al.* (2003). After desalting and dephosphorylation with CIP treatment, the resulting peptides were processed using a high sensitivity PRM method (Gallien *et al.*, 2014). For each peptide, retention times for the corresponding native and AQUA species as well as spectral contrast angles or ratio

dot product (Wan *et al.*, 2002) were controlled to ensure correct identification (Figure 2 a, b & Figure 2-Figure supplement 2-9). The absolute amount of native peptides were thus determined (Supplementary table 2).

We determined that the number of altMiD51 molecules in both HEK293 and HeLa cells was significantly higher than MiD51 molecules (Table 1).

Tableau 6.1 – **Copy number estimations of altMiD51 and MiD51 in HEK 293 and HeLa.**
Values show numbers of molecules per cell considering that each HeLa and HEK 293 cell contains 197 ± 21 pg and 114 ± 2 pg of protein respectively.

Cell line	altMiD51	MiD51
HEK 293	73,000 (± 4,000)	27,000 (± 2,000)
HeLa	82,000 (± 3,000)	14,000 (± 1,000)
Cas9 altMiD51-edited HeLa	-	31,000 (± 1,000)
Cas9 MiD51-edited HeLa	65,000 (± 4,000)	2,000 (± 600)

CRISPR-Cas9-mediated independent inactivation of altMiD51 or MiD51

As this is the first absolute quantification of a large and small protein encoded by two independent ORFs in the same gene, it is important to show that absolute amounts of MiD51 and altMiD51 are partially or completely obliterated by inactivating their respective coding sequences. Experimental modulation of altMiD51 expression independently of MiD51 expression using a knockdown approach is impossible since both proteins are coded by the same gene, and both coding sequences are present in the same transcripts. This is a general challenge for the study of small and large proteins coded in the same gene (Delcourt *et al.*, 2017). Thus, we implemented a CRISPR-Cas9 approach to independently prevent the expression of either altMiD51 or MiD51 (Figure 3 a & b) (Barrangou *et al.*, 2007; Garneau *et al.*, 2010; Ran *et al.*, 2013; Doench *et al.*, 2016; Stemmer *et al.*, 2015). Genome-edited clonal cell lines were validated by sequencing the targeted genomic region. The sequence of the PCR-amplified altMiD51 genomic region confirmed the homozygous insertion of an A at position 40 of exon 2, at the Cas9 cleavage site (Figure 3 c). For MiD51, the sequence electropherogram of the PCR-amplified genomic region showed overlapping peaks (Figure 3 c), indicating the presence of heterozygous mutations in the different alleles, and possibly the presence of remaining WT alleles.

AltMiD51 was competley undectable both by western blot (Figure 3 d) and absolute quantification (Figure 4 a, Welch's t-test p-value = 0.0013), confirming successful editing of the altMiD51 ORF (Figure 3 c). Remarkably, levels of MiD51 were significantly increased in

altMiD51-edited cells (Table 1 & Figure 4 a, Welch's t-test p-value = 0.0006) Although MiD51 was not detected by western blot in CRISP-Cas9-edited cells (Figure 3 d), PRM analyses showed a 86 % reduction in MiD51 levels (Table 1 & Figure 4 a, Welch's t-test p-value = 0.0004), suggesting that non-edited WT alleles remained. However, sequencing alleles of MiD51-edited HeLa (Figure 3 e) revealed that no WT sequence was detected, suggesting that signal from PRM experiments is due to the 6 nucleotides, and thus 2 amino acids, loss in MiD51 sequence, giving a truncated protein (Figure 3 e, blue bar). Overall, genome editing of altMiD51 and MiD51 conclusively validated the proteotypic peptides selected for absolute quantification, and the presence of two functional and physically independent coding information in the same gene.

Absolute amounts and ratio of altMiD51 to MiD51

We compared absolute quantities of altMiD51 and MiD51 in HEK 293, HeLa and human colon tissue samples. Unexpectedly, the stoichiometry indicated that the most abundant translation product from *MIEF1* is altMiD51 rather than the cannonical MiD51 protein. The ratio of altMiD51 to MiD51 is 2.71 in HEK 293 cells, 5.73 in HeLa cells, and 2.62 in Human colon tissue (Figure 4 b).

6.1.4 Discussion

Ribosome profiling and proteogenomics strongly support the translation of alternative protein products from altORFs in addition to the translation of canonical CDSs. Yet, the absolute quantification of a small and a large protein coded by the same gene is unknown. Here, we show that levels of the 70 amino acid altMiD51, a small protein encoded in an exon originally annotated as "non-coding" of *MIEF1/SMCR7L/MiD51* are two to six times higher than the levels of the canonical MiD51 protein in cells and in a human tissue. It is very likely that this is not a general feature of altORFs and that the expression levels of small and large proteins coded by the same genes are highly variable and gene-specific. Also, there is no correlation between protein abundance and functionality, and because the ratio altMiD51/MiD51 is > 2 does not mean that the function of altMiD51 is more significant than that of MiD51. However, this work illustrates that small proteins are important contributors of the proteome, and it is not because that altORFs and alternative proteins are not annotated, unlike large proteins, that they do not exist or have no function.

The ratio of altMiD51 to MiD51 may result from a better translation efficiency for altMiD51. Ribosome profiling data aligned to the *MIEF1* locus indicate that the density of elongating ribosomes is higher on the altORF compared to the CDS (Andreev *et al.*,

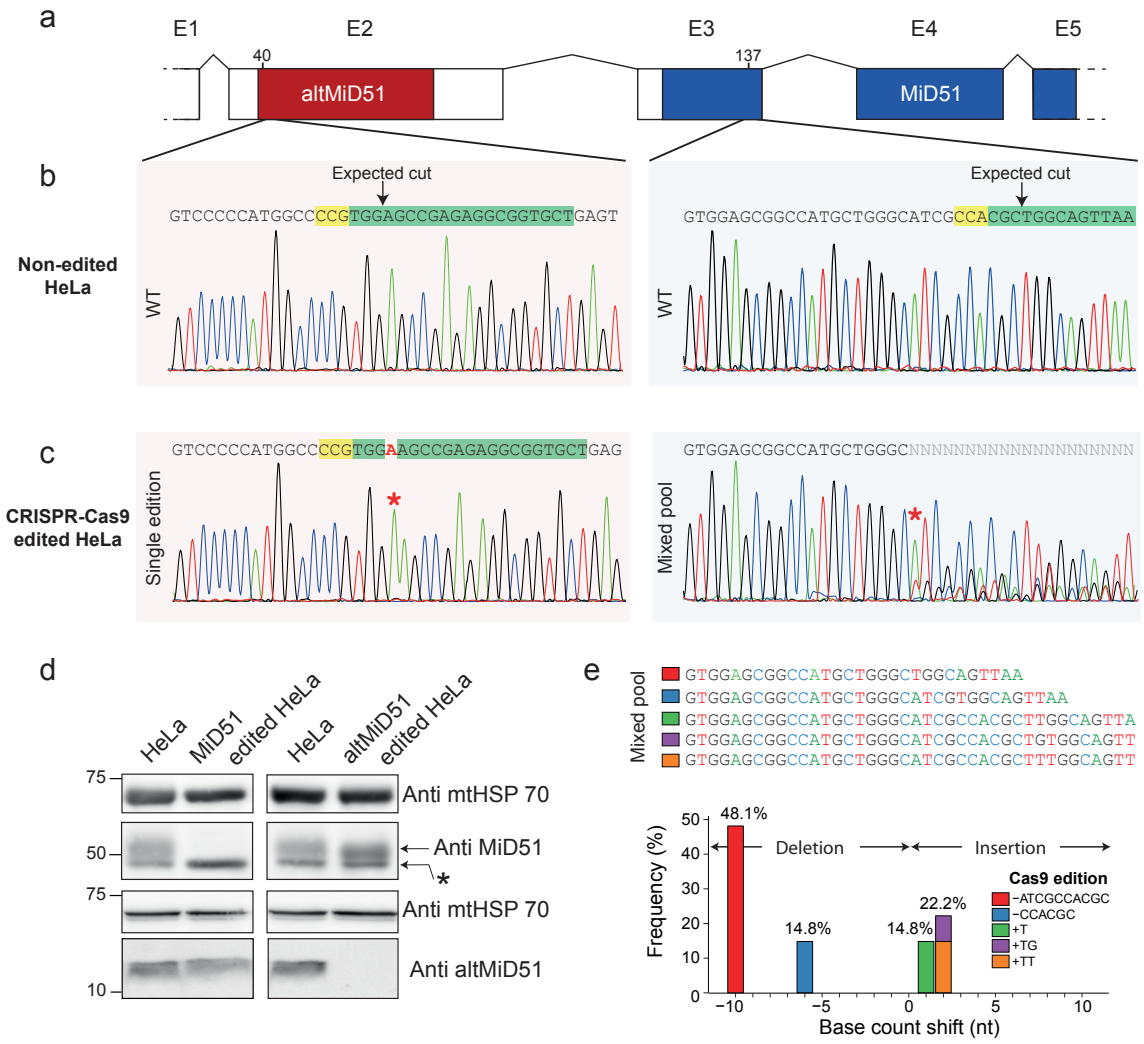


FIGURE 6.3 – **CRISPR-Cas9 editing of genomic altMiD51 and MiD51.**
(a) Schematic representation of CRISPR-Cas9 experiments strategy. For clarity, only 5 exons are shown. AltMiD51 within exon 2 is shown in red. MiD51 coding sequence (blue), overlaps exons 3, 4 and part of exon 5. (b) Genomic sequences around the programmed cut sites in non-edited HeLa cells and corresponding sequences. PAM sites are highlighted in yellow, and the genomic sequences targeted by the guide RNAs are highlighted in green. The programmed cut sites are also shown, at nucleotide 40 in exon 2 and nucleotide 137 in exon 3. (c) Genomic sequence around the programmed cut sites in CRISPR-Cas9-edited HeLa cells. In the altMiD51-edited clone, an adenine insertion (labeled in red, and red star above the electropherogram) occurred at the cut site. In the MiD51-edited clone, a mixture of different sequences are detected 10 nucleotides upstream the programmed cut site (red star), indicating the presence of different alleles. (d) Mitochondrial extracts from non-edited, MiD51-edited and altMiD51-edited HeLa cells were lysed and analyzed by western blot with antibodies against mtHSP70, MiD51 and custom altMiD51 antibodies, as indicated. (e) CRISPR-Cas9 MiD51 knock-out sequence analysis. Sequences are aligned with electropherogram of panel c.
* refers to a non-specific target of MiD51 antibodies (Osellame *et al.*, 2016)

2015a; Michel *et al.*, 2013), suggesting that ribosomes efficiently translate altMiD51. The RYMRMV**AUG**GC motif, where Y = U or C, M = A or C, R = A or G, and V = A, C, or G is known to enhance start codon recognition and translation efficiency (Noderer *et al.*, 2014). In particular, the -4, -3, -2, +4, and +5 positions (+1 denotes the first base of the start codon) are the most important for efficient start codon recognition. For MiD51, the -4A, -3G, -2C, +4G and +5C bases match the motif. For altMiD51, the -4C, -2C, +4G and +5C bases match the motif, but not the -3C base. Since the -3R position is the most important base for efficient initiation (Kozak, 1986, 1997), this elementary sequence analysis suggests that the RYMRMV**AUG**GC motif favors the translation of MiD51. According to the scanning model for translation initiation, the localization of altMiD51 upstream of MiD51 is the likely explanation for the ribosome profiling data discussed above.

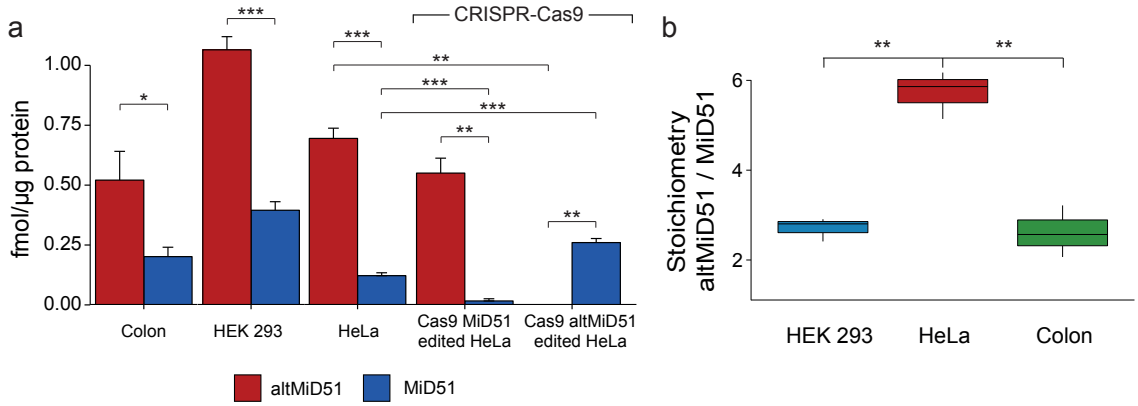


FIGURE 6.4 – **Absolute quantification of altMiD51 and MiD51.**
(a) Absolute quantification of altMiD51 and MiD51 in Colon tissue (technical triplicate), HEK 293, HeLa and CRISPR-Cas9 knock outs (biological triplicates). Error bars = standard deviations. Stoichiometry determination based on absolute quantities of altMiD51 and MiD51. Boxplots represent three biological (HEK 293 & HeLa) or technical (Colon) replicates (b).
Welch's t-test $p < 0.05$ (*), < 0.01 (**), < 0.001 (***)

A combination of several circumstances allowed small proteins to go unnoticed until recently : the lack of annotation of more than one CDS for each gene, their absence from protein sequence databases, the lack of detection tools such as specific antibodies, the technical challenges of studying small proteins and the idea that small proteins may not have functions as important as long proteins. First, according to current human annotations, protein-coding genes have a single CDS, generally the longest ORF (Dinger *et al.*, 2008). Thus, all efforts to find the physiological function or role in the pathology of a specific gene are invariably focused on the protein encoded by this CDS, or one of its variants generated by alternative splicing. Second, in the absence of annotation of non-canonical

ORFs, the protein sequence of the corresponding proteins cannot be routinely detected by MS-based proteomics approaches which rely on current protein databases containing the sequences of canonical proteins only. Third, the widely used western blot technique relies on specific antibodies, but antibodies have been raised and commercialized for canonical proteins only. Raising novel specific antibodies may take time and several attempts, thus delaying the investigations on small proteins. Fourth, the detection of small proteins by MS-based proteomics is more challenging than for large proteins. Typically, the proteome has to be fractionnated to enrich low molecular weight proteins, and the identification often relies on a single tryptic peptide (Ma *et al.*, 2014, 2016). In addition, there may be no sites for trypsin digestion and peptides exceeding 25 aa are rarely identified in bottom-up proteomics. Fifth, because they are short, small proteins are less likely to have known protein domains discovered in large proteins, or to display a specific structure. Thus, there might exist a biased perception that small proteins have minor functions compared to large proteins in biological mechanisms. Yet, many small proteins have essential functions in prokaryotes and eukaryotes (Storz *et al.*, 2014; Delcourt *et al.*, 2017).

Since *MIEF1* codes for at least two stable protein products and that the non-canonical altMiD51 protein is more abundant than MiD51, it will be important to update genome annotations according to recent proteogenomics studies. Indeed, the function of a dual-coding gene should not be inferred according to the molecular activity of the larger protein product only. In addition, the impact of mutations on gene function should not be analyzed in the conceptual frame of a single CDS, since mutations outside currently annotated CDSs may affect non-canonical ORFs and ultimately, gene function.

6.1.5 Methods and Materials

6.1.5.1 Tissue collection and ethics

This study was supported by the Biobanque des maladies digestives du Centre de recherche du CHUS, Centre intégré universitaire de santé et de services sociaux de l'Estrie - Centre hospitalier universitaire de Sherbrooke (CIUSSS de l'Estrie – CHUS), affiliated to the Réseau de recherche sur le cancer and a member of the Canadian Tissue Repository Network.

6.1.5.2 Cell culture

Cells were grown in Dulbecco's Modified Eagle Medium (DMEM, Wisent) supplemented with 10 % fetal bovine serum (FBS, Wisent) and antibiotic-antimycotic cocktail (Eurobio). For transfected samples, cells were grown in 100 mm petri dishes until 80 % confluent

and were transfected by adding 10 µg of plasmidic DNA in 2 mL of FBS/antibiotics-free DMEM and 10 µL of GeneCellIn (BioCellChallenge) and let to grow for 24 hours before cell lysis. For parallel reaction monitoring (PRM) experiments, cells were grown on 60 mm petri dishes until about 80 % confluent.

6.1.5.3 GFP-tagged and contextual mRNA DNA constructs

DNA constructs were obtained by Gibson assembly (Gibson *et al.*, 2009) of synthetic DNA (Gblocks, IDT) using the NEBuilder HiFi DNA Assembly Cloning Kit (New England BioLabs) according to manufacturer's recommendation. DNA blocks of C-terminally LAP (Cheeseman et Desai, 2005) tagged MiD51 and GFP tagged altMiD51 were inserted separately into pcDNA 3.1(-) expression vector (Invitrogen). The context construct was built on the assembly of the full 5' region containing the altMiD51 coding sequence with a C-terminal 2 FLAG tag and the canonical MiD51 coding sequence with a C-terminal HA tag (transcript NM_019008.4) into pcDNA 3.1(-) expression vector. DNA was transformed into *E. coli* on LB-Agar media under ampicillin selection pressure. Clones were selected after overnight incubation and amplified into LB-ampicillin media for 18h. Plasmidic DNA was then purified using midipreps columns purification (Invitrogen). DNA sequences were controlled by sequencing.

6.1.5.4 Mitochondrial extracts

For western blot analysis of HeLa and CRISPR-Cas9 HeLa clones as well as PRM optimizations, mitochondrial extract were performed according to Antonsson *et al.* (2001) with minor modifications. Cells were grown into three 100 mm dishes until 80 % confluent, rinsed twice with PBS and collected using a cell scrapper. Cells were pelleted by centrifugation at 500 g for 10 minutes at 4 °C. Supernatant was discarded and cells were suspended in mitochondrial buffer (210 mM mannitol, 70 mM sucrose, 1 mM EDTA, 10 mM HEPES-NaOH, pH 7.5, 2mg/ml Bovine Serum Albumin (BSA), 0.5 mM PMSF and Roche EDTA-free protease inhibitor) and passed through a 25G1 0.5 × 25 needle syringe 15 consecutive times on ice followed by a 3 minutes centrifugation at 2,000 g at 4 °C. Supernatant was collected and last step was repeated three times. All four supernatant were again passed through syringe needle in mito-buffer and centrifuged for 3 minutes at 2,000 g at 4 °C. Supernatants were collected and centrifuged for 10 minutes at 13,000 g at 4°C and the four mitochondrial pellets were pooled together into BSA-free mito-buffer. Mitochondria were washed again with two cycles of centrifugation and resuspension into BSA-free mito-buffer. Final supernatant was discarded and mitochondria were lysed using 250 µL SDS buffer (4% SDS, Tris-HCl 100 mM pH 7.6). After sonication, protein content

was assessed using BCA assay (Pierce).

6.1.5.5 Mass spectrometry sample preparation

Preliminary affinity-purification (AP)

Cells were rinsed twice with cold PBS and lysed in 1 mL of AP-buffer (NP-40 0.5 %, Tris-HCl 50 mM pH 7.5, NaCl 150 mM, EDTA-free Roche protease inhibitor 1X). Lysate was centrifuged to discard cell debris and insoluble parts and supernatant was collected. GFP-Trap agarose beads (ChromoTek) were conditioned by three consecutive phosphate buffer saline washes followed by three AP buffer washes. Lysate supernatant was mixed with beads and stored at 4 °C for 18 h on a rotating device. Beads were then washed with 3 consecutive washes of AP buffer and 5 consecutive washes of 50 mM NH_4HCO_3 (ABC) and supernatants were discarded. Digestion was performed on beads by adding 1 µg of trypsin (Promega) in 100 µL ABC at 37 °C overnight. Digestion was quenched with formic acid to a final concentration of 1 % and supernatant was collected. Beads were then washed once with acetonitrile/water/formic acid (1/1/0.01 v/v) and pooled with supernatant. Peptides were dried using a speedvac, were desalted using a C18 Zip-Tip and resuspended into 25 µL of 1 % formic acid in water prior to MS analysis.

PRM experiment

For mitochondrial extracts, mitochondrial pellet was lysed using SDS buffer. For whole cell lysates, cells were rinsed twice with cold PBS and lysed using SDS buffer. Tissue sample was homogeneized using a TissueRuptor (Qiagen) in SDS buffer. Lysates were sonicated to reduce viscosity followed by a 5 min centrifugation at 14,000 g to discard debris and insoluble parts. Protein content was assessed using BCA protein assay (Pierce). A total of 100 µg of protein and 1 µg of recombinant Glutathione S-transferase (GST, Schistosoma japonicum) were reduced by adding dithiothreitol to a final concentration of 50 mM and let for 15 minutes at 55 °C. Lysate were prepared according to the filter aided sample preparation protocol (FASP) with minor modifications (Wisniewski *et al.*, 2009). Lysates were diluted in 500 µL of 8 M urea solution and transferred into a 3 kDa centrifugation device (Amicon Ultra, Merck) and centrifuged for 30 mins at 14,000 g. After one 8 M urea wash and centrifugation, samples were diluted in 200 µL of 50 mM iodoacetamide in 8 M urea and let at room temperature in the dark for 30 mins. Samples were centrifuged followed by three consecutive washes and centrifugation of 8 M urea. Buffer was then exchanged for 50 mM ABC by three consecutive 200 µL wash. Final retentate was digested by adding 1 µg of trypsin (Gold, Promega) in 40 µL ABC and AQUA (Gerber *et al.*, 2003) peptides (pepoTec Ultimate, Thermo) were added to mixture prior to overnight incubation

at 37 °C. Tryptic peptides were collected by filter centrifugation followed by three ABC wash and centrifugation. Peptide-containing filtrate was concentrated using a speedvac and then acidified by formic acid to a final concentration of 1 %. Peptides were desalted using a C18 Zip-Tip (Merck) and dried using a speedvac.

Calf intestinal phosphatase treatment

Phosphorylation of peptides were eliminated using calf intestinal phosphatase (CIP) according to Wu *et al.* (2011) . Briefly, 5 µg of desalted peptides were resuspended with 10 units of CIP (New England Biolabs) in 50 µL of CIP buffer (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$ and 1 mM DTT ; pH 7.9) and incubated at 37 °C for 2 hours. Mixture was acidified by adding TFA to a final concentration of 0.5 %. Peptides were desalted using a C18 Zip-Tip (Merck), dried and resuspended into 25 µL of 1% formic acid in water.

6.1.5.6 nanoLC-MS/MS analysis

Instrument setup

A total of 12 μL of peptide mixture was loaded onto a trap column (Acclaim PepMap100 C18 column (0.3 mm id × 5 mm, Thermo Scientific)) at a constant flow rate of 4 μL/min. Peptides were separated on a PepMap C18 nano column (75 μm × 50 cm, Thermo Scientific) using a 0 – 35 % gradient (0-215 mins) of 90 % acetonitrile, 0.1 % formic acid at a flow rate of 200 nL/min followed by acetonitrile wash and column re-equilibration for a total gradient duration of 4 hours with a RSLC Ultimate 3000 (Dionex). Peptides were sprayed using an EASY-Spray source (Thermo) at 2 kV coupled to a quadrupole-Orbitrap (QExactive, Thermo) mass spectrometer.

Affinity purification

For preliminary AP-MS of GFP constructs, mass spectrometer was used in data dependent acquisition mode (DDA). Method consisted into a Full-MS spectra m/z 350-1600 mass range at 70,000 resolution, an AGC target of 1e6 and a maximum accumulation time (maximum IT) of 20 ms followed by the fragmentation (MS/MS) of the top ten ions detected in the Full-MS scan at 17,500 resolution, an automatic gain control (AGC) target of 5e5, a maximum IT of 60 ms with a fixed first mass of 50 within a 3 m/z isolation window at a normalized collision energy (NCE) of 25. Dynamic exclusion was set to 40 s.

Data-dependant protein identification of AP-MS samples

Mass spectrometry RAW files were searched with Andromeda (Cox *et al.*, 2011), search engine implemented in MaxQuant (Cox et Mann, 2008) 1.5.5.1. Trypsin/P was set as digestion mode with a maximum of two missed cleavages per peptides. Oxidation of methionine and acetylation of N-terminal were set as variable modifications. Carbamidomethylation of cysteine was set as fixed modification. Files were searched using a target-decoy approach (Elias et Gygi, 2007) against Uniprot (UniProt, 2017) 03/2017 release and GST (P08515) at a 1 % false discovery rate at peptide-spectrum-match, peptide and protein levels. Peptides sequences were recovered from MaxQuant output files.

PRM method refinement

First PRM method was defined with a large number of peptides in order to discriminate peptides that were detectable in overexpression as well as endogenous conditions in mitochondrial extracts. The peptide list consisted in 30 mass over charges corresponding to unique peptides of MiD51 (11 peptides), altMiD51 (4 peptides), HSP60 (4 peptides) and GST (5 peptides) at various charge states. Peptides were selected based on their high MS/MS count in AP experiments, high Andromeda scores, low number of miscleavages, high MS-1 and MS-2 intensities, and manual quality evaluation of MS/MS spectra. Method consisted in a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted-MS2 method with an AGC target of 5e5 ions, maximum IT of 130 ms, resolution of 17,500 with a 2 m/z isolation window and normalized collision energy of 27.

Second method was used to evaluate the signal recovered after CIP treatment on endogenous mitochondrial extracts. The peptide list consisted in 11 mass over charges based on previous PRM experiments corresponding to peptides of MiD51 (3 peptides), altMiD51 (2 peptides), HSP60 (3 peptides) and GST (3 peptides). Method consisted in a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted-MS2 method with an AGC target of 5e5 ions, maximum IT of 150 ms, resolution of 17 500 with a 2 m/z isolation window and normalized collision energy of 27. All method optimization files were processed using Skyline (MacLean *et al.*, 2010).

High sensitivity PRM

For endogenous CV analysis in whole cell extracts and mitochondrial extracts as well as absolute quantification experiments, mass spectrometer was set for highest sensitivity

according to Gallien *et al.* (2014). Method consisted into a Full-MS spectra acquisition with an AGC target of 3e6, maximum IT of 70 ms and a resolution of 70,000 followed by an unscheduled targeted-MS2 method with an AGC target of 1e6 ions, maximum IT of 250 ms resolution of 70,000 and normalized collision energy of 27. Isolation list contained one peptide from altMiD51, one for MiD51 and their AQUA standards, one peptide from GST spike-in as well as one peptide from HSP60 which were used as sample processing controls.

High sensitivity PRM sample analysis

Mass-spectrometry RAW files were analyzed using Xcalibur 2.2 (Thermo) by measuring area of each peptide monoisotopic transitions within a 3 ppm mass precision window. For AQUA peptide calibration curves, internal standards were spiked into a HeLa digest and analyzed *via* high sensitivity PRM in the same conditions as described previously. For each peptide, five precursor to fragment transitions starting from N-terminus within a mass deviation of 3 ppm were assessed for linearity and CV analysis, considering that a transition with a CV inferior to 20 % at a given concentration is quantifiable. For endogenous CV analysis, most quantifiable precursor to fragment transition was measured for each peptide within a 3 ppm precision window and two replicates were compared. For absolute quantification experiments, protein concentration was determined by comparing the ratio of endogenous peptide to spiked-in AQUA standard and its concentration with the same precursor to fragment transitions within a 3 ppm mass precision window. Peptide ratios were kept below 25. Spectral similarity was controlled by importing RAW files into Skyline and peptides were validated if their spectral contrast angles (Wan *et al.*, 2002) or ratio dot products were close to 1 as well as their retention times matching AQUA standards.

6.1.5.7 MiD51 and altMiD51 knockouts via CRISPR-Cas9

Knockouts clonal cell generation

MiD51 and altMiD51 KO HeLa cells were generated by CRISPR/Cas9 system according to Ran *et al.* (2013) with minor modifications. Briefly, sgRNAs were designed using the Broad Institute sgRNA Designer (CRISPRko) tool (http ://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design, Doench *et al.* (2016)) and additionnaly validated using the CCTOP tool (http ://crispr.cos.uni-heidelberg.de/, Stemmer *et al.* (2015)) for their informative output information, i.e. potential mismatch (MM) positions and off-target site genomic positions with respect to coding

regions, allowing to minimize imperfect match-sites. CRISPR-Cas9 and validation related nucleotide sequences are wrapped in Table 2. The sgRNA oligos inserts, containing an extra G in 5' required for the U6 RNA polymerase III promoter, were prepared by annealing the top and bottom oligos (Table 2) and cloned into the pSpCas9(BB)-2A-GFP plasmid (Addgene #48138, Ran *et al.* (2013)). The resulting plasmids were verified by sequencing. Enrichment for Cas9-2A-GFP expressing cells and isolation of clonal cell populations were performed 24 h after transfection by single-cell FACS sorting. The initial validation of genome editing was done by Mismatch-cleavage assay using T7 Endonuclease I (NEB), GenElute Mammalian Genomic DNA Miniprep Kit (Sigma) with missmatch assays primers (Table 2). The edited cells were confirmed by western blotting (Figure 3) and further by sequencing the PCR amplicons derived from the target sites.

Tableau 6.2 – **Oligonucleotide sequences used for CRISPR-Cas9 genome editing experiments.**

	altMiD51 knock out	**MiD51 knock out**
Genomic target site	5'-TGGAGCCGAGAGGCGGTGCT-3'	5'-CGCTGGCAGTTAAGCGGGTA-3'
Top oligonucleotide	5'-CACCGAGCACCGCCTCTCGGCTCCA-3'	5'-CACCGTACCCGCTTAACTGCCAGCG-3'
Bottom oligonucleotide	5'-AAACTGGAGCCGAGAGGCGGTGCTC-3'	5'-AAACCGCTGGCAGTTAAGCGGGTAC-3'
T7 endonuclease 1	5'-GGGGTCTCTGGAACTTGGAT-3'	5'-GGTCCCAGTACTTATGGCCG-3'
mismatch assays primers	5'-TCCTTTTCTCGGTCCCTTGC-3'	5'-CCACGCAGAAAATCTCAGGG-3'

Characterization of heterozygote MiD51 knockouts

Genomic DNA was amplified using T7 endonuclease 1 MiD51 mismatch assays primers with primer elongation allowing its insertion into linearized (EcoRI, BamHI, New England Biolabs) pcDNA 3.1(-) expression vector *via* Gibson assembly as mentioned above. After overnight ampicillin selection, *E. coli* clones were grown in LB-ampicillin media for 18h. Plasmidic DNA was purified using minipreps (Qiagen) and sequenced.

6.1.5.8 Western blotting

For each sample equivalent of 50 µg of mitochondrial protein extract was mixed 1/1 (v/v) with Laemmli buffer (4 % SDS w/v, 20 % glycerol v/v, Tris-HCl 100 mM pH 6.8, 5 % β-mercapto ethanol v/v) and heated at 95 °C for 15 minutes. For altMiD51, proteins were loaded onto a 4 % stacking, 15% acrylamide-bisacrylamide (37/1, w/w) resolving SDS-PAGE mini gel and separated for one hour at 200 V constant voltage using a glycine-buffer. For MiD51, proteins were loaded onto a 4 % stacking, 10 % acrylamide-bisacrylamide (32/1, w/w) resolving tricine SDS-PAGE (Palmer *et al.*, 2011; Schägger,

2006) gel (16×18 cm) and separated for 18 hours at 25 mA constant current using 0.2 M Tris-HCl pH 8.9 as anode buffer and 0.1 M Tris-HCl 0.1 M tricine 0.1 % SDS pH 8.25 as cathode buffer. Proteins were transferred onto polyvinyldiene difluoride membranes. The membranes were then blocked using a 5 % milk supplemented Tris-buffered saline 0.2 % Tween-20 (TBST). Membranes were probed with a custom anti altMiD51 rabbit antibody (Proteintech, see below), anti MiD51 rabbit antibody (Proteintech 20164-1-AP) and anti-mitochondrial HSP 70 mouse antibody (MA3-028, Pierce) at 4°C overnight. Membrane was then washed three times with TBST and probed with goat anti-mouse (sc-2005, Santa-Cruz Biotechnology) or goat anti-rabbit (7074S, Cell Signaling Technology) horseradish peroxidase antibody and revealed.

6.1.5.9 Raising altMiD51 antibody

Rabbit anti-altMiD51 affinity purified polyclonal antibody was raised against the complete 70 amino acids recombinant altMiD51 protein (Proteintech Group Inc.).

6.1.5.10 Statistics

All graphics and statistics were made using R (R Core Team, 2014) 3.3.2 and ggplot2 (Wickham, 2016) 2.2.1 or higher.

6.1.6 Acknowledgments

Authors are thankful to Michael T. Ryan and Laura Osellame for constructive exchanges, particularly on MiD51 detection *via* western-blot, François-Michel Boisvert for access to mass spectrometer.

6.1.7 Funding

Person	Organism	Reference
Xavier Roucou	Canada Research Chairs	Functional Proteomics and Discovery of New Proteins
Xavier Roucou	Canadian Institutes for Health Research	MOP-136962

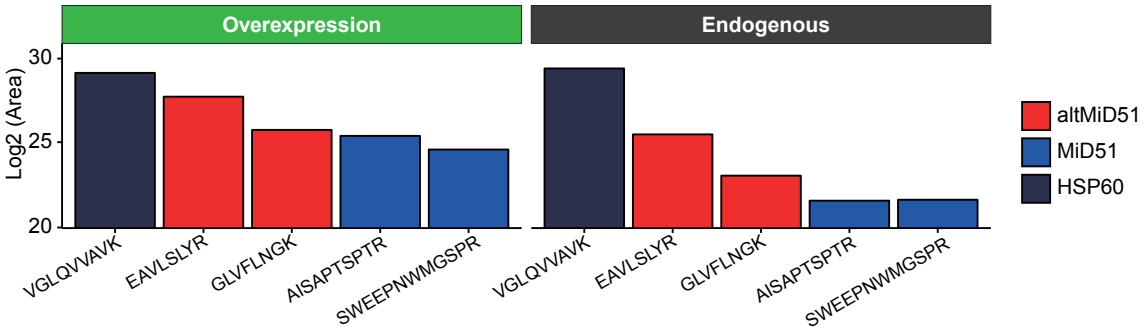


Figure 1–Figure supplement 1. Proteotypic peptides validation by low sensitivity PRM in mitochondrial extracts in transfected and endogenous HeLa cells. Endogenous HSP60 is used here as a control in both experiments.

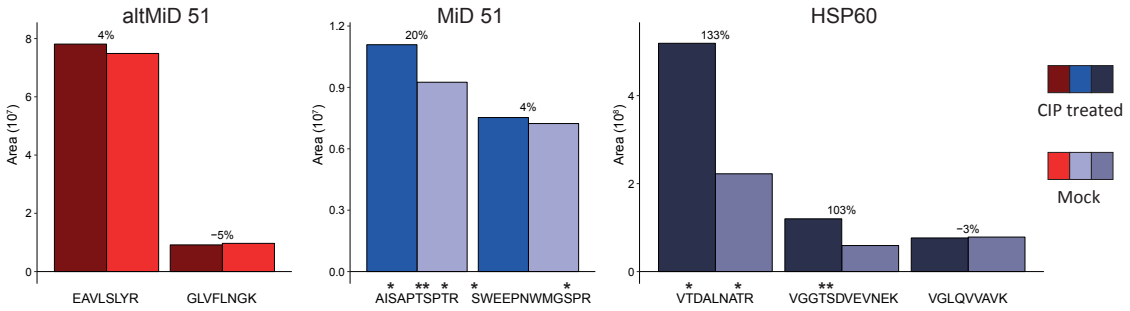


Figure 1–Figure supplement 2. Evaluation of CIP treatments for dephosphorylation of peptides. For each peptide, signal variation are indicated as percentage on top of histograms.
* : known phosphorylated residue (phosphosite.org)

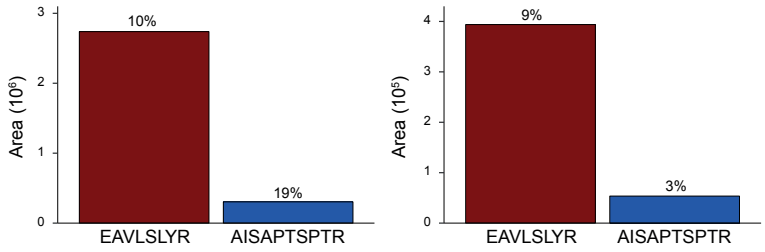


Figure 1–Figure supplement 3. CV analysis of endogenous altMiD51 (red) and MiD51 (blue) in mitochondrial extracts (left) and whole cell lysates (right). CV are indicated on top of histograms.

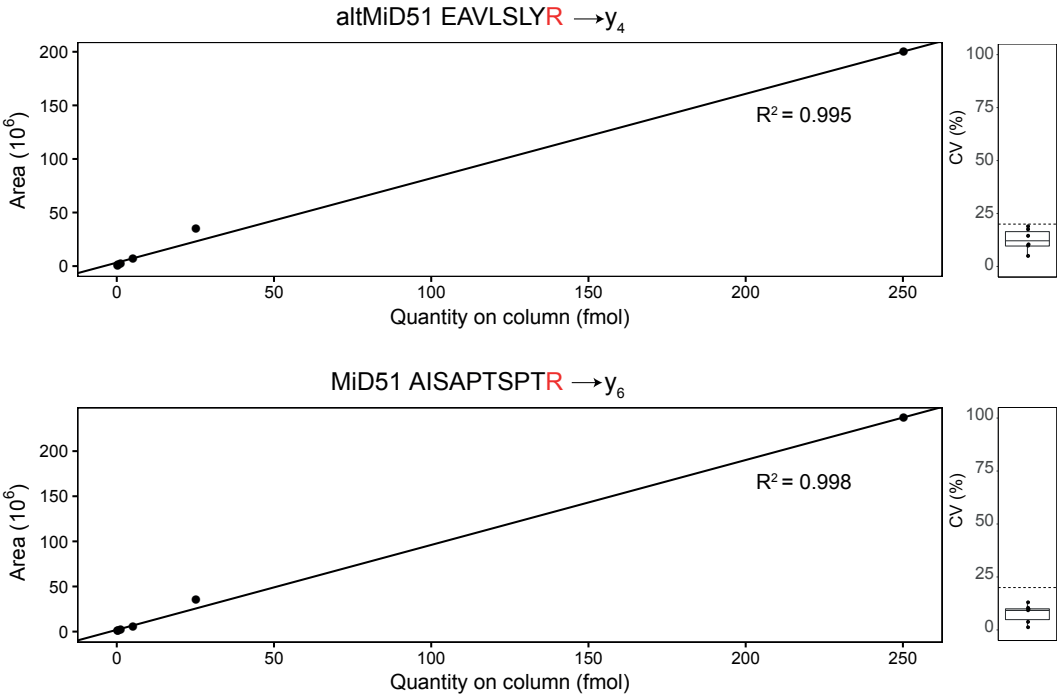


Figure 2–Figure supplement 1. Linear regression and CV analysis of stable isotope labeled peptides.

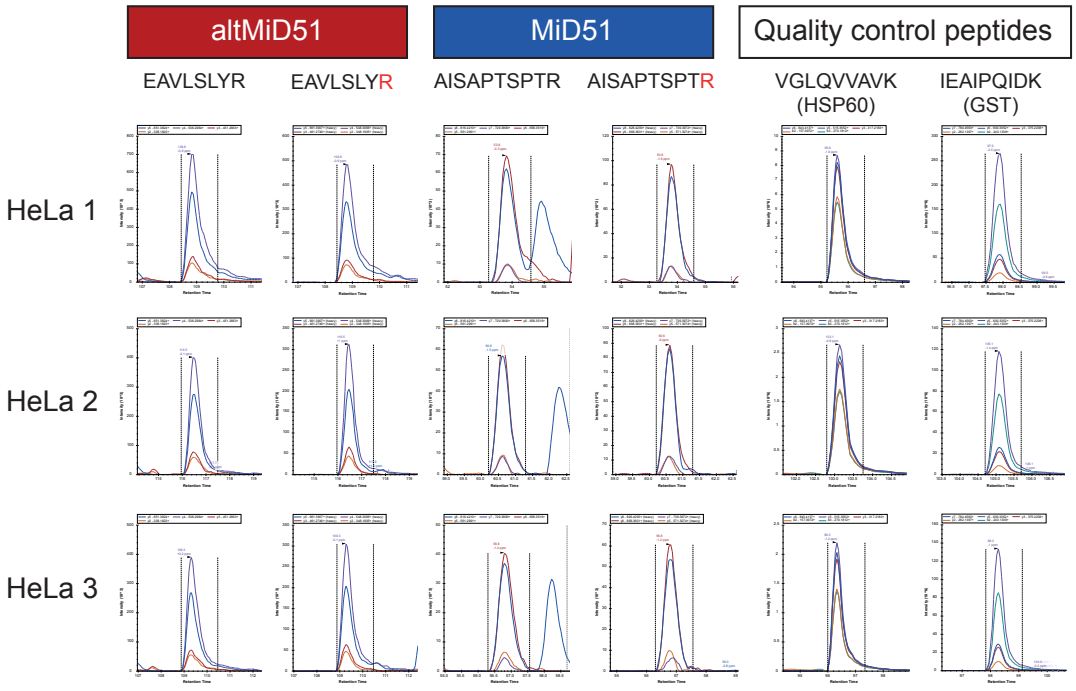


Figure 2–Figure supplement 2. Fragment ion traces in HeLa samples.

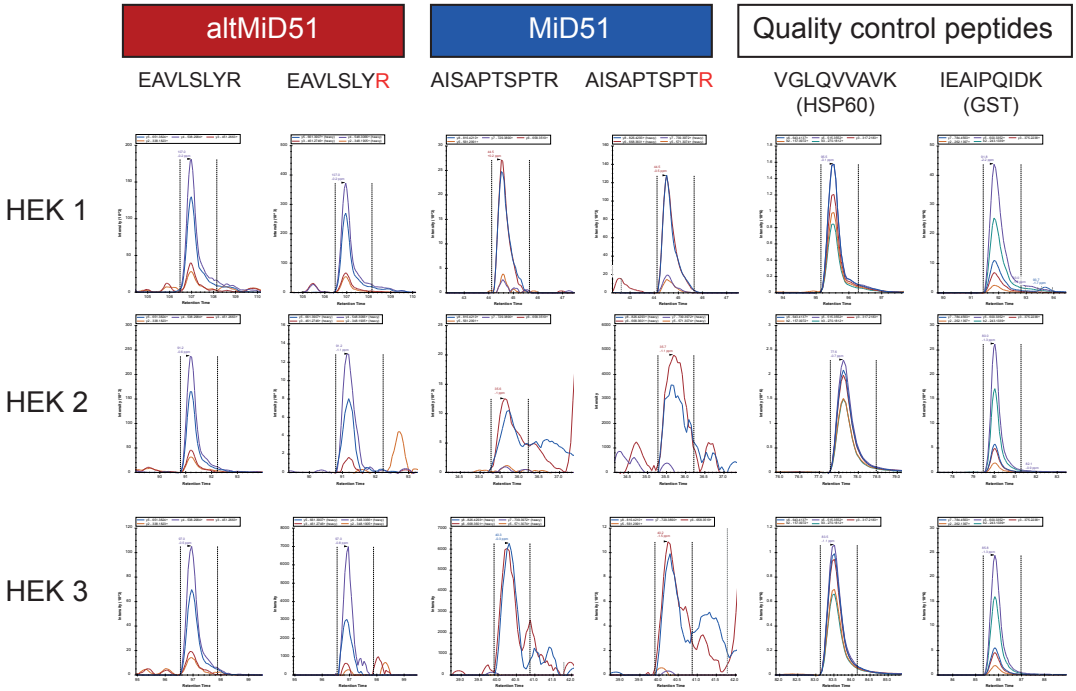


Figure 2–Figure supplement 3. Fragment ion traces in HEK 293 samples.

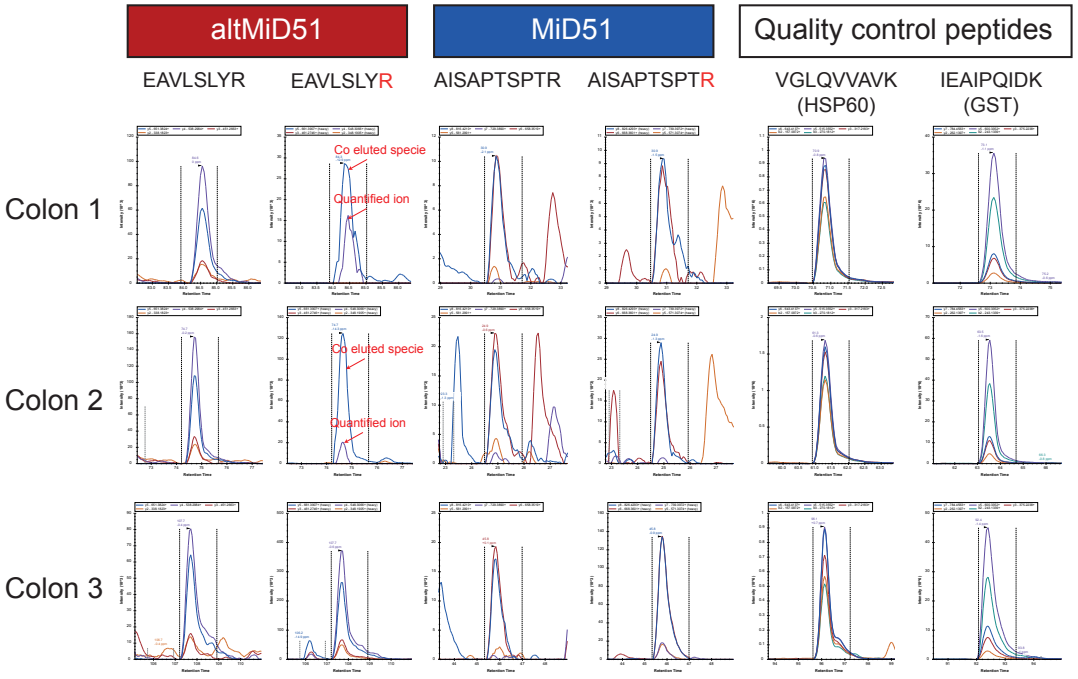


Figure 2–Figure supplement 4. Fragment ion traces in colon tissue samples.

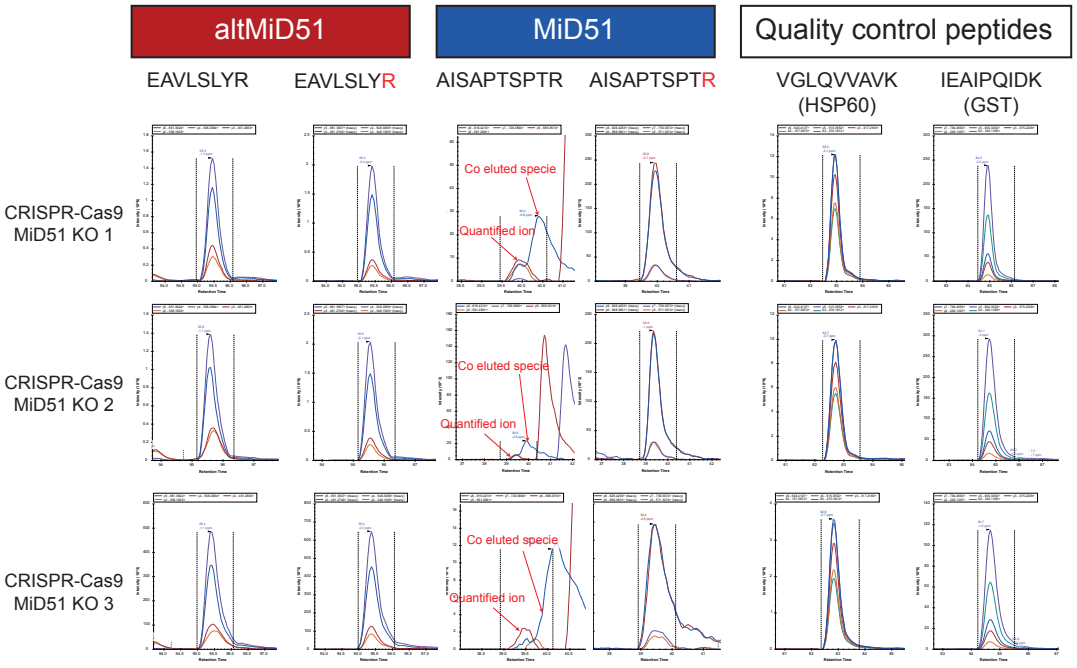


Figure 2–Figure supplement 5. Fragment ion traces in CRISPR-Cas9 MiD51 edited HeLa samples.

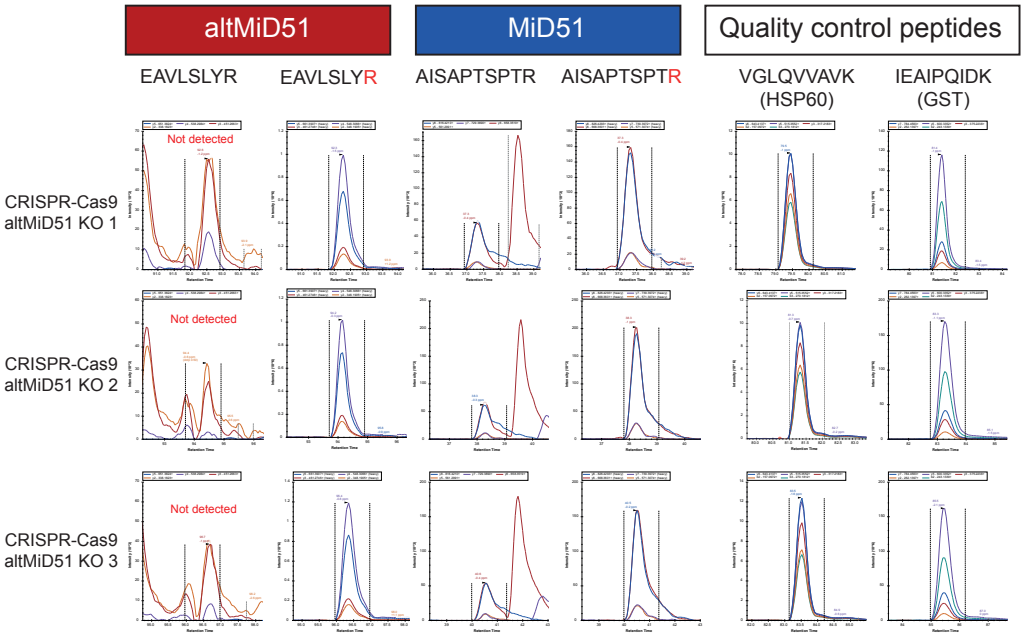


Figure 2–Figure supplement 6. Fragment ion traces in CRISPR-Cas9 altMiD51 edited HeLa samples.

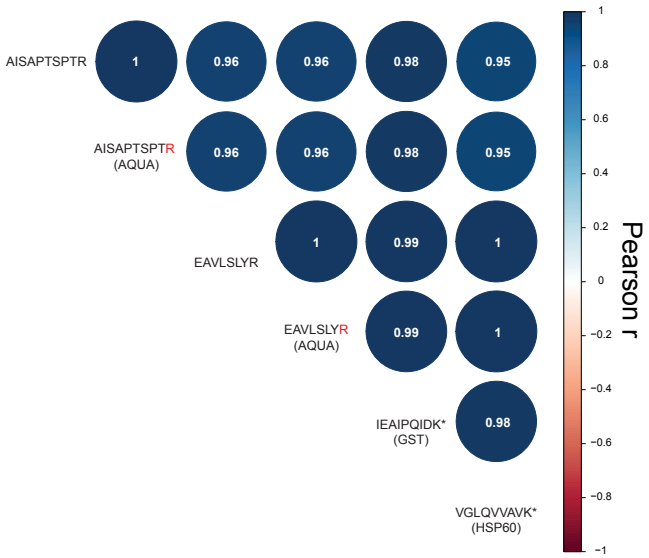


Figure 2–Figure supplement 7. Peptide retention time correlation across samples. Circle color and size indicate Pearson r value which is reported inside each circle. CRISPR-Cas9 altMiD51 samples were excluded as altMiD51 endogenous peptide EAVLSLYR was not detected. * Peptides from GST and HSP60 were used as sample-processing controls.

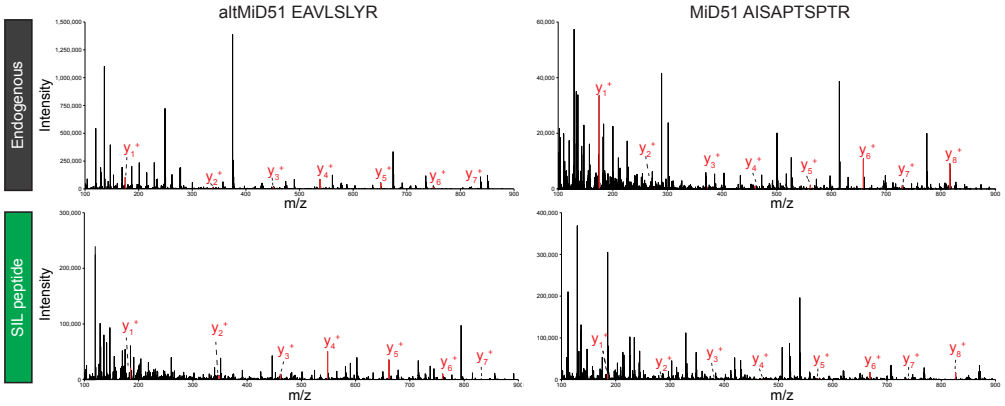


Figure 2–Figure supplement 8. PRM MS/MS spectra of EAVLSLYR and AISAPTSPTR peptides and their stable isotope labeled peptides in a HeLa sample.

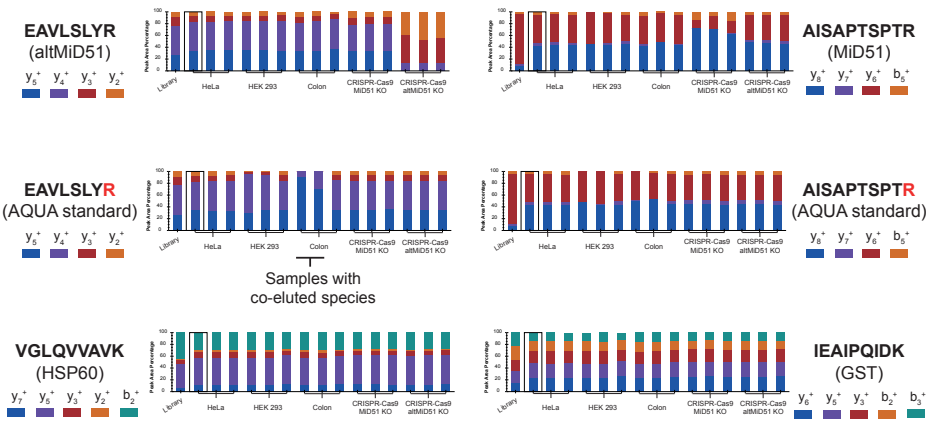


Figure 2–Figure supplement 9. Peptides fragmentation relative intensities across samples in absolute quanti1cation PRM experiments which were used to ensure correct identification of peptides.

6.2 Conclusion et perspectives

Les stratégies de profilage ribosomal et de protéogénomique par spectrométrie de masse ont démontré qu'un gène pouvait coder pour une protéine canonique et une ou des protéine(s) non-canonique(s). Malgré la robustesse de ces deux techniques, l'accès à des informations quantitatives quant au produit traductionnel d'un gène reste difficilement accessible. En effet, le profilage ribosomal permet d'estimer l'efficacité de traduction relative d'un cadre de lecture ouvert au sein d'un transcrit mais ne permet pas la quantification de la protéine effectivement traduite et donc ne tient pas compte de sa stabilité une fois traduite. Les approches de protéomique *bottom-up* par MS permettent quant à elles de quantifier les protéines de manière relative, en comparant des conditions expérimentales ou échantillons les uns par rapport aux autres, ou de manière absolue par le biais d'étalons internes protéiques ou peptidiques. Par l'emploi de méthodes de quantification absolue par MS, il est alors possible de déterminer les quantités absolues de protéines au sein d'échantillons biologiques tels que des lignées cellulaires ou tissus biologiques. Toutefois, bien que les techniques soient décrites et documentées, les niveaux d'expression absolus et la stœchiométrie de protéines canoniques et non canoniques exprimées au sein d'un même gène restent inconnus. Cette information peut s'avérer cruciale, étant donné que les protéines peuvent occuper des fonctions importantes au sein de la cellule, indépendamment de leur taille (Delcourt *et al.*, 2017).

Le gène *MIEF1* s'est distingué dès les premières descriptions de gènes multicodants tant par les études de ribosome profiling (Crappé *et al.*, 2014) que de protéogénomique par MS (Vanderperre *et al.*, 2013). En effet, ces études rapportent conjointement l'expression de deux protéines à partir de ce gène, la protéine canonique MiD51 et altMiD51, encodée à partir de la région 5'UTR du transcrit canonique du gène initialement ignorée lors de l'annotation du génome. De plus, MiD51 est un récepteur mitochondrial dont la fonction est associée au recrutement de Drp1, principale protéine du mécanisme de fission mitochondriale. De façon remarquable, altMiD51 est elle aussi une protéine mitochondriale également impliquée dans la fission mitochondriale. En effet, sa surexpression entraîne une augmentation significative de la fission des mitochondries (Samandi *et al.*, 2017). La détection par deux méthodes distinctes, mais aussi leur association fonctionnelle en font un gène modèle de choix pour la première détermination de la stœchiométrie de deux protéines encodées à partir d'un même gène. La quantification absolue par MS emploie fréquemment des méthodes de protéomique ciblée où l'expérimentateur programme le spectromètre de masse afin qu'il ne sélectionne que certains ions qui correspondent aux peptides dits « protéotypiques » des protéines. Les techniques de proteomique ciblée re-

groupent trois principales méthodes, les *single* ou *multiple reaction monitoring* (SRM / MRM) qui sont principalement employées sur des spectromètres de masse de type *triple quadrupole* et plus récemment le *parallel reaction monitoring* généralement utilisé sur des appareils à haute résolution de type *quadrupole-Orbitrap*. L'emploi du PRM couplé à la spectrométrie de masse à haute résolution offre l'avantage majeur d'obtenir la mesure de masse précise de l'analyte ciblé mais est toutefois plus sujet aux phénomènes de suppression d'ions, lorsque des ions précurseurs particulièrement abondants sont co-élués et co-fragmentés avec l'analyte ciblé.

Des études antérieures d'estimation de quantités absolues de protéines au sein d'échantillons biologiques par approche de protéomique *bottom-up* avec normalisation par rapport au signal des histones ont estimé que MiD51 faisait partie des 15 % des protéines les moins abondantes (Hein *et al.*, 2015). Après ce constat, et n'ayant pas d'information relative aux niveaux d'expression d'altMiD51, l'utilisation d'une méthode de protéomique ciblée a été choisie afin d'augmenter la sensibilité de détection de ces deux protéines. L'approche PRM a été privilégiée, compte tenu de sa capacité à mesurer précisément la masse des peptides ciblés, ce qui renforce la confiance d'identification. La détection de protéines par protéomique ciblée nécessite néanmoins des expériences préliminaires. En effet, il est nécessaire de déterminer les peptides protéotypiques de chacune des protéines c'est-à-dire les peptides offrant une détection reproductible avec une qualité de fragmentation suffisante pour en déduire l'identification sans ambiguïté. Pour ce faire, nous avons tout d'abord surexprimé indépendamment les deux protéines avec une étiquette GFP. Les purifications par affinité de chacune des protéines ont été réalisées. Les extraits ont ensuite été digérés à la trypsine et les peptides issus de cette digestion protéolytique ont été analysés par nanoLC-MS/MS selon le mode d'acquisition DDA. Cette analyse a mis en évidence d'une part les différents peptides accessibles pour ces deux protéines, mais également généré une banque spectrale utilisée par la suite pour identifier les peptides lors des expériences de protéomique ciblée. Après évaluation des différents peptides détectés et optimisation de la méthode PRM pour en augmenter la sensibilité, deux peptides ont été sélectionnés pour leur détectabilité à des niveaux d'expression endogènes. Ces deux peptides ont également été synthétisés avec incorporation d'acides aminés marqués par des isotopes stables, qui seront employés comme étalons internes lors des expériences de quantification absolue.

Les échantillons biologiques, qu'ils soient issus de lignées cellulaires ou de tissus biologiques, ont été préparés selon la méthode de d'échantillonage assistée par filtre avec des filtres à tamis moléculaires de 3 kDa, pour capter le petit protéome. En effet, la plupart des

méthodes de préparation d'échantillon en protéomique sont soit susceptibles d'éliminer les protéines de faible poids moléculaire, soit d'en perdre une partie. Une fois les protéines dénaturées, les peptides marqués furent ajoutés à l'échantillon, concomitamment à la trypsine plutôt qu'avant l'analyse MS (Gerber *et al.*, 2003). Cette distinction offre l'avantage de considérer les éventuelles dégradations, modifications ou encore la précipitation naturelle qui peuvent avoir lieu en solution (Shuford *et al.*, 2012). Il a également fallu adapter une méthode de déphosphorylation des peptides car le peptide sélectionné de la protéine MiD51 comporte 4 sites de phosphorylation référencés.

Les expériences de quantification absolue ont tout d'abord révélé que les quantités absolues de la protéine altMiD51 étaient systématiquement plus élevées que celles de la protéine MiD51 dans les lignées cellulaires HEK 293, HeLa et dans des tissus de colon. De plus, en rapportant les quantités absolues des deux protéines, nous avons pu démontrer que la relation stœchiométrique qui existe entre les deux protéines est spécifique de l'échantillon biologique étudié. Ce dernier résultat suggère que la régulation traductionnelle ou post-traductionnelle de ces deux protéines est différente, mais également spécifique de l'échantillon biologique étudié. La mesure de la stœchiométrie révèle qu'altMiD51 est deux à six fois plus exprimé que la protéine canonique MiD51. Ce constat soulève la question de l'efficacité de l'annotation du génome. En effet, la protéine majoritaire du gène *MIEF1*, ignorée lors des diverses annotations jusque récemment est en fait le produit majoritaire du gène. L'ensemble des expériences de protéomique par MS, qui emploient des bases de données issues des annotations génomiques et transcriptomiques, n'étaient pas en mesure d'identifier ni de quantifier altMiD51, ce qui a considérablement retardé sa découverte et sa caractérisation. De façon similaire, cette démonstration défait les idées généralement associées aux protéines issues de cadres de lecture non canoniques, qui sont souvent associées à des erreurs traductionnelles sans fonction biologique et généralement instables. Toutefois, il est peu probable que le produit majoritaire de chaque gène multicodant soit systématiquement une protéine alternative. De plus le fait qu'une protéine soit plus abondante qu'une autre n'implique pas que l'une ou l'autre soit plus importante pour la cellule.

Lors de cette publication, nous avons développé une méthode de quantification absolue de deux protéines, altMiD51 et MiD51, et pour la première fois déterminé la stœchiométrie d'une protéine canonique et non canonique issues d'un même gène. Cette technique pourrait être appliquée aux futures protéines alternatives en cours de caractérisation pour en valider l'expression endogène ce qui s'avère être complémentaire aux techniques employant des anticorps qui peuvent être couteux à développer et nécessiter de multiples

tentatives.

7 Discussion

7.1 La protéomique par MS révèle une part du protéome jusqu'alors restée inaperçue : les protéines alternatives

L'annotation systématique des génomes et la prédiction de gènes a considérablement étendu le champ des connaissances en biologie et ouvert la voie vers la médecine personnalisée.

Dans une ère post-génomique, de nombreux outils ont été développés pour l'étude des protéines à large échelle. Le profilage ribosomal a montré que la plasticité de la machinerie ribosomale, initialement exposée par Kozak (1987), s'appliquait à l'échelle du protéome (Ingolia *et al.*, 2011). Cette observation était tempérée par la détection indirecte de l'expression de protéines alternatives. Le profilage ribosomal démontre tout de même que le ribosome ne traduit pas seulement un ORF canonique au sein d'un transcrit annoté ARNm mais également certains altORFs d'ARNms ou d'ARNnc. Cependant, malgré l'importante profondeur de couverture du protéome offerte par le profilage ribosomal, cette technique ne permet pas d'évaluer si le produit de la traduction d'un ORF donne *in fine* une protéine stable, capable de remplir une fonction biologique.

Une part de ces protéines est instable et rapidement dégradée (Baboo et Cook, 2014), d'autres sont des éléments de régulation tels que les uORFs en amont de CDS canoniques. Toutefois, les approches de protéogénomique permettent l'identification d'un grand nombre de ces protéines, démontrant que certaines d'entre elles sont suffisamment stables pour être détectées et avoir des activités biologiques qui leur sont propres.

Dans ce nouveau domaine en pleine expansion, les découvertes se multiplient et recensent divers exemples de protéines encodées à partir de cadres de lecture alternatifs, que ce soit à large échelle ou centré sur un gène d'intérêt. Afin d'illustrer à quel point ce domaine est actif, la protéine altMiD51 qui avait encore le statut de protéine alternative au début de mon doctorat a été récemment annotée comme une protéine canonique dans les bases de données. Ainsi, le gène *MIEF1* est désormais un gène bicistronique. Certaines fonctions biologiques sont également décrites mais ces descriptions restent rares compte tenu du défi que représente la caractérisation d'une nouvelle protéine (Delcourt *et al.*, 2017). Une proportion significative de ces protéines, issues de gènes annotés comme codants, montre un lien fonctionnel avec leur protéine de référence (Samandi *et al.*, 2017). Ainsi,

les protéines, canoniques et alternatives, encodées par un même gène sont fréquemment en relation fonctionnelle. Cette relation est décrite chez les bactéries sous la forme d'opérons et pourrait également constituer une forme d'optimisation du génome chez les eucaryotes.

Les protéines alternatives se distinguent par leur taille généralement inférieure à celles des protéines de référence. En effet, à l'échelle du transcriptome, la prédiction de protéines alternatives indique une taille médiane de 45 acides aminés alors que celle des protéines de référence s'élève à 460 (Samandi *et al.*, 2017). Il est donc peu probable que la majorité de ces protéines aient des activités biologiques nécessitant de grands domaines protéiques. Il est plausible que celles-ci soient des petites protéines participant à la régulation de l'activité de plus grandes. Cependant, un nombre non négligeable des protéines alternatives prédites présentent des domaines protéiques connus, permettant d'émettre des hypothèses quant à leur fonction biologique et certaines sont également conservées à travers l'évolution.

La méthode de prédilection pour identifier un grand nombre de protéines par MS est l'approche *bottom-up*. De manière générale, la détection de protéines par cette stratégie reste limitée. En effet, afin de pouvoir identifier une protéine, il faut que celle-ci possède un ou plusieurs sites de coupure à l'enzyme protéolytique employée et que ceux-ci génèrent des peptides uniques. Ensuite, il faut que ces peptides soient compatibles avec les diverses méthodes de séparation, notamment en chromatographie liquide en phase inverse. Les peptides ne peuvent être ni trop hydrophobes pour pouvoir être élués, ni trop hydrophiles pour être retenus lors du couplage de chromatographie-MS qui est encore une technique à parfaire (Shishkova *et al.*, 2016). L'emploi d'autres techniques séparatives comme l'électrophorèse capillaire sont envisageables mais encore peu employées compte tenu des contraintes techniques et la nécessité de développements spécifiques. Ces peptides doivent être efficacement émis en tant qu'ions par la source du spectromètre de masse et avoir un signal suffisant pour être sélectionnés pour la fragmentation en DDA, méthode qui reste la plus répandue. Enfin, la qualité des produits de fragmentation doit être suffisante pour permettre l'identification. Les peptides qui remplissent ces conditions sont appelés protéotypiques et même si l'on considère une protéine d'une longueur importante avec de multiples peptides uniques, seuls quelques peptides remplissent ces conditions (Kuster *et al.*, 2005). Les difficultés de détection des protéines s'accumulent lorsqu'on s'intéresse aux protéines les moins abondantes car plus difficiles à détecter. S'il est délicat de détecter certaines protéines de référence de plusieurs centaines d'acides aminés, la difficulté est accrue pour détecter des protéines générant moins de peptides uniques. Cependant, il semble que le nombre de copies des protéines de référence par cellule soit inversement corrélé à leur

poids moléculaire (Wiśniewski *et al.*, 2014). Ainsi, plus la masse d'une protéine est importante, moins elle serait abondante. Pourtant, il apparaît que les protéines alternatives soient source de diversité génétique à travers l'évolution, ce qui est généralement associé à de faibles niveaux d'expression (Schlötterer, 2015).

Leur détection par approche *bottom up* reste possible (Slavoff *et al.*, 2013; Vanderperre *et al.*, 2013) mais il est probable qu'une proportion non négligeable d'entre elles ne le soit pas par l'emploi des techniques actuelles. Des méthodes de préparation spécifiques ont par ailleurs été développées impliquant le fractionnement des protéines en fonction de leur poids moléculaire. Néanmoins, les méthodes les plus performantes résultent généralement par la perte des protéines de plus haut poids moléculaire (Ma *et al.*, 2016).

Dans ce contexte, l'approche *top-down* apparaît comme la technique appropriée pour identifier les petites protéines et protéines alternatives. En effet, l'absence de traitement protéolytique et la prédisposition technique à détecter des petites protéines sont des critères à considérer pour l'étude des protéines alternatives.

Avec le développement d'approches spécifiques de microprotéomique par *top-down*, nous avons démontré qu'il était possible d'étudier les protéines entre 4 et 20 kDa sans traitement protéolytique préalable à partir de régions localisées d'un tissu biologique grâce à l'imagerie MALDI. La technique s'est montrée particulièrement efficace pour identifier le protéome de faible poids moléculaire de régions distinctes du cerveau de rat. Nous avons également remarqué des distributions différentielles de modifications post-traductionnelles en fonction des régions étudiées, impliquant une activité biologique modulée de ces protéines au sein de ces régions. Nous avons aussi constaté qu'un grand nombre de protéines identifiées étaient des protéines tronquées de plus grand poids moléculaire issues des régions N- et C- terminales ou internes des protéines canoniques. Ces fragments protéiques peuvent, dans le cadre d'une pathologie, constituer une source de biomarqueurs potentiels (Lemaire *et al.*, 2007a). De nouvelles protéines alternatives ont également été identifiées et pour la première fois par *top-down*. Toutefois, l'identification de protéines alternatives ne représente qu'une faible proportion des identifications totales. En effet, lors de la détection de protéines par approche *top-down*, plusieurs états de charges sont détectés pour le même précurseur protéique et leur nombre augmente avec la taille de la protéine. Le précurseur protéique est sélectionné puis fragmenté plusieurs fois, même si l'exclusion dynamique est activée. Ce fait est accentué par le phénomène de suppression d'ions, observé lors de l'élution de protéines ou peptides particulièrement abondants. Pour contourner le problème de sélection successive de précurseurs identiques, il serait utile d'incorporer au programme

de contrôle des spectromètres de masse une fonction de déconvolution automatique. Enfin, les protéines ou fragments de protéines peuvent être localisés de manière plus précise par imagerie MALDI de protéines. Si elle semble particulièrement efficace pour l'identification des petites protéines les plus abondantes, la technique développée permet aussi l'identification de protéines alternatives, supposées plus faiblement exprimées. Cette nouvelle méthode apporte des progrès significatifs avec une préparation d'échantillon simplifiée, une faible quantité de lysat et l'absence notable de précipitation protéique. Il est en effet possible, grâce à cette approche, d'analyser les protéines par *top-down* à partir d'un nombre limité de cellules. L'emploi de cette technique pour l'analyse *top-down* de plus grandes quantités protéiques pourrait à l'avenir offrir une couverture du protéome plus vaste que les méthodes qui nécessitent la précipitation.

L'application de cette approche dans un contexte pathologique a validé ses capacités de détection de biomarqueurs potentiels et d'identification de protéines alternatives. L'application de l'imagerie de métabolites par MALDI-MS a permis de mettre en évidence des régions moléculairement distinctes au sein d'une biopsie d'ovaire de haut grade séreux. La délimitation des régions mises en évidence par imagerie MALDI ont été confirmées par une évaluation pathologique et correspondent aux régions tumorales, nécrotiques et bénignes du tissu. Ce résultat démontre la puissance de l'imagerie MALDI pour la classification de tissus. L'analyse par approche de microprotéomique *top-down* a effectivement confirmé les résultats obtenus lors du développement de la méthode. Elle s'avère efficace pour caractériser le protéome de faible poids moléculaire et permet d'identifier un nombre non négligeable de potentiels biomarqueurs. De plus, ces protéines ont été identifiées à partir d'extractions localisées des tissus tumoraux et nécrotiques. Des protéines alternatives ont également été identifiées au sein de ces régions. Parmi celles-ci, AltGNL1 a particulièrement retenu notre attention. Elle est encodée à partir d'un altORF décalé du CDS de GNL1. GNL1 est une protéine détectée dans des cas de tumeur ovarienne mais non détectée dans des tissus normaux (Uhlén *et al.* (2015), proteinatlas.org). L'expression de ce gène pourrait être associée à la pathologie. Le clonage de la séquence codante de cette protéine de référence contenant l'altORF a permis de valider leur expression *via western-blot* et immunofluorescence. Cette validation démontre qu'il est possible d'identifier des protéines alternatives à partir de régions localisées d'un tissu par *top-down*.

Un autre aspect important relatif à l'étude des protéines alternatives est la relation qui existe entre protéine alternative et protéine de référence. Le gène *MIEF1* est un exemple frappant de gène codant pour deux protéines en relation fonctionnelle. En effet, MiD51, la protéine canonique de ce gène, est un récepteur de Drp1, acteur principal de la fission

mitochondriale (Osellame *et al.*, 2016). La protéine alternative altMiD51 est encodée à partir d'un altORF localisé dans la région 5'UTR. Elle est conservée à travers les vertébrés, contient un domaine protéique et est également impliquée dans la fission mitochondriale (Samandi *et al.*, 2017). Ce gène est, de ce fait, un modèle de choix pour l'étude des protéines alternatives. Cependant, aucune étude protéomique n'a évalué l'expression de ces deux protéines. De plus, la quantification absolue de deux protéines encodées par un même gène n'a jamais été réalisée. Dans ce contexte, nous avons développé une méthode de quantification absolue par protéomique ciblée. Nous avons constaté que le produit majoritaire de la traduction du gène *MIEF1* n'est pas la protéine canonique issue de l'annotation du génome mais la protéine alternative altMiD51. En effet, nous avons observé des ratios stœchiométriques altMiD51 / MiD51 compris entre 3 et 6 en fonction du type d'échantillon biologique analysé. Ce résultat implique que les deux protéines sont régulées différemment en fonction de l'échantillon biologique. L'annotation de ce gène devrait être modifiée en conséquence.

Par ces développements, nous avons démontré que la caractérisation des protéines alternatives est possible par approche *top-down* sur des régions localisées d'un tissu. Cette technique ouvre le champ de l'analyse de protéines intactes en MS à partir d'une quantité de matériel initial limité et permet d'identifier des protéines alternatives. Nous avons également rapporté que les protéines alternatives sont quantifiables par des approches de protéomique ciblée et de quantification absolue. Cette approche centrée sur le gène *MIEF1* indique qu'une protéine alternative peut être le produit majoritaire d'un gène.

La découverte de protéines alternatives par MS ne serait pas possible sans l'utilisation de bases de données modifiées contenant les séquences de protéines de référence et celles de protéines alternatives prédites à partir du transcriptome ou du génome. Ces bases de données sont à employer uniquement pour la découverte de protéines alternatives car leur utilisation entraîne des changements dans l'estimation du FDR (Guthals *et al.*, 2015; Nesvizhskii, 2014).

L'ensemble de ces résultats ainsi que les exemples référencés dans la bibliographie confirment qu'une part du protéome, dont la proportion est difficile à déterminer, reste méconnue. Il est nécessaire d'employer des méthodes adaptées tant pour la préparation d'échantillon que pour l'identification pour caractériser le protéome dans son intégralité.

7.2 Fonction des altORFs et des protéines alternatives

Dans mes travaux de thèse, je me suis surtout concentré sur la détection des protéines alternatives. Même si je n'ai pas eu l'opportunité d'étudier la fonction d'une protéine

alternative, il me paraît tout de même important de discuter brièvement de cet aspect. Dans le quatrième article, nous démontrons que le produit protéique principal d'un gène peut être une protéine alternative et non pas la protéine de référence. Une des critiques parfois émise dans la communauté scientifique est que ces protéines sont traduites suite à des erreurs du ribosome ; elles seraient donc le produit d'un « bruit traductionnel » et dépourvues de fonction biologique. Il est possible que ce soit le cas pour un certain nombre d'altORFs dont le codon d'initiation est détecté par le ribosome qui initie alors sa traduction. Cependant, dans le cas où le produit principal d'un gène est une protéine alternative et non pas la protéine annotée, il est peu probable que cela résulte uniquement de bruit traductionnel. Il est pour l'instant impossible de déterminer le nombre de protéines alternatives fonctionnelles, mais on peut quand même proposer certaines hypothèses.

Premièrement, il est clair que la grande majorité des protéines alternatives sont de très petite taille. Elles ne peuvent donc pas avoir les mêmes mécanismes d'action que des protéines plus longues qui peuvent et contenir des domaines protéiques complexes. Il se pourrait donc que ces petites protéines aient un rôle de régulation des grandes protéines (Andrews et Rothnagel, 2014). Selon cette hypothèse, il est possible qu'une petite protéine codée par un gène coopère fonctionnellement avec une grande protéine codée par le même gène. Ce concept est d'ailleurs apparu récemment dans la littérature avec plusieurs exemples de gènes multicodants (Mouilleron *et al.*, 2015).

Deuxièmement, il a été proposé que certains petits ORFs qui codent pour des protéines soient des précurseurs de futures séquences codantes qui permettront la production de protéines utiles pour la cellule. Ils seraient donc des précurseurs de futures protéines (Brar et Weissman, 2015). Selon cette hypothèse, un petit ORF apparaît d'abord suite à une substitution qui génère un codon d'initiation AUG. Dans un deuxième temps, cet ORF est traduit en peptide. Le codon STOP est par la suite substitué, augmentant la taille du peptide qui s'achève par le codon STOP suivant (Lee et Reinhardt, 2011; Andreatta *et al.*, 2015). Par l'augmentation de sa taille, le peptide peut acquérir des fonctions qui peuvent être utiles à la cellule. À partir de ce moment, la cellule peut conserver cette séquence codante. En quelque sorte, le bruit traductionnel pendant un temps évolutif permettrait à la cellule de tester de nouvelles protéines jusqu'à qu'elles deviennent utiles.

Troisièmement, certaines activités cellulaires sont effectuées par des protéines de petites tailles, et des protéines de faible taille peuvent avoir des rôles très importants. Par exemple, l'ATP-synthase mitochondriale, un moteur moléculaire essentiel à la vie de la cellule eucaryote, contient plusieurs protéines de taille inférieure à 100 acides aminés (Delcourt *et al.*, 2017). Il est donc probable que des protéines alternatives de petites tailles aient des rôles

importants. Plusieurs exemples de petites protéines non annotées et avec des rôles importants sont d'ailleurs apparus dans la littérature récente (Kondo *et al.*, 2010; Chng *et al.*, 2013; Pauli *et al.*, 2014; Anderson *et al.*, 2015; Nelson *et al.*, 2016).

8 PERSPECTIVES

8.1 Annotation et détection des protéines alternatives

La contribution réelle des petites protéines au protéome physiologique et pathologique reste encore inconnue, et des efforts supplémentaires devraient être faits pour améliorer leur détection. Par les développements analytiques réalisés pour l'étude des protéines alternatives, nous avons confirmé que les règles régissant l'annotation des données de séquençage des génomes devaient être redéfinies. Il est nécessaire d'évaluer les ORFs non canoniques pour leur potentiel codant, par la présence de séquences favorisant leur traduction (Kozak, 2002), l'étude de leur conservation et la recherche de signatures de domaines protéiques. Si un gène code pour plusieurs protéines, il devrait être annoté comme tel dans les diverses bases de données. L'expression devrait alors être classée selon l'approche expérimentale qui a permis sa détection, en commençant par le profilage ribosomal, puis la MS et enfin la validation biologique et la détection endogène par un anticorps spécifique. Si elle est déterminée, l'annotation devra comprendre également la stœchiométrie des différentes protéines du gène dans divers échantillons biologiques.

L'application de la stratégie de microprotéomique par approche *top-down*, guidée par la technique d'histologie moléculaire par imagerie MALDI, pourrait constituer à l'avenir une méthode performante pour détecter d'éventuels biomarqueurs issus de protéines de référence et alternatives. Elle permet également de détecter des protéines tronquées dont la caractérisation par *bottom-up* est complexe. Il serait également intéressant de tester ses capacités à l'aide d'un spectromètre de masse de dernière génération. Ces spectromètres de masse offrent des résolutions plus élevées pour un temps de *scan* plus court. De plus, ils sont parfois couplés à des méthodes d'activation d'ions différentes telles que la dissociation par transfert d'électrons, ou la photodissociation induite par rayonnement ultraviolet, qui offre la possibilité d'obtenir différentes séries d'ions fragments (Shaw *et al.*, 2013). Les scores associés aux identifications de protéines alternatives s'en trouveraient améliorés.

La détermination de la stœchiométrie de protéines encodées à partir d'un même gène pourrait être appliquée à tous les gènes multicodants si elle est détectable par MS. En effet, la fonction d'une protéine est souvent modulée par son niveau d'expression. Ainsi, un changement dans la stœchiométrie des protéines encodées à partir d'un gène pourrait

avoir des effets notables sur sa fonction. Cette information, jusqu'alors inconnue, s'avère cruciale pour la caractérisation de phénotypes cellulaires ou pathologiques.

Également, il serait intéressant de développer des stratégies d'enrichissement spécifiques des petites protéines. Même s'il existe des approches existantes employant des techniques de filtration, de précipitation ou de séparation par gel, elles restent imparfaites. En effet, les techniques de filtration ont une efficacité limitée pour séparer les protéines de haut poids moléculaire des petites protéines. Les méthodes de précipitation, notamment celles employant des solutions acides, font précipiter également des protéines de haut poids moléculaire (Ma *et al.*, 2016). Les méthodes de séparation par gel sont quant à elles limitées, étant donné les risques de diffusion des petites protéines lors des étapes de décoloration et dessalage. Enfin, même s'il est efficace de séparer les protéines par leur hydrophobicité par extraction en phase solide de type C8, son utilisation est également limitée par la perte des protéines de plus haut poids moléculaire (Ma *et al.*, 2016). Des techniques prometteuses pour l'analyse de composés de faibles poids moléculaires se développent. Elles sont notamment basées sur l'utilisation de silice superhydrophobe mésoporeuse (Bouamrani *et al.*, 2010; Zhang *et al.*, 2012). Ces techniques s'appuient sur l'exclusion de larges protéines alors que les protéines de faible poids moléculaires sont piégées à l'intérieur des pores puis éluées à l'aide d'un solvant apolaire. Elles pourraient être efficaces pour l'étude des protéines alternatives par protéomique.

Enfin, de nombreuses données d'analyses protéomiques par approche *bottom-up* sont stockées dans les bases de données telles que ProteomeXchange (Vizcaíno *et al.*, 2014). Un grand nombre de protéines alternatives pourraient être identifiées à partir des spectres MS/MS de haute qualité non associés à des peptides. De plus, des hypothèses concernant leur activité seraient émises grâce aux conditions expérimentales des différentes expériences de MS. On pourrait ainsi identifier une protéine alternative comme partenaire d'interaction de protéines de référence, impliquées dans la signalisation cellulaire ou voire même potentiels biomarqueurs.

Ce dernier aspect, déjà initié lors de la publication Samandi *et al.* (2017), a été appliqué pour l'analyse de larges jeux de données issus du *consortium Cancer Genome Atlas*. Ces données de protéomique sont obtenues par approche *bottom-up* suivie d'un fractionnement peptidique et d'un marquage pour la quantification avec ions rapporteurs MS2. Elles permettent la quantification relative des protéines dans des tumeurs mammaires (Mertins *et al.*, 2016) et d'ovaires (Zhang *et al.*, 2016a). Il est ainsi possible de quantifier l'expression de protéines alternatives à travers une cohorte de patientes et de les comparer grâce à un échantillon moyen. Des résultats préliminaires de ces analyses montrent

que certaines protéines alternatives sont susceptibles de constituer une source de biomarqueurs de diagnostic et/ou de pronostic. Par exemple, le gène *LINC00493*, initialement annoté comme non-codant et depuis renommé *SMIM26*, est surexprimé dans les cancers du sein de type luminaux-B (Figure 1 A). De même, une protéine alternative encodée dans la région 5' du gène *ASNSD1* est surexprimée dans les tumeurs ovariennes de haut grade de type proliférative (Figure 1 B). Enfin, la protéine alternative encodée à partir du 3'UTR du gène *EDARADD* (également référencée comme le pseudogène *ENO1P1*) pourrait être employée à des fins de pronostic (Figure 1 C). Un niveau élevé de son expression est associé à un pronostic vital défavorable pour les patientes atteintes de cancers ovariens de haut grade séreux.

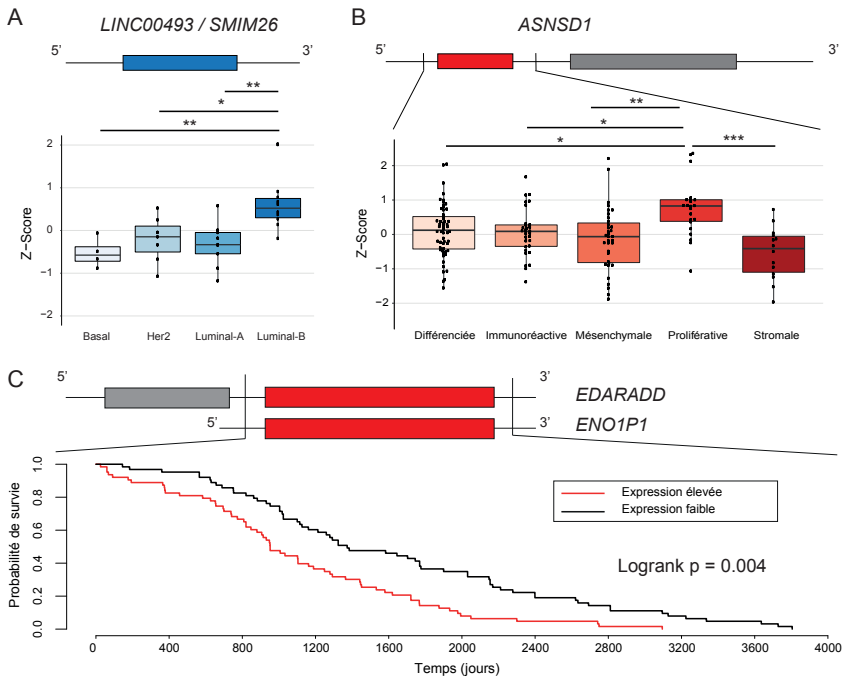


FIGURE 8.1 – **Les protéines alternatives comme biomarqueurs potentiels**
A. Représentation schématique du gène précédemment annoté comme non-codant *LINC00493* et récemment renommé *SMIM26* et quantification de cette protéine alternative dans une cohorte TCGA du cancer du sein. La protéine alternative est surexprimée dans les cancers luminaux-B. **B.** Représentation schématique de la protéine alternative encodée par le gène *ASNSD1* (rouge) et quantification de cette protéine alternative dans une cohorte TCGA du cancer de l'ovaire. La protéine alternative est surexprimée dans les tumeurs ovariennes de haut grade séreux de sous type prolifératif. **C.** Représentation schématique de la protéine alternative encodée par le gène *EDARADD/ENO1P1* (rouge) et analyse de survie par représentation Kaplan-Meier associée à l'expression de cette protéine alternative dans une cohorte TCGA du cancer de l'ovaire. Les groupes d'expression élevée et faible regroupent les patientes avec un niveau supérieur ou inférieur au niveau médian. La valeur statistique de p est déterminée par modèle de Cox (Cox, 1975).

8.2 Déterminer la fonction et la structure de protéines alternatives

Bien que la mesure de la contribution des protéines alternatives au protéome soit un objectif important, la recherche de la fonction de ces protéines est un enjeu majeur. Il est en effet probable que si la cellule a conservé et exprime ces protéines, cela signifie qu'elles ont des activités importantes dans des conditions physiologiques et pathologiques, tout comme les protéines de référence. Le coût énergétique associé à la traduction d'une protéine non-fonctionnelle serait trop important et aurait été éliminée au cours de l'évolution. Toutefois, étudier la fonction de protéines alternatives est un défi considérable puisque plusieurs milliers de protéines alternatives potentielles sont prédites. Deux types d'approches pourraient être envisagés.

Dans une approche ciblée pour une protéine alternative, il faut (1) sélectionner la protéine à étudier ; (2) développer de nouveaux outils comme des anticorps spécifiques ; (3) développer une stratégie d'inactivation de l'altORF forcément basée sur l'édition du génome. En effet il est délicat d'inactiver des ARNms par une approche d'interférence d'ARN étant donné que cela entraînerait l'inactivation simultanée de plusieurs séquences codantes. La sélection d'une protéine cible à partir de milliers de candidats peut se faire en établissant certains critères, comme la conservation à travers les espèces, les évidences et les niveaux d'expression, la présence de domaines ou motifs prédits, la présence de l'AltORF dans un gène d'intérêt, comme des gènes connus pour être impliqués dans le cancer par exemple. De plus, étant donné que de nombreuses protéines alternatives sont fonctionnellement liées à leur protéine de référence, il semble important que les connaissances sur cette dernière soient suffisamment documentées et précises. L'hypothèse de lien fonctionnel pourrait être testée.

Dans une approche à large échelle, il serait possible de tester l'implication de différentes protéines alternatives dans une fonction cellulaire particulière. Par exemple, dans le cas où des activités biologiques sont mesurables à grande échelle sur des plaques 96 ou 384 puits, comme la croissance cellulaire, l'apoptose, la signalisation cellulaire, ou la transformation cellulaire, il est envisageable de surexprimer un altORF dans chaque puits. Une telle approche nécessite des expertises et des équipements particuliers, mais elle est faisable et a l'avantage de faire progresser les connaissances rapidement avec de nombreuses protéines simultanément. En cas de variation, d'une ou plusieurs de ces activités biologiques, le ou les altORF(s) concernés seraient alors sélectionnés pour une caractérisation avancée des mécanismes sous-jacents.

Parallèlement à ces études fonctionnelles, il serait intéressant de déterminer la structure

de ces nouvelles protéines. Non seulement ces nouvelles structures aideront à déterminer le mécanisme d'action de ces protéines. De plus, la composition en acides aminés de certaines de ces protéines est unique. Par exemple, altPrP présente un nombre de résidus de tryptophanes très important compte tenu de sa longueur avec 17 résidus pour une longueur totale de 73 acides aminés (Vanderperre *et al.*, 2011). Il est également possible que de nombreuses protéines alternatives ne possèdent pas de structures complexes ordonnées et soient principalement désordonnées.

LISTE DES RÉFÉRENCES

Abramowitz, J., Grenet, D., Birnbaumer, M., Torres, H. N., et Birnbaumer, L. (2004) Xlαs, the extra-long form of the α-subunit of the gs g protein, is significantly longer than suspected, and so is its companion alex. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22) : 8366–8371.

Aebersold, R. et Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, 422(6928) : 198–207.

Aebersold, R. et Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620) : 347–355.

Agarwal, R., Whang, D. H., Alvero, A. B., Visintin, I., Lai, Y., Segal, E. A., Schwartz, P., Ward, D., Rutherford, T., et Mor, G. (2007) Macrophage migration inhibitory factor expression in ovarian cancer. *American journal of obstetrics and gynecology*, 196(4) : 348–e1.

Aigner, L., Arber, S., Kapfhammer, J. P., Laux, T., Schneider, C., Botteri, F., Brenner, H.-R., et Caroni, P. (1995) Overexpression of the neural growth-associated protein gap-43 induces nerve sprouting in the adult nervous system of transgenic mice. *Cell*, 83(2) : 269–278.

Akimoto, C., Sakashita, E., Kasashima, K., Kuroiwa, K., Tominaga, K., Hamamoto, T., et Endo, H. (2013) Translational repression of the mckusick–kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1830(3) : 2728–2738.

Alexandrov, T., Meding, S., Trede, D., Kobarg, J., Balluff, B., Walch, A., Thiele, H., et Maass, P. (2011) Super-resolution segmentation of imaging mass spectrometry data : solving the issue of low lateral resolution. *Journal of proteomics*, 75(1) : 237–245.

Anderson, D. M., Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., Kasaragod, P., Shelton, J. M., Liou, J., Bassel-Duby, R., *et al.* (2015) A micropeptide encoded by a putative long noncoding rna regulates muscle performance. *Cell*, 160(4) : 595–606.

Andreatta, M. E., Levine, J. A., Foy, S. G., Guzman, L. D., Kosinski, L. J., Cordes, M. H., et Masel, J. (2015) The recent de novo origin of protein c-termini. *Genome biology and evolution*, 7(6) : 1686–1701.

Andreev, D. E., O'Connor, P. B., Fahey, C., Kenny, E. M., Terenin, I. M., Dmitriev, S. E., Cormican, P., Morris, D. W., Shatsky, I. N., et Baranov, P. V. (2015a) Translation of 5' leaders is pervasive in genes resistant to eif2 repression. *Elife*, 4 : e03971.

Andreev, D. E., O'Connor, P. B., Zhdanov, A. V., Dmitriev, R. I., Shatsky, I. N., Papkovsky, D. B., et Baranov, P. V. (2015b) Oxygen and glucose deprivation induces widespread alterations in mrna translation within 20 minutes. *Genome biology*, 16(1) : 90.

Andrews, S. J. et Rothnagel, J. A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nature reviews. Genetics*, 15(3) : 193.

Antonsson, B., Montessuit, S., Sanchez, B., et Martinou, J.-C. (2001) Bax is present as a high molecular weight oligomer/complex in the mitochondrial membrane of apoptotic cells. *Journal of Biological Chemistry*, 276(15) : 11615–11623.

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., et Hartmann, E. M. (2014) Non-model organisms, a species endangered by proteogenomics. *Journal of proteomics*, 105 : 5–18.

Asai, M. M., Mayagoitia, L. L., García, D. D., Matamoros-Trejo, G.-T. G., Valdés-Tovar, M.-T. M., et Leff, P. P. (2007) Rat brain opioid peptides-circadian rhythm is under control of melatonin. *Neuropeptides*, 41(6) : 389–397.

Aspden, J. L., Eyre-Walker, Y. C., Philips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., et Couso, J.-P. (2014) Extensive translation of small orfs revealed by poly-ribo-seq. *Elife*, 3 : e03528.

Autio, K. J., Kastaniotis, A. J., Pospiech, H., Miinalainen, I. J., Schonauer, M. S., Dieckmann, C. L., et Hiltunen, J. K. (2008) An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and rna processing. *The FASEB Journal*, 22(2) : 569–578.

Bab, I., Smith, E., Gavish, H., Attar-Namdar, M., Chorev, M., Chen, Y.-C., Muhlrad, A., Birnbaum, M. J., Stein, G., et Frenkel, B. (1999) Biosynthesis of osteogenic growth peptide via alternative translational initiation at aug85 of histone h4 mrna. *Journal of Biological Chemistry*, 274(20) : 14474–14481.

Baboo, S. et Cook, P. R. (2014) "dark matter" worlds of unstable rna and protein. *Nucleus*, 5(4) : 281–286.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., et Horvath, P. (2007) Crispr provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819) : 1709–1712.

Basrai, M. A., Hieter, P., et Boeke, J. D. (1997) Small open reading frames : beautiful needles in the haystack. *Genome research*, 7(8) : 768–771.

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., Vejnar, C. E., Lee, M. T., Rajewsky, N., Walther, T. C., *et al.* (2014) Identification of small orfs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, page e201488411.

Beaino, W., Guo, Y., Chang, A. J., et Anderson, C. J. (2014) Roles of atox1 and p53 in the trafficking of copper-64 to tumor cell nuclei : implications for cancer therapy. *JBIC Journal of Biological Inorganic Chemistry*, 19(3) : 427–438.

Beaudoin, S., Vanderperre, B., Grenier, C., Tremblay, I., Leduc, F., et Roucou, X. (2009) A large ribonucleoprotein particle induced by cytoplasmic prp shares striking similarities with the chromatoid body, an rna granule predicted to function in posttranscriptional gene regulation. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1793(2) : 335–345.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., et Aebersold, R. (2011) The quantitative proteome of a human cell line. *Molecular systems biology*, 7(1) : 549.

Bentley, D. L. (2014) Coupling mrna processing with transcription in time and space. *Nature Reviews Genetics*, 15(3) : 163–175.

Bergeron, D., Lapointe, C., Bissonnette, C., Tremblay, G., Motard, J., et Roucou, X. (2013) An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *Journal of Biological Chemistry*, 288(30) : 21824–21835.

Birner-Gruenberger, R. et Breinbauer, R. (2015) Weighing the proteasome for covalent modifications. *Chemistry & biology*, 22(3) : 315–316.

Boddapati, N., Anbarasu, K., Suryaraja, R., Tendulkar, A. V., et Mahalingam, S. (2012) Subcellular distribution of the human putative nucleolar gtpase gnl1 is regulated by a novel arginine/lysine-rich domain and a gtp binding domain in a cell cycle-dependent manner. *Journal of molecular biology*, 416(3) : 346–366.

Bonnel, D., Longuespee, R., Franck, J., Roudbaraki, M., Gosset, P., Day, R., Salzet, M., et Fournier, I. (2011) Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in maldi-msi : application to prostate cancer. *Analytical and bioanalytical chemistry*, 401(1) : 149–165.

Bonnet, A., Lagarrigue, S., Liaubet, L., Robert-Granié, C., SanCristobal, M., et Tosser-Klopp, G. (2009) Pathway results from the chicken data set using gotm, pathway studio and ingenuity softwares. In *BMC proceedings*, volume 3, page S11. BioMed Central.

Bouamrani, A., Hu, Y., Tasciotti, E., Li, L., Chiappini, C., Liu, X., et Ferrari, M. (2010) Mesoporous silica chips for selective enrichment and stabilization of low molecular weight proteome. *Proteomics*, 10(3) : 496–505.

Bourguignon, L. Y., Zhu, H., Shao, L., et Chen, Y.-W. (2001) Cd44 interaction with c-src kinase promotes cortactin-mediated cytoskeleton function and hyaluronic acid-dependent ovarian tumor cell migration. *Journal of Biological Chemistry*, 276(10) : 7327–7336.

Bourmaud, A., Gallien, S., et Domon, B. (2016) Parallel reaction monitoring using quadrupole-orbitrap mass spectrometer : Principle and applications. *Proteomics*, 16(15-16) : 2146–2159.

Brar, G. A. et Weissman, J. S. (2015) Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature reviews. Molecular cell biology*, 16(11) : 651.

Bruand, J., Alexandrov, T., Sistla, S., Wisztorski, M., Meriaux, C., Becker, M., Salzet, M., Fournier, I., Macagno, E., et Bafna, V. (2011) Amass : algorithm for msi analysis by semi-supervised segmentation. *Journal of proteome research*, 10(10) : 4734–4743.

Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinović, S. M., Cheng, L.-Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., *et al.* (2015) Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Molecular & Cellular Proteomics*, 14(5) : 1400–1410.

Cabrera-Quio, L. E., Herberg, S., et Pauli, A. (2016) Decoding sorf translation–from small proteins to gene regulation. *RNA biology*, 13(11) : 1051–1059.

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., et Ohler, U. (2016) Detecting actively translated open reading frames in ribosome profiling data. *Nature methods*, 13(2) : 165.

Caprioli, R. M., Farmer, T. B., et Gile, J. (1997) Molecular imaging of biological samples : localization of peptides and proteins using maldi-tof ms. *Analytical chemistry*, 69(23) : 4751–4760.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, 309(5740) : 1559–1563.

Catherman, A. D., Durbin, K. R., Ahlf, D. R., Early, B. P., Fellers, R. T., Tran, J. C., Thomas, P. M., et Kelleher, N. L. (2013) Large-scale top-down proteomics of the human proteome : membrane proteins, mitochondria, and senescence. *Molecular & Cellular Proteomics*, 12(12) : 3465–3473.

Catherman, A. D., Skinner, O. S., et Kelleher, N. L. (2014) Top down proteomics : facts and perspectives. *Biochemical and biophysical research communications*, 445(4) : 683–693.

Celis, J. E., Celis, P., Palsdottir, H., Østergaard, M., Gromov, P., Primdahl, H., Ørntoft, T. F., Wolf, H., Celis, A., et Gromova, I. (2002) Proteomic strategies to reveal tumor heterogeneity among urothelial papillomas. *Molecular & Cellular Proteomics*, 1(4) : 269–279.

Chalick, M., Jacobi, O., Pichinuk, E., Garbar, C., Bensussan, A., Meeker, A., Ziv, R., Zehavi, T., Smorodinsky, N. I., Hilkens, J., *et al.* (2016) Muc1-arf—a novel muc1 protein that resides in the nucleus and is expressed by alternate reading frame translation of muc1 mrna. *PloS one*, 11(10) : e0165031.

Chappell, S. A., Edelman, G. M., et Mauro, V. P. (2006) Ribosomal tethering and clustering as mechanisms for translation initiation. *Proceedings of the National Academy of Sciences*, 103(48) : 18077–18082.

Cheeseman, I. M. et Desai, A. (2005) A combined approach for the localization and tan-

dem affinity purification of protein complexes from metazoans. *Sci. STKE*, 2005(266) : pl1.

Chen, H., Wang, L., Beretov, J., Hao, J., Xiao, W., et Li, Y. (2010) Co-expression of cd147/emmprin with monocarboxylate transporters and multiple drug resistance proteins is associated with epithelial ovarian cancer progression. *Clinical & experimental metastasis*, 27(8) : 557–569.

Cheng, H., Soon Chan, W., Li, Z., Wang, D., Liu, S., et Zhou, Y. (2011) Small open reading frames : current prediction techniques and future prospect. *Current Protein and Peptide Science*, 12(6) : 503–507.

Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., et Gygi, S. P. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology*, 33(7) : 743–749.

Chien, A., Edgar, D. B., et Trela, J. M. (1976) Deoxyribonucleic acid polymerase from the extreme thermophile thermus aquaticus. *Journal of bacteriology*, 127(3) : 1550–1557.

Chng, S. C., Ho, L., Tian, J., et Reversade, B. (2013) Elabela : a hormone essential for heart development signals via the apelin receptor. *Developmental cell*, 27(6) : 672–680.

Choi, D.-H., Kim, Y.-J., Kim, Y.-G., Joh, T. H., Beal, M. F., et Kim, Y.-S. (2011) Role of matrix metalloproteinase 3-mediated α-synuclein cleavage in dopaminergic cell death. *Journal of Biological Chemistry*, 286(16) : 14168–14177.

Choi, M., Eren-Dogu, Z. F., Colangelo, C., Cottrell, J., Hoopmann, M. R., Kapp, E. A., Kim, S., Lam, H., Neubert, T. A., Palmblad, M., *et al.* (2017) Abrf proteome informatics research group (iprg) 2015 study : Detection of differentially abundant proteins in label-free quantitative lc–ms/ms experiments. *Journal of proteome research*, 16(2) : 945–957.

Chu, P., Wu, E., et Weiss, L. M. (2000) Cytokeratin 7 and cytokeratin 20 expression in epithelial neoplasms : a survey of 435 cases. *Modern Pathology*, 13(9) : 962.

Chu, Q., Ma, J., et Saghatelian, A. (2015) Identification and characterization of sorf-encoded polypeptides. *Critical reviews in biochemistry and molecular biology*, 50(2) : 134–141.

Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S. K., et Nekrutenko, A. (2007) A first look at arfome : dual-coding genes in mammalian genomes. *PLoS computational biology*, 3(5) : e91.

Coscia, F., Watters, K., Curtis, M., Eckert, M., Chiang, C., Tyanova, S., Montag, A., Lastra, R., Lengyel, E., et Mann, M. (2016) Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status. *Nature communications*, 7.

Cox, D. R. (1975) Partial likelihood. *Biometrika*, 62(2) : 269–276.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., et Mann, M. (2014) Accurate

proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular & cellular proteomics*, 13(9) : 2513–2526.

Cox, J. et Mann, M. (2008) Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12) : 1367–1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., et Mann, M. (2011) Andromeda : a peptide search engine integrated into the maxquant environment. *Journal of proteome research*, 10(4) : 1794–1805.

Crappé, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., De Meester, E., De Meyer, T., Van Criekinge, W., Van Damme, P., *et al.* (2014) Proteoformer : deep proteome coverage through ribosome profiling and ms integration. *Nucleic acids research*, 43(5) : e29–e29.

Crappé, J., Van Criekinge, W., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., et Menschaert, G. (2013) Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sorfs. *BMC genomics*, 14(1) : 648.

Croce, C. M., Sozzi, G., et Huebner, K. (1999) Role of fhit in human cancer. *Journal of Clinical Oncology*, 17(5) : 1618–1618.

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., *et al.* (2014) Ensembl 2015. *Nucleic acids research*, 43(D1) : D662–D669.

Day, R. et Salzet, M. (2002) The neuroendocrine phenotype, cellular plasticity, and the search for genetic switches : redefining the diffuse neuroendocrine system. *Neuroendocrinology Letters*, 23(5-6) : 447–451.

de Klerk, E. et 't Hoen, P. A. (2015) Alternative mrna transcription, processing, and translation : insights from rna sequencing. *Trends in Genetics*, 31(3) : 128–139.

Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., et Roucou, X. (2017) Small proteins encoded by unannotated orfs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mrna. *Proteomics*.

Dilillo, M., Pellegrini, D., Ait-Belkacem, R., de Graaf, E. L., Caleo, M., et McDonnell, L. A. (2017) Mass spectrometry imaging, laser capture microdissection, and lc-ms/ms of the same tissue section. *Journal of Proteome Research*, 16(8) : 2993–3001.

Dinger, M. E., Pang, K. C., Mercer, T. R., et Mattick, J. S. (2008) Differentiating protein-coding and noncoding rna : challenges and ambiguities. *PLoS computational biology*, 4(11) : e1000176.

D'Lima, N. G., Ma, J., Winkler, L., Chu, Q., Loh, K. H., Corpuz, E. O., Budnik, B. A., Lykke-Andersen, J., Saghatelian, A., et Slavoff, S. A. (2017) A human microprotein that interacts with the mrna decapping complex. *Nature chemical biology*, 13(2) : 174.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F.,

Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al.* (2016) Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2) : 184–191.

Dournaud, P., Jazat-Poindessous, F., Slama, A., Lamour, Y., et Epelbaum, J. (1996) Correlations between water maze performance and cortical somatostatin mrna and high-affinity binding sites during ageing in rats. *European Journal of Neuroscience*, 8(3) : 476–485.

Dupuis-Sandoval, F., Poirier, M., et Scott, M. S. (2015) The emerging landscape of small nucleolar rnas in cell biology. *Wiley Interdisciplinary Reviews : RNA*, 6(4) : 381–397.

Durbin, K. R., Fornelli, L., Fellers, R. T., Doubleday, P. F., Narita, M., et Kelleher, N. L. (2016) Quantitation and identification of thousands of human proteoforms below 30 kda. *Journal of proteome research*, 15(3) : 976–982.

Elias, J. E. et Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature methods*, 4(3) : 207–214.

ElNaggar, M. S., Barbier, C., et Van Berkel, G. J. (2011) Liquid microjunction surface sampling probe fluid dynamics : computational and experimental analysis of coaxial intercapillary positioning effects on sample manipulation. *Journal of The American Society for Mass Spectrometry*, 22(7) : 1157.

Emory, J. F., Walworth, M. J., Van Berkel, G. J., Schulz, M., et Minarik, S. (2010) Direct analysis of reversed-phase high-performance thin layer chromatography separated tryptic protein digests using a liquid microjunction surface sampling probe/electrospray ionization mass spectrometry system. *European Journal of Mass Spectrometry*, 16(1) : 21–33.

Eng, J. K., McCormack, A. L., et Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11) : 976–989.

Ericson, M., Janes, M. A., Butter, F., Mann, M., Ullu, E., et Tschudi, C. (2014) On the extent and role of the small proteome in the parasitic eukaryote trypanosoma brucei. *BMC biology*, 12(1) : 14.

Faggioni, G., Pomponi, A., De Santis, R., Masuelli, L., Ciammaruconi, A., Monaco, F., Di Gennaro, A., Marzocchella, L., Sambri, V., Lelli, R., *et al.* (2012) West nile alternative open reading frame (n-ns4b/warf4) is produced in infected west nile virus (wnv) cells and induces humoral response in wnv infected individuals. *Virology journal*, 9(1) : 283.

Farrell, C. M., O'Leary, N. A., Harte, R. A., Loveland, J. E., Wilming, L. G., Wallin, C., Diekhans, M., Barrell, D., Searle, S. M., Aken, B., *et al.* (2013) Current status and new features of the consensus coding sequence database. *Nucleic acids research*, 42(D1) : D865–D872.

Feller, S. (2012) Microproteins (mips)–the next big thing. *Cell Communication and Signaling*, 10(1) : 42.

Fellers, R. T., Greer, J. B., Early, B. P., Yu, X., LeDuc, R. D., Kelleher, N. L., et Thomas, P. M. (2015) Prosight lite : graphical software to analyze top-down mass spectrometry data. *Proteomics*, 15(7) : 1235–1238.

Fickett, J. W. (1995) Orfs and genes : how strong a connection ? *Journal of Computational Biology*, 2(1) : 117–123.

Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S. A., Ingolia, N. T., *et al.* (2015) A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Molecular cell*, 60(5) : 816–827.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* (2016) The pfam protein families database : towards a more sustainable future. *Nucleic acids research*, 44(D1) : D279–D285.

Firth, A. E. et Brierley, I. (2012) Non-canonical translation in rna viruses. *Journal of General Virology*, 93(7) : 1385–1409.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., *et al.* (2011) Ensembl 2012. *Nucleic acids research*, 40(D1) : D84–D90.

Franck, J., Arafah, K., Elayed, M., Bonnel, D., Vergara, D., Jacquet, A., Vinatier, D., Wisztorski, M., Day, R., Fournier, I., *et al.* (2009a) Maldi imaging mass spectrometry state of the art technology in clinical proteomics. *Molecular & Cellular Proteomics*, 8(9) : 2023–2033.

Franck, J., El Ayed, M., Wisztorski, M., Salzet, M., et Fournier, I. (2009b) On-tissue n-terminal peptide derivatizations for enhancing protein identification in maldi mass spectrometric imaging strategies. *Analytical chemistry*, 81(20) : 8305–8317.

Franck, J., Quanico, J., Wisztorski, M., Day, R., Salzet, M., et Fournier, I. (2013) Quantification-based mass spectrometry imaging of proteins by parafilm assisted microdissection. *Analytical chemistry*, 85(17) : 8127–8134.

Frith, M. C., Forrest, A. R., Nourbakhsh, E., Pang, K. C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., et Grimmond, S. M. (2006) The abundance of short proteins in the mammalian proteome. *PLoS genetics*, 2(4) : e52.

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J., *et al.* (2012) Genome-wide search for novel human uorfs and n-terminal protein extensions using ribosomal footprinting. *Genome research*, 22(11) : 2208–2218.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A., et Couso, J. P. (2007) Peptides enco-

ded by short orfs control development and define a new eukaryotic gene family. *PLoS biology*, 5(5) : e106.

Gallien, S., Bourmaud, A., Kim, S. Y., et Domon, B. (2014) Technical considerations for large-scale parallel reaction monitoring analysis. *Journal of proteomics*, 100 : 147–159.

Gallien, S., Kim, S. Y., et Domon, B. (2015) Large-scale targeted proteomics using internal standard triggered-parallel reaction monitoring (is-prm). *Molecular & Cellular Proteomics*, 14(6) : 1630–1644.

Garneau, J. E., Dupuis, M.-E., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A. H., et Moineau, S. (2010) The crispr/cas bacterial immune system cleaves bacteriophage and plasmid dna. *Nature*, 468(7320) : 67–71.

Geiger, T., Velic, A., Macek, B., Lundberg, E., Kampf, C., Nagaraj, N., Uhlen, M., Cox, J., et Mann, M. (2013) Initial quantitative proteomic map of 28 mouse tissues using the silac mouse. *Molecular & Cellular Proteomics*, 12(6) : 1709–1722.

Geiger, T., Wehner, A., Schaab, C., Cox, J., et Mann, M. (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Molecular & Cellular Proteomics*, 11(3) : M111–014050.

Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., et Gygi, S. P. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem ms. *Proceedings of the National Academy of Sciences*, 100(12) : 6940–6945.

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., et Smith, H. O. (2009) Enzymatic assembly of dna molecules up to several hundred kilobases. *Nature methods*, 6(5) : 343–345.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J., Jacq, C., Johnston, M., *et al.* (1996) Life with 6000 genes. *Science*, 274(5287) : 546–567.

Goggin, K., Beaudoin, S., Grenier, C., Brown, A.-A., et Roucou, X. (2008) Prion protein aggresomes are poly (a)+ ribonucleoprotein complexes that induce a pkr-mediated deficient cell stress response. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1783(3) : 479–491.

Grenier, C., Bissonnette, C., Volkov, L., et Roucou, X. (2006) Molecular morphology and toxicity of cytoplasmic prion protein aggregates in neuronal and non-neuronal cells. *Journal of neurochemistry*, 97(5) : 1456–1466.

Griffiths, J. (2008) A brief history of mass spectrometry.

Guo, Y., Fu, P., Zhu, H., Reed, E., Remick, S. C., Petros, W., Mueller, M. D., et Yu, J. J. (2012) Correlations among ercc1, xpb, ube2i, egf, tal2 and ilf3 revealed by gene signatures of histological subtypes of patients with epithelial ovarian cancer. *Oncology reports*, 27(1) : 286–292.

Guthals, A., Boucher, C., et Bandeira, N. (2015) The generating function approach for

peptide identification in spectral networks. *Journal of Computational Biology*, 22(5) : 353–366.

Hagemann, T., Robinson, S. C., Thompson, R. G., Charles, K., Kulbe, H., et Balkwill, F. R. (2007) Ovarian cancer cell–derived migration inhibitory factor enhances tumor growth, progression, and angiogenesis. *Molecular cancer therapeutics*, 6(7) : 1993–2002.

Hagemann, T., Wilson, J., Kulbe, H., Li, N. F., Leinster, D. A., Charles, K., Klemm, F., Pukrop, T., Binder, C., et Balkwill, F. R. (2005) Macrophages induce invasiveness of epithelial cancer cells via nf-κb and jnk. *The Journal of Immunology*, 175(2) : 1197–1205.

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B., Zhang, D., Qin, Y., Yang, F., *et al.* (2017) Smprot : a database of small proteins encoded by annotated coding and non-coding rna loci. *Briefings in Bioinformatics*, page bbx005.

Harmer, D., Gilbert, M., Borman, R., et Clark, K. L. (2002) Quantitative mrna expression profiling of ace 2, a novel homologue of angiotensin converting enzyme. *FEBS letters*, 532(1-2) : 107–110.

He, G., Holcroft, C. A., Beauchamp, M.-C., Yasmeen, A., Ferenczy, A., Kendall-Dupont, J., Mes-Masson, A.-M., Provencher, D., et Gotlieb, W. H. (2012) Combination of serum biomarkers to differentiate malignant from benign ovarian tumours. *Journal of Obstetrics and Gynaecology Canada*, 34(6) : 567–574.

He, L. et Hannon, G. J. (2004) Micrornas : small rnas with a big role in gene regulation. *Nature reviews. Genetics*, 5(8) : 631.

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., et Minghim, R. (2015) Interactivenn : a web-based tool for the analysis of sets through venn diagrams. *BMC bioinformatics*, 16(1) : 169.

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I. A., Weisswange, I., Mansfeld, J., Buchholz, F., *et al.* (2015) A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3) : 712–723.

Hellens, R. P., Brown, C. M., Chisnall, M. A., Waterhouse, P. M., et Macknight, R. C. (2016) The emerging world of small orfs. *Trends in plant science*, 21(4) : 317–328.

Hinnebusch, A. G., Ivanov, I. P., et Sonenberg, N. (2016) Translational control by 5'-untranslated regions of eukaryotic mrnas. *Science*, 352(6292) : 1413–1416.

Hong-jun, S., Stevens, C. F., et Gage, F. H. (2002) Neural stem cells from adult hippocampus develop essential properties of functional cns neurons. *Nature neuroscience*, 5(5) : 438.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., et Skrzypek, E. (2014) Phosphositeplus, 2014 : mutations, ptms and recalibrations. *Nucleic acids research*, 43(D1) : D512–D520.

Hou, Q., Gao, X., Zhang, X., Kong, L., Wang, X., Bian, W., Tu, Y., Jin, M., Zhao, G., Li,

B., *et al.* (2004) Snap-25 in hippocampal ca1 region is involved in memory consolidation. *European Journal of Neuroscience*, 20(6) : 1593–1603.

Hudson, J. D., Shoaibi, M. A., Maestro, R., Carnero, A., Hannon, G. J., et Beach, D. H. (1999) A proinflammatory cytokine inhibits p53 tumor suppressor activity. *Journal of Experimental Medicine*, 190(10) : 1375–1382.

Ingolia, N. T. (2016) Ribosome footprint profiling of translation throughout the genome. *Cell*, 165(1) : 22–33.

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J., Jackson, S. E., Wills, M. R., et Weissman, J. S. (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports*, 8(5) : 1365–1379.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., et Weissman, J. S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924) : 218–223.

Ingolia, N. T., Lareau, L. F., et Weissman, J. S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4) : 789–802.

Jackson, R. J., Hellen, C. U., et Pestova, T. V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature reviews. Molecular cell biology*, 11(2) : 113.

Ji, Z., Song, R., Regev, A., et Struhl, K. (2015) Many lncrnas, 5'utrs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, 4 : e08890.

Johann Jr, D. J., Rodriguez-Canales, J., Mukherjee, S., Prieto, D. A., Hanson, J. C., Emmert-Buck, M., et Blonder, J. (2009) Approaching solid tumor heterogeneity on a cellular basis by tissue proteomics using laser capture microdissection and biological mass spectrometry. *Journal of proteome research*, 8(5) : 2310–2318.

Johansson, O., Hökfelt, T., et Elde, R. (1984) Immunohistochemical distribution of somatostatin-like immunoreactivity in the central nervous system of the adult rat. *Neuroscience*, 13(2) : 265–IN2.

Kachuk, C. et Doucette, A. A. (2017) The benefits (and misfortunes) of sds in top-down proteomics. *Journal of Proteomics*.

Karas, M. et Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry*, 60(20) : 2299–2301.

Kellie, J. F., Catherman, A. D., Durbin, K. R., Tran, J. C., Tipton, J. D., Norris, J. L., Witkowski, C. E., Thomas, P. M., et Kelleher, N. L. (2011) Robust analysis of the yeast proteome under 50 kda by molecular-mass-based fractionation and top-down mass spectrometry. *Analytical chemistry*, 84(1) : 209–215.

Kellie, J. F., Tran, J. C., Lee, J. E., Ahlf, D. R., Thomas, H. M., Ntai, I., Catherman, A. D., Durbin, K. R., Zamdborg, L., Vellaichamy, A., *et al.* (2010) The emerging process of

top down mass spectrometry for protein analysis : biomarkers, protein-therapeutics, and achieving high throughput. *Molecular BioSystems*, 6(9) : 1532–1539.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et Haussler, D. (2002) The human genome browser at ucsc. *Genome research*, 12(6) : 996–1006.

Kertesz, V. et Van Berkel, G. J. (2010) Fully automated liquid extraction-based surface sampling and ionization using a chip-based robotic nanoelectrospray platform. *Journal of mass spectrometry*, 45(3) : 252–260.

Kertesz, V. et Van Berkel, G. J. (2013) Automated liquid microjunction surface sampling-hplc–ms/ms analysis of drugs and metabolites in whole-body thin tissue sections. *Bioanalysis*, 5(7) : 819–826.

Kertesz, V. et Van Berkel, G. J. (2014) Sampling reliability, spatial resolution, spatial precision, and extraction efficiency in droplet-based liquid microjunction surface sampling. *Rapid Communications in Mass Spectrometry*, 28(13) : 1553–1560.

Kertesz, V., Weiskittel, T. M., et Van Berkel, G. J. (2015) An enhanced droplet-based liquid microjunction surface sampling system coupled with hplc-esi-ms/ms for spatially resolved analysis. *Analytical and bioanalytical chemistry*, 407(8) : 2117–2125.

Kim, J. H., Franck, J., Kang, T., Heinsen, H., Ravid, R., Ferrer, I., Cheon, M. H., Lee, J.-Y., Yoo, J. S., Steinbusch, H. W., *et al.* (2015) Proteome-wide characterization of signalling interactions in the hippocampal ca4/dg subfield of patients with alzheimer's disease. *Scientific reports*, 5 : srep11138.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., *et al.* (2014) A draft map of the human proteome. *Nature*, 509(7502) : 575–581.

Klemke, M., Kehlenbach, R. H., et Huttner, W. B. (2001) Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *The EMBO journal*, 20(14) : 3849–3860.

Koch, A., Gawron, D., Steyaert, S., Ndah, E., Crappé, J., De Keulenaer, S., De Meester, E., Ma, M., Shen, B., Gevaert, K., *et al.* (2014) A proteogenomics approach integrating proteomics and ribosome profiling increases the efficiency of protein identification and enables the discovery of alternative translation start sites. *Proteomics*, 14(23-24) : 2688–2698.

Koensgen, D., Mustea, A., Klaman, I., Sun, P., Zafrakas, M., Lichtenegger, W., Denkert, C., Dahl, E., et Sehouli, J. (2007) Expression analysis and rna localization of pai-rbp1 (serbp1) in epithelial ovarian cancer : association with tumor progression. *Gynecologic oncology*, 107(2) : 266–273.

Kondo, S., Lu, Y., Debbas, M., Lin, A. W., Sarosi, I., Itie, A., Wakeham, A., Tuan, J., Saris, C., Elliott, G., *et al.* (2003) Characterization of cells and gene-targeted mice deficient for

the p53-binding kinase homeodomain-interacting protein kinase 1 (hipk1). *Proceedings of the National Academy of Sciences*, 100(9) : 5431–5436.

Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., et Kageyama, Y. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mrna. *Nature cell biology*, 9(6) : 660.

Kondo, T., Plaza, S., Zanet, J., Benrabah, E., Valenti, P., Hashimoto, Y., Kobayashi, S., Payre, F., et Kageyama, Y. (2010) Small peptides switch the transcriptional activity of shavenbaby during drosophila embryogenesis. *Science*, 329(5989) : 336–339.

Kou, Q., Zhu, B., Wu, S., Ansong, C., Tolic, N., Pasa-Tolic, L., et Liu, X. (2016) Characterization of proteoforms with unknown post-translational modifications using the miscore. *Journal of proteome research*, 15(8) : 2422–2432.

Kozak, M. (1986) Point mutations define a sequence flanking the aug initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2) : 283–292.

Kozak, M. (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and Cellular Biology*, 7(10) : 3438–3445.

Kozak, M. (1995) Adherence to the first-aug rule when a second aug codon follows closely upon the first. *Proceedings of the National Academy of Sciences*, 92(7) : 2662–2666.

Kozak, M. (1997) Recognition of aug and alternative initiator codons is augmented by g in position+ 4 but is not generally affected by the nucleotides in positions+ 5 and+ 6. *The EMBO journal*, 16(9) : 2482–2492.

Kozak, M. (2001) Constraints on reinitiation of translation in mammals. *Nucleic Acids Research*, 29(24) : 5226–5232.

Kozak, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, 299(1) : 1–34.

Kracht, M. J., van Lummel, M., Nikolic, T., Joosten, A. M., Laban, S., van der Slik, A. R., van Veelen, P. A., Carlotti, F., de Koning, E. J., Hoeben, R. C., *et al.* (2017) Autoimmunity against a defective ribosomal insulin gene product in type 1 diabetes. *Nature medicine*, 23(4) : 501–507.

Krockenberger, M., Dombrowski, Y., Weidler, C., Ossadnik, M., Hönig, A., Häusler, S., Voigt, H., Becker, J. C., Leng, L., Steinle, A., *et al.* (2008) Macrophage migration inhibitory factor contributes to the immune escape of ovarian cancer by down-regulating nkg2d. *The Journal of Immunology*, 180(11) : 7338–7348.

Krönig, M., Walter, M., Drendel, V., Werner, M., Jilg, C. A., Richter, A. S., Backofen, R., McGarry, D., Follo, M., Schultze-Seemann, W., *et al.* (2015) Cell type specific gene expression analysis of prostate needle biopsies resolves tumor tissue heterogeneity. *Oncotarget*, 6(2) : 1302.

Kuster, B., Schirle, M., Mallick, P., et Aebersold, R. (2005) Scoring proteomes with proteotypic peptide probes. *Nature reviews Molecular cell biology*, 6(7) : 577–583.

Kwun, H. J., Toptan, T., Da Silva, S. R., Atkins, J. F., Moore, P. S., et Chang, Y. (2014) Human dna tumor viruses generate alternative reading frame proteins through repeat sequence recoding. *Proceedings of the National Academy of Sciences*, 111(41) : E4342–E4349.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) Initial sequencing and analysis of the human genome.

Landry, C. R., Zhong, X., Nielly-Thibault, L., et Roucou, X. (2015) Found in translation : functions and evolution of a recently discovered alternative proteome. *Current opinion in structural biology*, 32 : 74–80.

Laouirem, S., Le Faouder, J., Alexandrov, T., Mestivier, D., Léger, T., Baudin, X., Mebarki, M., Paradis, V., Camadro, J.-M., et Bedossa, P. (2014) Progression from cirrhosis to cancer is associated with early ubiquitin post-translational modifications : identification of new biomarkers of cirrhosis at risk of malignancy. *The Journal of pathology*, 234(4) : 452–463.

Lee, C.-f., Lai, H.-L., Lee, Y.-C., Chien, C.-L., et Chern, Y. (2014) The a2a adenosine receptor is a dual coding gene a novel mechanism of gene usage and signal transduction. *Journal of Biological Chemistry*, 289(3) : 1257–1270.

Lee, S., Liu, B., Lee, S., Huang, S.-X., Shen, B., et Qian, S.-B. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, 109(37) : E2424–E2432.

Lee, Y. C. G. et Reinhardt, J. A. (2011) Widespread polymorphism in the positions of stop codons in drosophila melanogaster. *Genome biology and evolution*, 4(4) : 533–549.

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124) : 168.

Lemaire, R., Ait Menguellet, S., Stauber, J., Marchaudon, V., Lucot, J.-P., Collinet, P., Farine, M.-O., Vinatier, D., Day, R., Ducoroy, P., *et al.* (2007a) Specific maldi imaging and profiling for biomarker hunting and validation : fragment of the 11s proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker. *Journal of proteome research*, 6(11) : 4127–4134.

Lemaire, R., Desmons, A., Tabet, J., Day, R., Salzet, M., et Fournier, I. (2007b) Direct analysis and maldi imaging of formalin-fixed, paraffin-embedded tissue sections. *Journal of proteome research*, 6(4) : 1295–1305.

Li, C., Goudy, K., Hirsch, M., Asokan, A., Fan, Y., Alexander, J., Sun, J., Monahan, P., Seiber, D., Sidney, J., *et al.* (2009) Cellular immune response to cryptic epitopes during therapeutic gene transfer. *Proceedings of the National Academy of Sciences*, 106(26) : 10770–10774.

Li, M., Yin, J., Mao, N., et Pan, L. (2013) Upregulation of phosphorylated cofilin 1 cor-

relates with taxol resistance in human ovarian cancer in vitro and in vivo. *Oncology reports*, 29(1) : 58–66.

Liu, J., Yosten, G. L., Ji, H., Zhang, D., Zheng, W., Speth, R. C., Samson, W. K., et Sandberg, K. (2014) Selective inhibition of angiotensin receptor signaling through erk1/2 pathway by a novel peptide. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 306(8) : R619–R626.

Liu, X., Hengel, S., Wu, S., Tolic, N., Pasa-Tolic, L., et Pevzner, P. A. (2013) Identification of ultramodified proteins using top-down tandem mass spectra. *Journal of proteome research*, 12(12) : 5830–5838.

Liu, X., Sirotkin, Y., Shen, Y., Anderson, G., Tsai, Y. S., Ting, Y. S., Goodlett, D. R., Smith, R. D., Bafna, V., et Pevzner, P. A. (2012) Protein identification using top-down spectra. *Molecular & cellular proteomics*, 11(6) : M111–008524.

Liu, Y., Beyer, A., et Aebersold, R. (2016) On the dependency of cellular protein levels on mrna abundance. *Cell*, 165(3) : 535–550.

Liu, Y., Han, X., et Gao, B. (2015) Knockdown of s100a11 expression suppresses ovarian cancer cell growth and invasion. *Experimental and therapeutic medicine*, 9(4) : 1460–1464.

Lomnytska, M., Dubrovska, A., Hellman, U., Volodko, N., et Souchelnytskyi, S. (2006) Increased expression of cshmt, tbx3 and utrophin in plasma of ovarian and breast cancer patients. *International journal of cancer*, 118(2) : 412–421.

Longuespée, R., Boyon, C., Castellier, C., Jacquet, A., Desmons, A., Kerdraon, O., Vinatier, D., Fournier, I., Day, R., et Salzet, M. (2012) The c-terminal fragment of the immunoproteasome pa28s (reg alpha) as an early diagnosis and tumor-relapse biomarker : evidence from mass spectrometry profiling. *Histochemistry and cell biology*, 138(1) : 141–154.

Lundby, A., Secher, A., Lage, K., Nordsborg, N. B., Dmytriyev, A., Lundby, C., et Olsen, J. V. (2012) Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. *Nature communications*, 3 : 876.

Ma, J., Diedrich, J. K., Jungreis, I., Donaldson, C., Vaughan, J., Kellis, M., Yates III, J. R., et Saghatelian, A. (2016) Improved identification and analysis of small open reading frame encoded polypeptides. *Analytical chemistry*, 88(7) : 3967–3975.

Ma, J., Ward, C. C., Jungreis, I., Slavoff, S. A., Schwaid, A. G., Neveu, J., Budnik, B. A., Kellis, M., et Saghatelian, A. (2014) Discovery of human sorf-encoded polypeptides (seps) in cell lines and tissue. *Journal of proteome research*, 13(3) : 1757–1765.

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., Mastrobuoni, G., Rajewsky, N., Kempa, S., Selbach, M., *et al.* (2015) Extensive identification and analysis of conserved small orfs in animals. *Genome biology*, 16(1) : 179.

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., et MacCoss, M. J. (2010) Skyline : an open source

document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7) : 966–968.

Magny, E. G., Pueyo, J. I., Pearl, F. M., Cespedes, M. A., Niven, J. E., Bishop, S. A., et Couso, J. P. (2013) Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science*, 341(6150) : 1116–1120.

Maier, S. K., Hahne, H., Gholami, A. M., Balluff, B., Meding, S., Schoene, C., Walch, A. K., et Kuster, B. (2013) Comprehensive identification of proteins from maldi imaging. *Molecular & Cellular Proteomics*, 12(10) : 2901–2910.

Makarov, A. (2000) Electrostatic axially harmonic orbital trapping : a high-performance technique of mass analysis. *Analytical chemistry*, 72(6) : 1156–1162.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., *et al.* (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nature biotechnology*, 25(1) : 125–131.

Mao, P. et Wang, D. (2014) Top-down proteomics of a drop of blood for diabetes monitoring. *Journal of proteome research*, 13(3) : 1560–1569.

Martin, F., Barends, S., Jaeger, S., Schaeffer, L., Prongidi-Fix, L., et Eriani, G. (2011) Cap-assisted internal initiation of translation of histone h4. *Molecular cell*, 41(2) : 197–209.

Massard, C., Oulhen, M., Le Moulec, S., Auger, N., Foulon, S., Abou-Lovergne, A., Billiot, F., Valent, A., Marty, V., Loriot, Y., *et al.* (2016) Phenotypic and genetic heterogeneity of tumor tissue and circulating tumor cells in patients with metastatic castrationresistant prostate cancer : a report from the petrus prospective study. *Oncotarget*, 7(34) : 55069.

Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K. I., Clohessy, J. G., et Pandolfi, P. P. (2017) mtorc1 and muscle regeneration are regulated by the linc00961-encoded spar polypeptide. *Nature*, 541(7636) : 228–232.

Menschaert, G. et Fenyö, D. (2015) Proteogenomics from a bioinformatics angle : A growing field. *Mass spectrometry reviews*.

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K., et Van Damme, P. (2013) Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics*, 12(7) : 1780–1790.

Mercer, T. R., Dinger, M. E., et Mattick, J. S. (2009) Long non-coding rnas : insights into functions. *Nature reviews. Genetics*, 10(3) : 155.

Mériaux, C., Franck, J., Park, D. B., Quanico, J., Kim, Y. H., Chung, C. K., Park, Y. M., Steinbusch, H., Salzet, M., et Fournier, I. (2014) Human temporal lobe epilepsy analyses by tissue proteomics. *Hippocampus*, 24(6) : 628–642.

Mertins, P., Mani, D., Ruggles, K. V., Gillette, M. A., Clauser, K. R., Wang, P., Wang, X., Qiao, J. W., Cao, S., Petralia, F., *et al.* (2016) Proteogenomics connects somatic mutations to signaling in breast cancer. *Nature*, 534(7605) : 55.

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., Pestova, T. V., Qian, S.-B., et Jaffrey, S. R. (2015) 5' utr m 6 a promotes cap-independent translation. *Cell*, 163(4) : 999–1010.

Michel, A. M. et Baranov, P. V. (2013) Ribosome profiling : a hi-def monitor for protein synthesis at the genome-wide scale. *Wiley Interdisciplinary Reviews : RNA*, 4(5) : 473–490.

Michel, A. M., Choudhury, K. R., Firth, A. E., Ingolia, N. T., Atkins, J. F., et Baranov, P. V. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome research*, 22(11) : 2219–2229.

Michel, A. M., Fox, G., M. Kiran, A., De Bo, C., O'Connor, P. B., Heaphy, S. M., Mullan, J. P., Donohue, C. A., Higgins, D. G., et Baranov, P. V. (2013) Gwips-viz : development of a ribo-seq genome browser. *Nucleic acids research*, 42(D1) : D859–D864.

Miettinen, T. P. et Björklund, M. (2014) Modified ribosome profiling reveals high abundance of ribosome protected mrna fragments derived from 3' untranslated regions. *Nucleic acids research*, 43(2) : 1019–1034.

Mise, N., Savai, R., Yu, H., Schwarz, J., Kaminski, N., et Eickelberg, O. (2012) Zyxin is a transforming growth factor-β (tgf-β)/smad3 target gene that regulates lung cancer cell motility via integrin $\alpha 5\beta 1$. *Journal of Biological Chemistry*, 287(37) : 31393–31405.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., *et al.* (2014) The interpro protein families database : the classification resource after 15 years. *Nucleic acids research*, 43(D1) : D213–D221.

Moore, M. J. (2005) From birth to death : the complex lives of eukaryotic mrnas. *Science*, 309(5740) : 1514–1518.

Morice, Y., Ratinier, M., Miladi, A., Chevaliez, S., Germanidis, G., Wedemeyer, H., Laperche, S., Lavergne, J.-P., et Pawlotsky, J.-M. (2009) Seroconversion to hepatitis c virus alternate reading frame protein during acute infection. *Hepatology*, 49(5) : 1449–1459.

Mouilleron, H., Delcourt, V., et Roucou, X. (2015) Death of a dogma : eukaryotic mrnas can code for more than one protein. *Nucleic acids research*, 44(1) : 14–23.

Mudge, J. M. et Harrow, J. (2016) The state of play in higher eukaryote gene annotation. *Nature Reviews Genetics*, 17(12) : 758–772.

Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Pääbo, S., et Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology*, 7(1) : 548.

Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P., et Makova, K. D. (2005) Oscillating

evolution of a mammalian locus with overlapping reading frames : an xlαs/alex relay. *PLoS genetics*, 1(2) : e18.

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., Reese, A. L., McAnally, J. R., Chen, X., Kavalali, E. T., *et al.* (2016) A peptide encoded by a transcript annotated as long noncoding rna enhances serca activity in muscle. *Science*, 351(6270) : 271–275.

Nesvizhskii, A. I. (2014) Proteogenomics : concepts, applications and computational strategies. *Nature methods*, 11(11) : 1114–1125.

Nicolardi, S., Switzar, L., Deelder, A. M., Palmblad, M., et van der Burgt, Y. E. (2015) Top-down maldi-in-source decay-fticr mass spectrometry of isotopically resolved proteins. *Analytical chemistry*, 87(6) : 3429–3437.

Nikolov, D. et Burley, S. (1997) Rna polymerase ii transcription initiation : a structural view. *Proceedings of the National Academy of Sciences*, 94(1) : 15–22.

Noderer, W. L., Flockhart, R. J., Bhaduri, A., de Arce, A. J. D., Zhang, J., Khavari, P. A., et Wang, C. L. (2014) Quantitative analysis of mammalian translation initiation sites by facs-seq. *Molecular systems biology*, 10(8) : 748.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., *et al.* (2015) Reference sequence (refseq) database at ncbi : current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1) : D733–D745.

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., et Menschaert, G. (2015) sorfs. org : a repository of small orfs identified by ribosome profiling. *Nucleic acids research*, 44(D1) : D324–D329.

Olexiouk, V. et Menschaert, G. (2016) Identification of small novel coding sequences, a proteogenomics endeavor. In *Proteogenomics*, pages 49–64. Springer.

Osellame, L. D., Singh, A. P., Stroud, D. A., Palmer, C. S., Stojanovski, D., Ramachandran, R., et Ryan, M. T. (2016) Cooperative and independent roles of the drp1 adaptors mff, mid49 and mid51 in mitochondrial fission. *J Cell Sci*, 129(11) : 2170–2181.

Ouelle, D. E., Zindy, F., Ashmun, R. A., et Sherr, C. J. (1995) Alternative reading frames of the ink4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest. *Cell*, 83(6) : 993–1000.

Paek, K. Y., Hong, K. Y., Ryu, I., Park, S. M., Keum, S. J., Kwon, O. S., et Jang, S. K. (2015) Translation initiation mediated by rna looping. *Proceedings of the National Academy of Sciences*, 112(4) : 1041–1046.

Palmer, C. S., Osellame, L. D., Laine, D., Koutsopoulos, O. S., Frazier, A. E., et Ryan, M. T. (2011) Mid49 and mid51, new components of the mitochondrial fission machinery. *EMBO reports*, 12(6) : 565–573.

Pan, S., Chen, R., Aebersold, R., et Brentnall, T. A. (2011) Mass spectrometry based

glycoproteomics—from a proteomics perspective. *Molecular & Cellular Proteomics*, 10(1) : R110–003251.

Park, S. B., Seronello, S., Mayer, W., et Ojcius, D. M. (2016) Hepatitis c virus frameshift/alternate reading frame protein suppresses interferon responses mediated by pattern recognition receptor retinoic-acid-inducible gene-i. *Plos one*, 11(7) : e0158419.

Pauli, A., Norris, M. L., Valen, E., Chew, G.-L., Gagnon, J. A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., *et al.* (2014) Toddler : an embryonic signal that promotes cell movement via apelin receptors. *Science*, 343(6172) : 1248636.

Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., et Coon, J. J. (2012) Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Molecular & cellular proteomics*, 11(11) : 1475–1488.

Picotti, P. et Aebersold, R. (2012) Selected reaction monitoring-based proteomics : workflows, potential, pitfalls and future directions. *Nature methods*, 9(6) : 555–566.

Pisarev, A. V., Skabkin, M. A., Pisareva, V. P., Skabkina, O. V., Rakotondrafara, A. M., Hentze, M. W., Hellen, C. U., et Pestova, T. V. (2010) The role of abce1 in eukaryotic posttermination ribosomal recycling. *Molecular cell*, 37(2) : 196–210.

Poirier, J., Baccichet, A., Dea, D., et Gauthier, S. (1993) Cholesterol synthesis and lipoprotein reuptake during synaptic remodelling in hippocampus in adult rats. *Neuroscience*, 55(1) : 81–90.

Popa, A., Lebrigand, K., Barbry, P., et Waldmann, R. (2016) Pateamine a-sensitive ribosome profiling reveals the scope of translation in mouse embryonic stem cells. *BMC genomics*, 17(1) : 52.

Prabakaran, S., Hemberg, M., Chauhan, R., Winter, D., Tweedie-Cullen, R. Y., Dittrich, C., Hong, E., Gunawardena, J., Steen, H., Kreiman, G., *et al.* (2014) Quantitative profiling of peptides from rnas classified as non-coding. *Nature communications*, 5 : 5429.

Pruitt, K. D. et Maglott, D. R. (2001) Refseq and locuslink : Ncbi gene-centered resources. *Nucleic acids research*, 29(1) : 137–140.

Pueyo, J. I., Magny, E. G., et Couso, J. P. (2016a) New peptides under the s (orf) ace of the genome. *Trends in biochemical sciences*, 41(8) : 665–678.

Pueyo, J. I., Magny, E. G., Sampson, C. J., Amin, U., Evans, I. R., Bishop, S. A., et Couso, J. P. (2016b) Hemotin, a regulator of phagocytosis encoded by a small orf and conserved across metazoans. *PLoS biology*, 14(3) : e1002395.

Quanico, J., Franck, J., Cardon, T., Leblanc, E., Wisztorski, M., Salzet, M., et Fournier, I. (2016a) Nanolc-ms coupling of liquid microjunction microextraction for on-tissue proteomic analysis. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*.

Quanico, J., Franck, J., Dauly, C., Strupat, K., Dupuy, J., Day, R., Salzet, M., Fournier, I., et Wisztorski, M. (2013) Development of liquid microjunction extraction strategy

for improving protein identification from tissue sections. *Journal of proteomics*, 79 : 200–218.

Quanico, J., Franck, J., Gimeno, J., Sabbagh, R., Salzet, M., Day, R., et Fournier, I. (2015) Parafilm-assisted microdissection : a sampling method for mass spectrometry-based identification of differentially expressed prostate cancer protein biomarkers. *Chemical Communications*, 51(22) : 4564–4567.

Quanico, J., Franck, J., Salzet, M., et Fournier, I. (2016b) On-tissue direct monitoring of global hydrogen/deuterium exchange by maldi mass spectrometry : Tissue deuterium exchange mass spectrometry (tdxms). *Molecular & Cellular Proteomics*, 15(10) : 3321–3330.

Quinn, J. J. et Chang, H. Y. (2016) Unique features of long non-coding rna biogenesis and function. *Nature Reviews. Genetics*, 17(1) : 47.

R Core Team (2014) *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., Stephens, M., Gilad, Y., et Pritchard, J. K. (2016) Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*, 5 : e13328.

Ramamurthi, K. S. et Storz, G. (2014) The small protein floodgates are opening ; now the functional analysis begins. *BMC biology*, 12(1) : 96.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., et Zhang, F. (2013) Genome engineering using the crispr-cas9 system. *Nature protocols*, 8(11) : 2281–2308.

Ribrioux, S., Brüngger, A., Baumgarten, B., Seuwen, K., et John, M. R. (2008) Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC genomics*, 9(1) : 122.

Rolfe, D. et Brown, G. C. (1997) Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological reviews*, 77(3) : 731–758.

Rosenberger, G., Koh, C. C., Guo, T., Röst, H. L., Kouvonen, P., Collins, B. C., Heusel, M., Liu, Y., Caron, E., Vichalkovski, A., *et al.* (2014) A repository of assays to quantify 10,000 human proteins by swath-ms. *Scientific data*, 1 : 140031.

Ruiz-Orera, J., Messeguer, X., Subirana, J. A., et Alba, M. M. (2014) Long non-coding rnas as a source of new peptides. *Elife*, 3 : e03523.

Saavedra, J. M., Fernandez-Pardal, J., et Chevillard, C. (1982) Angiotensin-converting enzyme in discrete areas of the rat forebrain and pituitary gland. *Brain research*, 245(2) : 317–325.

Saghatelian, A. et Couso, J. P. (2015) Discovery and characterization of smorf-encoded bioactive polypeptides. *Nature chemical biology*, 11(12) : 909–916.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis,

K. B., et Erlich, H. A. (1988) Primer-directed enzymatic amplification of dna with a thermostable dna polymerase. *Science*, 239(4839) : 487–491.

Salzet, M. (2001) Neuroimmunology of opioids from invertebrates to human. *Neuroendocrinology Letters*, 22(6) : 467–474.

Salzet, M., Vieau, D., et Day, R. (2000) Crosstalk between nervous and immune systems through the animal kingdom : focus on opioids. *Trends in neurosciences*, 23(11) : 550–555.

Samandi, S., Roy, A. V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M. C., Vanderperre, B., Breton, M.-A., Jacques, J.-F., Brunelle, M., *et al.* (2017) Deep transcriptome annotation suggests that small and large proteins encoded in the same genes often cooperate. *bioRxiv*, page 142992.

Samrat, S. K., Li, W., Singh, S., Kumar, R., et Agrawal, B. (2014) Alternate reading frame protein (f protein) of hepatitis c virus : paradoxical effects of activation and apoptosis on human dendritic cells lead to stimulation of t cells. *PloS one*, 9(1) : e86567.

Sarafian, T. A., Ryan, C. M., Souda, P., Masliah, E., Kar, U. K., Vinters, H. V., Mathern, G. W., Faull, K. F., Whitelegge, J. P., et Watson, J. B. (2013) Impairment of mitochondria in adult mouse brain overexpressing predominantly full-length, n-terminally acetylated human α-synuclein. *PloS one*, 8(5) : e63557.

Sarsby, J., Martin, N. J., Lalor, P. F., Bunch, J., et Cooper, H. J. (2014) Top-down and bottom-up identification of proteins by liquid extraction surface analysis mass spectrometry of healthy and diseased human liver tissue. *Journal of The American Society for Mass Spectrometry*, 25(11) : 1953–1961.

Schägger, H. (2006) Tricine–SDS-PAGE. *Nature Protocols*, 1(1) : 16–22.

Schlötterer, C. (2015) Genes from scratch–the evolutionary fate of de novo genes. *Trends in Genetics*, 31(4) : 215–219.

Semba, S., Han, S.-Y., Qin, H. R., McCorkell, K. A., Iliopoulos, D., Pekarsky, Y., Druck, T., Trapasso, F., Croce, C. M., et Huebner, K. (2006) Biological functions of mammalian nit1, the counterpart of the invertebrate nitfhit rosetta stone protein, a possible tumor suppressor. *Journal of Biological Chemistry*, 281(38) : 28244–28253.

Sendoel, A., Dunn, J. G., Rodriguez, E. H., Naik, S., Gomez, N. C., Hurwitz, B., Levorse, J., Dill, B. D., Schramek, D., Molina, H., *et al.* (2017) Translation from unconventional 5' start sites drives tumour initiation. *Nature*, 541(7638) : 494.

Sharma, K., D'Souza, R. C., Tyanova, S., Schaab, C., Wiśniewski, J. R., Cox, J., et Mann, M. (2014) Ultradeep human phosphoproteome reveals a distinct regulatory nature of tyr and ser/thr-based signaling. *Cell reports*, 8(5) : 1583–1594.

Shaw, J. B., Li, W., Holden, D. D., Zhang, Y., Griep-Raming, J., Fellers, R. T., Early, B. P., Thomas, P. M., Kelleher, N. L., et Brodbelt, J. S. (2013) Complete protein characterization using top-down mass spectrometry and ultraviolet photodissociation. *Journal of the American Chemical Society*, 135(34) : 12646–12651.

Shishkova, E., Hebert, A. S., et Coon, J. J. (2016) Now, more than ever, proteomics needs better chromatography. *Cell systems*, 3(4) : 321–324.

Shuford, C. M., Sederoff, R. R., Chiang, V. L., et Muddiman, D. C. (2012) Peptide production and decay rates affect the quantitative accuracy of protein cleavage isotope dilution mass spectrometry (pc-idms). *Molecular & Cellular Proteomics*, 11(9) : 814–823.

Sidrauski, C., McGeachy, A. M., Ingolia, N. T., et Walter, P. (2015) The small molecule isrib reverses the effects of eif2α phosphorylation on translation and stress granule assembly. *Elife*, 4 : e05033.

Sienel, W., Varwerk, C., Linder, A., Kaiser, D., Teschner, M., Delire, M., Stamatis, G., et Passlick, B. (2004) Melanoma associated antigen (mage)-a3 expression in stages i and ii non-small cell lung cancer : results of a multi-center study. *European journal of cardio-thoracic surgery*, 25(1) : 131–134.

Simpson, K. D., Templeton, D. J., et Cross, J. V. (2012) Macrophage migration inhibitory factor promotes tumor growth and metastasis by inducing myeloid-derived suppressor cells in the tumor microenvironment. *The Journal of Immunology*, 189(12) : 5533–5540.

Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A., et Saghatelian, A. (2014) A human short open reading frame (sorf)-encoded polypeptide that stimulates dna end joining. *Journal of Biological Chemistry*, 289(16) : 10950–10957.

Slavoff, S. A., Mitchell, A. J., Schwaid, A. G., Cabili, M. N., Ma, J., Levin, J. Z., Karger, A. D., Budnik, B. A., Rinn, J. L., et Saghatelian, A. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nature chemical biology*, 9(1) : 59–64.

Smith, E., Meyerrose, T. E., Kohler, T., Namdar-Attar, M., Bab, N., Lahat, O., Noh, T., Li, J., Karaman, M. W., Hacia, J. G., *et al.* (2005) Leaky ribosomal scanning in mammalian genomes : significance of histone h4 alternative translation in vivo. *Nucleic acids research*, 33(4) : 1298–1308.

Smith, J. E., Alvarez-Dominguez, J. R., Kline, N., Huynh, N. J., Geisler, S., Hu, W., Coller, J., et Baker, K. E. (2014) Translation of small open reading frames within unannotated rna transcripts in saccharomyces cerevisiae. *Cell reports*, 7(6) : 1858–1866.

Smith, L. M., Kelleher, N. L., Linial, M., Goodlett, D., Langridge-Smith, P., Goo, Y. A., Safford, G., Bonilla, L., Kruppa, G., Zubarev, R., *et al.* (2013) Proteoform : a single term describing protein complexity. *Nature methods*, 10(3) : 186.

Song, J., Tan, H., Perry, A. J., Akutsu, T., Webb, G. I., Whisstock, J. C., et Pike, R. N. (2012) Prosper : an integrated feature-based tool for predicting protease substrate cleavage sites. *PloS one*, 7(11) : e50300.

Staudt, A.-C. et Wenkel, S. (2011) Regulation of protein function by 'microproteins'. *EMBO reports*, 12(1) : 35–42.

Stein, S. E. et Scott, D. R. (1994) Optimization and testing of mass spectral library search

algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5(9) : 859–866.

Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J., et Mateo, J. L. (2015) Cctop : an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS one*, 10(4) : e0124633.

Stern-Ginossar, N. et Ingolia, N. T. (2015) Ribosome profiling as a tool to decipher viral complexity. *Annual review of virology*, 2 : 335–349.

Storz, G., Wolf, Y. I., et Ramamurthi, K. S. (2014) Small proteins can no longer be ignored. *Annual review of biochemistry*, 83 : 753–777.

Su, M., Ling, Y., Yu, J., Wu, J., et Xiao, J. (2013) Small proteins : untapped area of potential biological importance. *Frontiers in genetics*, 4.

Sugihara, Y., Taniguchi, H., Kushima, R., Tsuda, H., Kubota, D., Ichikawa, H., Fujita, S., et Kondo, T. (2013) Laser microdissection and two-dimensional difference gel electrophoresis reveal proteomic intra-tumor heterogeneity in colorectal cancer. *Journal of proteomics*, 78 : 134–147.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., *et al.* (2014) String v10 : protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1) : D447–D452.

Thomson, E., Ferreira-Cerca, S., et Hurt, E. (2013) Eukaryotic ribosome biogenesis at a glance.

Toby, T. K., Fornelli, L., et Kelleher, N. L. (2016) Progress in top-down proteomics and the analysis of proteoforms. *Annual Review of Analytical Chemistry*, 9 : 499–519.

Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E., Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., Li, M., *et al.* (2011) Mapping intact protein isoforms in discovery mode using top down proteomics. *Nature*, 480(7376) : 254.

Trexler, A. J. et Rhoades, E. (2012) N-terminal acetylation is critical for forming α-helical oligomer of α-synuclein. *Protein Science*, 21(5) : 601–605.

Tsiatsiani, L. et Heck, A. J. (2015) Proteomics beyond trypsin. *The FEBS journal*, 282(14) : 2612–2626.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., *et al.* (2015) Tissue-based map of the human proteome. *Science*, 347(6220) : 1260419.

Ulitsky, I. et Bartel, D. P. (2013) lincrnas : genomics, evolution, and mechanisms. *Cell*, 154(1) : 26–46.

UniProt (2017) Uniprot : the universal protein knowledgebase. *Nucleic acids research*, 45(D1) : D158–D169.

UniProt-Consortium *et al.* (2008) The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl 1) : D190–D195.

Van Berkel, G. J. et Kertesz, V. (2009) Application of a liquid extraction based sealing surface sampling probe for mass spectrometric analysis of dried blood spots and mouse whole-body thin tissue sections. *Analytical chemistry*, 81(21) : 9146–9152.

Van Berkel, G. J. et Kertesz, V. (2013) Continuous-flow liquid microjunction surface sampling probe connected on-line with high-performance liquid chromatography/mass spectrometry for spatially resolved analysis of small molecules and proteins. *Rapid Communications in Mass Spectrometry*, 27(12) : 1329–1334.

Vanderperre, B., Lucier, J.-F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F.-M., et Roucou, X. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PloS one*, 8(8) : e70698.

Vanderperre, B., Staskevicius, A. B., Tremblay, G., McCoy, M., O'Neill, M. A., Cashman, N. R., et Roucou, X. (2011) An overlapping reading frame in the prnp gene encodes a novel polypeptide distinct from the prion protein. *The FASEB Journal*, 25(7) : 2373–2386.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., et Luscombe, N. M. (2009) A census of human transcription factors : function, expression and evolution. *Nature reviews. Genetics*, 10(4) : 252.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) The sequence of the human genome. *science*, 291(5507) : 1304–1351.

Viale, G., Slaets, L., de Snoo, F. A., Bogaerts, J., Russo, L., van't Veer, L., Rutgers, E. J., Piccart-Gebhart, M. J., Stork-Sloots, L., Dell'Orto, P., *et al.* (2016) Discordant assessment of tumor biomarkers by histopathological and molecular assays in the eortc randomized controlled 10041/big 03-04 mindact trial breast cancer. *Breast cancer research and treatment*, 155(3) : 463–469.

Viallet, P. M. et Vo-Dinh, T. (2003) Monitoring intracellular proteins using fluorescence techniques : from protein synthesis and localization to activity. *Current Protein and Peptide Science*, 4(5) : 375–388.

Villa, N. et Fraser, C. S. (2014) Mechanism of translation in eukaryotes. In *Translation and Its Regulation in Cancer Biology and Medicine*, pages 7–37. Springer.

Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., *et al.* (2015) 2016 update of the pride database and its related tools. *Nucleic acids research*, 44(D1) : D447–D456.

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., *et al.* (2014) Proteomexchange provides glo-

bally coordinated proteomics data submission and dissemination. *Nature biotechnology*, 32(3) : 223–226.

Waldemarson, S., Krogh, M., Alaiya, A., Kirik, U., Schedvins, K., Auer, G., Hansson, K. M., Ossola, R., Aebersold, R., Lee, H., *et al.* (2012) Protein expression changes in ovarian cancer during the transition from benign to malignant. *Journal of proteome research*, 11(5) : 2876–2889.

Walther, T., Albrecht, D., Becker, M., Schubert, M., Kouznetsova, E., Wiesner, B., Maul, B., Schliebs, R., Grecksch, G., Furkert, J., *et al.* (2009) Improved learning and memory in aged mice deficient in amyloid β-degrading neutral endopeptidase. *PLoS One*, 4(2) : e4590.

Walworth, M. J., ElNaggar, M. S., Stankovich, J. J., Witkowski, C., Norris, J. L., et Van Berkel, G. J. (2011) Direct sampling and analysis from solid-phase extraction cards using an automated liquid extraction surface analysis nanoelectrospray mass spectrometry system. *Rapid Communications in Mass Spectrometry*, 25(17) : 2389–2396.

Walworth, M. J., Stankovich, J. J., Van Berkel, G. J., Schulz, M., Minarik, S., Nichols, J., et Reich, E. (2010) Hydrophobic treatment enabling analysis of wettable surfaces using a liquid microjunction surface sampling probe/electrospray ionization-mass spectrometry system. *Analytical chemistry*, 83(2) : 591–597.

Wan, J. et Qian, S.-B. (2013) Tisdb : a database for alternative translation initiation in mammalian cells. *Nucleic acids research*, 42(D1) : D845–D850.

Wan, K. X., Vidavsky, I., et Gross, M. L. (2002) Comparing similar spectra : from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry*, 13(1) : 85–88.

Wang, L.-N., Tong, S.-W., Hu, H.-D., Ye, F., Li, S.-L., Ren, H., Zhang, D.-Z., Xiang, R., et Yang, Y.-X. (2012) Quantitative proteome analysis of ovarian cancer tissues using a itraq approach. *Journal of cellular biochemistry*, 113(12) : 3762–3772.

Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A. A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., et Segal, E. (2016) Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science*, 351(6270) : aad4939.

Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M. A., et Leutz, A. (2013) uorfdb—a comprehensive literature database on eukaryotic uorf biology. *Nucleic acids research*, 42(D1) : D60–D67.

Whitehouse, C. M., Dreyer, R., Yamashita, M., et Fenn, J. (1989) Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, 246(4926) : 64–71.

Wickham, H. (2016) *ggplot2 : elegant graphics for data analysis*. Springer.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502) : 582.

Wiśniewski, J. R., Hein, M. Y., Cox, J., et Mann, M. (2014) A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Molecular & cellular proteomics*, 13(12) : 3497–3506.

Wisniewski, J. R., Zougman, A., Nagaraj, N., et Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nature methods*, 6(5) : 359.

Wisztorski, M., Desmons, A., Quanico, J., Fatou, B., Gimeno, J.-P., Franck, J., Salzet, M., et Fournier, I. (2016) Spatially-resolved protein surface microsampling from tissue sections using liquid extraction surface analysis. *Proteomics*, 16(11-12) : 1622–1632.

Wisztorski, M., Fatou, B., Franck, J., Desmons, A., Farré, I., Leblanc, E., Fournier, I., et Salzet, M. (2013) Microproteomics by liquid extraction surface analysis : Application to ffpe tissue to study the fimbria region of tubo-ovarian cancer. *PROTEOMICS-Clinical Applications*, 7(3-4) : 234–240.

Woolford, J. L. et Baserga, S. J. (2013) Ribosome biogenesis in the yeast saccharomyces cerevisiae. *Genetics*, 195(3) : 643–681.

Wu, B., Eliscovich, C., Yoon, Y. J., et Singer, R. H. (2016) Translation dynamics of single mrnas in live cells and neurons. *Science*, 352(6292) : 1430–1435.

Wu, R., Haas, W., Dephoure, N., Huttlin, E. L., Zhai, B., Sowa, M. E., et Gygi, S. P. (2011) A large-scale method to measure absolute protein phosphorylation stoichiometries. *Nature methods*, 8(8) : 677–683.

Xu, G., Paige, J. S., et Jaffrey, S. R. (2010a) Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nature biotechnology*, 28(8) : 868–873.

Xu, H., Wang, P., Fu, Y., Zheng, Y., Tang, Q., Si, L., You, J., Zhang, Z., Zhu, Y., Zhou, L., *et al.* (2010b) Length of the orf, position of the first aug and the kozak motif are important factors in potential dual-coding transcripts. *Cell research*, 20(4) : 445.

Xu, X., Yu, X., Deng, X., Yue, M., Zhang, J., Zhu, D., Zhou, Z., Zhai, X., Xu, K., et Zhang, Y. (2014) Hepatitis c virus alternate reading frame protein decreases interferon-α secretion in peripheral blood mononuclear cells. *Molecular medicine reports*, 9(2) : 730–736.

Yandell, M. et Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics*, 13(5) : 329.

Yang, X., Tschaplinski, T. J., Hurst, G. B., Jawdy, S., Abraham, P. E., Lankford, P. K., Adams, R. M., Shah, M. B., Hettich, R. L., Lindquist, E., *et al.* (2011) Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome research*, 21(4) : 634–641.

Ye, H., Mandal, R., Catherman, A., Thomas, P. M., Kelleher, N. L., Ikonomidou, C., et Li, L. (2014) Top-down proteomics with mass spectrometry imaging : a pilot study towards discovery of biomarkers for neurodevelopmental disorders. *PloS one*, 9(4) : e92831.

Yosten, G. L., Liu, J., Ji, H., Sandberg, K., Speth, R., et Samson, W. K. (2016) A 5'-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *The Journal of physiology*, 594(6) : 1601–1605.

Yuryev, A., Kotelnikova, E., et Daraselia, N. (2009) Ariadne's chemeffect and pathway studio knowledge base. *Expert opinion on drug discovery*, 4(12) : 1307–1318.

Yus-Najera, E., Munoz, A., Salvador, N., Jensen, B., Rasmussen, H., Defelipe, J., et Villarroel, A. (2003) Localization of kcnq5 in the normal and epileptic human temporal neocortex and hippocampal formation. *Neuroscience*, 120(2) : 353–364.

Zanet, J., Benrabah, E., Li, T., Pelissier-Monier, A., Chanut-Delalande, H., Ronsin, B., Bellen, H., Payre, F., et Plaza, S. (2015) Pri sorf peptides induce selective proteasome-mediated protein processing. *Science*, 349(6254) : 1356–1358.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518) : 382.

Zhang, H., Liu, T., Zhang, Z., Payne, S. H., Zhang, B., McDermott, J. E., Zhou, J.-Y., Petyuk, V. A., Chen, L., Ray, D., *et al.* (2016a) Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, 166(3) : 755–765.

Zhang, L., Wu, S., Li, C., et Yang, Q. (2012) Facile synthesis of hybrid hollow mesoporous nanospheres with high content of interpenetrating polymers for size-selective peptides/proteins enrichment. *Chemical Communications*, 48(35) : 4190–4192.

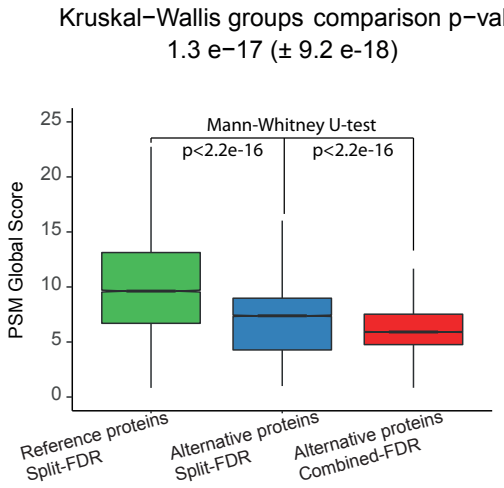
Zhang, Z., Liu, L., Wu, S., et Xing, D. (2016b) Drp1, mff, fis1, and mid51 are coordinated to mediate mitochondrial fission during uv irradiation–induced apoptosis. *The FASEB Journal*, 30(1) : 466–476.

Zheng, M., Seidah, N. G., et Pintar, J. E. (1997) The developmental expression in the rat cns and peripheral tissues of proteases pc5 and pace4 mrnas : comparison with other proprotein processing enzymes. *Developmental biology*, 181(2) : 268–283.

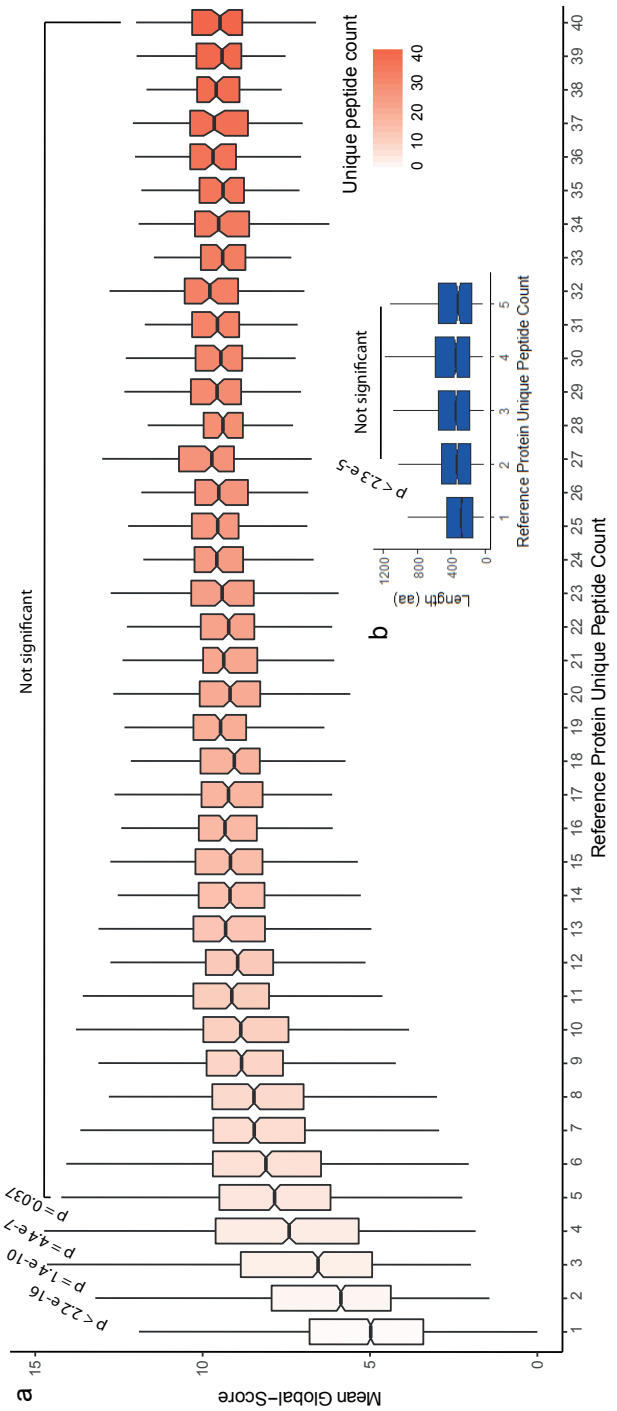
Zimmerman, T. A., Rubakhin, S. S., et Sweedler, J. V. (2011) Maldi mass spectrometry imaging of neuronal cell cultures. *Journal of the American Society for Mass Spectrometry*, 22(5) : 828.

Zolg, D. P., Wilhelm, M., Schnatbaum, K., Zerweck, J., Knaute, T., Delanghe, B., Bailey, D. J., Gessulat, S., Ehrlich, H.-C., Weininger, M., *et al.* (2017) Building proteometools based on a complete synthetic human proteome. *Nature Methods*.

ANNEXES



Distribution de scores de PSMs de protéines de référence (vert) et alternatives (bleu) par FDR séparé et alternatives par FDR combiné (rouge).



(a) Distribution de scores des protéines de référence en fonction du nombre de peptides uniques qui ont permis leur identification. (b) Distribution de la longueur en acides aminés des protéines de référence en fonction du nombre de peptides qui ont permis leur identification. Les p-values indiquées représentent la p-value de Mann-Whitney U tests de paire à paire entre les protéines identifiées avec un nombre de peptides donné et le niveau supérieur.