UNIVERSITÉ DE LILLE

ÉCOLE DOCTORALE DE BIOLOGIE-SANTÉ

CNU : Biologie Cellulaire

THÈSE DE DOCTORAT

En vue de l'obtention du grade de Docteur en Sciences de l'Université de Lille

Présenté par

TRISTAN CARDON

Relations, Interactions et Fonctions des Protéines Alternatives

À Lille le 10 octobre 2019

Présenté devant le jury composé de :

Président du jury	Mr Didier VIEAU	Professeur (Université de Lille)
Rapporteur	Mme Virginie REDEKER	Chargé de recherche Hors Classe (Inserm)
Rapporteur	Mr Jean ARMENGAUD	Directeur de recherche (CEA)
Examinateur	Mr Kris GEVAERT	Professeur (Université de Gent)
Directrice de Thèse	Mme Isabelle FOURNIER	Professeur (Université de Lille)
Co-encadrant	Mr Julien FRANCK	Maître de Conférences (Université de Lille)

> *“*
>
> *L'imagination est plus importante que le savoir, car si le savoir concerne tout ce qui existe, l'imagination concerne tout ce qui existera*
>
> *Albert Einstein*

Remerciements

Un grand merci aux Professeurs Michel Salzet et Isabelle Fournier, Directeur et co-Directrice du laboratoire, qui ont cru en moi et en ce projet ambitieux et un peu fou. Merci pour leur temps et investissement jusqu'aux dernières minutes de ce travail ainsi que pour leur implication dans la direction de ma thèse, dans l'espoir de futures collaborations.

Je dois évidemment souligner l'implication de mon co-encadrant, le Docteur Julien Franck, sans qui les projets n'auraient jamais autant avancé. Merci pour le soutien que tu m'as apporté durant ces 3 années, ainsi que pour tes enseignements … sans oublier les « After Work » du pub irlandais.

Je remercie la Professeure Redeker et le Professeur Armengaud qui m'ont fait l'honneur d'accepter d'être rapporteurs de mon travail, ainsi que les membres du jury le Professeur Vieau et le Professeur Gevaert pour avoir accepté de juger mon travail.

Merci à Maxence, qui m'a supporté les longues heures passées dans son bureau lors de mes débats et questions avec Julien (*ta patience est un modèle de maitrise*), à Annie qui m'a accompagné dans le meilleur et le pire de mes expériences à la paillasse, « *piètre élève est celui qui ne dépasse pas le maitre* » mon chemin est encore long, à Jusal mon premier maitre de stage au laboratoire merci d'avoir cru en moi.

Je souhaite remercier particulièrement ces personnes qui font du quotidien un moment exceptionnel, mes collègues devenus des ami(e)s au fil des ans, tout d'abord ceux devenus grand quand j'étais encore jeune : Marie, Stéphanie, Benoit, Khalil, Adriana, Tanina vous avez tous et toutes été à votre manières des exemples. Une mention spéciale pour les 3 filles : Mélanie, Flore et Lauranne ayant partagé et subi ma dernière année et rédaction de ce manuscrit (*Bénis soit la technologie Noise Cancelling*). Philippe toujours présent pour une expérience dingue et foireuse ou une discussion plus ou moins scientifique. Quentin mon alter ego, qui a affronté les épreuves de rédaction en même temps que moi.

Thank's to the Italian team of Daniele, who gives me the opportunities of a biological model and application, special thank's to Marina coming a friend, we

miss you in the cold region of the north of France, I expect the best for your future. Thank's to Nina who share this friendship, and for this advice.

Merci à Soulaimane ingénieur au laboratoire mais avant tout ami depuis mes débuts sur les bancs de l'université. Merci aux membres du laboratoire qui partagent mes journées : Irène, Raphaël, Christophe, Franck, Jean Pascal, Christelle, Pierre Eric et Jacopo.

Je voudrais remercier mes amis, trop nombreux pour être détaillés, de faire partie de ma vie et de m'avoir toujours soutenu même si vous n'avez sans doute jamais vraiment compris ce que je faisais.

Merci à ma mère Christine Panato, qui restera à jamais mon modèle de force, de patience et de conviction, elle a toujours défendu mes choix, jamais je n'aurais été capable de publier cet ouvrage sans toi, je ne te remercierai jamais assez (les batailles tard le soir pour mes dictées ont finalement fini par payer…ou presque). Merci à ma petite sœur loëssa, tu m'as vu lutter, galérer depuis le début pourtant tu ne m'as jamais conseillé d'arrêter.

Enfin je remercie le destin d'avoir mis sur ma route la femme de ma vie et mon épouse Antonella, sans toi chaque jour à mes côtés je n'aurais probablement pas été si fort. Merci de transformer mes échecs en victoires, mes difficultés en expériences et mes défauts en qualités. Tu m'as appris la patience et la maitrise de soi, je te dois ce travail, je ne remercierai jamais assez ton soutien au long de la rédaction de cette thèse.

Pour finir une pensée à la personne à qui je dois mon amour pour la science cette Professeure de CM2 dont la salle de classe était remplie d'insectes, de papillons, de pierres en tout genre exposés et suscitant la curiosité, merci Madame Napora

Productions scientifiques

Publications :

1. **Probing the Function of Alternative Proteins in Cell Reprogramming by Large Scale analyses**
   Tristan Cardon, Julien Franck, Marina Damato, Michele Maffia, Daniele Vergara, Isabelle Fournier, Michel Salzet
   (2019-en publication)

2. **Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins**
   Tristan Cardon, Flore Hervé, Vivian Delcourt, Xavier Roucou, Michel Salzet, Julien Franck, Isabelle Fournier
   (2019-en publication)

3. **MALDI MSI of Lipids in Experimental Model of Traumatic Brain Injury Detects Acylcarnitines as Injury Related Markers**,
   Khalil Mallah, Jusal Quanico, Antonella Raffo-Romero, Tristan Cardon, Soulaimane Aboulouard, David Devos, Firas Kobeissy, Kazem Zibara, Michel Salzet, Isabelle Fournier

   Analytical chemistry, 2019
   doi:10.1021/acs.analchem.9b02633

4. **Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation**,
   Tristan Cardon, Michel Salzet, Julien Franck, Isabelle Fournier

   Biochimica et Biophysica Acta (BBA)-General Subjects, 2019.
   doi:10.1016/J.BBAGEN.2019.05.009.

5. **Mapping spatiotemporal microproteomics landscape in experimental model of traumatic brain injury unveils a link to Parkinson's disease**'
Khalil Mallah, Jusal Quanico, Antonella Raffo-Romero, Tristan Cardon, Soulaimane Aboulouard, David Devos, Firas Kobeissy, Kazem Zibara, Michel Salzet and Isabelle Fournier

Molecular & Cellular Proteomics, 2019, p. mcp. RA119. 001604
doi: 10.1074/mcp.RA119.001604.

6. **NanoLC-MS coupling of liquid microjunction microextraction for on-tissue proteomic analysis**
Jusal Quanico, Julien Franck, Tristan Cardon, Eric Leblanc, Maxence Wisztorski, Michel Salzet, Isabelle Fournier

Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2017, vol. 1865, no 7, p. 891-900.
doi: 10.1016/j.bbapap.2016.11.002.

7. **The multiverse nature of epithelial to mesenchymal transition**
Pasquale Simeone, Marco Trerotola, Julien Franck, Tristan Cardon, Marco Marchisio, Isabelle Fournier, Michel Salzet, Michele Maffia, Daniele Vergara

Seminars in cancer biology. Academic Press, 2018
doi: 10.1016/J.SEMCANCER.2018.11.004.

8. **Distinct Protein Expression Networks are Activated in Microglia Cells after Stimulation with IFN-γ and IL-4**
Daniele Vergara, Annamaria Nigro, Alessandro Romano, Stefania De Domenico, Marina Damato, Julien Franck, Chiara Coricciati, Maxence Wistorski, Tristan Cardon, Isabelle Fournier, Angelo Quattrini, Michel Salzet, Roberto Furlan and Michele Maffia

Cells, 2019, vol. 8, no 6, p. 580.
doi: 10.3390/cells8060580.

Présentations Orales :

1. 13-14/10/2016 : **EURON Phd-Day** 2016, Lille, France,
   Can Alternative Proteins Bring New Insights to Spinal Cord Injury?

2. 20-24/03/2017 : Présentation au **Club Jeunes de la SFSM** (Société Française de Spectrométrie de Masse) – Trélon, France,
   Tracking biomarkers involved in neuro-paludisme, by mass spectrometry imaging and microextraction.

3. 05-07/04/2017 : Présentation au **Club Jeunes de la SFEAP** (Société Française d'Electrophorèse et d'Analyse Protéomique) - Montpellier, France,
   Can Alternative Proteins Bring New Insights to Spinal Cord Injury?

4. 25-26/10/2017 **EURON Phd-Day** 2017 - Maastricht, Pays-bas,
   Alternative Proteins the submerged part of proteomic

5. 18-20/04/2018 : Présentation au **Club Jeunes de la SFEAP** (Société Française d'Electrophorèse et d'Analyse Protéomique) – Rennes, France,
   Searching for Ghost Proteins Interactome

6. 16-20/06/2018 : Présentation au congrès **EUPA** (EUropean Proteomic Association) – Saint Jacques de Compostelle, Espagne,
   Searching for Ghost Proteins Interactome
7. 10/09/2018 : présentation **journée Andrée Verbert** Lille

Posters:

1. 13-14/10/2016 **EURON Phd-Day** – Lille, France,
   Can Alternative Proteins Bring New Insights to Spinal Cord Injury ?

2. 1/12/2016 **BIOFIT** – Lille, France,
   Alternative proteins: the hidden world of potential biomarkers on cancer

3. 05-07/10/2017 **SMMAP** – Paris, France,
   Alternative proteins: The hidden world of spinal cord injury?

4. 25-26/10/2017 **EURON Phd-Day** 2017 – Maastricht, Pays-bas,
   Alternative Proteins the submerged part of proteomic

Encadrements

- Flore Hervé 2017 stage de 6 mois
- Justine Fontaine 2018 stage de 6 mois
- Sylvain Osien 2019 stage de 6 mois
- Philipp Kaulich 2019 stage de 6 mois

Résumé :

Si en transcriptomique le dogme accepté par la communauté veut qu'un ARNm code pour une protéine unique, la protéomique vient de montrer l'inverse. Force est de constater que les ARNm peuvent traduire plusieurs protéines. Celles ne suivant pas le cadre de référence sont appelées protéines alternatives (AltProts) et forment le protéome caché ou fantôme. Ces AltProts nécessitent la mise en place de stratégies adaptées pour leur mise en évidence. Leurs caractéristiques physicochimiques spécifiques, telles que leur petite taille permet d'adapter les méthodes classiques de protéomique à leur étude. Dans cet objectif la mise en évidence des AltProts par différentes méthodes d'extraction, notamment adaptées des méthodes de peptidomique, a permis de mettre en évidence les conditions d'enrichissement avant une analyse *bottom-up.* Ces AltProts sont une nouvelle classe de protéines pour laquelle très peu d'informations fonctionnelles sont connues. Les prédictions de fonction avancées lors des premières constructions de bases de données, annonçaient des fonctions dans la régulation des ARN, de la synthèse de protéines et de la régulation d'expression des gènes par association avec des facteurs de transcription. Ces prédictions étaient basées sur les homologies de séquences entre les AltProts et les protéines de référence (RefProts). Cependant très peu d'études montrent le rôle de ces protéines de manière expérimentale. Afin de mettre en évidence les fonctions de ces AltProts, nous avons choisi de retrouver leurs partenaires d'interaction. À l'heure actuelle, plusieurs méthodes existent permettant d'étudier l'interactome des protéines, toutefois la majorité est dirigée vers une cible, nécessitant parfois des constructions biochimiques ou l'utilisation d'anticorps dirigés, rendant ces méthodes difficiles à mettre en place pour les AltProts. Seule la méthode de pontage chimique couplée à la spectrométrie de masse (XL-MS) permet d'observer des interactions cellulaires de manière non ciblée. Cette méthode de pontage chimique, bien que connaissant ses propres limitations, est applicable à la recherche des partenaires d'interaction des AltProts. Cet outil, associé aux logiciels de traitement des réseaux d'interaction, enrichi par les interactions connues entre RefProts dans la littérature, permet de replacer les AltProts dans ces réseaux. Ces réseaux peuvent ensuite être traités afin de mettre en évidence les voies de signalisation impliquant les RefProts et ainsi déduire les différentes voies de signalisation associées aux AltProts observées pontées aux RefProts.

Summary:

If in transcriptomics the dogma accepted by the community is that a single mRNA codes for a single protein, proteomics has just shown the opposite. It must be said that mRNAs can translate several proteins. These not following the reference framework are called alternative proteins (AltProts) and form the hidden or ghost proteome. These AltProts require the implementation of appropriate strategies to highlight them. Their specific physicochemical characteristics, such as their small size, make it possible to adapt classical proteomic methods to their study. With this objective in mind, the identification of AltProts by different extraction methods, particularly adapted to peptidomic methods, made it possible to highlight the enrichment conditions before a bottom-up analysis. These AltProts are a new class of proteins for which very little functional information is known. Advanced function predictions in the early database constructions announced functions in RNA regulation, protein synthesis and gene expression regulation by association with transcriptional factors. These predictions were based on sequence homologies between AltProts and reference proteins (RefProts). However, very few studies show the role of these proteins in an experimental way. In order to highlight the functions of these AltProts, we have chosen to find their interaction partners. At present, several methods exist to study the protein interactome, however the majority are directed towards a target, sometimes requiring biochemical constructs or the use of directed antibodies, making these methods difficult to implement for AltProts. Only the Crosslink method coupled with mass spectrometry (XL-MS) allows to observe cellular interactions in a non-targeted way. This chemical bridging method, although aware of its own limitations, is applicable to the search for AltProts interaction partners. This tool, combined with the software for processing interaction networks, enriched by the known interactions between RefProts in the literature, makes it possible to replace AltProts in these networks. These networks can then be processed to highlight the signaling pathways involving RefProts and thus deduce the different signaling pathways associated with the observed AltProts crosslinked to the RefProts.

Table des matières

Liste des figures :

Liste des Tables :

Liste des abréviations :

ACN	Acétonitrile
AltORF	ORF alternatif
AltProt	Protéine alternative
APEX	Ascorbate Peroxidase-catalyzed proximity labeling
AP-MS	Affinity Purification-Mass Spectrometry
APP	Amyloid Precursor Protein
ARNlnc	ARN long non codant
ARNlnc-IUR	ARNlnc imatinib-upregulated
ARNm	Acides ribonucléiques messagers
ARNmt	ARN mitochondriale
ARNnc	ARN non codant
Azide-A-DSBSO	Azide-A-Disuccinimidyl bis-sulfoxide
BAMG	Bis(succinimidyl)-3-azidomethyl glutarate
BioID	Proximity-dependent Biotin Identification
BS3	Bis(sulfosuccinimidylsuberate)
BuUrBu	Disuccinimidyl dibutyric urea
cAMP	Cyclic adenosine monophosphate
CCDS	Consensus CDS
CDS	Coding DNA Sequence
cHPP	Chromosome-centric Human Proteome Project
CID	Collision Induced Dissociation
coIP	Co-immunoprécipitation
DAU	1,3-diallylurea
DEST	Diethylsuberthioimidate
DSS	Disuccinimide suberate
DSSO	Disuccinimidyl sulfoxide
DTSSP	3,3'-Dithiobis(sulfosuccinimidylpropionate)
EM	Electron Microscopy
EMT	Epithelial–Mesenchymal Transition
EMT-TFs	EMT Transcription Factors
EST	Expressed Sequence Tags
ETD	Electron Transfer Dissociation
FLOSS	Fragment Length organization similarity score
FRET	Forster Resonance Energy Transfer
FSPP	smORF-encoded peptides predictor
GO	Gene Ontology
HN	Humanin
IDP	Intrinsically Disordered Proteins
IMAC	Immobilized Metal Affinity Chromatography
IP-MS	Immunoprecipitation - Mass Spectrometry
IRM	Imagerie par résonnance magnétique
LC-MS	Liquid Chromatography –Mass Spectrometry

LESA	Liquid Extraction Surface Analysis
LFQ	Label Free Quantification
LPS	Lipopolysaccharides
MET	Mesenchymal–epithelial transition
MeOH	Methanol
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MSI	Mass Spectrometry Imaging
NHS	N-hydroxysuccinimide
OMS	Organisme Mondiale de la Santé
ORF	Open Reading Frame
PAM	Parafilm Assisted Microdissection
PLA	Proximity Ligation Assay
PPI	Protein-Protein Interaction
PrP	Prion Protein
PS1&2	Presenilin 1 et 2
PTM	Post Translation Modification
RefORF	Reference Open Reading Frame
RefProt	Reference Protein
RNA-seq	RNA sequencing
SCX	Strong Cation Exchange Chromatography
SDS	Sodium dodecyl sulfate
SEC	Size Exclusion Chromatography
SEPs	sORF-encoded polypeptides
sORF, smORF	short-ORF, small-ORF
SPE	Solid Phase Extraction
Sulfo-SBED	Sulfo-N-hydroxysuccinimidyl-2-(6-[biotinamido]-2-(p-azido benzamido)-hexanoamido) ethyl-1,3'-dithioproprionate
TCA	Trichloroacetic acide
upORF	Upstream ORF
UTR	Untranslated Regions
UVPD	ultraviolet photodissociation
VLP	Viral Like Particule
XL-MS	Cross-Linking-Mass Spectrometry
Y2H	Yeast Two-Hybrid
ZEB1&2	Zinc Finger E-box binding homeobox 1 & 2

PARTIE I
INTRODUCTION

I. La mise en évidence d'un protéome fantôme

Les stratégies d'analyses protéomiques à grande échelle offrent l'accès à l'identification et à la quantification relative de plusieurs milliers de protéines. Aujourd'hui il est possible d'identifier et de quantifier de manière relative plus de 10 000 protéines en moins de 2h [1]. Les analyses protéomiques grande échelle d'échantillons complexes dites *shotgun* sont en particulier devenues non seulement rapides mais aussi très robustes. Elles constituent à l'heure actuelle les stratégies les plus couramment utilisées en protéomique. Cependant, les stratégies protéomiques à grande échelle reposent sur l'identification des protéines via l'interrogation en banques de données. Ainsi, seules les protéines présentes dans ces banques peuvent être identifiées. Elles ne permettent donc pas la découverte et la mise en évidence de nouvelles protéines. Cependant, les analyses *shotgun* montrent toujours un pourcentage non négligeable (>10 % environ [2,3]) de données qui ne trouvent pas de correspondance au sein des banques de données, alors que les données de spectrométrie de masse en tandem (MS/MS) présentent une qualité suffisante pour confirmer la présence d'une séquence protéique identifiée.

Ainsi en 2010, dans le cadre d'une collaboration entre le laboratoire PRISM et le professeur *Roucou de l'Université de Sherbrooke*, cette constatation a poussé les deux équipes à rechercher l'existence de protéines encore non répertoriées dans les banques de données traditionnelles telles que Uniprot. Le professeur *Roucou* ayant mis en évidence la capacité du gène PRNP à coder pour deux protéines non homologues et de structures primaires différentes, [4], il a donc développé sur la base de ses travaux une nouvelle base de données, nommée HaltORF [5]. Cette base de données répertorie toutes les prédictions de traductions protéiques issues d'un même acide ribonucléique messager (ARNm) mature, y compris les protéines ne suivant pas les règles de traduction de la séquence Kozak (nommées protéines alternatives ou AltProts). Cette nouvelle base de données a permis de combler les lacunes des bases de données traditionnelles couramment utilisées en protéomique.

En effet, les bases de données usuelles sont établies suivant des règles strictes ne considérant que les protéines de taille supérieure à 100 acides aminés (soit 300 nucléotides), et respectant le contexte Kozak optimal [6–9]. Ainsi, seuls les cadres de lecture ouverts (ORF) les plus longs sont prédits comme référence (RefORF). Ces règles négligent donc les protéines de moins de 100 acides aminés issues de petits ORF (sORF, smORF) ainsi que celles ne respectant pas le contexte Kozak (décalage du cadre de lecture, chevauchement avec une région non codante, cadre de lecture situé sur les régions 3' et 5' non codantes ou encore celles qui ne suivent pas la règle d'un codon START AUG). Ces ORF non pris en compte sont nommés ORF alternatifs (encore notés AltORF). Ces protéines alternatives (AltProts) constituent un nouveau protéome, dit protéome caché ou protéome fantôme [12], dont les fonctions des protéines qui le constitue restent à découvrir.

« Les AltProts remettent-elles en question un dogme fondamental sur les règles de traduction des protéines : un ARN messager mature code pour une protéine unique ? »

En parallèle de ces travaux, d'autres équipes ont mis en évidence par des techniques de génomique et de transcriptomique, la capacité du ribosome à se fixer sur des régions non codantes ou en décalage sur la séquence codante appelée CDS (*Coding DNA Sequence*) inclus dans le cadre de lecture ouvert. Ces mécanismes permettent la production d'AltProts dans les régions 3' et 5' UTR, ainsi que de protéines chevauchant la région codante du CDS et les régions 3' ou 5' UTR [10,11]. Les AltProts obtenues par ces mécanismes diffèrent par leurs séquences en acides aminés des protéines de référence (RefProt) initialement décrites pour les ARNm matures. Les ARN longs non codants (ARNlnc) sont un autre exemple de mécanisme de traduction non conventionnel. En effet la découverte de sites de fixation du ribosome et la présence de séquences d'initiation en l'absence de séquences Kozak idéales, laissent supposer que ces ARNlnc peuvent produire des protéines qui pourraient jouer un rôle dans la régulation des voies de signalisation cellulaire [12].

L'interrogation des données de protéomique via les banques AltProts permet de confirmer que ces mécanismes conduisent effectivement à la traduction de protéines. Les données de spectrométrie de masse (MS) issues des stratégies *shotgun* [3,13] et *top-down* [14,15] combinées aux nouvelles bases de données fondées sur la prédiction des AltProts à partir des ARNm de RefProt, a permis de mettre en évidence 1259 nouvelles protéines [3]. Une étude de quantification déterminant le niveau d'expression des protéines traduites à partir du gène MIEF1 dont la RefProt MiD51 et l'AltProt AltMiD51 dans deux lignées cellulaires et dans les tissus du colon, a révélé une expression deux fois plus importante de AltMiD51 comparé à sa RefProt MiD51 [16]. Ces données renforcent l'idée que les AltProts peuvent jouer un rôle important dans les voies de signalisation. De plus, la mise en évidence d'AltProts dans le sécrétome cellulaire ou dans les fluides biologiques conduit également à émettre l'hypothèse qu'elles participent activement à la régulation des fonctions biologiques. Ces protéines ne sont pas uniquement l'apanage des mammifères. En effet, les AltProts ont également été mises en évidence chez les bactéries [17–20], la drosophile [21], les plantes [22] et d'autres eucaryotes (oursin, drosophile, rat) [23]. La diversité de ces protéines montre que celles-ci ne sont pas exprimées uniquement lors de conditions pathologiques ou de stress, mais seraient impliquées dans la régulation physiologique des systèmes. Quelques fonctions ont déjà été décrites pour ces AltProts notamment dans la réparation de l'ADN, le décapage des ARN [24], la voie de signalisation mTor [25], la régulation de l´homéostasie calcique [26], la formation de myoblastes [27], la performance musculaire [27], la fission mitochondriale et la régulation de la traduction et du ribosome [28]. Des expressions d'AltProt spécifiques ont également été mises en évidence dans des contextes physiopathologiques notamment lors d'études sur le cancer de l'ovaire [15], les gliomes [13] ou de la lésion de la moelle épinière (données non publiées).

Cependant, les fonctions de ces protéines restent encore largement méconnues. L'étude de l'expression de ces protéines est rendue difficile par le fait qu'aucun anticorps n'existe commercialement pour ces protéines. En conséquence, l'utilisation de stratégies à base d'anticorps ne peut se réaliser que

sur un nombre limité de candidats. L'inhibition de AltProts est également une stratégie envisageable pour comprendre l'effet d'une protéine cible sur les voies de signalisation intracellulaires. Ces études peuvent être réalisées, par exemple, par la technique de CRISPR-Cas9 si le gène peut-être ciblé ou bien par siRNA en ciblant le transcrit. L'impact de cette inhibition sur le phénotype cellulaire, les voies de signalisation, ou encore la régulation de l'expression d'autres protéines est une stratégie également pertinente mais qui reste ciblée et donc limitée à un nombre de candidats restreints et d'intérêt. Notre volonté de comprendre à plus large échelle, dans un premier temps, la fonction des AltProts, nous a entraînée à rechercher des stratégies non ciblées à plus grande échelle. Ainsi notre intérêt s'est porté sur les stratégies permettant la recherche des partenaires d'interaction protéine-protéine (PPI ; *protein-protein interaction*).

Diverses approches sont proposées pour identifier les PPIs. La plus répandue reste la méthode ciblée de purification par affinité telle que la co-immunoprécipitation (coIP). D'autres stratégies telles que l'Apex, le BioID et la Vitrotrap basées sur le co-marquage de la protéine cible (ou protéine appât) et de ses partenaires sont particulièrement efficaces et connaissent une application croissante. Cependant, toutes ces méthodes sont ciblées et spécifiques d'une protéine, nécessitant soit un anticorps de capture, soit une modification de la protéine cible permettant le marquage de ses partenaires. En revanche, les stratégies de pontage chimique des protéines couplées à la spectrométrie de masse (pontage chimique analysé par mass spectrométrie ou XL-MS) permettent de rechercher les PPIs de façon non ciblée à partir d'une faible quantité de matériel biologique et à grande échelle dans des mélanges complexes. La stratégie XL-MS se base sur le pontage chimique entre les chaînes latérales des acides aminés de deux protéines proches dans l'espace. Ce pontage fige les systèmes et permet ensuite après analyse protéomique basée sur la MS d'identifier les partenaires d'interaction en interaction ainsi que de déterminer le site de l'interaction. Cette méthode, outre l'identification des partenaires d'interaction, fournit des données permettant l'étude structurale des protéines seules ou en complexe en combinaison avec des données d'autres modalités

comme les RMN, la cristallographie et la microscopie électronique de type cryo-EM.

L'objectif de ma thèse était de rechercher les partenaires d'interaction des AltProts par des approches XL-MS afin de les replacer au sein des voies de signalisation, puis d'appréhender ces interactions de façon dynamique afin de mieux comprendre les mécanismes induits par les processus physiologiques et physiopathologiques. Cet objectif m'a amené à m'intéresser au développement de stratégies optimisées de protéomique pour l'analyse des protéines alternatives ainsi qu'au développement de stratégies de type XL-MS. Dans le but de répondre à cette problématique trois objectifs ont été traités comme suit.

Le premier objectif de ma thèse a donc consisté à développer des méthodes permettant d'extraire et d'enrichir de façon optimale les AltProts sachant que celles-ci représentent environ 10-15% des protéines totales et sont en moyenne de petites tailles comparativement aux protéines conventionnelles (certaines étant de la taille d'un neuropeptide). En particulier, ces études ont porté sur la comparaison des méthodes d'extraction ainsi que sur les méthodes permettant d'enrichir les AltProts.

Le second objectif de ma thèse a été focalisé sur le développement de la stratégie de XL-MS non ciblée en combinaison avec l'identification des AltProts. Les développements réalisés ont porté sur les différentes étapes de la stratégie telle décrite **Figure 1** et notamment sur l'optimisation de l'étape de pontage chimique à l'échelle de l'ensemble des protéines, la digestion enzymatique des protéines pontées en peptides, l'enrichissement des peptides pontés générés par chromatographie échangeuse de cations (SCX) ou d'exclusion stérique (SEC) puis l'analyse par LC-MS et l'identification des protéines en interaction à l'aide de solutions dédiées.

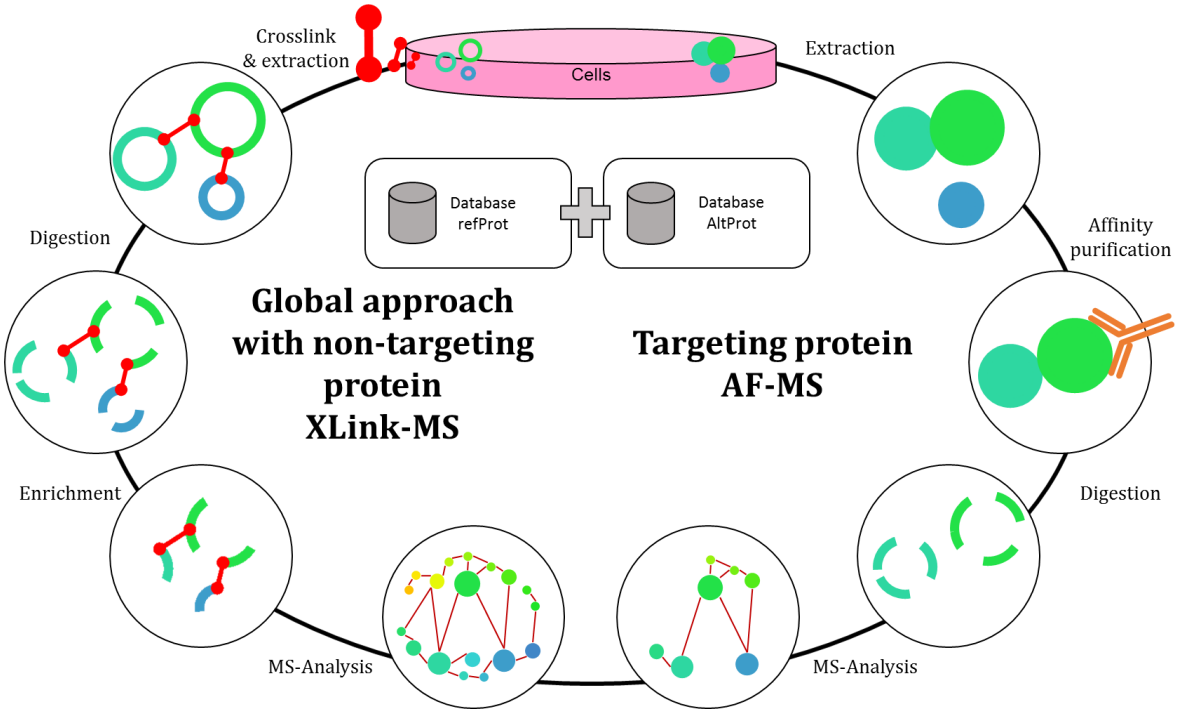


Figure 1 : ***Schéma récapitulatif de la méthodologie de pontage chimique (XI-MS) appliquée et la recherche des partenaires des AltProts suivant une approche ciblée et non ciblée****. L'utilisation de la stratégie XL-MS peut être réalisée selon deux voies. La première, ciblée, nécessite la connaissance de la cible afin de réaliser une purification par affinité, le complexe purifié est alors analysé. Dans une deuxième voie l'ensemble des interactions protéine-protéine d'une cellule sont fixées puis analysées.*

Comme présentée **Figure 1**, cette stratégie de XL-MS peut être appliquée lors d'une approche ciblée et non ciblée. La stratégie non ciblée a pour but de considérer l'ensemble de l'interactome sans cible protéique avec un pontage chimique à grande échelle. La deuxième est une stratégie ciblée reposant sur la production d'une AltProt spécifique porteuse d'un « FLAG » permettant ainsi de pallier l'absence d'anticorps anti AltProt, ceci permettant de purifier l'interactome des AltProts via une expérience de co-immunoprecipitation (coIP) couplée à la MS (AP-MS) [15].

Enfin, le troisième objectif de ma thèse a consisté à appliquer les méthodologies développées dans un contexte physiopathologique et plus précisément dans le cadre de la reprogrammation de cellules cancéreuses. Une étude de a donc été conduite sur des cellules humaines de glioblastome de type NCH82 dans des conditions natives et sous l'effet d'une stimulation à la Forskoline. La Forskoline est un activateur de la voie de signalisation de

l'adénylate cyclase (cAMP) et est notamment décrite comme induisant le changement phénotypique de cellules de glioblastome vers une transition métastatique. Cette étude nous a conduit à mettre en évidence le rôle des AltProts dans la signalisation liée à la reprogrammation des cellules de NCH82 par stimulation via la cartographie des PPI par stratégies XL-MS pour les RefProts tout comme les AltProts.

Ainsi l'application dans un contexte physiopathologique des développements méthodologiques réalisés au cours de ce travail nous amène à préciser la fonction des protéines alternatives et représente une étude princeps à de futures études.

PARTIE II
État de l'Art

I. Introduction aux AltProts

1. Notion de traduction protéique

A. La traduction protéique : une machinerie complexe

Chez les eucaryotes l'ADN contient les gènes constitués de 4 nucléotides (A,C,T,G), ces nucléotides sont le support de la transcription en ARNm également composé de 4 nucléotides (A,U,G,C) formant par groupe de 3 lettres un codon. C'est cet ARNm qui par traduction permet la production de protéines composées majoritairement de 20 acides aminés différents. La traduction est une étape complexe de décodage de l'ARNm, ceci est réalisé par le Ribosome. Les ribosomes sont formés de deux sous unités organisées à partir de différents types moléculaires, on y trouve des protéines et de l'ARN ribosomique (ARNr). La petite sous unité permet le décodage de la séquence ARNm, tandis que la grande réalise la synthèse protéique par catalyse des liaisons entre acides aminés. Cette synthèse nécessite également l'implication des ARN de transfert (ARNt) qui apportent les acides aminés au ribosome. Les ARNt sont intégrés via 3 sites sur le ribosome, le site A où les ARNt se fixent sur le codon correspondant à leur acide aminé, le site P qui contient l'ARNt rattaché à la chaîne polypeptidique en construction et enfin le site E dans lequel l'ARNt est déacylé puis éjecté du ribosome. Le mécanisme de traduction est décrit en 4 étapes majeures : l'initiation, l'élongation, la terminaison et le recyclage. Le mécanisme d'initiation 5' dépendant est crucial dans la sélection du premier acide aminé constituant la protéine traduite. Elle représente également le moment de sélection de la séquence nucléotidique traduite et ainsi l'ORF. On note différentes étapes dans cette phase d'initiation (**Figure 2**).

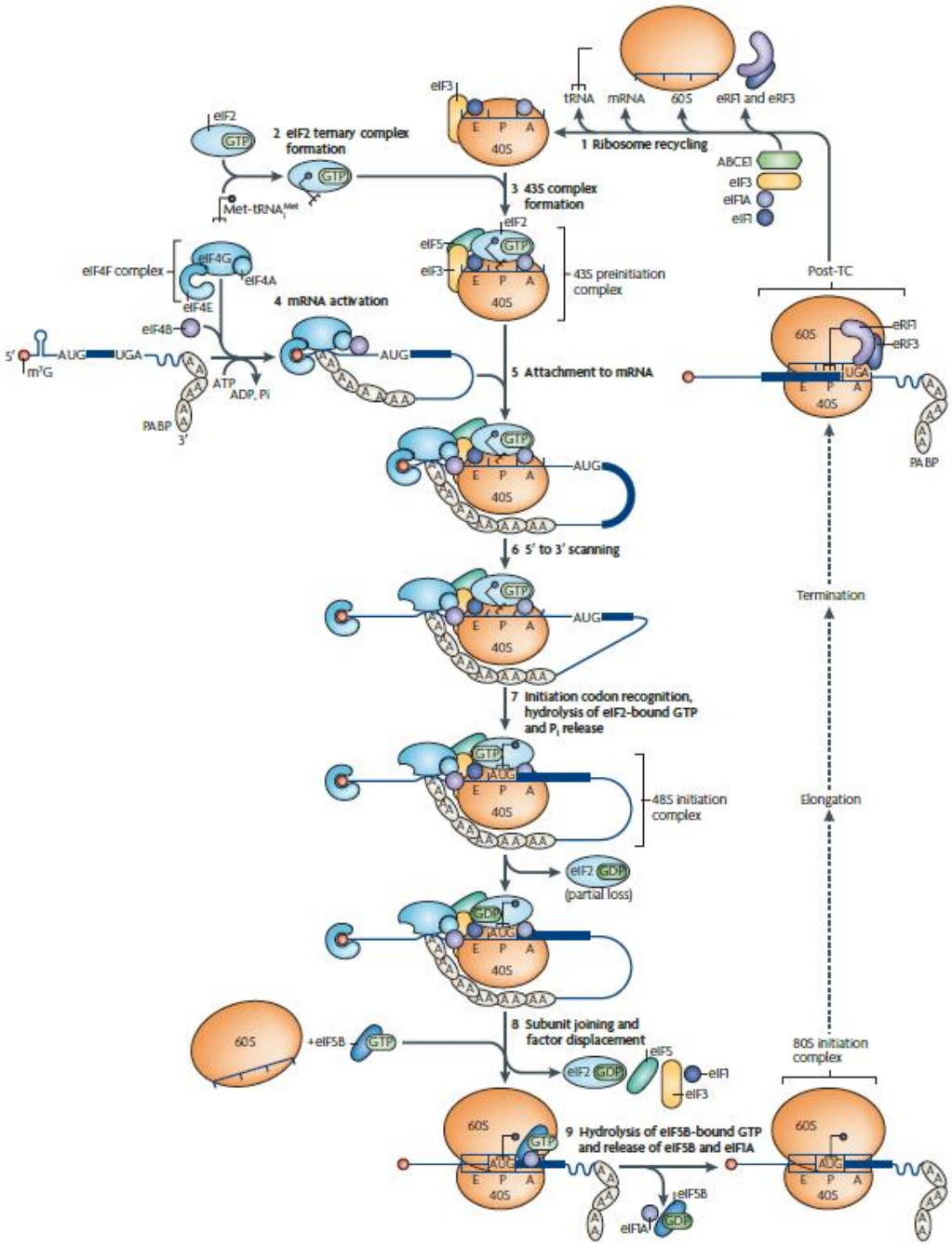


Figure 2 : ***Mise en place des différentes étapes d'initiation de la traduction****, formation du complexe ribosomique 80S puis élongation et enfin recyclage. L'initiation est présentée divisée en 9 étapes, brièvement : 1-recyclage des constituants après élongation et terminaison. 2/3-Formation du complexe eIF2 puis constitution du complexe 43S. 4- activation de l'ARNm par eIF4. 5/6-fixation du complexe 43S sur l'ARNm et scan 5' vers 3'. 7- reconnaissance du codon Start. 8/9-fixation de la sous unité 60S du ribosome sur la 40S formant le complexe 80S après libération de facteurs de traduction. (Jackson & al., 2009* [29])

a. Formation du complexe de pré-initiation 43S

La formation du ribosome et la traduction protéique sont des étapes cycliques dans lesquelles les éléments sont recyclés. Ainsi la formation du premier complexe est issue du recyclage de la petite sous unité ribosomique, le ribosome 40S, auquel s'ajoutent des facteurs d'initiation (eIF) : eIF1, eIF3 et eIF1A [30] et formant le complexe 43S. Celui-ci est ensuite complété par l'ajout de eIF5 et de eIF2 lui-même transportant l'ARNt-MET, permettant la reconnaissance du premier codon, le codon START et la fixation du premier acide aminé : la méthionine.

b. Préparation de l'ARNm

En parallèle à la formation du complexe 43S de pré-initiation, l'ARNm est associé à d'autres facteurs d'initiation : le complexe eIF4F (constitué d'eIF4A, eIF4E et eIF4G) et eIF4B. eIF4E reconnaît la coiffe en 5' de l'ARNm permettant au reste du complexe eIF4F de linéariser la séquence de l'ARNm [31]. Des protéines de fixation de la queue poly-A (PABP PolyA Binding Proteins) présentes à l'extrémité 3' s'apparient avec le complexe eIF4F circularisant l'ARNm.

c. Etape de scan

Le complexe 43S de pré-initiation peut alors se fixer à l'ARNm, débute l'étape de scan de l'extrémité 5' vers la 3' permettant de trouver le codon d'initiation (codon START) AUG. Cette étape est la plus critique puisqu'elle détermine le début de la séquence de la protéine. Elle est dirigée par la reconnaissance du contexte Kozak étant décrit comme la région la plus favorable à un début de traduction. eIF1 est le garant du maintien de la fidélité traductionnelle afin d'éviter les erreurs de traduction [29,32]. La reconnaissance du codon d'initiation permet alors la fixation de l'ARNt-MET associé à eIF2-GTP qui devient alors par hydrolyse, eIF2-GDP. La déphosphorylation d'eIF2 a pour effet la formation de la sous unité 48S du ribosome. Toutefois l'hydrolyse de eIF2 a pour conséquence de diminuer l'affinité de la protéine avec l'ARNt fixé sur l'ARNm, on observe alors une libération partielle de eIF2-GDP.

d. Regroupement des sous-unités du ribosome

La libération d'eIF2-GDP est poussée par la compétition avec eIF5B-GTP qui se fixe à sa place. La libération complète d'eIF2-GTP est réalisée lors de la fixation de la sous-unité 60S sur la sous-unité 40S, formant le ribosome 80S. Il y a également libération des facteurs d'initiation : eIF1, eIF3 et eIF5. L'hydrolyse d'eIF5B-GTP en eIF5B-GDP provoque sa libération ainsi que celle d'eIF1A, laissant alors le ribosome 80S fixé sur l'ARNm à la position du codon START.

Le ribosome 80S est capable de démarrer l'élongation pendant laquelle les ARNt transportant les acides aminés se fixeront un à un pour former la séquence de la protéine. Enfin à la détection du codon STOP « UGA » les facteurs eRF1 et eRF3 sont recrutés engendrant la séparation des sous-unités 40S et 60S, ainsi que la séparation avec l'ARNm. Les sous unités ainsi que les facteurs d'initiation peuvent alors être recyclés (**Figure 2**).

B. Régulation de la traduction ribosomique

La fixation des sous-unités ribosomiques et des facteurs de transcription ainsi que la reconnaissance du codon Start sont soumises à différentes régulations. Ces régulations ont pour effet de modifier la protéine traduite par un changement de cadre de lecture, voire d'inhiber totalement sa production. Ces régulations sont divisées en deux branches : La première impacte les constituants du ribosome tels que la modification de l'hydrolyse d'eIF2 entrainant une modification du site de fixation du premier acide aminé [29]. Cela a pour effet de modifier la protéine issue de l'ARNm. La phosphorylation d'eIF4E ou d'autres facteurs de traduction tels qu'eIF1, eIF3 et eIF5 peuvent modifier la traduction, ayant parfois des conséquences pathologiques. La deuxième branche de régulation est la modification de l'ARNm. Parmi ces modifications on observe la fixation de protéines, celles-ci fixées en 5'UTR empêchent la fixation du complexe 43S et donc le début de l'initiation. D'autres fixées en 3'UTR forment un ARNm circulaire attachant la protéine de la coiffe à la queue poly(A), empêchant toute fixation de ribosome [33]. Enfin il est possible d'observer la fixation de micro-ARN (miRNA) sur des régions de l'ARNm. Les miRNA sont de petites séquences nucléotidiques, pouvant se fixer sur l'ARNm par

complémentarité de séquence. Cette fixation a pour conséquence de rendre indisponible la région de fixation des facteurs de traduction [33].

2. Le dogme de la protéine unique

Bien que chez les eucaryotes les ARNm matures aient longtemps été considérés comme monocistroniques, c'est à dire conduisant à la traduction d'une seule et unique protéine, appelée protéine de référence, le protéome est cependant en réalité bien plus complexe.

La généralisation du dogme de la protéine unique a été conduite par la généralisation de travaux effectués sur l'origine des protéines et sur la manière dont on peut prédire leurs séquences. Une solution a été apportée par la mise en évidence du contexte Kozak, permettant de prédire les bases de données actuelles de manière fiable [7]. Néanmoins, aujourd'hui il apparaît nécessaire de remettre en question ce dogme et de se pencher sur les oublis survenus lors de la constitution de la majeure partie des bases de données et donc omises lors des analyses protéomiques.

A. Le contexte Kozak

Le cadre de lecture ouvert de référence (ou *reference Open Reading Frame*, RefORF) contenu dans la séquence codante (CDS) des ARNm matures conduit à la traduction d'une protéine dite de référence (RefProt). Cette région CDS est définie par la séquence nucléotidique la plus longue présentant un codon START (AUG) et un codon STOP (UAA, UGA ou UAG) tout en respectant le contexte Kozak. En effet en 1999 *Marilyn Kozak* rédige un article intitulé : « *initiation of translation in prokaryotes and eukaryotes* »[34], celui-ci a pour objectif de décrire de manière simple et abordable par des non spécialistes de la génomique, les principes de la traduction des ARNm en protéines. En s'appuyant sur son propre travail et celui de ses pairs, *M. Kozak* pose les bases du mécanisme nécessaire à la réalisation de la traduction des protéines. Parmi les généralités abordées, le mécanisme d'initiation de la traduction par la fixation de

la petite sous-unité du ribosome (40S) sur l'extrémité de la région non codante 5'UTR est décrit. Une fois fixée, la petite sous-unité est rejointe par des facteurs de transcription tels que eIF2, et migre le long de la région 5'UTR jusqu'à rencontrer un codon START « AUG ». La sous-unité 40S arrêtée sur un codon Start est ensuite rejointe par la sous-unité 60S. Si cet assemblage est permis, alors le codon fixé devient le site d'initiation de la traduction protéique. Dans ce contexte, et de manière expérimentale, il a été montré que le premier site « AUG » rencontré est naturellement le site majoritaire du début de la traduction. Ces régions sont par ailleurs soumises à d'importantes régulations, la présence de boucles en épingle sur la région 5'UTR, ou de protéines de régulations réduit l'accessibilité des codons START de l'ARNm et de ce fait entraîne une modification du site de traduction de la protéine. D'autres facteurs tels que les régions riches en G+C sont décrits comme étant des promoteurs internes pour la traduction, entraînant un ralentissement du scanning de la séquence ARNm. Toutefois, le facteur le plus connu reste le « contexte Kozak », celui-ci est décrit comme étant une séquence entourant de part et d'autre le premier codon AUG et permettant l'arrêt du scanning par la sous-unité 40S. Ce contexte est majoritairement de type « GCCACC-AUG-G », avec un important taux de conservation retrouvé pour la position -3 représentée par une purine, le plus souvent une adénine. La position en +4 de la guanine est également décrite comme dirigeant la traduction.

Cependant, *Kozak* précise que la régulation de la transcription des ARNm est spécifique de chaque ARNm, nécessitant donc de rester critique sur l'utilisation du codon START majoritaire pour prédire la protéine exprimée par un ARNm. *Kozak* consacre ainsi une partie entière de son article à la manière dont le mécanisme de traduction est capable de contourner la règle du premier AUG. Celui-ci peut notamment être déjoué par les mécanismes de « *reinitiation* » et de « *leaky scanning* » (**Figure 3**).

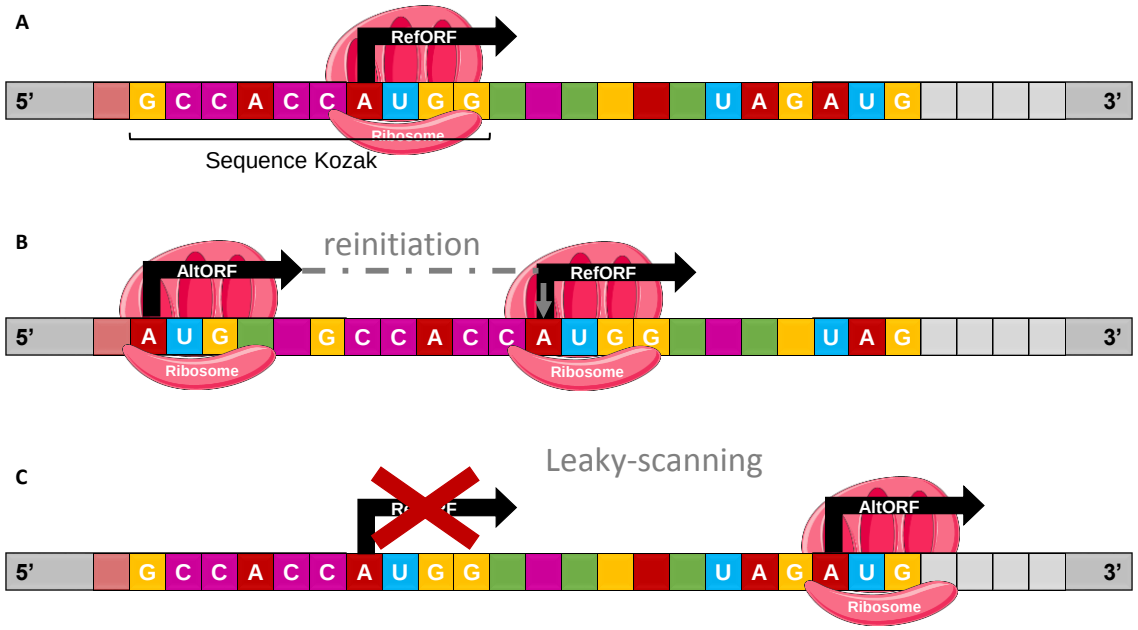


Figure 3 : ***Représentation de la théorie de contexte Kozak*** *A. Dans le cas de l'expression de RefProts suivant le cas le plus courant à partir du premier codon AUG et encadré par la séquence de Kozak B. Le principe de reinitiation, ou le ribosome est recruté sur un second AUG après avoir traduit la RefProt C. Le principe de leaky scanning ou le premier codon AUG est temporairement mais complétement remplacé par la fixation sur l'AUG suivant.*

e. Le mécanisme de « *reinitiation* »

Le mécanisme de « *reinitiation* » (**Figure 3B**) est dû à la formation du ribosome 80S en amont du site AUG le plus probable. Le ribosome commence alors une traduction donnant naissance à un peptide décrit comme une protéine non complète. Cette petite région, autrefois nommée *upstream* ORF (upORF) est décrite comme ne dépassant pas 30 codons [35]. En effet, si sa longueur est plus élevée (plus de 120 codons) elle ne permet plus l'enchaînement avec l'ORF principal et donc la « *reinitiation* ». La région upORF permet la régulation de la transcription de la protéine en aval, toutefois le rôle de la petite protéine produite est encore très mal compris, et est généralement attribué à un peptide non fonctionnel. Cependant, certaines études notamment chez *Saccharomyces cerevisiae* montrent la présence de multiples upORF présents dans la région 5'UTR, indispensables à l'expression de l'ORF de référence car traduisant un facteur de transcription pour d'autres gènes [36].

f. Le mécanisme de « *Leaky-scanning* »

Contrairement au mécanisme de « *reinitiation* », le « *leaky-scanning* » (**Figure 3C**) permet d'outre passer la règle du premier codon « AUG ». C'est alors le second ou le troisième codon *Start* qui sera utilisé. Le mécanisme serait causé par l'absence d'un contexte optimal d'initiation. En effet, si le contexte Kozak autour du codon AUG n'est pas assez fort, la sous unité 40S ne s'arrête pas sur le codon *Start* attendu et ne traduit que la région suivante. Grâce au « *Leaky-scanning* », les ARNm eucaryotes sont donc capables d'exprimer plusieurs protéines différentes à partir d'une séquence unique. Les protéines plus ou moins longues proviennent de régions complètes et chevauchantes dans le CDS de la protéine de référence. Toutes les deux sont alors décrites comme fonctionnelles.

L'annotation des banques de données majoritairement utilisées reste basée sur l'ancien dogme reposant sur les règles dictées par *Kozak*, applicables pour la majorité des ARNm. En effet, les autres mécanismes de traduction mis en évidence n'étaient pas considérés comme pouvant conduire à la traduction d'une protéine fonctionnelle. Ainsi, les banques de données utilisées pour la protéomique sont restées basées sur la traduction d'une seule protéine par ARNm.

B. Les ARN non codants

Le dogme actuel en biologie veut qu'une séquence ADN exprime un ou plusieurs transcrits ARN qui permettent pour certains de traduire une protéine. Ainsi, les transcrits ARN ne conduisant pas à une traduction en protéines sont définis comme non codants (ARNnc). Toutefois il est démontré depuis plusieurs années que les ARN possèdent des fonctions plus étendues que la simple production de protéines. En effet, basé sur l'existence d'ARN non codants (ARNnc) et parfois même longs non codants (ARNlnc) lorsque faisant plus de 200 nucléotides, des fonctions alternatives à la production de protéines ont été démontrées [37]. Les ARNnc peuvent également être divisés en plusieurs groupes : ceux issus des introns, des régions intergéniques, bidirectionnels, anti-sens ou chevauchants les exons codants un ARNm. Parmi les fonctions

alternatives des ARN, on note la régulation épigénétique par modulation de l'accessibilité à la chromatine par la polymérase, et la régulation de la méthylation des histones [38]. La régulation transcriptomique est une autre de ces fonctions qui impacte la production de protéines. Il est possible de citer à titre d'exemple l'inhibition de l'expression de la Cycline D1 par modulation de l'activité acétyle transférase des protéines CREB et p300. D'autres fonctions ont été mises en évidence telles que l'assemblage d'un ARNlnc avec la protéine TLS [39] en tant que cofacteur modulant les facteurs de transcription et donc l'expression de protéines [40] ou en influençant le choix du site de fixation de l'ARN polymérase II [41]. Les ARNlnc sont également décrits comme étant capables de régulation post-transcriptionnelle, comme dans le cas de ZEB2. La fixation d'un ARNlnc sur la région 5'UTR provoque une modification de l'épissage et la conservation d'un site de fixation pour le ribosome permettant la production de ZEB2 [42]. Ces fonctions alternatives des ARNlnc peuvent être associées à un contexte pathologique comme l'ARNlnc imatinib-upregulated (ARNlnc-IUR) récemment décrit comme un suppresseur tumoral par inhibition de la voie de signalisation STAT5-CD71 (**Figure 4**) [41].

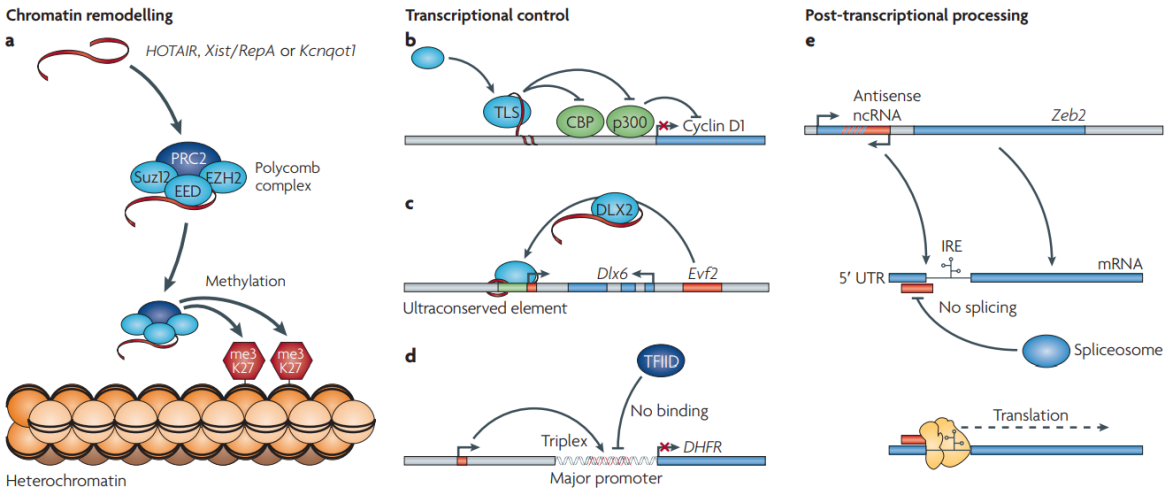


Figure 4 : ***Description de la fonction des ARN non codants****, a. dans le modelage de la chromatine et la régulation épigénétique, b. dans le contrôle de la transcription et de l'expression des gènes, c. dans la régulation de la traduction. (Mercer & al., 2009* [37]*)*

Toutefois, si ces fonctions sont aujourd'hui avérées et démontrées, les règles de traduction appliquées lors des prédictions protéiques pour les ARNlnc, souffrent

des mêmes limitations que celles pour les ARNm. De ce fait, la mise en évidence de smORF dans les régions non codantes des ARNm a également soulevé la question de la présence de smORF dans les ARNlnc. Les analyses de suivi de l'empreinte ribosomique ou « *ribosome profiling* » ont ainsi conduit à démontrer que 90% des ARNlnc possèdent des sites de traduction pouvant produire une protéine [43].

Ainsi tout comme pour les ARNm, les ARNlnc présentent des fonctions alternatives de régulation transcriptomique, post-transcriptomique et protéomique, mais également pour certains d'entre eux sont la source de production d'AltProts [44].

C. Constitution des banques de données protéiques

Le séquençage à haut débit du génome et du transcriptome a montré que le nombre d'ORFs prédit est largement sous-estimé. Ceci a été montré par la mise en évidence d'ORFs permettant la traduction de protéines de moins de 100 acides aminés dans les régions non codantes des ARNm ou dans la région CDS suivant un cadre de lecture +2 ou +3. Il a également été démontré que le nombre de protéines traduites est bien supérieur aux prédictions actuelles [16,23,45,46].

Cependant, les bases de données actuellement disponibles en ligne, et utilisées dans la majorité des études protéomiques sont basées sur les règles généralistes décrites par Kozak. C'est ainsi qu'UniProtKB/Swiss-Prot [47], une des bases de données les plus répandues, suit des règles basées sur le fait qu'un ARNm ne code que pour une seule et unique protéine, concept soutenu par l'idée qu'une séquence canonique unique correspond à la séquence la plus représentée.

La constitution de la base de données est complétée par la prédiction de séquences protéiques encore non observées de manière expérimentale grâce à l'algorithme de prédiction TrEMBL [48]. Cela permet de compléter et de découvrir de nouvelles protéines dans les voies de signalisation et dans les mécanismes actuellement décrits. Ces bases de données prennent également en compte la formation d'isoformes. Toutefois UniProtKB/Swiss-Prot reste incomplète et

nécessite du temps et des ressources afin d'être enrichies et complétées au fur et à mesure des découvertes.

UniProt nous permet cependant de poser la définition de la protéine de référence ou RefProt comme étant la protéine issue du « Gold Standard CDS » décrit par le Consensus CDS (CCDS) et ayant l'alignement le plus important avec la traduction du génome de référence [7]. Malgré cette définition, il est connu que 5% des protéines du CCDS ne sont pas encore incluses dans les bases de données UniProt consolidées, car elles représentent des protéines issues de cas complexes ou de nouvelles séquences n'ayant pas encore fait l'objet d'une observation expérimentale. Toutefois, UniProt reconnaît l'existence des petits ORF dans des régions actuellement décrites comme non codantes telles que les régions 5' et 3' UTR, nommées smORF. Cependant, la base de données intègre uniquement les protéines issues de ces régions après une validation expérimentale stricte.

La mise en retrait des prédictions de séquences protéiques dans les bases de données standards, telle qu'UniProt ne permet pas la mise en évidence de ces séquences qui existent pourtant bien dans nos cellules. Ainsi la mise en évidence des protéines issues de régions décrites comme « non codantes » nécessite la réalisation de bases de données à façon.

3. Mise en évidence des AltProts

A. Les transcrits non codants et leurs produits protéiques

En 2010, l'équipe de *X. Roucou*, qui travaille alors sur la protéine Prion (PrP), met en évidence la capacité du gène PRNP à exprimer une protéine en +3 de la région CDS et décrite comme possédant un contexte Kozak optimal pour la traduction d'une protéine alors nommée AltPrP, présentant entre 64 et 81 acides aminés [4]. Afin de démontrer la présence de cette protéine dans différentes lignées cellulaires, des études par PCR, modification et suivi par TAG ont été réalisées. Il a pu être montré que le gène PRNP était capable d'exprimer deux protéines comportant deux séquences complètement différentes et ayant des

localisations spécifiques, AltPrP étant majoritairement mitochondriale tandis que PrP étant observée au niveau de la membrane. Concernant leur fonction, encore très peu de choses sont alors décrites. Cependant, une interrogation importante est alors soulevée: les fonctions attribuées à PrP dans la littérature pourraient alors être partagées avec AltPrP. Ainsi lors d'une inhibition du gène codant pour PrP il peut y avoir inhibition de la RefProt mais aussi de l'AltProt entraînant des modifications qui ne sont pas spécifiques de PrP [4]. En parallèle, le laboratoire PRISM s'interroge sur les nombreuses données générées par LC-MS lors d'analyses shotgun et ne trouvant pas d'identification dans les bases de données. En effet, une partie des données reste sans identification bien que les données structurales acquises (spectres MS/MS) présentent une qualité suffisante (nombre de fragments caractéristiques) pour aboutir à l'identification d'un peptide associé à une protéine. Les discussions entre les deux équipes conduisent à l'application d'une base de données créée par l'équipe de X. Roucou, et recensant l'ensemble des AltProts non prises en compte dans les bases de données disponibles, aux données de protéomique acquises par PRISM en 2011. Ces études permettent de démontrer que les données de protéomique de qualité restant sans identification, après interrogation dans Uniprot, trouvent une identification dans la banque de donnée d'AltProt prédites. En outre, les expériences permettent de démontrer que les AltProts sont un mécanisme conservé à l'ensemble des eucaryotes et que des AltProts spécifiques sont retrouvées aussi bien dans un contexte physiologique que pathologique. Enfin, l'étude des données protéomiques sur le cancer de l'ovaire conduit à identifier notamment une protéine alternative à un interactant de BRCA1, qui présente une localisation identique à BRCA1 et semble également pouvoir interagir avec cette dernière. Ces études constituent la première démonstration de l'existence de protéines alternatives à grande échelle.

En 2013, d'autres études réalisées par *S. A. Slavoff* et coll. [45] confirment l'observation de protéines alternatives via la mise en évidence des « sORF-Encoded Polypeptides » (SEPs). En effet, si Kozak emploie le terme de smORF (*small ORF*) afin d'évoquer les régions exprimant de petites séquences de l'ARNm possédant des codons *Start* et *Stop*, mais ne correspondant pas à la

séquence de la RefProt, on retrouve également le terme de sORF (*short ORF*) dans la littérature. Ces sORFs sont décrits comme permettant la production de protéines alors appelées « sORF-Encoded Polypeptides ». En effet, partant du constat de l'existence de SEPs dans des espèces telles que les bactéries, les virus ou encore les plantes, et appuyé par des expériences démontrant la fonction de ces polypeptides dans ces espèces, il a été proposé que les bases de données utilisées en protéomique sous-estimaient grandement l'importance et la fonction de ces SEPs. Afin de répondre à la question d'existence des SEPs dans les cellules humaines K562 (leucémie humaine), une stratégie a été mise en place permettant dans un premier temps d'enrichir les protéines/polypeptides de faible poids moléculaire afin d'augmenter leur détection en chromatographie couplée à la MS (LC-MS). Ensuite, afin d'identifier ce qui n'est pas annoté dans les bases de données de protéomique disponibles, une base de données à façon est réalisée par le séquençage d'ARN (RNA-seq) à large échelle sur le modèle étudié. Cette base de données unique à l'échantillon analysé a conduit à la mise en évidence de 90 SEPs [45] dont 86 qui n'avaient jamais été décrits et 4 qui précédemment découverts par *M. Oyama* en 2004 [49].

En 2014, l'équipe de *N. Ingolia* a observé que le ribosome avait la capacité de se fixer sur des régions non codantes. Par une méthode de *ribosome profiling* combinée à la prédiction des sites de fixation du ribosome sur l'ARN, il est alors possible de définir un score de similarité entre la prédiction et l'observation, le score « *Fragment Length organization similarity score* » (FLOSS) [43]. La mesure de ce score donne alors un poids supplémentaire à la probabilité de produire des protéines dans les régions, 5', 3'UTR et provenant des ARNlnc. Il est également décrit une différence importante entre le démarrage de traduction en position +1 classique et les positions : +2 et +3 dans l'ORF. La position +1 étant la plus représentée cela explique que l'on en ait fait la position la plus favorable à la traduction protéique et à la prédiction de régions traduites. Cependant, cette différence est moins importante sur les régions 5' et 3' UTR où une expression importante de la position +2 dans les régions 3'UTR des ARNm codants est observée. Cela est à corréler avec l'absence de contexte Kozak dans

ces régions, et suppose une régulation ainsi que des mécanismes de traduction différents.

B. Développement des bases de données AltORF/AltProt

a. HaltORF

La base de données « *Human alternative Open Reading Frame »* HaltORF, est décrite comme la première base de données, large échelle et en ligne permettant la recherche d'AltProts chez l'homme.

La constitution de la banque s'est faite en plusieurs étapes. La première fut le téléchargement d'un maximum d'ARNm connus. Ces ARNm sont ensuite associés à leur RefProt pour ensuite rechercher *in silico* les protéines prédites à partir des régions CDS en +2 et +3(+1 correspondant au cadre de lecture de référence). Toutefois afin de garder une taille de données raisonnable une limitation aux protéines de plus de 24 acides aminés est réalisée permettant également de faciliter la confirmation des protéines observées par les méthodes protéomiques classiques. La deuxième étape est l'annotation des AltProts prédites en prenant en compte :

- Leur localisation sur l'ARN
- La présence d'un contexte Kozak fort ou faible
- Leur présence dans les bases de données classiques ou non

De cette manière, 17 096 séquences d'AltProts ont été prédites à partir de 31 422 ARNm soit 8 744 gènes, avec une répartition majoritairement en CDS+2 pour 83% et en CDS+3 pour les 17% restants [5].

Toutefois, cette base de données omet encore beaucoup de possibilités, puisqu'elle ne prenait alors pas en compte ni les traductions dans les régions 5' et 3' UTR, ni les codons START alternatifs ou les ARNnc. La collaboration en 2010 avec le laboratoire PRISM Inserm U1192, qui développait des méthodes d'analyses protéomiques par MS, a permis l'utilisation de bases de données très importantes. De ce fait une base de données plus exhaustive cumulant un maximum d'informations a pu être réalisée à partir des bases transcriptomiques NCBI RefSeq [50] et Ensembl [9] en conservant les régions non codantes telles que les parties 5', 3' UTR formant ainsi 3 groupes de AltORF notés AltORF$^{5'UTR}$,

AltORF^{CDS} [3,46] et AltORF$^{3'UTR}$ mais également ceux présents dans les ARNnc (**Figure 5**). Les protéines de plus faible poids moléculaire avec des tailles inférieures à 24 acides aminés sont également conservées dans cette nouvelle version de la base de données.

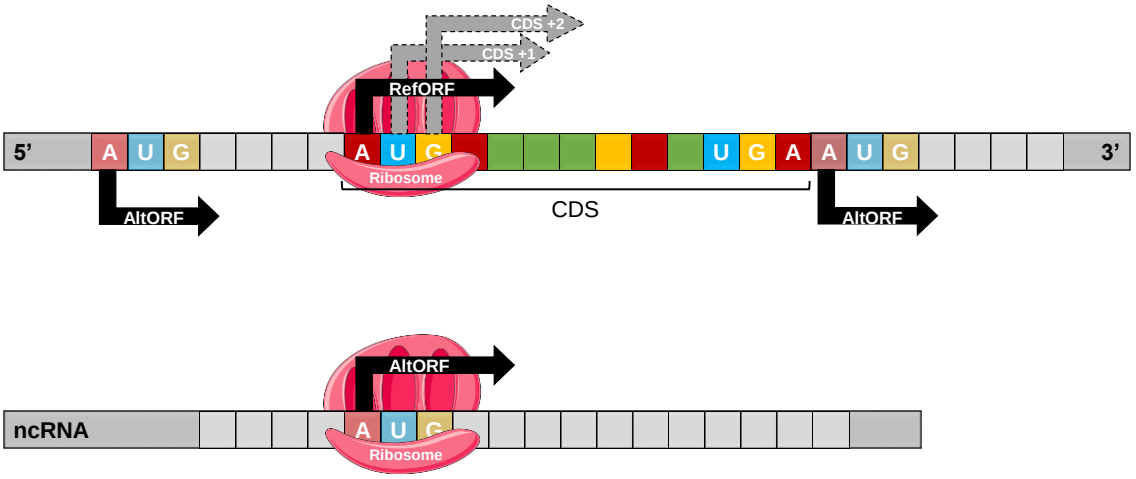


Figure 5 : ***Schéma de l'expression d'AltProt à partir d'un ARNm ou d'un ARNnc****. Région refCDS code pour la RefProt tandis que toute autre région code pour une AltProt. On retrouve des sites d'expression en 5'UTR, 3'UTR ou en CDS+1 et +2. Les AltProts sont également originaires des ARNnc, ne possédant pas de RefProt.*

L'utilisation de cette base de données et des méthodes MS a permis de caractériser les AltProt [23]. On sait ainsi aujourd'hui qu'elles sont en moyenne de longueur de séquence plus courtes que les RefProt issues du cadre de référence, avec une moyenne de 57 acides aminés contre 344 pour une protéine issue d'un cadre de lecture canonique [3]. 54% sont retrouvées dans la région 3'UTR, dont 7% chevauchant le CDS ; 5% sont issues de la région 5'UTR dont 2% chevauchant le CDS et enfin les décalages de cadre de lecture +2 et +3 dans le CDS représentent 41% des AltProts prédites. Cette répartition est confirmée par les identifications de AltProt obtenues dans les cellules HeLa [3].

b. smPROT

En 2017 et sur la base des précédents travaux de *N. Ingolia de 2014* [43], l'équipe de *Hao* s'interroge également sur la pauvreté des banques de données protéomiques et plus particulièrement pour les petites protéines [11]. Comme décrit précédemment, la majeure partie de ces bases de données ne considère

pas les protéines de moins de 100 acides aminés. De plus, la communauté protéomique tend à négliger le rôle et la fonction de ces petites protéines. Ces petites protéines même avec un nombre restreint d'acides aminés présentent une activité biologique dans les cellules. Afin de répondre aux lacunes observées dans les bases de données disponibles telles qu'UniProt et CCDS, il a été proposé la création d'une base de données combinant protéines prédites et protéines observées expérimentalement. Cette base de données est réalisée à partir de 291 types cellulaires et tissus provenant de huit espèces différentes dont l'Homme, la souris, le rat et le poisson-zèbre. Cette banque repose notamment sur des données bibliographiques, des analyses en *ribosome profiling,* mais aussi sur des analyses obtenues en MS issues d'un mélange entre leur propre analyse, celle de EMBL-EI et de ENCODE. La combinaison de ces informations permet de réaliser la première base de données des petites protéines : smPROT [11]. Les outils permettant l'analyse des séquences en acides aminés des RefProts appliqués à ces petites protéines ont permis de réunir les premières informations : carte génomique, code d'identification, origine des informations, mais surtout les premières prédictions de fonctions. Cette base regroupe un total de 255010 smProt dont 167785 smProt humaines.

c. Openprot

Afin de rendre accessibles les bases de données élargies, OpenProt (https://openprot.org/) a été mis en ligne en 2019. Cette base de données reprend l'ensemble des prédictions obtenues pour les AltProts dans l'ensemble des régions 5' et 3' UTR, mais aussi les ARNnc, à partir de différentes bases de données de transcrit : RefSeq et Ensembl. Cette base de données fait également la différence entre les AltProts observées comme ayant une homologie de séquence forte avec les RefProts, décrivant une putative nouvelle isoforme et non une AltProt. De cette manière plus de 450.000 nouvelles AltProts ont pu être prédites chez l'homme. OpenProt prédit également les AltProts chez plusieurs espèces eucaryotes telles que l'Homme, la souris, le rat et le poisson zèbre pour en citer quelques unes [51].

Toutefois, cette base de données conserve une limitation puisqu'elle ne prédit que les protéines issues d'un minimum de 30 codons, donc aucune protéine inférieure à 30 acides aminés. Les sites de traduction ne commençant pas par un codon Start AUG ne sont pas non plus pris en compte dans cette base de données, contrairement aux bases de données issues des stratégies de protéogénomique et appliquant une expérience de « *ribosome profilling* » telles qu'appliquées par *S. Slavoff* et all.

d. La stratégie protéogénomique

La stratégie que nous avons choisie est une stratégie globale passant par la constitution et l'utilisation d'une base de données totale pour une espèce issue des prédictions des régions aujourd'hui décrites comme non codantes. Cependant une autre approche existe, la protéogénomique, qui est une approche spécifique à l'échantillon. En protéogénomique l'objectif est de réaliser une base de données protéique, à partir des données de transcriptomique et de génomique. Ainsi les expériences de séquençage ARN, de *ribosome profilling*, de prédiction de traduction du génome, de séquençage d'ADN et d'expression de séquences taguées (expressed sequence tags : ESTs) réalisées sur un même échantillon permettent de définir une base de données de protéines propres à l'échantillon. Parmi ces prédictions des protéines peuvent alors être observées dans des régions non codantes ou des ARNlnc, on identifie alors la présence de nouvelles protéines qui sont des AltProts [52,53]. Cette stratégie peut être mise en place dans des études multiomics, où une étude génomique et une étude transcriptomique sont réalisées préalablement à l'étude protéomique. Ces études permettent de mettre en évidence des protéines spécifiques aux échantillons et dans le cas de pathologies d'identifier de nouveaux biomarqueurs [54].

Toutefois si cette méthode permet la découverte de nouvelles protéines par la mise en place de bases de données très spécifiques aux échantillons, elle nécessite des moyens techniques importants et très spécifiques. Ainsi les études de protéogénomique se réalisent souvent au travers de consortia tel que le consortium « *chromosome-centric Human Proteome Project (cHPP)* » porté dans le cadre du projet de protéome humain [55].

C. Détection des AltProts par MS

a. Stratégie shotgun

L'utilisation de banques de données spécifiques à un échantillon (étude de protéogénomique) ou plus larges (telle que HaltORF) incluant la prédiction de nouvelles protéines permet l'application des stratégies d'analyses de protéomique à grande échelle. Actuellement, la stratégie la plus répandue reste l'analyse par approche bottom-up de type « *shotgun* ». Dans cette approche, les protéines sont digérées après extraction puis les peptides de digestion séparés et analysés par LC-MS. Le traitement des données consiste alors à reconstituer les peptides grâce aux séquences en acides aminés obtenues par MS/MS, puis à retrouver la protéine associée à partir des peptides identifiés et par homologie entre les protéines présentes dans la base de données et les peptides identifiés qui les constituent [56]. Les AltProts étant en moyenne plus petites que les RefProts les stratégies shotgun sont clairement plus adaptées que les statégies bottom-up classiques passant par une séparation des protéines sur gel d'électrophorèse bidimensionnelle préalablement à leur digestion après excision des spots. Ainsi, cette méthode a été la première mise en place pour l'analyse des AltProts par le laboratoire [3]. Le développement de cette méthode permet même aujourd'hui d'identifier jusqu'à 10 000 protéines avec un gradient LC de 100min [1]. Cette stratégie a permis de mettre en évidence de nombreuses AltProt [23].

Toutefois, appliquée aux AltProts cette méthode souffre d'un désavantage certain. En effet, elle nécessite une digestion optimale des protéines en peptides pour identifier la protéine. Cependant, les AltProts, par leur petite taille produisent souvent moins de 3 peptides de digestion. Ceci force l'identification à partir de peptides de moins de 6 acides aminés et souvent avec un minimum de 1 peptide unique par protéine. Cela peut facilement représenter une couverture de séquence de 30 à 40% de l'AltProt. Cette limitation oblige pour le moment à utiliser des seuils de traitements statistiques bas lors de l'analyse des données issues de la LC-MS, qui remettent souvent en question la réelle présence de l'AltProt, possiblement considérable comme un faux positif.

b. Stratégie Top-Down

La méthode « Top-Down » représente la stratégie inverse du « Bottom-Up ». En effet, en Top-Down les protéines ne sont pas digérées mais identifiées directement en MS après séparation sous leur forme intacte. La stratégie Top-Down est-elle, plus difficile à mettre en œuvre car elle nécessite l'utilisation de spectromètres de très hautes performances possédant une grande sensibilité, un fort pouvoir résolutif, une haute résolution spectrale ainsi qu'une grande précision de mesure [57]. Elle nécessite également l'utilisation de méthodes d'activation des ions particulières permettant une fragmentation efficace des protéines intactes telles que l'HCD, l'ETD, l'EThCD ou l'UVPD [58]. À contrario ces stratégies présentent un certain nombre d'avantages parmi lesquels la possibilité de mesurer le poids moléculaire des protéines intactes avec une grande précision et la possibilité de remonter à l'identification de séquences en acides aminés plus longues. Ces avantages permettent d'une part d'accéder plus facilement à l'identification des modifications post-traductionnelles et d'autre part de mettre en évidence les différentes protéoformes des protéines i.e. les isoformes, les protéines tronquées, les mutations de séquences [59]… Dans ce contexte, la stratégie Top-Down est une approche intéressante pour les AltProts qui sont de petite taille et peuvent présenter un chevauchement de séquence partiel avec des RefProts.

Ainsi, les stratégies Top-Down peuvent être appliquées à l'étude des AltProts dans le cadre d'études de protéogénomiques [60].En 2017 et 2018, il a également été démontré par Delcourt et al [14,15] que la stratégie Top-Down réalisée de façon localisée sur les tissus en mode dit « Spatially-Resolved » permettait l'identification de plusieurs centaines de protéoformes dont des AltProts. Ces stratégies dites « Spatially-Resolved » permettent d'extraire des protéines d'une région restreinte d'une section mince de tissu (20 µm d'épaisseur) à l'échelle du diamètre du cône d'une pipette et ainsi d'identifier des protéines spécifiques à des régions particulières d'un tissu (tumorale, saine, marge). Ces régions d'intérêts peuvent-être déduites des données d'imagerie par spectrométrie de masse (MSI), technique qui permet de localiser un ou un ensemble de composés à la surface d'une section mince de tissu. Cette

combinaison de détection des régions d'intérêts par MSI, puis extraction de ces régions par micro extraction (ou excision du tissu déposé sur un parafilm) couplée à une analyse Top-Down a conduit à l'identification d'AltProts dans différentes régions de coupes de cerveau de rat [14] et dans des coupes de cancer de l'ovaire [15].

c. Enrichissement des petites protéines

Les AltProts étant des protéines de faible poids moléculaire, il est important de pouvoir les enrichir afin de garantir une bonne efficacité d'identification. Il a été montré que l'enrichissement en protéines de petites tailles permettait effectivement d'augmenter grandement le nombre de AltProt identifiées à partir de cellules K562 (leucémie myéloïde) et de cellules A549 (carcinome de poumon humain) [61]. Les auteurs ont utilisé différentes stratégies allant de l'extraction sur phase solide (SPE) C8 et C18, de la précipitation en milieu acide et de l'utilisation de membranes avec une limite à 30 KDa. Les résultats ont montré que l'enrichissement était plus efficace avec les colonnes SPE C18 et une précipitation à l'acide acétique, pour des approches habituellement utilisées pour les peptides. D'autres techniques d'enrichissement ont été étudiées telles que l'enrichissement des protéines de faible poids moléculaire sur gel d'acrylamide [62]. En effet, D'Lima *et al.* [63] ont utilisé cette stratégie pour isoler deux AltProts (YmcF et YnfQ), traduites à partir de codons START non canoniques ATT à partir des gènes *ymcF* et *ynfQ*, sous-jacents aux gènes cold shock (*cspG* et *cspI*) [63].Les AltProt de par leur petites tailles, sont proches des polypeptides dont les plus connus sont les neuropeptides et les hormones. Lors de l'étude des neuropeptides des optimisations d'extraction ont été réalisées notamment afin d'augmenter le taux d'identification de ces composés [64]. Les résultats montrent que l'extraction à partir d'eau bouillante et de précipitation est bien adaptée à ces peptides. Notre objectif a donc été d'adapter ces méthodes à l'extraction et l'identification des AltProts dans nos analyses, cela a donné lieu au premier papier présenté dans cette thèse.

Les neuropeptides jouent des rôles physiologiques variés tels que la communication cellule-cellule, régulation de l'anxiété, dépression, mémoire, et sensation de douleur[65]. Si ces neuropeptides sont présentés comme un produit

issu de peptidases et de dégradation de protéines, les AltProts pourraient avoir un rôle et une régulation similaires, mais en étant un produit de transcrits ARN. Ainsi elles pourraient être un nouveau type de biomarqueurs dans des pathologies, mais également de nouvelles cibles thérapeutiques [66].

D. Implication pathologique et fonction des AltProts

Les études associées aux AltProts ne cessent de croître tant dans des conditions physiologiques normales que pathologiques. Par exemple dans le cadre du cancer de l'ovaire, dans une étude réalisée par approche Top-Down en mode « Spatially-Resolved » via extraction par jonction liquide de surface (LESA) [67] et dissection assistée par Parafilm (PAM) [68] sur coupes de tissus après identification de régions d'intérêt grâce à l'imagerie par MS (MSI) ; 15 AltProts ont été mises en évidence dont 4 provenant de la région tumorale dans le cancer de l'ovaire : AltCMBL, AltGNL1, AltRP11-576E20.1, AltCSNK1A1L [15]. De manière intéressante, GNL1 est décrit comme une protéine exprimée dans le cancer de l'ovaire, ainsi l'expression d'une AltProt issue du même transcrit ARNm mature n'est pas incohérente. L'étude montre de plus que la RefProt GNL1 et son AltProt AltGNL1 ne sont pas localisées dans les mêmes régions. AltGNL1 est retrouvée principalement dans le noyau des cellules, tandis que GNL1 est décrite et observée dans le cytoplasme [15]. Cela laisse supposer que probablement l'AltProt et la RefProt issues du même ARNm mature peuvent avoir des rôles différents dans la cellule y compris dans un contexte pathologique. Toutefois il ne faut pas oublier que pour établir la localisation d'AltGNL1 la protéine a été surexprimée avec un tag permettant son identification dans la cellule. Ce marquage peut entraîner une modification de la localisation notamment sur une protéine de faible poids moléculaire telle que les AltProts.

Une autre AltProt, AltMRVI1, a été mise en évidence dans le cancer de l'ovaire lors d'analyses protéomiques shotgun standards. Celle-ci est traduite à partir du gène *MRVI1* qui présente une séquence homologue à la protéine BRCA1 IP3. BRCA1 IP3 est un partenaire d'interaction de la protéine BRCA1 connue pour son implication dans le cancer du sein et de l'ovaire. Le gène *BRCA1* appartient à la famille des gènes suppresseurs de tumeurs et la protéine associée BRCA1 est directement impliquée dans la réparation de l'ADN [69]. Si

le gène *BRCA1* (ou *BRCA2*) est muté, alors la protéine traduite associée n'est plus active ce qui conduit à la prolifération des cellules anormales et par conséquent à la progression de la tumeur. BRCA1 et AltMRVI1 se co-localisent dans le noyau. Leur interaction a été démontrée par des expériences de coIP et de western blot. Ces résultats suggèrent que AltMRVI1 pourrait être impliquée dans le même réseau moléculaire que BRCA1 montrant ainsi une possible implication dans le cancer de l'ovaire [3]. Un résultat similaire est obtenu pour la protéine AtlAKT2, présente en 5' UTR de la protéine de référence AKT2, qui est connue comme une protéine présente dans le cancer de l'ovaire [70]. En effet, AltAKT2 semble réguler la phosphorylation de la protéine AKT2 (Delcourt, données non publiées).

Si ces études n'apportent que des indices préliminaires quant aux rôles tenus par ces AltProts dans la pathologie du cancer de l'ovaire, elles permettent de mettre en évidence de manière très spécifique le type de tissu dans lequel ces AltProts sont retrouvées, dont certaines semblent être liées à des gènes connus comme étant impliqués dans la pathologie.

On peut souligner le rôle de ces AltProts dans la transition épithéliale à mésenchymale (EMT) mais aussi dans la transition mésenchymale à épithéliale (MET). Parmi les acteurs connus dans ces changements EMT-MET, le rôle des facteurs de transcription (EMT-TFs) est crucial et le rôle de SNAI1&2, TWIST1&2 est avéré. On retrouve également l'implication des protéines à doigts de zinc fixant les régions E-Box 1 & 2 (ZEB1&2) [71]. Deux AltProts en +2 et +3 du CDS de ZEB1 ont été observées en MS/MS [3], et depuis peu enregistrées dans la base de données UniProtKB respectivement sous les identifiants : L0R5E9_HUMAN, and L8EAF3_HUMAN. Même si le rôle de ces deux AltProts n'est pas actuellement connu, elles semblent participer à des changements impliqués lors des EMT et MET, et par conséquent dans les mécanismes associés tels que la formation de métastases.

L'AltProt, désignée comme le polypeptide Humanin (HN), est issue de la recherche d'un suppresseur des gènes impliqués dans la mort cellulaire des neurones résultant de la pathologie d'Alzheimer, les FAD telle que la protéine précurseur de l'Amyloïde (APP) et la Presenilin 1 et 2 (PS1&2). Lors de la

recherche d'un suppresseur de ces gènes, un motif polypeptidique de 24 acides aminés a été mis en évidence dans les cellules résistantes à la mort cellulaire. Cette AltProt pourrait être impliquée dans la régulation des gènes APP et PS1&2 [72]. Cette même AltProt a également été retrouvée dans la mutation de sORF de l'ARN mitochondrial (ARNmt), spécifique de la population japonaise et pouvant être impliqué dans la longévité de la population asiatique [73].

De même, Il a été montré que le mécanisme cellulaire de réponse à une infection virale, a pour impact l'augmentation du nombre d'ARNlnc dans la cellule. La recherche d'AltProts via les ARNlnc dans un contexte d'infection a permis de mettre en évidence de nouvelles identifications. De manière intéressante, l'augmentation du nombre d'ARNlnc correspond à l'augmentation des AltORF dans les cellules infectées. Cette augmentation du nombre d'AltORF a donc pour conséquence la traduction d'AltProts, de manière spécifique à l'infection virale [74]. Toutefois, l'implication et la localisation dans les voies de signalisation de ces AltProts restent encore une fois inconnues.

Aujourd'hui, l'observation des AltProts est un fait. Bien qu'ayant soulevé un grand débat initialement, leur existence est de mieux en mieux acceptée. Cependant leur rôle dans la cellule et leur implication dans des conditions pathologiques restent incompris et encore inexplorés. L'objectif de la thèse est de mettre en évidence les voies de signalisation dans lesquelles les AltProts sont impliquées. Placer et localiser les AltProts dans les réseaux et les voies de signalisation permettraient de donner une première hypothèse quant au rôle de celles-ci dans la cellule et par conséquent dans les pathologies impliquant ces voies de signalisation. Cette méthodologie permet d'attribuer, pour la première fois, un Gène Ontologie (GO) aux AltProts. L'utilisation des GO-terms pour prédire la fonction de protéines est notamment utilisée lors d'approches statistiques basées sur l'observation des PPIs [75].

La recherche des PPIs est essentielle pour mettre en évidence les partenaires et les voies de signalisation impliquées pour des protéines encore peu décrites. De multiples techniques permettent cette recherche de partenaires, nécessitant parfois des constructions biomoléculaires telle qu'en BioID [76], ou des recherches dépendantes d'un ligand anticorps pour une protéine cible en coIP.

Le retraitement de données dans ces expériences est également un point important et des logiciels d'analyse de réseau sont aujourd'hui disponibles (*e.g.*: Cytoscape [77], ClueGo [78], STRING [79]). Ceux-ci permettent de replacer les voies de signalisation, les GO-terms ou encore les types de liens référencés pour les protéines observées.

II. Analyse des interactions Protéine-Protéine

1. Mise en évidence des interactions protéine-protéine

A. L'interactome dans la compréhension du rôle des protéines

La mise en évidence des partenaires d'interaction permet d'obtenir des informations importantes notamment sur le réseau auquel appartient la protéine d'intérêt permettant ainsi de connaître les voies de signalisation dans lesquelles elle participe. De plus, la mise en évidence des partenaires directs permet alors d'émettre des hypothèses quant à la fonction possible de la protéine au sein des voies de signalisation identifiées. En effet, dans la cellule les protéines interagissent entre elles et un enchaînement de réactions peut débuter par la fixation d'un ligand sur un récepteur. Cette fixation peut entraîner la phosphorylation d'un partenaire du récepteur et par conséquent le transfert de phosphorylation et de proche en proche la modification des acteurs de la voie de signalisation. Dans une voie de signalisation, on trouvera donc des activateurs, mais aussi des inhibiteurs, des partenaires de contact, des enzymes,… L'enchaînement des modifications protéiques mène à des changements au niveau phénotypique de la cellule. C'est également la modification de ces voies de signalisation qui sont à l'origine de pathologies. Identifier les partenaires d'une voie de signalisation permet donc d'identifier la fonction de ces partenaires, et donne un indice supplémentaire afin d'identifier l'origine d'une pathologie.

B. Les méthodes d'étude de l'interactome

À ce jour, de nombreuses techniques ont été développées afin de détecter les interactions protéines-protéines (PPIs). Certaines comme la double hybridation en levure (Y2H) [80] ou encore le transfert d'énergie entre molécules

fluorescentes (FRET) [81] requièrent la construction de protéines fusionnées à partir des protéines cibles d'intérêt. D'autres comme le marquage de proximité (PLA) sont basées sur l'utilisation d'anticorps. Ces stratégies sont pour la plupart limitées à la détection d'un couple de protéines et non d'un complexe protéique impliquant plus de deux protéines. Plus récemment, le développement de la protéomique au travers de la MS a largement contribué au développement de nouvelles méthodes de détection des PPI basées sur la protéomique telles que l'AP-MS, l'Apex, le BioID, le Virotrap ou le XL-MS La stratégie la plus répandue reste néanmoins la purification d'affinité couplée à la MS (AP-MS) [82].

a. La double hybridation dans la levure (Y2H)

Le système classique Y2H est une méthode déjà ancienne [83] qui repose sur l'interaction de deux protéines fusionnées A & B. La première A est fusionnée avec une protéine de liaison à un domaine ADN, cette fusion lui permet de se fixer sur l'ADN un gène rapporteur. Dans un deuxième temps, la protéine B suspectée d'être en interaction avec la cible A, est fusionnée avec le domaine d'activation de l'ADN. Lors de l'interaction entre A & B, il y a reconstruction d'un facteur de transcription fonctionnel, permettant le recrutement de l'ARN polymérase II traduisant le gène rapporteur. Les gènes rapporteurs les plus utilisés sont HIS3 pour sélectionner les levures sur un milieu sans Histidine ou le gène *lacZ* qui conduit à un changement de couleur de la colonie (**Figure 6**). Les systèmes sont choisis pour fournir une lecture rapide et quantifiable.

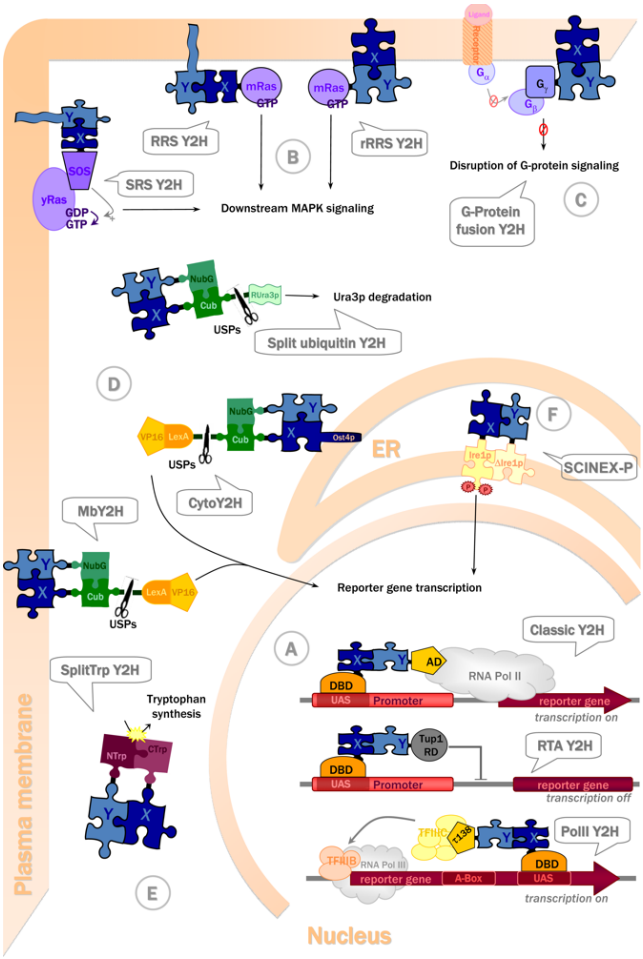


Figure 6 : ***Description des différentes applications de la méthode Y2H.*** *Application de la stratégie Y2H permet par différentes optimisations de la méthode de cibler différentes protéines ayant par exemple leurs propres spécificités de localisation. Application à la recherche de cible localisée telle que des partenaires cytoplasmiques, nucléaires ou membranaires, combiné à divers type de rapporteur expression de gènes et voies de signalisation spécifique. (Brückner & al., 2009* [84]*)*

Cette méthode a ensuite été détournée afin de réaliser des marquages de protéines de manière moins spécifique telles que l'ensemble des protéines provenant d'un compartiment cellulaire comme le cytosol, le réticulum endoplasmique, la membrane ou encore liées à une voie de signalisation comme la voie des RhoA-Rho kinases ou encore l'activation des protéines-G.

Cette méthode a plusieurs inconvénients. Elle est développée chez la levure, un organisme qui possède des différences biochimiques avec l'organisme d'étude de la protéine cible. Ceci peut entraîner notamment de fausses interactions (faux positifs), particulièrement si les protéines sont exprimées dans des compartiments cellulaires et à des moments différents entre l'hôte et le système

étudié. L'utilisation de la levure peut également entraîner la production de modifications post-traductionnelles (PTM) de la protéine cible différentes de l'organisme initial. S'il existe aujourd'hui des systèmes doubles hybrides dans les cellules mammifères (piège à interactions protéine-protéine MAPPIT) ceux-ci restent limités à deux protéines cibles et ne permettent pas la détection de larges complexes [85].

b. Le test de ligation entre protéines (PLA)

La méthode PLA dépend de la reconnaissance anticorps-antigène, la protéine d'intérêt et son partenaire d'interaction étant ciblés par des anticorps primaires différents produits chez des espèces différentes. La partie constante de ces anticorps primaires est ensuite reconnue par un anticorps secondaire lui-même couplé à un brin d'ADN. Si les deux protéines cibles sont proches dans l'espace (<40 nm) donc en interaction, il y a hybridation des brins d'ADN liés aux anticorps secondaires. L'ADN circulaire ainsi formé est ensuite amplifié, et détecté par l'ajout d'une sonde nucléotidique marquée par fluorescence permettant alors d'observer la formation de spots fluorescents s'il y a eu interaction entre les deux anticorps primaires (**Figure 7**) [86].

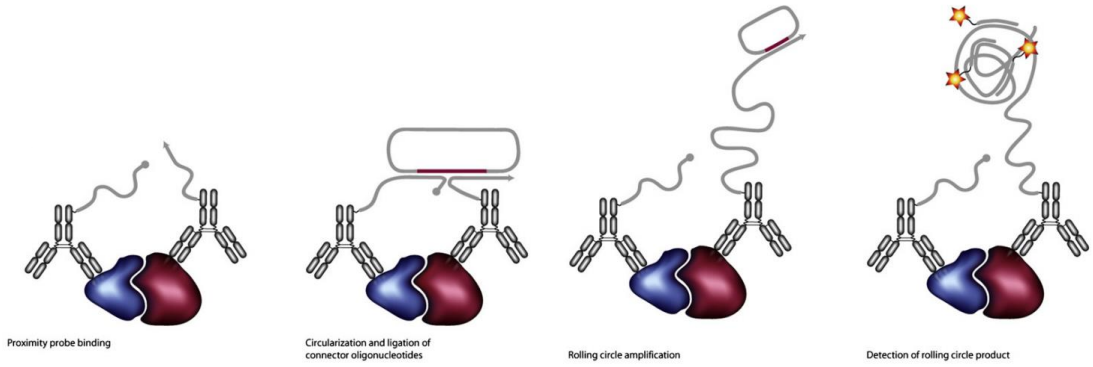


Figure 7 : ***Représentation des différentes étapes de la détection d'interaction par PLA****. 1- interaction entre les protéines reconnues par les anticorps combinés à une sonde, 2- circularisassions et ligation avec la sonde d'oligonucléotides, 3-4- amplification du produit oligonucléotidique et détection.(Söderberg & al., 2008* [87]*)*

Cette méthode permet de mettre en évidence une ou plusieurs PPIs en même temps, cependant elle nécessite de disposer des anticorps dirigés contre les cibles. Elle ne peut être utilisée que pour des partenaires d'interaction supposés. La spécificité, la distance d'interaction et la précision des informations (site

d'interaction) ne peuvent pas non plus être déterminées par cette méthode. Son coût mais aussi la difficulté de production d'anticorps spécifiques dirigés contre les AltProts, rendent cette stratégie difficilement adaptable aux AltProts.

c. Purification d'affinité couplée à la spectrométrie de masse (AP-MS)

La purification d'affinité couplée à la MS (AP-MS) est basée sur l'expression d'une protéine marquée, et la purification par capture d'affinité du marqueur de la cible et de ses partenaires. L'identification des protéines retenues est réalisée par LC-MS/MS. Les méthodes les plus représentées sont le TAP-TAG et le FLAG-TAG [88–90].

1. Le TAP-TAG

Cette stratégie utilise un plasmide afin de modifier le génome de la cellule cible. L'objectif est d'y ajouter une protéine A et un peptide de fixation à la calmoduline (CBP) tous deux séparés par un site de coupure enzymatique TEV protéase. Cette architecture permet dans un premier temps de récupérer la protéine cible et ses partenaires d'interaction fixés sur une colonne de billes IgG par affinité entre la ProtA et les IgG. On peut alors couper le site TEV protéase afin de libérer la partie CBP qui sera purifiée sur billes de calmoduline [89] (**Figure 8**).

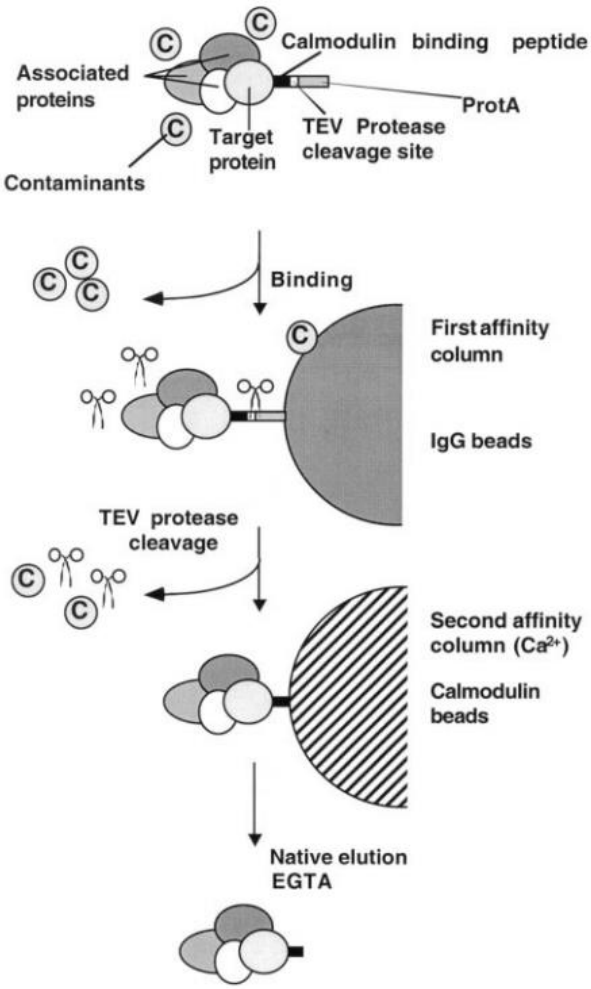


Figure 8 : ***Description des différentes étapes permettant de purifier un complexe protéique par stratégie TAP-TAG****. La protéine cible est fusionnée avec une construction contenant une protéine de fixation à la calmoduline et une protéine A, séparées par un site TEV de coupure protéase. La protéine fusionnée produite entre en interaction avec ses partenaires, les complexes formés sont retenus grâce à la protéine A sur colonne d'IgG. La coupure du site TEV est ensuite réalisée libérant de la colonne la cible et ses partenaires. Le complexe est alors purifié sur une colonne de calmoduline grâce à la protéine de fixation à la calmoduline puis élué. (Puig & al.,* [89])

2. Le FLAG-TAG

Le FLAG est l'ajout d'un octapeptide (DYKDDDDK)[91] sur la protéine cible. Ce peptide est un épitope reconnu par des anticorps M1. La protéine fusionnée ainsi que ses partenaires d'interaction peuvent de ce fait être retenus sur une colonne d'anticorps monoclonaux puis élués par une modification du pH ou en utilisant un agent chélateur tel que l'EDTA [92]. La réalisation de la fusion entre le tag et la protéine cible est réalisée par modification du génome ou insertion d'un plasmide porteur du tag et de la protéine.

Toutefois, si la fusion de ces marqueurs avec la protéine cible n'entraine pas de modification des PTM et permet l'expression de la protéine cible à un niveau physiologique pour la méthode TAP-TAG, elle est à l'origine d'un encombrement stérique important notamment lors de l'étude de petites protéines. De plus, la purification de complexes maintenus par des interactions faibles est difficile. Ainsi, lors de la capture par affinité les étapes de lavage peuvent entraîner des pertes d'interactions faibles et transitoires. Cette méthode ne permet pas non plus de différencier les partenaires directs ou indirects en interaction avec la protéine cible. Toutefois, cette stratégie permet encore aujourd'hui de mettre en évidence l'implication de composés dans des voies de signalisation. Récemment, une étude réalisée a montré l'implication des composés du ginseng dans la voie de signalisation de RAS, et donc de ses possibles effets anti tumoraux [93].

La nécessité de fusionner un marquage à la protéine cible provoque une surexpression de celle-ci pour certains cas de FLAG-TAG dans la cellule, induisant des rencontres de partenaires non spécifiques et une modification de l'expression des voies de signalisation impliquant la cible. Parmi les complications engendrées par la production d'une protéine fusion on notera également que le poids et l'encombrement induits par la protéine servant de marquage peut également entrainer une modification de la localisation de la protéine. Si l'AltProt pouvait passer la membrane nucléaire, le fait d´être associée avec une autre protéine peut empêcher ce déplacement, avec une observation finalement nucléaire quand elle devrait être cytoplasmique.

La solution afin d'éviter cette surexpression non désirée dans les cellules, est d'utiliser des méthodes d'immunoprécipitation (IP-MS), toutefois elles présentent également un grand nombre d'inconvénients liés à l'utilisation d'anticorps dirigés contre la cible possédant souvent un taux de spécificité assez bas et donc, une fixation non spécifique. De plus, l'utilisation d'un anticorps ciblant une protéine limite cette méthode aux protéines connues et pour lesquelles des anticorps existent. De ce fait, elles restent difficilement réalisables pour la recherche des partenaires d'AltProt.

d. Le marquage de proximité

Afin de s'affranchir des limites de l'AP-MS, de nouvelles stratégies ont récemment été développées basées sur le marquage des voisins proches dans l'espace d'une protéine cible. Parmi ces méthodes, on retrouve l'identification biotine dépendante de proximité (BioID) [94] et l'identification de partenaire de proximité ascorbate peroxydase dépendante (APEX) [95].

Le BioID est basé sur l'utilisation de la biotine ligase, variant de la biotine ligase BirA d'*E. Coli* au niveau de son domaine d'activation (R118G). BirA utilise l'ATP et la biotine pour synthétiser la biotinyl-5'-AMP, un composé réactif avec les amines primaires. Dans des conditions physiologiques la biotinyl-5'-AMP est liée aux protéines reconnues par le domaine R118G qui lui confère une spécificité. Toutefois, la mutation du site R118G de BirA (BirA*) entraîne sa déstabilisation. De ce fait la biotinyl-5'-AMP se fixe sur toutes les protéines environnantes. C'est la fusion entre BirA* et une protéine d'intérêt dans un milieu enrichi en biotine qui permet la biotinylation des protéines proches de la cible (**Figure 9**). Les partenaires ainsi biotinylés peuvent être récupérés par affinité via la streptavidine puis identifiés par LC-MS/MS [96]

Cependant, cette méthode souffre également de quelques limitations. En effet, la fusion de la protéine BirA peut provoquer des modifications de localisation de la protéine cible, un encombrement stérique modifiant les PPIs ou encore une modification de l'activité. La biotinylation des partenaires est également dépendante de la présence d'amine libre (principalement les lysines), ayant également comme conséquence une possible modification des PTMs du site. La méthode permet de mettre en évidence tous les voisins proches dans un rayon de 20nm ce qui n'implique pas nécessairement que tous ces partenaires soient en interaction. Cela peut conduire à l'identification de faux positifs et des validations sont donc nécessaires. Afin de réduire le temps de marquage (18-24H initialement) et permettre des études dynamiques plus fines, des optimisations ont été réalisées réduisant le temps de biotinylation (TurboID) permettant un marquage en 10 min [97] ou en réduisant la taille de la biotine ligase (MiniTurboID) avec un temps de marquage proche de la version classique

du bioID mais avec une sensibilité plus importante [97], permettant d'observer des protéines faiblement exprimées.

La stratégie APEX-MS repose sur une biotinylation de proximité via l'ascorbate peroxydase (APEX) qui catalyse l'oxydation de la biotine-phénol en biotine-phénoxyle en présence d'H_2O_2. La biotine-phénoxyle est quant à elle réactive avec les acides aminés riches en électrons tels que la tyrosine, le tryptophane, la cystéine ou encore l'histidine (**Figure 9**). De cette manière, la protéine cible est fusionnée avec l'APEX puis incubée 30 min avec de la biotine-phénol avant d'être exposée 1min à l'H_2O_2 induisant la biotinylation. Cette méthode efficace en environ 1 min est plus rapide que la version BioID classique et donc souvent représentée dans des études temps dépendantes. Cette stratégie à l'origine utilisée en microscopie électronique, car les radicaux libres provoqués par l'H_2O_2 polymérisent et précipitent, les rendant observables, est aujourd'hui utilisée dans diverses applications d'études interactomiques telles que les protéines transmembranaires du réticulum endoplasmique [98]. Des optimisations de la méthode par des sélections de versions mutées de l'enzyme ont abouti au développement de l'APEX2. Cette version mutée de l'enzyme ascorbate peroxydase permet d'augmenter la sensibilité de la méthode, ainsi avec peu d'enzyme le taux de biotinylation est plus élevé, facilitant la détection. Cette nouvelle version permet donc de mettre en évidence les partenaires de protéines faiblement exprimées.

Si l'encombrement de la fusion avec APEX (27 kDa) est moindre qu'avec BirA* (35 kDa), elle représente quand même des masses très importantes comparées aux cibles que sont les AltProts (en moyenne 50 acides aminés) pouvant de ce fait modifier les PPIs, les localisations ou encore les fonctions de la cible. La méthode APEX nécessite également une activation à l'aide d'H_2O_2, toxique pour les cellules et donc difficilement utilisable *in vivo ;* tandis que la biotinylation permet de garder des organismes en vie durant l'expérience.

e. Le piège à protéine VIROTRAP

Le principe de la VIROTRAP est fondé sur le mécanisme de production de particules virales (VLP) par la cellule lors d'une infection [99]. Dans ce mécanisme de multiplication du virus, une protéine peut déclencher la réaction

de production des VLP : p55 GAG. GAG et ses partenaires sont alors recrutés à la membrane. L'accumulation de GAG à la membrane entraîne une multimérisation de GAG. Cette formation a pour effet de former une excroissance de la membrane, qui finit par donner la VLP qui sera libérée dans le milieu cellulaire. Cette VLP contient normalement tout le matériel nécessaire à l'infection virale d'autres cellules [99].

Issue de cette observation la production, dans une cellule non infectée, d'une protéine cible fusionnée avec la protéine GAG permet alors le recrutement de la protéine et de ses partenaires d'interaction à la membrane puis leur libération dans le milieu cellulaire sous forme de VLP. Ainsi la purification des VLP et leur analyse permettent d'identifier les partenaires de la protéine cible [100].

Cette méthode a pour avantage de capturer l'ensemble des partenaires de la protéine cible sans restriction de distance ni sélectivité, elle permet un enrichissement en se libérant de la présence des autres protéines de la cellule. Toutefois elle ne permet pas une étude non ciblée, à large échelle et comme pour les méthodes de BioID et d'APEX, elle nécessite la production d'une protéine cible fusionnée avec une autre protéine (**Figure 9**). L'utilisation d'une construction pour la production d'une protéine peut entraîner des variations de quantité d'expression, et donc des modifications des partenaires d'interaction retrouvés. L'utilisation de deux protéines fusionnées modifie également l'encombrement et l'accessibilité aux partenaires d'interaction potentiels. La présence d'une autre protéine telle que GAG au sein de la cellule peu également engendrer des modifications, en influençant d'autres voies de signalisation. Si la démonstration de la méthode est réalisée dans des cellules HEK293T, il est précisé que la présence de particules virales peut engager une réaction antivirale et peut nécessiter une modification de la voie de réponse antivirale de la cellule avant utilisation, suivant l'étude réalisée [100].

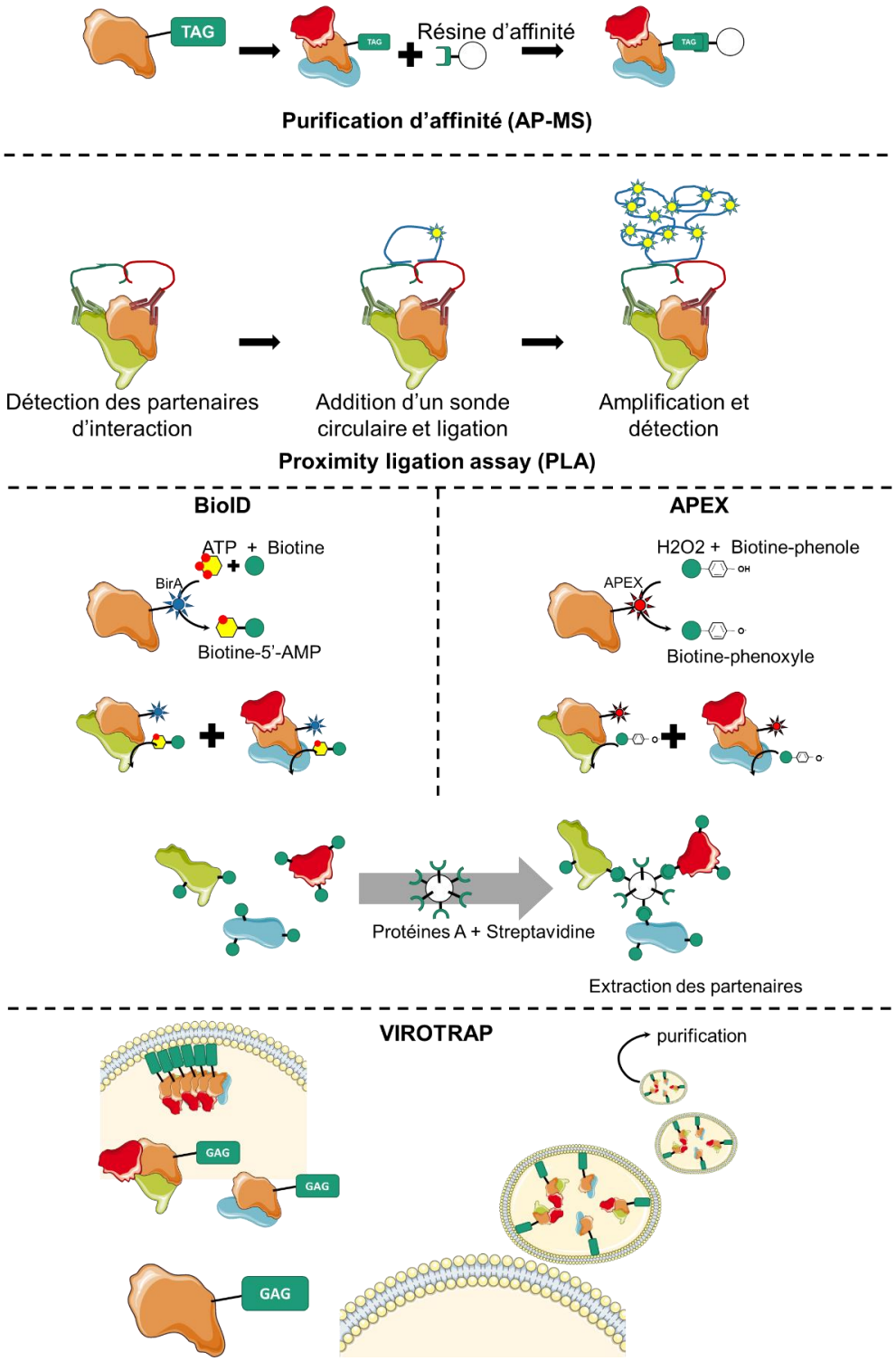


Figure 9 : ***Représentation des différentes méthodes d'identification de l'interactome d'une protéine*** *: AP-MS, PLA, BioID, APEX et VIROTRAP. Chaque méthodologie présentée a ses avantages et inconvénients, toutefois toutes nécessitent de connaître une cible protéique. Ces stratégies sont donc limitantes dans la mise en évidence d'interactions pour de nouvelles protéines telles que les AltProts.*

2. La méthode de XL-MS

Il n'existe à ce jour qu'une méthode non ciblée permettant la mise en évidence de PPIs dans une cellule, c'est la méthode de XL-MS. Cette stratégie repose sur la formation de liaisons covalentes intra et inter protéines nommées « *crosslink* » (pontage chimique ou XL) à l'aide d'une molécule homo ou hétéro-bifonctionnelle (agent de pontage ou *crosslinker*) réagissant avec les chaînes latérales des acides aminés proches dans l'espace, permettant ainsi de figer les interactions [101]. Cette méthode a été historiquement utilisée en protéomique structurale en supplément de méthodes comme la microscopie électronique (EM), la RMN et la cristallographie. Cette technique permet d'identifier des distances de résolution plus faibles que les autres méthodes tout en apportant des données structurales sur l'arrangement des complexes (résidus proches dans l'espace). De plus, elle présente l'avantage de pouvoir être appliquée sur des complexes de grandes tailles. Ainsi, la méthode de pontage chimique couplée à l'analyse par MS est devenue un nouvel outil d'étude structurale en biologie. Bien que cette technique ait été longtemps utilisée pour des complexes protéiques purifiés, les avancées dans le domaine de l'analytique et de la bioinformatique permettent désormais de réaliser ces approches sur des milieux complexes.

La stratégie XL-MS conventionnelle et que nous avons mise en place au cours de ces travaux de thèse, présentée **Figure 10,** se décompose en plusieurs étapes. Premièrement, afin de faciliter la détection des protéines pontées la cellule est dé-complexifiée (séparation noyau/cytoplasme) puis les protéines extraites et pontées. La deuxième étape consiste à réaliser la digestion enzymatique dans un filtre (de cut-off 50 kDa) et permettant d'éliminer les protéines <50kDa qui ne seraient pas pontées. Enfin, les peptides pontés sont enrichis par exemple sur colonne SCX, avant d'être analysés en LC-MS.

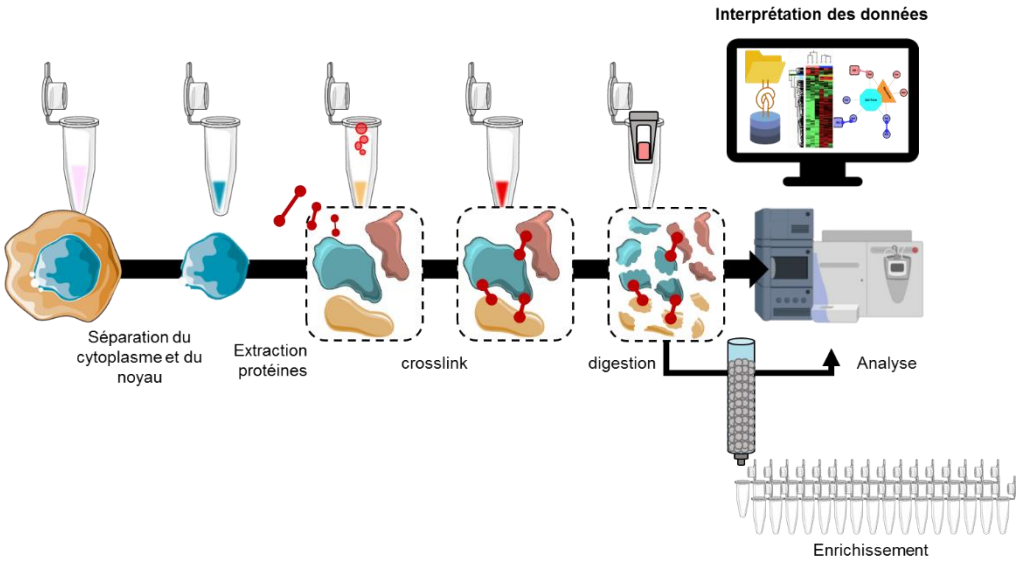
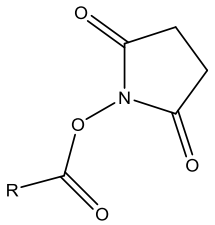
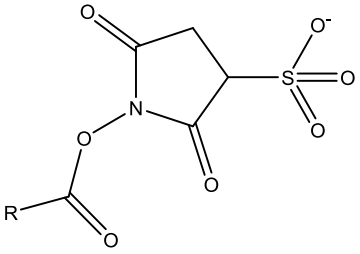
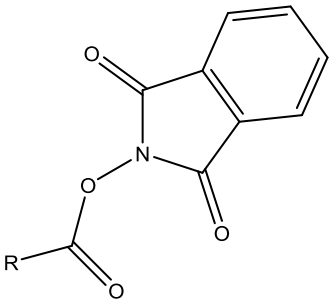
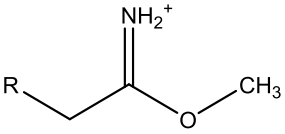


Figure 10 : ***Stratégie suivie lors de la réalisation de la méthode XL-MS****, séparation du noyau et du cytoplasme afin de dé-complexifier l'échantillon, extraction des complexes protéiques du noyau puis pontage des protéines. Réalisation de la digestion de l'échantillon, suivi de l'analyse. L'analyse peut être précédée d'une phase d'enrichissement des peptides pontés par chromatographie SEC ou SCX.*

A. Les agents de pontages

Un agent de pontage (ou *crosslinker*) est défini par la nature chimique de ses extrémités et sa longueur, qui conditionnent les acides aminés ciblés sur les protéines et la distance maximale d'interaction. La majeure partie des analyses XL-MS réalisées utilise des agents de pontages homo-bifonctionnels, réagissant avec les amines primaires des chaînes latérales de la protéine. La fonction chimique réagissant avec les amines primaires, la plus répandue est le N-hydroxysuccinimide ester (NHS). Il existe d'autres fonctions réactives pouvant également être utilisées tels que le N-hydroxyphthalimide (BDP-NHP), le hydroxybenzotriazole, ou encore le 1-hydroxy-7-azabenzotriazole (**Table 1**).

Table 1 : Présentation des fonctions chimiques couramment retrouvées dans les agents de pontages

Fonction chimique	Exemple de d'agent de pontage	formule
N-hydroxysuccinimide ester (NHS)	DSS, DSSO	
Sulfo-NHS	BS3	
N-hydroxyphthalimide	BDP-NHP	
Imidoester	DEST	

Actuellement, un des challenges de la communauté utilisant le XL-MS est de trouver des fonctions réactives permettant une plus grande efficacité et hétérogénéité de la fixation à la surface des protéines, afin de garantir un maximum d'information sur les interactions présentes dans le milieu. Les fonctions imidates peuvent ainsi être utilisées à condition de rester à pH physiologique sans modifier la basicité de la protéine. Une optimisation de cette fonction chimique a permis de mettre en évidence les PPIs impliquant 50 partenaires ribonucléoprotéines chez *E.Coli*. Le diethyl suberthioimidate (DEST),

agent de pontage utilisé dans cette étude montre notamment des avantages physicochimiques. Il permet, en effet, de garder les charges sur les résidus de lysine et ainsi de faciliter la séparation des peptides pontés par rétention en chromatographie d'échange cationique fort [102]. Ces dernières années, l'utilisation de groupements Diazo a également été démontrée, permettant la formation d'agents de pontages réagissant avec les résidus acides des protéines comme l'acide aspartique et l'acide glutamique [103]. Les groupements cystéines peuvent également être ciblés. Classiquement la réaction est réalisée par un pontage au maléimide, mais plus récemment l'utilisation du 1,3-diallylurea (DAU) a été décrite. Le DAU est un agent pontant de type photo-thiol réactif sous UV-A [104] et défini comme zéro distance (ou *zero-length*), car présentant une longueur entre 9 et 10 Å. Le formaldéhyde, un des plus anciens agents pontant, connu pour maintenir la localisation des protéines dans les tissus fixés, est le plus répandu de la famille des « *zero-length* ». L'utilisation de ce genre de d'agent pontant permet d'identifier des interactions très proches dans l'espace et entre protéines. En effet, les cellules sont perméables à ces agents de pontages et ils permettent une fixation *in vivo* plus aisée. Toutefois, il a été montré que la distance entre deux lysines d'une protéine ou d'un complexe protéique est souvent comprise entre 35 et 40 Å [105]. Il faut donc prendre en compte ces différents paramètres et adapter la fonction chimique réagissant sur le résidu d'acide aminé avec la longueur possible et acceptable entre les deux résidus dans l'espace recherché.

Certains agents pontant sont activables notamment par photo-réactivité. En effet, en mélange avec des protéines leurs fonctions sont activées par une exposition aux UV, comme c'est le cas des composés diazirine qui sont activés par des UV-A. Cette activation a pour effet de former un composé diazo qui réagira avec les acides aminés asparagines et glutamines ainsi qu'avec la partie C-terminale des protéines en formant un ester. Mais le diazirine activé peut également former un carbène qui alors peut s'associer à tous les acides aminés [103].

Hormis les paramètres de fonctions, réactions et de distances qui viennent d'être évoqués, les agents de pontages sont divisés en deux grandes familles,

dépendamment de leurs réactivités en analyse MS. On parle alors d'agents pontant non clivables ou clivables en MS. L'introduction d'agents pontant clivables en MS a été motivée par la difficulté d'interprétation des données obtenues à partir des agents de pontage non clivables.

a. Les agents pontant non clivables

Parmi les non clivables les plus répandu, on peut citer le disuccinimide suberate (DSS) qui possède deux groupements NHS à chaque extrémité lui permettant de créer des liaisons covalentes avec les protéines à travers les amines primaires (i.e. lysine). Il possède une chaîne carbonée de 8 atomes pour une longueur totale de 11,4 Å et est perméable aux membranes cellulaires. Au contraire, le bis(sulfosuccinimidyl)suberate (BS3 ou Sulfo-DSS) qui a une structure proche du DSS mais une polarité différente est un agent de pontage imperméable aux membranes qui est par conséquent utilisé pour les études portant sur les protéines extracellulaires, les protéines membranaires ou les protéines en solution. L'avantage du BS3 par rapport au DSS est sa capacité à être solubilisé dans l'eau et non dans un solvant organique, ceci grâce à ses fonctions sulfonates qui modifient la polarité malgré une structure principale identique.

Cette classe d'agents pontant présente toutefois une limite très importante. En effet, lors de l'analyse en MS les peptides pontés issus de la digestion des protéines en interaction sont non labiles. Ainsi, le pont ne sera pas un site de fragmentation préférentiel lors des analyses MS/MS. Les spectres générés sont alors complexes et nécessitent la prédiction des partenaires d'interactions potentiels. Pour ce faire, les logiciels de traitement de données doivent générer l'ensemble des peptides possibles puis confronter chacun de ces peptides entre eux s'ils possèdent un acide aminé permettant la fixation de l'agent pontant. De ce fait, la base de données ainsi générée s'agrandit de manière exponentielle rendant impossible l'analyse de mélanges complexes. L'utilisation de ce type d'agents de pontage est donc plus répandue dans le cadre d'échantillons déjà purifiés et lors d'analyses structurales.

b. Les agents pontant clivables en MS

Parmi les agents de pontages dits clivables en MS, on trouve de manière répandue et commerciale le disuccinimidyl sulfoxide (DSSO) [106] et le Disuccinimidyl Dibutyric Urea (BuUrBu ou DSBU) [107]. Ils ont l'avantage d'être clivables en MS/MS sous dissociation induite par collision (CID), générant des doublets caractéristiques après fragmentation (**Figure 11**). Cette empreinte spécifique de peptides pontés permet d'automatiser les identifications par prédiction du décalage de masse associé aux peptides. De ce fait, la modification réalisée par le pontage est déterminée comme un décalage de masse du peptide comme lors de l'identification d'une PTM. L'identification est réalisée grâce à un pic majoritaire non ponté et un pic minoritaire ponté possédant un décalage de masse. Dans le cas du DSSO, la différence de masse est de 158 Da. L'identification par MS/MS est ensuite corrélée à la MS pour confirmer les deux partenaires identifiés et isolés. De plus, certains de ces agents de pontages tels que le DSS sont perméables aux cellules permettant ainsi les analyses *in vivo*. On peut noter l'exemple de l'utilisation de la stratégie « protein interaction reporter (PIR) » utilisant un agent de pontage perméable aux membranes cellulaires et portant un tag présentant une signature spécifique en MS [108].

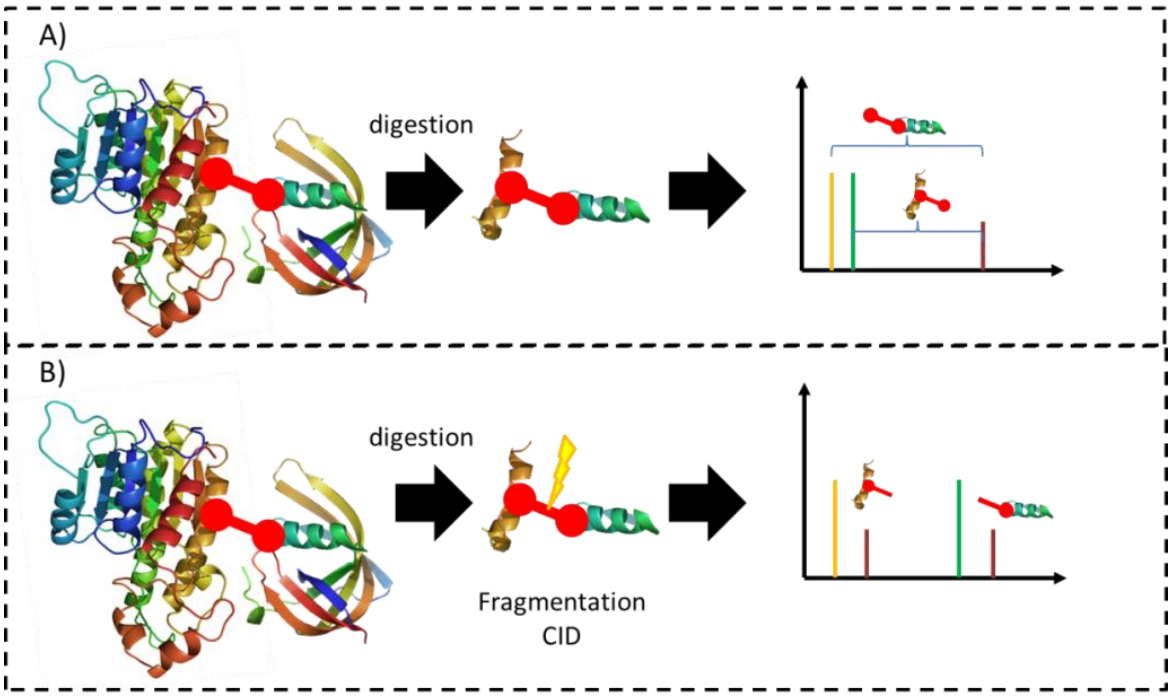


Figure 11 : ***Différences entre l'utilisation de stratégies non clivable et clivable en MS****. A) stratégie utilisant un agent de pontage simple, le pic de masse supplémentaire observé correspond aux peptides liés avec l'agent de pontage. B) strategie CID clivable : la fragmentation rompt l'agent de pontage libérant un fragment correspondant au peptide additionné d'une partie du pontage.*

Malgré le progrès considérable pour contrôler les réactions de pontage, un agent pontant peut ne pas réagir totalement, créer des pontages inter protéines ou intra protéines. Dans ce contexte, une nomenclature a été établie afin de pouvoir différencier les différentes classes de pontages pouvant être observées par MS.

c. Les différents types de pontage

La nomenclature établie est présentée **Figure 12**.

Parmi les différents pontages, il est possible d'observer :

- Le pontage « *dead-end* » ou type 0: désignant un agent pontant n'ayant réagi que d'un seul côté avec une protéine
- Le pontage intra-peptidique ou type 1: désignant un pontage sur la même protéine, établi entre deux parties proches dans l'espace.
- le pontage inter-peptidique ou type 2 : désignant une interaction entre deux protéines identifiées par chacun des peptides pontés.

La formation de ces pontages entraine des différences de masse spécifiques ou une identification de peptides issus de la même protéine permettant d'attribuer à un pontage son type (0, 1 ou 2). Une limitation existe cependant dans le cas des complexes homo-multimèriques où il devient impossible de distinguer un pontage intra (type 1) d'un pontage inter (type 2). Seule la confrontation entre la distance connue de l'agent pontant aux points de fixation et la modélisation de la protéine sous forme multimèrique peut permettre de différencier ces deux cas.

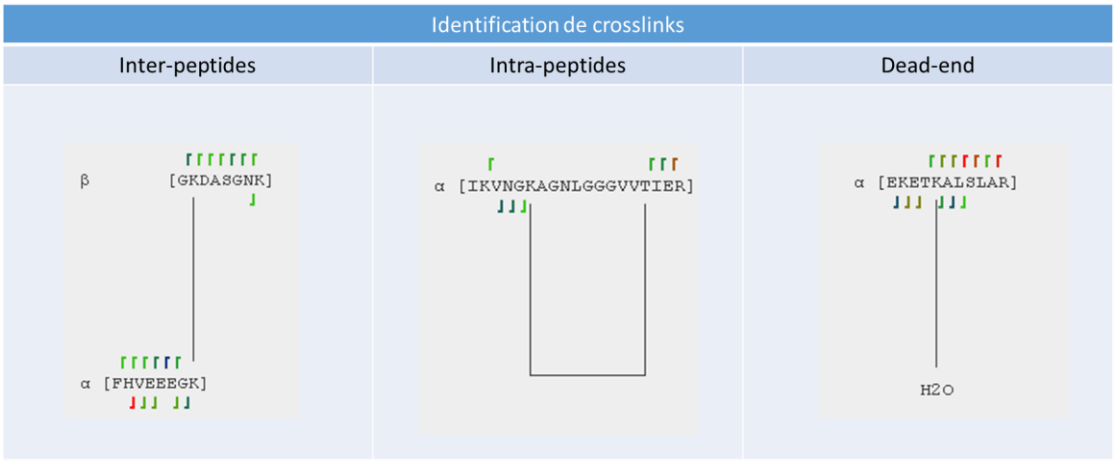


Figure 12 : ***Description des différents types de liaison.*** *Ces formations de fixation de l'agent pontant sur une ou des protéines sont observables sur les peptides de digestion et identifiables après analyse XL-MS.*

d. La quantification des analyses XL-MS

Les méthodes de quantification traditionnelles en protéomique ont également été explorées en XL-MS. C'est le qXL-MS. Ces stratégies permettent d'obtenir des informations supplémentaires quant à la dynamique de la structure protéique, la variation du nombre d'interactions et de partenaires en fonction du temps. Plusieurs études utilisent le marquage isotopique par isotopes stables, comme le deutérium, des agents de pontages pour suivre cette dynamique. Ainsi les expériences sont réalisées en utilisant un échantillon marqué par l'agent pontant normal dit léger et l'agent de pontage marqué ou encore nommé lourd. Le marquage isotopique du BS3 a permis de mettre en évidence que la variation de la phosphorylation induit un changement de conformation du chloroplaste ATP synthase [109]. La méthode qXL-MS a également été utilisée dans des études à large échelle dans des cellules d'*E. coli* [110]. Toutefois l'utilisation de marquage

isotopique introduit des complications dans l'analyse des données. En effet le spectre est complexifié par la séparation des signaux provenant de chaque paire de peptides pontés lourds et légers. Une autre difficulté est liée au décalage de temps de rétention causé par l'utilisation d'un composé marqué. Cela a pour effet de compliquer la détection et l'intégration des pics. Afin de palier à ces difficultés, des méthodes en LFQ sont actuellement en développement [111].

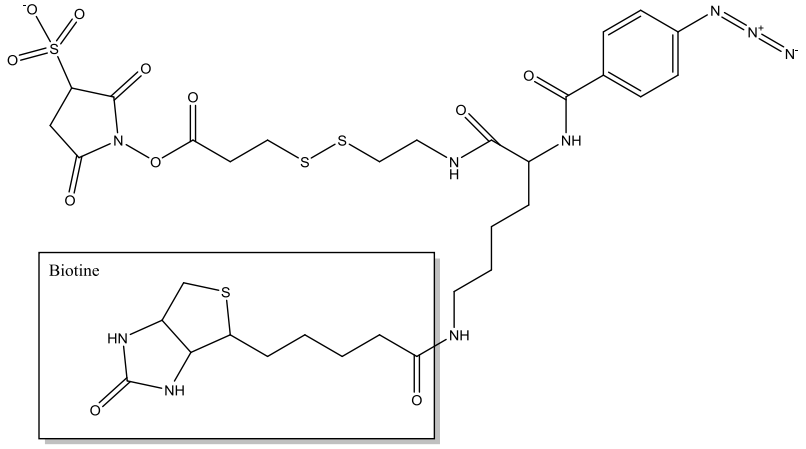
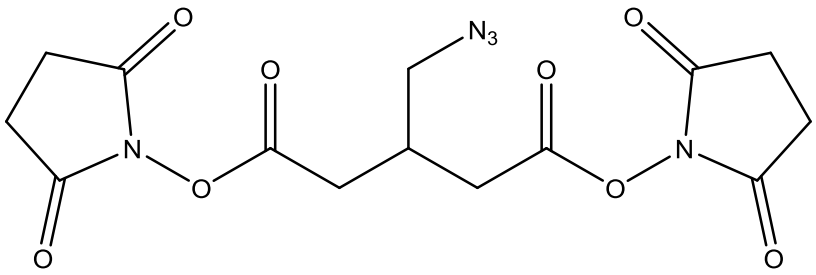
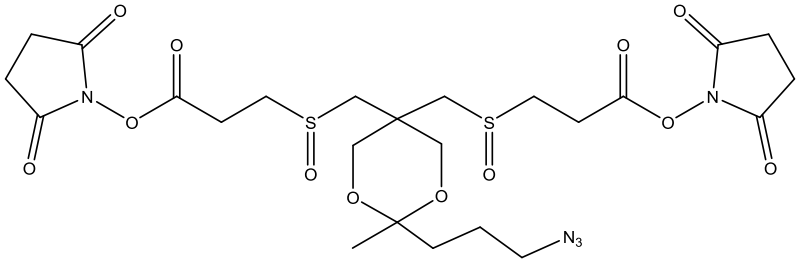
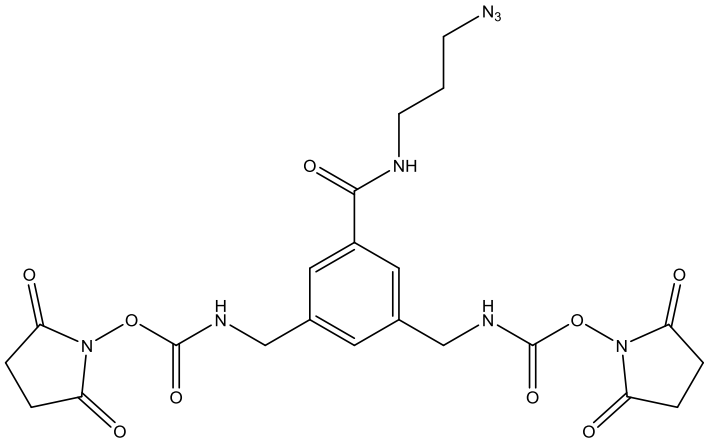
Un autre challenge de l'utilisation des méthodes XL-MS est lié à l'enrichissement en peptides pontés avant l'analyse MS. En effet après digestion les peptides pontés ne sont pas majoritaires par rapport aux peptides non pontés. Ainsi lors de l'analyse MS, les peptides non pontés seront principalement identifiés au détriment des peptides pontés d'intérêt, limitant le nombre de protéines pontées identifiées. Ainsi, il est nécessaire de mettre en place des méthodes d'enrichissement. Certains agents de pontages peuvent être fonctionnalisés pour faciliter cette étape d'enrichissement.

B. Les méthodes d'enrichissement

a. Les agents pontant fonctionnalisés

L'agent permettant le pontage des protéines peut être porteur d'une fonction chimique spécifique, permettant ainsi son enrichissement avant analyse par MS (**Table 2**). On trouve notamment le sulfo-N-hydroxysuccinimidyl-2-(6-[biotinamido]-2-(p-azido benzamido)-hexanoamido) ethyl-1,3'-dithioproprionate (Sulfo-SBED) possédant une biotine, le bis(succinimidyl)-3-azidomethyl glutarate (BAMG) possédant une fonction azido ou encore le disuccinimidyl bis-sulfoxide (Azide-A-DSBSO) possédant un bras azide permettant leur enrichissement.

Table 2 : Représentation des structures des différents agents de pontages fonctionnalisés

Sulfo-SBED	
BAMG	
Azide-A-DSBSO	
NNP9	

La fonction chimique la plus répandue reste le couplage avec une biotine. L'utilisation d'un agent de pontage biotinylé facilite grandement l'enrichissement des peptides pontés après digestion. Cependant, la biotine entraîne un encombrement stérique non négligeable lorsque l'on utilise un pontage de quelques Angströms. Récemment l'utilisation d'un agent trifonctionnel, le NNP9 permettant une réaction de *« click chemistry »* sur sa troisième fonction, a été décrit [112]. Le NNP9 est inséré dans l'extrait protéique, fixant les PPIs de l'échantillon, puis la réaction de « click » est réalisée. L'incorporation de la biotine est faite avant la digestion trypsique. Les peptides de digestions pontés peuvent alors être enrichis grâce à l'affinité entre la biotine et la streptavidine ou la neutravidine [113]. Cette méthode a pour avantage de ne pas perturber la réaction de pontage et règle les problèmes d'encombrement stérique tout en permettant l'intégration de la biotine, après réaction avec l'agent de pontage. Cependant, l'utilisation de fixation avec la biotine peut entraîner une modification de la fragmentation en phase gazeuse. En effet à ce jour peu d'études ont porté sur la fragmentation de ces agents de pontages en MS [114].

Dans une recherche d'optimisation de l'enrichissement, la communauté met en place de nouvelles technologies. Récemment, un nouvel agent pontant a été synthétisé possédant une fonction acide phosphonique (PhoX). Cette fonction a pour intérêt de pouvoir être enrichie par chromatographie d'affinité métallique (IMAC). Elle présente des résultats plutôt satisfaisants avec 97% de spécificité [115].

b. L'enrichissement par chromatographie d'exclusion stérique (SEC)

Les peptides pontés ont un poids moléculaire supérieur aux peptides non pontés Cette augmentation de masse est la cible de l'enrichissement par chromatographie d'exclusion stérique (SEC). Les peptides issus de la digestion des protéines en interaction pontées sont injectés en SEC pour une séparation en fonction du poids moléculaire et la réalisation d'un fractionnement. Les peptides de masse plus élevée seront élués en premier sauf pour les peptides de poids trop élevé et dépendamment de la perméation de la colonne utilisée qui sortiront au volume mort. Cette séparation en fonction du volume de rétention dans la colonne permet de fractionner l'échantillon contenant les peptides pontés

des peptides non pontés. En effet, deux peptides pontés ensemble ont en moyenne un poids moléculaire de 3000 Da alors qu'un peptide simple sans pontage présente en moyenne un poids moléculaire de 1500 Da. Cependant, s'il est possible d'appliquer cette stratégie sur une protéine ou un complexe protéique purifié, il peut être difficile de l'utiliser dans un mélange complexe issu d'un extrait cellulaire. En effet, si le rendement de digestion enzymatique de l'échantillon est faible (i.e. sites de coupure manqués par l'enzyme) alors la différence de masse entre les peptides pontés et non pontés n'est plus existante et l'enrichissement des peptides pontés devient complexe.

c. L'enrichissement par chromatographie d'échange cationique fort (SCX)

La modification de la taille des peptides pontés entraîne dans le même temps, une augmentation de la charge présente sur le complexe. Basé sur cette observation, l'enrichissement utilisant la SCX a pour objectif de séparer les peptides de digestion les plus chargés ayant une interaction plus forte avec la phase stationnaire de la colonne chargée négativement et donc un temps de rétention plus important. Cette méthode a montré son efficacité sur des mélanges protéiques purifiés [116] mais aussi sur des mélanges complexes issus d'extractions cellulaires [117]. La SCX peut être utilisée de différentes manières. De manière conventionnelle, le chargement de la colonne nécessite un minimum de 200 µg d'échantillon, avec un optimum de 500 µg [118], tandis que d'autres systèmes miniaturisés sous forme de « *stage-tip* » nécessitent moins d'échantillons [119]. La lysine est un acide aminé chargé positivement à pH<10,53 (pKa=10,53 NH2 de la chaîne latérale, pKa=8,95 pour le NH2 terminal). Or cet acide aminé est impliqué dans l'interaction avec l'agent de pontage, diminuant la charge des peptides pontés. Les peptides non pontés sont donc plus retenus que les peptides pontés en SCX. La structure formée par le pontage et la taille de l'agent pontant peuvent cependant également influencer la charge globale des peptides pontés nécessitant des optimisations de la méthode SCX. Pour limiter la perte de charge sur les peptides pontés, il est possible d'utiliser un agent de pontage possédant lui-même des charges positives et offrant alors la possibilité d'une double séparation en SCX, avant et après

induction de la charge sur l'agent de pontage [120] ou un agent possédant une charge positive à pH acide tel que le diethyl suberthioimidate (DEST) [121]

C. Stratégies MS

Certains paramètres des peptides pontés sont à prendre en compte lors de l'analyse en MS. En effet, la présence du pontage peut provoquer une perte de certains sites de coupure par l'enzyme de digestion utilisée, généralement la trypsine. Cela a pour conséquence de former des peptides pontés plus longs que les peptides linéaires de digestion classiques. La répartition de masse pour des peptides pontés est ainsi présentée comme étant à 98% supérieure à 1300 Da alors que cette masse ne représente que 67% des peptides linéaires. De même, la charge portée par le précurseur est influencée par sa taille. De ce fait, l'état de charge des peptides pontés est à 98% supérieur à +3 [122]. Ces observations ont permis de mettre en place des paramètres de filtres à appliquer lors de l'analyse MS pour les stratégies XL-MS, notamment le rejet des ions précurseurs chargés +2.

La fragmentation en MS de deux peptides pontés est également affectée par la présence de ce pontage. Lors de la réalisation de la fragmentation, un peptide α et un peptide β sont définis, dépendamment du taux de fragmentation du peptide, α étant le plus fragmenté et β le moins. Toutefois on observe un écart important entre la fragmentation des peptides α et β. 78% des fragments obtenus à partir des 10 pics les plus intenses sont attribués au peptide α dans le cas d'un pontage de type BS3 sous CID [122]. La fragmentation des peptides se décompose en plusieurs séries d'ions ; soit la fragmentation N terminale : ions a (ion aldimine), ions b (ion acylium) et ions c (ion amino) ; soit C terminale : ions x (ion acylium), ions y (ion amino) et ions z (carbocation). En analyse XL-MS la répartition des ions-y et ions-b n'est pas équivalente. Les ions-y sont beaucoup plus intenses que les ions-b. Lors d'une analyse la fragmentation de 10 pics les plus intenses aboutit à 8 fragments de type y alors que 2 seulement seront des ions-b (**Figure 13**).

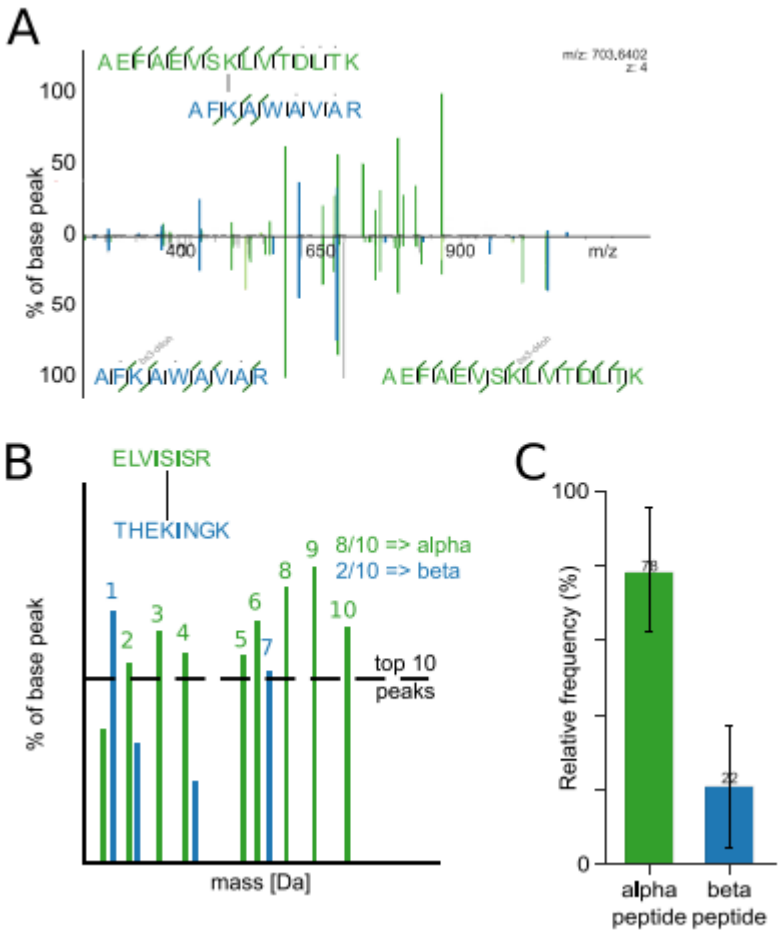


Figure 13 : ***Répartition de la fragmentation de peptides pontés lors d'une analyse XL-MS****. A. Représentation de la fragmentation de deux peptides pontés par du BS3. B. Ensemble des pics pris en compte lors d'une analyse en Top10 sur deux peptides pontés, on observe une intensité nettement plus intense pour les ions alpha, ce qui explique qu'ils soient les plus représentés en analyse Top10. C. On retrouve cette disparité dans l'analyse de la fréquence (Giese & al., 2016* [122]*)*

Afin de compenser la difficulté de fragmentation du peptide β d'autres méthodes d'activation des ions et de dissociation peuvent être employées comme la dissociation par transfert d'électron (ETD).

L'utilisation d'agents pontant clivables en MS permet aussi de s'affranchir de certaines limites permettant d'identifier les peptides α et β comme deux peptides linaires partageant une modification. Afin d'augmenter la qualité de l'identification des partenaires d'interaction, de nouvelles stratégies d'analyse MS ont été proposées. Celles-ci couplent différentes méthodes de fragmentation. Une étude utilisant les informations générées par la fragmentation CID et ETD du même ion précurseur en MS1 a permis l'identification de 2000 pontage dans

un extrait total de cellules HeLa [117]. Cette méthode a ensuite été couplée à la fragmentation en MS^3, permettant d'enrichir l'information de fragmentation obtenue en analyse MS. La réalisation d'une double fragmentation en parallèle du même ion précurseur de la MS1 par les méthodes CID-MS^2 et ETD-MS^2 puis par la fragmentation d'une deuxième fois en CID-MS^3, permet de mettre en évidence jusqu'à quatre fois plus de pontage qu'une méthode CID-MS2 classique (**Figure 14**) [123].

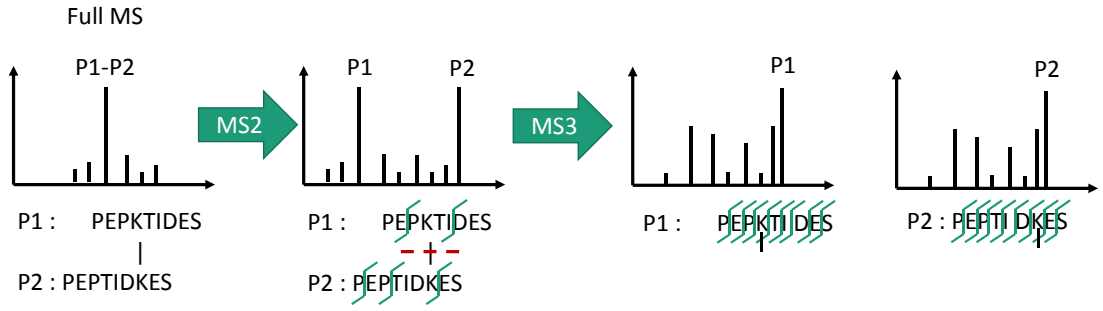


*Figure 14 : **Schématisation de la fragmentation des ions XL-MS**. Description de la méthode d'activation de deux peptides pontés par un agent de pontage clivable, fragmentation en MS2 et MS3.*

D. Les outils informatiques en XL-MS

Le retraitement informatique des données MS est une étape cruciale pour l'interprétation des résultats. La mise en place du pontage, influençant la fragmentation ainsi que les différentes stratégies de fragmentations comme précédemment présentées, nécessitent des algorithmes dédiés. De nombreux logiciels ont été développés tels que XQuest [124], Plink [125], StavroX [126], MeroX [127], XlinkX [117]. Certains permettent notamment de comparer les données de MS à des banques de protéines complètes. La limitation première des algorithmes utilisés réside dans le traitement du nombre de possibilités. Lors de l'utilisation d'un agent de pontage non clivable, le nombre de possibilités d'interaction pour un ion précurseur identifié est exponentiel. Ainsi pour n peptides la valeur estimée de possibilités de combinaisons de pontage est de $n^2/2+n$ [114], cette prédiction réduit énormément les capacités de calcul. Avec l'apparition des pontages clivables, l'identification du pontage peut se faire sur la base d'une modification post traductionnelle, impliquant un décalage de masse, d'un peptide linéaire, spécifique à la masse de l'agent pontant par exemple 158,003 Da pour le DSSO. XLinkX couplé à Proteome Discoverer

(ThermoScientific) permet par exemple de traiter des échantillons utilisant des agents de pontages non clivables, avec une base de données limitée à 600 entrées, mais propose également des stratégies identifiant des agents de pontages clivables en MS2, MS2_MS2 et MS3 et ainsi de traiter des échantillons complexes avec des bases de données complètes.

Aujourd'hui, les perspectives en termes d'analyse de données XL-MS sont surtout centrées sur la recherche de significativité, par l'amélioration de l'identification et par la diminution du taux de faux positifs. Les méthodes de quantification sont également en émergence afin de pouvoir observer dans la cellule à différents temps ou sous différentes conditions les modifications de la quantité d'interactions présentes.

E. Stratégies XL-MS et AltProts

Notre objectif dans la mise en évidence des AltProts est d'observer de manière globale les interactions présentes dans la cellule en y impliquant les AltProts. Les AltProts représentent la partie cachée du protéome de la cellule. Afin d'obtenir un maximum d'information vis-à-vis de ce protéome fantôme nous recherchons des stratégies protéomiques à large échelle pour observer aussi largement que possible les modifications d'interactions dans différentes conditions physiologiques sur l'ensemble du protéome. Ainsi les stratégies XL-MS sont-elles parfaitement adaptées à l'objectif recherché. Cependant, le choix des agents pontant et des méthodes d'enrichissement doit être réfléchi en prenant en compte les particularités des AltProts. Ainsi, la mise en évidence des AltProts au sein des cellules et la recherche de leurs partenaires RefProts nécessitent l'interrogation des deux banques de données en simultanée. La compilation de ces deux banques peut atteindre jusqu'à 200.000 protéines. Compte-tenu de la taille des banques, il est impératif d'utiliser un pontage clivable en MS.

3. XL-MS et futurs développements ?

De nombreux développements et améliorations ont vu le jour ces dernières années afin de développer et rendre accessible cette méthode, historiquement réservée à l'étude de la conformation en MS. Le type d'agent pontant par sa taille et ses groupements fonctionnels, permet aujourd'hui de s'adapter à n'importe quelle problématique. Quelques barrières restent encore à franchir, concernant les méthodes d'enrichissement, d'analyse et d'interprétation, toutefois la communauté utilisant et développant cette technologie s'agrandie de plus en plus. Si aujourd'hui beaucoup se focalisent sur la technologie, bientôt les limites apparaîtront sur l'application notamment avec la mise en place de pontage *in vivo, in cellulo* ou *in situ* pour des applications cliniques [128].

PARTIE III

Optimisation des Stratégies Protéomiques pour l'Identification des AltProts

Le premier objectif de mon travail a consisté à mettre en place une méthodologie permettant l'identification à grande échelle des AltProts par protéomique. Les AltProts ont en moyenne un poids moléculaire bien inférieur à celui des RefProts. Une optimisation de la préparation des échantillons pré-analyse est donc nécessaire. Avec une taille moyenne de 50 acides aminés les AltProts ne sont ni extraites, ni enrichies de la même manière que les RefProts. Ainsi, une partie de ces protéines peut être perdue lors de l'utilisation des méthodes conventionnelles d'extraction protéique et/ou de purification. De plus, lors de l'étape de digestion la petite taille de ces protéines ne permet pas toujours de produire un nombre important de peptides de digestion. Par conséquent les règles d'analyses basées sur des paramètres d'identification stricts habituellement préconisées pour les analyses protéomiques, telles que deux peptides minimums par protéine dont au moins un peptide unique, sont alors difficilement applicables dans le cas des AltProts.

I. Enrichissement en AltProt

Les AltProts de par leurs caractéristiques, peuvent être enrichies de manière à augmenter leur abondance et donc leur détection en analyse MS. Pour ce faire, différentes approches existent et sont basées sur les caractéristiques physico-chimiques des AltProts. En effet, étant de faible poids moléculaire, les techniques couramment utilisées en peptidomique peuvent alors être appliquées incluant l'extraction en milieu acide et à l'eau bouillante. Également, les techniques conventionnelles d'extraction de protéines peuvent être appliquées en y incluant des étapes de précipitation ou de fractionnement en gel de polyacrylamide afin de récupérer les protéines de faible poids moléculaire.

1. L'enrichissement par limite de taille

Par la taille généralement observée de ces petites protéines, il est possible d'envisager l'utilisation d'une limite de taille notamment par l'utilisation d'un filtre de porosité contrôlée tel que les filtres AMICON ayant une porosité de 30 kDa. Le préfractionnement via l'usage d'un filtre 30 kDa permet de soustraire toute

protéine ayant une taille supérieure au sein de l'échantillon. De plus cette méthode couplée à une modification spécifique des cystéines a permis de mettre en évidence 16 nouvelles AltProts, également nommées dans ce cas les « *cysteine-containing human sORFencoded polypeptides* » (ccSEPs) [129]. Cette méthode permet de cibler une famille d'AltProts. Cependant l'utilisation de filtres de 30 kDa a toutefois montré ses limites en termes de quantité d'AltProts enrichies, comparativement à d'autres méthodes.

2. Enrichissement par précipitation

Un des paramètres nécessaire pour la séparation des protéines est la précipitation. En effet en milieu acide les protéines les plus grosses sont les plus chargées. Elles peuvent alors former des complexes et être précipitées par centrifugation douce, laissant ainsi libres en solution les protéines de plus bas poids moléculaire telles que les AltProts et les peptides. La précipitation acide la plus répandue en protéomique utilise l'acide trichloroacétique (TCA). Toutefois cette précipitation peut également se réaliser avec des acides tels que l'acide acétique. L'utilisation d'acétonitrile (ACN) couplée à des sels permet également de changer l'environnement des protéines. L'ajout de ces sels permet d'augmenter la force ionique du milieu et par conséquent de masquer les charges de surface des protéines provocant ainsi leur agglomération et les rendant donc moins solubles. Elles précipitent alors par centrifugation douce laissant libres en solution les petites protéines ayant une masse inférieure à 15kDa [62].

3. Enrichissement par extraction sur phase solide

Une troisième approche est de retenir les protéines sur une colonne fonctionnalisée avec, par exemple, des chaines 8 atomes de carbone (C8) afin de retenir les protéines les plus grandes, possédant donc un pouvoir de rétention plus important sur ce type de colonne. Des colonnes C18, avec 18 atomes de carbone ont un pouvoir de rétention des protéines plus important et peuvent donc également être envisagées. Ces méthodes de pré-purification sur phases solides (SPE), montrent des résultats d'enrichissement des petites protéines assez efficaces [61].

II. Objectif

L'enrichissement est une étape cruciale dans la mise en évidence des AltProts. L'utilisation des diverses méthodes présentées précédemment est indispensable à la caractérisation de l'ensemble des AltProts d'un mélange complexe. C'est pourquoi j'ai choisi de comparer ces différentes méthodes d'extraction : extraction au MeOH et à l'eau bouillante connues pour l'extraction des peptides et des neuropeptides [64]. D'autres méthodes ont été envisagées comme les extractions aux détergents tels que le RIPA et le SDS combinées à une filtration sur Amicon 30kDa, une séparation par électrophorèse ou encore avec une précipitation à l'acide acétique et au TCA. Chacune de ces approches nous permet d'obtenir un résultat en complétant un autre. Si certaines méthodes permettent de mettre en évidence un plus grand nombre d'AltProts, chacune permet d'en identifier de manière spécifique.

Dans un tel contexte, j'ai appliqué ces différentes méthodes d'enrichissement sur un cas biologique, le gliome de grade 4. Pour ce faire, j'ai utilisé la lignée humaine de gliome NCH82 à disposition au laboratoire. J'ai pu ainsi mettre en évidence les limites de chacune des approches et les meilleures conditions pour avoir accès au protéome fantôme. Ces travaux font l'objet de l'article présent ci-dessous.

Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins

Tristan Cardon[1]*, Flore Hervé[1]*, Vivian Delcourt[1,2], Xavier Roucou[1,2], Michel Salzet[1], Julien Franck[1]** and Isabelle Fournier[1]**

[1]Université de Lille, Inserm, U1192 - Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM), F-59000 Lille, France

[2]Department of Biochemistry, Université de Sherbrooke, Quebec, Canada

ABSTRACT: Large scale proteomic strategies rely on database interrogation. Thus, only referenced proteins can be identified. Recently, Alternative Proteins (AltProts) translated from non-annotated Alternative Open reading frame (AltORFs) were discovered using customized databases. Because of their small size which confers them peptide-like physico-chemical properties, they are more difficult to detect using standard proteomics strategies. In this study, we tested different preparation workflows for improving the identification of AltProts in NCH82 human glioma cell line. The highest number of identified AltProts was achieved with RIPA buffer or boiling water extraction followed by acetic acid precipitation.

A functional open reading frame (ORF) or coding sequence (CDS) is a continuous stretch of codons that begins with a start codon (usually AUG) and ends with a stop codon (UAA, UAG or UGA)[1]. In conventional genome annotations in eukaryotes, a single ORF per coding gene or coding transcript is annotated, generally the longest one. Moreover, genes or transcripts with no ORFs longer than 100 codons are annotated as non-coding genes or RNAs (ncRNAs), unless the corresponding small protein coded by such small ORF was previously characterized [2]. Yet, recent proteogenomics, including ribosome profiling and proteomics with customized protein databases have demonstrated that these rules prevent the detection and functional characterization of novel proteins, particularly small proteins. Indeed, mRNAs may harbour several functional ORFs in addition to the currently annotated CDS, and ncRNAs may also contain functional ORFs coding for proteins smaller than 100 amino acids[3–8]. Here, we define as alternative ORF or AltORF any unannotated coding sequence from any reading frame of an mRNA or an allegedly ncRNA. Within mRNAs, AltORFs are found upstream, downstream, or overlapping CDSs in a different reading frame[9] (**Figure1**).

Proteins translated from AltORFs are termed alternative proteins (AltProts) and constitute the so called ghost proteome as it can be deciphered only with customized protein databases[9]. For clarity, canonical proteins annotated in current protein databases such as UniProtKB are termed reference proteins (RefProts). Since AltORFs are obligatorily smaller than canonical CDSs, AltProts are also smaller than RefProts. In human, the median size of predicted AltProts is 45 amino acids compared to 460 for RefProts[5]. However, in contrast to small proteins, microproteins or small-encoded peptides (SEPs)[10], AltProts can be longer than 100 amino acids. Hence, AltProts include proteins both smaller and larger than 100 amino acids, but the majority are less than 100 amino acid long.

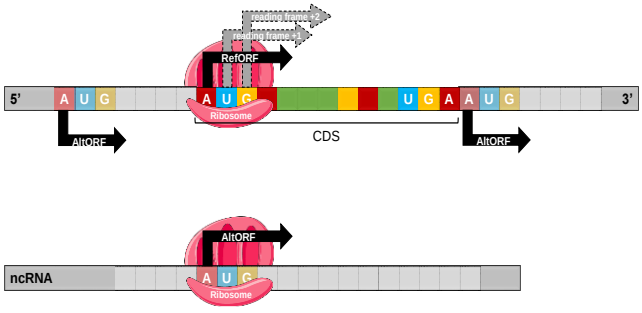


Figure 1. Diagram of the translation of alternative proteins (AltProts). AltProts are coded by AltORFs that may be present within an mRNA or an RNA annotated as ncRNA. Top: AltORFs may localize in 5' or 3' UTRs, or overlap the CDS (or RefORF) in a different reading frame. Bottom: AltORFs may also be present within ncRNAs. RefProts are translated from annotated RefORFs and AltProts are translated from AltORFs.

Mass spectrometry (MS) based proteomics with customized protein databases has emerged as a powerful tool to detect small proteins[4]. This strategy is advantageous compared to ribosome profiling since it allows the detection of the final translation product. The shotgun approach is certainly the widely used technique in large scale proteomics analyses. The first identifications of AltProts by means of large-scale MS based proteomics was described by Vanderperre et al. 2013[11]. Recently, it was shown that a proteogenomics approach lead to a significant improvement of AltProts identification by the development of a customized RNAseq-based database containing the amino acids sequence of SEPs[7]. Example of this strategy was presented by Slavoff *et al*. They were able to identify 90 human SEPs from K562 cell line[7] while Oyama et al. were able to identify only 4 SEPs from K562 cell line[12]. The refseq protein and mRNA database was used to compare the MS/MS spectra demonstrating the importance of a well annotated database for the investigation of new functional SEPs. Delcourt *et al*. have

demonstrated by targeted absolute quantitative MS based proteomics that the main translational product of MIEF1 is not the canonical MiD51 protein but the AltProt AltMiD51[13]. Altogether, these studies confirm that the single CDS dogma for mRNAs and the coding potential of alleged "non-coding" RNAs needs to be reevaluated.

Even if proteogenomics offers new possibilities to identify novel AltProts[14], conventional large scale proteomics analysis can suffer from a lack of sensitivity towards small proteins. Indeed, considering the shotgun strategy, proteins are identified by taking into account at least two peptides including one unique peptide. However, AltProts may be smaller than 50 amino acid long and the number of enzymatic peptides generated from AltProts may be very low compared to conventional proteins. Recently, the top down proteomics approach was shown as a very effective technique for the detection of AltProts[4,15]. The technique offers many advantages including the detection of proteins into their intact or truncated forms and the possibility to detect proteoforms[16]. Considering that some AltProts produce few or no proteolytic peptides, their identification by shotgun proteomics approach is problematic and therefore the top down strategy can overcome this major drawback. A recent spatially resolved top-down proteomics study resulted in the identification of 8 AltProts from specific regions of rat brain tissue sections including the hippocampus, the corpus callosum and the medulla oblongata[10]. Among these AltProts, one is translated from a transcript annotated as a non-coding RNA, illustrating the importance of updating the annotation of transcripts. Similarly, 15 novel AltProts were detected in ovarian cancer, including 5 in the tumor region[4]. One of them, AltGNL1, is encoded in an ORF overlapping the CDS in a frameshifted reading frame. This small AltProt translated from a gene described to be involved in the human major histocompatibility complex class I region and potentially affecting the immune response, could not have been identified using a conventional bottom-up strategy. Small proteins are involved in a variety of functions[17], including in physiopathological processes such as cancer[4] or neurodegenerative diseases[18].

Whether a bottom-up or a top down strategy is used, the detection and identification of AltProts remains complex considering their low abundance and their small size. In this context, the implementation of a low molecular weight protein enrichment step prior to proteomics analyses is of great importance[19]. For example, Schwaid et al. used a 30Kda filter followed by a cysteine enrichment to improve the detection of novel cysteine containing SEPs (ccSEPs) and 16 novel ccSEPS from K562 cell line were identified[20]. They compared enrichment procedures including acid precipitation, 30kDa cut off filter and C8 SPE cartridge. These enrichment steps were coupled to different SEPs extraction methods such as boiling water, lysis buffer, acetic acid extraction and HCl (1N) extraction. Thus, the combination of lysis buffer and C8 SPE gave the best recovery of SEPs from A549 cell line. Cassidy et al. enriched low molecular weight proteins using a GelFree system allowing the fractionation of intact proteins and their recovery in a condensed phase[21]. Low mass recovered proteins are then subjected to enzymatic digestion prior to their analysis by LC-MS allowing the identification of 17 novel SEPs from the archaea *Methanosarcina mazei*. The same group demonstrated that an acetonitrile-based precipitation method depleted proteins above 15kDa and resulted in a better identification of small proteins including SEPs from *Methanosarcina mazei*[22].

Thus, AltProts analyses still require improved sample preparation protocols to enrich small proteins prior to their LC-MS analysis. Here, we tested several protocols to optimize the detection of AltProts in the human NCH82 glioblastoma cell line. This cell line was selected considering the high potential of AltProts as novel biomarkers. Different extraction methods coupled to small proteins or peptides enrichment strategies were then tested and compared to establish the best protocol for the detection of AltProts.

MATERIAL AND METHODS

Reagents: Dulbecco's modified Eagle's medium (DMEM), fœtal bovin serum (FBS), L-glutamine, penicillin, streptomycin, phosphate-buffered saline (PBS) were obtained from Thermo Fisher Scientific (Les Ulis, France). Formic acid (FA), HPLC grade water, trifluoroacetic acid (TFA), acetonitrile (ACN) methanol (MeOH), ethanol (EtOH), acetone and trichloroacetic acid (TCA) were purchased from Biosolve BV (Dieuze, France). DL-dithiothreitol (DTT), iodoacetamide (IAA), chloroform, dimethylsulfoxyde (DMSO), amonium bicarbonate (AmBic) were obtained from Sigma Aldrich (Saint-Quentin Fallavier, France). Tris and SDS were purchased from Bio-Rad (Steenvoorde, France). Trypsin was obtained by Promega (Charbonnières-les-Bains, France).

Cell culture. Human NCH82 stage IV glioma cells stage IV were supplied by obtained from Régnier Vigouroux and grown at 37°c under an atmosphere of 5% CO2. The cells were grown in high glucose Dulbecco's Modified Eagle's Medium (D-MEM, Thermofisher), supplemented with 10% fetal bovine serum and 100 U/ml penicilin/streptomicin antibiotic. For AltProts extraction, the cells were grown in cell culture plate 6 wells. Two wells were used for protein extraction.

Protein extraction methods. Prior to cell lysis and enrichment of AltProts, culture plates were placed on ice, media were removed from all the plates and the cells were washed twice with Dulbecco's phosphate-buffered saline DPBS 1x Mg^{2+}, Ca^{2+} . The AltProt extraction was performed scraping the cells from the wells with 150 µL of extraction buffer. The extracts were then collected into 1.5 mL low binding tubes. . Four different extraction buffers were compared from 1.8×10^6 cells/mL. Three different methods were used for peptides and proteins extraction: **(1)** SDS 4% buffer (SDS 4%, Tris-HCl 0.1 M, DTT 0.1 M, pH: 7.6), **(2)** RIPA lysis buffer (150 mM NaCl, 50 mM Tris, 50mM EGTA, 2 mM EDTA, NP40, 100 mM sodium pyrophosphate, IPEGAL 1%), **(3)** MeOH acid buffer (1:9:90 acetic acid /H_2O/MeOH). Samples were incubated 30s in liquid nitrogen, then 30 s in boiling water. This was repeated 4 times. Samples were then sonicated at level 3 for 20 bursts and placed on ice between every 5 pulses for 30 s. The homogenate was then centrifuged at 17,000 g for 20 min at 4°C The supernatant was collected and directly used, or dried under vacuum and stored at -80°C The fourth method was investigated only for peptides extraction: **(4)** Boiling Water. The extracts were washed three times with PBS by centrifugation at 1,000 g for 1 min, pelleted and then stored at -80°C. Boiling water (500 µl) was directly added to the frozen pellet and the sample was then boiled for 20 min to stop the proteolytic activity and maintain the integrity of the peptidome. Once the samples were brought to room temperature (RT), they were sonicated on ice for 20 bursts. Acetic acid in water (1:4, v/v) was added to the cell lysate and the samples were centrifuged at 17,000 g for 20 min at 4°C. The supernatant was directly used or dried under nitrogen and stored at -80°C.

AltProts enrichment methods. Three AltProts enrichment methods were compared: **(1)** Gel fractionation, **(2)** acetic acid precipitation, **(3)** trichloroacetic acid precipitation. **(1)** Gel fractionation: Stored protein pellets were solubilized in Laemmli1X

buffer and then loaded onto a 4-12 % acrylamide of SDS-PAGE gel. Proteins were separated at 70V for 20 minutes and at 170 V for 60 min. The gel was stained with Instant blue (Expedeon) and cut into 5 pieces below the 50 kDa marker as follows: 50-40, 40-25, 25-15, 15-10, 10-1 kDa **(2)** Acetic acid precipitation (AA): AA in water (1:4, v/v) was added to the supernatant followed by the centrifugation at 15,000 g for 20 min at 4 °C. This step precipitates larger proteins to reduce the complexity of the supernatant and enriches low molecular weight proteins. The supernatant was then collected. **(3)** Trichloroacetic Acid (TCA) precipitation: 10 volumes of cold acetone/TCA. (9:1, v/v) were added to the supernatant. The solution was mixed and stored overnight at -20°C followed by a centrifugation at 15,000 g for 10 min. The supernatant was removed, and 1 volume of acetone was added to the pellet, mixed and then stored for 10 min at -20°C. After the additional centrifugation for 5min at 15,000 g the pellets and the supernatant were split in two samples, and dried under vacuum.

Digestion and sample preparation for nLC-MS/MS. For the gel fractionation method (1), each gel slice was washed with 300 ml of distilled deionized water for 15 min, ACN for 15 min, and 100 mM NH_4HCO_3 for 15 min followed by a mix of NH_4HCO_3/ACN (1:1, v/v) for 15 min and ACN for 5 min. Gel slices were vacuum dried for 5 min. The reduction of cysteine residues was performed with 10 mM DTT in 50 mM NH_4HCO_3. Gel slices were incubated at 56°C for 30min. The alkylation of cysteines was performed with 50 mM of IAA and 100 mM NH_4HCO_3. After incubation at RT in the dark for 15 min, gel slices were washed a second time and vacuum dried for 5 min. Digestion was then performed at 37°C overnight with 20µg/mL trypsin in 50 mM NH_4HCO_3. Peptides were extracted twice on a shaking platform with 1% FA for 20 min each; then 150 mL of ACN for 10 min. The supernatant was transferred into a new tube and vacuum dried. For enrichment methods (2) and (3), the pellets were directly reduced by the addition of 20 µL of 50 mM DTT in 100mM NH_4HCO_3, incubated for 30 min at 37°C then alkylated with 20µL of 50 mM IAA in 100 mM NH_4HCO_3 for 15 min at RT in the dark. The samples were digested overnight using 10 µL of trypsin (20µg/ml) in 50mM NH_4HCO_3. The digestion was stopped by the addition of 1% FA.

LC MS analysis. The preparation for the nanoLC-MS/MS analysis was similar for all extracts. Dried samples were re-suspended with 20 µL of 0.1% TFA and desalted on a C18 Ziptip (Millipore,Saint-Quentin-enYvelines, France). The samples were then vacuum dried and finally re-suspended in ACN/ 0.1% FA (2:98, v/v).The tryptic peptides were separated with a nanoAcquity (Waters) chromatography equipped with a C18 pre-column (180 µm × 20mm, 5µm DP, Waters) and BEA peptide column (25 cm, 75 µmID, 1.7 µL DP, Waters) using a gradient of ACN from 5% to 30% in 2H at 300 nL/min. The tryptic peptides from gel fractionation were separated using a gradient of ACN from 5% to 30% in 70 min at 300 nL/min. A Thermo Scientific Q-Exactive mass spectrometer was used for MS acquisition. The instrument was set to acquire the ten most intense precursors in data-dependent acquisition mode, with a voltage of 2.8 kV. The survey scans were set to a resolving power of 70 000 at FWHM (*m/z* 400), in positive mode in a scan range of 300 to 1600 *m/z* and using an AGC target of 3E+6. For the MS/MS, 1 microscan was obtained at 17,500 FWHM and dynamic exclusion was enabled. The instrument was set to perform MS/MS only from >+2 and <+8 charge states.

Data analyses. RAW data obtained by nanoLC-MS/MS analysis were analyzed using Proteome Discoverer V2.2 (Thermo Scientific) with the following parameters: Trypsin as enzyme, 2 missed cleavages, methionine oxidation as variable modification and carbamidomethylation of cysteines as static modification, Precursor Mass Tolerance: 10 ppm and Fragment mass tolerance: 0.6 Da. The validation was performed using Percolator with a FDR set to 1%. A consensus workflow was then applied for the statistical arrangement, using the high confidence protein identification. The protein database was uploaded from Openprot (https://openprot.org/) and included RefProt, novel isoforms and AltProts predicted from both Ensembl and RefSeq annotations (GRCh38.83, GRCh38.p7).

RESULTS

Methods for AltProts identification. Different methods of extraction (Figure 2) were investigated to improve the detection and identification of AltProts. Some are widely used in proteomic strategies and others are known to be more suitable for peptidomics or neuroproteomics[23,24]. Considering that AltProts are rather small proteins, they are expected to display physicochemical features similar to peptides.

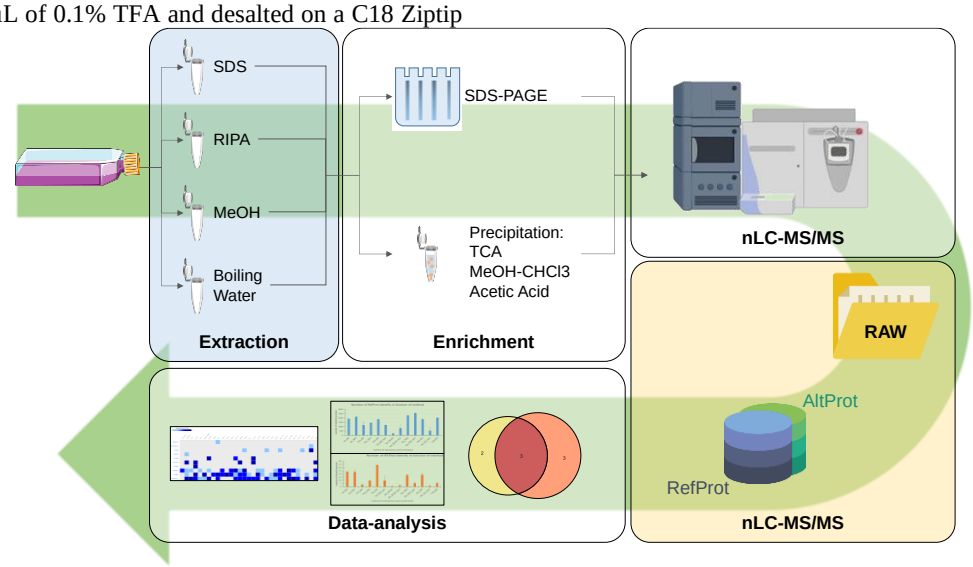


Figure 2. Designed workflow including various extractions and enrichment methods for the analysis of AltProts. Combination of methods increase the possibility to identify AltProts from a complex sample.

As summarized in the experimental workflow designed for these experiments (**Figure 2)**, we tested four extraction methods followed by three enrichment methods in order to improve the enrichment of the lower molecular weight proteins were also implemented and compared. The extraction methods included boiling water (BW), RIPA or SDS 1% solubilization, and methanol (MeOH) extraction. The enrichment methods included SDS-PAGE fractionation, acetic acid and trichoroacetic acid precipitation. An in-gel digestion in the stacking gel was used as reference for all extracting methods.

Identified protein as a function of extraction and enrichment methods. Comparison of the different extraction methods (**Supplementary Figure 1A**) shows that protein extraction with RIPA buffer provided the highest number of identified RefProts with 3,036 identification. Protein extraction with MeOH and SDS were much less effective with 1,429 and 2,381 RefProts identifications, respectively. When combining RIPA extraction with an enrichment step, a mean of 2292 (mean standard deviation = 55) proteins were identified after gel fractionation, 1445 (mean standard deviation = 69) after TCA precipitation, and 2143 (mean standard deviation = 25) after AA precipitation. For RIPA extraction, further enrichment by precipitation or fractionation did not improve the number of identified proteins. In contrast, the enrichment post-BW extraction really improved the number of identified RefProts with an average increase of 1.5 except for TCA for which the number of identification remained unchanged. The enrichment step led to a decrease in RefProts identifications, likely because they display a wide-range of molecular weights and thus of physico-chemical features, and none of the tested methods may be optimal for all of them. For AltProts, the results were clearly different (**Supplementary Figure 1B**). The best extraction method without fractionation was BW with a mean of 10 (mean standard deviation = 7) proteins identified compared to RIPA which only leads to about a half of identified proteins. In contrast to RefProts, the enrichment step resulted in a significant improvement in AltProts identification. This observation may be explained by the narrower spectrum of physico-chemical properties of the AltProts compared to the RefProts. For BW extraction, >3 folds more proteins are identified after gel fractionation and >2 folds for the AA precipitation. For RIPA extraction, a mean of 6 (mean standard deviation = 0.4) proteins only were identified with no enrichment. This number increased by >3 folds after gel fractionation and >4 folds after AA precipitation. In general, TCA precipitation did not perform well as an enrichment method. MeOH extraction resulted in poor identification both for RefProts and AltProts. SDS extraction alone was also not very efficient unless it was used in combination with gel fractionation. Then, the number of identified AltProts is multiplied by more than a factor 10. Thus, RIPA, SDS and BW extraction followed by gel fractionation or AA precipitation are the most efficient strategies to improve the identification of AltProts. This is clearly illustrated by comparing the percentage of identified AltProts with respect to identified RefProts (**Supplementary Table 1**).

Molecular weight distribution of identified protein. We further analyzed the efficiency of the various methods for small proteins in general compared to large proteins using a 15kDa identification cutoff and determining the ratio of identified RefProts below and above 15kDa (**Table 1**). BW extraction without fractionation resulted in the best ratio (9.5%) (**Table 1A**). RIPA/Gel fractionation and SDS/Gel fractionation also resulted in a significant ratio of 8.23% and 7.42%, respectively. Then we compared AltProts to RefProts <15kDa in order to compare proteins that bear similar physico-chemical properties with respect to extraction and enrichment. This shows that up to 75.39% and 71.54% of AltProts are identified for RIPA and BW extraction respectively both using AA enrichment (**Table 1B**). Most of the other extraction and enrichment provide less than 10% AltProts identification. Most of identified AltProts are under 15 kDa ranging from 8 to 14 kDa with MeOH extraction retrieving the lower MW proteins (<8 kDa) and the SDS the highest MW (<14 kDa) in line with their physico-chemical properties (**Supplementary Figure 1C**). For the AA precipitation with BW and RIPA extraction, 14 AltProts are common to these 2 conditions within all the replicates and not found with the TCA precipitation, 13 being exclusive to the BW and 9 to the RIPA. The fact that only 4 protein in total are common to the 2 extractions (RIPA, BW) and the 2 enrichment (AA, TCA) demonstrates the importance of choosing the right combination of extraction and enrichment. On the total of 52 AltProts identified, 36 are found for AA combining RIPA and BW which represents about 70% of AltProts (**Supplementary Figure 1D**).

Table 1. Ratio of the small proteins (≤ 15 kDa) to the total proteins. **(A)** ratio between ≤15 kDa RefProts and total RefProts and **(B)** ratio between ≤15 kDa AltProts and total RefProts.

A		Mean ratio RefProts<15kDa/RefProts in percent	Associated Mean Standard deviation
extraction	enrichment		
RIPA	NoF	2.94	0.19
	Gel	8.23	0.43
	AA	2.74	0.25
	TCA	6.79	0.71
BW	NoF	9.51	0.35
	Gel	7.37	0.16
	AA	2.86	0.20
	TCA	4.68	1.29
MeOH	NoF	4.35	0.86
	Gel	4.07	1.83
SDS	NoF	2.83	1.50
	Gel	7.42	0.67
B		Mean ratio AltProt/RefProt<15kDa in percent	Associated Mean Standard deviation
extraction	enrichment		
RIPA	NoF	7.81	0.85
	Gel	9.03	0.84
	AA	75.39	16.41
	TCA	7.20	0.51
BW	NoF	6.91	4.92
	Gel	21.72	0.63
	AA	71.54	3.92
	TCA	6.51	4.36
MeOH	NoF	3.89	2.59
	Gel	6.86	1.68
SDS	NoF	38.10	41.27
	Gel	10.72	1.12

We further analyzed the distribution of MW of AltProts in the different experimental conditions **(Figure 3**). Overall, the proportion of AltProts >20 kDa was very low (**Supplementary Figure 2**) but could be improved with an enrichment step. The best methods for the identification of AltProts >20 kDa were SDS/Gel fractionation (16%) and RIPA/Gel fractionation (13%) (Figure 3). The best methods for the identification of AltProts with a MW in the 15-20 kDa range and 10-15 kDa range were RIPA/Gel fractionation (19%) and BW/Gel fractionation, respectively (Figure 3). A large fraction ofthe identified AltProts were below 10 kDa and the best methods were BW/AA or BW/TCA.

With respect to reproducibility, BW extraction with gel fractionation gives 7.9% variation in the identified RefProts between the triplicates. 8.4% is observed for RIPA with gel fractionation but up to 29% for MeOH extraction with gel fractionation (**Supplementary Table 2, Supp Data 1**). Thus, the reproducibility is overall good for all the extractions and enrichments expect for the MEOH extraction. For AltProts, the reproducibility in the proteins identified is clearly less and does not show any peculiar trend in relation to the extraction or enrichment method used (**Supplementary Data 4**). The low number of identified AltProts and the lower reproducibility in the identification can be related to the dynamic of translation of AltProts and their susceptibility to enzymatic degradation similarly to neuropeptides as recently demonstrated[10]. Without enrichment the variation rate of AltProts is similar to what was previously found, confirming that AltProt are less stable than RefProts (**Supplementary Table 2, Supplementary Data 1**). Nevertheless, some AltProts such as IP_584395, IP_587085, IP_595471 and IP_759887 are identified in any case independently of the enrichment procedure used (**Supplementary Table 3**). Among the 89 identified AltProts, 1.2% are common to all of the 8 extraction/enrichment combination, 3.3% to 7 and 6 them, 2.2% to 5 and 4, 16.8% to 3 and 15.7% to 2 and 55.3% in a single enrichment procedure (**Supplementary Table 3**). Three AltProts (IP_591792, IP_592880 and IP_734708) were commonly identified to any extraction with gel fractionation and precipitation enrichments; 4 (IP_655967, IP_602798, IP_557834, IP_572421) only for precipitation and 9 ((IP_688853, IP_624545, IP_689114, IP_592855, IP_662403, IP_593099, IP_691726, IP_723386, IP_737334) only for gel fractionation strategy. This establish the fact that the physicochemical properties of some AltProts make them easily identifiable independently of the extraction/enrichment methods when others are clearly only observed for certain extraction or enrichment condition. Examination of the isoelectric point (IP) of identified AltProts show a specific distribution according to the enrichment method (**Figure 4**). Indeed, for AA precipitation the average IP is >8, around 7-8 for the gel fractionation and 5-6 for the TCA extraction. This clearly demonstrates that it is not possible to get the whole panel of AltProts by a single extraction/enrichment combination although by combining BW and RIPA it is possible to retrieve all AltProts present in the sample.

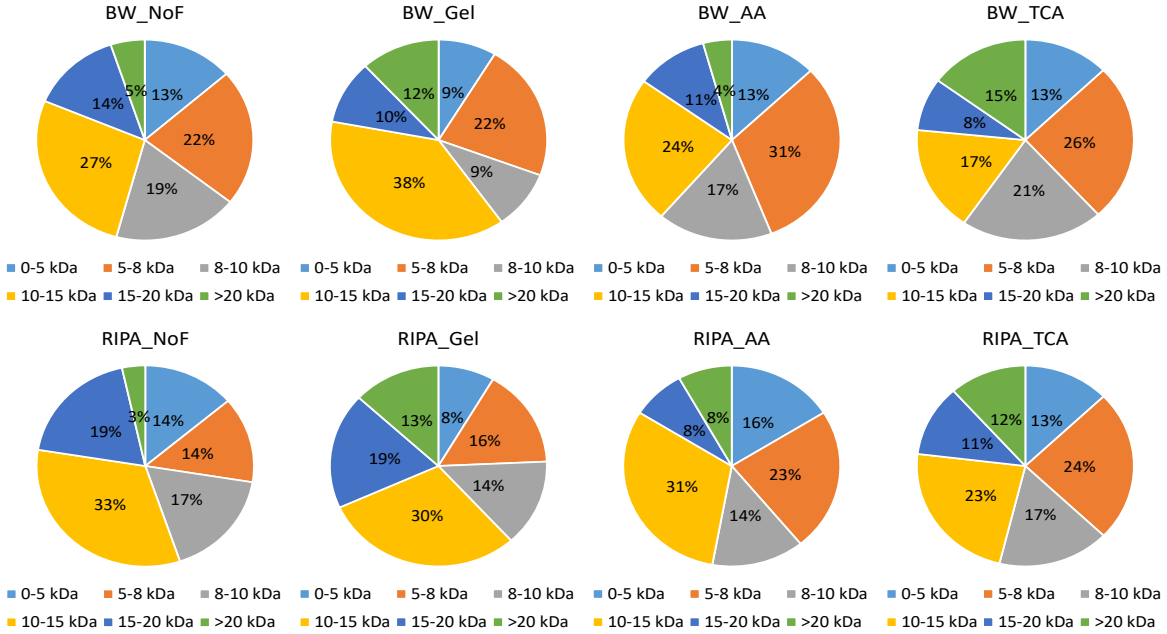


Figure 3. AltProts MW distribution. Identified AltProts with a MW below 20 kDa are largely overrepresented compared to AltProts above 20 kDa independently of the methods used.

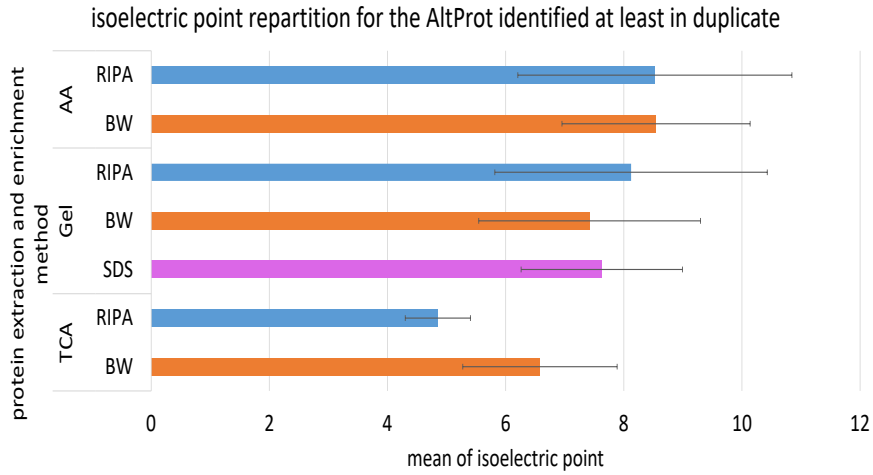


Figure 4. Mean of isoelectric point of AltProt according to protein extraction and enrichment methods. Repartition show the possibility to separate different isoelectric propriety in function of methodology.

Characteristics of the identified AltProts. AltProts were found to be specific to both the extraction and enrichment steps. Interestingly, for gel fractionation and AA precipitation, shared AltProts identification show to be translated from pseudogenes-derived lncRNAs (**Supp. Table 1**)

DISCUSSION

We have tested different methods to extract and enrich AltProts prior to their identification by MS. Overall, the best methods combines BW or RIPA extraction with gel fractionation or AA precipitation. This is interesting considering that hot or boiling water in acidic condition is widely used in the field of peptidomics, more specifically for neuropeptides analysis[23–26]. The majority of AltProts are small proteins and may they display physico-chemical features similar to peptides. In agreement with physicochemical features of peptides, we also observed that RIPA and BW extractions followed by AA precipitation provide the highest recovery of AltProts. In contrast, MeOH extraction neither enhanced the identification of AltProts or RefProts (**Figure 3**). We observed that some AltProts can be identified with all methods (**Supplementary Table 3**). These results indicate that some AltProts can be recovered with different methods whilst others require specific extraction methods.

Reproducibility of identification is clearly an issue for AltProts, as already observed for the identification of peptides compared to proteins [27]. Even if the number of AltProts identified generally improves after BW and RIPA extraction followed by AA precipitation, reproducibility remains low compared to RefProts. This significantly decreases confidence in identification and three main reasons may be put forward: i) posttranslational processing into bioactive peptides; ii) stability; and iii) physico-chemicals parameters. Indeed, peptides such as neuropeptides are generally processed into smaller peptides by aminopeptidases and endopeptidases prior to binding specific receptors and inducing a unique biological response[28]. Hence, AltProts may also be postranslationally processed to generate biologically active small peptides. The detection of such cleaved alternative peptides issued from AltProts would require a top-down strategy. However, peptide identification remains difficult due to a lack of libraries based on experimental observation completed by *in silico* prediction deduced from enzymatic cleavage sites. Some AltProts directly interact with RefProts, including ribosomal proteins[29] and other cytoplasmic and membrane complexes[17]. As regulators of RefProts, we speculate that AltProts may be relatively labile and quickly degraded.

Most of the identified AltProts from the NCH82 glioblastoma cell line are encoded in alleged "lncRNA" and only a few from mRNA despite that AltProts can be issued from different classes of transcripts including mRNA, lncRNA and even pri-miRNAs, (**Supplementary Table 3, 4**). Among those, a majority are encoded in pseudogenes-derived lncRNAs, including three AltProts that were detected independently of the extraction method. Pseudogenes are considered non-functional genes [30]. Some pseudogenes termed unprocessed pseudogenes result from gene duplication events and accumulate inactivating mutations, yet they still possess the regulatory elements to promote transcription and are transcribed into RNA. As such, 16,840 and 1474 pseudogene-derived ncRNAs are currently annotated within Ensembl and RefSeq databases, respectively. Expression of specific pseudogenes in diseases may provide useful signatures for diagnosis and prognosis [31]. In particular, several pseudogenes capable of predicting survival in lower grade glioma patients were recently identified [25,32]. In a first large human proteome catalogue, 107 pseudogenes-derived proteins were detected [33]. This result is in agreement with our data and with the observation that many pseudogenes are translated [5,34]. Among the AltProt-derived pseudogenes identified in this study, five of them (IP_591792 AltACTBP8, IP_557834 AltACTBP1, IP_737334 AltACTBP7, IP_593099 AltTUBB2BP1 and IP_572421 AltTUBBP1) were already described in the literature[35] and are expressed in different cell lines referenced in the Expression Atlas repository of gene expression[36]. For other AltProt-derived pseudogenes, very little information is available. We could retrieve information for 10 of the pseudogenes encoding AltProts detected in our study in the Expression Atlas repository. The most represented in glioma cells are AC008481.2 and ACTBP7 which were identified in 22 cell lines, and TUBB2BP1 which is found in 18 cell lines (**SuppData 3**). The AltProts we identified could be potential diagnostic or prognostic markers and could be combined to previously established pseudogenes signatures. It would be therefore interesting to further search for these proteins in the body fluids from patient samples. AltEDARADD (IP_079312) could be an interesting prognostic marker. AltEDARADD is encoded in a small ORF located in the 3'UTR of two EDARADD mRNA isoforms, EDARADD-202 and EDARADD-204 (**Figure 5**). Although AltEDARADD was not identified in all replicates in NCH82 cells, we previously identified the protein in biopsies from patients with high grade ovarian serous cancer[4]. AltEDARADD is now annotated by Uniprot with the accession number *L0R849*. Very interestingly, a high level of expression of AltEDARADD protein is associated with a poorer prognosis for patients suffering from high grade serous ovarian cancer (**Figure 6**). We hypothesize that AltEDARADD levels may also correlate with survival of GBM patients. Of note, EDARADD is a factor involved in the anti-viral immune response[37]. Both Glioma[38] and High grade serous ovarian cancer[39] were shown to be suspected of probable viral origin.

CONCLUDING REMARKS

This study has demonstrated that although reproducibility remains an issue in the detection of AltProts, these small proteins are better detected with traditional peptidomics preparation approaches such as BW or RIPA extraction followed by acidified water precipitation such as acetic acid. AltProts are regulators of translation, involved in the regulation of protein-protein interactions identified large fraction of detected AltProts are translated from pseudogenes-derived transcripts annotated as lncRNAs. This supports the new idea that the coding potential of pseudogenes has been overlooked. More interestingly, we have also identified AltEDARADD, a protein translated from the 3'UTR of *EDARADD* mRNAs. We previously detected this protein in biopsies of patients with high grade serous ovarian cancer and we show here that its levels correlate poor survival for patients with ovarian cancer. It would be interesting to investigate AltProts in body fluids as a source of new potential diagnostic and prognostic markers especially for GBM which suffers from a real absence of blood markers.

ASSOCIATED CONTENT

Supporting Information

Supplementary data 1 (Word File): extra figure of representing value and supplementary table1, 2, 3 and 4
Supplementary data 2 (Excel File) extra table
Supplementary data 3 (Excel File) extra table
Supplementary data 4 (Excel File) extra table

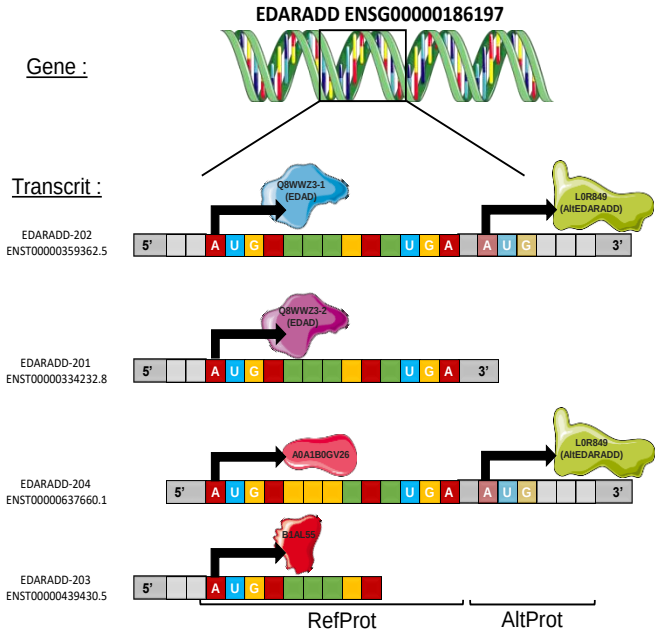


Figure 5. Representation of the AltEDARADD (IP_079312), now annotated in Uniprot under the accession number L0R849. Four major transcripts are shown; two of them display both the EDARADD coding sequence and AltEDARADD open reading frame.

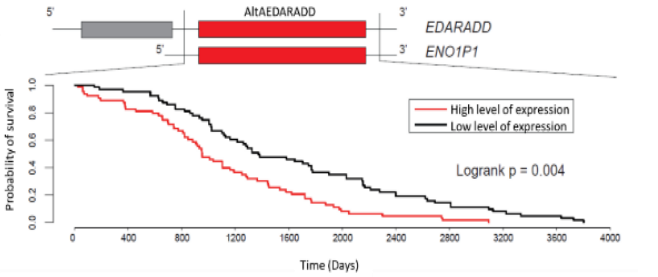


Figure 6. Survival analysis based on Kaplan-Meier representation associated to AltEDARADD expression from a cohort of ovarian cancer patients taken form the TCGA. AltEDARADD/ENOP1P open reading frames are represented in red.

AUTHOR INFORMATION

****Corresponding authors**

isabelle.fournier@univ-lille.fr and julien.franck@univ-lille.fr

ORCID : Isabelle Fournier: 0000-0003-1096-5044 & Julien Franck: 0000-0002-7443-1706

Author Contributions

*Authors TC and FH have equal contribution. Conceptualization, I.F., J.F.; Methodology, T.C., F.H., J.F. and I.F. Validation, T.C. and F.H.; Formal Analysis, I.F., J.F., M.S., F.H., T.C..; Investigation, I.F., F.H., J.F., T.C. and V.D. Resources, I.F. and M.S.; Data curation, I.F.,J.F., M.S., X.R., T.C. and F.H.; Writing – Original Draft, T.C., J.F., I.F. and M.S. ; Writing - Review & Editing, T.C., J.F., I.F., X.R.., and M.S. ; Supervision, I.F. and J.F..; Project Administration, I.F. and J.F.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This research was supported by funding from Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), Institut National de la Santé et de la Recherche Médicale (Inserm) and Université de Lille.

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga

REFERENCES

(1) Kozak, M. Regulation of Translation in Eukaryotic Systems. *Annual Review of Cell Biology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA November 28, 1992, pp 197–225. https://doi.org/10.1146/annurev.cb.08.110192.001213.

(2) Basrai, M. A.; Hieter, P.; Boeke, J. D. Small Open Reading Frames: Beautiful Needles in the Haystack. *Genome Research*. August 1, 1997, pp 768–771. https://doi.org/10.1101/gr.7.8.768.

(3) Mouilleron, H.; Delcourt, V.; Roucou, X. Death of a Dogma: Eukaryotic MRNAs Can Code for More than One Protein. *Nucleic Acids Research*. Oxford University Press January 8, 2016, pp 14–23. https://doi.org/10.1093/nar/gkv1218.

(4) Delcourt, V.; Franck, J.; Leblanc, E.; Narducci, F.; Robin, Y. M.; Gimeno, J. P.; Quanico, J.; Wisztorski, M.; Kobeissy, F.; Jacques, J. F.; et al. Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine* **2017**, *21*, 55–64. https://doi.org/10.1016/j.ebiom.2017.06.001.

(5) 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.; et al. Deep Transcriptome Annotation Enables the Discovery and Functional Characterization of Cryptic Small Proteins. *Elife* **2017**, *6*, e27860. https://doi.org/10.7554/eLife.27860.

(6) Brunet, M. A.; Roucou, X. Mass Spectrometry-Based Proteomics Analyses Using the Openprot Database to Unveil Novel Proteins Translated from Non-Canonical Open Reading Frames. *J. Vis. Exp.* **2019**, *2019* (146), 59589. https://doi.org/10.3791/59589.

(7) 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.; Saghatelian, A. Peptidomic Discovery of Short Open Reading Frame-Encoded Peptides in Human Cells. *Nat. Chem. Biol.* **2013**, *9* (1), 59–64. https://doi.org/10.1038/nchembio.1120.

(8) Aspden, J. L.; Eyre-Walker, Y. C.; Phillips, R. J.; Amin, U.; Mumtaz, M. A. S.; Brocard, M.; Couso, J.-P. Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq. *Elife* **2014**, *3*, e03528.

(9) Brunet, M. A.; Brunelle, M.; Lucier, J. F.; Delcourt, V.; Levesque, M.; Grenier, F.; Samandi, S.; Leblanc, S.; Aguilar, J. D.; Dufour, P.; et al. OpenProt: A More Comprehensive Guide to Explore Eukaryotic Coding Potential and Proteomes. *Nucleic Acids Res.* **2019**, *47* (D1), D403–D410. https://doi.org/10.1093/nar/gky936.

(10) Delcourt, V.; Staskevicius, A.; Salzet, M.; Fournier, I.; Roucou, X. Small Proteins Encoded by Unannotated ORFs Are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an MRNA. *Proteomics* **2018**, *18* (10), 1700058. https://doi.org/10.1002/pmic.201700058.

(11) Vanderperre, B.; Lucier, J.-F.; Bissonnette, C.; Motard, J.; Tremblay, G.; Vanderperre, S.; Wisztorski, M.; Salzet, M.; Boisvert, F.-M.; Roucou, X.; et al. Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS One* **2013**, *8* (8), e70698. https://doi.org/10.1371/journal.pone.0070698.

(12) Oyama, M.; Kozuka-Hata, H.; Suzuki, Y.; Semba, K.; Yamamoto, T.; Sugano, S. Diversity of Transplantation Start Sites May Define Increased Complexity of the Human Short ORFeome. *Mol. Cell. Proteomics* **2007**, *6* (6), 1000–1006. https://doi.org/10.1074/mcp.M600297-MCP200.

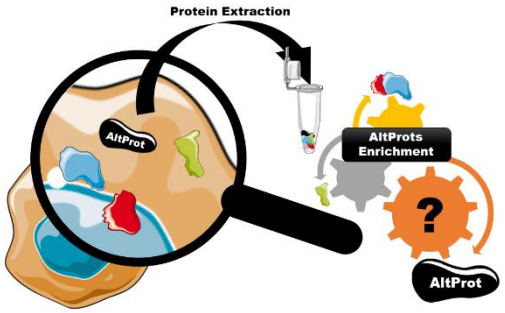
(13) Delcourt, V.; Brunelle, M.; Roy, A. V.; Jacques, J.-F.; Salzet, M.; Fournier, I.; Roucou, X. The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol. Cell. Proteomics* **2018**, *17* (12), 2402–2411. https://doi.org/10.1074/mcp.ra118.000593.

(14) Olexiouk, V.; Menschaert, G. Identification of Small Novel Coding Sequences, a Proteogenomics Endeavor. In *Advances in Experimental Medicine and Biology*; Springer, Cham, 2016; Vol.

926, pp 49–64. https://doi.org/10.1007/978-3-319-42316-6_4.
(15) Delcourt, V.; Franck, J.; Quanico, J.; Gimeno, J. P.; Wisztorski, M.; Raffo-Romero, A.; Kobeissy, F.; Roucou, X.; Salzet, M.; Fournier, I. Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Mol. Cell. Proteomics* **2018**, *17* (2), 357–372. https://doi.org/10.1074/mcp.M116.065755.
(16) Kelleher, N. L. Peer Reviewed: Top-Down Proteomics. *Anal. Chem.* **2004**, *76* (11), 196 A-203 A. https://doi.org/10.1021/ac0415657.
(17) Couso, J.-P.; Patraquim, P. Classification and Function of Small Open Reading Frames. *Nat. Publ. Gr.* **2017**, *18* (9), 575–589. https://doi.org/10.1038/nrm.2017.58.
(18) Le Rhun, E.; Duhamel, M.; Wisztorski, M.; Gimeno, J. P.; Zairi, F.; Escande, F.; Reyns, N.; Kobeissy, F.; Maurage, C. A.; Salzet, M.; et al. Evaluation of Non-Supervised MALDI Mass Spectrometry Imaging Combined with Microproteomics for Glioma Grade III Classification. *Biochim. Biophys. Acta - Proteins Proteomics* **2017**, *1865* (7), 875–890. https://doi.org/10.1016/j.bbapap.2016.11.012.
(19) Ma, J.; Diedrich, J. K.; Jungreis, I.; Donaldson, C.; Vaughan, J.; Kellis, M.; Yates, J. R.; Saghatelian, A.; Saghatelian, A. Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* **2016**, *88* (7), 3967–3975. https://doi.org/10.1021/acs.analchem.6b00191.
(20) Schwaid, A. G.; Shannon, D. A.; Ma, J.; Slavoff, S. A.; Levin, J. Z.; Weerapana, E.; Saghatelian, A. Chemoproteomic Discovery of Cysteine-Containing Human Short Open Reading Frames. *J. Am. Chem. Soc.* **2013**, *135* (45), 16750–16753. https://doi.org/10.1021/ja406606j.
(21) Cassidy, L.; Prasse, D.; Linke, D.; Schmitz, R. A.; Tholey, A. Combination of Bottom-up 2D-LC-MS and Semi-Top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon Methanosarcina Mazei. *J. Proteome Res.* **2016**, *15* (10), 3773–3783. https://doi.org/10.1021/acs.jproteome.6b00569.
(22) Cassidy, L.; Kaulich, P. T.; Tholey, A. Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J. Proteome Res.* **2019**, *18* (4), 1725–1734. https://doi.org/10.1021/acs.jproteome.8b00948.
(23) Salzet, M.; Wattez, C.; Verger-Bocquet, M.; Beauvillain, J. C.; Malecha, J. Oxytocin-like Peptide: A Novel Epitope Colocalized with the FMRFamide-like Peptide in the Supernumerary Neurons of the Sex Segmental Ganglia of Leeches—Morphological and Biochemical Characterization; Putative Anti-Diuretic Function. *Brain Res.* **1993**, *601* (1–2), 173–184. https://doi.org/10.1016/0006-8993(93)91708-Z.
(24) Salzet, M.; Bulet, P.; Wattez, C.; Verger-Bocquet, M.; Malecha, J. Structural Characterization of a Diuretic Peptide from the Central Nervous System of the Leech Erpobdella Octoculata: Angiotensin II Amide. *J. Biol. Chem.* **1995**, *270* (4), 1575–1582. https://doi.org/10.1074/jbc.270.4.1575.
(25) Gao, K. M.; Chen, X. C.; Zhang, J. X.; Wang, Y.; Yan, W.; You, Y. P. A Pseudogene-Signature in Glioma Predicts Survival. *J. Exp. Clin. Cancer Res.* **2015**, *34* (1), 23. https://doi.org/10.1186/s13046-015-0137-6.
(26) SALZET, M.; BULET, P.; WATTEZ, C.; MALECHA, J. FMRFamide-related Peptides in the Sex Segmental Ganglia of the Pharyngobdellid Leech Erpobdella Octoculata Identification and Involvement in the Control of Hydric Balance. *Eur. J. Biochem.* **1994**, *221* (1), 269–275. https://doi.org/10.1111/j.1432-1033.1994.tb18738.x.
(27) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A. J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; et al. Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography-Tandem Mass Spectrometry. *J. Proteome Res.* **2010**, *9* (2), 761–776. https://doi.org/10.1021/pr9006365.
(28) Hallberg, M. Neuropeptides: Metabolism to Bioactive Fragments and the Pharmacology of Their Receptors. *Med. Res. Rev.* **2015**, *35* (3), 464–519. https://doi.org/10.1002/med.21323.
(29) Cardon, T.; Salzet, M.; Franck, J.; Fournier, I. Nuclei of HeLa Cells Interactomes Unravel a Network of Ghost Proteins Involved in Proteins Translation. *Biochim. Biophys. Acta - Gen. Subj.* **2019**. https://doi.org/10.1016/j.bbagen.2019.05.009.
(30) Tutar, Y. Review Article Pseudogenes. *Comp. Funct. Genomics* **2012**, *2012*, 4. https://doi.org/10.1155/2012/424526.
(31) Poliseno, L.; Marranci, A.; Pandolfi, P. P. Pseudogenes in Human Cancer. *Frontiers of Medicine*. Frontiers Media SA 2015. https://doi.org/10.3389/fmed.2015.00068.
(32) Liu, B.; Liu, J.; Liu, K.; Huang, H.; Li, Y.; Hu, X.; Wang, K.; Cao, H.; Cheng, Q. A Prognostic Signature of Five Pseudogenes for Predicting Lower-Grade Gliomas. *Biomed. Pharmacother.* **2019**, *117*, 109116. https://doi.org/10.1016/j.biopha.2019.109116.
(33) 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. A Draft Map of the Human Proteome. *Nature* **2014**, *509* (7502), 575–581. https://doi.org/10.1038/nature13302.
(34) Ji, Z.; Song, R.; Regev, A.; Struhl, K. Many LncRNAs, 5′UTRs, and Pseudogenes Are Translated and Some Are Likely to Express Functional Proteins. *Elife* **2015**, *4* (DECEMBER2015). https://doi.org/10.7554/eLife.08890.
(35) Wenda, S.; Dauber, E. M.; Schwartz, D. W. M.; Jungbauer, C.; Weirich, V.; Wegener, R.; Mayr, W. R. ACTBP2 (Alias ACTBP8) Is Localized on Chromosome 6 (Band 6q14). *Forensic Sci. Int.* **2005**, *148* (2–3), 207–209. https://doi.org/10.1016/j.forsciint.2004.05.006.
(36) Petryszak, R.; Keays, M.; Tang, Y. A.; Fonseca, N. A.; Barrera, E.; Burdett, T.; Füllgrabe, A.; Fuentes, A. M. P.; Jupp, S.; Koskinen, S.; et al. Expression Atlas Update - An Integrated Database of Gene and Protein Expression in Humans, Animals and Plants. *Nucleic Acids Res.* **2016**, *44* (D1), D746–D752. https://doi.org/10.1093/nar/gkv1045.
(37) Li, J.; McGettigan, J. P.; Faber, M.; Schnell, M. J.; Dietzschold, B. Infection of Monocytes or Immature Dendritic Cells (DCs) with an Attenuated Rabies Virus Results in DC Maturation and a Strong Activation of the NFκB Signaling Pathway. *Vaccine* **2008**, *26* (3), 419–426. https://doi.org/10.1016/j.vaccine.2007.10.072.
(38) Ochsner, F. [Contamination of a Glioma by the Herpes Virus]. *Schweiz. Arch. Neurol. Neurochir. Psychiatr.* **1981**, *129* (1), 19–30.
(39) Longuespée, R.; Boyon, C.; Desmons, A.; Vinatier, D.; Leblanc, E.; Farré, I.; Wisztorski, M.; Ly, K.; D'Anjou, F.; Day, R.; et al. Ovarian Cancer Molecular Pathology. *Cancer and Metastasis Reviews*. Springer US December 23, 2012, pp 713–732. https://doi.org/10.1007/s10555-012-9383-7.

Graphical Abstract

Schematic representation of the research of Alternative Proteins (AltProts). Enrichment strategy is needed to highlight AltProts in complex mixture. Workflow combined different extraction strategies and some specific steps for the AltProt enrichment.



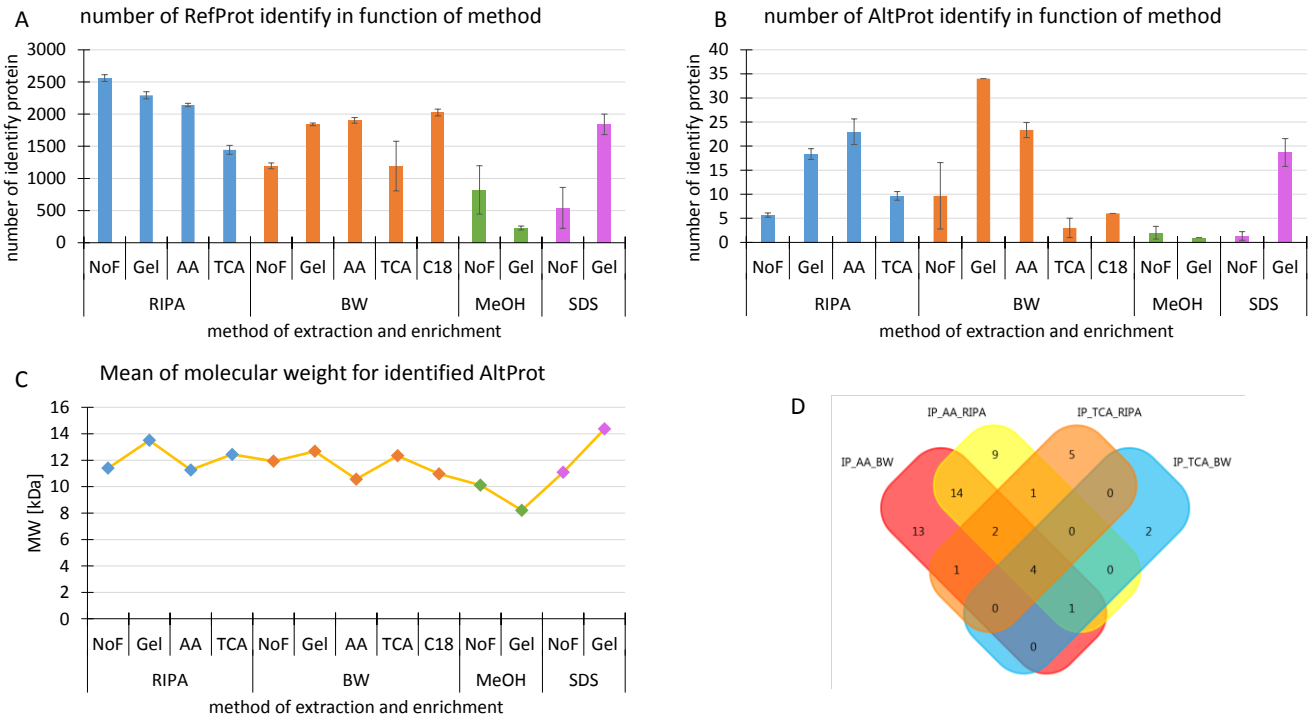
Highlights

- Alternative proteins are issued from Genome annotation missing
- To highlight the identification of AltProts in a complex mixtures' enrichment is need
- Enrichment workflow adjust extraction methods but also add some specifics steps
- NCH82, glioma cell line, present a high number of AltProts issue to ncRNA
- ncRNA expressed Altprots
- AltProts could be involved as ncRNA is translation or transcription modulation

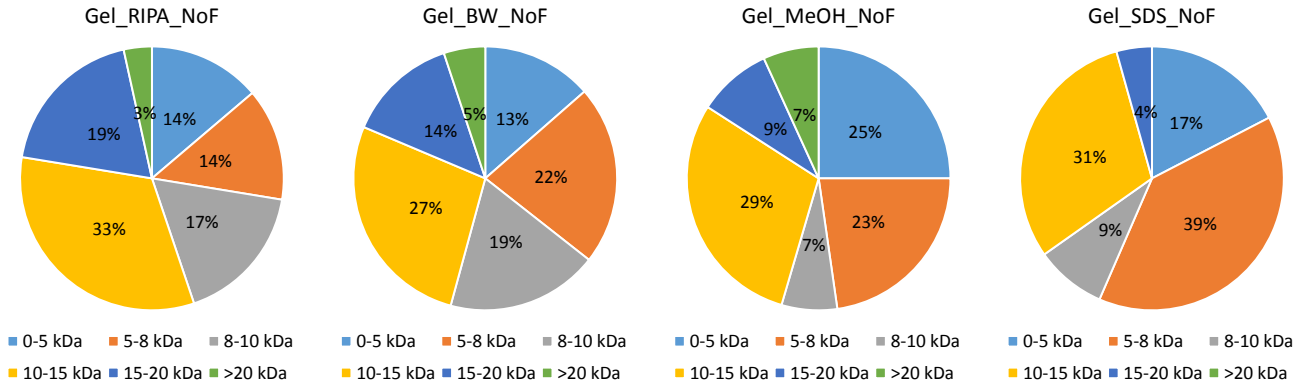
In brief: We established the workflow to characterized Alternative Proteins (AltProts) translated from non-annotated Alternative Open reading frame (AltORFs from glioma cell line NCH82. We identified 107 pseudogenes-derived proteins issued from lncRNA.

Supplementary data

Supplementary FIGURES



SuppFigure.1. A- number of RefProt identify discriminate for each methods of extraction and enrichment using, B- the same for the AltProt identify C- representation of the weight mean distribution for each kind of methods D- comparison of AA and TCA precipitation method in the recovering of AltProt identification, 4 are find in common between all, value based on at least two identifications in a triplicate for each methods



SuppFigure.2 : weight repartition of the AltProt identify for each extraction without fragmentation

Supplementary TABLES :

Extraction	Enrichment	mean in percent	standard deviation of the mean in percent
RIPA	NoF	0,22	0,02
	Gel	0,80	0,03
	AA	1,08	0,14
	TCA	0,67	0,03
BW	NoF	0,81	0,59
	Gel	1,85	0,02
	AA	1,23	0,10
	TCA	0,22	0,08
	C18	0,30	0,01
MeOH	NoF	0,22	0,15
	Gel	0,45	0,05
SDS	NoF	0,29	0,25
	Gel	1,03	0,21

SuppTable 1. Ratio of the number of AltProt identified in comparison to all the RefProt identified for all the experimental condition, for the triplicate a mean was do and the standard variation calculate.

Extraction	Enrichment	mean in percent	standard deviation of the mean in percent
RIPA	NoF	9,23	1,42
	Gel	8,37	2,85
	AA	11,24	4,47
	TCA	9,19*	1,87

BW	**NoF**	11,20	0,64
	Gel	7,91	1,16
	AA	10,57	1,71
	TCA	15,88	2,82
	C18	7,79	1,91
MeOH	**NoF**	18,48	6,02
	Gel	28,91*	9,77
SDS	**NoF**	8,30*	1,07
	Gel	11,80	0,58

SuppTable 2. Variation in percentage of RefProt identification in the triplicate for each kind of extraction and enrichment methods (* n=2)

protein accession	protein length (a.a.)	molecular weight (kDa)	isoelectric point	gene symbol	type	AA_BW	AA_RIPA	Gel_BW	Gel_MeOH	Gel_RIPA	Gel_SDS	TCA_BW	TCA_RIPA
IP_587085	148	16,62	11,73	RP11-475C16.1	ncRNA	x	x	x	x	x	x	x	x
IP_595471	204	24,17	12,13	RPL15P3	ncRNA	x	x	x	x	x		x	x
IP_759887	254	28,85	7,2	RP11-490H24.5	ncRNA	x	x	x		x	x	x	x
IP_591792	153	17,24	4,62	ACTBP8	ncRNA	x	x	x		x	x	x	x
IP_746392	266	30,11	11,44	RPL7AP6	ncRNA	x		x		x	x	x	x
IP_781237	145	17,27	11,11	RPL26P30	ncRNA	x	x	x			x	x	x
IP_584395	160	18,59	11,17	RPL21P75	ncRNA	x	x	x		x	x		x
IP_734708	71	7,77	6,48	RP11-24M17.3	ncRNA	x	x	x		x			x
IP_592880	90	10,1	10,08	RP3-486D24.1	ncRNA	x	x	x		x	x		
IP_736003	264	29,97	10,47	RPS3AP6	ncRNA	x	x					x	x
IP_688853	58	6,26	7,51	RP11-15H20.6	ncRNA			x	x	x	x		
IP_637160	255	28,4	9,5	RP11-395L14.17	ncRNA	x						x	x
IP_602798	92	10,25	6,79	RP11-509M23.1	ncRNA		x	x					x
IP_572421	229	25,27	4,43	TUBBP1	ncRNA	x	x						x
IP_559683	115	13,04	11,63	GS1-184P14.2	ncRNA	x	x						x
IP_613138	44	5,06	4,72	TMSB4XP8	ncRNA					x	x	x	
IP_592855	165	17,99	7,41	PPIAP9	ncRNA			x		x	x		
IP_662403	141	15,58	4,28	NPM1P19	ncRNA			x		x	x		
IP_556680	58	6,62	10,39	USMG5P1	ncRNA			x		x	x		

IP_603809	189	20,68	10,41	RP11-79P5.10	ncRNA			x		x	x		
IP_691726	74	8,32	10,53	CTC-398G3.1	ncRNA			x		x	x		
IP_624545	108	12,55	10,57	AC092798.2	ncRNA			x		x	x		
IP_723386	88	10,2	10,98	RPL18P13	ncRNA			x		x	x		
IP_565092	115	13,02	11,35	RPS26P3	ncRNA			x		x	x		
IP_557390	109	12,26	6,77	RP5-878I13.1	ncRNA	x	x			x			
IP_565117	92	10,68	8,46	SNRPEP2	ncRNA	x	x			x			
IP_641478	101	11,42	7,67	HNRNPA1P66	ncRNA						x		x
IP_592933	64	7,11	10,33	BTF3P7	ncRNA			x					x
IP_670480	76	8,43	9,49	HNRNPA1P68	ncRNA	x							x
IP_612504	136	15,23	11,61	H3F3AP6	ncRNA	x							x
IP_648412	86	9,11	7,81	ERCC3	ncRNA					x	x		
IP_737334	106	12,03	9,82	ACTBP7	ncRNA					x	x		
IP_556923	142	16,18	11,25	H3F3AP5	ncRNA					x	x		
IP_593099	225	24,94	4,31	TUBB2BP1	ncRNA			x			x		
IP_595290	130	14,48	8,68	ASS1P1	ncRNA			x			x		
IP_709935	222	23,77	9,41	UBE2SP1	ncRNA			x			x		
IP_634654	129	15,44	10,88	AC009302.2	ncRNA		x			x			
IP_689114	394	44,7	6,33	AC078899.1	ncRNA			x		x			
IP_557834	87	9,73	4,09	ACTBP1	ncRNA	x	x						
IP_655967	83	9,42	10,6	CTA-243E7.4	ncRNA	x	x						
IP_612062	215	23,4	4,13	NACA3P	ncRNA								x
IP_613753	74	8,75	4,53	TUBB4BP5	ncRNA								x

IP_572422	212	24,18	4,64	TUBBP1	ncRNA								x
IP_673532	144	16,46	4,73	EIF1AXP1	ncRNA								x
IP_667681	160	18	6,23	RP11-193H5.1	ncRNA								x
IP_2264871	84	9,92	4,62	LOC105372741	ncRNA							x	
IP_776097	132	14,6	6,9	RP11-864N7.2	ncRNA							x	
IP_649269	74	8,41	8,22	NCOA1	**mRNA**							x	
IP_660647	73	7,16	4,93	UBASH3A	**mRNA**						x		
IP_671420	212	23,3	5,57	PSAT1P3	ncRNA						x		
IP_737074	85	9,48	7,6	EEF1A1P22	ncRNA						x		
IP_755869	335	36,71	8,07	GAPDHP69	ncRNA						x		
IP_761518	123	13,79	8,1	RPS4XP1	ncRNA						x		
IP_671464	249	26,94	8,21	TPI1P1	ncRNA						x		
IP_765292	56	6,51	10,91	RPL21P136	ncRNA						x		
IP_592773	135	14,92	9,85	RP11-378G13.2	ncRNA			x					
IP_592972	95	10,67	5,18	PKMP5	ncRNA		x						
IP_619702	154	17,24	3,74	RPSAP69	ncRNA					x			
IP_721455	187	21,62	4,32	TUBB8P7	ncRNA					x			
IP_602542	239	27,24	4,78	RP11-274E7.2	ncRNA					x			
IP_2390011	57	6,48	8,23	LOC105372901	ncRNA					x			
IP_774832	168	17,95	8,66	GAPDHP70	ncRNA					x			
IP_2364146	35	3,83	9,61	CFI	**mRNA**					x			
IP_623043	80	8,83	9,61	EEF1A1P8	ncRNA					x			

IP_669889	100	11,58	10,47	RP11-389O22.4	ncRNA					x			
IP_639671	108	12,68	10,87	RP11-416L21.1	ncRNA					x			
IP_639834	149	17,59	10,94	RPL7P13	ncRNA					x			
IP_559678	225	24,93	4,34	EEF1B2P3	ncRNA			x					
IP_563312	44	5,11	4,72	TMSB4XP4	ncRNA			x					
IP_755940	36	4,11	4,87	HNRNPA1P30	ncRNA			x					
IP_671453	306	34,92	6,64	RP11-181C21.4	ncRNA			x					
IP_789374	103	11,09	7,15	GAPDHP21	ncRNA			x					
IP_580245	131	14,49	7,44	AC091654.7	ncRNA			x					
IP_592506	112	12,48	8,25	TUBBP9	ncRNA			x					
IP_587041	271	29,39	8,29	LDHAL6FP	ncRNA			x					
IP_624042	58	7	10,96	NDUFB1P1	ncRNA			x					
IP_579441	41	4,63	11,55	EEF1GP1	ncRNA			x					
IP_624921	50	5,66	4,19	ACTG1P12	ncRNA		x						
IP_593685	128	14,51	5,7	EEF1A1P42	ncRNA		x						
IP_591881	151	16,44	7,57	GAPDHP63	ncRNA		x						
IP_590800	228	24,4	9,7	GAPDHP72	ncRNA		x						
IP_641652	92	10,26	10,32	RP11-33O4.2	ncRNA		x						
IP_623047	111	12,14	10,69	EEF1A1P8	ncRNA		x						
IP_622873	152	17,65	11,51	RP11-234A1.1	ncRNA		x						
IP_756756	135	15,15	5	FABP5P2	ncRNA	x							
IP_2286622	60	6,67	8,11	CDYL2	**mRNA**	x							

IP_612311	236	25,52	8,76	RTN3P1	ncRNA	x							
IP_736298	70	8,09	9,01	LINC01579	ncRNA	x							
IP_173179	82	9,52	11,85	NAA35	**mRNA**	x							

SuppTable 3. List of AltProts found in triplicate (red font protein accession) or in duplicate (black font protein accession), according to this MW in kDa, this isoelectric point prediction, the gene origin of the transcript, the type of RNA produced the AltProt (mRNA or ncRNA) and finally the extraction and enrichment methods where the AltProt was identified. In colored row the identification attributed specifically to one kind of extraction and enrichment method.

	Protein Accession	Type	Transcript Accession	Gene	Gene Description	chromosome
Gel/Precipitation	IP_591792	ncRNA	ENST00000403258	ACTBP8	Actin, Beta Pseudogene 8	6
	IP_592880	ncRNA	ENST00000401715	AL136226.1	ribosomal protein L7A (RPL7A) pseudogene	6
	IP_734708	ncRNA	ENST00000567565	PPIAP47	peptidylprolyl isomerase A pseudogene 47	15
Precipitation	IP_655967	ncRNA	ENST00000623888	AL022323.3	novel transcript	22
	IP_602798	ncRNA	ENST00000425840	PPIAP79	peptidylprolyl isomerase A pseudogene 79	5
	IP_557834	ncRNA	ENST00000417985	ACTBP1	ACTB pseudogene 1	X
	IP_572421	ncRNA	ENST00000248151, ENST00000518096	TUBBP1	tubulin beta pseudogene 1	8
GEL	IP_688853	ncRNA	ENST00000598599	RP11-15H20.6		19
	IP_624545	ncRNA	ENST00000458542	AC092798.1	ribosomal protein L32 (RPL32) pseudogene	3
	IP_689114	ncRNA	ENST00000521432	AC078899.1	ARP2 actin-related protein 2 homolog (yeast) (ACTR2) pseudogene	19
	IP_592855	ncRNA	ENST00000403866	PPIAP9	peptidylprolyl isomerase A pseudogene 9	6
	IP_662403	ncRNA	ENST00000450070	NPM1P19	nucleophosmin 1 pseudogene 19	20
	IP_593099	ncRNA	ENST00000404155	TUBB2BP1	tubulin beta 2B class IIb pseudogene 1	6
	IP_691726	ncRNA	ENST00000483614	AC008481.2	ribosomal protein L18a (RPL18A) pseudogene	19
	IP_723386	ncRNA	ENST00000478088	RPL18P13	ribosomal protein L18 pseudogene 13	16
	IP_737334	ncRNA	ENST00000418351	ACTBP7	Actin, Beta Pseudogene 7	15

SuppTable 4. AltProt identified in commune for Gel and precipitation and them specific to one or the other condition.

Each are finding in triplicate in the extraction method and at least three of the four condition of precipitation (AA_BW, AA_RIPA, TCA_BW, TCA_RIPA) and same in gel condition (Gel_SDS, Gel_MeOH, Gel_RIPA, Gel_BW).

III. Conclusion

Cette étude permet de poser plusieurs constats, le premier lié au fait qu'une seule et unique méthode ne permet pas encore d'extraire spécifiquement les AltProts. En effet parmi les différentes méthodes utilisées dans cette étude, l'extraction peptidique à l'eau bouillante est celle qui montre le meilleur résultat, toutefois en combinaison avec une séparation sur gel ou une précipitation à l'acide acétique on améliore également le rendement d'identification. Le fait est qu'à la suite des différentes extractions très peu de protéines sont retrouvées en commun. Cela est dû à une grande variété de AltProts, qui comme les RefProts font partie de différentes familles de protéines, avec des propriétés physicochimiques particulières, ayant pour conséquence des réponses différentes aux méthodes d'extractions ou de séparations.

Le deuxième constat est l'origine très majoritaire des AltProts identifiées dans les cellules (NCH82) étudiées. En effet après filtration des AltProts afin qu'elles soient retrouvées dans au moins deux des trois expériences réalisées pour chaque méthodes, 89 AltProts sont identifiées dont plus de 90% sont issues de ARNlnc. De manière intéressante les ARNlnc sont de plus en plus étudiés dans le cas de pathologies. Si aujourd'hui on leurs attribues un grand nombre de fonctions, comme il a été présenté précédemment, leur statut de transcrit non codant reste inchangé. Cependant lors de cette étude, des produits protéiques de ces transcrits, prédits dans une base de données ne suivant pas le code de traduction des RefProts, ont été identifiés par analyse MS. Cette observation montre une fois de plus les lacunes imposées par l'utilisation de bases de données disponibles, et la nécessité de se tourner vers des méthodes de protéogénomique.

Le troisième constat est l'origine des transcrits ARNlnc, en effet tous sont issus de pseudogènes. Les pseudogènes sont des séquences ADN copier-coller sur des régions différentes de l'ADN voir sur d'autres chromosomes. Toutefois lors de cette manœuvre la séquence a été légèrement modifiée provocant l'inhibition de la RefProt normalement produite. Cela est produit par une modification de la séquence ADN, modifiant le transcrit ARN qui alors soit n'est

pas produit, soit possède une modification de son codon START ne permettant plus la traduction de la RefProt. Cette perte de la traduction de la RefProt est liée à l'appellation non codante de l'ARN. Toutefois si cette modification affecte la RefProt initialement induite par le transcrit ARN les AltProts ne sont pas nécessairement touchées. Plus encore, la modification du transcrit peut amener à la production d'un nouveau codon START et à la traduction d'une AltProt. Ces AltProts possèdent parfois des chevauchements de séquences avec la partie non modifiée du transcrit et donc avec la partie codant pour la RefProt. Ceci a pour conséquence la présence de domaines connus sur des protéines non décrites dans les bases de données.

La présence de domaines connus grâce à la reconnaissance de séquences en acides aminés spécifiques, et faisant lieu de site actif sur les RefProts, permet de proposer des fonctions pour ces AltProts. En effet l'enrichissement des extraits en AltProts permet de mettre en évidence de nouveaux biomarqueurs dans les tissus pathologiques, et peut-être dans les fluides biologiques. À l'image des neuropeptides qui sont sécrétés par la cellule et construisent le circuit de la communication intercellulaire, les AltProts ont également une chance de se trouver dans le milieu extracellulaire. Toutefois ces prédictions de fonction basées sur la reconnaissance de domaines issus de séquences en acides aminés, nécessitent d'être validées.

Afin de comprendre de manière plus spécifique la fonction des AltProts qui peuvent-être mises en évidence dans les études réalisées, nous avons cherché à développer une stratégie permettant de combiner l'analyse large échelle de protéines avec l'identification et l'attribution de voies de signalisation. La recherche des partenaires d'interaction par XL-MS était alors une solution adaptée à cette problématique mais pour laquelle aucune étude n'a jamais été publiée. C'est pourquoi avant de nous lancer dans l'utilisation de la méthode XL-MS appliquée à la recherche de fonctions des AltProts, nous avons analysé des données publiques utilisant le XL-MS afin de mettre en évidence la capacité à identifier des AltProts dans les réseaux obtenus.

PARTIE IV

Le protéome fantôme un acteur important des réseaux d'interaction Protéines-Protéines : Mise en évidence par Stratégie XL-MS

I. Protéomique et réutilisation des données

Depuis plusieurs années, la protéomique à large échelle est en plein essor. Il est de plus en plus facile d'obtenir un grand nombre d'identifications, jusqu'à 10.000 protéines identifiées par analyse, avec une séparation de 100 min [1]. Désormais l'ensemble du protéome est identifié et accessible, mais l'analyse de ces données d'un point de vue fonctionnel reste un véritable challenge que l'émergence de l'intelligence artificielle devrait grandement aider. De ce fait, en partie par manque d'outils bioinformatiques dédiés ou intégrés, les analyses large échelle sont rarement exploitées à 100%. Pour répondre à ce problème et par transparence éthique sur la publication de résultats, des banques de données brutes ont vu le jour. Ces banques donnent accès aux données de MS non retraitées, permettant à chacun de télécharger les fichiers d'analyses et de les retraiter par ses propres outils. Parmi ces banques de données, on retrouve PRIDE [130], GPMDB [131], PeptideAtlas [132], ou encore le projet Chorus (https://chorusproject.org), chacun permettant de télécharger les fichiers bruts de divers types d'analyses associés à une publication. La mise en place de ces banques permet notamment des retraitements statistiques de grande envergure. Ces données publiques offrent également la possibilité de tester de nouvelles méthodologies bioinformatiques afin de prouver la faisabilité d'un système ou d'une méthode [133].

Ces banques de données nous ont permis d'avoir accès à des jeux de données provenant d'analyses XL-MS réalisées par le groupe de Heck [117] pour lequel un grand nombre de PPI avait été identifié. Cette publication avait pour but de présenter une méthodologie de XL-MS robuste et permettant d'identifier un maximum de partenaires en interaction en une seule expérience. En effet, les auteurs ont obtenu l'identification de 2,426 pontages uniques avec un FDR de 5% (2 013 intraprotéines et 413 interprotéines). De manière intéressante, l'identification de ces protéines en interaction permet de les replacer dans des voies de signalisation communes et ainsi d'avoir une idée de leurs fonctions. Les analyses réalisées ne prennent évidemment pas en compte les AltProts. Rechercher les AltProts à partir de ces données nous permet de savoir si ces dernières interagissent et dans quelle mesure avec des RefProts puis de

retrouver leurs partenaires (RefProts et/ou AlProts). Cette première approche *in silico* nous permet de déterminer si la stratégie XL-MS est réellement une stratégie pertinente pour l'étude de la fonction des AltProts.

II. Le Protéome caché dans les données de PPIs

1. Approche des fonctions des AltProts par analyse *in silico*

Par le passé, la ré-analyse de données issues des banques publiques, a permis de mettre en évidence des caractéristiques intrinsèques aux AltProts. Notamment, ces réanalyses ont permis de préciser la proportion des AltProts dans différentes espèces, la distribution de leur localisation sur l'ARNm (CDS, 5' et 3' UTR) mais aussi, par exemple, de définir leur capacité à être phosphorylées et parfois même d'approcher leurs partenaires d'interaction (**Figure 15**) [23] ; tout en gardant à l'esprit que les études reprises ne sont pas optimisées pour l'analyse des AltProts.

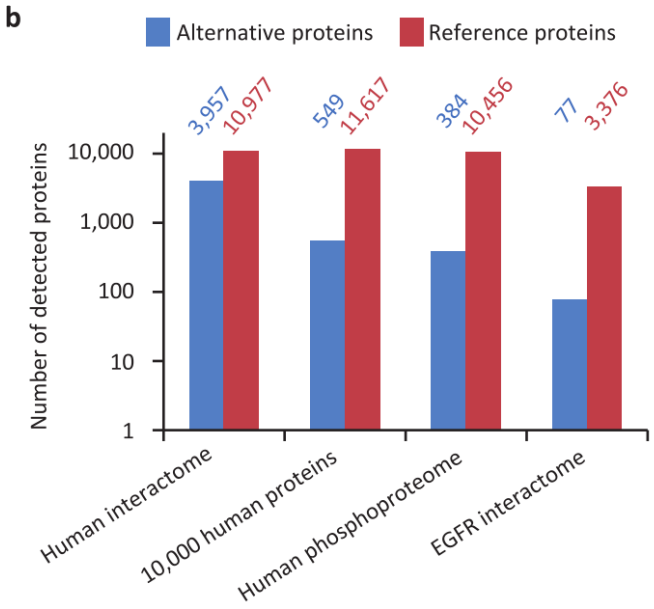


Figure 15 : ***Représentation de la répartition des AltProts dans différents contextes*** *: dans des données d'interactome total réalisé chez l'homme, parmi les partenaires d'interaction des phosphoprotéines, ou encore de manière ciblée dans l'interactome d'EGFR (Samandi & al., 2017* [23]*)*

Cependant, la compréhension de la fonction de ces AltProts reste limitée souvent par la mise en évidence d'un partenaire RefProt ciblé faisant apparaître la

présence d'AltProts, comme dans la description utilisant l'interactome de l'EGFR. Les données obtenues sur PRIDE (PXD000788), issues entre autre d'analyses d'IP dirigées contre EGFR, ont permis la mise en évidence de 77 AltProts [23]. L'étude à large échelle des interactions formées par les AltProts nécessite toutefois le développement de nouvelles méthodologies.

Malgré une utilisation encore très réduite dans la communauté protéomique, la méthode XL-MS, appliquée à des échantillons complexes n'échappe pas à la règle et permet pour quelques publications d'avoir accès aux fichiers d'analyses. Dans le cadre de ma thèse, je me suis basé sur la publication de référence «*Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry*» de *F.Liu & al.* [117], présentant l'application du XL-MS à partir de l'extraction de noyaux de cellules HeLa, dont on peut trouver les données sur CHORUS sous l'identifiant de projet numéro 890. Cette étude menée par le groupe d'Albert Heck a permis de mettre en évidence 2179 pontages uniques dont 1,665 intra-protéines et 514 inter-protéines. Afin de vérifier l'hypothèse que les AltProts pourraient être impliquées dans la régulation des réseaux de protéines ou directement impliquées en tant que ligands, une ré-analyse des données disponibles a donc été réalisée.

2. Prédiction 3D et fonctions des protéines

La prédiction d'un modèle 3D est une étape critique mais devenue incontournable après l'analyse par XL-MS. En effet, l'identification d'une interaction par XL-MS conduit à connaître la distance d'interaction via la longueur restrictive de l'agent pontant (30 Å pour le DSSO et 11 Å pour le DSS et le BS3). Ainsi, lorsqu'un modèle d'interaction entre deux protéines est généré sans à priori de distance, le résultat du « *docking* » doit alors prédire une distance équivalente à celle observée en XL-MS. De cette manière, l'observation faite en XL-MS est confirmée par la prédiction des modèles. Pour les RefProts les prédictions sont facilitées par le fait que les modélisations 3D de certaines protéines sont déjà disponibles sur les bases de données telles que PDB [134] ou obtenues de manière expérimentale par cristallographie ou RMN. Pour les

AltProts, l'absence de données expérimentales et de modèles en banque requiert la réalisation de la prédiction du modèle sur la base de la séquence en acides aminés et de leur réaction dans l'eau. Cette prédiction est possible grâce à des algorithmes tel que I-Tasser [135] qui se base sur l'homologie de séquence avec des RefProts ainsi que sur le pouvoir hydrophobe et électrostatique des acides aminés de la séquence de l'AltProt.

III. Objectif

L'objectif est de mettre en évidence la présence d'AltProts qui pourraient interagir avec des RefProts et qui n'auraient pas été considérées dans l'étude menée par *F.Liu & al.* [117]. La méthodologie XL-MS permet l'étude à large échelle des interactions, toutefois même s'il est connu que les AltProts peuvent interagir avec les RefProts, rien ne laisse supposer que la méthode XL-MS permette également d'observer la présence d'AltProts. La méthode décrite portait sur l'utilisation de la méthodologie XL-MS appliquée à un mélange complexe constitué des noyaux de cellules HELA. Si la majeure partie des informations a été exploitée dans la publication, les AltProts n'avaient pas été recherchées. En analysant de nouveau ces données avec la base de données AltProts, un grand nombre d'informations inexploitées ont été mises en lumière. Parmi les points importants, nous avons pu mettre en évidence l'implication d'AltProts dans la régulation du ribosome et donc, dans la régulation de la biosynthèse des protéines (**Figure 16**). Cet article princeps permet de valider les résultats préliminaires obtenus [23], annonçant une implication des AltProts dans les processus biologiques de la biosynthèse, du métabolisme des acides nucléiques, du métabolisme protéique ou encore dans le transport de celles-ci.

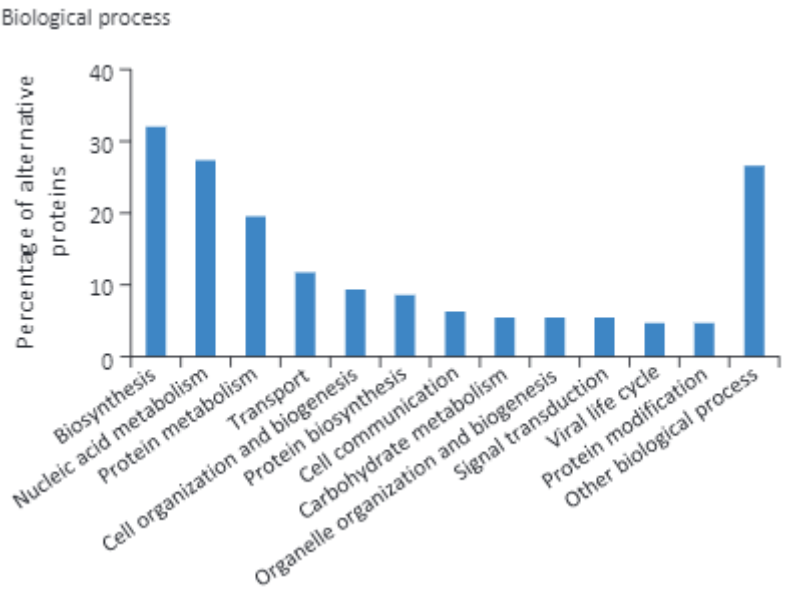


Figure 16 : ***Prédiction de l'implication des AltProts dans les processus biologiques****. Les voies de signalisation majoritairement représentées sont la biosynthèse de composés, le métabolisme des acides nucléiques et le métabolisme des protéines.(Samandi & al., 2017* [23]*)*

Avec la capacité de la méthode XL-MS à trouver les partenaires des AltProts, les fonctions de celles-ci deviennent alors accessibles. La méthode XL-MS nous montrera plus tard que par la modification des interactions et l'observation des partenaires RefProts reliés aux annotations de gènes (GO-Term), la fonction des AltProts devient alors accessible. Ici la fonction ne sera abordée que pour une AltProt : AltATAD2, retrouvée en interaction avec RPL10 et AUF1. Ces deux RefProts sont décrites dans la composition et la régulation du ribosome et dans la synthèse des protéines. La prédiction de la structure tertiaire d'AltATAD2 et l'utilisation d'un logiciel de « *docking* » permet alors de décrire l'agencement de l'AltProt sur les RefProts. Cette étape permet à la fois de confirmer les distances d'interactions obtenues par XL-MS et de proposer un mécanisme de fonction pour AltATAD2.

BBA - General Subjects xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

BBA - General Subjects

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

Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation

Tristan Cardon, Michel Salzet*, Julien Franck*, Isabelle Fournier*

Inserm, U1192 – Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM), Université de Lille, F-59000 Lille, France

ARTICLE INFO

Keywords:
Cross-linking mass spectrometry
Alternative proteins
Ghost proteins
Alternative open reading frame
Translation
Disuccinimidyl sulfoxide

SUMMARY

Ghost proteins are issued from alternative Open Reading Frames (ORFs) and are missing a genome annotation. Indeed, historical filters applied for the detection of putative translated ORFs led to a wrong classification of transcripts considered as non-coding although translated proteins can be detected by proteomics. This Ghost (also called Alternative) proteome was neglected, and one major issue is to identify the implication of the Ghost proteins in the biological processes. In this context, we aimed to identify the protein-protein interactions (PPIs) of the Ghost proteins. For that, we re-explored a cross-link MS study performed on nuclei of HeLa cells using cross-linking mass spectrometry (XL-MS) associated with the HaltOrf database. Among 1679 cross-link interactions identified, 292 are involving Ghost Proteins. Forty-Four of these Ghost proteins are found to interact with 7 Reference proteins related to ribonucleoproteins, ribosome subunits and zinc finger proteins network. We, thus, have focused our attention on the heterotrimer between the RE/poly(U)-binding/degradation factor 1 (AUF1), the Ribosomal protein 10 (RPL10) and AltATAD2. Using I-Tasser software we performed docking models from which we could suggest the attachment of AUF1 on the external part of RPL10 and the interaction of AltATAD2 on the RPL10 region interacting with 5S ribosomal RNA as a mechanism of regulation of the ribosome. Taken together, these results reveal the importance of Ghost Proteins within known protein interaction networks.

1. Introduction

Advances in mass spectrometry (MS) instrumentation and bioinformatics tools have led to an exponential increase of MS-based proteomics strategy performances. It is now possible to get the identification and relative quantification of > 10,000 proteins in 100 min as recently published by Meier et al. [1]. These MS-based shot-gun strategies were also extended to structural characterization of the identified proteins. Various approaches were proposed over the past 15 years in proteomics for measuring protein-protein interactions (PPIs). Among these are the affinity capture [2], proximity labeling methods such as Apex [3,4], BioID [5–7] and Virotrap [8]; and the cross-linking mass spectrometry (XL-MS) [9]. XL-MS is advantageously non-targeted, providing a global onset for systems biology. However, all these strategies rely on protein database interrogation. Therefore, only referenced proteins can be identified [10,11]. This is a clear limitation for discovery if the databases are not complete. Public databases are built on both measured proteins and predicted ones. Predicted proteins are deduced from genome information accordingly to well-defined rules of annotation but are not all experimentally validated. The rules used for predicting protein sequences include the number of codons, the type of sequence and the Kozak context, which predicts the ribosome binding capacity on an mRNA [10,12–14]. Indeed, only the longest open reading frame (ORF) (so-called reference ORF, RefORF) or protein-coding sequences (CDSs) is considered per transcript in the databases (e.g. Ensembl [15] and GENCODE [16]), other ORFs being excluded from annotation [17]. In particular, short ORFs (sORFs) or small ORFs (smORFs) that do not respect the 100 codons (300 nucleotides) cut-off rule or the Kozak code, alternative ORFs (AltORF) remain unannotated. However, the proteome is more complex than initially expected and with recent advances in the field of genomics and MS-based proteomics with the high throughput sequencing technologies, it has been shown that traditional computational genome annotation algorithms have underestimated the number of coding sequences leaving out alternative promotors [18], alternative splicing [19], alternative polyadenylation [20] and ribosomal frameshifting [21]. There is, thus, a major challenge for genome annotation to reference all these new ORFs which are left out despite leading to proteins presenting biological activities. One

* Corresponding authors at: Faculté des Sciences, Laboratoire Réponse Inflammatoire et Spectrométrie de Masse (PRISM), Inserm U1192 - Université de Lille, Campus Cité Scientifique, Bât SN3, 1er étage, F-59655 Villeneuve d'Ascq Cedex, France.

E-mail addresses: Michel.salzet@univ-lille.fr (M. Salzet), julien.franck@univ-lille.fr (J. Franck), isabelle.fournier@univ-lille.fr (I. Fournier).

https://doi.org/10.1016/j.bbagen.2019.05.009
Received 19 December 2018; Received in revised form 18 April 2019; Accepted 14 May 2019

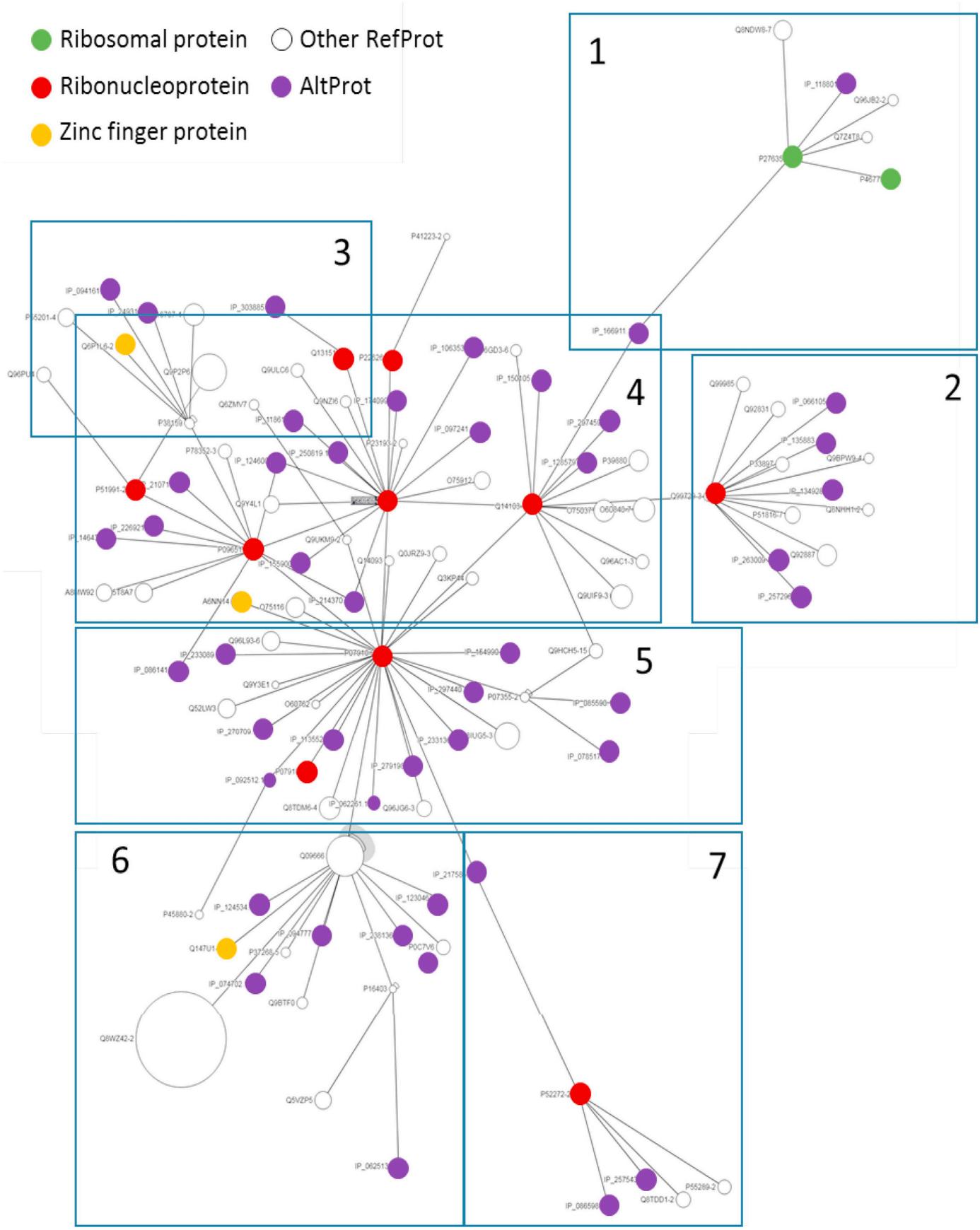


Fig. 1. Interaction network obtained from the XL-MS experiments using a RefProt/AltProt database for data interrogation issued from the combination of the RefProt (Uniprot) and the AltProt (HaltProt) databases. 44 AltProts are found in the network to be interacting with 10 ribonucleoproteins, 3 zinc finger proteins and 2 ribosomal proteins. The network is subdivided into 7 fractions allowing the annotation of AltProts and RefProts in interactions (see Table 2).

difficulty, is to be able to distinguish in this rising number of ORFs, the ORFs which are translated into functional proteins (such as micro-proteins, micropetides or SEPs) from the small ORFs that are randomly present but not translated. The smORFs/sORFs/AltORFs are often distinguished from the RefORFs because they are shorter size leading to the translation of small proteins (< 30 kDa). In average proteins translated from AltORFs are 57 amino acids in size, when by contrast the RefORFs proteins are 344 amino acids [22–26]. Importantly, these microproteins are not proteoforms of annotated proteins but have a different primary structure. Different computational approaches were used to identify these novel coding ORFs and create new databases including the predicted "alternative" transcripts (HaltORF [27], OpenProt [23], smProt [28]). The proteins issued from these AltORFs are called Ghost or Alternative Proteins (AltProts). Interestingly, proteomics has largely contributed to experimentally evidence and validate the existence of AltProts. Indeed, RefProts and AltProts were both detected from various studies by bottom-up [29,30] and top-down [31,32] proteomics. Interestingly, these proteins were identified within the 15% of proteomics data remaining unmatched after database interrogation despite a good quality MS/MS spectra; thus bridging the gap between experimental and predicted data.

If the discovery of these AltProts was definitively a revolution in the approach to systems biology, there is a clear unmet goal to find out the functions of these proteins. Absolute quantification by stable isotope-labeling and parallel reaction monitoring (PRM) was used to determine the levels of the two MIEF1 gene translational products, the reference MiD51 and the alternative MiD51 (AltMiD51) proteins, in two human cells lines and human colon tissues. This study has revealed a twofold higher expression of AltMiD51 compared to MiD51 [22] reinforcing the conviction that AltProts are major players in the regulation of biological systems. Studies have, indeed, demonstrated that AltProts can be important regulators in many fundamental events such as DNA repair [33], RNA decapping [34], calcium homeostasis metabolism [35], mTor signaling pathway [36], muscle performance [37], myoblast formation [38] and mitochondria fission [22]. Recently, it was shown that unannotated Heat Shock Protein [39] and Cold Shock Protein [40] were identified in *E. coli* by means of MS based proteomics. Specific AltProts were also found to be involved in physiopathological mechanisms including cancer and Spinal Cord Injury [30,32,41]. One step forwards the function of AltProts, is the identification of their interactome, by measuring PPI to gather information on the signaling pathways they are involved in [42]. Several studies have recently highlighted the adequacy of large scale interactomics XL-MS method as a discovery tool for new interactions [43,44]. In this context, we were willing to re-explore, using the HaltOrf database [27], a dataset of XL-MS from HeLa cells nuclei previously published by Heck group [44]. This has lead us to demonstrate the ability of XL-MS technique to discover previously unrevealed AltProt-RefProt interactions.

2. Material and methods

2.1. Ghost protein databases

The study was carried out using HaltORF database named "HS_GRCh38_altorf_20170421". This database is derived from the predicted *H. sapiens* alternative proteins (release hg38, Assembly: GCF_000001405.26) which contains 182,709 entries. This database is a computer compilation of all putative proteins from noncoding regions of mRNA and ncRNA. Additional online databases such as "Ensembl" (https://www.ensembl.org) and "ref Seq" (https://www.ncbi.nlm.nih.gov/refseq) were also used to trace back the origin of the identified AltProts after HaltORF data interrogation. The AltProts originate from either the 5 'and 3' UTR parts or from +2 or +3 reading frame shifts in the CDS of mature RNA; not following the Kozak frame despite the presence of a START and STOP codon. The HaltORF database was used in combination with the conventional RefProts database obtained from "UNIPROT".

2.2. Cell culture

The cells used in the analysis are derived from the HeLa line (ATCC). To summarizethe protocol described in the publication by F. Liu and al. ("*Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry*") [44], the cells are cultured in modified Dulbecco's Eagle environment with 10% fetal calf serum and 1% penicillin-streptomycin up to 80% confluence. The cells were then harvested by trypsinization and washed three times with PBS. After separation in the lysis membrane buffer and centrifugation, only the remaining nuclei were kept for the XL-MS. This nuclei fraction was then cross-linked with DSSO (1 mM) with a 100 fold excess of cross-linker with respect to the protein quantity. The cross-linked proteins were then reduced, alkylated and digested by Lys-C/Trypsin mixture. The resulting cross-linked peptides were desalted on a Sep-Pak C18, dried and further enriched using SCX as previously described [44].

2.3. Cross-link workflow

Data were extracted from the Chorus data repository (https://chorusproject.org) project I.D. number 890 and re-analyzed using Proteome Discoverer 2.2 (PD2.2) with the XLinkX node [45]. Interrogation of data was performed accordingly to the following workflow: first spectra were selected and DSSO was defined as cross-linker (characteristic mass 158.003765 Da). Then the workflow was divided into two paths. The first is dedicated to the cross-link identifications using the XLinkX Search as parameters Precursor Mass Tolerance: 10 ppm, FTMS fragment: 20 ppm, ITMS Fragment: 0.5 Da, search (database compiled AltProt + RefProt) and the validation was performed using percolator with a FDR set to 0.01. The second path is the total protein identification using SequestHT considering the following parameters: Trypsin as enzyme, 2 missed cleavages, methionine oxidation as variable modification, DSSO hydrolyzed and carbamidomethylation of cysteins as static modification, Precursor Mass Tolerance: 10 ppm and Fragment mass tolerance: 0.6 Da. The validation was performed using Percolator with a FDR set to 0.01. A consensus workflow was then applied for the statistical arrangement. A de-isotope and TopX filter were used to determine the *m/z*-error with a selectivity around 10% FDR. The protein-protein interaction identifiers were displays in the xiNET software (http://crosslinkviewer.org) [46] and Cytoscape3.7.1 allowing for visualization of the partners and the number of recurrences of the same interaction.

2.4. Modeling and prediction of interactions

Structure modeling of Ghost Proteins (AltProts) and Reference Proteins (RefProts), were performed with the I-Tasser software [47] when protein structures were not available on Protein Data Bank (PDB) [48]. For both RefProts and AltProts the most stable models (C-Score between −5 and +2) were retained. Within the set of best predictions, only models which are in line with the distances expected for the DSSO cross-linker were considered and further examined. The prediction of protein-protein interactions were performed with the ClusPro software [49]. The RefProt was identified as a receiver and the AltProt as a ligand. The interaction model was carried out by docking the ligand on the receiver without cross-link restriction. ClusPro then generates multiple interaction models ranked in the order of stability. The selected models are still part of the Top5 "balanced" models taking into account the best compromise of stability. The selected interactions were then recreated with Chimera [50] to measure the distance between the atoms observed during the cross-link. The model is split between the ligand and the receptor to form two independent chains, the lysines found to be involved in interactions on PD2.2 and xiNET were then designated in order to identify the distance between the two points of the model. For example, the AltProt AltATAD2 model was generated from its amino acid sequence since it was never previously described

the structural data could not be predicted by sequence homology and nature of these amino acids. The model was thus generated by I-Tasser with a C-Score of −3.66 in accordance with recommendations [−5; 2]. It was observed that, AltATAD2 has a secondary structure composed of 4 alpha helices generating a tubular tertiary structure. Similarly, the AUF1 reference protein had no experimental model and needed to be carried out on I-Tasser. The generated model has a C-Score of −2.81 in agreement with the recommendations [−5; 2]. The second RefProt in interaction with AltATAD2, RPL10, has a public model which was obtained on PDB (reference number: 5aj0) [51]. The structure of RPL10 was performed by cryo electron microscopy. RPL10 is found in interaction with several ribosome proteins, forming the 60S subunit. In this model we also found the presence of several messengers and ribosomal RNAs. Thus, from this model, RPL10 could be isolated in order to generate the AltATAD2-RPL10 interaction. However, once this interaction has been obtained, the entire 60S ribosome model is used to correlate the position of AltATAD2 and to hypothesize the function.

Key resources table

Resource	Source	Identifier
CellLine		
HeLa		
Chemical		
Amino acid		
Amino acids		
Cysteins		
FTMS		
ITMS		
Lysines		
Methionine		
Penicillin-streptomycin		
ProteinPeptide		
Protein		

3. Results

3.1. Ghost proteins revealed in nuclei of HeLas cells by XL-MS

Reprocessing of the PPIs from the nuclei of HeLas cells by XL-MS, revealed 1679 cross-link interactions (Supplementary Data 1). Each of these interaction was determined with a minimum score of 20 and a cross-link workflow with FDR of 0.01, limiting the number of false positives. Among these 1679 cross-link interactions, 292 were found to involve Ghost Proteins (Supplementary Data 1, colored Ghost Proteins) including 4 Ghost-Ghost proteins interactions (Table 1). In order to get a visual interpretation, the protein networks were generated under xiNET. To ease the data mining, it was possible to separate the interactions of two, three or more partners. Our interest is to focus on networks involving more than three partners thus facilitating the understanding of the involved signaling pathways. One of the most important network identified was highlighted, which represents the observed interactions between ribonucleoproteins, ribosome subunits, zinc finger proteins, in which RefProts and AltProts interact each other's (Fig. 1). 44 Ghost Proteins in interaction with 7 ribonucleoproteins are observed in this specific network, reflecting the importance of Ghost Proteins in such interactions. Each Ghost Protein is identified by an “IP_” accession number and can be correlated to its transcript number and its associated gene (Table 1). This type of annotation facilitates the identification of the RefProts associated with the mRNA presenting the translated AltProts. We were specifically interested in the networks where Ghost proteins interact with at least two RefProts. Among these, the ghost protein AltATAD2 was found to be in interaction with the RE/poly(U)-binding/degradation factor 1 (AUF1) and the Ribosomal protein 10 (RPL10).

3.2. Comparison of the identified networks for RefProts versus RefProts/AltProts

To assess the influence of the AltProts on the identified cross-links and the protein networks, the identified interactions were compared with the interrogation of the RefProt database alone and with the combined RefProt/AltProt databases (Fig. 2). This comparison shows that a large part of identified interactions are found both after using RefProts database alone and using the combined RefProts/AltProts database (yellow) and correspond to RefProts. It is also observed that a large number of protein interactions are added when the AltProt database is considered which is expected since the AltProt database is larger in size than the RefProt one (green). Finally, a non-negligible portion of RefProts that were identified with the RefProt database alone are not observed anymore when using the combination of the two databases (red). From these data, two main features are derived. The first is that in few cases, proteins initially identified as RefProts become attributed to AltProts by combination of the two databases. The second is that some of the RefProts identified are no longer observed with the combined database interrogation (Fig. 3A). This highlight two important issues. One, is that somehow the current bioinformatics tools seems not to be well-suited to such large databases as the combination of RefProt/AltProt. Indeed the AltProt database has 182,709 entries when the RefProt is only 42,335 entries. In that situation, some of the RefProts fail to pass the FDR threshold. The second is that because some RefProts and AltProts can share a part of their amino acid sequences making proper identification of one or the other difficult. Indeed, if the peptides considered for the identification are only in the common region to the two proteins, and because the AltProt sequences are much smaller by comparison to the RefProts one, the identification weight in favor of the AltProts due to better sequence coverage. The representation of the number of interaction identified per score range (Fig. 3B) shows that interaction that were identified with both databases (RefProt/AltProt) are more confident that those identified only with one of the database (RefProt). These not surprisingly correspond to the proteins that are involved in larger network (Fig. 3A) and identified with a larger number of peptides and interaction. The others (only identified in one interaction) present a relative similar score range. This correspond to proteins identified by only a single interaction. However, in general (Fig. 3) the addition of the AltProt database bring a lot new information to the picture. To assess the veracity of the identify interactions, we have extracted some MS/MS spectra corresponding to the network involving the AltATAD2 protein. Fig. 4 provides examples of MS/MS spectra for two different interactions. For each interaction the CID and the ETD spectra are displayed with the proteins ID, the amino acid sequences and the cross-link sites. More MS/MS spectra can be found in the Supplementary Data 2. The first interaction presented (Fig. 4A) is an interaction between an AltProt and a RefProt which was identified with a score of 40.04. The CID spectrum mainly provides the exact mass of the two peptide chains after the CID cleavage of the DSSO. The annotation of the ETD spectrum show that both cleavages in the two peptide chains are observed and enable confident attribution of the cross-link site. The second example (Fig. 4B and C) presents a case for which an interaction of the Q14103-4 (HNRNPD) RefProt is truly identified but the identification fails to provide the interacting partner with confidence. Indeed, this protein is found to interact with either an AltProt (IP_128579.1) (Fig. 4B) or a RefProt (Q8TF62 i.e. ATP8B4) (Fig. 4C) with scores passing the threshold (> 20) on the two cases. Again, CID spectra provide the exact mass of the 2 peptide chains after CID cleavage of the cross-linker. The careful examination of the ETD spectra show that only 2–3 fragmentations (only 1 for the AltProt) are observed for the peptide chain which is not confidently identified. Despite the two proteins have no sequence homology (Fig. 4D) the peptide MFMVDTKR (Mw_{mono} = 1026.50 Da) of the AltProt with an oxydation of Methionine (+16) has the same molecular weight as the DLDDKYFK peptide (Mw_{mono} = 1042.50 Da) of the RefProt. In that

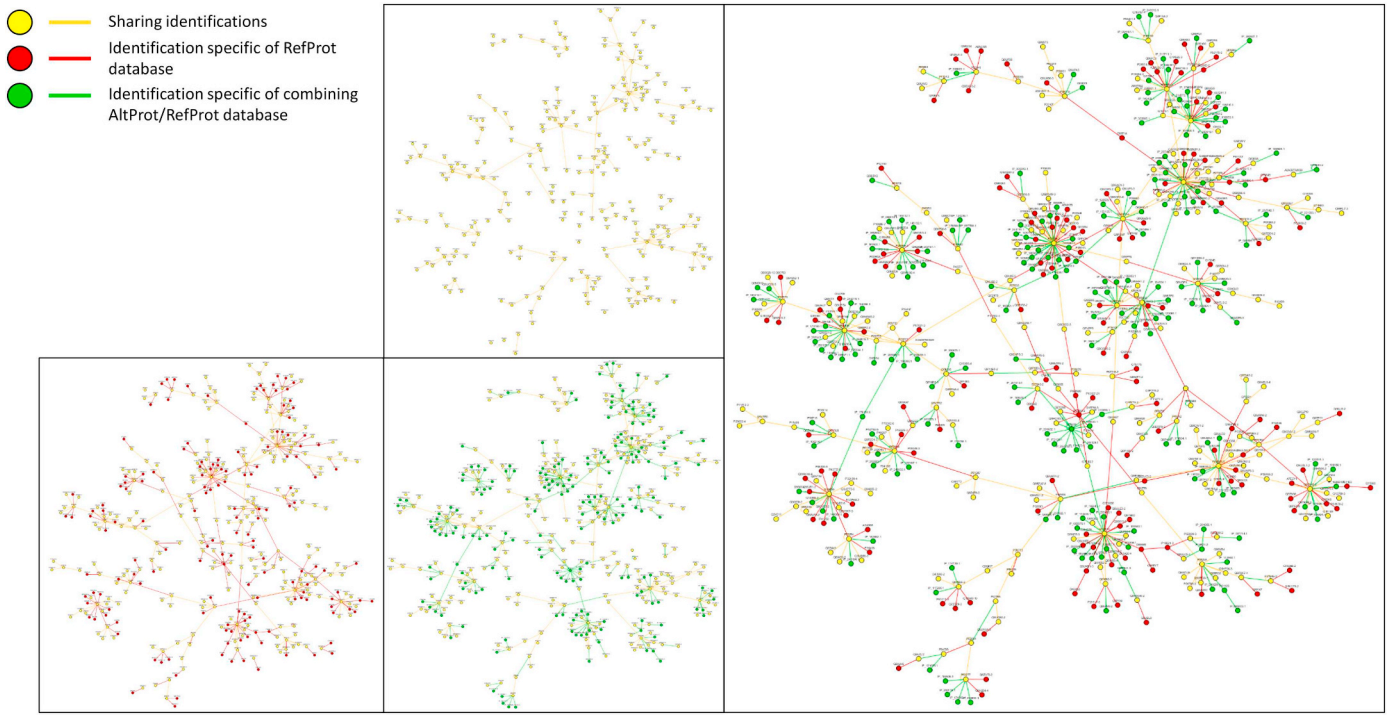


Fig. 2. Cytoscape description of the interaction map obtained from XL-MS data. Data analysis comparison for the RefProt or the combined RefProts/AltProts databases using DyNet apps. In green are the nodes and edges found using the combined RefProts/AltProts databases, in red the identification specific to the RefProt database and in yellow identifications obtained with both the RefProt and the combined RefProts/AltProts databases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 3. Identified interaction obtained from XL-MS data by data interrogation with the RefProt database alone or the combination of the RefProts/AltProts databases. (A) Global mapping of all interactions. Red indicates interactions identified with the RefProt database alone, green with the combined RefProt/AltProt alone and yellow with both RefProt and RefProt/AltProt Databases. (B) Distribution of the identification scores for each cross-link as a function of the number of identified interaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

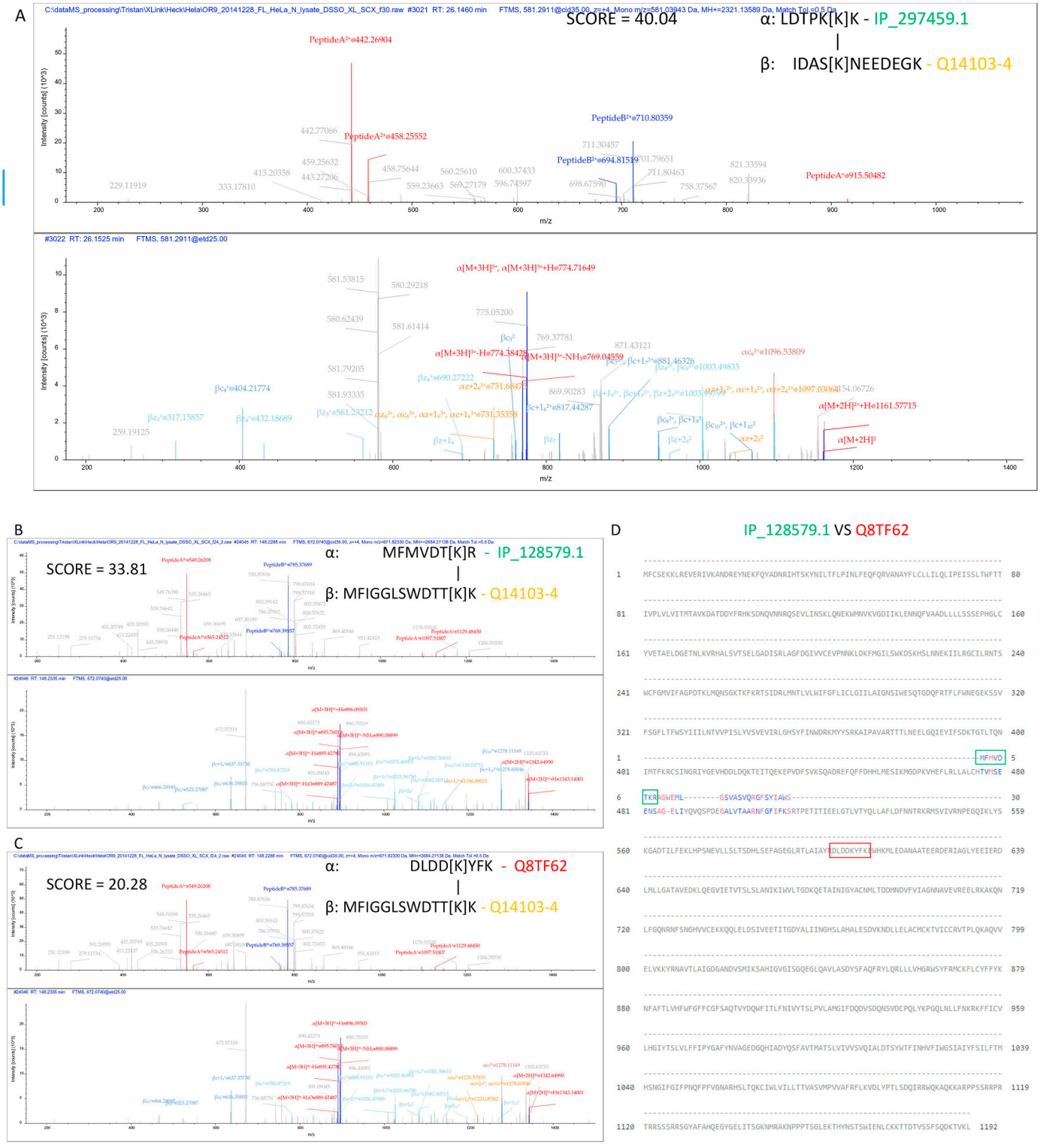


Fig. 4. MS/MS spectra (CID and ETD) with their annotation for identified interaction between RefProt and AltProt. CID/ETD MS^2 spectra of the identified interaction of (A) the RefProt Q14103-4 (HNRNPD) with the AltProt IP_297459.1, (B) the RefProt Q14103-4 (HNRNPD) with the AltProt IP_128579 and (C) the RefProt Q14103-4 (HNRNPD) with the RefProt Q8TF62 (ATP8B4). (D) Sequences alignment of the RefProt Q14103-4 (HNRNPD) and the AltProt IP_128579 showing that the 2 proteins do not share sequence homology.

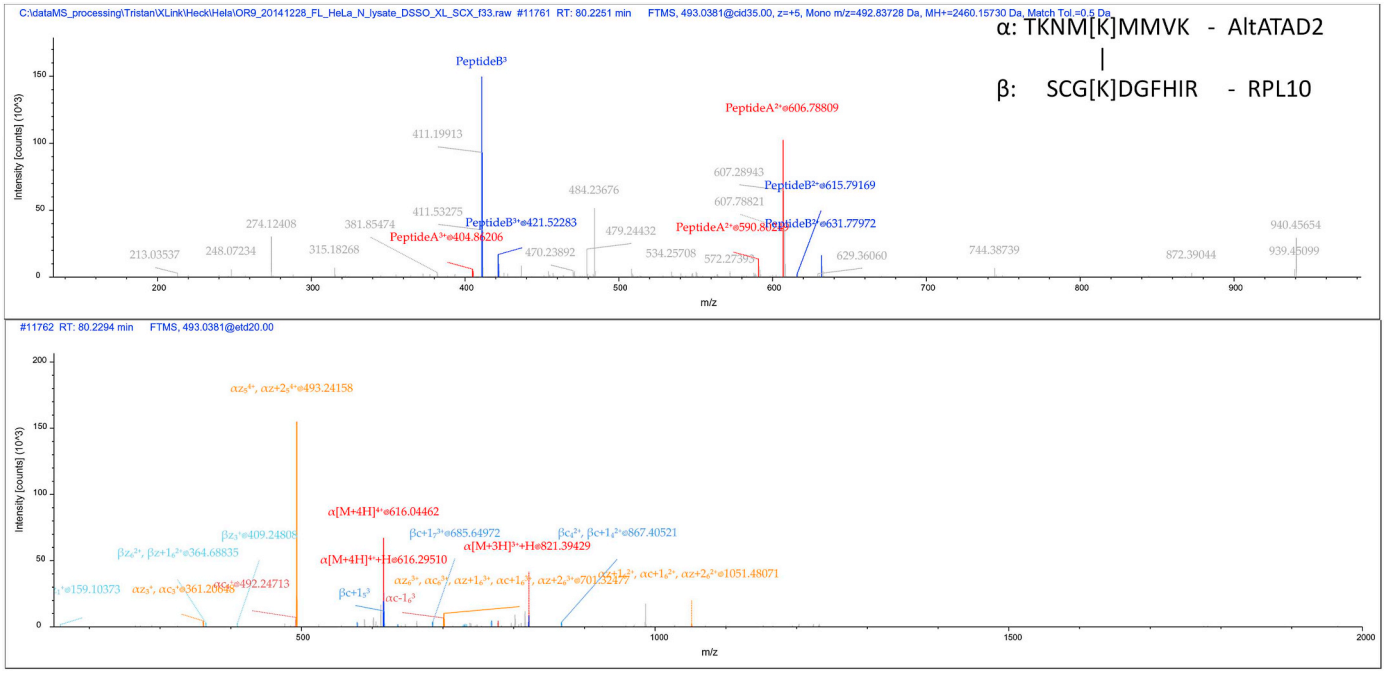


Fig. 5. CID/ETD MS^2 spectra and their annotation of the identified interaction between AltATAD2 and this interacted protein RPL10.

Table 1
Identification of inter-cross-links Ghost Proteins, with a maximum identification score of 50.91 and a minimum of 26.54 these identifications are found among the RefProt-RefProt/Ghost Proteins interactions.

Score	Protein1	PepPos1	PepSeq1	LinkPos1	Protein2	PepPos2	PepSeq2	LinkPos2
50.91	IP_243260.1	5	APRPGNWKWQRR	8	IP_202369.1	28	VGNKSR	4
28.74	IP_222735.1	61	RENKVCVSTWQK	4	IP_093889.1	46	QRAKS	4
28.73	IP_145224.1	2	TIKTKHMIK	5	IP_177042.1	8	LTSRKR	5
26.54	IP_210743.1	24	GGLKTSRDSR	4	IP_138860.1	1	MPATDGKCK	7

case, because no specific fragments are found by ETD on that peptide chain the interacting peptide is only identified by its exact mass. Since the AltProt sequence is much shorter than the RefProt, this positively weight on the identification score in favor of the AltProt (33.81) and lead to its preferential identification. Except for these rare cases, most interaction were found to be trustworthy. For example, Fig. 5 presents the MS^2 spectra for the AltATAD2-RPL10 interaction. Here, the presence of fragments in the two peptides chains give better reliability to the identification.

3.3. AltATAD2 partners

AltATAD2 is found in the CDS with a +2 ORF frame shift and presents a sequence of 139 amino acid residues for a theoretical molecular weight of 17,077 Da (Fig. 6). The structure of this Ghost Protein, was predicted by I-Tasser, based on its amino acid sequence (Fig. 7). The model with the best C-score was retained and used when performing docking by ClusPro2.0 (Fig. 7). AltATAD2 is observed to interact with ARE/poly(U)-binding/degradation factor 1 (AUF1) and Ribosomal proteins (RPL10), two RefProts described in the literature to be involved in different signaling pathways. Docking was carried out between the AltATAD2-RPL10 and AltATAD2-AUF1 proteins, the Ghost Protein being always designated as the ligand of the refprot due to their size difference. For the refprot RPL10 and AUF1 the models were known from previous experiments thanks to structural studies and were retrieved from PDB. The *in-silico* interaction between AltATAD2 and RPL10 mainly shows, two binding sites for AltATAD2 on RPL10 (Fig. 7A). The first binding sites is in the cavity of RPL10 and the second one at the periphery as part of the top 5 best electrostatic structures. These two models were chosen in the best generated models but also taking into account the molecular distance derived from the XL-MS using the DSSO cross linker which is < 50 Å. Similarly, the interaction between AltATAD2 and AUF1 gave two possible interaction sites between the partners i.e. one with the best electrostatic characteristics and the second with the best hydrophobic parameters and considering the distance XL-MS imposed by the cross-linker (Fig. 7B). Finally, AltATAD2 is observed in interaction with these two refProts by fixing different regions. When assembling the docking of AtlATAD2-AUF19RPL10, AltATAD2-RPL10 and AltATAD2-AUF1 by "Match Making" of Chimera, the simultaneous fixation of AltATAD2 and AUF1/RPL10 was found to be feasible (Fig. 7C) resulting in a possible heterodimer biological active complex.

4. Discussion

AltORFs were shown to lead to the translation of AltProts as demonstrated by their observation in the large scale proteomics data [30–32] when using appropriate databases. Very interestingly, the AltProts are also evidenced in the large scale XL-MS data and are found to be interacting with their RefProts counterparts. Observing the AltProts in their interacting network is definitely an approach to get closer to the function of these proteins. Indeed, large scale approach such as XL-MS will provide a global picture for many of these novel proteins without the requirement of developing antibody for each of these proteins as required by the antibody-based strategies. The PPIs highlighted for AltATAD2 is a good example. We have describe an interaction of AltATAD2 with both RPL10 and AUF1XL-MS. AUF1 is a heterogeneous nuclear ribonucleoprotein D (hnRNPD) which was among

Table 2
List of AltProt-RefProt interactions identify in the network (color code is the same as in Fig. 1). For each AltProt the transcript number and gene name from Ensembl database associated with the RNA is given. Each interaction observed in the subdivisions of Fig. 1 is identified.

1				TR	GN	Gene Description	Gene Name
RefProt		RPL10				60S ribosomal protein L10	
		IP_118801.1		NM_001144756.1	10886	neuropeptide FF receptor 2	NPFFR2
		IP_166911.1		NM_014109.3	29028	ATPase family, AAA domain containing 2	ATAD2
2				TR	GN	gene name	
RefProt		HNRNPAB				heterogeneous nuclear ribonucleoprotein A/B	
		IP_066105.1		NM_001002912.4	127254	glutamate rich 3	ERICH3
		IP_135883.1		XR_427728.1	102724275		
		IP_134928.1		NM_014594.1	30832	Zinc finger protein 354C	ZNF354C
		IP_263009.1		NR_028337.1	400624	long intergenic non-protein coding RNA 1973	LINC01973
		IP_257296.1		NM_001160423.1	10642	insulin like growth factor 2 mRNA binding protein 1	IGF2BP1
3				TR	GN	gene name	
RefProt		RBMX				RNA binding motif protein X-linked	
		IP_094161.1		NM_001204.6	659	bone morphogenetic protein receptor type 2	BMPR2
		IP_249315.1		NM_018146.2	55178	RNA methyltransferase like 1	RNMTL1
RefProt		HNRNPA0				heterogeneous nuclear ribonucleoprotein A0	
		IP_303885.1		NM_001163280.1	27336	HIV-1 Tat specific factor 1	HTATSF1
4				TR	GN	gene name	
RefProt		HNRNPA1				heterogeneous nuclear ribonucleoprotein A1	
	HNRNPA2B1					heterogeneous nuclear ribonucleoprotein A2/B1	
		IP_210711.1		NM_022658.3	3224	homeobox C8	HOXC8
		IP_226921.1		NM_001281734.1	53349	zinc finger FYVE-type containing 1	ZFYVE1
		IP_146439.1		NM_001002255.1	387082	small ubiquitin-like modifier 4	SUMO4
		IP_086141.1		NM_032208.2	84168	ANTXR cell adhesion molecule 1	ANTXR1
	IP_124600.1			NR_034075.1	100499177	THAP9 antisense RNA 1	THAP9-AS1
	IP_214370.1			NM_001286262.1	255394	t-complex 11 like 2	TCP11L2
	IP_118616.1			NM_214711.3	401137	proline rich 27	PRR27
	IP_250819.1			NM_198154.1	339168	Transmembrane protein 95	TMEM95
	IP_155900.1			NM_019042.3	54517	Pseudouridylate synthase 7 homolog	PUS7
	IP_174099.1			NM_014290.2	23424	tudor domain containing 7	TDRD7
	IP_097241.1			NM_022817.2	8864	period circadian regulator 2	PER2
RefProt		HNRNPD				heterogeneous nuclear ribonucleoprotein D	
		IP_150105.1		NM_000535.5	5395	postmeiotic segregation increased 2	PMS2
		IP_297459.1		NM_000381.3	4281	midline 1	MID1
		IP_128579.1		NM_000046.3	411	Arylsulfatase B	ARSB
5				TR	GN	gene name	
RefProt		HNRNPC				heterogeneous nuclear ribonucleoprotein C (C1/C2)	
	VDAC2					voltage dependent anion channel 2	
			HNRNPM			heterogeneous nuclear ribonucleoprotein M	
		IP_233089.1		NM_001194998.1	22995	centrosomal protein 152	CEP152
		IP_270709.1		NM_003437.3	7695	Zinc finger protein 136	ZNF136
		IP_297440.1		NM_001256944.1	1183	chloride voltage-gated channel 4	CLCN4
		IP_154990.1		NM_001287054.1	7586	zinc finger with KRAB and SCAN domains 1	ZKSCAN1
		IP_113552.1		NR_103821.1	442075	EMC3 antisense RNA 1	EMC3-AS1
		IP_062261.1		NM_032884.4	84970	chromosome 1 open reading frame 94	C1orf94
		IP_279198.1		NM_052925.2	114823	Leukocyte receptor cluster (LRC) member 8	LENG8
		IP_233136.1		NM_001193489.1	9728	SECIS binding protein 2 like	SECISBP2L
	IP_092512.1			NM_003659.3	8540	alkylglycerone phosphate synthase	AGPS
		IP_217585.1		XR_429051.1	102724196		
			IP_257543.1	XM_006721752.1	201191		
			IP_086598.1	NM_015470.2	26056	RAB11 family interacting protein 5	RAB11FIP5
RefProt		ANXA2				annexin A2	
		IP_085590.1		NM_001122964.2	57223	protein phosphatase 4 regulatory subunit 3B	PPP4R3B
		IP_078517.1		NM_001024226.1	375	ADP ribosylation factor 1	ARF1
6				TR	GN	gene name	
RefProt		AHNAK				AHNAK nucleoprotein	
		IP_123046.1		NM_001166373.1	55016	membrane associated ring-CH-type finger 1	MARCH1
		IP_238136.1		NR_026647.1	791115	Prader-Willi region non-protein coding RNA 2	PWRN2
		IP_094777.1		NM_001039538.1	4133	microtubule associated protein 2	MAP2
		IP_074702.1		NM_000721.3	777	calcium voltage-gated channel subunit alpha1 E	CACNA1E
		IP_124534.1		NR_046377.1	728040	long intergenic non-protein coding RNA 2499	LINC02499
7				TR	GN	gene name	
RefProt		HIST1H1C				histone cluster 1 H1 family member c	
		IP_062513.1		NM_012199.2	26523	Argonaute RISC catalytic component 1	AGO1

the first identified ARE-specific binding proteins (AUBPs) [52]. The AUBPs are complexes of proteins which are involved in the regulation of the AU-rich element (ARE) containing mRNAs. One of the limitations of the experiment here is the possible correlation between a found interaction and the time at which this interaction takes place. Here, XL-MS exhibits a global picture of the protein interaction network in the cell, not enabling to determine when an interaction occurs. As a result, the graphical representation obtained on xiNET gives a common interaction between the three proteins but fails to clarify if they are all together interacting at the same time. To access this information, one would need to phase the cells and performed XL-MS time course analyses. Therefore, several interpretations to this trimer interaction can be advanced. The first one is related to an independent interaction between AltATAD2-RPL10 and AltATAD2-AUF1. The AltATAD2-RPL10 interaction observed by XL-MS using DSSO is confirmed by 3D protein modeling and docking of AltATAD2 on RPL10. RPL10 structure was extracted via the online public model on PDB: 5aj0 from the study of Behrmann et al. [51]. The docking performed on ClusPro highlights several possible fixation sites of AltATAD2 on RPL10; however, only two of them are redundant and in line with the distance limits imposed by the DSSO cross-linker. The first model attaches AltATAD2 at the periphery of RPL10, far from the region fixing the 5S ribosomal RNA. However, it has been shown that RPL10, by its external location on the ribosome, allows the grouping of the subunits and the formation of an active ribosome. Moreover, its interaction with the 60S ribosomal export protein NMD3 would also be responsible for the migration of the peri-ribosome from the nucleus to the cytoplasm [53]. Thus, in this case the RPL10 interaction with AltATAD2 can be directly involved in this peri-ribosome migration. The second model locates AltATAD2 within the ribosome, and more precisely within the region of RPL10 interacting with the 5S ribosomal RNA. In that case, the protein could be involved in the regulation of the binding of the 5S ribosomal RNA (Fig. 8). A previous study by cryoelectron microscopy (cryoEM) has demonstrated that the interaction of RPL10 participates in the ribosome constitution, integrating the proteins RPL5 and RPL11. However, it was shown that RPL10 was not essential for the ribosome formation and functionality. The attachment of AltATAD2 on RPL10 could explain RPL10 regulation function by blocking the 5S rRNA binding site. Another hypothesis is the possible cooperation of the interacting partners with the formation of a co-interaction between RPL10, AltATAD2 and

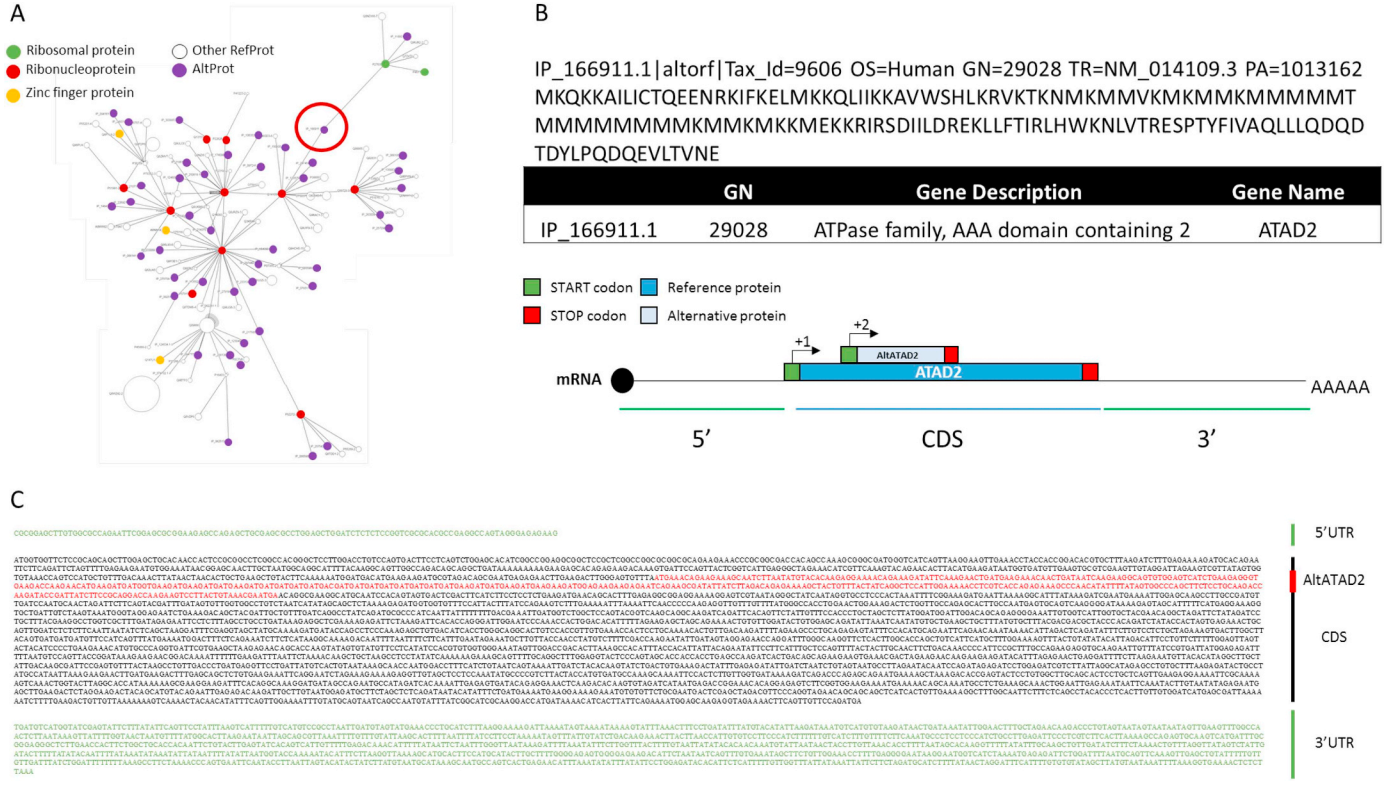


Fig. 6. Focus on AltATAD2 protein. (A) In the interactome network, AltATAD2 is observed to be in interaction with 2 different partners. (B) AltATAD2 amino acid sequence and schematic representation of its sequence location in the mature RNA. AltATAD2 is found in the CDS with a +2 ORF shift. (C) Nucleic acid sequence encoding AltATAD2 within the ATAD2 mRNA.

AUF1. In this scenario, the interaction of AUF1 and RPL10 is not without consequence. Indeed AUF1, was previously described to have a dual function. It is an initiator of the mRNA degradation but as well a protein fixing to the ARE regions of the 3'UTR. On the other hand, RPL10 role is opposite to AUF1. RPL10 is involved in the ribosome assembly and, thus, in the regulation of the translation of mRNAs into proteins (Fig. 9). The co-interaction of AltATAD2 with AUF1 and RPL10 could be the first description of a regulation of the expression of the Ghost proteins resulting from non-coding regions such as the 3'UTR. Since Ghost proteins were not considered before, this could explain why this mechanism was not demonstrated before. Finally, we could hypothesize that the formation of the hetero-trimer, with the attachment of AUF1 on the external part of RPL10, could be involved in a mechanism of regulation of the ribosome. In this case, the RNA5S would fix onto the 60S subunit of the ribosome and activate the transcription. This mechanism would regulate ribosome activation by recruitment of AltATAD2 at the periphery of RPL10 via AUF1 leading to a fine regulation of protein translation. Recently, the so-called Ghost Protein "Nobody", derived from the non-coding RNA: LINC01420/LOC550643, was shown to be involved in the mRNA decapping signaling pathway by interacting with the decapping proteins 4 (EDC4) by multiple techniques including APEX (ascorbate enzymes peroxidase), Photo-cross-link and co-immunoprecipitation [34]. In summary, this study confirms the involvement of Ghost Proteins in the regulation of mRNA expression. The demonstration of an interaction between AltProts and RefProts is a first clue demonstrating their effective role and function in cells. Ghost Proteins are active compounds actively participating to the cell regulation as the RefProts [30].

The proteomic community must widen its field of view to a world, going against the known Kozak dogma of the expression of the proteins, but existing and influencing the known and described models of today. Demonstrating the interaction of these proteins and their involvement in the signaling pathways within the cells is an important step forwards in understanding their functions. Herein, we demonstrate that the XL-MS non targeted large scale approach is useful in this demonstration by revealing the importance of the Ghost Proteins within the interaction networks of RefProts. Although various developments remain to be performed to improve cell interactomic inclusive to AltProts. As demonstrated here for a few percentage of proteins, the size of the total database used by aggregating the AltProt (182,709 entries) to the RefProt (42,335 entries) database, show some limitations of the actual interrogation tools due to the size of this database. Moreover, due to the important number of sequence in this total database and in the cases where only few fragments are observed for one of the peptide chains there is a not unneglectable probability that different peptides match with the same exact mass. This clearly highlight the importance, for such large scale data using such large database for interrogation, to be more stringent on the identification of the interaction and yet to manually check the MS/MS spectra corresponding to the interaction of interest. In general the rate of false positive interactions remain more elevated in the XL-MS approach and confident identification can only be obtained if strictly respecting specific guidelines as reported by t

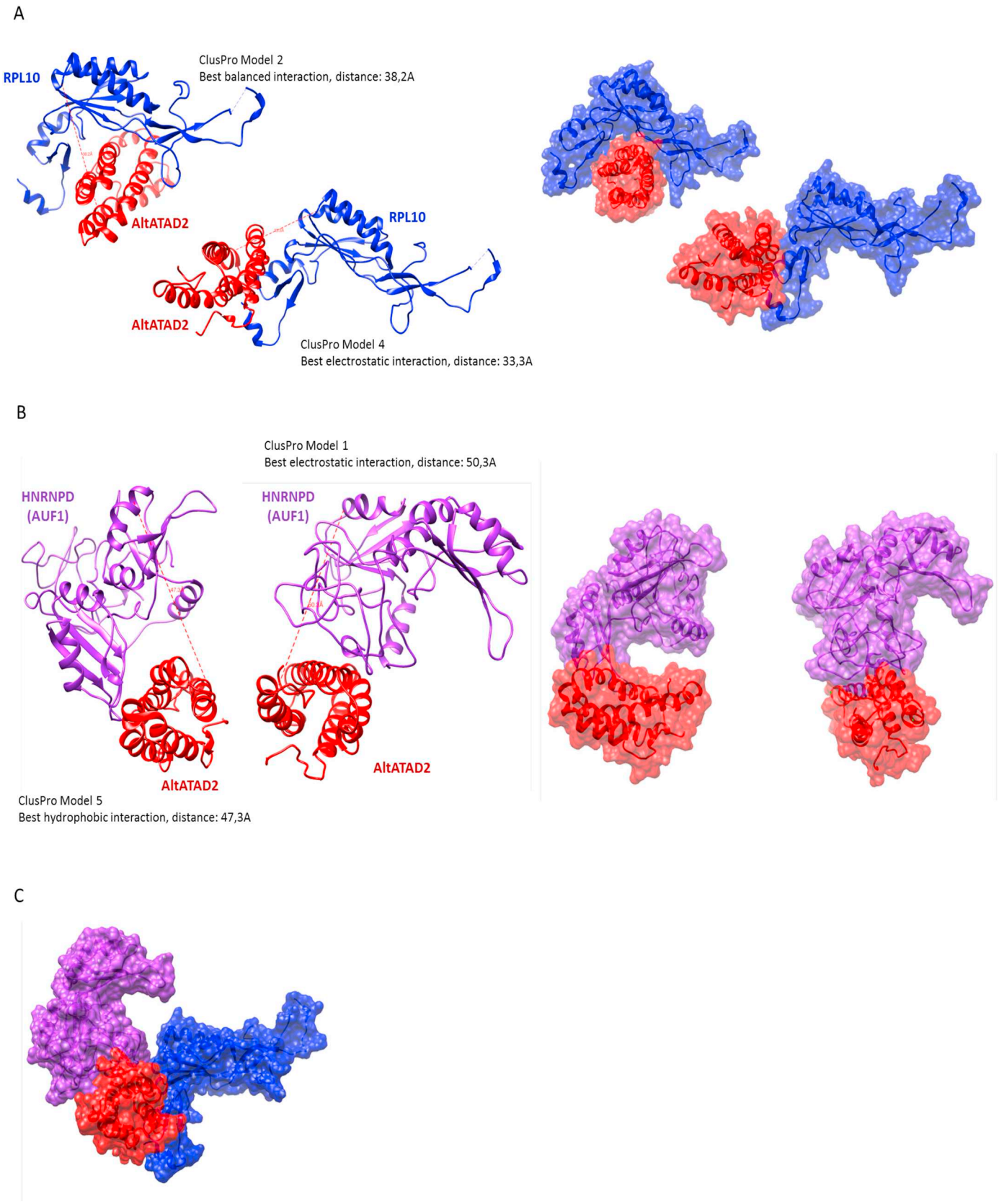


Fig. 7. 3D modeling of the interactions between AltATAD2 and the RefProts AUF1 or RPL10. (A) Models predicted by ClusPro2.0 for the AltATAD2/RPL10 interaction. These two models are part of the TOP5 predictions and are in agreement with the distance restrictions imposed by the XL-MS. Surface modeling was also performed to manually control the likelihood of the result. (B) Predicted models and 3D surface presentations for the AltATAD2/AUF1 interaction selecting predictions with the highest scores in good agreement with XL-MS. (C) 3D model of the co-interaction between AUF1-AltATAD2-RPL10. 3D modeling was used to check that AUF1 and RPL10 are not confused in space.

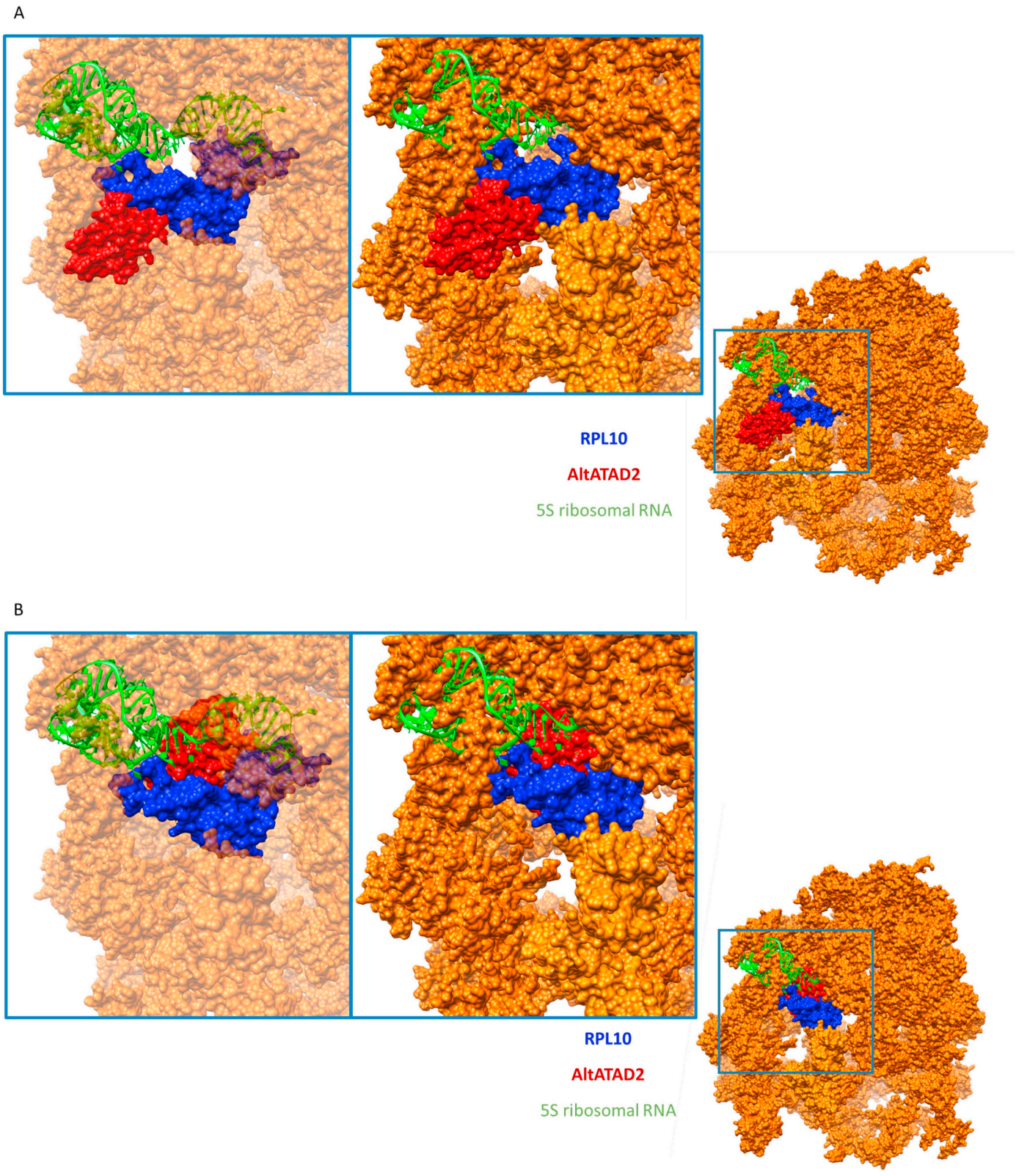


Fig. 8. Implementation of AltATAD2 on the 3D modeling of RPL10 and ribosome 60S obtained by cryoEM. (A) AltATAD2 is found to be interacting at the periphery of RPL10, thus meeting no other subunit of the 60S ribosome or 5S rRNA. (B) On the second position AltATAD2 is observed in the space used by the 5S rRNA. However AltATAD2 does not merge with the position of other subunit of the ribosome 60S. This confirms the ability of AltATAD2 to get into this position.

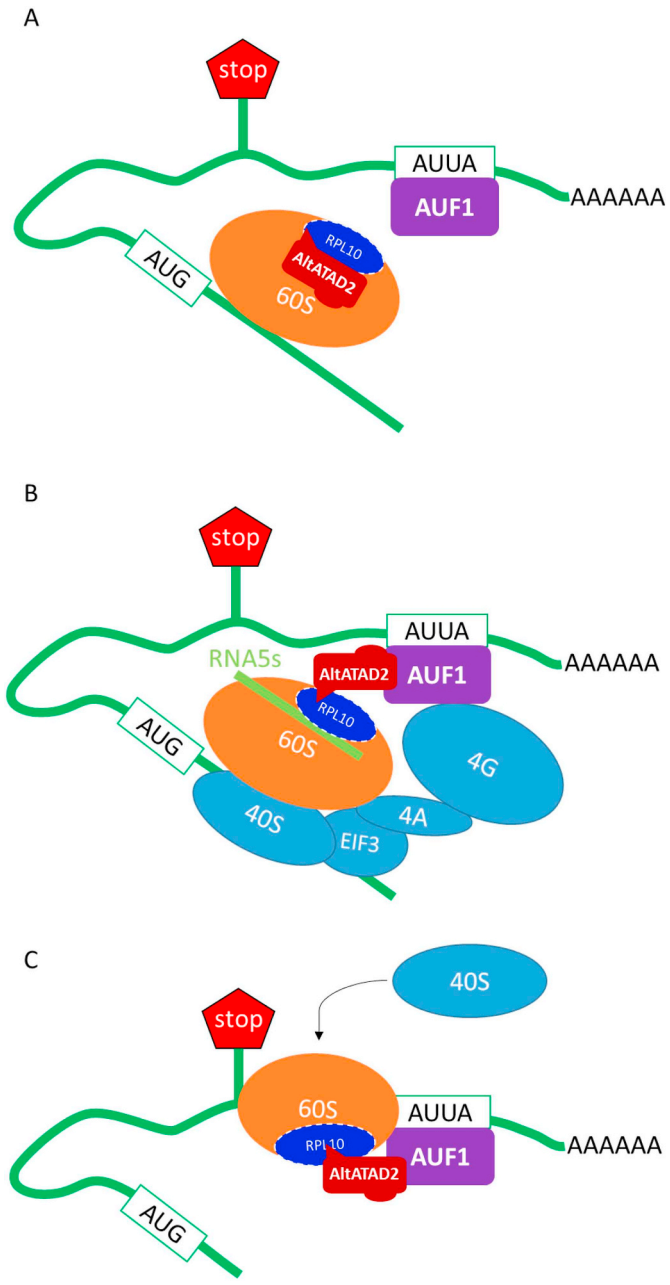


Fig. 9. Schematic representation of the different hypothesized configurations for the co-interaction of RPL10-AltATAD2 and AUF1. All these steps could sequentially exist at different time point to regulate the transcription and the translation. (A) AltATAD2 in internal position on RPL10 prevents the binding of the ribosomal RNA5S. A decrease in binding of RNA5S on the 60S subunit of the ribosome leads to a decrease in the protein translation. (B) AltATAD2 at the outer position on RPL10 allows the formation of the RPL10-AltATAD2-AUF1 complex. In this configuration the RNA5S can fix onto the 60S subunit of the ribosome and activate the transcription. This mechanism would regulate ribosome activation by recruitment of AltATAD2 at the periphery of RPL10 by AUF1 leading to a fine regulation of protein translation. (C) Lastly the interaction of the RPL10-AltATAD2-AUF1 complex takes place in the 3'UTR region and leads to the recruitment of the sub-unit 60S at the ARE to activate the translation of AltProts present in this region.

Iacobucci et al. [54]. However, despite these few limitations, it is clear that large scale interactomics of AltProt will open the way to more complete systems biology pictures [43,55].

Acknowledgements

This research was supported by funding from Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), Institut National de la Santé et de la Recherche Médicale (Inserm) and Université de Lille.

Author contributions

Conceptualization, I.F., J.F. and M.S.; Methodology, I.F., J.F., T.C. and M.S.; Software, T.C.; Validation, I.F., J.F., T.C. and M.S.; Formal analysis, T.C.; Investigation, I.F., J.F., T.C., and M.S.; Resources, I.F. and M.S.; Data curation, T.C.; Writing - original draft T.C. and M.S. Writing - review & editing, I.F. and M.S.; Supervision, I.F., J.F. and M.S.; Project administration, I.F. and M.S.; Funding acquisition, I.F., and M.S.

Declaration of interests

The authors declare no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagen.2019.05.009.

References

[1] F. Meier, P.E. Geyer, S. Virreira Winter, J. Cox, M. Mann, BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes, Nat. Methods 15 (2018) 440–448.
[2] A.-C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature 415 (2002) 141–147.
[3] J.D. Martell, T.J. Deerinck, Y. Sancak, T.L. Poulos, V.K. Mootha, G.E. Sosinsky, M.H. Ellisman, A.Y. Ting, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy, Nat. Biotechnol. 30 (2012) 1143–1148.
[4] S.S. Lam, J.D. Martell, K.J. Kamer, T.J. Deerinck, M.H. Ellisman, V.K. Mootha, A.Y. Ting, Directed evolution of APEX2 for electron microscopy and proximity labeling, Nat. Methods 12 (2015) 51–54.
[5] K.J. Roux, D.I. Kim, M. Raida, B. Burke, A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells, J. Cell Biol. 196 (2012) 801–810.
[6] D.I. Kim, S.C. Jensen, K.A. Noble, B. KC, K.H. Roux, K. Motamedchaboki, K.J. Roux, An improved smaller biotin ligase for BioID proximity labeling, Mol. Biol. Cell 27 (2016) 1188–1196.
[7] E. Coyaud, C. Ranadheera, D. Cheng, J. Gonçalves, B.J.A. Dyakov, E.M.N. Laurent, J. St-Germain, L. Pelletier, A.-C. Gingras, J.H. Brumell, P.K. Kim, D. Safronetz, B. Raught, Global Interactomics uncovers extensive organellar targeting by Zika virus, Mol. Cell. Proteomics 17 (2018) 2242–2255.
[8] S. Eyckerman, K. Titeca, E. Van Quickelberghe, E. Cloots, A. Verhee, N. Samyn, L. De Ceuninck, E. Timmerman, D. De Sutter, S. Lievens, S. Van Calenbergh, K. Gevaert, J. Tavernier, Trapping mammalian protein complexes in viral particles, Nat. Commun. 7 (2016) 11416.
[9] A. Leitner, M. Faini, F. Stengel, R. Aebersold, Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines, Trends Biochem. Sci. 41 (2016) 20–32.
[10] K. Verheggen, H. Raeder, F.S. Berven, Lennart Martens, H. Barsnes, M. Vaudel, Anatomy and Evolution of Database Search Engines-A Central Component of Mass Spectrometry Based Proteomic Workflows, Mass Spectrom. Rev. (2017) 1–15.
[11] UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212.
[12] M. Kozak, Rethinking some mechanisms invoked to explain translational regulation in eukaryotes, Gene 382 (2006) 1–11.
[13] M. Kozak, Initiation of translation in prokaryotes and eukaryotes, Gene 234 (1999) 187–208.
[14] M. Kozak, Regulation of translation in eukaryotic systems, Annu. Rev. Cell Biol. 8 (1992) 197–225.
[15] B.L. Aken, P. Achuthan, W. Akanni, M.R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, L. Gil, C.G. Girón, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, T. Juettemann, S. Keenan, M.R. Laird, I. Lavidas, et al., Ensembl 2017, Nucleic Acids Res. 45 (2017) D635–D642.
[16] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, R. Guigo, GENCODE: producing a reference annotation for ENCODE, Genome Biol. 7 (2006) S4.
[17] D. Thierry-Mieg, J. Thierry-Mieg, AceView: a comprehensive cDNA-supported gene

and transcripts annotation, Genome Biol. 7 (Suppl. 1) (2006) S12.1–14.
[18] R.V. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T.H.-M. Huang, The functional consequences of alternative promoter use in mammalian genomes, Trends Genet. 24 (2008) 167–177.
[19] T.W. Nilsen, B.R. Graveley, Expansion of the eukaryotic proteome by alternative splicing, Nature 463 (2010) 457–463.
[20] D.C. Di Giammartino, K. Nishida, J.L. Manley, Mechanisms and consequences of alternative polyadenylation, Mol. Cell 43 (2011) 853–866.
[21] N.M. Wills, J.F. Atkins, The potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases, RNA 12 (2006) 1149–1153.
[22] V. Delcourt, M. Brunelle, A.V. Roy, J.-F. Jacques, M. Salzet, I. Fournier, X. Roucou, The protein coded by a short open Reading frame, not by the annotated coding sequence, is the Main gene product of the dual-coding gene *MIEF1*, Mol. Cell. Proteomics 17 (2018) 2402–2411.
[23] M.A. Brunet, M. Brunelle, J.-F. Lucier, V. Delcourt, M. Levesque, F. Grenier, S. Samandi, S. Leblanc, J.-D. Aguilar, P. Dufour, J.-F. Jacques, I. Fournier, A. Ouangraoua, M.S. Scott, F.-M. Boisvert, X. Roucou, OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes, Nucleic Acids Res. 47 (2019) D403–D410.
[24] S. Samandi, A.V. Roy, V. Delcourt, J.-F. Lucier, J. Gagnon, M.C. Beaudoin, B. Vanderperre, M.-A. Breton, J. Motard, J.-F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D.J. Hunting, A.A. Cohen, C.R. Landry, M.S. Scott, X. Roucou, Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins, Elife (6) (2017) e27860.
[25] H. Mouilleron, V. Delcourt, X. Roucou, Death of a dogma: eukaryotic mRNAs can code for more than one protein, Nucleic Acids Res. 44 (2016) 14–23.
[26] V. Delcourt, A. Staskevicius, M. Salzet, I. Fournier, X. Roucou, Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA, Proteomics 18 (2018) e1700058.
[27] B. Vanderperre, J.-F. Lucier, X. Roucou, HAltORF: a database of predicted out-of-frame alternative open reading frames in human, Database (Oxford) 2012 (2012) bas025.
[28] Y. Hao, L. Zhang, Y. Niu, T. Cai, J. Luo, S. He, B. Zhang, D. Zhang, Y. Qin, F. Yang, R. Chen, SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci, Brief. Bioinform. 19 (2018) 636–643.
[29] E. Le Rhun, M. Duhamel, M. Wisztorski, J.-P. Gimeno, F. Zairi, F. Escande, N. Reyns, F. Kobeissy, C.-A. Maurage, M. Salzet, I. Fournier, C. Henkel, H. Peter, Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification, (2016), Biochim. Biophys. Acta - Proteins Proteomics 1865 (2017) 875–890.
[30] B. Vanderperre, J.-F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F.-M. Boisvert, X. Roucou, H. Steen, M. Mann, D. Licatalosi, R. Darnell, R. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T. Huang, T. Nilsen, et al., Direct detection of alternative open reading frames translation products in human significantly expands the proteome, PLoS One 8 (2013) e70698.
[31] V. Delcourt, J. Franck, J. Quanico, J.-P. Gimeno, M. Wisztorski, A. Raffo-Romero, F. Kobeissy, X. Roucou, M. Salzet, I. Fournier, Spatially-resolved top-down proteomics bridged to MALDI MS imaging reveals the molecular physiome of brain regions, Mol. Cell. Proteomics 17 (2018) 357–372.
[32] V. Delcourt, J. Franck, E. Leblanc, F. Narducci, Y.-M. Robin, J.-P. Gimeno, J. Quanico, M. Wisztorski, F. Kobeissy, J.-F. Jacques, X. Roucou, M. Salzet, I. Fournier, Combined mass spectrometry imaging and top-down microproteomics reveals evidence of a hidden proteome in ovarian cancer, EBioMedicine 21 (2017) 55–64.
[33] S.A. Slavoff, J. Heo, B.A. Budnik, L.A. Hanakahi, A. Saghatelian, A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining, J. Biol. Chem. 289 (2014) 10950–10957.
[34] N.G. D'Lima, J. Ma, L. Winkler, Q. Chu, K.H. Loh, E.O. Corpuz, B.A. Budnik, J. Lykke-Andersen, A. Saghatelian, S.A. Slavoff, A human microprotein that interacts with the mRNA decapping complex, Nat. Chem. Biol. 13 (2017) 174–180.
[35] C. Lee, J. Zeng, B.G. Drew, T. Sallam, A. Martin-Montalvo, J. Wan, S.-J. Kim, H. Mehta, A.L. Hevener, R. de Cabo, P. Cohen, The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance, Cell Metab. 21 (2015) 443–454.
[36] A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K.I. Nakayama, J.G. Clohessy, P.P. Pandolfi, mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide, Nature 541 (2017) 228–232.
[37] D.M. Anderson, K.M. Anderson, C.-L. Chang, C.A. Makarewich, B.R. Nelson, J.R. McAnally, P. Kasaragod, J.M. Shelton, J. Liou, R. Bassel-Duby, E.N. Olson, A micropeptide encoded by a putative long noncoding RNA regulates muscle performance, Cell 160 (2015) 595–606.
[38] P. Bi, A. Ramirez-Martinez, H. Li, J. Cannavino, J.R. McAnally, J.M. Shelton, E. Sánchez-Ortiz, R. Bassel-Duby, E.N. Olson, Control of muscle formation by the fusogenic micropeptide myomixer, Science 356 (2017) 323–327.
[39] P. Yuan, N.G. D'Lima, S.A. Slavoff, Comparative membrane proteomics reveals a nonannotated *E. coli* heat shock protein, Biochemistry 57 (2018) 56–60.
[40] N.G. D'Lima, A. Khitun, A.D. Rosenbloom, P. Yuan, B.M. Gassaway, K.W. Barber, J. Rinehart, S.A. Slavoff, Comparative proteomics enables identification of non-annotated cold shock proteins in *E. coli*, J. Proteome Res. 16 (2017) 3722–3731.
[41] E. Le Rhun, M. Duhamel, M. Wisztorski, F. Zairi, C.A. Maurage, I. Fournier, N. Reyns, M. Salzet, METB-07classification of high grade glioma using matrix-assisted laser desorption/ionization mass spectrometry imaging (Maldi MSI): interim results of the gliomic study, Neuro-Oncology 17 (2015) v136.3–v136.
[42] Z. Ning, B. Hawley, C.-K. Chiang, D. Seebun, D. Figeys, Detecting Protein–Protein Interactions/Complex Components Using Mass Spectrometry Coupled Techniques, Methods in Molecular Biology (Methods and Protocols), 1164 Humana Press, New York, NY, 2014, pp. 1–13.
[43] O. Klykov, B. Steigenberger, S. Pektaş, D. Fasci, A.J.R. Heck, R.A. Scheltema, Efficient and robust proteome-wide approaches for cross-linking mass spectrometry, Nat. Protoc. 13 (2019) 2964–2990.
[44] F. Liu, D.T.S. Rijkers, H. Post, A.J.R. Heck, Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry, Nat. Methods 12 (2015) 1179–1184.
[45] H. Li, B. Lei, W. Xiang, H. Wang, W. Feng, Y. Liu, S. Qi, Differences in protein expression between the U251 and U87 cell lines, Turk. Neurosurg. 27 (2017) 894–903.
[46] C.W. Combe, L. Fischer, J. Rappsilber, xiNET: cross-link network maps with residue resolution, Mol. Cell. Proteomics 14 (2015) 1137–1147.
[47] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinforma. 9 (2008) 40.
[48] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[49] S.R. Comeau, D.W. Gatchell, S. Vajda, C.J. Camacho, ClusPro: a fully automated algorithm for protein-protein docking, Nucleic Acids Res. 32 (2004) W96–W99.
[50] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF chimera? A visualization system for exploratory research and analysis, J. Comput. Chem. 25 (2004) 1605–1612.
[51] E. Behrmann, J. Loerke, T.V. Budkevich, K. Yamamoto, A. Schmidt, P.A. Penczek, M.R. Vos, J. Bürger, T. Mielke, P. Scheerer, C.M.T. Spahn, Structural snapshots of actively translating human ribosomes, Cell 161 (2015) 845–857.
[52] G. Brewer, An A + U-Rich Element RNA-Binding Factor Regulates c-myc mRNA Stability In Vitro, (1991).
[53] A.W. Johnson, E. Lund, J. Dahlberg, Nuclear export of ribosomal subunits, Trends Biochem. Sci. 27 (2002) 580–585.
[54] C. Iacobucci, A. Sinz, To be or not to be? Five guidelines to avoid misassignments in cross-linking/mass spectrometry, Anal. Chem. 89 (2017) 7832–7835.
[55] F. Liu, P. Lössl, R. Scheltema, R. Viner, A.J.R. Heck, Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification, Nat. Commun. 8 (2017) 15473.

Conclusion

Grâce à la ré-analyse des données, nous avons mis en évidence la capacité de la méthode XL-MS à identifier les partenaires d'interaction des AltProts dans la cellule. Ainsi sur 1679 interactions pontages identifiées, 292 contiennent une AltProt. La réanalyse des données montre l'importance des AltProts dans les réseaux d'interactions et comment ces protéines complètent les voies de signalisation connues. En revanche, cette étude démontre également les limites des outils bioinformatiques actuels qui sont bien adaptés aux banques protéiques disponibles mais dont les limites commencent à être atteintes pour des banques de données plus conséquentes telles que celle obtenue par combinaison de la banque conventionnelle Uniprot et celle des AltProts. Toutefois cette stratégie qui ne nécessite pas l'utilisation d'anticorps dirigés vers une protéine est un atout pour la compréhension de la fonction des AltProts pour lesquelles aucun anticorps n'est encore disponible.

L'étude du cas particulier d'AltATAD2 est intéressante pour démontrer la capacité de ces approches à fournir des informations sur la fonction des AltProts. AltATAD2 est retrouvée comme étant en interaction avec deux RefProts différentes, RPL10 et AUF1. Ces interactions distinctes permettent alors de connecter deux groupes dans le réseau d'interaction. De plus, les fonctions des deux partenaires d'AltATAD2 sont décrites dans la littérature. AUF1 est une « *heterogeneous nuclear ribonucleoprotein* » (hnRNP), capable de fixer les régions riches en AU des ARNm. Elle fait donc partie de la famille des « *AU Binding Proteins* » (AUBP) qui régulent l'expression des ARNm et leur traduction. La deuxième RefProt identifiée dans l'interaction est RPL10, soit la protéine ribosomale L10, constituant le ribosome. Cette protéine n'est pas directement connue pour être impliquée dans son interaction avec l'ARN ribosomal. Toutefois sur la structure 3D obtenue par cristallographie et disponible en ligne sur PDB, elle est représentée dans une cavité, contenant de l'ARNr 5S. Cette proximité spatiale avec l'ARNr 5S et la modélisation de l'interaction avec AltATAD2, montre une imbrication des modèles. En effet, deux positions ont pu être prédites pour AltATAD2 sur RPL10 avec un score positif et en accord avec les distances

contraintes par l'agent pontant. La première est extérieure à RPL10 et la deuxième est à l'intérieure de RPL10. De ce fait dans le premier cas, AltATAD2 est décrite dans la cavité intégrant l'ARNr 5S et ne permettrait pas le recrutement de cet ARN. Dans la deuxième position, l'interaction de l'ARNr 5S avec le ribosome est possible, ce qui nous laisse penser que comme précédemment prédit [23,71] les AltProts interviendraient dans la régulation de l'expression des RefProt.

Cependant, l'interprétation des résultats obtenus à partir des prédictions de structure ne peut fournir une information complète sur la fonction de la protéine. En effet, les structures des RefProt sont obtenues par cristallographie et les interactions observées sont identifiées par analyse XL-MS. Pour les Altprots, il s'agit d'une prédiction de la structure basée sur sa séquence en acides aminés et une prédiction de l'interaction par *docking* qui peut ensuite être confirmée par XL-MS et par les distances imposées par l'agent de pontage. Ces études permettent de fournir une 1ère indication sur la fonction des AltProts et de repérer des AltProts d'intérêt mais ces stratégies ne peuvent évidemment pas se substituer à des études fonctionnelles plus poussées par des approches ciblées impliquant la répression ou la surexpression de la protéine.

Mon objectif était donc de mettre en place et de valider une méthodologie permettant, dans une étude globale d'identifier les AltProts mais également de mettre en évidence leurs fonctions. Cet objectif est possible grâce à la méthode de XL-MS couplée aux analyses bioinformatiques permettant de replacer ces protéines dans leur GO-terme et aux corrélations avec les analyses structurales via les simulations. En effet, l'identification à large échelle permet de détecter dans un premier temps, tout ou partie, des modifications survenues suite à un traitement, une stimulation ou encore dans un contexte physiopathologique. À cette étude est alors ajoutée l'identification des AltProts associées. Bien que cette étape ne nous donne pas directement accès à la fonction de la protéine, connaître la voie de signalisation dans laquelle elle est impliquée est la première étape pour parvenir à sa fonction. De plus, les résultats peuvent être cross-validés par corrélation aux prédictions établies. Cette approche permet notamment de confirmer que certaines AltProts sont impliquées dans la

régulation de la traduction des protéines et seraient donc un niveau supplémentaire dans la régulation de l'expression des protéines. C'est pourquoi, nous avons par la suite associé la méthodologie présentée dans ce chapitre, à l'interprétation, l'enrichissement des réseaux d'interactions par les GO-Terms associés. Cet enrichissement permet une lecture globale des résultats d'interaction XL-MS obtenus en y intégrant les AltProts et les RefProts.

PARTIE V

Application des Stratégies XL-MS à l'Identification des fonctions des AltProts dans le cadre de la reprogrammation des cellules cancéreuses

I. Mise en évidence d'une fonction protéique

1. Fonction associée à l'homologie de séquence

Afin de mettre en évidence la fonction d'une protéine nouvellement identifiée, plusieurs méthodes peuvent-être utilisées. Pour prédire cette fonction, alors que peu d'études existent sur la cible recherchée, il est possible de réaliser une recherche de domaines spécifiques. La réalisation d'un alignement de séquence avec l'ensemble des protéines connues permet de mettre en évidence, par un jeu de couverture de séquence et d'identité entre les séquences, la mise en évidence d'une homologie entre la protéine cible et les protéines comparées. Ceci est réalisable grâce à des algorithmes tel que : HMMER [136] et BLAST [137]. Ces homologies permettent de trouver des domaines représentatifs d'une famille de protéines. Ces domaines sont généralement responsables des fonctions des protéines, car ils représentent les acides aminés impliqués dans les interactions protéiques ayant pour effet de modifier l'un des deux partenaires comme la phosphorylation. Ils peuvent être liés à une fonction enzymatique ou encore à une autre fonction spécifique. Certains outils permettent de chercher ces domaines tels que Prosite [138], Pfam [139] ou encore InterProScan [140].

La découverte de la présence d'un domaine commun entre une cible inconnue et une famille de protéines connue, permet de supposer le rôle de la cible. Toutefois, la recherche d'une fonction à partir de la découverte d'un domaine, reste limitée pour les AltProts. Au vu de leurs petites tailles l'identification de domaine n'est parfois pas possible ou parfois incomplète. En effet les domaines identifiés sur les RefProt ont une taille pouvant varier de 50 à 500 acides aminés, avec un maximum entre 100 et 150 [141]. Les AltProts avec une longueur moyenne de 55 acides aminés ne présentent en général qu'une toute petite partie des domaines connus. Ainsi, l'utilisation de domaines connus pour des fonctions particulières obtenus par homologie de séquence est souvent de peu d'aide et bien loin d'être suffisante pour comprendre la fonction des AltProts.

2. Fonction associée au réseau

L'utilisation des réseaux PPIs, est une approche démontrée pour la découverte de fonction de nouvelles protéines [75,142,143]. Plusieurs méthodes d'analyses de données peuvent être utilisées, mais toutes reposent sur la même base de mise en évidence de connexions entre des protéines identifiées. Sur la connaissance de ces liens entre protéines, deux approches peuvent être alors utilisées :

- L'approche directe qui tient compte de l'existence d'une interaction entre deux partenaires afin d'associer une fonction. Ainsi dans un réseau, plus deux protéines sont proches, plus elles sont similaires et donc possèdent une fonction comparable voire identique.
- L'approche assistée dans laquelle le réseau est analysé dans son ensemble. C'est la formation de groupes de protéines par des liens plus ou moins proches qui vont définir la fonction.

Une analyse spatiale de la répartition est effectuée afin de réaliser un regroupement selon des paramètres définis par l'expérimentateur. La limite réside donc dans les paramètres choisis pour l'algorithme de regroupement et d'association des protéines identifiées entre elles. Cette méthode possède l'avantage de pouvoir prédire la fonction de protéines sans lien direct avec la fonction identifiée [142] (**Figure 17**).

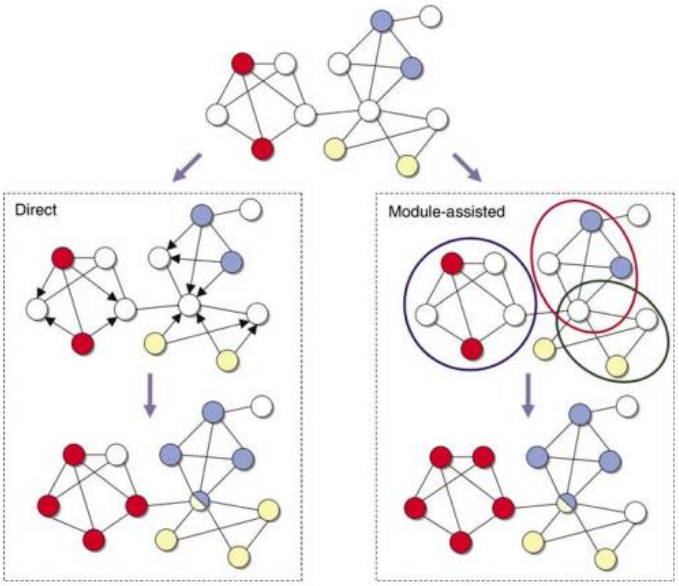


Figure 17 : ***Méthode d'analyse de réseaux d'interactions.*** *Deux stratégies sont décrites dans l'étude spatiale d'un réseau, la première directe regroupe dans une même classe les partenaires du réseau ayant une connexion établie et orienté vers le même nœud. La deuxième étudie la forme et la proximité des partenaires dans l'espace afin de leur attribuer une classe (Sharan & al., 2007* [142])

II. Prédiction de la fonction des AltProts

Peu d'études traitent de la fonction attribuée aux AltProts et aux SEPs. Récemment une méthode proposant un outil de prédiction de la fonction des AltProts a été proposée, le « *smORF-encoded peptides predictor* » (FSPP) [144]. Cet outil utilise la détection des AltProts dans différentes expériences, regroupant ainsi l'identification en MS, mais combinant aussi les informations de transcriptomique et de génomique. L'ensemble de ces informations forment un réseau permettant de rapprocher les AltProts détectées des RefProt identifiées et ainsi leurs prédire une fonction. Cet outil présente une réussite de prédiction sur 3 études décrivant la présence d'AltProts : AltHER2 (HER2_uORF), AltMKKS (MKKS_uORF) et AltIFRD1 (IFRD1_uORF), corrélant la prédiction faite par FSPP avec les observations des auteurs originaux des AltProts citées. Cette méthode est toutefois limitée pour les études à grande échelle. Elle nécessite un grand nombre d'informations *i.e.* les données de protéomique avec les analyses MS des échantillons, les bases de données adaptées, la connaissance de la localisation subcellulaire des AltProts identifiées et des RefProts afin de réaliser

le réseau. Elle nécessite également des informations génomiques, décrivant encore une fois la localisation spatiale des ARNs dans la cellule, le séquençage des cellules étudiées et les prédictions de transcription et de traduction.

L'utilisation de la méthode XL-MS et des outils de traitement de réseaux cytoscape et ClueGo, permettent l'observation d'interactions reliées aux voies de signalisation de manière spécifique et à large échelle. Cette combinaison nous permet ainsi, grâce à l'observation des interactions dans la cellule par XL-MS, et par l'analyse de réseaux enrichis par les interactions connues dans la littérature, d'être entre la prédiction étendue de fonction par méthode assistée et la méthode directe permettant d'apporter une estimation fiable de la fonction des AltProts identifiées.

III. Objectif

La mise en place d'une nouvelle méthodologie rendant accessible la fonction des AltProts, par la combinaison de l'analyse XL-MS et du retraitement de données des réseaux obtenus, nous a permis de mettre en évidence la présence d'AltProts dans différentes voies de signalisation. Cette stratégie a été appliquée à l'étude d'une lignée cellulaire humaine de gliomes de grade IV (lignée NCH82) dans un contexte de reprogrammation des cellules cancéreuses. L'observation d'AltProts dans des réseaux d'interactions tels que la constitution du cytosquelette, m'a amené à observer les modifications de ce réseau et des partenaires d'interaction dans des conditions d'induction de la transition métastasique des cellules gliomales induite par la Forskoline. Peu d'informations moléculaires sont disponibles pour la lignée NCH82, comparée à la lignée U87, plus répandue. Cependant, cette lignée est issue d'un glioblastome caractérisé tandis que les U87 sont d'origine inconnue [145]. Ces cellules ne présentent pas de modifications phénotypiques particulières sous Forskoline. Toutefois, on observe un changement au niveau moléculaire par protéomique de la lignée après traitement. L'application de la méthodologie présentée au chapitre précédent dont la méthode XL-MS permet alors d'identifier un grand nombre d'AltProts spécifiques soit dans les conditions contrôles, soit sous traitement à la Forskoline. L'identification de pontage entre les AltProts et les RefProts permet

alors de rattacher l'identification de la fonction des RefProts pour en déduire la fonction des AltProts. Un enrichissement grâce à la base de données STRING m'a notamment permis de corréler des groupes de protéines entre elles afin d'augmenter l'information obtenue par XL-MS. Cela nous a permis de mettre en évidence l'implication des voies de signalisation des ARNt et des protéines associées sous stimulation à la Forskoline dans la transition métastasique. Ainsi pour la première fois une fonction a pu être attribuée à des AltProts, via la présence d'interactions spécifiques avec des RefProts, dans un contexte pathologique.

Alternative Proteins are Functional Regulators in Cell Reprogramming by PKA Activation

Tristan Cardon[1§], Julien Franck[1§], Marina Damato[1,2], Michele Maffia[2], Daniele Vergara[2], Isabelle Fournier[1*] and Michel Salzet[1*]

[1]Université de Lille, Inserm, U1192, Laboratoire Protéomique, Réponse Inflammatoire et Spectrométrie de Masse (PRISM), F-59000 Lille, France
[2]Department of Biological and Environmental Sciences and Technologies, University of Salento, 73100 Lecce, Italy

* To whom correspondence should be addressed. Michel Salzet and Isabelle Fournier, Tel: +33 320 43 41 94; Fax: +33 320 43 40 54; Email: michel.salzet@univ-lille.fr and isabelle.fournier@univ-lille.fr

§The authors wish it to be known that, in their opinion, the first x authors should be regarded as joint First Authors

ABSTRACT

It has been recently shown that many proteins are lacking from reference databases due to the fact that they are translated from alternative ORFs (AltORFs) opposing the rules decreed for the genome annotation for protein translation from mRNA. Of interest, the function of these Alternative Proteins (AltProts) remains largely unknown. Here, we are investigating the function of these AltProts in the context of cancer cell reprogramming. We have developed a large scale approach based on shot-gun proteomics and cross-linking mass spectrometry (XL-MS) to understand and decipher AltProts regulation, their functions and interaction partners in reference to proteins (RefProts). The study was performed on NCH82 human glioma cells which were stimulated by the protein kinase A activator Forskolin at different time points of 16H, 24H and 48H to induce cell differentiation and epithelial-mesenchymal transition. The data have shown enabled us to trace back the function of the AltProts and regulation achieved by combining experimental data to in silico analysis using Cytoscape and ClueGo determining gene ontology annotation and pathways enrichment with STRING analysis. Interestingly, results from this work has indicated that many AltProts demonstrate functionality related to the regulation of tRNA through their interaction with aaRS proteins and of cellular mobility.

Key Words: Alternative proteins, Proteomics, Cross-linking mass spectrometry, Cell reprogramming

INTRODUCTION

It is conventionally accepted that eukaryotes mature messenger ribonucleic acids (mRNAs) are monocistronic, leading to the translation of a single protein. According to the rules decreed by Kozak (1) the coding DNA sequence (CDS) region corresponds to the longest nucleotide sequence flanked by a START and a STOP codon which defines the reference open reading frame (RefORF) that is translated into a protein. Nevertheless, the proteome was shown to be more complex than initially expected. Indeed, more than 10% of proteomic data remain unmatched by reference databanks interrogation although the quality of the MS/MS data (number of characteristic fragment ions) is sufficient to lead to protein identification. These unmatched data were used to demonstrate that proteins from alternative open reading frames (AltORFs) are translated from mRNAs in addition to the predicted proteins from RefORF (2). These AltORFs lead to the so-called Alternative proteins (AltProts) that altogether can be considered as a hidden or ghost proteome. AltProts can translated from different parts of the mRNAs including 5'UTR, 3'UTR, overlapping regions between 5'UTR and CDS or CDS and 3'UTR as well as from +2 and +3 frameshifts in the CDS. Therefore, they are separated into 3 groups named AltORF5'UTR, AltORFCDS and AltORF3'UTR. It is expected that about 59% of AltORFs and about 41% of RefORFs coexist post-translation which was confirmed by proteomic data (2, 3). On the other hand, high-throughput genome and transcriptome sequencing have led to validate an important number of small ORFs (sORFs) containing less than 100 codons which were arbitrary considered as non-coding. However, the translation of these sORFs into peptides or small-proteins was demonstrated by different strategies such as ribosome profiling (4, 5) and combined peptidomics to massively parallel RNA-seq (6). Since the protein databanks used in large scale proteomic approaches such as Uniprot (7) or NCBI (8) are deduced from genomic data, they miss these AltProts. This explains why these proteins were ignored so far. With the evidence of this hidden proteome, the genome annotation and protein translation dogma has to be reconsidered and the sequence of the AltProts need to be predicted and included in the different databases. Over the past decade, there have been several initiatives to predict the AltProts, and recently the Openprot database was publicly released which can be used to search for AltProts from proteomic large scale data (9). On average the AltProts are of lower molecular weight in comparison to their reference counterpart with ~57 against ~344 amino acids for the RefProt issued from the canonical reading frame (2, 10). Before the advent of advanced proteomics and the emergence of databases, few AltProt findings were described; however, they were considered as an epiphenomenon; despite the fact that they show central biological functions (11). With the use of large scale proteomics in combination to AltProts database interrogation, it was possible to demonstrate the central role of these proteins in physiological and physiopathological signaling pathways (e.g. ovarian cancer (12), brain physiome (13) and viral infection (14)). It was also shown that AltProts could be present at higher levels than their reference counterparts but with shorter expression half life (10).

However, if many AltProts have been identified, their functions remain largely unknown; indeed, very few studies report on the function of AltProts. Recently, it has been shown that the "Humanin" AltProt could be involved in the longevity of the Asian population (15) while other AltProts could play a role in the regulation of the metabolism (16). The acquired data and the small size of AltProts indicate a potential role in the regulation of the protein expression but this hypothesis must be confirmed. But

searching for the function of these AltProts is clearly not an easy task since no antibody raised against them are commercially available. Therefore, conventional strategies such as co-immunoprecipitation (coIP) cannot be used. Moreover, candidate by candidate strategy is very time consuming and can not depict the global picture of AltProts. In that sense, large scale approaches are more favored. In particular, finding the interaction partners of AltProts and including them in known signaling pathways is one approach to deduce their functions.

Over the past decade, various methods have been developed to measure protein-protein interaction (PPI). MS-based proteomics strategies are particularly well-suited for measuring PPIs at larger scale from a limited amount of complex mixture (17). This includes affinity-purification (18) and tandem-affinity purification (19, 20), proximity labeling (21) such as APEX (22) and BioID (23–25), viral particle sorting approach (Virotrap) (26) and cross-linking MS (XL-MS) (27, 28). XL-MS is based on the formation of covalent bonds using a chemical linker of defined length to freeze the interaction between partners. The complex can then be submitted to enzymatic digestion and the peptides further analyzed by LC-MS. The MS/MS information gathered during the LC-MS run is then used to identify the interacting peptides and deduce the PPIs. Both intra-protein and inter-protein crosslink are observed depending on the protein conformation and interaction and the length of the cross-linker. This information can be used not only to decipher interacting partners but to predict the structure of the protein in combination with data from other modalities such as molecular modeling, X-Ray crystallography, NMR and cryoEM. If XL-MS appears to be a straightforward strategy with very wide application range, the strategy development has faced several challenges. The first challenge is related to the difficulty of interpretation of the generated MS/MS data from two peptides cross-linked together. This can be overcomed by using CID cleavable cross-linkers such as the commercialized DSSO (29) and DSBU (30). Another issue was found to be low abundance of cross-linked peptides compare to non-crosslinked ones limiting their measurement during the LC-MS run. Various protocols have been proposed for the enrichment in crosslinked peptides including size exclusion chromatography (SEC), ion-exchange using either cartridges or columns (31) or affinity purification for tagged crosslinkers (32–34). Finally, the data analysis is more complex than for conventional shot-gun proteomics and requires dedicated software solution and platforms (35–40). With respect to data analysis, finding PPIs from all identified proteins versus the entire protein database is still not possible due to the lack of computational power but searching for the partners of candidates is feasible. Another aspect of the data analysis is to be careful about the possible misassignments (41). More advanced tools are developed such as *XLinkX* (42) computing node which can be included within the usual proteomic identification tool *Proteome Discoverer 2.2*. Robust XL-MS workflows are now in use by different groups (43–46). Along the same line, It was shown that XL-MS could be used *in cellulo* by choosing appropriate cross-linkers which can cross cell or cell compartment membranes (47, 48). For AltProts, XL-MS has the great advantage of being totally untargeted without even the requirement of starting from a candidate used as a bait, as for most of the other proteomics-based PPIs strategies. We recently analyzed HeLa cells data (49) to demonstrate the importance of AltProts in the signaling pathways. This analysis confirmed the potential of XL-MS strategies to get new insights into AltProts functions (49).

In this work, we applied this approach to study the role of the AltProts during cancer cells reprogramming. To address this question, we stimulated glioma cells with the protein kinase A (PKA) activator Forskolin. Activation of PKA has been shown to promote cell differentiation by the activation of epithelial-mesenchymal transition (EMT) or its reversal process mesenchymal to epithelial transition (MET) as previously demonstrated (50, 51). Combined shot-gun proteomics and XL-MS were used to identify the AltProts, their interaction partners, and the signaling pathways. An XL-MS workflow was optimized from the NCH82 human grade IV glioma cells to enable XL-MS to be performed separately form cell nucleus and cell cytoplasm, enabling the identification of AltProts and their interacting partners. The resulting network can then be enriched based on the interactions described in the literature. By comparison of these networks and their combination, the functions can then be assigned to the identified AltProts as a result of the biological process in which they are involved. For the first time, the demonstration of AltProts changes in response to PKA activation is provided.

MATERIAL AND METHODS

Chemicals & Materials

Disuccinimidyl sulfoxide (DSSO) Dulbecco's modified Eagle's medium (DMEM), fœtal bovine serum (FBS), L-glutamine, penicillin, streptomycin, phosphate-buffered saline (PBS) were obtained from Thermo Fisher Scientific (Les Ulis, France). Formic acid (AF), HPLC grade water, trifluoroacetic acid (TFA), acetonitrile (ACN) methanol (MeOH), ethanol (EtOH), acetone and trichloroacetic acid (TCA) were all purchased from Biosolve (Dieuze, France). DL-dithiothreitol (DTT), iodoacetamide (IAA), chloroform, dimethylsulfoxide (DMSO) ammonium bicarbonate (AB) 4-(2-Hydroxyethyl)piperazine-1-ethane sulfonic acid, N-(2-Hydroxyethyl)piperazine-N-(2-ethane sulfonic acid) (HEPES), Sodium Chloride (NaCl), magnesium chloride (MgCl) were obtained from Sigma Aldrich. Tris was purchased from Bio-Rad (Steenvoorde, France). Extraction Illustra triplePrep Kit was from GE Healthcare. LysC/Trypsin was obtained by Promega (Charbonnières-les-Bains, France). Amicon centrifugal filters and C18 ZipTip pipette tips were from MERCK Millipore (Merck KGaA, Darmstadt, Germany).

Cell culture

NCH82 and U87-MG human glioma cell lines were cultured as monolayers until 80 – 90% of confluence in high glucose Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% heat-inactivated Foetal Bovine Serum (FBS), 100 U/ml penicillin, 100 µg/ml streptomycin and 2mM L-glutamine (Sigma-Aldrich), before harvesting and passing in a new flask. The cells were kept at 37 ℃ in humidified air containing 5% CO2. NCH82 cells (1 x 106 / T75 flask) and U87MG (2 x 106 / T75 flask) were stimulated in DMEM complete medium for 48h supplemented with 50 µM of Forskolin (BIOTREND, Chemicals AG), a cell-permeable activator of adenylyl cyclase, that leads to an increase in intracellular concentration of cAMP and, consequently, to a Protein Kinase A (PKA) stimulation. Forskolin was prepared as 50 mM stock solution in DMSO.

Shotgun proteomics

For whole cells analysis, protein extraction was carried out using the Illustra triplePrep Kit (GE Healthcare 28-94259-44) to separate DNA, RNA and proteins. The isolated protein fraction was kept for large scale shot-gun proteomics. The protein extraction, reduction/alkylation and enzymatic digestion was performed using the FASP method (51). Briefly, the sample was taken up in 30 µL of 8M urea in 0.1 M Tris / HCl, pH 8.5 (UA buffer) and an equivalent volume of 100 mM in UA DTT. The sample was then incubated for 40 minutes at 56°C. Total proteins were loaded onto 10 kDa Amicon filters, supplemented with 200 µL of UA buffer and centrifuged for 15 min at 14,000 g. Then, 100 µL of a 0.05 M IAA in AU were added and incubated for 20 min in the dark before centrifugation for 15 min at 14,000 g. Finally, a 0.05 M ammonium bicarbonate solution in water (AB) was added and centrifuged again for 15 min at 14,000 g twice. For the digestion, 50 µL LysC/Trypsin at 20 µg/mL in AB buffer was added and incubated at 37°C overnight. The digested peptides were then recovered after centrifugation for 15 min at 14,000 g after transferring the filter into new tubes, reconstitution in 50 µL of AB buffer followed by a second centrifugation step for 15 min at 14,000 g. The eluted peptides were then acidified with 10 µL of 0.1% TFA and vacuum dried.

Cellular cross-linking

The separation between cytoplasm and nuclei was carried out according to the method described by Liu and al. (19). Briefly, the cell pellet was recovered in 100 µL of stabilizing buffer (10 mM HEPES, 10 mM KCl, 1.5 mM MgCl 2, 0.5 mM DTT, 0.4% NP-40, pH 7.8) containing protease inhibitors (AEBSF 2mM, phosphoramidon 1 µM, Bestatin 130 µM, 14 µM E-64, 1 µM Leupeptin, 0.2 µM Aprotinin, 10 µM pepstatin A); incubated on ice for 10 min and centrifuged at 3200 g for 10 min. The supernatant was discarded, and the pellet was taken up in 100 µL of buffer (20 mm HEPES, 150 mm NaCl, 1.5 mm MgCl 2, 0.5 mm DTT, pH 7.8) containing the protease inhibitor. The cells were lysed by sonication performing 3 cycles of sonication of 30 seconds each at 50% of the maximum power on the ice. The extract was then centrifuged for 20 min at 13,800 g to remove cell debris. 500mM stock solution of DSSO cross-linker was prepared in DMSO. The cross-linking reaction was performed on the nuclear fraction at a concentration of 1mM DSSO cross-linker by addition of 2 µL of the DSSO stock solution to 100 µL of the sample, in estimation of 100-fold excess in crosslinker for the proteins. The reaction was then carried out at RT by gently stirring the solution and stopped after 1H by adding 2 µL of 500 mM Tris pH 8.5 solution and gentle stirring for 20 min, for quenching the NHS function of the crosslinker. After the reaction was stopped, the sample was vacuum dried. Crosslinked proteins were suspended in 30 µL of 8M urea, loaded onto 30 kDa Amicon filters and processed according to the FASP procedure described in the previous section.

LC-MS/MS analysis

The samples were reconstituted in 20 µL of a 0.1% TFA and desalted using a C18 ZipTip (Millipore, Saint-Quentin-en-Yvelines, France). After elution with 20 µL of 80% ACN/0, 1% TFA from the ZipTip, the sample was vacuum dried. For the LC-MS, samples were then reconstituted in 0.1% FA in water /ACN (98:2, v/v), and separated by reverse-phase liquid chromatography (RPLC) using a nanoAcquity UPLC equipped with a C18 pre-column (180 µm ID × 20 mm length, 5 µm PD, Waters) and a Peptide

BEA C18 column (25 cm length, 75 µm ID, 1.7µm PD, Waters). Separation was performed using a linear gradient starting at 95% solvent A (0.1% FA in water) and 5% solvent B (0.1% FA in ACN) up to 70% solvent A and 30% solvent B for 120 min at 300 nL/min. The LC system was coupled onto a Thermo Scientific Q-Exactive mass spectrometer set to acquire the ten most intense precursors in data-dependent acquisition mode, with a voltage of 2.8 kV. The survey scans were set to a resolving power of 70 000 at FWHM (m/z 400), in positive mode and using an AGC target of 3E+6. For the shot-gun proteomics, the instrument was set to perform MS/MS only from >+2 and <+8 charge state but for XL-MS were larger peptides are measured only >+3 charge state ions were selected excluding unassigned load states, +1, +2 and > +8 .

Shot-gun proteomics data analysis

RAW data obtain from the nLC-MS/MS run were treated using MaxQuant V1.6.1.0 using the LFQ annotation of the protein identified. UniProtKB database for reviewed human of April 2018 containing 20303 protein sequences was used. Statistical analyses were carried out using Perseus software after filtering for “reverse”, and “contaminants” proteins. For the comparison between control and Forskolin-treated groups, t-test was performed with a permutation - based FDR of 0.05, and p values less than 0.05 were considered to be statistically significant. A heat-map of differentially expressed proteins across the two different groups was also generated. Gene ontology (GO) analysis was performed using ClueGO on Cytoscape v3.7.1. AltProts peptide lists were searched against the human AltProt database HaltORF (reference name "HS_GRCh38_altorf_20170421"), since the conventional UniprotKB does not contain data about AltProts. This database is derived from the predicted H. sapiens alternative proteins (release hg38, Assembly: GCF_000001405.26) which contains 182,709 entries. This is a database for annotated long non-coding RNAs (lncRNAs), non-coding RNAs (ncRNAs), and mRNA uncoding regions. For unbiased analysis, the HaltORF database was used in combination with UniprotKB database which contains the RefProts for a total of a bit more than 203012 entries. Additional online databases such as "Ensembl" (https://www.ensembl.org) and "ref Seq" (https://www.ncbi.nlm.nih.gov/refseq) were also used to trace back the origin of the identified AltProts after HaltORF data interrogation.

Cross-link data-analysis

Data were analyzed using Proteome Discoverer 2.2 (PD2.2) implemented with the XLinkX node (48). Interrogation of data was performed accordingly to the following workflow: first spectra were selected and DSSO was defined as cross-linker (characteristic mass 158.003765 Da). Then the workflow was divided into two paths. The first was dedicated to the cross-link identifications using the XLinkX with the following search parameters: Precursor Mass Tolerance: 10 ppm, FTMS fragment: 20 ppm, ITMS Fragment: 0.5 Da, and searching a compiled database comprising both HaltORF and UniProtKB. The validation was performed using percolator with an FDR set to 0.01. The second path was dedicated to the shot-gun protein identification using SequestHT and considering the following parameters: Trypsin as an enzyme, 2 missed cleavages, methionine oxidation as variable modification, DSSO hydrolyzed and carbamidomethylation of cysteines as static modification, Precursor Mass Tolerance: 10 ppm and

Fragment mass tolerance: 0.6 Da. The validation was performed using Percolator with an FDR set to 0.01. A consensus workflow was then applied for the statistical arrangement. A de-isotope and TopX filter were used to determine the m/z-error with a selectivity around 10% FDR. The protein-protein interaction identifiers were displays in Cytoscape 3.7.1, allowing for visualization of the partners and the number of recurrences of the same interaction.

Modeling and prediction of interactions

Structure modeling of AltProts and RefProts were performed with the I-Tasser software (53). For both RefProts and AltProts, the most stable models (C-Score between -5 and +2) were retained. The prediction of PPIs was performed with the ClusPro software (54). The RefProt was identified as a receiver and the AltProt as a ligand. The interaction model was carried out by docking, the ligand on the receiver without crosslink restriction. ClusPro then generates multiple interaction models ranked in the order of stability. The selected models are still part of the Top5 "balanced" models taking into account the best compromise of stability. The selected interactions were then illustrated with Chimera (55) to measure the distance between the atoms observed during XL-MS analysis. The model is split between the ligand and the receptor to form two independent chains, the lysine found to be involved in interactions on PD2.2 are designated in order to identify the distance between the two points of the model.

RESULTS

Protein regulation under cell reprogramming by Forskolin

To assess the proteomic changes of the PKA-induced reprogramming, we performed large scale protein identification with Label-Free Quantification (LFQ) of NCH82 human glioma cells treated with Forskolin for 48H (51) (**Figure 1**). A total of 3363 proteins were identified, among which 41 are exclusive to the non-treated cells and 201 to the Forskolin condition (**Supp. Data1**). Among these 201 proteins, 148 are organelles, and some are known to be involved in cell reprogramming such as TGFβ1 or the Death-inducer Obliterator 1 (DIDO1), Mitogen-activated protein kinase 4 (MAP4K4) and Protein Hook homolog 3 (HOOK3) which are known to be involved in regulating self-renewal of embryonic stem cells (52, 53). Cytoscape combined with the ClueGO application was then used to retrieve the signaling pathways associated with these specific proteins (**Figure 1A**). We identified two main signaling pathways associated with Forskolin treatment; one being related to the mRNA splicing and metabolism and the second to intracellular trafficking including Golgi vesicle transport. Other pathways deal with non-integrin membrane ECM interaction and nucleotide excision repair. A T-Test with a significant threshold ($p<0.05$) was then applied to generate the Heatmap representing the over-and under-expressed proteins. 1797 proteins are significantly over- and under-expressed and distributed between 2 main clusters, one (991 proteins) associated with control cells and the other one (806 proteins) with Forskolin treatment (**Figure 1B**). Over-expressed proteins are involved in mitochondrial tRNA processing, valine metabolic process and the negative regulation of the 5'-3' RNA directed polymerase activity. In contrast, under-expressed proteins are related to nucleotide biosynthesis, rRNA 3'-end processing and tRNA aminoacetylation. PKA activation by Forskolin signals through a complex network of cellular pathways

including nuclear proteins (54). To elucidate molecular events that occur in the nucleus as cells are treated with Forskolin, we isolated nuclei from control and treated cells and analysed them by shot-gun proteomics (**Figure 2**). From the nuclei fraction, 936 proteins were identified with 69 exclusives to Forskolin condition and 11 to the control. Signalling pathways associated to the Forskolin specific proteins are divided in 4 main pathways including translation, axogenesis and neuritogenesis, centrosome and oxidative stress linked to apoptotic pathways. Heatmaps showing hierarchical clustering of differentially expressed proteins revealed a clear separation between the two conditions (**Figure 2B**). This analysis revealed a set of signalling events associate with Forskolin treatment. For instance, we identified proteins involved in the regulation of alternative splicing, the translation and the modulation of the translation by miRNA. On the contrary, networks related to neurotransmitter receptor transport post-synaptic membrane to endosome and clearance of the nuclear envelope membrane were down-regulated after treatment (**Figure 2C**).

To better understand the timing of signaling events associated with Forskolin reprogramming, a time-course study at 16H, 24H and 48H treatment was performed using nuclear proteins of control and treated cells (**Figure 3**). Each treatment dataset point was compared to control cells to identify specific changes in the protein regulation. Notably, specific clusters of up- and down-regulated proteins were identified after analysis (**Figure 3A**) suggesting that significant changes occur at the nuclear level after treatment. Temporal profiles of signaling pathways within each cluster revealed specific molecular events modified in a temporal and not temporal way (**Figure 3B**). More in detail, Go-Terms analysis of identified proteins showed a significant altered enrichment of pathways after 16H of Forskolin treatment (**Figure 3B**). Consistent with the role of PKA in transcriptional and translational regulation, we identified proteins related to translation initiation and regulation, ribosome assembly and nucleic acid metabolism. Moreover, pathways involving protein degradation through ubiquitin tagging, regulation of protein depolymerization and enzymatic activity such as aminopeptidase were also identified. Within 24H after PKA activation, only two pathways were modulated, threonine-type endopeptidase activity and collagen-activated tyrosine kinase receptor signaling pathway, while no specific pathways were enriched after 48H (**Figure 3B**). The other identified pathways did not show a specific temporal regulation (**Supp Figure 2**). Taken together, this analysis revealed a dynamic modulation of signaling pathways that underlies PKA activation.

Time-course PPIs interaction identification

To validate pathways identified *in silico,* we searched for PPIs by XL-MS large scale strategy upon the time course of Forskolin treatment. Nuclear fractions were submitted to XL-MS using the membrane-permeable DSSO cross-linker, and samples processed using the FASP method (**Figure 4**). In this time-course analysis, a total of 20 cross-links were detected after XL-MS. The interaction networks of these proteins enriched from the literature by STRING interrogation is presented in **Figure 5**. Cytoscape and ClueGo application were applied to correlate the resulting identifications to known signaling pathways using Reactome and GO-term databases. Most proteins identified at 16H are on the signaling pathways of ATP synthesis coupled to electron transport. Based on literature prediction, CDK1, ACTA2 and ACTB represent a possible link with proteins identified by XL-MS. At 24H, identified proteins are linked to the

hydrolysis of ATP by Myosin pathway according to Reactome. CALD1 that was identified by XL-MS with an internal cross-link, appears as a node controlling pathways activated at 16H and 24H, respectively. The modulation of these signaling pathways clearly shows the role of PKA activation in the regulation of ATP and its link to Myosin and Caldesmon at 24H. Hydrolysis of ATP by Myosin is known to stimulate elongation of actin filaments. This effect is particularly well-described in neurons in relation to the elongation of the filopodia (55). RPL5, which was identified at 24H and 48H, represents a node-specific to proteins modified after 48H including 2 AltProts. Notably, a precise temporal modulation of AltProts was observed. Indeed, specific AltProts were detected at 48H (AltDHTKD1, AltCRADD, AltLNC00675 and Alt LOC101927348) compared to 16H (AltSPTBN2, AltLATS2) and 24H (AltSIDT1, AltCFLAR). Overall, this analysis provided important insights into the timing and composition of PPI networks modulated after PKA stimulation and revealed that AltProts might have a role in the regulation of these.

PPIs interaction network at 48H Forskolin treatment

Since Forskolin is known to induce a complete reprogramming after 48H treatment (51), a focus was made on this time point. XL-MS analysis was performed from the nuclear fraction of NCH82 glioma cells leading to the identification of a total of 219 cross-links including 138 specifics to the controls and 81 to Forskolin (**Figure 6**). Five networks were identified from the cross-links plus a few outsider proteins that were not assigned to any known network. The main network (**numbered 6 in Figure 6**) was related to cell mobility and cytoskeleton reorganization (**Figure 7, boxes A and D**). The second and third networks (4, 5) were related to tRNA amino acylation for protein translation and response to interleukin-12 (**Figure 7, box C**). tRNA amino acylation for protein translation was found in control cells. Four genes were found to be directly related to this signaling pathway: SARS, NARS, AARS, and IARS. These four proteins are described in the translation of mRNAs into protein, involved in the attachment of amino acids (aa) to transfer RNAs (tRNAs), aminoacyl-tRNA synthetases (aaRS) enzymes are essential for a proper translation. In the amino acid binding reaction in tRNA, there is an ATP consumption by phosphate transfer. This group comprises as well two AltProts interacting with these aaRS, respectively AltSETD1B and AltLINC00624. If AltLINC00624 is derived from a non-coding RNA, AltSETD1B is from the mRNA encoding the SETD1B histone-lysine N-methyltransferase. Based on the observation of the crosslink interaction between these two AltProts and the aaRS (**Supp Figure 1**), one may assume that these 2 AltProts have a central role in this signaling pathway. Once again, AltProts have been described to being involved in regulating the translation of other proteins. Here indirectly by participating in the assembly of aa-tRNA (**Figure 7, box B**). Network 5 shows proteins that switch from the control (LMNA, LMNB2 and P4HB) to the Forskolin condition (LMNB1, STXBP1, ARHCEF28). ARHGEF28 is directly connected to vimentin (VIM) which is involved in the production of microtubules and therefore cell mobility and cytoskeletal formation. STXBP1 is also linked to the regulation of vesicle fusion. The scheme is enriched by CORO1As and EEF2 proteins which are only found with intramolecular crosslinks and thanks to STRING analysis link all the networks together. Indeed, CORO1A links the GO terms of IL-12 regulation, vesicular fusion and cytoskeletal structuring. EEF2 is known to be involved in the regulation of t-RNA, but it is also described in cell mobility and cytoskeletal

regulation. Network 3 is centered on ALB. By contrast, proteins in network 2 were only found with intra-crosslinks by they are known to be connected according to the literature. Among the proteins not known to be involved in specific networks, not surprisingly 4 are AltProts. One of them was found to interact with LIN9 which was shown to be involved in embryonic stem cell reprogramming and exerting an antitumor effect (56, 57).

The main network (network 6) is centred on the formation of microtubules and in particular the formation of actin filaments. This network presents 15 common proteins (yellow) common to controls and Forskolin stimulation. The main interaction node of proteins for this network are TPM4, TPM3, TPM2, TPM1, and CGNL1. These interactions observed by XL-MS enable to identify under both conditions IQGAP1 and RGL1 connected to TPM4 as well as SMIM17, TMEM68 and MAPKAPK2 connected to TPM3. Conditions-specific interactions are also observed with 10 AltProts and 6 RefProts (green) for the control condition and 3 AltProts plus 3 RefProts (red) for the Forskolin condition (**Table 1**).

TPM network reveals hidden AltProt interactions

In Network 6 (**Figure 6**), TPMs are the major protein family for which, PPis were measured by XL-MS. Interestingly, TPM1, TPM2, TPM3 and TPM4 are in direct interaction in both control and Forskolin. However, some of the PPI partners are specific to one of the two conditions. In particular, in the control condition, TPM4 is interacting with AltProts such as AltTRAU1AP, AltMAP2 and AltEPHA5. The other members TPM3 and PLEC also show to interact with AltProts (**Table 2**). Two of these identified AltProts were previously included in transcript databases but were removed (XR_428143.1 and XM_006723305.1). 8 others identified AltProts were issued either from an overlapping between CDS and 3'UTR (CDS-3'UTR) (2 of them), 3'UTR (1), 5'UTR (CDS-5'UTR) (1), +2 on the CDS (3 of them) and an ncRNA (1) (**Figure 8A**). For the 3 AltProts interacting with the TPM4, AltTRNAU1AP is translated from the 3'UTR part of the transcript 201 of the TRNAU1AP gene. AltEPHA5 is coming from an overlap between the CDS domain and the 3'UTR of the transcript EPHA5-204 also coding for an EPHA5 RefProt. Finally, the sequence coding for AltMAP2 can be issued from the transcripts 201,202 and 205 of MAP2 as an overlapping of the CDS and the 3'UTR (**Figure 8B**).

To model the interaction observed by XL-MS between TPM4 and the 3 AltProts, a 3D model was reconstructed. First, the protein structures were predicted in I-Tasser and only the best conformations for each of them, according to the C-score were kept (**Supp Figure 3**). Then ClusPro was used to predict the possible interaction between these partners. Not to bias the prediction, the models were predicted without taking into account the crosslinker length. Docking analysis was realized between TPM4 and the 3 AltProts. The best interaction model was visualized in Chimera (**Figure 9A**) and the distances between the amino acids found to be cross-linked were then measured. These distances were in good agreement with what expected for DSSO cross-linker (30 Å maximum). Indeed, the distance between TPM4 and the 3 AltProts was found to be in the range of 10 to 30 Å. Therefore, the prediction comforts the existence of an interaction between TPM4 and the 3 AltProts i.e. AltMAP2 and AltTRNAU1AP and AltEPHA5 **(Figure 9A)**. The crosslinks obtained transposed on the model to confirm the probability of interaction leads the possibility that AltEPHA5 (crosslink distance 21Å) could interact

with the middle of the TPM4 sequence while AltMAP2 (crosslink distance= 18.7 Å) and AltTRNAU1AP (crosslink distance= 26.3 Å) are located at the TPM4 ends.

DISCUSSION

The initiation of translation from an alternative reading frame (AltORF) different from the RefORF of the same mRNA and encoding different proteins in terms of the primary structure has been demonstrated in viruses and bacteria four decades ago (58). Some examples of the identification of alternative proteins were then described in human cells (59), however the general consensus for such a mechanism seemed anecdotal in eukaryotes. Nevertheless, due to the increase of the large scale studies in proteomic and transcriptomic, we (2, 13, 60) and other teams have (61, 62) established the presence of proteins derived from translation from AltORFs. The main question for this proposed "ghost" proteome is related to the exact functions of such small proteins and, how to analyze and predict the functions of these AltProts a larger scale. Thus, our objective was to attribute the function by correlating the Go-Term to the AltProt in interaction with RefProt. We demonstrated the capacity of the large-scale XL-MS study in a physio-pathological model, the NCH82 human glioma. We treated these cells in time-course study with the Forskolin at 16, 24 and 48 hours in order to reprogram them. In the whole proteins extract, we demonstrated the presence of specific proteins due to Forskolin treatment, which are known to be involved in cell reprogramming especially embryonic stem cells like TGFβ1 or DIDO1, MAP4K4 and HOOK3. Forskolin can impact cancer stem cells (CSC) present in glioma and we confirm it. Moreover, CSCs are well-known to stimulate growth arrest by cell cycle regulators, p53, p21, p27 and phase-specific cyclins, and neural differentiation, in contrast to Forskolin, seems to induce growth arrest and neural differentiation via cAMP/CREB signalling pathway. We confirm this hypothesis in NCH82 model using XL-MS strategy by focusing our study at the nuclei level. We were able to identify specific networks and pathways by combining in silico analyses using STRING, Reactome and GO-Term to PPIs obtained by XL-MS. We highlighted the ATP pathway at 16h post Forskolin treatment in these glioma cells. STRING analysis predicted the presence of CDK1, ACTA2 and ACTB which can be linked to the proteins identified by XL-MS present in ATP synthesis node. In time course a switch of this network from ATP synthesis to the ATP hydrolysis by Myosin was observed. Hydrolysis of ATP by Myosin is a mechanism already described in neurons during the development of cellular protrusions including filopodia (63, 64). Such type of morphological modification under Forskolin treatment in U87 glioma cells has also been reported (27). Forskolin stimulates expression of cAMP-related protein CREB and pCREB as well as apoptosis-related proteins. Thus, this molecule will inhibit the proliferation as well as invasion and promote the apoptosis of U87. This is in agreement with our PPI data that highlighted the prominent role of signaling pathways controlling cytoskeletal organization and remodeling (**Figure 5**). At 48h, we identified 5 networks from the cross-links plus a few outsider proteins, not assigned to any known network. The main network was enriched for proteins related to cell mobility and cytoskeleton reorganization and it is also linked to the cAMP pathway. For example, DIXDC1 protein, found in this network has been shown to be involved in Wnt5 signalling pathway, known to modulate the increase of Ca^{2+} under Forskolin treatment and to stimulate the cAMP pathway. ARID4B is known to interact with MBD2, a regulator of transcription under the influence of cAMP signalling

pathway. The GPR82 orphan receptor seems also to be under the regulation of factors involved in the cAMP pathway. The other identified networks are related to tRNA amino acylation for protein translation and response to interleukin-12. All these 6 networks are linked together through STXBP1, CORO1As and EEF2. CORO1A links the GO terms of IL-12 regulation, vesicular fusion and cytoskeletal structuring. EEF2 is known to be involved in the regulation of t-RNA but it is also involved in cell mobility and cytoskeletal regulation. Nevertheless, the main network is centered on the formation of microtubules and the formation of actin filaments which is related to tumor phenotype switching in neuronal profile as represented by the axon guidance proteins such like protein canopy homolog 2 (CNPY2), debrin, plectin, and synaptopodin.

Among these networks, we identified several ghost proteins (81) and we confirmed their involvement in all the 6 networks. RefProts and AltProts are contained in the same networks, in same pathways and we expect that they would have the same function. For example, AltProts AltSMIM13, AltLOC101927356, and AltCPXM2, were specific to Forskolin stimulation and potentially representing important mediators of its mechanism of action. In detail, AltSMIM13 and AltLOC101927356 were directly connected to TPM4, while AltCPXM2 showed a specific interaction with PLEC that is also connected to cellular mobility. Other identified AltProts were shown to be involved in protein synthesis regulation. For example, AltSETD1B and AltLINC00624 are connecting to the tRNA aminoacylation for protein translation. This pathway was correlated with protein-based networks identified by total proteome analysis (**Figure 1**). This result may reflect a functional correlation between RefProts and AltProts in modulating PKA signaling in glioblastoma cells. Based on the AltATAD2 model, AltProts AltSETD1B and AltLINC00624 may play a role in the regulation of protein expression, corroborating the hypothesis that multiple layers of regulation are involved in the regulation of cellular reprogramming (65). In this context, the role of AltProts as possible drivers of this complexity is still unexplored thus offering novel opportunities to identify therapeutic targets and biomarkers. Even today, the description of the interactions of AltProts and the assignment of a GO-Term allowed us to highlight the involvement of these AltProts in the regulation of tRNA and protein synthesis.

The identification by XL-MS of different interaction networks observed allowed to connect AltProts to a specific GO-term. An example of this is what we demonstrated for the signaling pathway of cytoskeletal and intracellular transport, which was confirmed by 3D modeling. TPM4 association between 2 AltProts highlighted the role of these AltProts in the cellular response to external stress. Interestingly, prediction of the binding sites of the TPMs inhibitor, TR100 (2-Cyano-3-[1-[3-(dimethylamino)propyl]-2-methyl-1H-indol-3-yl]-N-octyl-2-propenamide) performed by docking using Chimera predicts that TR100 could interact at the same location than the 2 AltProts AltTRNAU1AP and AltMAP2. TR100 could also bind close by the position of AltEPHA5 (**Figure 9B**). In this context, no cross-link with these 3 AltProts is observed with TR100 which reinforced the hypothesis of their interaction with TPM4. The observations obtained by XL-MS show consistency with enriched signaling pathways identified after total proteomics.

Moreover, in this study, two major pathways were identified by the XL-MS methodology: (1) the protein translation, by the interaction with the aaRS proteins involved in the protein's biosynthesis and (2) the

cell mobility, which is significantly modulated after Forskolin stimulation. These two pathways are in agreement with the previous prediction of the AltProt function (66). Based on this prediction, the first biological process predicted for the AltProt was the Biosynthesis, the second the nucleic acid metabolism, these two predictions are in line with the involvement of AltSETD1B and AltLINC00624 and the protein synthesis by interaction with aaRS proteins. Thus, we confirmed that the predicted function of AltProt could be retrieved by a combination of XL-MS and AltProt database.

Taken together, by this study, we permit to place AltProts in the proteomic landscape of the cell, through a non-targeted strategy using a large-scale analysis approach. Moreover, by using an integrated analysis of total proteomics and XL-MS, we demonstrated that AltProts have functional importance specific to PKA activation by Forskolin. Overall, these results provided the rationale for investigating the role of ghost proteomes in the regulation of mechanisms of cell reprogramming. Moreover, this analysis allowed also to highlight the response of NCH82 glioma cells to Forskolin stimulation thus providing a comprehensive molecular description of its effects at the proteome level. We showed a significant effect of the quantitative variation of cellular proteins after PKA stimulation, and the molecular switch of glioma cells into a neuron phenotype after stimulation. Moreover; the identification of signaling pathways regulating the production of proteins associated to tRNA, rRNA and gene regulation showed a deep change in the protein landscape of these cells after stimulation. We also established that AltProts as RefProt, when they are in the same networks, would have the same functions, but we cannot determine if they are involved as positive or negative regulators. Further deeper biological studies will be necessary to respond to this question, but we can now envisage in which global functions involved members of the “Ghost proteome” in this context of cell are reprogramming after Forskolin treatment.

AVAILABILITY

Proteomic datasets including MaxQuant files and annotated MS/MS datasets were uploaded to the ProteomeXchange Consortium via the PRIDE database, and then assigned the dataset identifier PXD014642

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

Authors contribution. Conceptualization, I.F., J.F. and M.S.; Methodology, I.F., J.F., T.C. and M.S.; Software, T.C.; Validation, I.F., J.F., T.C. and M.S.; Formal Analysis, T.C.; Investigation, I.F., J.F., T.C., and M.S.; Resources, I.F. M.M. and M.S.; Data curation, T.C.; Writing - Original Draft T.C. and M.S.

Writing - Review & Editing, I.F. and M.S.; Supervision, I.F., J.F. and M.S.; Project Administration, I.F. and M.S.; Funding Acquisition, I.F., M.M. and M.S.

FUNDING

This research was supported by funding from Ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation (MESRI), Institut National de la Santé et de la Recherche Médicale (Inserm) and University of Lille

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.
2. Vanderperre,B., Lucier,J.-F., Bissonnette,C., Motard,J., Tremblay,G., Vanderperre,S., Wisztorski,M., Salzet,M., Boisvert,F.-M., Roucou,X., *et al.* (2013) Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS One*, **8**, e70698.
3. Mouilleron,H., Delcourt,V. and Roucou,X. (2016) Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.*, **44**, 14–23.
4. Aspden,J.L., Eyre-Walker,Y.C., Phillips,R.J., Amin,U., Mumtaz,M.A.S., Brocard,M. and Couso,J.-P. (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife*, **3**.
5. Ji,Z., Song,R., Regev,A. and Struhl,K. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, **4**.
6. 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. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
7. Bateman,A., Martin,M.J., O'Donovan,C., Magrane,M., Alpi,E., Antunes,R., Bely,B., Bingley,M., Bonilla,C., Britto,R., *et al.* (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
8. 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.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
9. Brunet,M.A., Brunelle,M., Lucier,J.-F., Delcourt,V., Levesque,M., Grenier,F., Samandi,S., Leblanc,S., Aguilar,J.-D., Dufour,P., *et al.* (2018) OpenProt: a more comprehensive guide to explore

eukaryotic coding potential and proteomes. *Nucleic Acids Res.*, 10.1093/nar/gky936.

10. Delcourt,V., Staskevicius,A., Salzet,M., Fournier,I. and Roucou,X. (2018) Small Proteins Encoded by Unannotated ORFs are Rising Stars of the Proteome, Confirming Shortcomings in Genome Annotations and Current Vision of an mRNA. *Proteomics*, **18**, e1700058.
11. Hashimoto,Y., Niikura,T., Tajima,H., Yasukawa,T., Sudo,H., Ito,Y., Kita,Y., Kawasumi,M., Kouyama,K., Doyu,M., *et al.* (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 6336–41.
12. Delcourt,V., Franck,J., Leblanc,E., Narducci,F., Robin,Y.-M., Gimeno,J.-P., Quanico,J., Wisztorski,M., Kobeissy,F., Jacques,J.-F., *et al.* (2017) Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine*, **21**, 55–64.
13. Delcourt,V., Franck,J., Quanico,J., Gimeno,J.-P., Wisztorski,M., Raffo-Romero,A., Kobeissy,F., Roucou,X., Salzet,M. and Fournier,I. (2018) Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions. *Mol. Cell. Proteomics*, **17**, 357–372.
14. Razooky,B., Obermayer,B., O'May,J. and Tarakhovsky,A. (2017) Viral Infection Identifies Micropeptides Differentially Regulated in smORF-Containing lncRNAs. *Genes (Basel).*, **8**, 206.
15. Fuku,N., Pareja-Galeano,H., Zempo,H., Alis,R., Arai,Y., Lucia,A. and Hirose,N. (2015) The mitochondrial-derived peptide MOTS-c: a player in exceptional longevity? *Aging Cell*, **14**, 921–3.
16. Couso,J.-P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Publ. Gr.*, **18**.
17. Bensimon,A., Heck,A.J.R. and Aebersold,R. (2012) Mass Spectrometry–Based Proteomics and Network Biology. *Annu. Rev. Biochem*, **81**, 379–405.
18. Dunham,W.H., Mullin,M. and Gingras,A.C. (2012) Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics*, **12**, 1576–1590.
19. Gavin,A.-C., Bösche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.-M., Cruciat,C.-M., *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
20. Maeda,K., Poletto,M., Chiapparino,A. and Gavin,A.C. (2014) A generic protocol for the purification and characterization of water-soluble complexes of affinity-tagged proteins and lipids. *Nat. Protoc.*, **9**, 2256–2266.
21. Li,P., Li,J., Wang,L. and Di,L.J. (2017) Proximity Labeling of Interacting Proteins: Application of BioID as a Discovery Tool. *Proteomics*, **17**.
22. Lam,S.S., Martell,J.D., Kamer,K.J., Deerinck,T.J., Ellisman,M.H., Mootha,V.K. and Ting,A.Y. (2015) Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nat. Methods*, **12**, 51–54.
23. Roux,K.J., Kim,D.I., Raida,M. and Burke,B. (2012) A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.*, **196**, 801–10.
24. Li,P., Meng,Y., Wang,L. and Di,L.J. (2019) BioID: A proximity-dependent labeling approach in

proteomics study. In *Methods in Molecular Biology*.Vol. 1871, pp. 143–151.

25. Roux,K.J., Kim,D.I., Burke,B. and May,D.G. (2018) BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. protein Sci.*, **91**.
26. Eyckerman,S., Titeca,K., Van Quickelberghe,E., Cloots,E., Verhee,A., Samyn,N., De Ceuninck,L., Timmerman,E., De Sutter,D., Lievens,S., *et al.* (2016) Trapping mammalian protein complexes in viral particles. *Nat. Commun.*, **7**, 11416.
27. Yu,C. and Huang,L. (2017) Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.*, 10.1021/acs.analchem.7b04431.
28. Chavez,J.D. and Bruce,J.E. (2019) Chemical cross-linking with mass spectrometry: a tool for systems structural biology. *Curr. Opin. Chem. Biol.*, **48**, 8–18.
29. Kao,A., Chiu,C., Vellucci,D., Yang,Y., Patel,V.R., Guan,S., Randall,A., Baldi,P., Rychnovsky,S.D. and Huang,L. (2011) Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics*, **10**, M110.002212.
30. Müller,M.Q., Dreiocker,F., Ihling,C.H., Schäfer,M. and Sinz,A. (2010) Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS. *Anal. Chem.*, **82**, 6958–6968.
31. Fritzsche,R., Ihling,C.H., Götze,M. and Sinz,A. (2012) Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. *Rapid Commun. Mass Spectrom.*, **26**, 653–8.
32. Rey,M., Dupré,M., Lopez-neira,I., Duchateau,M., Chamot-rooke,J., Rey,M., Dupré,M., Lopez-neira,I., Duchateau,M. and Chamot-rooke,J. (2018) eXL-MS : An enhanced Cross-Linking Mass Spectrometry Workflow to Study Protein Complexes eXL-MS : An enhanced Cross-Linking Mass Spectrometry Workflow to Study Protein Complexes . 10.1021/acs.analchem.8b00737.
33. Nury,C., Redeker,V., Dautrey,S., Romieu,A., van der Rest,G., Renard,P.-Y., Melki,R. and Chamot-Rooke,J. (2015) A Novel Bio-Orthogonal Cross-Linker for Improved Protein/Protein Interaction Analysis. *Anal. Chem.*, **87**, 1853–1860.
34. Burke,A.M., Kandur,W., Novitsky,E.J., Kaake,R.M., Yu,C., Kao,A., Vellucci,D., Huang,L. and Rychnovsky,S.D. (2015) Synthesis of two new enrichable and MS-cleavable cross-linkers to define protein–protein interactions by mass spectrometry. *Org. Biomol. Chem.*, **13**.
35. Riffle,M., Jaschob,D., Zelter,A. and Davis,T.N. (2016) ProXL (protein cross-linking database): A platform for analysis, visualization, and sharing of protein cross-linking mass spectrometry data. *J. Proteome Res.*, 10.1021/acs.jproteome.6b00274.
36. Götze,M., Pettelkau,J., Schaks,S., Bosse,K., Ihling,C.H., Krauth,F., Fritzsche,R., Kühn,U. and Sinz,A. (2012) StavroX—A Software for Analyzing Crosslinked Products in Protein Interaction Studies. *J. Am. Soc. Mass Spectrom.*, **23**, 76–87.
37. Müller,F., Fischer,L., Chen,Z.A., Auchynnikava,T. and Rappsilber,J. (2018) On the Reproducibility of Label-Free Quantitative Cross-Linking/Mass Spectrometry. *J. Am. Soc. Mass Spectrom.*, **29**, 405–412.
38. Leitner,A., Walzthoeni,T. and Aebersold,R. (2013) Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet

software pipeline. *Nat. Protoc.*, **9**, 120–137.

39. Walzthoeni,T., Joachimiak,L.A., Rosenberger,G., Röst,H.L., Malmström,L., Leitner,A., Frydman,J. and Aebersold,R. (2015) XTract: Software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nat. Methods*, **12**, 1185–1190.
40. Combe,C.W., Fischer,L. and Rappsilber,J. (2015) xiNET: Cross-link Network Maps With Residue Resolution. *Mol. Cell. Proteomics*, **14**, 1137–1147.
41. Iacobucci,C. and Sinz,A. (2017) To Be or Not to Be? Five Guidelines to Avoid Misassignments in Cross-Linking/Mass Spectrometry. *Anal. Chem.*, **89**, 7832–7835.
42. Du,X., Chowdhury,S.M., Manes,N.P., Wu,S., Mayer,M.U., Adkins,J.N., Anderson,G. a and Smith,R.D. (2011) Xlink-Identifier: An automated data analysis platform for confident identifications of chemically cross-linked peptides using tandem mass spectrometry. *J. Proteome Res.*, **10**, 923–931.
43. Leitner,A., Faini,M., Stengel,F. and Aebersold,R. (2016) Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. 10.1016/j.tibs.2015.10.008.
44. Klykov,O., Steigenberger,B., Pektaş,S., Fasci,D., Heck,A.J.R. and Scheltema,R.A. Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.*, 10.1038/s41596-018-0074-x.
45. Sinz,A. (2014) The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. *Expert Rev. Proteomics*, **11**, 733–743.
46. Chen,Z.A. and Rappsilber,J. (2019) Quantitative cross-linking/mass spectrometry to elucidate structural changes in proteins and their complexes. *Nat. Protoc.*, **14**, 171–201.
47. Kaake,R.M., Wang,X., Burke,A., Yu,C., Kandur,W., Yang,Y., Novtisky,E.J., Second,T., Duan,J., Kao,A., *et al.* (2014) A new in vivo cross-linking mass spectrometry platform to define protein-protein interactions in living cells. *Mol. Cell. Proteomics*, **13**, 3533–43.
48. Liu,F., Rijkers,D.T.S., Post,H. and Heck,A.J.R. (2015) Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods*, **12**.
49. Cardon,T., Salzet,M., Franck,J. and Fournier,I. (2019) Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. *Biochim. Biophys. Acta - Gen. Subj.*, 10.1016/J.BBAGEN.2019.05.009.
50. Pattabiraman,D.R., Bierie,B., Kober,K.I., Thiru,P., Krall,J.A., Zill,C., Reinhardt,F., Tam,W.L. and Weinberg,R.A. (2016) Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. *Science*, **351**, aad3680.
51. Xing,F., Luan,Y., Cai,J., Wu,S., Mai,J., Gu,J., Zhang,H., Li,K., Lin,Y., Xiao,X., *et al.* (2017) The Anti-Warburg Effect Elicited by the cAMP-PGC1α Pathway Drives Differentiation of Glioblastoma Cells into Astrocytes. *Cell Rep.*, **18**, 468–481.
52. Liu,Y., Kim,H., Liang,J., Lu,W., Ouyang,B., Liu,D. and Songyang,Z. (2014) The Death-inducer Obliterator 1 (Dido1) Gene Regulates Embryonic Stem Cell Self-renewal. *J. Biol. Chem.*, **289**, 4778–4786.

53. Saito,S., Lin,Y.-C., Nakamura,Y., Eckner,R., Wuputra,K., Kuo,K.-K., Lin,C.-S. and Yokoyama,K.K. (2019) Potential application of cell reprogramming techniques for cancer research. *Cell. Mol. Life Sci.*, **76**, 45–65.
54. Shabb*,J.B. (2001) Physiological Substrates of cAMP-Dependent Protein Kinase. 10.1021/CR000236L.
55. Mitchison,T. and Kirschner,M. (1988) Cytoskeletal dynamics and nerve growth. *Neuron*, **1**, 761–772.
56. Boheler,K.R. (2009) Stem cell pluripotency: a cellular trait that depends on transcription factors, chromatin state and a checkpoint deficient cell cycle. *J. Cell. Physiol.*, **221**, 10–7.
57. Gagrica,S., Hauser,S., Kolfschoten,I., Osterloh,L., Agami,R. and Gaubatz,S. (2004) Inhibition of oncogenic transformation by mammalian Lin-9, a pRB-associated protein. *EMBO J.*, **23**, 4627–4638.
58. Normark,S., Bergstrom,S., Edlund,T., Grundstrom,T., Jaurin,B., Lindberg,F.P. and Olsson,O. (1983) Overlapping Genes. *Annu. Rev. Genet.*, **17**, 499–525.
59. Rong-Fu Wang,B., Kawakami,Y., Robbins,P.F. and Rosenberg,S.A. Utilization of an Alternative Open Reading Frame of a Normal Gene in Generating a Novel Human Cancer Antigen.
60. Delcourt,V., Franck,J., Leblanc,E., Narducci,F., Robin,Y.-M., Gimeno,J.-P., Quanico,J., Wisztorski,M., Kobeissy,F., Jacques,J.-F., *et al.* (2017) Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine*, **21**, 55–64.
61. Menschaert,G., Van Criekinge,W., Notelaers,T., Koch,A., Crappé,J., Gevaert,K. and 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. *Mol. Cell. Proteomics*, **12**, 1780–90.
62. 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. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.*, **9**, 59–64.
63. Wang,F.S., Wolenski,J.S., Cheney,R.E., Mooseker,M.S. and Jay,D.G. (1996) Function of myosin-V in filopodial extension of neuronal growth cones. *Science*, **273**, 660–3.
64. Berg,J.S. and Cheney,R.E. (2002) Myosin-X is an unconventional myosin that undergoes intrafilopodial motility. *Nat. Cell Biol.*, **4**, 246–250.
65. Simeone,P., Trerotola,M., Franck,J., Cardon,T., Marchisio,M., Fournier,I., Salzet,M., Maffia,M. and Vergara,D. (2018) The multiverse nature of epithelial to mesenchymal transition. *Semin. Cancer Biol.*, 10.1016/J.SEMCANCER.2018.11.004.
66. 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., *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *Elife*, **6**, e27860.

TABLE AND FIGURES LEGENDS

Tables

Table 1. List of the AltProts identified to be in direct interaction with the TPM RefProt family members after 48H Forskolin treatment of the HCH82 human glioma cell. The nucleus fraction of the treated cells is collected and submitted to XL-MS after protein extraction using DSSO as cross-linker. The cross-links are identified from the shot-gun proteomics analysis using XlinX in addition to Proteome Discover 2.2.

Table 2. Sequence, gene entry and name, transcript number and origin of the AltProts identified to be in direct interaction with the TPM RefProt family members after 48H Forskolin treatment of the HCH82 human glioma cell.

Figures

Figure 1. Identified proteins and their related signalling pathways from NCH82 glioma total cells upon or not Forskolin treatment. **(A)** GO-Terms and signalling pathways associated to the proteins identified as unique to the Forskolin treated cells. **(B)** Heatmap representation generated after raw nLC-MS/MS data interrogated by MaxQuant with LFQ and further processed in Perseus using a T-Test showing the proteins over-and under-expressed upon Forskolin treatment. **(C)** GO-Terms and signalling pathways associated to the over-and under-expressed proteins obtained by Cytoscape with the application ClueGo. In cytoscape, the proteins overexpressed under Forskolin stimulation are represented in red and the under-expressed ones in green.

Figure 2. Identified proteins and their related signaling pathways from NCH82 glioma cell nuclei fraction upon or not Forskolin treatment. **(A)** GO-Terms and signaling pathways associated to the proteins identified as unique to the Forskolin treated cells. **(B)** Heatmap representation generated after raw nLC-MS/MS data interrogated by MaxQuant with LFQ and further processed in Perseus using a T-Test showing the proteins over-and under-expressed upon Forskolin treatment. (C) GO-Terms and signaling pathways associated to the over-and under-expressed proteins obtained by Cytoscape with the application ClueGo. In cytoscape, the proteins overexpressed under Forskolin stimulation are represented in red and the under-expressed ones in green.

Figure 3. Identified over-and under-expressed proteins and their related signaling pathways from NCH82 glioma cells nucleus fraction for a time course treatment by Forskolin (16H, 24H and 48H) compared to the control condition. **(A)** Heatmaps of the LFQ variation for 16H (A.1), 24H (A.2) and 48H (A.3) treatment with Forskolin, each compared to the control condition, generated after raw nLC-MS/MS data interrogated by MaxQuant with LFQ and further processed in Perseus using a T-Test showing the proteins over-and under-expressed **(B)** GO-Terms and signaling pathways associated to the proteins identified upon Forskolin treatment time course. The control condition presents the under-expression of the regulation of mRNA processing which was observed for stimulation times. At 16H stimulation, an over-expression of proteins in specific pathways such as the initiation of the translation and the

nucleotide metabolic process appears. At 24H, 2 specific pathways are found related to the collagen formation and the cytoskeletal expression. At 48H, no specific pathways are observed. All the other pathways are common to all Forskolin time of treatment (so-called time-independent).

Figure 4. Schematic representation of the XL-MS workflow used in the study. After removing cell cytoplasm by cell lysis using appropriated buffers and centrifugation, the cell nuclei proteins are extracted and cross-linked. Cross-linked proteins are then digested and peptides are analyzed by LC-MS/MS. Recorded data are processed to get the LFQ variations of the cross-linked proteins using either Uniprot or a combination of Uniprot and AltProt databases. Networks of proteins and their associated pathways are then built in Cytoscape and associated to GO-terms with ClueGo. The measured interactions are as well enrich from the bibliographic data by STRING.

Figure 5. Network of proteins and their associated Go-Term generated from the XL-MS data and enriched from String and known pathways for the NCH82 cell nucleus fraction upon time course treatment by Forskolin for 16H (purple), 24H (red) and 48H (green). The proteins are given different colors depending the time point at which they were observed. Data were interrogated using a combined databased between RefProts and AltProts. Circles correspond to the identified RefProts and squares to the AltProts. The networks were enriched by addition of string network to the identified RefProts using ClueGO application on Cytoscape.

Figure 6. Network of proteins generated using cytoscape and their associated Go-Term generated from the XL-MS data and enriched from String and known pathways for the NCH82 cell nucleus fraction for 48H treatment by Forskolin by comparison to the control condition (non treated cells) . Proteins were identified by interrogating a combined database between RefProts and AlProts. The networks were enriched by addition of String network to the identified RefProts using ClueGO application on Cytoscape. The Green circles correspond to the proteins only found in the control, the red circles the proteins specific to 48H Forskolin treatment and the yellow the proteins both found in the control and the 48H incubation with Forskolin. The dark grey lines represent the measured inter molecular cross-links and the small black circles the intra-molecular ones. The pink lines correspond to the known connections obtained by String. The global network was subdivided into 6 groups. In certain of these groups various AltProts are identified. The attribution of GO-Terms to the RefProts enable to connect certain networks involving AltProts to RefProts providing a first information on the function of these AltProts.

Figure 7. Schematic representation of the pathway over-regulated upon Forskolin stimulation for 48H. The AltProts have been located in this pathway. The Forskolin stimulates the cAMP pathway. cAMP pathway is implicating different factors such as ARIB4B, DIXDC1 and GPR82 as previously described. Is was found that the cell mobility pathway seems to tightly connected to the cAMP one. The tRNA modulation and the regulation of proteins synthesis are also downstream pathways of cAMP. AltProts are implication at different level, both in the cAMP pathway and the tRNA modulation and protein synthesis with the involvement of AltProt AltSETD1B and AltLINC00624. The RefProt SETD18

associated to AltSETD1B was described to be involved in the epigenetic control of chromatin structure and gene expression.

Figure 8. (A) Focus on the TPMs protein network and the different PPI interactions found by the XL-MS experiment from the NCH82 glioma cell nucleus fraction upon 48H stimulation with Forskolin by comparison to the control condition (non treated cells). The green color corresponds to the proteins only found in the control condition, the red to the proteins over-expressed for the 48H treatment with Forskolin and the yellow represents the proteins observed in both conditions. The circles are the RefProts and the square the AlProts. Map of the interactions obtained after crosslink experiments performed in triplicates (set of files N0, N1, N2, N3) were grouped together in order to get the significant interactions. **(B)** Representation of the location of the coding regions on the mRNA for the AltProts in interaction with TPM4. AltTRNAU1AP is issued from the 3'UTR of the transcript 201 coding of TRNAU1AP. AltEPHA5 is issued from the overlapping between the CDS and the 3'UTR region of the 204 transcript coding for EPHA5. AltMAP2 is issued from overlapping between the CDS and the 3'UTR for the 3 transcripts 201, 202 and 203 coding for MAP2.

Figure 9. (A) ClusPro docking representation of the interaction between TPM4 and the 3 AltProts that were measured by XL-MS to be TMP4 interaction partners. The folded model is the one that best allows the AltProts to be positioned. It is observed that AltEPHA5 interacts with the middle of the TPM4 sequence while AltMAP2 and AltTRNAU1AP are located at the TPM4 ends. The interaction models are generated without a priori on the crosslinks found. Then the crosslinks obtained are transposed on the model to confirm the probability of interaction. AltEPHA5 crosslink distance= 21Å, AltTRNAU1AP crosslink distance= 26.3 Å, AltMAP2 crosslink distance= 18.7 Å. **(B)** Prediction of the binding sites of the TPMs inhibitor. Performed docking using Chimera predicts that TR100 could interact at the same location than the 2 AltProts AltTRNAU1AP and AltMAP2. TR100 could also bind close by the position of AltEPHA5.

Table 1

Control		Forskolin	
AltMAP2	LEO1	AltLOC101927356	ARID4B
AltEPHA5	PLEKHH2	AltSMIM13	DIXDC1
AltTRNAU1AP	HP	AltCPXM2	GPR82
AltPHF6	CBX2		
AltGTF2IP13	DHCR24		
AltPLEKHA2	GADD45GIP1		
AltPSPC1			
IP_274074.1			
AltLELP1			
AltITGB2-AS1			

Table 2

AltProt	sequence	gene	gene name	transcrit number	origine
TPM4					
IP_061498.1	MGFIYFHELQHLHIQGLDHFRINILTREK AAQKGLRKIAQSQSRPSV	54952	TRNAU1AP	NM_017846.4	3'UTR
IP_118417.1	MARGNQDGPVYRDFHGKWIQFNGRC GSGDLGVSNFSDNFYIIAGARKIVQKPI WMR	2044	EPHA5	NM_001281767.1	CDS-3'UTR
IP_094793.1	MLKPVWTMGLRSLHSPQADPAWHHPD DSAMSPRLEASTCSNLLSLPLWLRMSL LHSLSRACEYFSFSIEIIIFRHELLAGVGS EQLLYSFFINHKINNLIPKL	4133	MAP2	NM_001039538.1	CDS-3'UTR
TPM3					
IP_218514.1	MHLDGRLLMKWKSSSVSRLIETSEKPK RNWRQKWKQLGMNTN	55269	PSPC1	NM_001042414.2	CDS+2
IP_159909.1	MLVKNRKLCLAWPRREDPSGQEQPSN MAQPHLCSPTSDPQLLEPKGSRLQ	101930567	removed	XR_428143.1	/
IP_274074.1	MGIHVVFIKTLMYTRKFIMRSSINVRNTE GPLKELEKLLHFKEFMMVRNTLNAHSV GNPLECMHNLLDIRKSILMRKLTNVWNV ARTSDFIHSLPNIREFILVRNPTNVCTVR RFLELVHSSLNIREFTLVRNLMHVRNVG RLLEYVENLLVIREFILENTVDGFNR	57711	removed	XM_006723305.1	/
IP_163557.1	MRNVINPEKRAEGRERRRGHWGRPER GEREARRVWQADPEIPGARRTRRPEG RPRPM	59339	PLEKHA2	NM_021623.1	5'UTR (NCBI)
IP_303632.1	MIKLKYERNLHKEFTWSIAENTRKLHITP KQLI	84295	PHF6	NM_032335.3	CDS+2
PLEC					
IP_070624.1	MIKVNQMTPRLSPRTAIPSVNKSVSPNA SPAV	149018	LELP1	NM_001010857.2	CDS+2
IP_291098.1	MERRNQPRRIKRWEQARKPGLSLENN EEGQKENTILLHPLMT	100505746	ITGB2-AS1	NR_038311.1	ncRNA

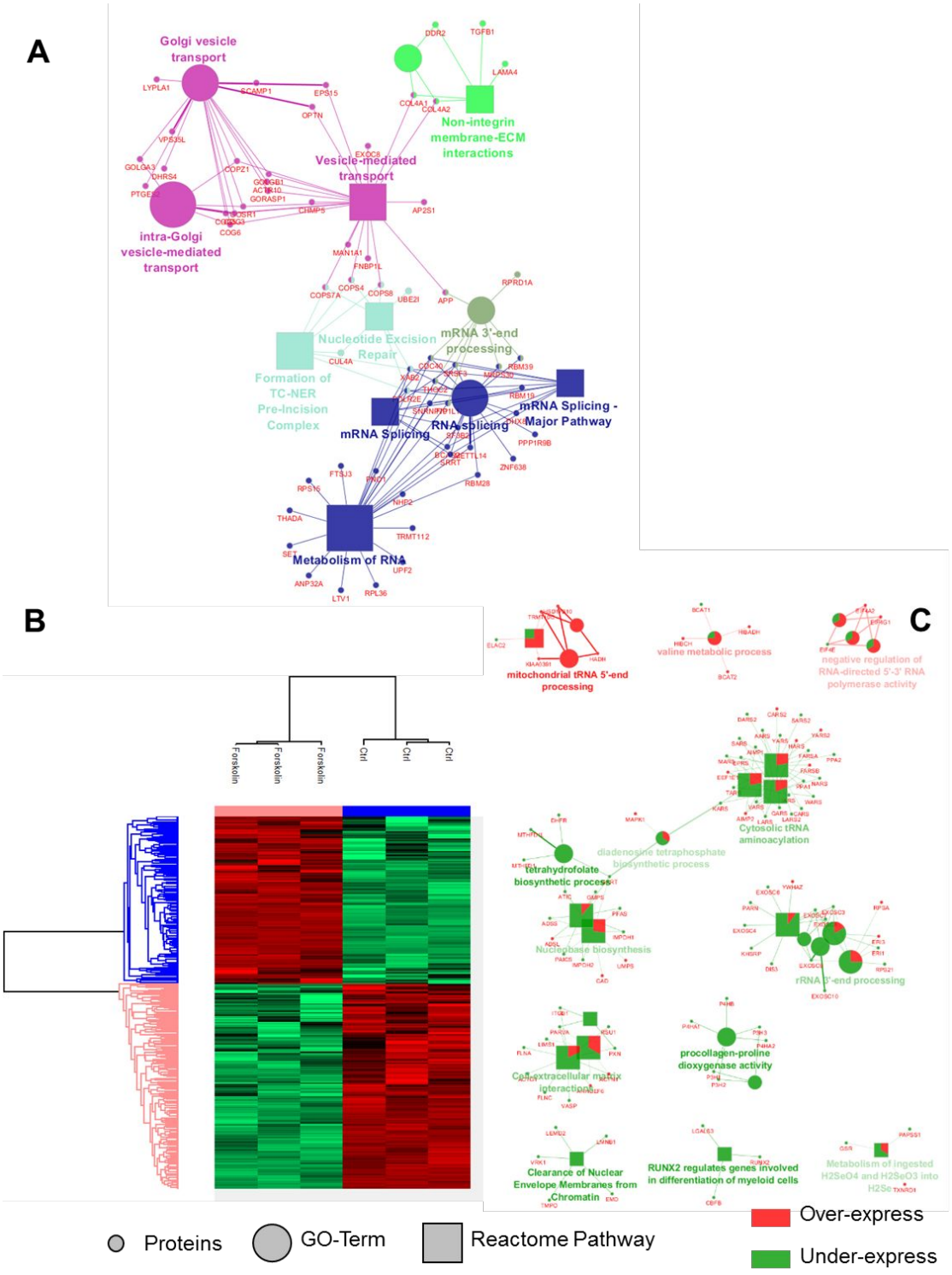


Figure 1

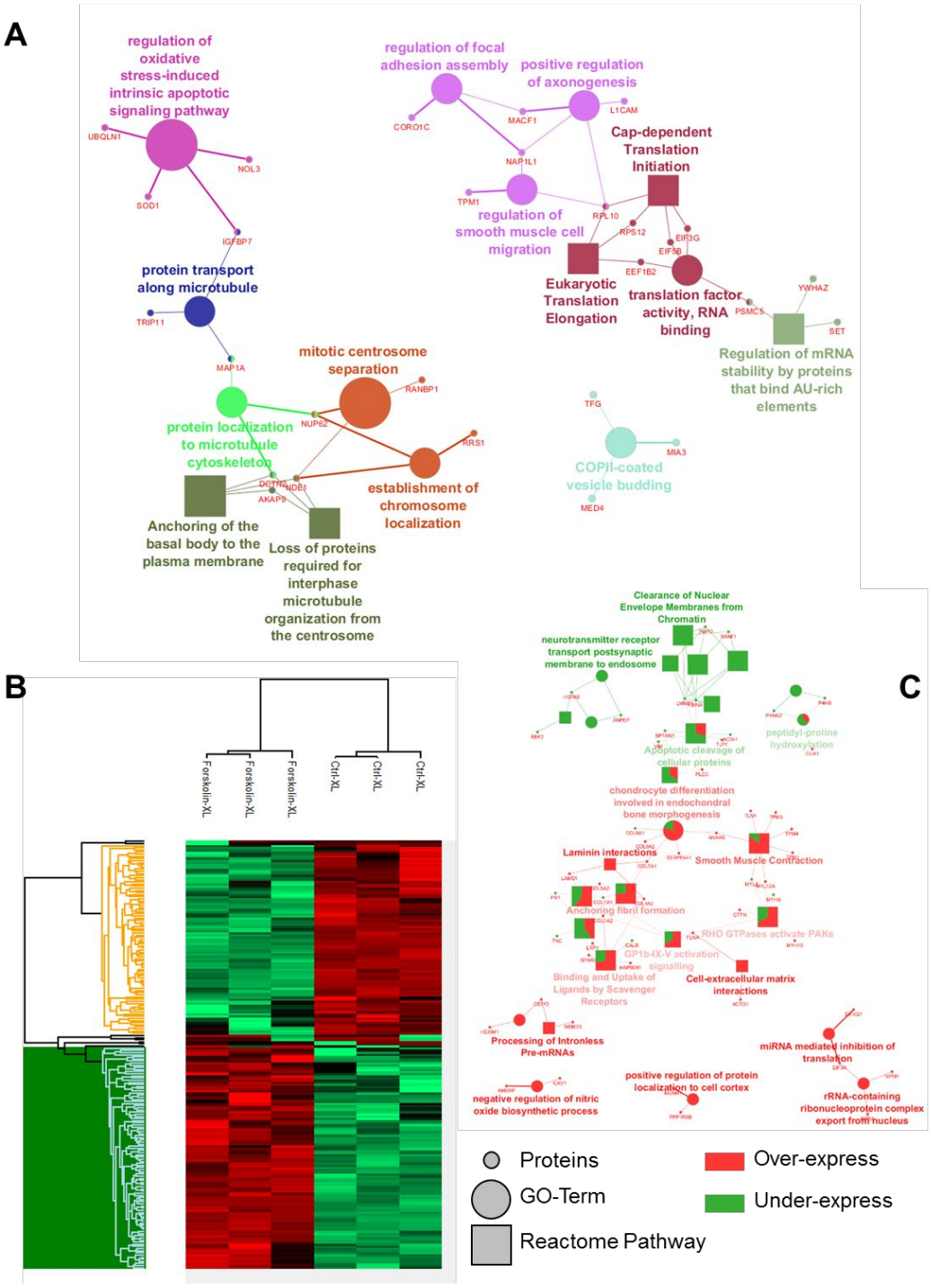


Figure 2

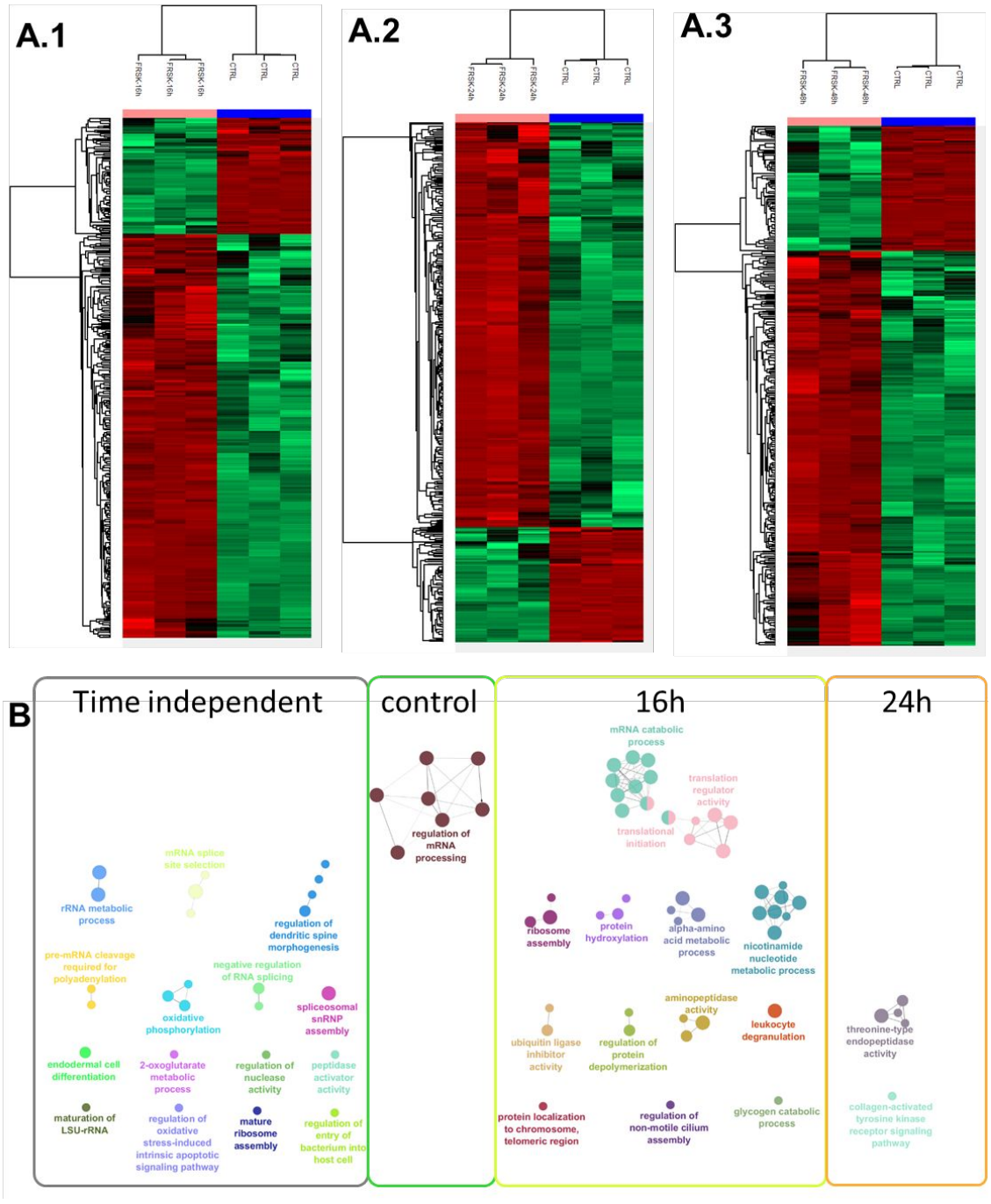


Figure 3

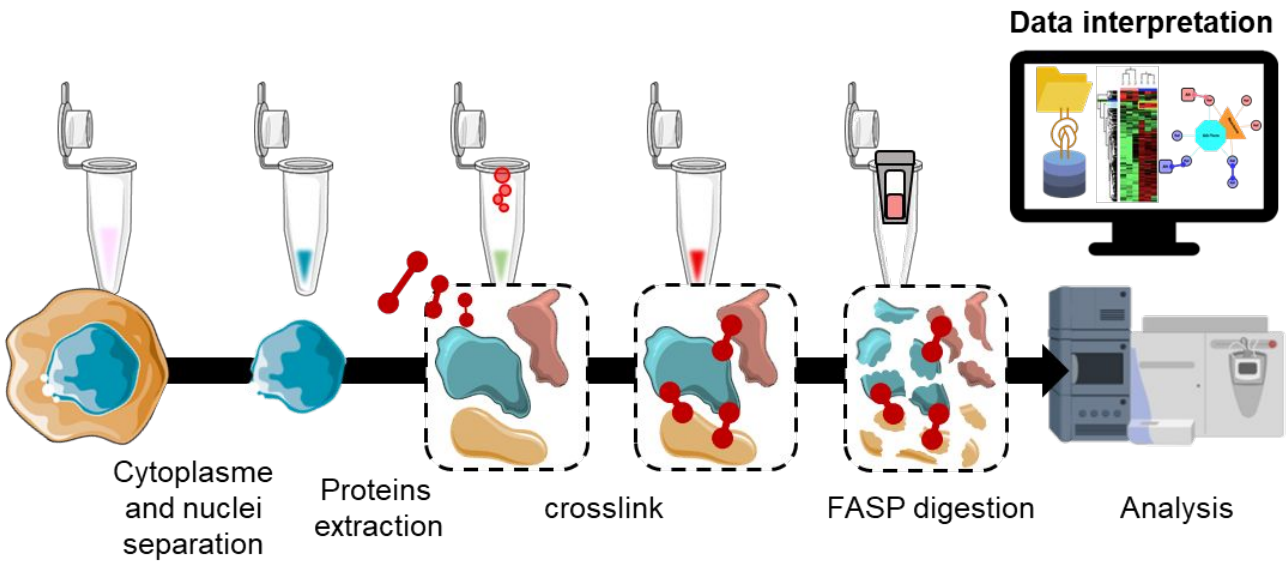


Figure 4

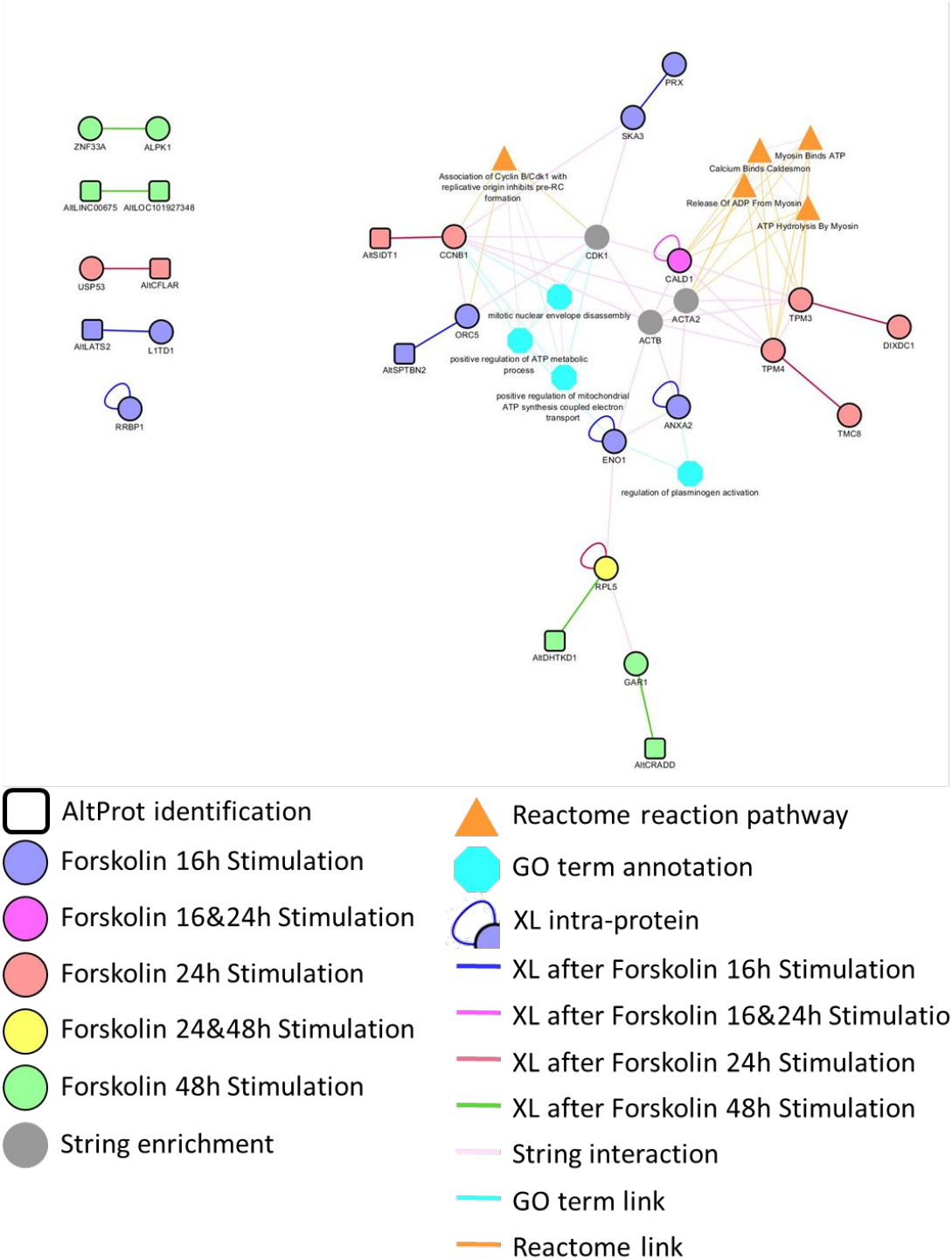


Figure 5

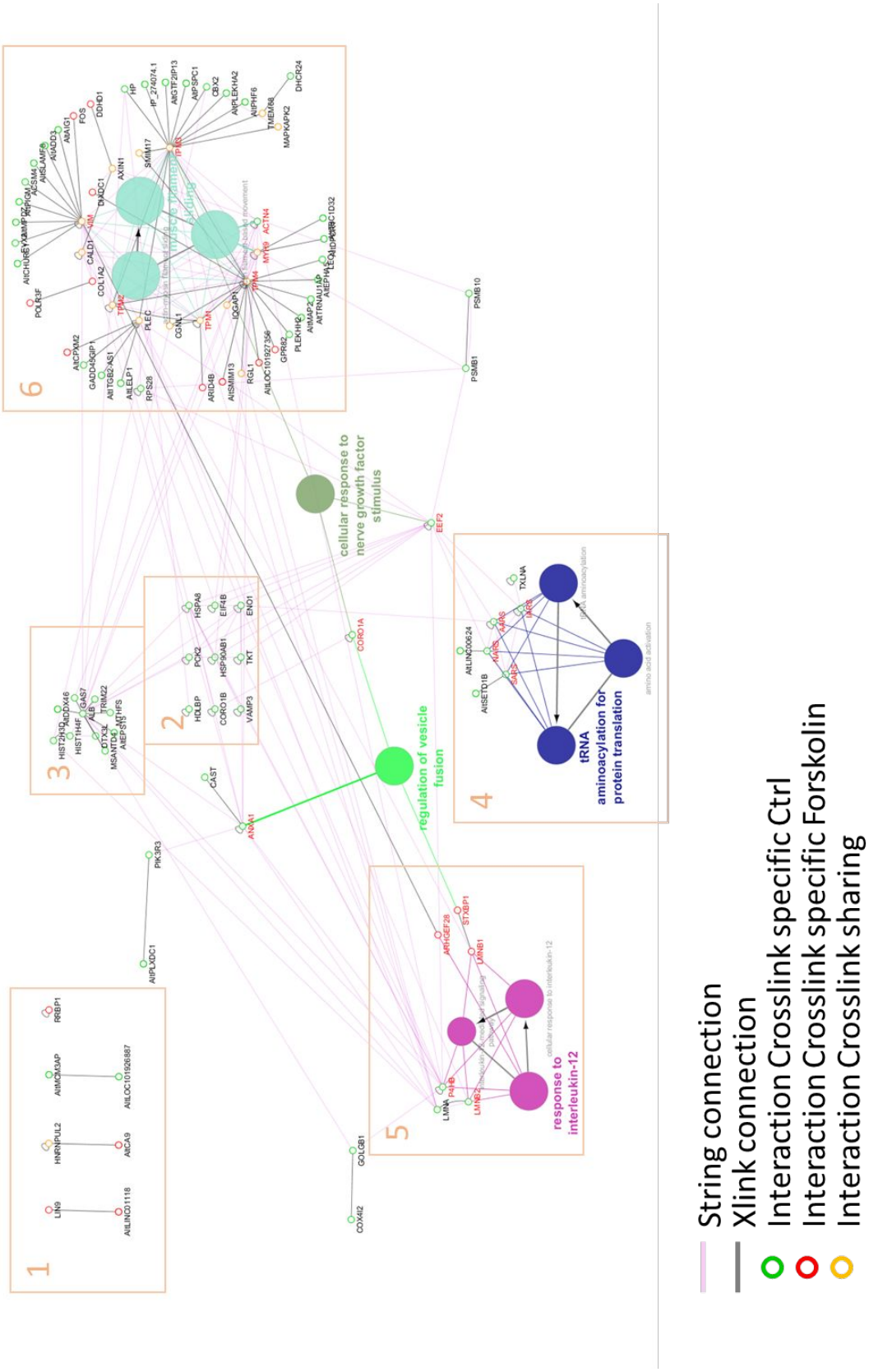


Figure 6

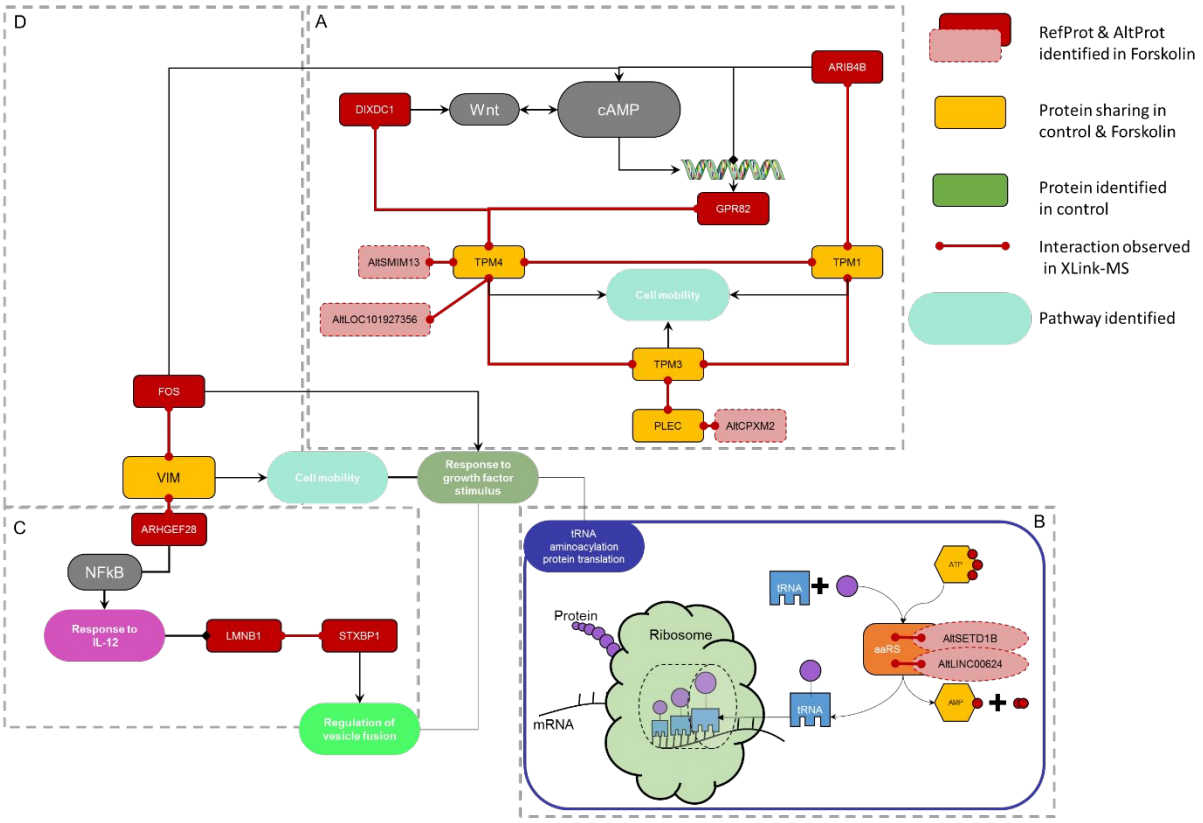
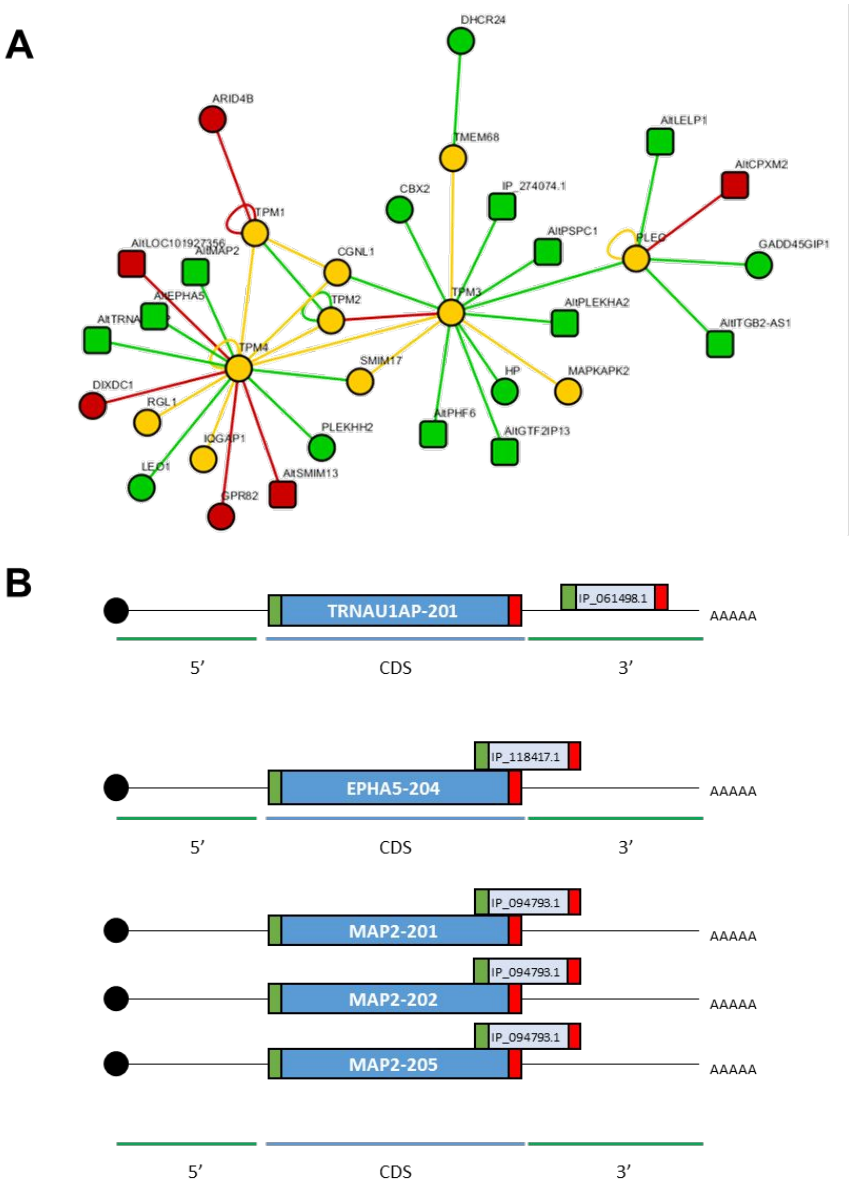


Figure 7



>IP_061498.1|altorf|Tax_Id=9606 OS=Human GN=54952 TR=NM_017846.4
MGFIYFHELQHLHIQGLDHFRINILTREKAAQKGLRKIAQSQSRPSV

>IP_118417.1|altorf|Tax_Id=9606 OS=Human GN=2044 TR=NM_001281767.1
MARGNQDGPVYRDFHGKWIQFNGRCGSGDLGVSNFSDNFYIIAGARKIVQKPIWMR

>IP_094793.1|altorf|Tax_Id=9606 OS=Human GN=4133 TR=NM_001039538.1 NM_002374.3
NM_031845.2 NM_031847.2
MLKPVWTMGLRSLHSPQADPAWHHPDDSAMSPRLEASTCSNLLSLPLWLRMSLLHSLSRACEYFSFSIEIIIFRHELLAGVGS
EQLLYSFFINHKINNLIPKL

Figure 8

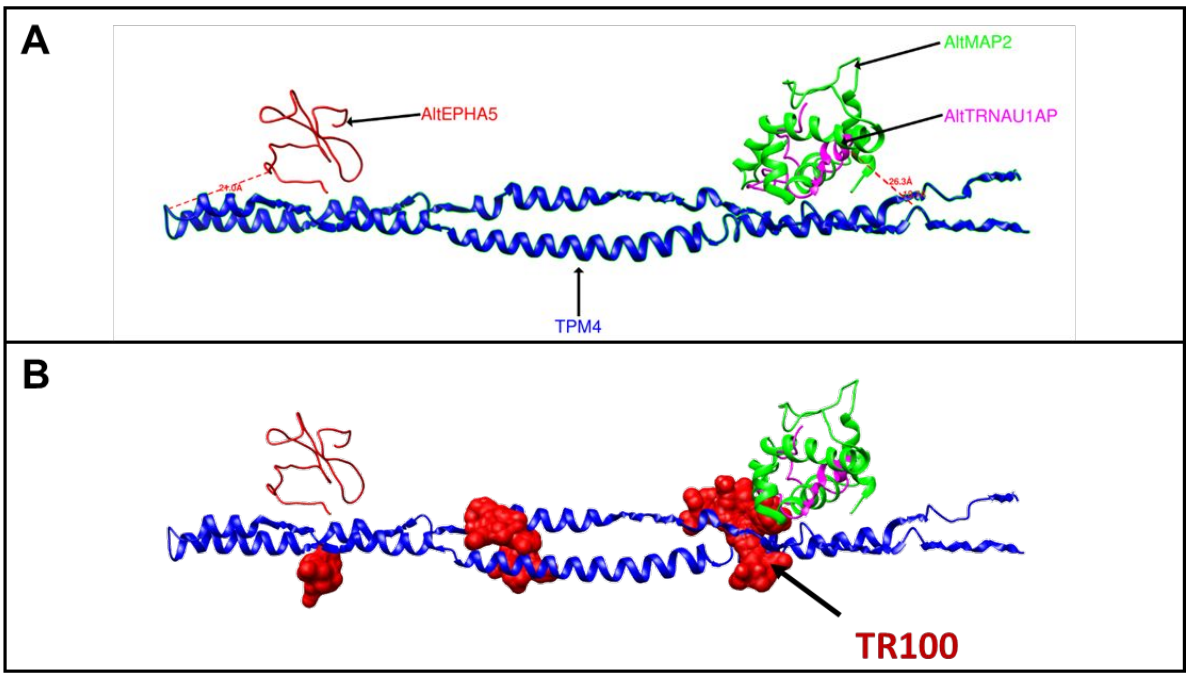


Figure 9

Supplementary data

Supp Figure 1.

Spectra of a crosslinking between a RefProt and an AltProt obtain by interrogation in Proteome Discovered 2.2 with the XLinkX node for DSSO. Here AltLINC0624 is connected to NARS and AltSETD1B is connected to SARS, this two proteins are involved on the tRNA regulation and the protein synthesis.

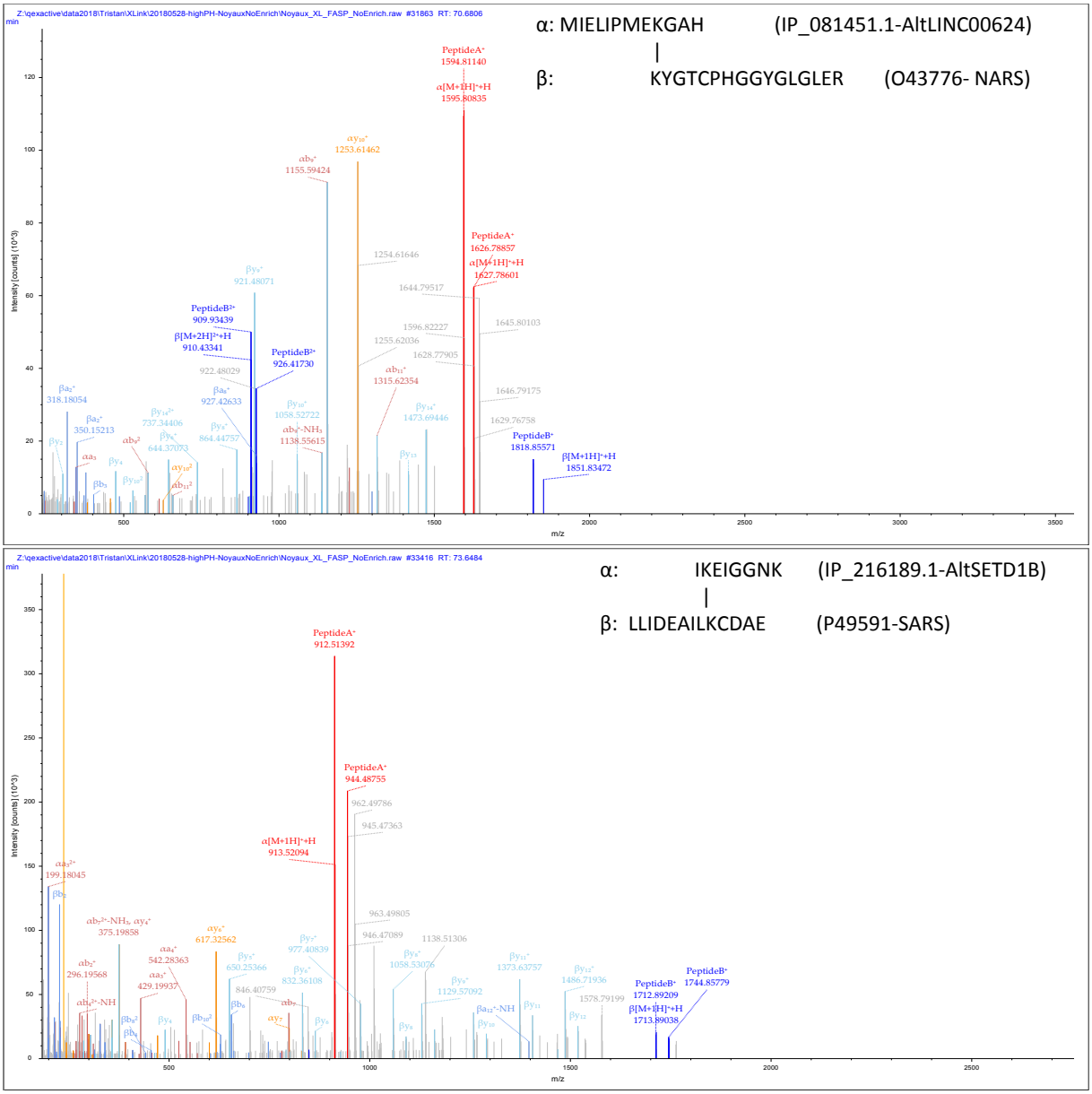
Supp Figure 2.

Cytoscape representation of the pathway connected to the protein discovered overexpress after 16h (light green) 24h (orange) and 48h (red) of stimulation, each pathway are divided function of the number of protein attributed to a time identification. Three group are visible name in grey no time dependent, hard green control, light green 16h and orange 24h specific pathway (ref. Figure 3).

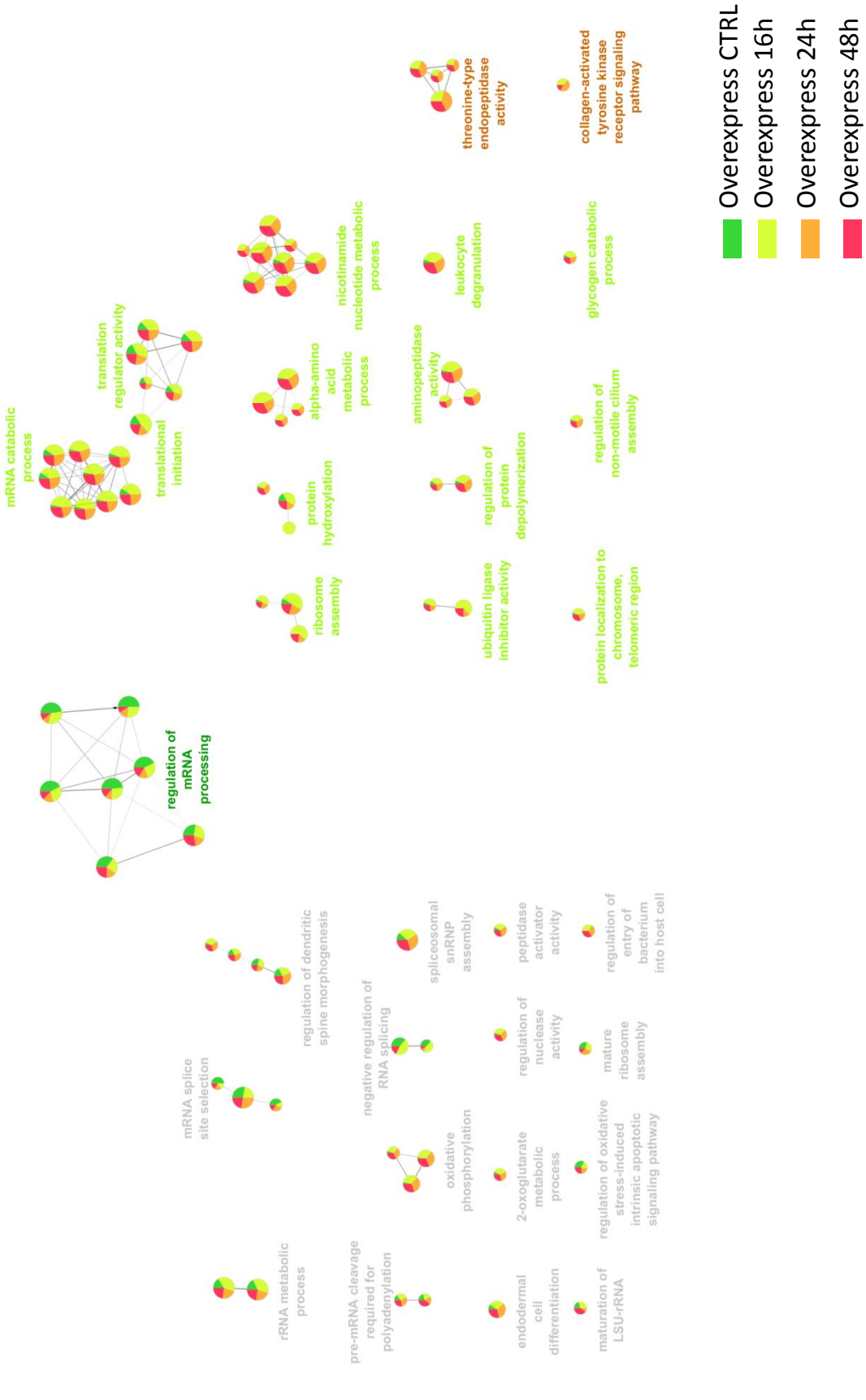
Supp Figure 3.

A- I-Tasser 3D model of TPM4 obtained from the Uniprot sequence of isoform 1, P67396-1. The best folded model were selected from all the models generated. B- I-Tasser Modeling of the three AltProts in interaction with TPM4 representation obtained with the sequences present in the database "HS_GRCh38_altorf_20170421". The best models are retained with a C-Score between [-5 and 2].

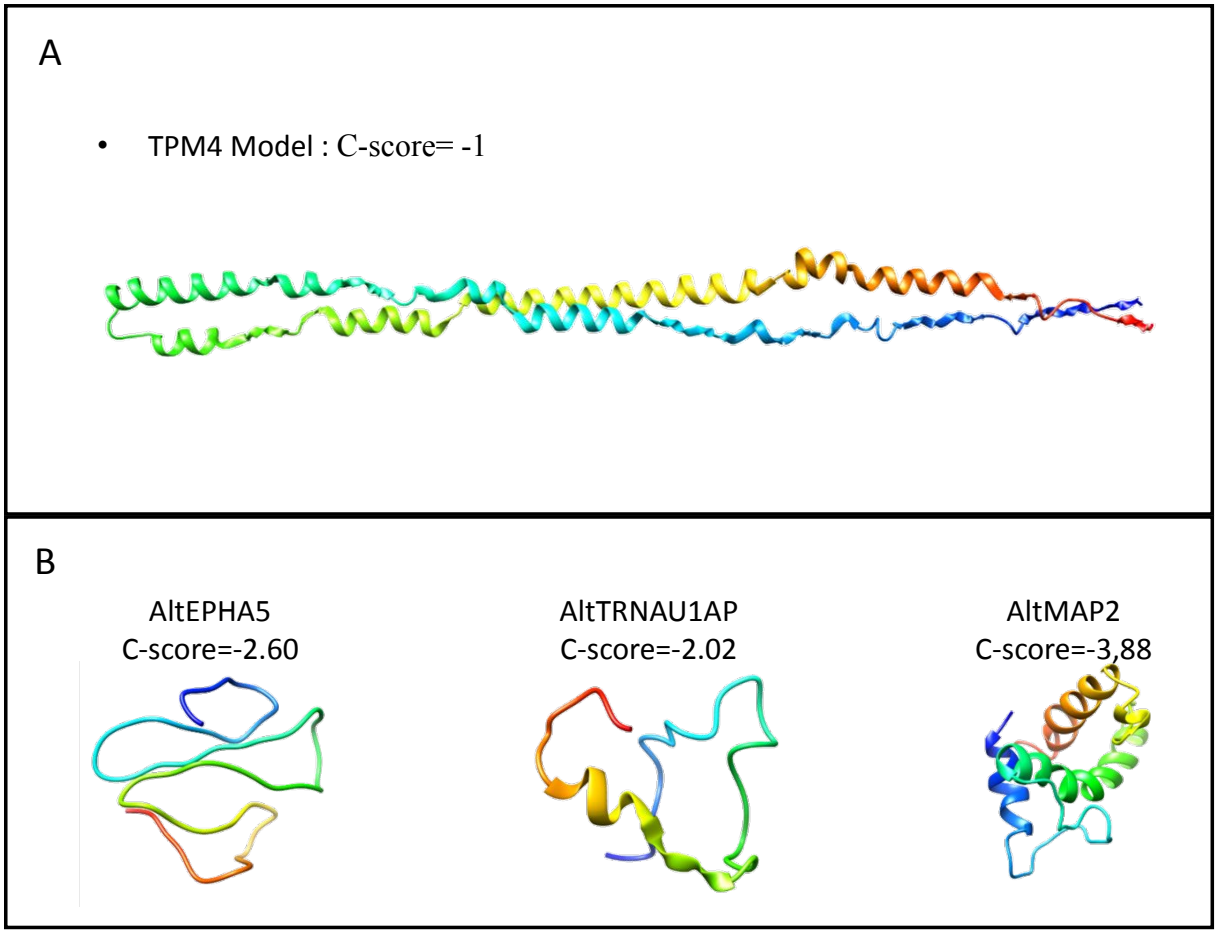
Supp Figure 1.



Supp Figure 2.



Supp Figure 3.



IV. Conclusion

Si quelques études décrivent depuis peu le rôle de quelques AltProts spécifiques [45,146], aucune étude à grande échelle n'avait encore été appliquée afin de mettre en évidence leurs fonctions. En effet, la connaissance limitée sur ces protéines ne permet que de réaliser des prédictions basées sur des homologies de séquences. Le développement de notre méthodologie est basé sur l'identification des partenaires d'interaction des AltProts par XL-MS. Cet outil nous permet, une fois les partenaires de l'AltProt identifiés de corréler les RefProts et leurs voies de signalisation. L'implication du partenaire d'une AltProt dans une voie de signalisation particulière, permet alors de compléter cette voie de signalisation par la présence de cette AltProt. L'AltProt identifiée peut alors être assignée à cette voie de signalisation. Bien évidemment l'implication d'une interaction ne répond pas à toutes les questions sur la fonction de cette AltProt dans la voie de signalisation : activation, inhibition, quel type de régulation et d'impact ? Les réponses à ces questions nécessiteraient une étude ciblée sur une AltProt d'intérêt, toutefois l'application de cette méthodologie est le premier pas vers la découverte de cibles d'intérêts.

L'utilisation de cette méthode dans le contexte pathologique de la reprogrammation de cellules cancéreuses de glioblastome de grade IV, sous l'effet de la Forskoline nous a permis de mettre en évidence des voies de signalisation d'intérêt. La Forskoline est un activateur des protéines kinases, en activant la voie cAMP elle provoque la consommation d'ATP par l'adénylate cyclase et la phosphorylation des protéines. Cette phosphorylation est à l'origine de nombreux mécanismes cellulaires dont la régulation de l'expression protéique. La Forskoline est présentée comme un composé aux vertus anti-tumorales, sur les cellules de gliome U87 elle est décrite comme permettant une transition phénotypique. Sous stimulation 48H à la Forskoline, les cellules U87 passent d'un phénotype astrocytaire à un phénotype neuronal [147]. De plus l'étude de l'environnement protéomique de la cellule montre la présence de protéines spécifiques des neurones telles que la protéine neuronale du noyau NEUN, ainsi que la diminution des marqueurs astrocytaires tels que la *Glial*

fibrillary acidic protein GFAP. Ces modifications décrivent clairement un changement phénotypique. Dans le cas des NCH82, aucune modification morphologique n'est visible. Toutefois l'étude du protéome total, montre un changement de l'environnement protéique des cellules sous stimulation. Les mêmes changements ont pu être mis en évidence dans l'étude protéomique des noyaux de ces cellules. L'utilisation de la méthodologie XL-MS GO-Terms assignation, a dans un premier temps permis d'étudier la modification dynamique du réseau avec notamment la transition entre 16H et la synthèse d'ATP vers l'hydrolyse de celle-ci par la myosine à 24H. Cette transition en accord avec l'effet de la Forskoline sur les cellules, permet d'identifier plusieurs AltProts dont AltSPTBN2, AltLATS2 à 16H et AltSIDT1, AltCFLAR à 24H. L'étude approfondie des voies de signalisation impliquées à 48h de stimulation montrent un nombre important d'AltProts impliquées dans diverses voies de signalisation : mobilité cellulaire, synthèse protéique, réponse à l'interleukine 12. L'observation de l'interaction des AltProts AltSETD1B et AltLINC00624, avec les protéines de régulation des ARNt de transcription et la synthèse de protéines, tend à confirmer l'implication des AltProts dans la régulation des ARNr [148], mais à un niveau de régulation différent. Comme énoncé précédemment [71], les AltProts peuvent-être des régulateurs d'expression des gènes, des ARN et des protéines, représentant un nouveau niveau de régulation d'expression. Dans cette étude les AltProts sont également présentes en grand nombre dans l'interaction avec le cytosquelette et la mobilité cellulaire. Dans un modèle de transition EMT ces informations grossissent un peu plus les données de compréhension de ces phénomènes. Le réseau formé par les TPMs présente quelques-uns de ses partenaires AltProts, comme AltTRNAU1P, AltEPH5A et AltMAP2 retrouvés en interaction avec TPM4. Les modèles de prédiction de l'effet d'un inhibiteur de TPMs, le TR100 montre la possibilité de rompre ces interactions comme entre une TPMs et une RefProt. Cela permet aussi de montrer que les AltProts peuvent se fixer sur le site actif d'une protéine.

Dans cette étude la démonstration de la méthodologie combinant l'identification des partenaires d'interaction par XL-MS à l'interprétation par les interactions décrites dans la littérature STRING, et la liaison aux GO-Term nous

permet d'appréhender les données protéomiques d'un nouveau point de vue. Il a ainsi été possible de mettre en évidence une communication protéine-protéine initiant le changement phénotypique et la transition EMT des cellules de gliome sous Forskoline, mais également la protéine LIN9 impliquée dans la reprogrammation des cellules souches embryonnaires. Couplée aux bases de données adaptées, cette stratégie nous permet de mettre en évidence les partenaires cachés de ces RefProts, tel que AltLINC01118 avec LIN9.

Cette méthodologie est facile à mettre en place et permet de compléter l'information protéomique obtenue par analyses *shot-gun*. Toutefois elle souffre de certaines limitations, dans un premier lieu il n'est pas possible de connaître la nature de l'interaction et l'effet de l'AltProt identifiée dans la voie de signalisation découverte. Une autre limitation est la quantité d'interactions identifiées, afin de rendre la méthode accessible aucun enrichissement n'est réalisé, toutefois un grand nombre d'interactions est perdu.

De manière paradoxale, bien que notre objectif dans une étude large échelle soit l'obtention d'un maximum d'informations, une autre limitation de cette stratégie réside dans la quantité d'informations produites. En effet un ensemble de protéines est observé en interaction auquel s'ajoute le protéome fantôme ainsi que l'enrichissement d'interaction STRING, l'identification des GO-term qui peuvent être réalisés à différents niveaux de représentation du GO-term par les protéines identifiées. L'ensemble du réseau ainsi formé doit alors être décrit, détaillé afin de permettre la compréhension. L'accès aux bases de données telles que GO-term et Reactome devient de plus en plus simple, et des logiciels tels que Cytoscape permettent aujourd´hui d'intégrer ces outils à l'étude des réseaux. Dans un avenir proche les méthodes d'étude de l'interactome pourront être mises en place en parallèle des méthodes classiques dans une vision de « combined shotgun analysis » accessible à tous.

PARTIE VI

Conclusion & Perspectives

I. Conclusion générale

Durant mes trois années de thèse, il m'a été possible de mettre en place de nouvelles stratégies d'analyses pour rechercher les fonctions des AltProts, à partir d'études à large échelle. Je me suis en particulier focalisé sur la mise en place et l'optimisation des stratégies XL-MS en association avec les bases de données AltProts. Le développement de cette stratégie et la combinaison avec des logiciels de traitement des réseaux d'interactions protéiques associés à un enrichissement des interactions connues tels que STRING ou Cytoscape ont également contribué à ces études.

Ce travail a permis de remettre en question plusieurs grands principes en biologie *i.e.* le fait qu'un ARNm ne peut coder que pour une seule protéine fonctionnelle. En effet, si les AltProts ont été mises en évidence depuis plusieurs années, de nombreux scientifiques restent convaincus que les protéines produites par ces mécanismes alternatifs sont non fonctionnelles. La mise en évidence des AltProts et leur corrélation avec des GO-term permettant de leur attribuer une voie de signalisation, remet définitivement en question ce dogme en démontrant que les protéines produites ont une fonction réelle. Leur existence est donc à présent une certitude et leur fonction indéniable ; les études ciblées permettent de confirmer leurs interactions avec les protéines de référence ouvrant vers une nouvelle vision des processus physiologiques et physiopathologiques.

La principale limitation liée à l'étude des AltProts résidait dans l'absence d'anticorps dirigés contre ces protéines. Pour cela les stratégies ciblées restent limitées à quelques candidats d'intérêt tout particulier pour lesquels des anticorps doivent être produits à façon. Il a donc fallu développer de nouvelles stratégies avec les outils d'analyses à large échelle existants. La méthode XL-MS s'est avérée parfaitement adaptée. Cette stratégie permet de figer l'ensemble des PPIs présentes dans la cellule, et d'identifier de manière spécifique un grand nombre d'interactions issues d'un échantillon complexe.

L'identification d'interactions localisées sur les protéines, nous offre la possibilité de reconstruire le réseau d'interactions entre ces protéines, bien que la méthode souffre d'une faible dynamique temporelle et ne permette pas de déterminer précisément à quel moment les partenaires sont fixés entre eux. Par conséquent dans l'observation d'une interaction à trois partenaires il est impossible de différencier si l'interaction a lieu simultanément entre les trois ou en cascade un par un.

L'analyse des réseaux ainsi formés permet toutefois de regrouper les protéines suivant leur voie de signalisation, et ainsi par l'identification de partenariats encore inconnus de décrire des nœuds entre les différentes voies de signalisation. De plus l'utilisation des bases de données AltProts, identifiant des interactions AltProt-RefProt permet de compléter les voies de signalisation connues avec ces nouvelles protéines pouvant être impliquées dans des mécanismes essentiels pour la cellule.

1. Rôle et fonction des AltProts

Après la mise en évidence des AltProts dans les cellules, la question principale a été de déterminer la fonction de ces protéines fantômes.

A. Un niveau de régulation supplémentaire pour les gènes

Les premières prédictions de fonctions par analyse d'homologie de séquence avec des RefProts, ont montré une forte probabilité d'implication dans les voies de régulation de l'expression des protéines [23]. Nous avons avancé le même rôle pour les AltProts présentes sur l'ARNm codant pour ZEB1/2 et d'autres marqueurs de la transition épithéliale-mésenchymateuse, régulant ainsi la différenciation cellulaire et la reprogrammation cellulaire [71]. Lors des études, j'ai également pu observer à plusieurs reprises un rôle des AltProts dans la régulation de l'expression protéique. Par exemple, l'interaction de AltATAD2 avec RPL10 semble inhiber la fixation de l'ARNr 5S dans les cellules HeLa [148] avec AUF1 qui est un régulateur des régions riches en AU des ARNm ; ou encore avec l'interaction observée d'AltProts avec des protéines aaRS impliquées dans la synthèse protéique dans les cellules NCH82 sous Forskoline. L'ensemble de ces

observations apportent des arguments supplémentaires à l'hypothèse d'un rôle fort des AltProts dans la régulation de la transcription et la traduction des protéines, un niveau de régulation encore peu étudié.

B. Un système de « secours » pour la cellule

L'observation d'importantes homologies de séquence entre des AltProts et des RefProts issues de régions et de gènes différents, laisse à penser que ces protéines permettraient également la mise en place d'un système de « *secours* », afin de compenser la perte fonctionnelle d'une protéine lors d'une mutation. Cette hypothèse permettrait d'expliquer pourquoi l'inhibition d'une protéine dans un but thérapeutique conduit souvent à des résultats moindres qu'attendus. Ainsi, l'AltProt ayant une homologie de séquence pourrait alors compenser l'inhibition de la RefProt en la remplaçant. Cette hypothèse est intéressante mais risque de soulever un problème majeur lors de l'édition du génome comme traitement. En effet, ces dernières années, l'édition génétique est de plus en plus souvent évoquée comme solution de traitement car considérée comme une solution efficace et précise pour la modification d'une cible protéique. Toutefois la modification d'un gène, entraîne une modification de l'ARNm avec une substitution, une addition ou encore une délétion d'un ou plusieurs nucléotides ayant pour conséquence la disparition du codon « *Start* », l'apparition d'un codon « *Stop* » prématuré ou encore la modification du code modifiant la séquence en acides aminés de la protéine résultante. L'ensemble de ces modifications peut également engendrer l'expression de nouvelles AltProts, une modification dans leur séquence ou encore leur inhibition. Les systèmes d'édition du génome doivent donc être minutieusement contrôlés afin de vérifier que l'impact de la modification protéique soit bien localisé à un seul transcrit et ne touche aucune autre protéine qu'elle soit RefProt ou AltProt.

Une étude est actuellement menée au laboratoire afin d'invalider une protéine alternative observée chez le rat nommée « Heimdall », celle-ci sera présentée un peu plus tard. L'application de la méthode CRISPR y sera décrite, avec son impact sur le protéome global et les AltProts.

2. Limitation de la méthode XL-MS

Si à l'heure actuelle la méthode XL-MS est de plus en plus répandue et décrite dans la littérature, elle reste une méthode, pour laquelle un grand nombre d'optimisations sont encore nécessaires. La principale limite de cette stratégie réside dans le taux de peptides pontés présents après digestion d'un mélange complexe ponté. Le nombre de peptides pontés dans le mélange est largement minoritaire comparé aux peptides libres. Afin de résoudre ce problème des méthodes d'enrichissement sont décrites (colonne SEC et SCX). Toutefois, si la mise en place de ces méthodes telles que décrites dans la littérature semble simple, en réalité il en est tout autre. D'importants efforts ont été mis en œuvre afin d'obtenir un enrichissement satisfaisant, celui-ci permet notamment d'augmenter jusqu'à trois fois le taux d'identification de 10 peptides pontés contre 30 après un enrichissement sur StageTips SCX [119] pour de la BSA pure. Une étude récemment réalisée pour un extrait total entre les fractions cytoplasmiques de cellules NCH82 sous pontage au DSSO pour différents types d'enrichissements (séparation sur gel, StageTips C18 et StageTips SCX) montre une nette amélioration du nombre de d'interaction détectés (**Figure 18**). En effet, 190 pontages ont été détectés après StageTips C18+SCX dans la fraction cytoplasme contre 45 en extraction de gel et 38 sur StageTips C18. Ces résultats sont très encourageants mais des méthodes à haut débit restent encore à développer.

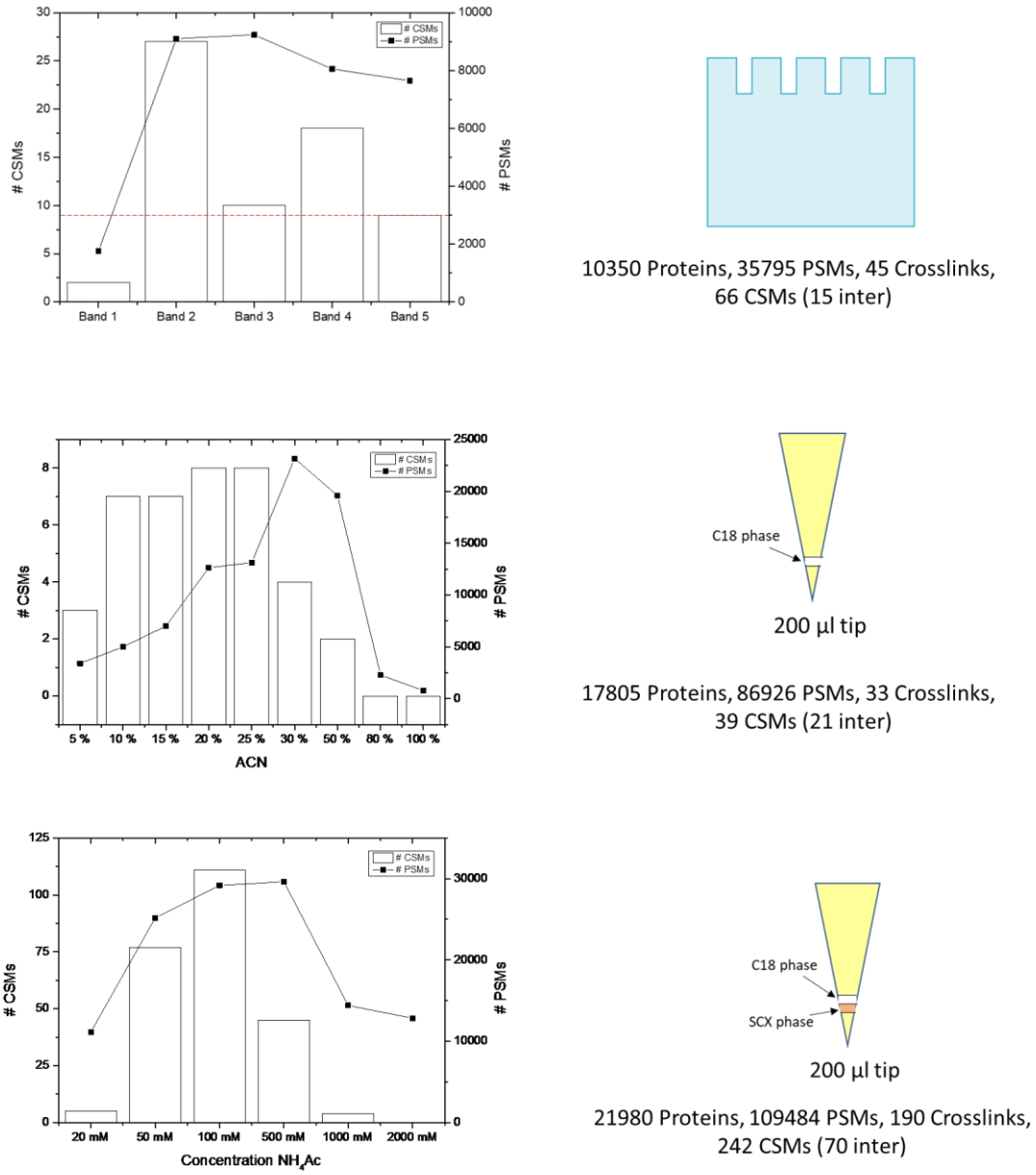


Figure 18 : ***Application de méthodes d'enrichissement des peptides pontés sur mélange complexe.*** *Comparaison de différentes méthodes d'enrichissement des peptides pontés issus du pontage de la fraction cytoplasmique des cellules NCH82. Comparaison des stratégies fractionnement sur gel SDS-PAGE, Stage-Tip C18 et Stage-Tip C18+SCX*

Notre choix était de dé-complexifier l'échantillon avant pontage. Dans ce contexte, nous avons réalisé nos études sur des extraits cellulaires après séparation de la fraction cytoplasmique du reste des organites (fraction noyau) pour une analyse globale, permettant quand même de mettre en évidence des variations d'interactions et les voies de signalisation en fonction du traitement ou non. Des efforts doivent encore être réalisés afin de permettre un enrichissement

stable et spécifique des peptides ou protéines pontés. Ceci pourra être réalisé par le design de nouveaux agents de pontages fonctionnalisés mais n'entraînant pas d'encombrement stérique ou d'autres phénomènes physico-chimiques déstabilisant l'interaction protéique (agents pontant dont le groupement fonctionnel est ajouté post réaction par chimie clic).

II. Perspectives

Si aujourd'hui la fonction des AltProts a été approchée dans le noyau par la méthode XL-MS il est évident que la suite de ce projet est de pouvoir appliquer cette méthodologie à l'ensemble de la cellule.

1. La fonction cachée du protéome fantôme

Les AltProts sont des protéines de petite taille facilement comparables aux neuropeptides. Les neuropeptides sont connus pour leur rôle de messager, permettant la communication entre neurones et le transfert d'informations par activation de voies de signalisation. Ces peptides sont décrits pour être sécrétés puis retrouvés couplés à des récepteurs généralement aux protéines G. Les neuropeptides sont aussi connus pour avoir un temps de vie court dans la cellule. L'étude des AltProts sécrétées par les cellules et notamment par des neurones n'a pas encore été réalisée. Toutefois, tout laisse à penser que les AltProts pourraient jouer les mêmes fonctions régulatrices que les neuropeptides. Les doutes de la communauté envers l'existence et la fonction des AltProts reposant parfois sur un temps de vie, jugé trop court pour exprimer une fonction, seraient alors également réfutés. Si tel est le cas, il est alors envisageable que les AltProts soient finalement les ligands de récepteurs comme les récepteurs à la protéine G orphelins. En effet si ces récepteurs sont connus et caractérisés, aucune protéine n'a encore pu être identifiée comme leur activateur. Pour ce faire il serait intéressant de réaliser une étude d'identification des AltProts à la surface des membranes cellulaires, là où se trouvent ces récepteurs. Une étude XL-MS, intégrant un agent de pontage non perméable aux membranes comme le BS3 pourra être envisagée. Afin de faciliter l'analyse, un enrichissement de la fraction membranaire après pontage avec un agent clivable comme le 3,3'-

Dithiobis(sulfosuccinimidylpropionate) (DTSSP) peut également être réalisé. L'étude du sécrétome cellulaire et de l'interactome à la membrane permettra de compléter l'éventail des fonctions déjà observées.

2. Stratégie ciblée : Heimdall

En parallèle de mes recherches d'identification des voies de signalisation à large échelle, une étude, est conduite à l'heure actuelle afin de mettre en évidence la fonction d'une AltProt identifiée comme étant surexprimée à 12h après lésion de la moelle épinière chez le rat. Cette étude fait actuellement l'objet d'un dépôt de brevet. Toutefois, succinctement l'AltProt identifiée, nommée « Heimdall », a été retrouvée dans les données d'une étude conduite au laboratoire à partir des données de protéomique obtenues de manière spatio-temporelle après lésion de la moelle épinière de rat [149,150]. Cette étude a clairement décrit les modifications protéomiques survenant après lésion, mais a également montrée la capacité des astrocytes à produire des immunoglobulines. Lors de la recherche d'AltProts dans ces données, il nous est apparu qu'une AltProt était surexprimée spécifiquement dans la région de la lésion à 12H puis au cours du temps au niveau des régions rostrales et caudales les plus proches de la lésion, 3 jours après le traumatisme pour être répartie sur l'ensemble de la moelle épinière 10 jours après lésion (**Figure 19** et **20**).

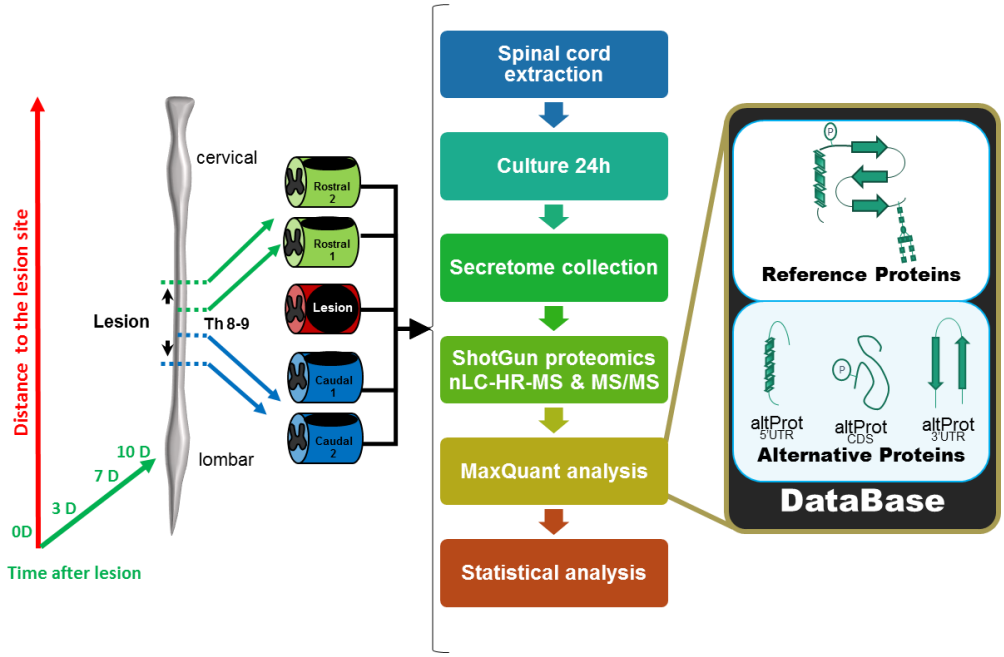


Figure 19 : ***Schématisation de l'étude spatio-temporelle sur la moelle épinière de rat lésé****. Intégration de l'utilisation des bases de données d'AltProts, lors de la ré-analyse des données générées précédemment.*

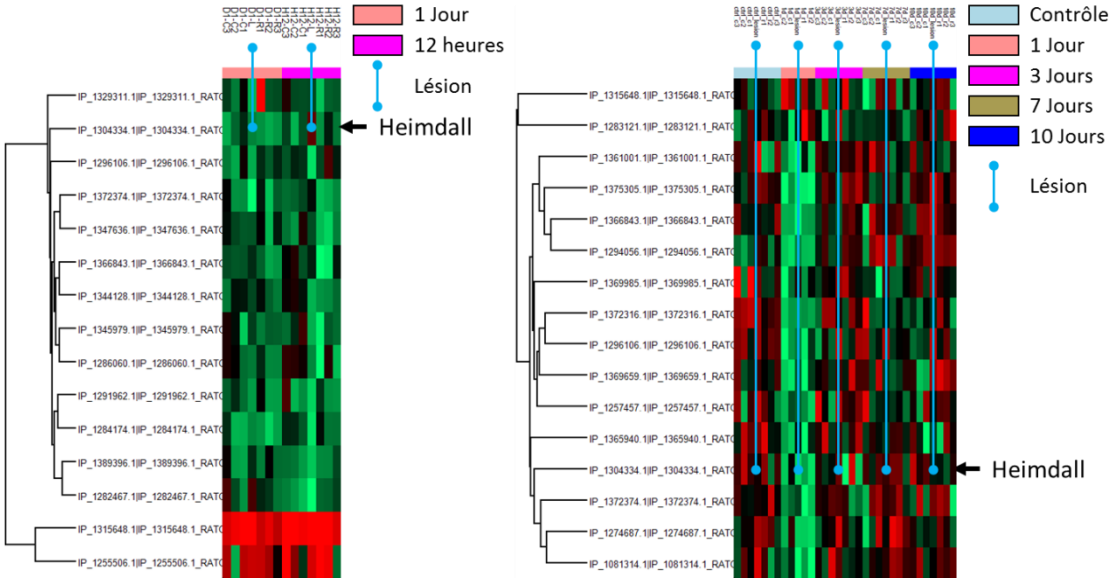


Figure 20 : ***Heatmap représentant les variations d'expression d'AltProts à différent temps après lésion.*** *À 12h, 1Jour, 3 jours, 7 jours et 10 jours, en fonction des régions caudales et rostrales entourant le site de lésion. Heimdall est observée surexprimée dans la lésion à 12h et présente des variations d'expression dans les régions environnant la lésion à différents temps.*

Heimdall, présente la particularité d'être une protéine issue d'un ARNnc. L'étude de sa séquence indique la présence d'un domaine IgG. Sa séquence indique une homologie avec celle de la chaîne légère Kappa et plus particulièrement la partie variable (**Figure 21.A**). De manière encore plus

intéressante, la séquence possède des motifs connus (**Figure 21.C**) présents au sein des protéines intrinsèquement désordonnées (IDP) décrites pour être à l'origine de pathologies telles que la maladie de Bence Jones dans laquelle ces parties non conformées forment d'hyper structures moléculaires (**Figure 21.B**), donnant lieux à des fibrilles dans les tissus. Ainsi Heimdall, par sa proximité de séquence pourrait avoir un effet similaire.

A

Rattus norvegicus immunoglobulin kappa variable 12S38

	from	to	length	matches	mismatches	gaps	identity(%)
FR1-IMGT	21	46	26	26	0	0	100
CDR1-IMGT	47	52	6	6	0	0	100
FR2-IMGT	53	69	17	17	0	0	100
CDR2-IMGT	70	72	3	3	0	0	100
FR3-IMGT	73	108	36	36	0	0	100
CDR3-IMGT (germline)	109	115	7	7	0	0	100
Total			95	95	0	0	100

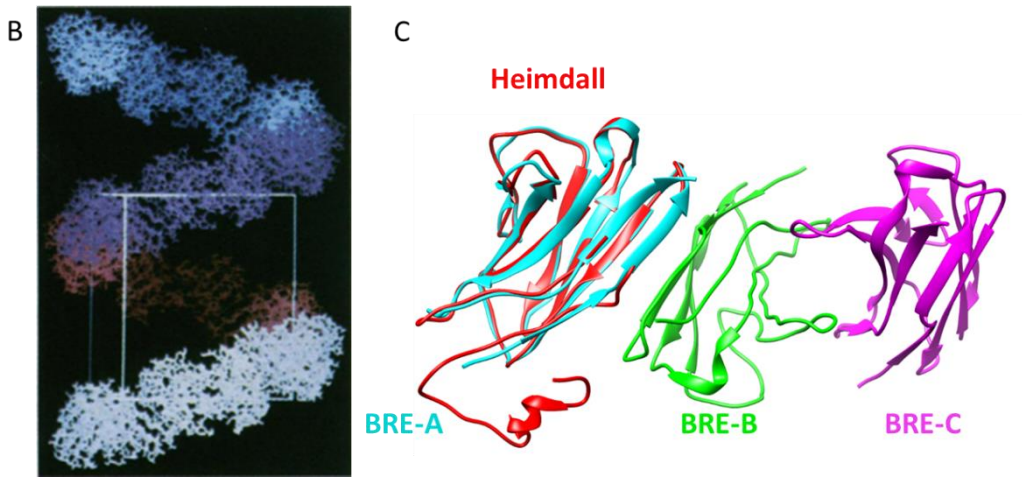


Figure 21 : ***Prédiction de la conformation de Heimdall et de son impact dans la cellule*** *A. Résultat de la recherche d'homologie entre Heimdall et les IgG de rat, la partie 21 à 115 de Heimdall est similaire à la partie variable de l'Ig Kappa 12S38. B. Représentation de Schormann & al. en 1995 représentant la cristallisation des Immunoglobulines BRE dans la pathologie de Bence Jones et d'amyloidose. (Copyright (1995) National Academy of Sciences) C. Prédiction structurale de Heimdall, orienté par la comparaison avec BRE, celle-ci semble pouvoir présenter la même structure et donc les mêmes régions intrinsèquement désordonnées (IDP).*

Cette AltProt semble présenter une double fonction. Elle possède un domaine IgG, lui permettant d'interagir au sein de la cellule avec d'autres protéines la rapprochant des fonctions jusque-là décrites pour les AltProts telles que la régulation d'expression de gènes. Mais la présence dans sa séquence de

structures IDP laisse penser qu'elle peut être sécrétée et jouer un rôle dans la communication, la structure des tissus ou encore la défense de l'organisme.

Un anticorps basé sur les séquences spécifiques identifiées en nLC-MS/MS de l'AltProt a été produit. Cet anticorps nous permet aujourd'hui d'aller plus loin dans l'étude de la fonction d'Heimdall, notamment au travers d'études par Western-Blot, coIP et coIP-XL-MS (sujet faisant actuellement l'objet d'une thèse) sur des cellules astrocytaires de rat (DITNC1). Cet anticorps nous a également permis de mettre en évidence que lorsque les astrocytes de rat sont incubés 24H avec l'anticorps, inhibant les interactions entre Heimdall et les autres protéines, les astrocytes se dédifférencient en un phénotype neuronal. La même observation semble être confirmée avec les premiers tests réalisés après inhibition de Heimdall par CRISPR-Cas9 (expériences en cours).

Pour ce faire l'ADN est coupé afin d'engendrer une modification sur l'ARN résultant. La modification créée par l'utilisation du CRISPR a pour effet l'inhibition de la fonction de Heimdall par l'insertion d'un codon STOP prématuré (**Figure 22**). Lorsque l'on étudie la séquence nucléotidique de cet ARNnc (**Figure 23**), il est possible d'observer plusieurs séquences de traduction issues des différents cadres de lecture : *5' Frame1, 2 et 3*. La prédiction de ces séquences protéiques, avant et après CRISPR (**Tables 3** et **4**) nous montre que Heimdall est significativement réduite. Cependant, on observe également un grand nombre de modifications intervenant sur les séquences suivant le site de coupure de l'ARNnc.Cette modification d'expression par CRISPR a été confirmée par Western-Blot ciblant l'AltProt et observant la disparition de la détection de la protéine après CRISPR.

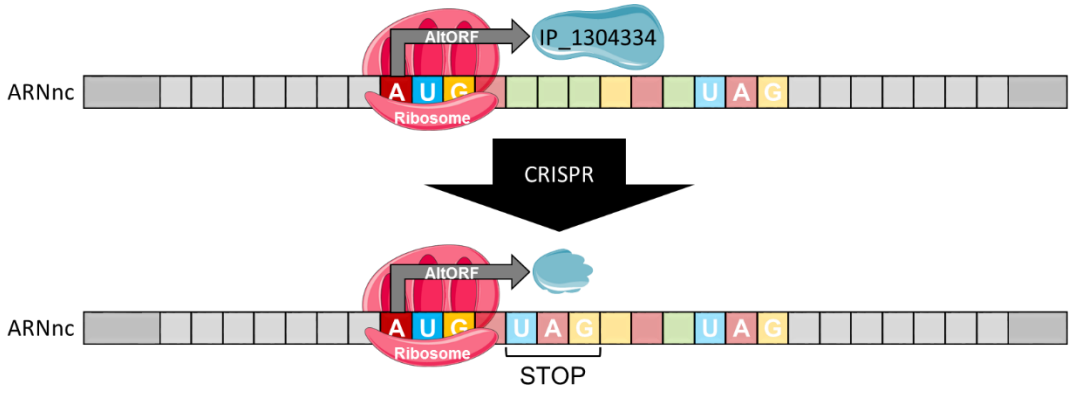


*Figure 22 : **Design de la méthode CRISPR-Cas9 appliqué à l'AltProt Heimdall.** La méthode CRISPR est utilisée afin d'insérer un codon « STOP » dans la séquence de l'ARNnc, codant normalement pour l'AltProt IP_1304334, celle-ci est alors inhibée.*

>AABR07051592.2-201 ENSRNOE00000542240 exon:NOVEL_lincRNA

GTGTGTGTGCATGCATAGCATATAAGTTCATAGGAAAAATGGATAAAGAAACCGTGAATTTGAAAAAGAACAGGGACAACTATGGGG
GAGACTGTTAGTAAGGAAAGGTAAGAGAAAAATGTTCTAATTAAAATTCGGTCTCACAAACATATGGAATGGAAAAAAAAGCAGAAG
AAGCTGTATGGTAACAATGACATCTACACCTAACTTTGGAAGGATAACGGAACCAGTTATGTACTGCCTAGAACAGTACGACATATGA
AAAACCTGGACATGGAGGAAAACACATGGCAACTCTGAACAGAACATGATAAAAATCAAATAATGTTTTACTCATAATCTTTGGCCCTT
GTTTGTTGTTCAAACAAGAGCACTATCTCTACATATAAACCTGTACTGTGTCTTAACTGGAGAATTAAGAGAGCCACAGGTCTATAATG
AGTACTGACCTAATGAACCCTGCAGCTCTGCCTACCCTTTGCTGATTTGCATATATCCAAAGCACTTACATGAGGACTTCTTCATAATCA
GGTCACACCCTGTGTTGGAATCAGCCTCATAGAGATCACACACAGAC**ATGGGTGTGCCCACTCAGCTCCTGGGGTTGTTGCTGCTGTG**
GATAACAGATGGCATATGCGACATCCAGATGACACAGTCTCCAGCTTCCCTGTCTGCATCTCTGGGAGAAACTATCTCCATCGAATG
TCGAGCAAGTG**AGGACATTTACAGTAATTTAGCGTGGTATCAGCAGAAGTCAGGGAAATCTCCTCAGCTCCTGATCTATGCTGCAAA**
TAGGTTGCAAGATGGGGTCCCATCACGGTTCAGTGGCAGTGGATCTGGCACACAGTATTCTCTCAAGATCAGCGGCATGCAACCTG
AAGATGAAGGGGATTATTTCTGTCTACAGGGTTCCAAGTTTCCTCCCACAGTGATTCAAGCCATGACATAAACCATGCAGGGAAGCA
GAAGTGAGAGTACAGGGTGCCCCAGC

*Figure 23 : **Séquence de l'ARNnc codant Heimdall**. La séquence codante pour l'AltProt IP_1304334 est identifiée en gras et en vert la partie de séquence ciblée par les sondes coupées par CRISPR.*

Table 3 : Prédiction des séquences protéiques possiblement exprimées par l'ARNnc origine de l'AltProt IP_1304334

Prédiction de transcription ENSRNOE00000542240		
5'3' Frame 1	**5'3' Frame 2**	**5'3' Frame 3**
MGETVSKER	**M**HSI	**M**DKETVNLKKNRDNYGGDC
MEWKKKQKKLYGNNDIYT	**M**F	**M**EKKAEEAVW
MKNLD**M**EENTWQL	**M**VT**M**TSTPNFGRITEPV**M**YCLEQYDI	**M**ATLNRT
MFYS	**M**IKIK	**M**STDL**M**NPAALPTLC
MRHPDDTVSSFPVCISGR NYLHR**M**SSK	**IP_1304334 :** **M**GVPTQLLGLLLLWITDGICDIQ**M**TQSPASLSASLGETISIE CRASEDIYSNLAWYQQKSGKSPQLLIYAANRLQDGVPSRFS GSGSGTQYSLKISG**M**QPEDEGDYFCLQGSKFPPTVIQA**M**T	**M**RTSS
	MQGSRSESTGCPS	**M**AYATSR
		MLQIGCK**M**GSHHGSVAVDLAH SILSRSAACNLK**M**KGIISVYRVPSF LPQ

Table 4 : Prédiction des séquences protéiques possiblement exprimées par l'ARNnc origine de l'AltProt IP_1304334 après application de la modification par CRISPR, en bleu les séquences identiques avant et après CRISPR, soulignées les séquences modifiées

Prédiction de transcription ENSRNOE00000542240+CRISPR		
5'3' Frame 1	5'3' Frame 2	5'3' Frame 3
MGETVSKER	MHSI	MDKETVNLKKNRDNYGGDC
MEWKKKQKKLYGNNDIYT	MF	MEKKAEEAVW
MKNLDMEENTWQL	MVTMTSTPNFGRITEPVMYCLEQY DI	MATLNRT
MRHPDDTVSSFPVCISGRNYLRTFT VI	MIKIK	MSTDLMNPAALPTLC
MLQIGCKMGSHHGSVAVDLAHSIL SRSAACNLKMKGIISVYRVPSFLPQ	MGVPTQLLGLLLLWITDGICDIQMT QSPASLSASLGETISGHLQ	MRTSS
		MAYATSR
		MQPEDEGDYFCLQGSKFPPTVIQA MT
		MQGSRSESTGCPS

Suite à la modification des cellules par CRISPR, l'étude des modifications protéiques entre cellules contrôles et cellules modifiées montre des changements significatifs dans les voies de signalisation impliquées (**Figure 24.A**). La même expérience a été réalisée sous stimulation au lipopolysaccharide (LPS) un agent bactérien stimulant la voie TLR4 dépendante de la réponse inflammatoire. Heimdall a notamment été décrite dans les cellules astrocytaires de rat pour être surexprimée lors de cette stimulation. Dans l'étude des voies de signalisation dédiées aux protéines exclusives aux conditions contrôles et KO, on observe que sans stimulation l'effet du KO reste relativement restreint. On peut tout de même observer des modifications dans l'expression de la voie de transport des vésicules synaptiques antérogrades. Ceci est corrélé avec le changement phénotypique observé d'astrocytes en neurones (**Figure 24.A**). Dans la condition de stimulation sous LPS, les protéines exclusives au KO sont beaucoup plus importantes couvrant des voies de signalisation plus nombreuses (**Figure 24.B**). On notera par exemple de nombreuses protéines impliquées dans les processus métaboliques de l'ARN. Cette voie de signalisation modifiée après inhibition d'une AltProt (Heimdall) est en accord avec les fonctions observées jusqu'à maintenant, de régulation de la transcription et de la traduction de

RefProts. La voie de la réponse à l'interféron I, va dans le sens d'une fonction d'activateur des récepteurs membranaires tels des neuropeptides.

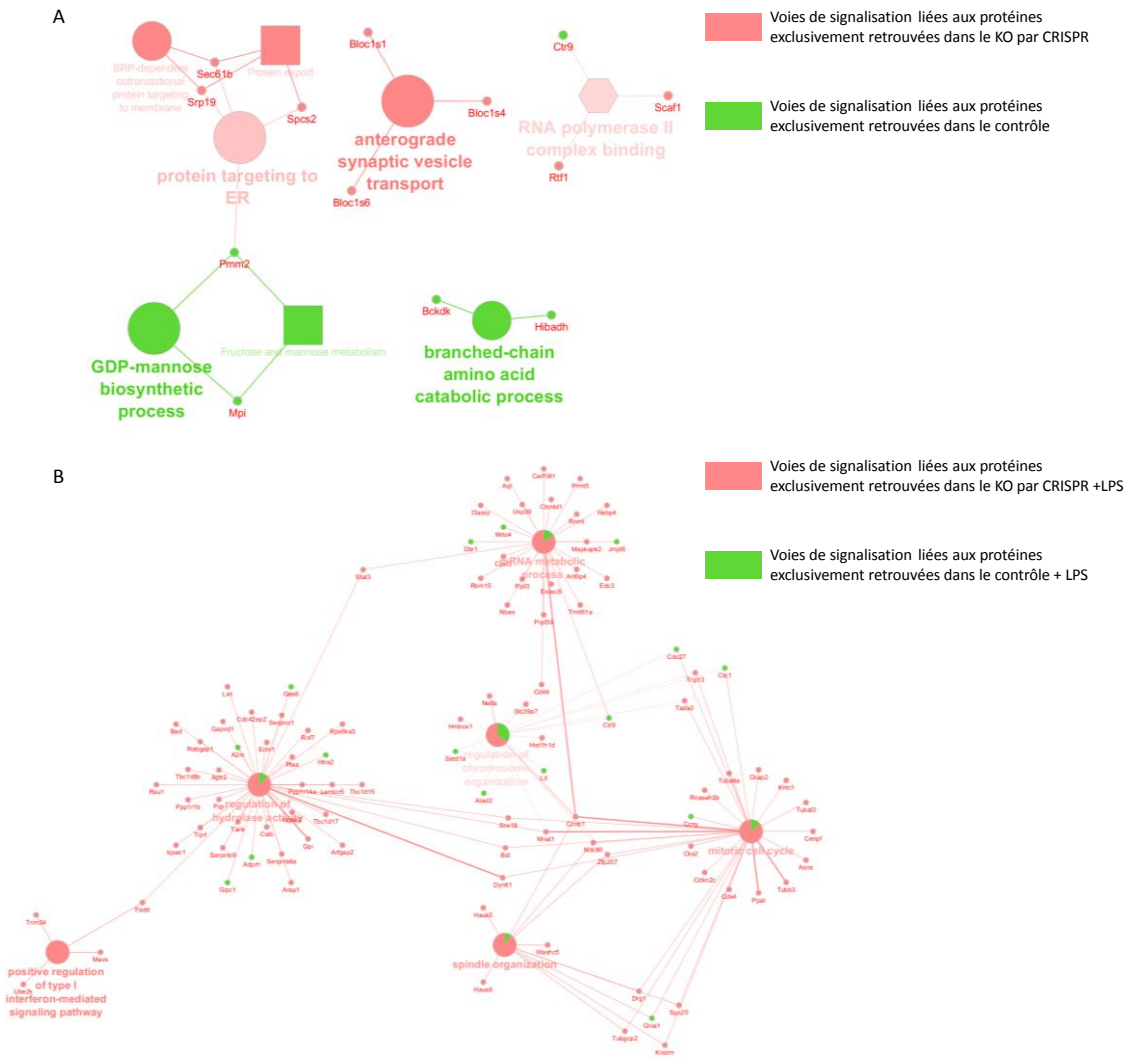


Figure 24 : ***Description des modifications des voies de signalisation après inhibition de l'AltProt Heimdall.*** *Voies de signalisation impliquant les protéines exclusives identifiées, Avant et Après CRISPR ciblant Heimdall, dans les cellules DITNC1 astrocytes de rat A. sans stimulation, B. avec une stimulation de 48h au LPS. En rouge les voies de signalisations observées exclusivement dans les cellules KO Heimdall, en vert les voies de signalisations exclusives aux cellules contrôles. On peut observer l'implication de Heimdall dans des processus biologiques au métabolisme d'ARN mais aussi au cytosquelette.*

Cependant il est impossible de savoir si les modifications observées sont issues de l'inhibition seule de Heimdall, ou si elles sont également liées à la modification des séquences suivantes. Afin de déterminer l'implication des AltProts dans ce contexte une étude du protéome par XL-MS est en cours de

réalisation. L'objectif étant de mettre en évidence l'implication de protéines spécifiquement produites après la réalisation du CRISPR.

Cette étude permettra également de mettre en évidence les limites encore incomprises de l'utilisation de la méthode CRISPR-Cas9. Cibler l'ADN afin de modifier à terme l'expression de protéines, néfastes dans certaines pathologies est une chose. Cependant il semble clair que l'utilisation de modifications géniques de ce genre, même si elles ne touchent qu'un seul gène provoquent des répercussions sur l'ARN résultant de l'ADN. Ces modifications de l'ARN ne posent pas de problèmes dans un système monocistronique, toutefois prenant compte de toutes les séquences possiblement traduites, cette méthode peut entraîner des modifications beaucoup plus profondes.

Ces résultats préliminaires semblent montrer un rôle de régulation des AltProts dans le processus de neurogénèse après un traumatisme. Ces hypothèses et observations demandent encore à être validées. Toutefois Heimdall semble essentielle dans le mécanisme de différenciation après lésion de la moelle épinière chez le rat.

3. Transfert sur tissu

Un des savoir-faire en protéomique du laboratoire PRISM réside dans l'analyse d'images réalisées par spectrométrie de masse permettant de déterminer la distribution des composés endogènes (ou exogènes) dans des sections minces de tissus et en déduire la localisation de régions d'intérêts présentant un phénotype cellulaire différent. L'utilisation de méthodes de micro-extraction des protéines, permet l'analyse à large échelle des régions d'intérêts ainsi identifiées par imagerie MS. Depuis plusieurs années le laboratoire développe des méthodes de micro-protéomique ou encore nommée Spatially-Resolved Proteomics telles que la microdissection assistée par Parafilm (PAM) [151], la microextraction liquide de surface (LESA) [67], l'extraction liquide de surface par capillaire (FAMOS) [152], ou encore l'ablation laser [153] afin d'obtenir une résolution spatiale de plus en plus fine tout en conservant un maximum d'informations sur l'identification des protéines. Cette méthode est bien évidemment compatible avec l'identification des AltProts. La mise en place de

méthodes permettant le pontage de protéines présentes dans le tissu permettrait d'obtenir des informations sur les variations d'interaction dans les régions d'intérêts, puis de comparer une micro-extraction sans modification à une micro-XL-MS de la même zone afin de confirmer les identifications d'interactions et les partenaires de ces protéines (**Figure 25**).

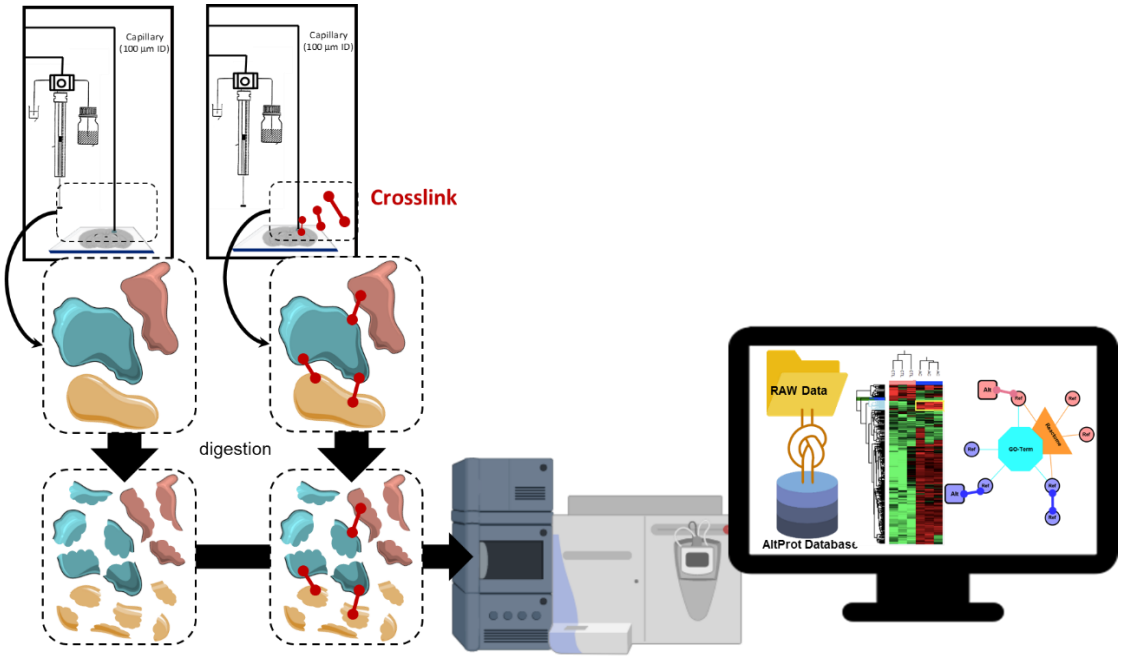


Figure 25 : ***Stratégie d'application de la méthode XL-MS couplée à la micro extraction de surface.*** *Stratégie décrivant l'application du pontage par méthode de micro dépôt et extraction de surface, en prévision d'une utilisation permettant d'identifier les partenaires protéiques ainsi que les modifications des réseaux à large échelle sur des coupes de tissus pathologiques.*

Ceci permettrait une cartographie moléculaire de la répartition des protéines dans des régions pathologiques en tenant compte des variations de la quantité de protéines, mais également des modifications touchant les réseaux d'interactions protéiques, tout en considérant les RefProts et les AltProts.

La dernière perspective porte sur l'application *in vivo* du pontage. En effet, le pontage chimique n'a pas encore fait ses preuves *in vivo* cependant l'idée a déjà émergée [128,154]. La réalisation d'une cartographie des interactions dynamiques dans la cellule au cours du temps après la prise d'une drogue ou à différents temps d'une stimulation permet de complétement changer la vision des résultats obtenus. Pendant très longtemps l'étude protéomique a été basée sur

la présence ou l'absence de protéine identifiée. Puis avec l'accès aux méthodes de quantification sans marquage (LFQ), des études de la variation d'expression ont été rendues possibles. J'espère qu'un jour, l'étude de l'interaction entre protéines soit aussi accessible afin d'enrichir l'information obtenue d'un échantillon.

Ces deux derniers développements, la mise en place de pontage sur tissus et *in vivo* seront les nouveaux axes de recherche de l'unité pour ces prochaines années.

Références

[1] F. Meier, P.E. Geyer, S. Virreira Winter, J. Cox, M. Mann, BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes, Nat. Methods, 15 (2018) 440–448.

[2] J. Crappé, E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, E. De Meester, T. De Meyer, W. Van Criekinge, P. Van Damme, G. Menschaert, PROTEOFORMER: Deep proteome coverage through ribosome profiling and MS integration, Nucleic Acids Res., 43 (2015) e29–e29.

[3] B. Vanderperre, J.-F. Lucier, C. Bissonnette, J. Motard, G. Tremblay, S. Vanderperre, M. Wisztorski, M. Salzet, F.-M. Boisvert, X. Roucou, H. Steen, M. Mann, D. Licatalosi, R. Darnell, R. Davuluri, Y. Suzuki, S. Sugano, C. Plass, T. Huang, T. Nilsen, et al., Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome, PLoS One, 8 (2013) e70698.

[4] B. Vanderperre, A.B. Staskevicius, G. Tremblay, M. McCoy, M.A. O'Neill, N.R. Cashman, X. Roucou, An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein, FASEB J., 25 (2011) 2373–2386.

[5] B. Vanderperre, J.F. Lucier, X. Roucou, HAltORF: Adatabase of predicted out-of-frame alternative open reading frames in human, Database, 2012 (2012) bas025.

[6] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, et al., UniProt: A hub for protein information, Nucleic Acids Res., 43 (2015) D204–D212.

[7] R.A. Harte, C.M. Farrell, J.E. Loveland, M.M. Suner, L. Wilming, B. Aken, D. Barrell, A. Frankish, C. Wallin, S. Searle, M. Diekhans, J. Harrow, K.D.

Pruitt, Tracking and coordinating an international curation effort for the CCDS Project, Database, 2012 (2012).

[8] J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J.G.R. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, R. Guigo, GENCODE: producing a reference annotation for ENCODE, Genome Biol., 7 Suppl 1 (2006) S4.

[9] B.L. Aken, P. Achuthan, W. Akanni, M.R. Amode, F. Bernsdorff, J. Bhai, K. Billis, D. Carvalho-silva, L. Gordon, C. Cummins, P. Clapham, L. Gil, C. Garc, T. Hourlier, S.E. Hunt, S.H. Janacek, T. Juettemann, S. Keenan, M.R. Laird, I. Lavidas, et al., Ensembl 2017 ıa Gir on, Nucleic Acids Res., 45 (2017) 1–8.

[10] N.T. Ingolia, Ribosome profiling: New views of translation, from single codons to genome scale, Nat. Rev. Genet., 15 (2014) 205–213.

[11] Y. Hao, L. Zhang, Y. Niu, T. Cai, J. Luo, S. He, B. Zhang, D. Zhang, Y. Qin, F. Yang, R. Chen, SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci, Brief. Bioinform., 19 (2018) 636–643.

[12] A. Saghatelian, J.P. Couso, Discovery and characterization of smORF-encoded bioactive polypeptides, Nat. Chem. Biol., 11 (2015) 909–16.

[13] E. Le Rhun, M. Duhamel, M. Wisztorski, J.P. Gimeno, F. Zairi, F. Escande, N. Reyns, F. Kobeissy, C.A. Maurage, M. Salzet, I. Fournier, Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification, Biochim. Biophys. Acta - Proteins Proteomics, 1865 (2017) 875–890.

[14] V. Delcourt, J. Franck, J. Quanico, J.P. Gimeno, M. Wisztorski, A. Raffo-Romero, F. Kobeissy, X. Roucou, M. Salzet, I. Fournier, Spatially-Resolved Top-down Proteomics Bridged to MALDI MS Imaging Reveals the Molecular Physiome of Brain Regions, Mol. Cell. Proteomics, 17 (2018) 357–372.

[15] V. Delcourt, J. Franck, E. Leblanc, F. Narducci, Y.M. Robin, J.P. Gimeno,

J. Quanico, M. Wisztorski, F. Kobeissy, J.F. Jacques, X. Roucou, M. Salzet, I. Fournier, Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer, EBioMedicine, 21 (2017) 55–64.

[16] V. Delcourt, M. Brunelle, A. V. Roy, J.-F. Jacques, M. Salzet, I. Fournier, X. Roucou, The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1, Mol. Cell. Proteomics, 17 (2018) 2402–2411.

[17] J. Baek, J. Lee, K. Yoon, H. Lee, Identification of unannotated small genes in Salmonella, G3 Genes, Genomes, Genet., 7 (2017) 983–989.

[18] M. Lluch-Senar, J. Delgado, W.-H. Chen, V. Llorens-Rico, F.J. O'Reilly, J.A. Wodke, E.B. Unal, E. Yus, S. Martinez, R.J. Nichols, T. Ferrar, A. Vivancos, A. Schmeisky, J. Stulke, V. van Noort, A.-C. Gavin, P. Bork, L. Serrano, Defining a minimal cell: essentiality of small ORFs and ncRNAs in a genome-reduced bacterium, Mol. Syst. Biol., 11 (2015) 780–780.

[19] M.R. Hemm, B.J. Paul, T.D. Schneider, G. Storz, K.E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models, Mol. Microbiol., 70 (2008) 1487–1501.

[20] C.S. Wadler, C.K. Vanderpool, A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide, Proc. Natl. Acad. Sci., 104 (2007) 20454–20459.

[21] J.P. Albuquerque, V. Tobias-Santos, A.C. Rodrigues, F.B. Mury, R.N. Da Fonseca, small ORFs: A new class of essential genes for development, Genet. Mol. Biol., 38 (2015) 278–283.

[22] K. Hanada, M. Higuchi-Takeuchi, M. Okamoto, T. Yoshizumi, M. Shimizu, K. Nakaminami, R. Nishi, C. Ohashi, K. Iida, M. Tanaka, Y. Horii, M. Kawashima, K. Matsui, T. Toyoda, K. Shinozaki, M. Seki, M. Matsui, Small open reading frames associated with morphogenesis are hidden in plant genomes, Proc. Natl. Acad. Sci., 110 (2013) 2395–2400.

[23] S. Samandi, A. V Roy, V. Delcourt, J.F. Lucier, J. Gagnon, M.C. Beaudoin,

B. Vanderperre, M.A. Breton, J. Motard, J.F. Jacques, M. Brunelle, I. Gagnon-Arsenault, I. Fournier, A. Ouangraoua, D.J. Hunting, A.A. Cohen, C.R. Landry, M.S. Scott, X. Roucou, Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins, Elife, 6 (2017) e27860.

[24] N.G. D'Lima, J. Ma, L. Winkler, Q. Chu, K.H. Loh, E.O. Corpuz, B.A. Budnik, J. Lykke-Andersen, A. Saghatelian, S.A. Slavoff, A human microprotein that interacts with the mRNA decapping complex, Nat. Chem. Biol., 13 (2017) 174–180.

[25] A. Matsumoto, A. Pasut, M. Matsumoto, R. Yamashita, J. Fung, E. Monteleone, A. Saghatelian, K.I. Nakayama, J.G. Clohessy, P.P. Pandolfi, mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide, Nature, 541 (2016) 228–232.

[26] E.G. Magny, J.I. Pueyo, F.M.G. Pearl, M.A. Cespedes, J.E. Niven, S.A. Bishop, J.P. Couso, Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames, Science (80-.)., 341 (2013) 1116–1120.

[27] Q. Zhang, A.A. Vashisht, J. O'Rourke, S.Y. Corbel, R. Moran, A. Romero, L. Miraglia, J. Zhang, E. Durrant, C. Schmedt, S.C. Sampath, S.C. Sampath, The microprotein Minion controls cell fusion and muscle formation, Nat. Commun., 8 (2017).

[28] J.-P. Couso, P. Patraquim, Classification and function of small open reading frames, Nat. Publ. Gr., 18 (2017) 575–589.

[29] R.J. Jackson, C.U.T. Hellen, T. V. Pestova, The mechanism of eukaryotic translation initiation and principles of its regulation, Nat. Rev. Mol. Cell Biol., 11 (2010) 113–127.

[30] L.A. Passmore, T.M. Schmeing, D. Maag, D.J. Applefield, M.G. Acker, M.A.A. Algire, J.R. Lorsch, V. Ramakrishnan, The Eukaryotic Translation Initiation Factors eIF1 and eIF1A Induce an Open Conformation of the 40S Ribosome, Mol. Cell, 26 (2007) 41–50.

[31] T. Von Der Haar, J.D. Gross, G. Wagner, J.E.G. McCarthy, The mRNA cap-binding protein eIF4E in post-transcriptional gene expression, Nat. Struct. Mol. Biol., 11 (2004) 503–511.

[32] T.F. Donahue, Genetic Approaches to Translation Initiation in Saccharomyces cerevisiae, in: Transl. Control Gene Expr., 2000: pp. 487–502.

[33] N. Sonenberg, A.G. Hinnebusch, Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets, Cell, 136 (2009) 731–745.

[34] M. Kozak, Initiation of translation in prokaryotes and eukaryotes, Gene, 234 (1999) 187–208.

[35] B.G. Luukkonen, W. Tan, S. Schwartz, Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance, J. Virol., 69 (1995) 4086–94.

[36] C. Vilela, B. Linz, C. Rodrigues-Pousada, J.E.G. Mccarthy, <yap1 uORF.pdf>, Nucleic Acids Res., 26 (1998) 1150–1159.

[37] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: Insights into functions, Nat. Rev. Genet., 10 (2009) 155–159.

[38] J.L. Rinn, M. Kertesz, J.K. Wang, S.L. Squazzo, X. Xu, S.A. Brugmann, L.H. Goodnough, J.A. Helms, P.J. Farnham, E. Segal, H.Y. Chang, Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs, Cell, 129 (2007) 1311–1323.

[39] X. Wang, S. Arai, X. Song, D. Reichart, K. Du, G. Pascual, P. Tempst, M.G. Rosenfeld, C.K. Glass, R. Kurokawa, Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription, Nature, 454 (2008) 126–130.

[40] J. Feng, C. Bi, B.S. Clark, R. Mady, P. Shah, J.D. Kohtz, The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator, Genes Dev., 20 (2006)

1470–1484.

[41] I. Martianov, A. Ramadass, A. Serra Barros, N. Chow, A. Akoulitchev, Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript, Nature, 445 (2007) 666–670.

[42] M. Beltran, I. Puig, C. Peña, J.M. García, A.B. Álvarez, R. Peña, F. Bonilla, A.G. De Herreros, A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition, Genes Dev., 22 (2008) 756–769.

[43] N.T.T. Ingolia, G.A.A. Brar, N. Stern-Ginossar, M.S.S. Harris, G.J.S.J.S. Talhouarne, S.E.E. Jackson, M.R.R. Wills, J.S.S. Weissman, Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes, Cell Rep., 8 (2014) 1365–1379.

[44] K. Verheggen, P.J. Volders, P. Mestdagh, G. Menschaert, P. Van Damme, K. Gevaert, L. Martens, J. Vandesompele, Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products, J. Proteome Res., 16 (2017) 2508–2515.

[45] S.A. Slavoff, A.J. Mitchell, A.G. Schwaid, M.N. Cabili, J. Ma, J.Z. Levin, A.D. Karger, B.A. Budnik, J.L. Rinn, A. Saghatelian, Peptidomic discovery of short open reading frame-encoded peptides in human cells, Nat. Chem. Biol., 9 (2013) 59–64.

[46] H. Mouilleron, V. Delcourt, X. Roucou, Death of a dogma: Eukaryotic mRNAs can code for more than one protein, Nucleic Acids Res., 44 (2016) 14–23.

[47] L. Breuza, S. Poux, A. Estreicher, M.L. Famiglietti, M. Magrane, M. Tognolli, A. Bridge, D. Baratin, N. Redaschi, The UniProtKB guide to the human proteome, Database, 2016 (2016) bav120.

[48] A. V. Kochetov, Alternative translation start sites and hidden coding potential of eukaryotic mRNAs, BioEssays, 30 (2008) 683–691.

[49] M. Oyama, H. Kozuka-Hata, Y. Suzuki, K. Semba, T. Yamamoto, S. Sugano, Diversity of transplantation start sites may define increased

complexity of the human short ORFeome, Mol. Cell. Proteomics, 6 (2007) 1000–1006.

[50] F. Wikipedia, RefSeq RefSeq categories, Nucleic Acids Res., 33 (2015) 5–7.

[51] M.A. Brunet, M. Brunelle, J.F. Lucier, V. Delcourt, M. Levesque, F. Grenier, S. Samandi, S. Leblanc, J.D. Aguilar, P. Dufour, J.F. Jacques, I. Fournier, A. Ouangraoua, M.S. Scott, F.M. Boisvert, X. Roucou, OpenProt: A more comprehensive guide to explore eukaryotic coding potential and proteomes, Nucleic Acids Res., 47 (2019) D403–D410.

[52] A.I. Nesvizhskii, Proteogenomics: Concepts, applications and computational strategies, Nat. Methods, 11 (2014) 1114–1125.

[53] J. Armengaud, Proteogenomics and systems biology: Quest for the ultimate missing parts, Expert Rev. Proteomics, 7 (2010) 65–77.

[54] H.D. Shukla, J. Mahmood, Z. Vujaskovic, Integrated proteo-genomic approach for early diagnosis and prognosis of cancer, Cancer Lett., 369 (2015) 28–36.

[55] G.S. Omenn, The strategy, organization, and progress of the HUPO Human Proteome Project, J. Proteomics, 100 (2014) 3–7.

[56] M.L. Fournier, J.M. Gilmore, S.A. Martin-Brown, M.P. Washburn, Multidimensional separations-based shotgun proteomics, Chem. Rev., 107 (2007) 3654–3686.

[57] N.L. Kelleher, Peer Reviewed: Top-Down Proteomics, Anal. Chem., 76 (2004) 196 A-203 A.

[58] R. R. Julian, The Mechanism Behind Top-Down UVPD Experiments: Making Sense of Apparent Contradictions, J. Am. Soc. Mass Spectrom., 28 (2017) 1823–1826.

[59] L. V. Schaffer, R.J. Millikin, R.M. Miller, L.C. Anderson, R.T. Fellers, Y. Ge, N.L. Kelleher, R.D. LeDuc, X. Liu, S.H. Payne, L. Sun, P.M. Thomas, T. Tucholski, Z. Wang, S. Wu, Z. Wu, D. Yu, M.R. Shortreed, L.M. Smith, Identification and Quantification of Proteoforms by Mass Spectrometry,

Proteomics, 19 (2019) 1970085.

[60] M. Kolmogorov, X. Liu, P.A. Pevzner, SpectroGene: A Tool for Proteogenomic Annotations Using Top-Down Spectra, J. Proteome Res., 15 (2016) 144–151.

[61] J. Ma, J.K. Diedrich, I. Jungreis, C. Donaldson, J. Vaughan, M. Kellis, J.R. Yates, A. Saghatelian, A. Saghatelian, Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides, Anal. Chem., 88 (2016) 3967–3975.

[62] L. Cassidy, P.T. Kaulich, A. Tholey, Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes, J. Proteome Res., 18 (2019) 1725–1734.

[63] N.G. D'Lima, A. Khitun, A.D. Rosenbloom, P. Yuan, B.M. Gassaway, K.W. Barber, J. Rinehart, S.A. Slavoff, N.G. D'lima, A. Khitun, A.D. Rosenbloom, P. Yuan, B.M. Gassaway, K.W. Barber, J. Rinehart, S.A. Slavoff, Comparative proteomics enables identification of non-annotated cold shock proteins in E coli, J. Proteome Res., 16 (2017) 3722–3731.

[64] F.Y. Che, X. Zhang, I. Berezniuk, M. Callaway, J. Lim, L.D. Fricker, Optimization of neuropeptide extraction from the mouse hypothalamus, J. Proteome Res., 6 (2007) 4667–4676.

[65] F.L. Strand, Neuropeptides: General characteristics and neuropharmaceutical potential in treating CNS disorders, Prog. Drug Res., 61 (2003) 1–37.

[66] M. Hallberg, Neuropeptides: Metabolism to bioactive fragments and the pharmacology of their receptors, Med. Res. Rev., 35 (2015) 464–519.

[67] M. Wisztorski, A. Desmons, J. Quanico, B. Fatou, J.P. Gimeno, J. Franck, M. Salzet, I. Fournier, Spatially-resolved protein surface microsampling from tissue sections using liquid extraction surface analysis, Proteomics, 16 (2016) 1622–1632.

[68] J. Franck, J. Quanico, M. Wisztorski, R. Day, M. Salzet, I. Fournier,

Quantification-based mass spectrometry imaging of proteins by parafilm assisted microdissection, Anal. Chem., 85 (2013) 8127–34.

[69] D. Cortez, Y. Wang, J. Qin, S.J. Elledge, Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks, Science, 286 (1999) 1162–6.

[70] M.J. Arboleda, J.F. Lyons, F.F. Kabbinavar, M.R. Bray, B.E. Snow, R. Ayala, M. Danino, B.Y. Karlan, D.J. Slamon, Overexpression of AKT2/protein kinase Bβ leads to up-regulation of β1 integrins, increased invasion, and metastasis of human breast and ovarian cancer cells, Cancer Res., 63 (2003) 196–206.

[71] P. Simeone, M. Trerotola, J. Franck, T. Cardon, M. Marchisio, I. Fournier, M. Salzet, M. Maffia, D. Vergara, The multiverse nature of epithelial to mesenchymal transition, Semin. Cancer Biol., (2018).

[72] Y. Hashimoto, T. Niikura, H. Tajima, T. Yasukawa, H. Sudo, Y. Ito, Y. Kita, M. Kawasumi, K. Kouyama, M. Doyu, G. Sobue, T. Koide, S. Tsuji, J. Lang, K. Kurokawa, I. Nishimoto, A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta, Proc. Natl. Acad. Sci. U. S. A., 98 (2001) 6336–41.

[73] N. Fuku, H. Pareja-Galeano, H. Zempo, R. Alis, Y. Arai, A. Lucia, N. Hirose, The mitochondrial-derived peptide MOTS-c: A player in exceptional longevity?, Aging Cell, 14 (2015) 921–923.

[74] B. Razooky, B. Obermayer, J. O'May, A. Tarakhovsky, Viral Infection Identifies Micropeptides Differentially Regulated in smORF-Containing lncRNAs, Genes (Basel)., 8 (2017) 206.

[75] S. Letovsky, S. Kasif, Predicting protein function from protein/protein interaction data: A probabilistic approach, in: Bioinformatics, Narnia, 2003: pp. i197–i204.

[76] D.I. Kim, S.C. Jensen, K.A. Noble, B. KC, K.J.K.H. Roux, K. Motamedchaboki, K.J.K.H. Roux, An improved smaller biotin ligase for BioID proximity labeling, Mol. Biol. Cell, 27 (2016) 1188–1196.

[77] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: A software Environment for integrated models of biomolecular interaction networks, Genome Res., 13 (2003) 2498–2504.

[78] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.H. Fridman, F. Pagès, Z. Trajanoski, J. Galon, ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, Bioinformatics, 25 (2009) 1091–1093.

[79] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K.P. Tsafou, M. Kuhn, P. Bork, L.J. Jensen, C. Von Mering, STRING v10: Protein-protein interaction networks, integrated over the tree of life, Nucleic Acids Res., 43 (2015) D447–D452.

[80] D. Miura, Y. Fujimura, H. Wariishi, In situ metabolomic mass spectrometry imaging: recent advances and difficulties, J. Proteomics, 75 (2012) 5052–60.

[81] A.K. Kenworthy, Imaging protein-protein interactions using fluorescence resonance energy transfer microscopy, Methods, 24 (2001) 289–296.

[82] J.H. Morris, G.M. Knudsen, E. Verschueren, J.R. Johnson, P. Cimermancic, A.L. Greninger, A.R. Pico, Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions, Nat. Protoc., 9 (2014) 2539–2554.

[83] S. Fields, O.K. Song, A novel genetic system to detect protein-protein interactions, Nature, 340 (1989) 245–246.

[84] A. Brückner, C. Polge, N. Lentze, D. Auerbach, U. Schlattner, Yeast two-hybrid, a powerful tool for systems biology, Int. J. Mol. Sci., 10 (2009) 2763–2788.

[85] S. Eyckerman, A. Verhee, J. Van der Heyden, I. Lemmens, X. Van Ostade, J. Vandekerckhove, J. Tavernier, Design and application of a cytokine-receptor-based interaction trap, Nat. Cell Biol., 3 (2001) 1114–1119.

[86] O. Söderberg, M. Gullberg, M. Jarvius, K. Ridderstråle, K.J. Leuchowius, J. Jarvius, K. Wester, P. Hydbring, F. Bahram, L.G. Larsson, U. Landegren, Direct observation of individual endogenous protein complexes in situ by proximity ligation, Nat. Methods, 3 (2006) 995–1000.

[87] O. Söderberg, K.J. Leuchowius, M. Gullberg, M. Jarvius, I. Weibrecht, L.G. Larsson, U. Landegren, Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay, Methods, 45 (2008) 227–232.

[88] T. Berggård, S. Linse, P. James, Methods for the detection and analysis of protein-protein interactions, Proteomics, 7 (2007) 2833–2842.

[89] O. Puig, F. Caspary, G. Rigaut, B. Rutz, E. Bouveret, E. Bragado-Nilsson, M. Wilm, B. Séraphin, The tandem affinity purification (TAP) method: A general procedure of protein complex purification, Methods, 24 (2001) 218–229.

[90] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, B. Seraphin, B. Séraphin, A generic protein purification method for protein complex characterization and proteome exploration, Nat. Biotechnol., 17 (1999) 1030–1032.

[91] T.P. Hopp, K.S. Prickett, V.L. Price, R.T. Libby, C.J. March, D.P. Cerretti, D.L. Urdal, P.J. Conlon, A short polypeptide marker sequence useful for recombinant protein identification and purification, Bio/Technology, 6 (1988) 1204–1210.

[92] C.L. Young, Z.T. Britton, A.S. Robinson, Recombinant protein expression and purification: A comprehensive review of affinity tags and microbial applications, Biotechnol. J., 7 (2012) 620–634.

[93] Z. Wang, U. Kim, Y. Jiao, C. Li, Y. Guo, X. Ma, M. Jiang, Z. Jiang, Y. Hou, G. Bai, Quantitative Proteomics Combined with Affinity MS Revealed the Molecular Mechanism of Ginsenoside Antitumor Effects, J. Proteome Res., 18 (2019) 2100–2108.

[94] K.J. Roux, D.I. Kim, M. Raida, B. Burke, A promiscuous biotin ligase fusion

protein identifies proximal and interacting proteins in mammalian cells, J. Cell Biol., 196 (2012) 801–10.

[95] J.D. Martell, T.J. Deerinck, Y. Sancak, T.L. Poulos, V.K. Mootha, G.E. Sosinsky, M.H. Ellisman, A.Y. Ting, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy, Nat. Biotechnol., 30 (2012) 1143–1148.

[96] L. Trinkle-Mulcahy, Recent advances in proximity-based labeling methods for interactome mapping [version 1; referees: 2 approved], F1000Research, 8 (2019).

[97] T.C. Branon, J.A. Bosch, A.D. Sanchez, N.D. Udeshi, T. Svinkina, S.A. Carr, J.L. Feldman, N. Perrimon, A.Y. Ting, Efficient proximity labeling in living cells and organisms with TurboID, Nat. Biotechnol., 36 (2018) 880–898.

[98] J. Jing, L. He, A. Sun, A. Quintana, Y. Ding, G. Ma, P. Tan, X. Liang, X. Zheng, L. Chen, X. Shi, S.L. Zhang, L. Zhong, Y. Huang, M.Q. Dong, C.L. Walker, P.G. Hogan, Y. Wang, Y. Zhou, Proteomic mapping of ER-PM junctions identifies STIMATE as a regulator of Ca 2+ influx, Nat. Cell Biol., 17 (2015) 1339–1347.

[99] B.M. Peterlin, P.A. Luciw, Replication of the human immunodeficiency virus: Sirategies for inhibition, Bio/Technology, 6 (1988) 794–799.

[100] S. Eyckerman, K. Titeca, E. Van Quickelberghe, E. Cloots, A. Verhee, N. Samyn, L. De Ceuninck, E. Timmerman, D. De Sutter, S. Lievens, S. Van Calenbergh, K. Gevaert, J. Tavernier, Trapping mammalian protein complexes in viral particles, Nat. Commun., 7 (2016) 11416.

[101] A. Leitner, M. Faini, F. Stengel, R. Aebersold, Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines, Trends Biochem. Sci., 41 (2016) 20–32.

[102] M.A. Lauber, J.P. Reilly, Structural analysis of a prokaryotic ribosome using a novel amidinating cross-linker and mass spectrometry, J. Proteome Res., 10 (2011) 3604–3616.

[103] C. Iacobucci, M. Götze, C. Piotrowski, C. Arlt, A. Rehkamp, C. Ihling, C. Hage, A. Sinz, Carboxyl-Photo-Reactive MS-Cleavable Cross-Linkers: Unveiling a Hidden Aspect of Diazirine-Based Reagents, Anal. Chem., 90 (2018) 2805–2809.

[104] C. Iacobucci, C. Piotrowski, A. Rehkamp, C.H. Ihling, A. Sinz, The First MS-Cleavable, Photo-Thiol-Reactive Cross-Linker for Protein Structural Studies, J. Am. Soc. Mass Spectrom., 30 (2019) 139–148.

[105] E.D. Merkley, S. Rysavy, A. Kahraman, R.P. Hafen, V. Daggett, J.N. Adkins, Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine-lysine distances, Protein Sci., 23 (2014) 747–759.

[106] A. Kao, C. Chiu, D. Vellucci, Y. Yang, V.R. Patel, S. Guan, A. Randall, P. Baldi, S.D. Rychnovsky, L. Huang, Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes, Mol. Cell. Proteomics, 10 (2011) M110.002212.

[107] M.Q. Mü, F. Dreiocker, C.H. Ihling, M. Schä, A. Sinz, M.Q. Müller, F. Dreiocker, C.H. Ihling, M. Schäfer, A. Sinz, Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS, Anal. Chem., 82 (2010) 6958–6968.

[108] X. Tang, J.E. Bruce, A new cross-linking strategy: Protein interaction reporter (PIR) technology for protein-protein interaction studies, Mol. Biosyst., 6 (2010) 939–947.

[109] C. Schmidt, C. V Robinson, A comparative cross-linking strategy to probe conformational changes in protein complexes, Nat. Protoc., 9 (2014) 2224–2236.

[110] X. Zhong, A.T. Navare, J.D. Chavez, J.K. Eng, D.K. Schweppe, J.E. Bruce, Large-Scale and Targeted Quantitative Cross-Linking MS Using Isotope-Labeled Protein Interaction Reporter (PIR) Cross-Linkers, J. Proteome Res., 16 (2017) 720–727.

[111] J.D. Chavez, J.E. Bruce, Chemical cross-linking with mass spectrometry: a

tool for systems structural biology, Curr. Opin. Chem. Biol., 48 (2019) 8–18.

[112] C. Nury, V. Redeker, S. Dautrey, A. Romieu, G. Van Der Rest, P.Y. Renard, R. Melki, J. Chamot-Rooke, A novel bio-orthogonal cross-linker for improved protein/protein interaction analysis, Anal. Chem., 87 (2015) 1853–1860.

[113] M. Rey, M. Dupré, I. Lopez-Neira, M. Duchateau, J. Chamot-Rooke, EXL-MS: An Enhanced Cross-Linking Mass Spectrometry Workflow to Study Protein Complexes, Anal. Chem., 90 (2018) 10707–10714.

[114] H.M. Barysz, J. Malmström, Development of Large-scale Cross-linking Mass Spectrometry, Mol. Cell. Proteomics, 17 (2018) 1055–1066.

[115] B.A. Steigenberger, R.J. Pieters, A.J.R. Heck, R.A. Scheltema, PhoX - an IMAC-enrichable Crosslinking Reagent, BioRxiv Biochem., (2019) 556688.

[116] R. Fritzsche, C.H. Ihling, M. Götze, A. Sinz, Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis, Rapid Commun. Mass Spectrom., 26 (2012) 653–658.

[117] F. Liu, D.T.S.S. Rijkers, H. Post, A.J.R.R. Heck, Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry, Nat. Methods, 12 (2015) 1179–1184.

[118] O. Klykov, B. Steigenberger, S. Pektaş, D. Fasci, A.J.R. Heck, R.A. Scheltema, Efficient and robust proteome-wide approaches for cross-linking mass spectrometry, Nat. Protoc., (n.d.).

[119] R. Schmidt, A. Sinz, Improved single-step enrichment methods of cross-linked products for protein structure analysis and protein interaction mapping, Anal. Bioanal. Chem., 409 (2017) 2393–2400.

[120] L. De Jong, E.A. De Koning, W. Roseboom, H. Buncherd, M.J. Wanner, I. Dapic, P.J. Jansen, J.H. Van Maarseveen, G.L. Corthals, P.J. Lewis, L.W. Hamoen, C.G. De Koster, In-Culture Cross-Linking of Bacterial Cells Reveals Large-Scale Dynamic Protein-Protein Interactions at the Peptide Level, J. Proteome Res., 16 (2017) 2457–2471.

[121] M.A. Lauber, J.P. Reilly, Novel amidinating cross-linker for facilitating analyses of protein structures and interactions, Anal. Chem., 82 (2010) 7736–7743.

[122] S.H. Giese, L. Fischer, J. Rappsilber, A Study into the Collision-induced Dissociation (CID) Behavior of Cross-Linked Peptides, Mol. Cell. Proteomics, 15 (2016) 1094–1104.

[123] F. Liu, P. Lössl, R. Scheltema, R. Viner, A.J.R. Heck, Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification, Nat. Commun., 8 (2017) 15473.

[124] A. Leitner, T. Walzthoeni, R. Aebersold, Lysine-specific chemical cross-linking of protein complexes and identification of cross-linking sites using LC-MS/MS and the xQuest/xProphet software pipeline, Nat. Protoc., 9 (2014) 120–137.

[125] S.B. Fan, J.M. Meng, S. Lu, K. Zhang, H. Yang, H. Chi, R.X. Sun, M.Q. Dong, S.M. He, Using pLink to analyze cross-linked peptides, Curr. Protoc. Bioinforma., 2015 (2015) 8.21.1-8.21.19.

[126] M. Götze, J. Pettelkau, S. Schaks, K. Bosse, C.H. Ihling, F. Krauth, R. Fritzsche, U. Kühn, A. Sinz, StavroX-A software for analyzing crosslinked products in protein interaction studies, J. Am. Soc. Mass Spectrom., 23 (2012) 76–87.

[127] M. Götze, J. Pettelkau, R. Fritzsche, C.H. Ihling, M. Schäfer, A. Sinz, Automated assignment of MS/MS cleavable cross-links in protein 3d-structure analysis, J. Am. Soc. Mass Spectrom., 26 (2014) 83–97.

[128] J.D. Chavez, C.F. Lee, A. Caudal, A. Keller, R. Tian, J.E. Bruce, Chemical Crosslinking Mass Spectrometry Analysis of Protein Conformations and Supercomplexes in Heart Tissue, Cell Syst., 6 (2018) 136-141.e5.

[129] A.G. Schwaid, D.A. Shannon, J. Ma, S.A. Slavoff, J.Z. Levin, E. Weerapana, A. Saghatelian, Chemoproteomic discovery of cysteine-containing human short open reading frames, J. Am. Chem. Soc., 135 (2013) 16750–16753.

[130] J.A. Vizcaíno, R.G. Côté, A. Csordas, J.A. Dianes, A. Fabregat, J.M. Foster, J. Griss, E. Alpi, M. Birim, J. Contell, G. O'Kelly, A. Schoenegger, D. Ovelleiro, Y. Pérez-Riverol, F. Reisinger, D. Ríos, R. Wang, H. Hermjakob, The Proteomics Identifications (PRIDE) database and associated tools: Status in 2013, Nucleic Acids Res., 41 (2013) D1063-9.

[131] R. Craig, J.P. Cortens, R.C. Beavis, Open source system for analyzing, validating, and storing protein identification data, J. Proteome Res., 3 (2004) 1234–1242.

[132] F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S.N. Loevenich, R. Aebersold, The PeptideAtlas project, Nucleic Acids Res., 34 (2006) D655-8.

[133] M. Vaudel, K. Verheggen, A. Csordas, H. Ræder, F.S. Berven, L. Martens, J.A. Vizcaíno, H. Barsnes, Exploring the potential of public proteomics data, Proteomics, 16 (2016) 214–225.

[134] H.M. Berman, The Protein Data Bank / Biopython, Presentation, 28 (2000) 235–242.

[135] Y. Zhang, I-TASSER server for protein 3D structure prediction, BMC Bioinformatics, 9 (2008) 40.

[136] S.R. Eddy, A new generation of homology search tools based on probabilistic inference, Genome Inform., 23 (2009) 205–11.

[137] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol., 215 (1990) 403–410.

[138] C.J.A. Sigrist, E. De Castro, L. Cerutti, B.A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Res., 41 (2013) D344.

[139] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. Finn, The Pfam protein families database, Nucleic Acids Res., 40 (2012) D290-301.

[140] P. Jones, D. Binns, H.Y. Chang, M. Fraser, W. Li, C. McAnulla, H.

McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.Y. Yong, R. Lopez, S. Hunter, InterProScan 5: Genome-scale protein function classification, Bioinformatics, 30 (2014) 1236–1240.

[141] D. Xu, R. Nussinov, Favorable domain size in proteins, Fold. Des., 3 (1998) 11–17.

[142] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Mol. Syst. Biol., 3 (2007) 1–13.

[143] W. Zhu, J. Hou, Y.P. Phoebe Chen, Semantic and layered protein function prediction from PPI networks, J. Theor. Biol., 267 (2010) 129–136.

[144] H. Li, L. Xiao, L. Zhang, J. Wu, B. Wei, N. Sun, Y. Zhao, FSPP: A tool for genome-wide prediction of smORF-encoded peptides and their functions, Front. Genet., 9 (2018) 96.

[145] M. Allen, M. Bjerke, H. Edlund, S. Nelander, B. Westermark, Origin of the U87MG glioma cell line: Good news and bad news, Sci. Transl. Med., 8 (2016) 354re3.

[146] G. Menschaert, W. Van Criekinge, T. Notelaers, A. Koch, J. Crappé, K. Gevaert, P. Van Damme, 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, Mol. Cell. Proteomics, 12 (2013) 1780–1790.

[147] F. Xing, Y. Luan, J. Cai, S. Wu, J. Mai, J. Gu, H. Zhang, K. Li, Y. Lin, X. Xiao, J. Liang, Y. Li, W. Chen, Y. Tan, L. Sheng, B. Lu, W. Lu, M. Gao, P. Qiu, X. Su, et al., The Anti-Warburg Effect Elicited by the cAMP-PGC1α Pathway Drives Differentiation of Glioblastoma Cells into Astrocytes, Cell Rep., 18 (2017) 468–481.

[148] T. Cardon, M. Salzet, J. Franck, I. Fournier, Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation, Biochim. Biophys. Acta - Gen. Subj., (2019).

[149] S. Devaux, D. Cizkova, J. Quanico, J. Franck, S. Nataf, L. Pays, L.

Hauberg-Lotte, P. Maass, J.H. Kobarg, F. Kobeissy, C. Mériaux, M. Wisztorski, L. Slovinska, J. Blasko, V. Cigankova, I. Fournier, M. Salzet, L. Hauberg-Lotteţ, P. Maassţ, J.H. Kobarg, et al., Proteomic Analysis of the Spatio-temporal Based Molecular Kinetics of Acute Spinal Cord Injury Identifies a Time- and Segment-specific Window for Effective Tissue Repair, Mol. Cell. Proteomics, 15 (2016) 2641–70.

[150] S. Devaux, D. Cizkova, K. Mallah, M.A. Karnoub, Z. Laouby, F. Kobeissy, J. Blasko, S. Nataf, L. Pays, C. Meriaux, I. Fournier, M. Salzet, RhoA inhibitor treatment at acute phase of spinal cord injury may induce neurite outgrowth and synaptogenesis, Mol. Cell. Proteomics, 16 (2017) 1394–1415.

[151] J. Quanico, J. Franck, J.P. Gimeno, R. Sabbagh, M. Salzet, R. Day, I. Fournier, Parafilm-assisted microdissection: a sampling method for mass spectrometry-based identification of differentially expressed prostate cancer protein biomarkers, Chem. Commun. (Camb)., 51 (2014) 4564–7.

[152] J. Quanico, J. Franck, T. Cardon, E. Leblanc, M. Wisztorski, M. Salzet, I. Fournier, NanoLC-MS coupling of liquid microjunction microextraction for on-tissue proteomic analysis, Biochim. Biophys. Acta - Proteins Proteomics, 1865 (2017) 891–900.

[153] F. Donnarumma, K.K. Murray, Laser ablation sample transfer for localized LC-MS/MS proteomic analysis of tissue, J. Mass Spectrom., 51 (2016) 261–8.

[154] A. Sinz, Crosslinking Mass Spectrometry Goes In-Tissue, Cell Syst., 6 (2018) 10–12.

Droits des Figures:

Figure 2:

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Aug 21, 2019

This Agreement between Université science et technologie de Lille -- Tristan Cardon ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4652941166526
License date	Aug 20, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	The mechanism of eukaryotic translation initiation and principles of its regulation
Licensed Content Author	Richard J. Jackson, Christopher U. T. Hellen, Tatyana V. Pestova
Licensed Content Date	Feb 1, 2010
Licensed Content Volume	11
Licensed Content Issue	2
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Relations, Interactions et Fonctions des Protéines Alternatives
Institution name	Université Sciences et Technologies de Lille, INSERM
Expected presentation date	Aug 2019
Order reference number	figure 1
Portions	Figure 1 Model of the canonical pathway of eukaryotic translation initiation

Requestor Location	Université science et technologie de Lille Bât SN3, 1er étage, porte 109 Villeneuve d'ascq, haut de france 59655 France Attn: Université science et technologie de Lille
Total	0.00 EUR

Figure 4:

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Aug 21, 2019

This Agreement between Université science et technologie de Lille -- Tristan Cardon ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4652950109315
License date	Aug 20, 2019
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Long non-coding RNAs: insights into functions
Licensed Content Author	Tim R. Mercer, Marcel E. Dinger, John S. Mattick
Licensed Content Date	Mar 1, 2009
Licensed Content Volume	10
Licensed Content Issue	3
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	Relations, Interactions et Fonctions des Protéines Alternatives
Institution name	Université Sciences et Technologies de Lille, INSERM
Expected presentation date	Aug 2019
Order reference number	2
Portions	Figure 2 Functions of long non-coding RnAs (ncRnAs)
Requestor Location	Université science et technologie de Lille Bât SN3, 1er étage, porte 109 Villeneuve d'ascq, haut de france 59655

	France Attn: Université science et technologie de Lille
Total	0.00 EUR

Figure 6:

Figure 7:

ELSEVIER LICENSE
TERMS AND CONDITIONS

Aug 21, 2019

This Agreement between Université science et technologie de Lille -- Tristan Cardon ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4652950470906
License date	Aug 20, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	Methods
Licensed Content Title	Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay
Licensed Content Author	Ola Söderberg,Karl-Johan Leuchowius,Mats Gullberg,Malin Jarvius,Irene Weibrecht,Lars-Gunnar Larsson,Ulf Landegren
Licensed Content Date	Jul 1, 2008
Licensed Content Volume	45
Licensed Content Issue	3
Licensed Content Pages	6
Start Page	227
End Page	232
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	2.A
Original figure numbers	Fig. 2. (A) schematic presentation of in situ PLA. Dual binding by a pair of proximity probes (antibodies with attached DNA strands)
Title of your thesis/dissertation	Relations, Interactions et Fonctions des Protéines Alternatives
Publisher of new work	Université Sciences et Technologies de Lille, INSERM

Expected completion date	Aug 2019
Estimated size (number of pages)	1
Requestor Location	Université science et technologie de Lille Bât SN3, 1er étage, porte 109 Villeneuve d'ascq, haut de france 59655 France Attn: Université science et technologie de Lille
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR

Figure 8:

ELSEVIER LICENSE
TERMS AND CONDITIONS

Aug 21, 2019

This Agreement between Université science et technologie de Lille -- Tristan Cardon ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4652950868481
License date	Aug 20, 2019
Licensed Content Publisher	Elsevier
Licensed Content Publication	Methods
Licensed Content Title	The Tandem Affinity Purification (TAP) Method: A General Procedure of Protein Complex Purification
Licensed Content Author	Oscar Puig,Friederike Caspary,Guillaume Rigaut,Berthold Rutz,Emmanuelle Bouveret,Elisabeth Bragado-Nilsson,Matthias Wilm,Bertrand Séraphin
Licensed Content Date	Jul 1, 2001
Licensed Content Volume	24
Licensed Content Issue	3
Licensed Content Pages	12
Start Page	218
End Page	229
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	1.B
Original figure numbers	Figure1 (B) Overview of the TAP purification strategy
Title of your thesis/dissertation	Relations, Interactions et Fonctions des Protéines Alternatives
Publisher of new work	Université Sciences et Technologies de Lille, INSERM
Expected completion date	Aug 2019

Estimated size (number of pages)	1
Requestor Location	Université science et technologie de Lille Bât SN3, 1er étage, porte 109 Villeneuve d'ascq, haut de france 59655 France Attn: Université science et technologie de Lille
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR

Figure 13:

- **Order detail ID:**71985293
- **ISSN:**1535-9484
- **Publication Type:**e-Journal
- **Volume:**
- **Issue:**
- **Start page:**
- **Publisher:**AMERICAN SOCIETY FOR BIOCHEMISTRY AND MOLECULAR BIOLOGY
- **Author/Editor:**American Society for Biochemistry and Molecular Biology
- **Permission Status: Granted**

- **Permission type:**Republish or display content
- **Type of use:**Thesis/Dissertation
-

Order License Id: 4652960364378

Order ref number: figure 13

- Hide details
-

Requestor type	Academic institution
Format	Print, Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
The requesting person/organization	Tristan CARDON
Title or numeric reference of the portion(s)	FIG. 3. A,B,C
Title of the article or chapter the portion is from	A study into the CID behavior of cross-linked peptides
Editor of portion(s)	Molecular & Cellular Proteomics
Author of portion(s)	Sven H. Giese, Lutz Fischer and Juri Rappsilber
Volume of serial or monograph	vol. 15
Page range of portion	p. 1094-1104.
Publication date of portion	December 30, 2015
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	Relations, Interactions et Fonctions des Protéines Alternatives

Institution name	Université Sciences et Technologies de Lille, INSERM
Expected presentation date	Aug 2019
Order reference number	figure 13

Figure 17 :

JOHN WILEY AND SONS LICENSE TERMS AND CONDITIONS

Aug 21, 2019

This Agreement between Université science et technologie de Lille -- Tristan Cardon ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	4652960884179
License date	Aug 20, 2019
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	Molecular Systems Biology
Licensed Content Title	Network-based prediction of protein function
Licensed Content Author	Ron Shamir, Igor Ulitsky, Roded Sharan
Licensed Content Date	Mar 13, 2007
Licensed Content Volume	3
Licensed Content Issue	1
Licensed Content Pages	13
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Original Wiley figure/table number(s)	Figure 2
Will you be translating?	No
Title of your thesis / dissertation	Relations, Interactions et Fonctions des Protéines Alternatives
Expected completion date	Aug 2019
Expected size (number of pages)	1
Requestor Location	Université science et technologie de Lille Bât SN3, 1er étage, porte 109 Villeneuve d'ascq, haut de france 59655 France Attn: Université science et technologie de Lille

Publisher Tax ID	EU826007151
Total	0.00 EUR

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.
- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.
- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,**You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.
- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto
- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.
- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.
- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.
- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.
- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.
- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.
- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.

- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.
- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.
- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.
- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.
Further details can be found on Wiley Online Library http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions: v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Résumé :

Si en transcriptomique le dogme accepté par la communauté veut qu'un ARNm code pour une protéine unique, la protéomique vient de montrer l'inverse. Force est de constater que les ARNm peuvent traduire plusieurs protéines. Celles ne suivant pas le cadre de référence sont appelées protéines alternatives (AltProts) et forment le protéome caché ou fantôme. Ces AltProts nécessitent la mise en place de stratégies adaptées pour leur mise en évidence. Leurs caractéristiques physicochimiques spécifiques, telles que leur petite taille permet d'adapter les méthodes classiques de protéomique à leur étude. Dans cet objectif la mise en évidence des AltProts par différentes méthodes d'extraction, notamment adaptées des méthodes de peptidomique, a permis de mettre en évidence les conditions d'enrichissement avant une analyse *bottom-up.* Ces AltProts sont une nouvelle classe de protéines pour laquelle très peu d'informations fonctionnelles sont connues. Les prédictions de fonction avancées lors des premières constructions de bases de données, annonçaient des fonctions dans la régulation des ARN, de la synthèse de protéines et de la régulation d'expression des gènes par association avec des facteurs de transcription. Ces prédictions étaient basées sur les homologies de séquences entre les AltProts et les protéines de référence (RefProts). Cependant très peu d'études montrent le rôle de ces protéines de manière expérimentale. Afin de mettre en évidence les fonctions de ces AltProts, nous avons choisi de retrouver leurs partenaires d'interaction. À l'heure actuelle, plusieurs méthodes existent permettant d'étudier l'interactome des protéines, toutefois la majorité est dirigée vers une cible, nécessitant parfois des constructions biochimiques ou l'utilisation d'anticorps dirigés, rendant ces méthodes difficiles à mettre en place pour les AltProts. Seule la méthode de pontage chimique couplée à la spectrométrie de masse (XL-MS) permet d'observer des interactions cellulaires de manière non ciblée. Cette méthode de pontage chimique, bien que connaissant ses propres limitations, est applicable à la recherche des partenaires d'interaction des AltProts. Cet outil, associé aux logiciels de traitement des réseaux d'interaction, enrichi par les interactions connues entre RefProts dans la littérature, permet de replacer les AltProts dans ces réseaux. Ces réseaux, peuvent ensuite être traités afin de mettre en évidence les voies de signalisation impliquant les RefProts et ainsi déduire les différentes voies de signalisation associées aux AltProts observées pontées aux RefProts.

Mots clés : Protéine Alternative ; Protéomique ; Interaction Protéine-Protéine ; Spectrométrie de Masse ; Transcriptomique

Summary:

If in transcriptomics the dogma accepted by the community is that a single mRNA codes for a single protein, proteomics has just shown the opposite. It must be said that mRNAs can translate several proteins. These not following the reference framework are called alternative proteins (AltProts) and form the hidden or ghost proteome. These AltProts require the implementation of appropriate strategies to highlight them. Their specific physicochemical characteristics, such as their small size, make it possible to adapt classical proteomic methods to their study. With this objective in mind, the identification of AltProts by different extraction methods, particularly adapted to peptidomic methods, made it possible to highlight the enrichment conditions before a bottom-up analysis. These AltProts are a new class of proteins for which very little functional information is known. Advanced function predictions in the early database constructions announced functions in RNA regulation, protein synthesis and gene expression regulation by association with transcriptional factors. These predictions were based on sequence homologies between AltProts and reference proteins (RefProts). However, very few studies show the role of these proteins in an experimental way. In order to highlight the functions of these AltProts, we have chosen to find their interaction partners. At present, several methods exist to study the protein interactome, however the majority are directed towards a target, sometimes requiring biochemical constructs or the use of directed antibodies, making these methods difficult to implement for AltProts. Only the Crosslink method coupled with mass spectrometry (XL-MS) allows to observe cellular interactions in a non-targeted way. This chemical bridging method, although aware of its own limitations, is applicable to the search for AltProts interaction partners. This tool, combined with the software for processing interaction networks, enriched by the known interactions between RefProts in the literature, makes it possible to replace AltProts in these networks. These networks can then be processed to highlight the signaling pathways involving RefProts and thus deduce the different signaling pathways associated with the observed AltProts crosslinked to the RefProts.

Keywords : Alternative protein ; Proteomic ; Protein-Protein Interaction ; Mass Spectrometry ; Transcriptomic