Université de Sherbrooke

Les cadres ouverts de lecture alternatifs contribuent significativement au protéome des eucaryotes

Par
Benoît Vanderperre
Programme de Biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de *Philosophiae Doctor* (Ph.D.)
en Biochimie

Sherbrooke, Québec, Canada
Septembre 2013

Membres du jury d'évaluation :

Dr Xavier Roucou (Biochimie)
Dr François Bachand (Biochimie)
Dr Benoit Chabot (Microbiologie et Infectiologie)
Dr Benoit Coulombe (Département de Biochimie et médecine moléculaire, Faculté de Médecine, Université de Montréal)

Your file Votre référence
ISBN: 978-0-499-00428-4

Our file Notre référence
ISBN: 978-0-499-00428-4

A mes parents, Marie-Christine et Jean-Luc, pour m'avoir inculqué travail, humilité et altruisme comme valeurs inébranlables.

« C'est ce que nous pensons déjà connaître qui nous empêche souvent d'apprendre. »

Claude Bernard

RÉSUMÉ

Les cadres ouverts de lecture alternatifs contribuent significativement au protéome des eucaryotes

Par Benoît Vanderperre
Programme de Biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de *Philosophiae Doctor* (Ph.D.) en Biochimie

Un défi majeur de l'ère post-génomique est de définir l'ensemble des protéines encodées par le génome : le protéome. Un ARNm mature est en général associé à un seul cadre ouvert de lecture (ORF, *open reading frame* en anglais) de référence (RefORF) codant pour une protéine. Des ORFs alternatifs (AltORFs) sont cependant présents dans les régions non-traduites (UTRs), ou chevauchant le RefORF dans les cadres de lecture alternatifs +2 et +3. Les AltORFs offrent le potentiel d'augmenter la diversité protéique, mais leur réelle contribution au protéome est peu caractérisée.

Par des techniques de biologie moléculaire, de biochimie, et de biologie cellulaire, j'ai tout d'abord mis en évidence chez plusieurs mammifères supérieurs, l'expression endogène d'une protéine alternative appelée AltPrP à partir du gène *PRNP*. La découverte d'AltPrP devrait améliorer notre compréhension des rôles pathologiques et physiologiques de ce gène. Suite à la découverte d'AltPrP, et basé sur ce modèle d'AltORF chevauchant un RefORF, j'ai entrepris d'investiguer l'étendue de l'utilisation de ces AltORFs comme source de diversité protéique, chez l'humain en particulier. Par des méthodes computationnelles, j'ai participé à la création d'une base de données d'AltORFs prédits dans les ARNm humains (HAltORF, pour Human Alternative ORFs). HAltORF est consultable et interrogeable en ligne. Elle facilitera et accélérera la découverte et l'étude des AltORFs. J'ai ensuite mis au point une approche protéomique afin d'apporter des preuves expérimentales de l'utilisation à grande échelle des AltORFs. Une base de données d'AltORFs, mise à jour pour inclure ceux chevauchant en tout ou partie les UTRs, a été créée et utilisée pour déterminer par spectrométrie de masse la contribution des protéines alternatives au protéome humain. J'ai validé l'expression de 1259 protéines alternatives prédites à travers différents échantillons. J'ai aussi démontré que l'expression d'AltORFs impliquait d'importants biais dans les dessins expérimentaux (transfections ou criblage de banques d'ADNc par exemple). Enfin, un grand nombre de protéines alternatives semblent conservées à travers l'évolution, suggérant leur importance fonctionnelle.

En conclusion, mes travaux de doctorat ont permis de mettre en évidence que les AltORFs conduisent à l'expression de nouvelles protéines jusqu'alors ignorées. Ces résultats redéfinissent notre vision du protéome, remettent en question notre compréhension de la structure et de la fonction des gènes eucaryotes, et ouvrent la voie vers l'étude fonctionnelle des protéines alternatives.

Mots clés : initiation alternative de la traduction, cadres ouverts de lecture alternatifs, gènes multicodants, ARNm polycistroniques, protéome, spectrométrie de masse, bases de données protéiques, protéine prion.

TABLE DES MATIERES

LISTE DES FIGURES

INTRODUCTION

DISCUSSION

LISTE DES ABRÉVIATIONS

ORF	Cadre ouvert de lecture *open reading frame*
RefORF	Cadre ouvert de lecture de référence *Reference open reading frame*
AltORF	Cadre ouvert de lecture alternatif *Alternative open reading frame*
UTR	Région non-traduite *Untranslated region*
AltPrP	Protéine prion alternative *Alternative prion protein*
AA	Acides aminés *Amino acids*
CR-APA	Poly-adénylation alternative dans la région codante *Alternative poly-adenylation in the coding region*
UTR-APA	Poly-adénylation alternative dans les UTRs *Alternative poly-adenylation in the UTRs*
CDS	Séquence codante *Coding sequence*
Met-ARNt$_i^{Met}$	ARNt-Met initiateur *initiator Met-tRNA*
MFC	Complexe multi-factoriel *Multifactorial complex*
43S PIC	Complexe de pré-initiation 43S *43S pre-initiation complex*
SD	Séquence Shine-Dalgarno *Shine-Dalgarno sequence*
IRES	Site d'entrée interne des ribosomes *Internal ribosome entry site*
CITE	Facilitateur traductionnel coiffe-indépendant *Cap independent translational enhancer*
uORF	Cadre ouvert de lecture en amont *Upstream open reading frame*
PP	Peptide paralytique *Paralytic peptide*
nt	Nucléotides *Nucleotides*
CSE	Cellule souche embryonnaire *Embryonnic stem cell*
TIS	Site d'initiation de la traduction *Translation initiation site*
MS	Spectrométrie de masse *Mass spectrometry*
LC-MS/MS	Chromatographie liquide couplée à de la spectrométrie de masse en tandem *Liquid chromatography coupled tandem mass spectrometry*
RNA-SEQ	Séquencage d'ARN *RNA sequencing*

Y2H	Crible doubre hybride chez la levure *Yeast two hybrid screening*
ARNi	ARN interférent *Interfering RNA*
NMD	Dégradation des ARNms non-sens *Nonsense mRNA decay*
RE	Réticulum endoplasmique *Endoplasmic reticulum*
eGFP	Protéine fluorescente verte (améliorée) *(enhanced) Green fluorescent protein*
TSE	Encéphalopathie spongiforme transmissible *Transmissible spongiform encephalopathy*
PMCA	Amplification cyclique du mauvais repliement protéique *Protein misfolding cyclic amplification*

INTRODUCTION

1. Diversité protéique et mécanismes associés

1.1. Importance de la diversité protéique

Au cours de l'évolution, les organismes vivants ont développé une complexité de plus en plus grande, estimée par la complexité de l'architecture cellulaire, et par le nombre total de cellules et de types cellulaires différents qu'ils contiennent (Hedges et al, 2004, Vogel & Chothia, 2006). Dans l'ère post-génomique actuelle, l'un des objectifs majeurs est d'identifier les mécanismes ayant mené à cette augmentation de complexité, et d'établir leur contribution respective. L'un d'eux est l'augmentation de la diversité protéique (Schluter et al, 2009). Dans cette partie, je soulignerai l'importance des protéines dans les processus biologiques, et le fort intérêt scientifique qui en résulte. Ensuite, j'expliquerai comment la diversité protéique semble influencer la complexité des organismes vivants. Enfin, je discuterai de l'importance fonctionnelle des protéines de petite taille, dont l'étendue de leur contribution au protéome* (ensemble des protéines d'un organisme) commence seulement à être appréciée.

1.1.1. Les protéines comme acteurs majeurs du vivant

Les protéines forment un lien primordial entre l'information contenue dans les gènes et leur fonction biologique. Ce sont des acteurs moléculaires majeurs qui permettent la régulation d'une multitude de réactions chimiques nécessaires à la vie. Catalysant un nombre restreint de réactions au départ, les protéines se sont diversifiées au cours de l'évolution et constituent à présent un immense catalogue d'outils moléculaires extrêmement spécialisés. Elles peuvent cependant être regroupées en plusieurs classes en fonction des rôles qui leurs sont attribués (Figure 1)(Lodish et al, 2005).

*protéome : initialement défini comme l'ensemble des protéines encodées par un génome, le protéome est maintenant plutôt considéré comme l'ensemble des protéines exprimées dans un système biologique particulier (cellule, organisme, structure sub-cellulaire), dans un état particulier (développement, pathologie ...)

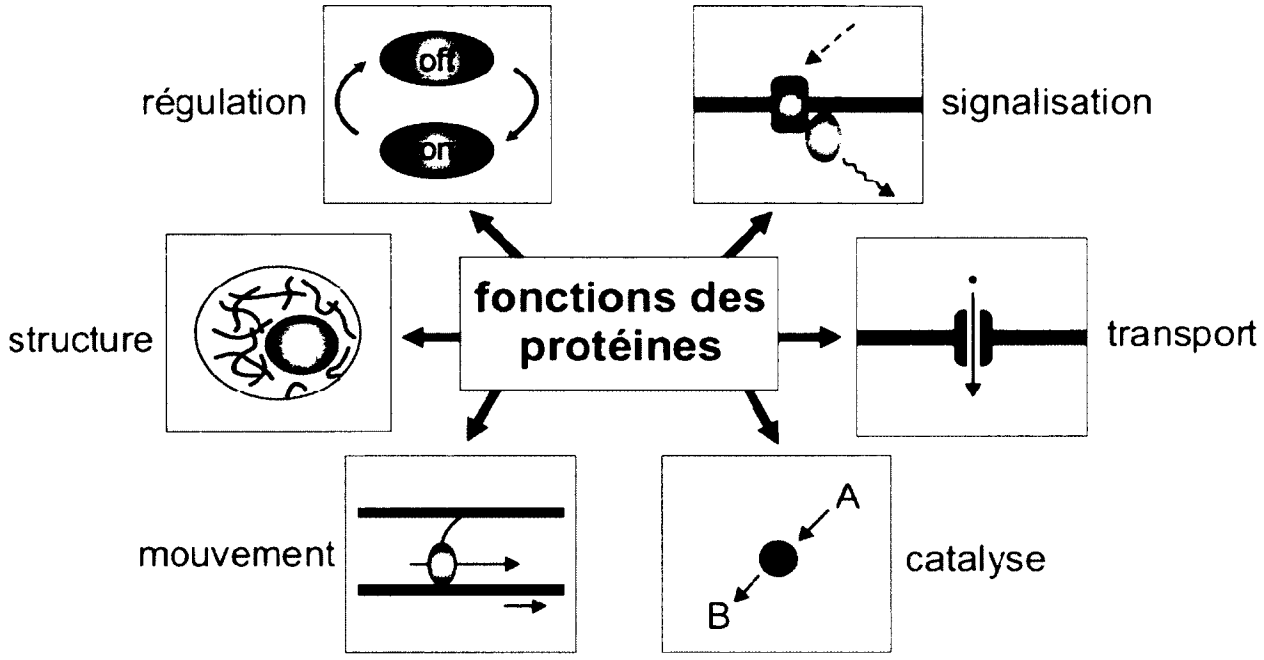


Figure 1. Importance fonctionnelle des protéines dans les processus biologiques. Adapté de (Lodish et al, 2005). Les protéines régulent les processus cellulaires nécessaires au fonctionnement d'une cellule et d'un organisme, tels que la réponse aux conditions environnementales (signalisation), le transport de composés entres compartiments, la catalyse de réactions chimiques, la génération de forces pour les mouvements, l'organisation structurale des cellules et de ses composants, et la régulation des activités d'acteurs moléculaires (protéines, ARNs).

Etant donné l'implication des protéines dans le bon fonctionnement des processus biologiques, des perturbations dans leur niveau d'expression, leur séquence, leur structure, ou dans tout autre caractéristique cruciale pour leur fonction en conditions physiologiques peut aboutir au développement de nombreuses pathologies. Ainsi, la caractérisation des protéines et de leurs fonctions a suscité un grand intérêt de la part des biologistes moléculaires et cellulaires, de manière à comprendre comment les grands processus biologiques sont effectués, et affectés dans le cadre de maladies. Conjointement, un effort particulier a été, et est toujours déployé afin de dresser une liste aussi exhaustive que possible des protéomes d'organismes de plus en plus nombreux.

1.1.2. La diversité protéique : un facteur de complexité des organismes vivants

Au vu de l'importance fonctionnelle des protéines dans les processus biologiques, un moyen instinctif de prédire la diversité protéique d'un organisme est d'évaluer le nombre de gènes codant pour des protéines dans son génome. Chez les procaryotes, la complexité biologique est bien corrélée au nombre de gènes codant pour des protéines (Giovannoni et al, 2005, Mira et al, 2001). Suite au séquençage des génomes de plusieurs eucaryotes, y compris *Homo sapiens*, une estimation du nombre de gènes codant pour des protéines a été effectuée. Face au nombre réduit de gènes codant pour des protéines prédits chez l'humain suite au *Human Genome Project* (~20 300 contre 100 000 estimés au début du projet)(Pruitt et al, 2007), il est apparu que la corrélation entre nombre de gènes et complexité macroscopique (taille, nombre de types cellulaires différents) d'un organisme eucaryote est mauvaise (Figure 2)(Gregory, 2002, Lynch & Conery, 2003, Pertea & Salzberg, 2010, Vogel & Chothia, 2006). Une explication amenée est que l'expansion au niveau génique (nombre de gènes) de certaines familles de protéines existantes permet cette complexification (régulation de l'expression génique, processus extracellulaires), alors que l'expansion d'autres familles affecte peu la complexité globale de l'organisme en question (Lespinet et al, 2002, Vogel et al, 2003, Vogel & Chothia, 2006).

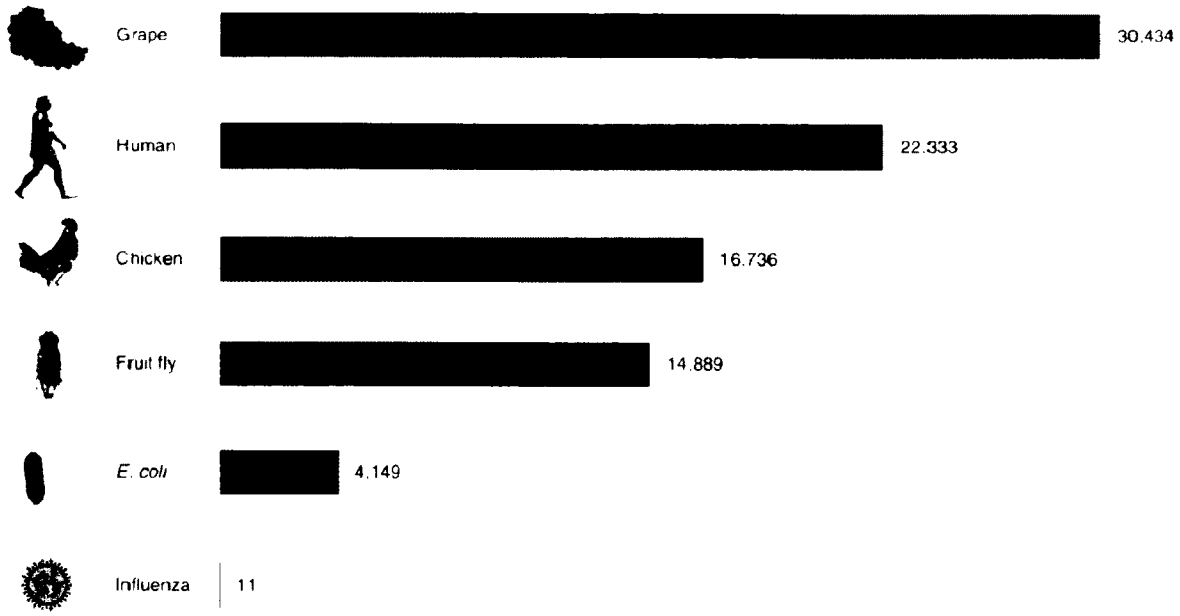


Figure 2. Nombre de gènes dans plusieurs espèces. Adapté de (Pertea & Salzberg, 2010). Le nombre de gène porté par une espèce ne semble que peu corrélé à sa taille ou à sa complexité morphologique. Les plantes et les animaux n'ont par exemple que deux à dix fois plus de gènes que la bactérie *Escherichia coli*.

Une autre explication pourrait être l'augmentation de la variabilité au niveau protéique, et non au niveau du nombre de gènes (Schluter et al, 2009). En effet, le produit protéique d'un seul cadre ouvert de lecture (ORF*) peut être trouvé sous de nombreuses formes appelées **protéoformes** (Figure 3) (Smith et al, 2013). Par exemple, une forme glycosylée et son équivalente non glycosylée d'une même protéine représentent deux protéoformes différentes. De nombreux mécanismes (transcriptionnels, post-transcriptionnels, traductionnels et post-traductionnels) participent à augmenter le nombre de protéoformes d'une même protéine chez les eucaryotes en comparaison aux procaryotes. Ces mécanismes seront détaillés dans le paragraphe 1.2. A titre d'exemple, la majorité des pré ARN messagers eucaryotes subissent un épissage alternatif, contrairement aux ARNm procaryotes, produisant plusieurs ARNm matures différents à partir d'un seul gène (Brett et al, 2002).

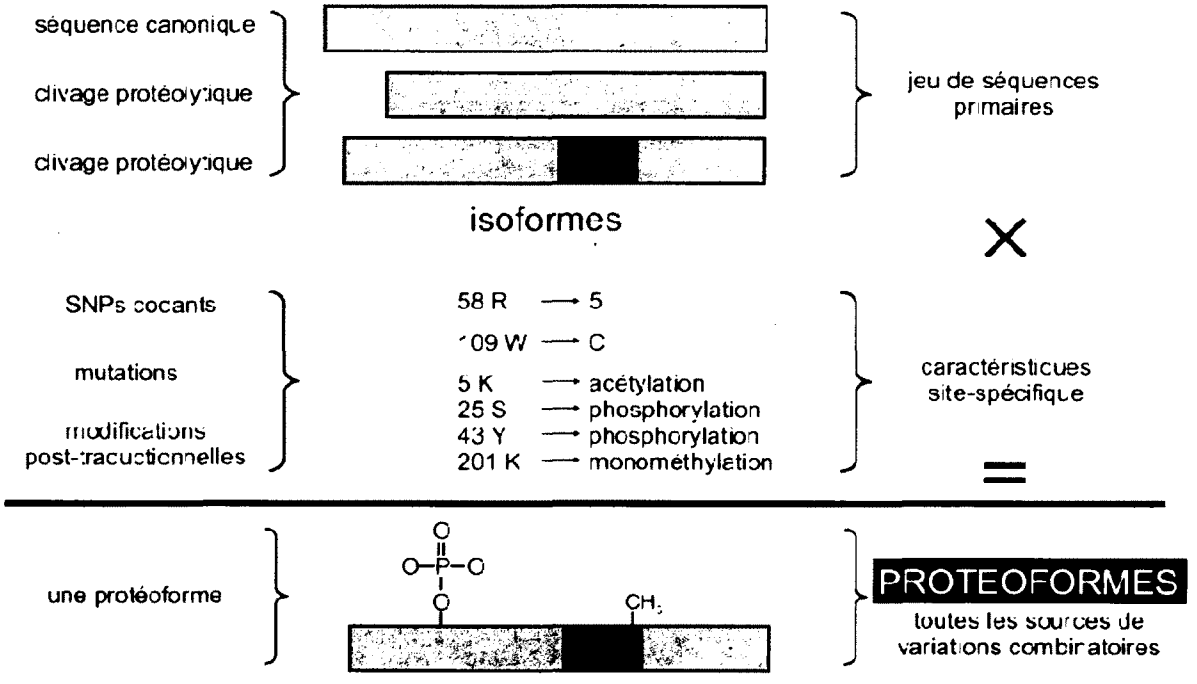


Figure 3. Protéoforme : un terme pour décrire la variabilité au niveau de la structure primaire des protéines. Adapté de (Smith et al, 2013). Par différents mécanismes, un seul gène peut produire plusieurs isoformes dont la séquence en acides aminés varie. Une variété de caractéristiques site-spécifique ajoute un niveau de complexité supplémentaire au nombre possible d'entités protéiques distinctes (protéoformes) produites depuis un ORF donné à un locus donné.

*ORF (pour *open reading frame* en anglais) : séquence nucléotidique délimitée par un codon d'initiation de la traduction et un codon stop.

L'augmentation du nombre de **groupes de protéoformes** (un groupe de protéoforme étant l'ensemble des produits protéiques issus d'un ORF donné) produits à partir d'un seul gène peut aussi expliquer l'augmentation de la diversité protéique au cours de l'évolution. En effet, les ARNm eucaryotes matures semblent de plus en plus reconnus comme capables de guider la production de plusieurs groupes de protéoformes indépendants par l'utilisation de plusieurs ORFs. Cela sera évoqué dans les paragraphes 1.2.4, 2.2, et 3.

Il est clair que, au vu de l'importance fonctionnelle des protéines dans le vivant, l'augmentation de la diversité protéique permet d'amener de nouvelles possibilités de régulation des processus biologiques, et donc de complexité phénotypique. Ainsi, l'objectif de l'ère post-génomique d'évaluer la diversité au niveau protéomique des organismes devrait permettre de déterminer sa contribution à la complexification des organismes vivants observée au cours de l'évolution.

1.1.3. Importance des petites protéines

Bien que le génome de nombreuses espèces soit séquencé à ce jour, l'identification adéquate de petits ORFs codants est un problème récurrent (Basrai et al, 1997). Lorsque l'on observe la distribution de la taille des ORFs répertoriés dans les bases de données protéiques actuelles, une baisse bien visible du nombre de protéines référencées apparaît à partir de 100 AA et moins (Frith et al, 2006). Il n'y a pas de raison *a priori* pour laquelle une discontinuité aussi marquée apparaisse, mais plusieurs explications peuvent être avancées. Tout d'abord, les algorithmes de prédiction de régions génomiques codantes peuvent être mis en cause (Mathe et al, 2002) : en diminuant la taille minimale de prédiction, la possibilité d'observer un ORF par chance et donc le nombre de faux positifs augmentent. Une taille minimale de 100 codons est donc utilisée dans la majorité des cas (Carninci et al, 2005). Une seconde explication est liée au dogme central de la biologie moléculaire, aujourd'hui remis en cause, « un gène – une protéine ». La nécessité de ne prédire qu'une seule séquence codante par gène a amené à ne retenir, par défaut, que le plus grand ORF. Ceci introduit un biais vers le référencement des ORFs les plus longs. Basé sur ces faits, il est logique que les fonctions des ORFs les plus longs aient été étudiées en priorité. Ainsi, le nombre réduit de protéines de petite taille prédites a certainement amené à

une étude moins assidue de leur fonction, et indirectement à l'idée que les petites protéines ne portent peu ou pas de fonctions importantes.

Cette vision est en pleine évolution face au nombre croissant de contre-exemples à cette assomption. Certains produits protéiques de petite taille ont été associés à des fonctions biologiques importantes. Par exemple, les chimiokines, dont la taille varie entre 8 et 10 kDa, sont capables de jouer des rôles dans la chimiotaxie des leucocytes, le développement, l'angiogenèse, la tumorigenèse, la métastase, la réponse immunitaire et le contrôle d'infections microbiennes (Le et al, 2004). Un peptide de 24 acides aminés (AA) appelé humanin a une fonction anti-apoptotique (Guo et al, 2003) et pourrait avoir un rôle neuroprotecteur dans la maladie d'Alzheimer (Matsuoka, 2011). La protéine ISD11 (91 AA chez l'humain) est essentielle pour la biogenèse des centres fer-soufre et l'homéostasie du fer, de la levure à l'humain (Shi et al, 2009, Wiedemann et al, 2006). Certains peptides peuvent réguler la fonction de facteurs de transcription, affectant ainsi le développement d'un organisme entier (Kondo et al, 2010). Des petits ORFs codant pour des peptides d'environ 30 à 50 AA conservés depuis 550 millions d'années régulent la contraction des muscles cardiaques des insectes jusqu'à l'humain (Magny et al, 2013).

Face à la multiplication des exemples de petites protéines aux fonctions biologiques primordiales, des approches à large échelle pour les prédire et en valider l'expression ont été mises en place. Remarquant la discontinuité, en dessous de 100 AA, des protéines référencées dans les bases de données protéiques actuelles (UniProt, International Protein Index (projet maintenant fermé)), Frith *et al.* ont entrepris d'évaluer, à partir d'une banque de données d'ADNc murins, le nombre de régions codant pour des petites protéines dans un génome de mammifère (Frith et al, 2006). Ils ont montré que la discontinuité peut être perdue lorsqu'aucun filtre de taille minimale n'est appliqué pour la prédiction d'ORFs, et que le potentiel codant est évalué par alignement de séquences avec d'autres espèces. Ainsi, ils prédisent environ trois fois plus d'ORFs de moins de 100 codons que dans les bases de données habituelles en ne considérant qu'un seul ORF par ARNm. Une autre approche a été de prédire et d'inclure dans une base de données (tsORFdb) l'ensemble des petits ORFs du génome de plusieurs espèces, de la levure à l'humain, afin d'aider à l'identification de nouvelles protéines par spectrométrie de masse (Heo et al, 2010). Chez les bactéries, les produits protéiques issus de petits ORFs sont également de plus en plus étudiés pour leurs

fonctions diverses (Hobbs et al, 2011). Ainsi, l'utilisation de la génomique comparative (comparaison de séquences afin d'évaluer le niveau de conservation d'une région génomique donnée) a permis d'identifier 35 petits ORFs bactériens (<50 AA, expression validée expérimentalement) parmi les 200 ORFs prédits avec le plus de confiance, et 6 petits ORFs parmi les 10 meilleurs (Samayoa et al, 2011).

Certaines études à large échelle ont apporté des preuves expérimentales de l'expression et de la fonction de produits protéiques issus de petits ORFs. Chez la levure *Saccharomyces cerevisiae*, l'application de méthodes computationnelles et expérimentales ont permis de déterminer que ~5% des ORFs annotés contenaient moins de 100 codons (299 au total, dont 170 découverts dans l'étude en question)(Kastenmayer et al, 2006). 184 de ces petits ORFs sont conservés entre la levure et d'autres organismes, soulignant leur probable importance fonctionnelle. Aussi, par analyse phénotypique de mutants de délétion, les auteurs ont lié la fonction de 22 nouveaux petits ORFs à la croissance dans plusieurs conditions de culture (haploïde, haute température, dommages à l'ADN, ...). Chez la plante modèle *Arabidopsis thaliana*, près de 8 000 ORFs de 30 à 100 codons ont été prédits dans les régions intergéniques, et l'expression de 2 099 d'entre eux a été validée expérimentalement (évidences au niveau de l'ARNm). 571 sont conservés chez d'autres plantes. La surexpression de 49 petits ORFs (sur 473 testés) a provoqué des changements phénotypiques visibles, une proportion 7 fois plus élevée que celle attendue par hasard, indiquant que beaucoup de petits ORFs semblent être impliqués dans la morphogenèse (Hanada et al, 2013). Chez la mouche *Drosophila melanogaster,* une étude génomique a identifié 401 petits ORFs potentiellement fonctionnels (Ladoukakis et al, 2011). Des études protéomiques (Oyama et al, 2004, Oyama et al, 2007) ont également contribué à amener des évidences expérimentales généralisant l'expression de protéines de petite taille, dont l'importance biologique a été sous-estimée auparavant.

1.2. Mécanismes eucaryotes pour la diversité protéique

Une grande diversité protéique semble être favorable pour la complexification des organismes vivants et de nombreux mécanismes biologiques, que je vais évoquer dans cette partie, sont disponibles pour y contribuer.

1.2.1. Patrimoine génétique

Le premier niveau contribuant à (et limitant) la quantité finale de protéines différentes produites par un organisme est constitué par le nombre de gènes codant pour des ARNs potentiellement traduits que porte son génome. Les mécanismes permettant l'apparition de nouveaux gènes au cours de l'évolution a donc été le sujet d'un intérêt scientifique particulier (revu dans Kaessmann, 2010). Tout d'abord la duplication de gènes est un contributeur majeur à l'apparition de nouveaux gènes. Elle peut apparaître comme conséquence d'une duplication d'un segment chromosomique contenant des parties de gènes ou des gènes entiers. Une recombinaison sujette à erreur lors de la méïose explique cela en grande partie (Kaessmann, 2010). La duplication de gènes peut aussi être une conséquence d'une duplication entière du génome par polyploïdisation (Conant & Wolfe, 2008, Van de Peer et al, 2009). La rétroduplication (transcription inverse d'un ARNm mature, puis insertion de l'ADN complémentaire (ADNc) dans le génome) participe aussi à l'apparition de nouveaux gènes (Kaessmann et al, 2009). La formation de gènes chimères, par juxtaposition suite à duplication de fragments de gènes (Bailey et al, 2002, Paulding et al, 2003), ou par couplage de transcription et épissage intergéniques (Akiva et al, 2006, Parra et al, 2006), amène également à de nouvelles combinaisons de séquences protéiques possibles. Enfin, un mécanisme qui semble plus rare (Jacob, 1977) est l'apparition de nouveaux gènes codant pour des protéines à partir de séquences précédemment non-codantes. Mécanistiquement, c'est le couplage (dans un ordre ou un autre) entre mutations agrandissant l'ouverture d'un cadre de lecture primitif (suppression de codons stops) et son activation transcriptionnelle (apparition d'un promoteur en amont) qui mène à l'apparition de nouveaux gènes (Knowles & McLysaght, 2009). Des exemples évolutivement favorables de gènes apparus à travers ces différents mécanismes existent (Kaessmann, 2010), soulignant l'importance de l'apparition de nouveaux gènes et des nouvelles protéines qu'ils encodent dans l'augmentation de la complexité du vivant.

1.2.2. Mécanismes transcriptionnels

Pour un seul gène donné, de multiples ARNm peuvent être transcrits, augmentant le potentiel codant de ce gène avant même la maturation des ARNm (cf paragraphe 1.2.3). Deux mécanismes sont ici possibles : l'utilisation de sites d'initiation ou de terminaison alternatifs de la transcription (Davuluri et al, 2008, Pal et al, 2011). Les variations entre les différentes espèces d'ARNm produites peuvent se situer au niveau de la région codante du gène, affectant le nombre de protéoformes d'une protéine produite par ce gène (Goossens et al, 2007, Tappe & Kuner, 2006). Les régions régulatrices dites non-traduites (UTR, de l'anglais *untranslated region*) peuvent aussi être touchées par ces mécanismes, affectant l'expression protéique en aval (Blaschke et al, 2003, Pozner et al, 2000). Nous verrons plus loin (paragraphes 2.2 et 3) que, les UTRs portant parfois des ORFs traduits en protéines, la diversité protéique issue d'un gène peut également être affectée par les variations des UTRs entre plusieurs isoformes d'ARNm produites par ces mécanismes transcriptionnels. Il est important de noter, en particulier pour l'initiation alternative de la transcription, que ce n'est pas un phénomène marginal. Dans le cervelet embryonnaire et adulte chez la souris, il a même été proposé que les évènements alternatifs de transcription dépassent le déterminant post-transcriptionnel majeur de la diversité transcriptomique (l'épissage alternatif, cf paragraphe 1.2.3)(Pal et al, 2011).

1.2.3. Maturation des ARN messagers

Chez les procaryotes, un ARNm est généralement mature dès la fin de la transcription, et ne subit donc pas de modifications affectant son potentiel codant. La traduction est d'ailleurs couplée à la transcription chez les procaryotes. En revanche les pré ARNm eucaryotes subissent des étapes de maturations importantes au noyau avant d'être exportés au cytoplasme pour y être traduits.

L'ajout de la structure coiffe à l'extrémité 5' a un effet stabilisateur en protégeant l'ARNm de la dégradation, et favorise l'initiation de la traduction en augmentant le recrutement des ribosomes sur l'ARNm (Shuman, 2002). Cette modification ne change pas *a priori* le potentiel codant d'un ARNm, car elle n'en affecte pas la séquence primaire.

Etant donné la structure composée d'introns et d'exons des gènes eucaryotes, une fois qu'un pré ARNm est transcrit, il doit subir un évènement appelé épissage (Figure 4). Brièvement, le mécanisme d'épissage met en jeu le complexe d'épissage (appelé *spliceosome* en anglais), qui permet d'exciser des régions du pré ARNm en reliant des jonctions exon/intron (dits sites d'épissages) entre eux. On distingue les sites d'épissage 5', constitués par l'extrémité 3' d'un exon et l'extrémité 5' de l'intron qui suit, aux sites d'épissage 3', formés par l'extrémité 3' d'un intron et l'extrémité 5' de l'exon qui suit. D'un point de vue simpliste, l'épissage permet d'exciser les introns, et après ligation des exons, ne laisse que ces derniers dans la séquence de l'ARNm mature (Figure 4 A). L'utilisation de combinaisons variables de sites d'épissage dans un pré ARNm permet d'augmenter le nombre de combinaisons d'ARNm matures produits et donc de séquences à traduire à partir d'un pré ARNm unique : c'est le concept d'épissage alternatif (Graveley, 2001, Pajares et al, 2007)(Figure 4 B-F). Certains exons sont dits constitutifs car ils sont présents dans toutes les isoformes d'ARNm matures produites, contrairement aux exons dits alternatifs qui ne sont pas systématiquement conservés après l'épissage. De plus certains exons sont mutuellement exclusifs, et parfois, un exon donné peut avoir plusieurs sites d'épissage 5' ou 3', ce qui peut changer en partie la séquence de cet exon. Enfin, des introns sont parfois conservés dans les ARNm matures, et même traduits par la suite. Au niveau de l'impact fonctionnel de l'épissage alternatif, outre les effets régulateurs dus aux modifications des régions UTRs (Palaniswamy et al, 2010, Wu et al, 2013), ceci aboutit là encore à la production possible de multiples protéoformes d'une même protéine, voire à de multiples groupes de protéoformes si plusieurs ORFs sont encodés dans un locus donné. Des approches à l'échelle du génome indiquent que ~95% des gènes humains qui contiennent plusieurs exons semblent sujet à de l'épissage alternatif, ce qui en fait une source majeure de diversité protéique (Pan et al, 2008).

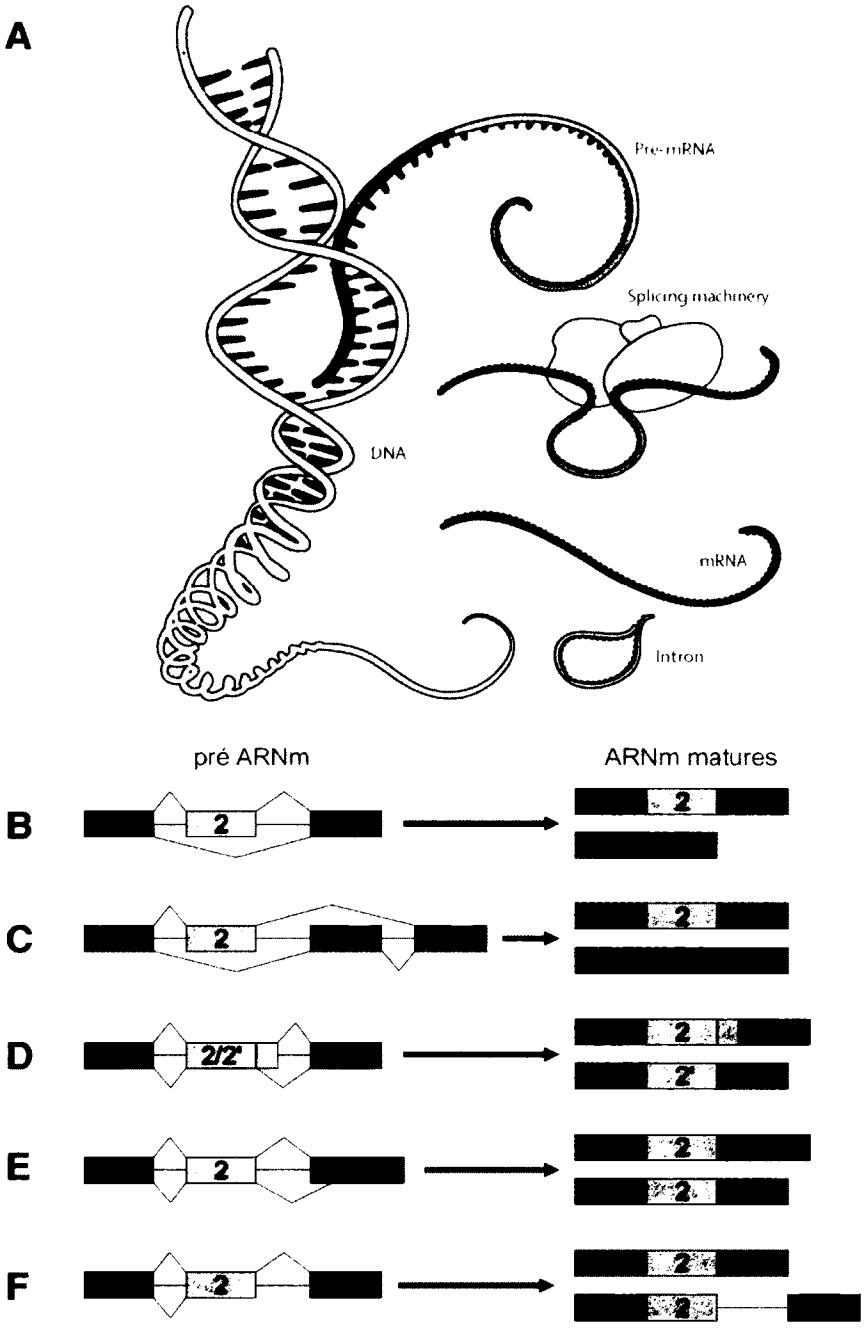


Figure 4. L'épissage alternatif en cis génère de la diversité protéique. Adapté de (Pajares et al, 2007). Noter que les pré ARNm présentés dans chaque cas de figures sont différents les uns des autres. (**A**) Illustration du mécanisme d'épissage. Une fois transcrit, le pré ARNm contenant introns et exons est pris en charge par la machinerie d'épissage (*splicing machinery* en anglais) qui permet, de manière générale, de retirer les introns et de ne conserver que les exons dans l'ARNm mature. Les étapes de clivage et ligation nécessaires à ce processus ont lieu aux sites d'épissage (jonctions exons/introns). L'épissage en cis d'un pré ARNm peut suivre les patrons suivants : (**B**) l'inclusion d 'un exon, (**C**) l'inclusion mutuellement exclusive d'exons, (**D**) l'utilisation de sites d'épissage 5' alternatifs, (**E**) l'utilisation de sites d'épissage 3' alternatifs, (**F**) la rétention d'un intron.

La diversité protéique produite est illustrée par les combinaisons variables de séquences maintenues dans les ARNm matures. Les lignes horizontales noires représentent les introns. Les boîtes grises et oranges représentent les exons. Les lignes en diagonales indiquent les différentes combinaisons possibles dans l'utilisation des sites d'épissage.

Un épissage entre deux molécules distinctes d'ARN peut parfois survenir (Konarska et al, 1985, Solnick, 1985, Yang & Walsh, 2005). Cet évènement appelé *trans*-épissage semble rare chez les mammifères bien qu'il y ait été observé (Caudevilla et al, 1998, Flouriot et al, 2002). Il est plus courant chez des organismes comme le trypanosome, le nématode, ou la drosophile (Yang & Walsh, 2005). Cela peut aboutir à l'association d'exons provenant de pré ARNm issus de loci différents dans une même molécule d'ARNm mature unique. Après traduction, des chimères protéiques sont alors synthétisées (Figure 5 A). Ce mécanisme est devenu une option thérapeutique de choix afin de remplacer des exons défectueux dans plusieurs pathologies, en exprimant simplement l'exon sauvage dans les cellules cibles (Puttaraju et al, 1999, Yang & Walsh, 2005). Une variante du *trans*-épissage permet à deux molécules d'ARNm matures d'être produites à partir d'une seule molécule de pré ARNm (Blumenthal, 2004, Spieth et al, 1993). Cette variante consiste en un clivage d'un pré ARNm dans une région intronique, ce qui produit deux brins d'ARN, l'un contenant les premiers exons et introns, l'autre les suivants. Le fragment en aval du site de clivage subit alors l'ajout d'une région *leader* à son extrémité 5' (typiquement l'ARN SL2 chez *Caenorhabditis elegans*). Le fragment en amont est simplement cis-épissé. Après épissage, deux ARNm matures sont obtenus, contenant chacun deux ORFs distincts (Figure 5 B). Ce mécanisme est très utilisé chez *C. elegans*, où environ 1000 opérons pourraient être présents, impliquant ~15% du total des gènes (Blumenthal et al, 2002). Plusieurs groupes de protéoformes peuvent ainsi être formés à partir d'un pré ARNm unique.

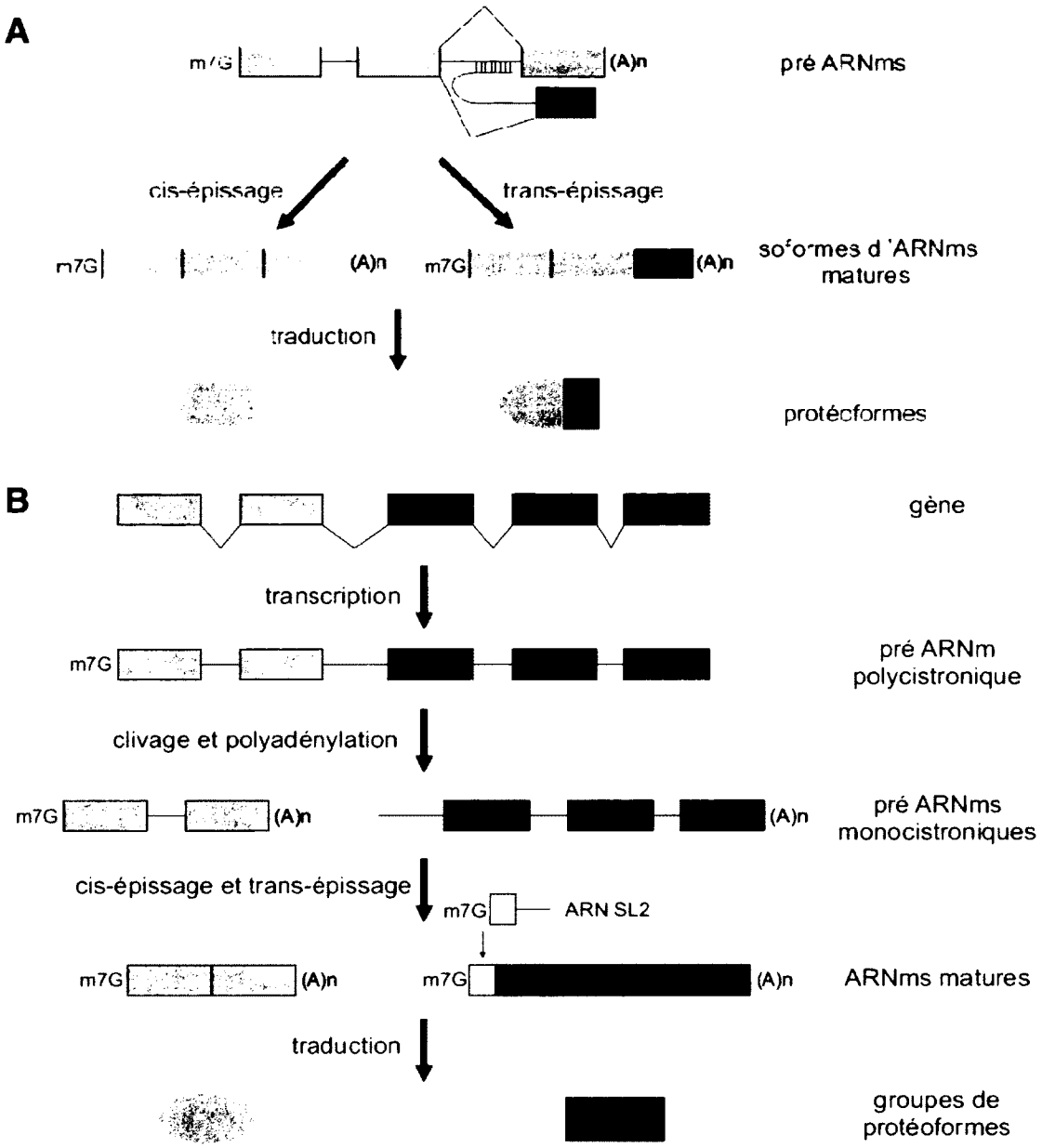


Figure 5. Mécanisme et conséquences du trans-épissage. (**A**) Adapté de (Blumenthal, 2004). De manière similaire au cis épissage, le trans-épissage permet d'obtenir des combinaisons d'exons variables dans les ARNm matures, mais cette fois à partir de plusieurs pré ARNm. Plusieurs protéoformes de la même protéine sont ainsi produites, combinant des exons issus de loci différents. Les boîtes oranges représentent les exons d'un pré ARNm, la boîte noire l'exon d'un autre pré ARNm. (**B**) Adapté de (Yang & Walsh, 2005). Le trans-épissage est particulièrement utilisé chez C. elegans, où il permet l'expression de groupes de protéoformes indépendants à partir d'un pré ARNm unique. Le pré ARNm subit un clivage et une polyadénylation en deux sites distincts. Le brin correspondant à la région la plus en aval du pré ARNm subit alors une ligation avec une région *leader* coiffée, le plus souvent correspondant à l'ARN SL2. Les deux ARNm matures ne partagent pas de séquence codante en commun, et dirigent donc l'expression de groupes de protéoformes distincts (exons bleus *vs* exons oranges).

L'ajout d'une queue poly-A en 3' des ARNm eucaryotes a un effet protecteur contre les 3'-5' exoribonucléases (stabilisant les ARNm), et favorise l'efficacité de traduction (Zhang et al, 2010). Cependant, puisque son ajout nécessite un clivage de l'ARNm à un site de polyadénylation déterminé par la séquence primaire de l'ARNm, l'utilisation de sites de polyadénylation alternatifs peut permettre d'inclure ou d'exclure une partie des régions transcrites des ARNm matures (Figure 6). Les conséquences sont variables en fonction de l'emplacement des différents sites de polyadénylation possibles (Di Giammartino et al, 2011). Si les différents sites sont tous dans le 3'UTR du gène, les éléments régulateurs qu'il porte sont possiblement affectés. Ceci a des effets sur la stabilité et/ou la traduction des ARNm matures, par exemple en modulant la présence de sites de reconnaissance pour les mécanismes d'interférence par l'ARN. En revanche si les différents sites de polyadénylation sont distribués dans différents exons/introns, alors cela peut affecter la séquence codante lors des évènements d'épissage subséquents. Le nombre de protéoformes possibles pour une même protéine se trouve alors augmenté (Di Giammartino et al, 2011). La polyadénylation alternative semble toucher de nombreux gènes chez les eucaryotes (10-15% des gènes chez la levure (Nagalakshmi et al, 2008), ~54% chez l'humain (Tian et al, 2005)), et est peu conservée entre les espèces (Ara et al, 2006).

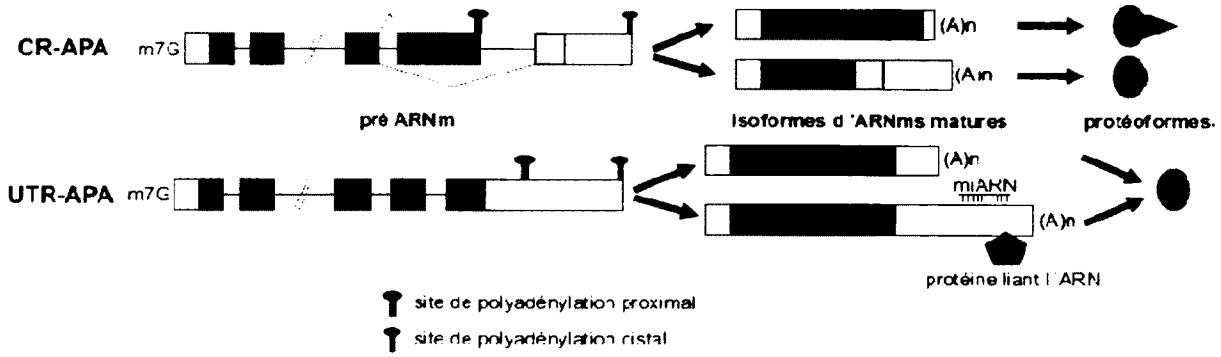


Figure 6. Conséquences fonctionnelles de l'utilisation de sites alternatifs de polyadénylation. Adapté de (Di Giammartino et al, 2011). Au sein d'un gène, on distingue les sites alternatifs de polyadénylation selon qu'ils affectent soit la séquence codante (CR-APA), soit les UTRs (UTR-APA). Dans la CR-APA, un même pré ARNm est épissé alternativement, donnant deux isoformes d'ARNm qui varient au niveau de leur dernier exon codant. Cela se répercute au niveau de la diversité protéique, puisque deux protéoformes différentes d'une même protéine sont générées, possédant des portions C-terminales distinctes. L'UTR-APA amène, après clivage au sein du même exon final du pré ARNm, à l'incorporation ou non de sites de liaison pour des éléments régulateurs (transcriptionnels/traductionnels) au sein des différentes isoformes d'ARNm, tels que miARNs ou protéines liant l'ARN. CR-APA : de l'anglais *coding region - alternative polyadenylation*. UTR-APA : de l'anglais *untranslated region - alternative polyadenylation*. m7G : structure coiffe. miARN : micro-ARN. Les boîtes blanches représentent les séquences des exons qui constituent les UTRs. Les boîtes colorées représentent les séquences codantes des exons.

Enfin, les ARNm eucaryotes peuvent subir de l'édition, c'est-à-dire l'ajout, la délétion ou la modification d'un nucléotide (Figure 7). Les évènements d'édition les plus utilisés sont les modifications de nucléotides : déamination par des enzymes d'édition d'une cytosine (C) en uracile (U), ou, plus fréquemment, d'une adénine (A) en inosine (I, lu comme une guanine (G) par la machinerie traductionnelle). Bien que tous les ARNm ne soient pas sujets à édition, cette modification post-transcriptionnelle eucaryote-spécifique n'en est pas moins importante fonctionnellement, puisqu'un *knock-out* partiel ou complet de gènes impliqués dans l'édition peut être létal chez les mammifères (Keegan et al, 2001). Les sites subissant l'édition sont ciblés, mais peuvent être présents dans toutes les régions des pré ARNm (exons, introns, régions codantes, UTRs, sites d'épissage, codon stop)(Keegan et al, 2001). Ceci a des conséquences directes sur les protéoformes codées par les ARNm ainsi édités, tels une substitution d'acide aminé (Seeburg, 1996), un changement de cadre de lecture (*frameshift* en anglais)(Benne et al, 1986), un épissage alternatif (Rueter et al, 1999), l'apparition ou la suppression d'un codon initiateur ou stop (Koslowsky et al, 1990) ou des effets parfois indéterminés sur les UTRs (Morse & Bass, 1999).

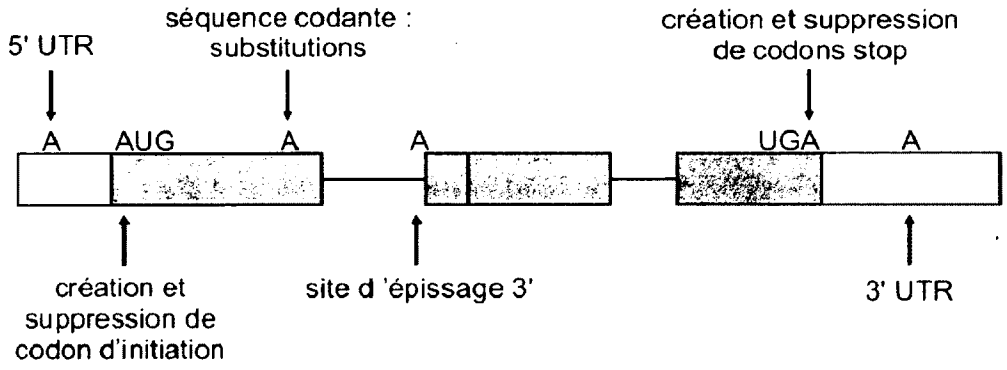


Figure 7. Conséquences possibles de l'édition des ARNm sur l'expression protéique. Adapté de (Keegan et al, 2001). L'exemple de l'édition A→I est utilisé ici. Des modifications dans la séquence des régions 5' ou 3' UTRs peuvent survenir bien que leurs conséquences soient inconnues à ce jour. La séquence codante peut subir de l'édition, créant des substitutions d'acides aminés. Des sites d'initiation ou d'arrêt de la traduction peuvent être créés ou abolis. Enfin, des séquences importantes pour l'épissage des pré ARNm peuvent être modifiées, ce qui peut avoir un impact sur la diversité des isoformes d 'ARNm matures produits, et sur les protéoformes qu'ils encodent.

1.2.4. Mécanismes traductionnels

Après avoir subi les étapes de maturation, les ARNm matures sont exportés au cytoplasme pour y être traduits en protéines par les ribosomes. A la différence de la régulation transcriptionnelle, la régulation de l'expression protéique au niveau de la traduction permet de produire des changements rapides dans les niveaux d'expression des protéines codées par les ARNm préexistants. La régulation traductionnelle est ainsi importante au cours de stress cellulaires, du développement et de la différenciation, du fonctionnement du système nerveux, dans le développement de maladies comme le cancer ainsi que dans leur traitement (Sonenberg & Hinnebusch, 2009). De manière simpliste, le dogme central implique que chaque ARNm ne dirige la synthèse que d'une seule protéine, par la traduction d'un unique ORF. L'ORF de chaque ARNm était alors identifié grâce à plusieurs règles incluant une taille minimale (l'ORF le plus long étant retenu), une utilisation de codons biaisée (fréquence d'utilisation de codons synonymes variable selon les espèces) (Bulmer, 1987), et un codon initiateur AUG (si possible le premier de l'ARNm) (Brent, 2005). Cette vision a été largement revisitée ces dernières années, indiquant que chaque ARNm mature peut en réalité contenir plusieurs ORFs et ainsi permettre la synthèse de plusieurs protéoformes, voire groupes de protéoformes (voir paragraphe 3) (Atkins & Gesteland, 2010, Kochetov, 2006, Kochetov, 2008). La traduction peut être modulée au niveau des quatre étapes qu'elle comporte : initiation, élongation, terminaison et recyclage des ribosomes. Cette dernière étape permet la dissociation du ribosome de l'ARNm afin de lui permettre d'effectuer à nouveau un cycle de traduction, et à ce jour aucune évidence n'indique qu'elle participe à la génération de diversité protéique. En revanche, la modulation des trois autres étapes participe à l'augmentation du nombre de produits protéiques générés à partir d'un seul ARNm mature (Atkins & Gesteland, 2010).

La majorité de la régulation du processus de traduction a lieu au niveau de l'initiation. L'initiation de la traduction consiste au recrutement d'un ribosome sur l'ARNm, suivi de la sélection du codon d'initiation de la traduction, s'achevant lorsque le ribosome est prêt à l'élongation (c'est-à-dire à former le premier pont peptidique). Différents mécanismes possibles d'initiation existent, et ces multiples mécanismes offrent de nombreuses possibilités quant au choix du codon initiateur. Tout d'abord, concernant l'identité du

codon d'initiation, AUG est considéré comme le codon initiateur par excellence, historiquement (Brent, 2005) et dans l'étendue de son utilisation (Ingolia et al, 2011, Lee et al, 2012). Cependant, il apparaît de plus en plus clairement que de très nombreux autres codons (en particulier variant d'un nucléotide par rapport à AUG) peuvent servir efficacement de codon initiateur, soit dans quasiment 50% des cas (Ingolia et al, 2011, Lee et al, 2012). Un autre facteur de diversité dans l'initiation concerne la position des codons initiateurs. En raison du mécanisme majoritaire de reconnaissance du codon d'initiation qui consiste en un balayage depuis l'extrémité 5' par la sous-unité 40S du ribosome (Kozak, 1978), la plupart des évènements d'initiation répertoriés se situent assez proche (au plus quelques centaines de nucléotides) de cette extrémité 5'. Néanmoins, une certaine souplesse dans ce mécanisme ainsi que l'existence de mécanismes alternatifs d'initiation permettent dans certains cas la sélection de sites d'initiation plus en aval, et donc la traduction d'ORFs qui n'auraient pas été traduits en protéine si le mécanisme canonique avait été utilisé (Malys & McCarthy, 2011, Sonenberg & Hinnebusch, 2009). Je mentionnerai simplement ici la notion conceptuelle que l'initiation en des sites différents permet d'inclure ou d'exclure certaines régions de l'ARNm de la ou des zone(s) traduite(s), et également de décoder la séquence de l'ARNm dans des cadres de lectures différents, affectant la diversité de séquences protéiques produites. Les possibilités offertes par les différents mécanismes d'initiation de la traduction eucaryotes en termes de diversité protéique seront discutées en détail dans la partie 2.2.

Lors de l'élongation de la traduction, un évènement appelé *frameshift* (changement de cadre de lecture) peut survenir, provoquant un saut en avant (+1) ou plus fréquemment un retour (-1) d'un nucléotide du ribosome et affectant ainsi la séquence en acides aminés traduite subséquemment (Farabaugh, 1996). La protéine ainsi produite est une chimère entre les séquences en acides aminés décodées depuis les différents cadres de lecture. Un *frameshift* peut apparaître spontanément par erreur, mais c'est un évènement très rare (fréquence de $3x10^{-5}$ par codon) qui peut toucher tous les ARNms en traduction (Farabaugh, 1996). En d'autres occasions, et beaucoup plus efficacement, le *frameshift* peut être programmé, ne touchant que certains ARNms contenant les signaux en *cis* nécessaires à induire le décalage. Cette programmation a été observée chez les procaryotes, les virus, et les eucaryotes (de la levure à l'humain). Mécanistiquement, les sites de changement de

cadre de lecture des *frameshifts* -1 sont constitués par certains heptamères nucléotidiques (de nature répétitive) favorisant le glissement du ribosome (Jacks et al, 1988), surtout si sa vitesse d'élongation est ralentie par une structure telle un pseudo-nœud dans la séquence de l'ARNm située juste en aval (Tu et al, 1992). Le *frameshift*, même programmé, n'est pas très efficace (5,1% pour l'ARNm du virus de l'immunodéficience humaine (Gendron et al, 2008)). Il permet la traduction de deux protéoformes différentes à partir d'un seul ARNm mature (Figure 8).

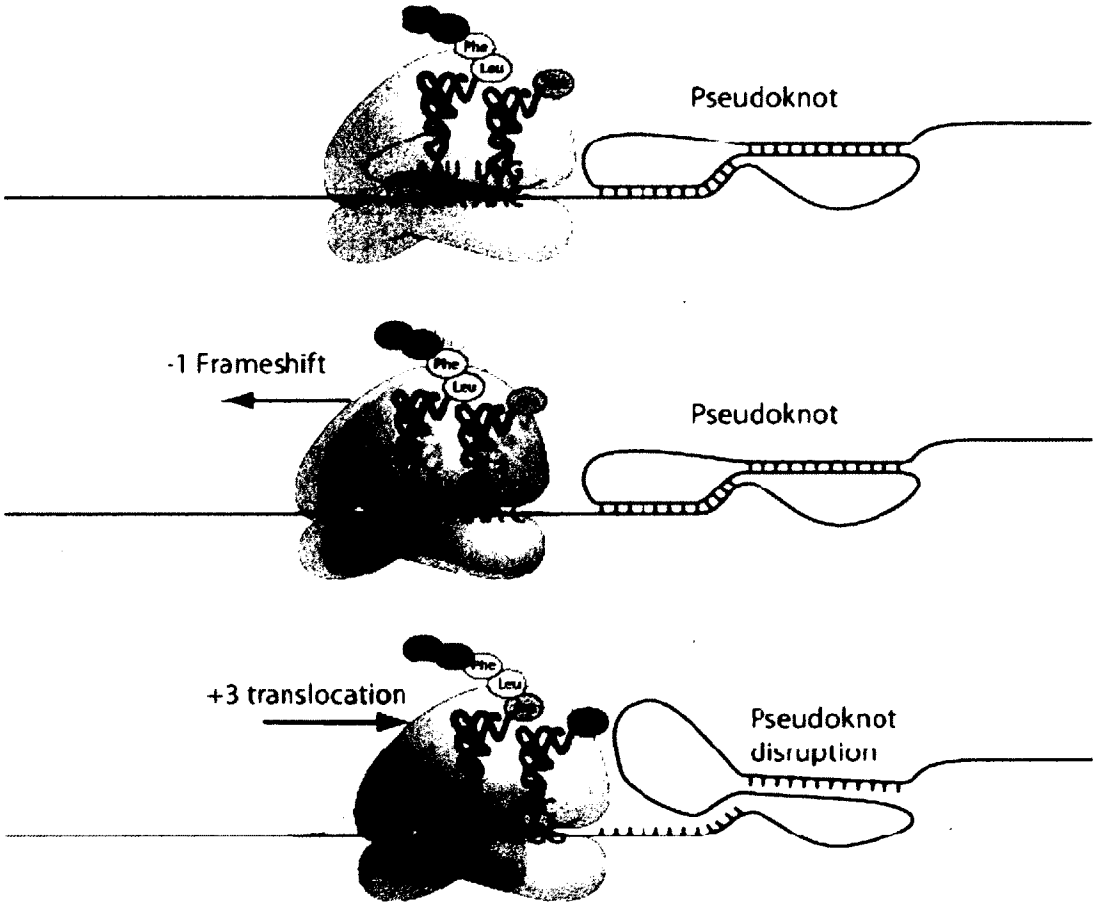


Figure 8. Le mécanisme de *frameshift* -1. Au cours de l'élongation de la traduction, la présence de certaines séquences nucléotidiques de nature répétitive induit un glissement du ribosome (retour en arrière, d'un nucléotide). Ceci est favorisé par la présence d'une structure stable en aval, telle qu'un pseudo-noeud (*pseudoknot*), qui ralentit l'élongation à proximité des séquences « glissantes ». Lorsque l'élongation de la traduction reprend, le décalage de cadre de lecture provoqué par le *frameshift* résulte en la synthèse d'une protéine chimérique entre les séquences en acides aminés décodées depuis les différents cadres de lecture. Adapté de http://viralzone.expasy.org/all_by_protein/860.html.

Au niveau de la terminaison de la traduction, il est parfois possible qu'un codon stop ne soit pas reconnu en tant que tel, mais qu'un acide aminé soit incorporé lors du décodage de ce codon par le ribosome en élongation (mécanisme appelé *stop codon readthrough* en anglais) (Doronina & Brown, 2006) (Figure 9). L'élongation continue alors jusqu'au prochain codon stop reconnu, provoquant l'ajout d'une portion C-terminale supplémentaire à la protéine produite. Dans le cas des sélénoprotéines, une sélénocystéine peut être ajoutée à un codon UGA, ce qui est favorisé par une séquence d'insertion de sélénocystéine dans le 3'UTR (Berry et al, 1993, Fixsen & Howard, 2010). Dans d'autres cas, la reconnaissance non canonique du codon stop par un ARN de transfert (ARNt) chargé permet l'ajout de l'acide aminé qu'il porte (Bonetti et al, 1995) avec une efficacité pouvant aller jusqu'à 5% (Namy et al, 2001). Des modulations dans le niveau d'expression ou dans le statut de maturation post-traductionnelle des protéines du facteur de libération du ribosome peuvent affecter de manière générale l'efficacité du *readthrough*, sur l'ensemble des ARNm (von der Haar & Tuite, 2007). Bien que dans la majorité des cas, un autre codon stop sera rencontré par le ribosome peu après le *readthrough*, n'affectant que légèrement la partie C-terminale de la protéine produite, l'ajout d'une portion supplémentaire pour une partie des protéines produites peut procurer certains avantages. C'est le cas des versions allongées de certaines protéines produites en condition de stress chez *S. cerevisiae*, où le *readthrough* est utilisé comme vecteur évolutif (True & Lindquist, 2000). Par exemple, certains allèles du gène *SKY1* portent des polymorphismes favorables dans la région traduite lors du *readthrough* facilité en présence de diamide ou de peroxyde d'hydrogène (deux agents oxydants), bien que le mécanisme moléculaire expliquant l'effet bénéfique de la protéine kinase encodée n'ait pas été élucidé (Torabi & Kruglyak, 2012). Chez la drosophile, la conservation évolutive du potentiel codant des régions directement en aval du codon stop de l'ORF principal indique que 283 gènes sont sujets au *readthrough*, qui semble particulièrement répandu chez les insectes et les crustacés (Jungreis et al, 2011).

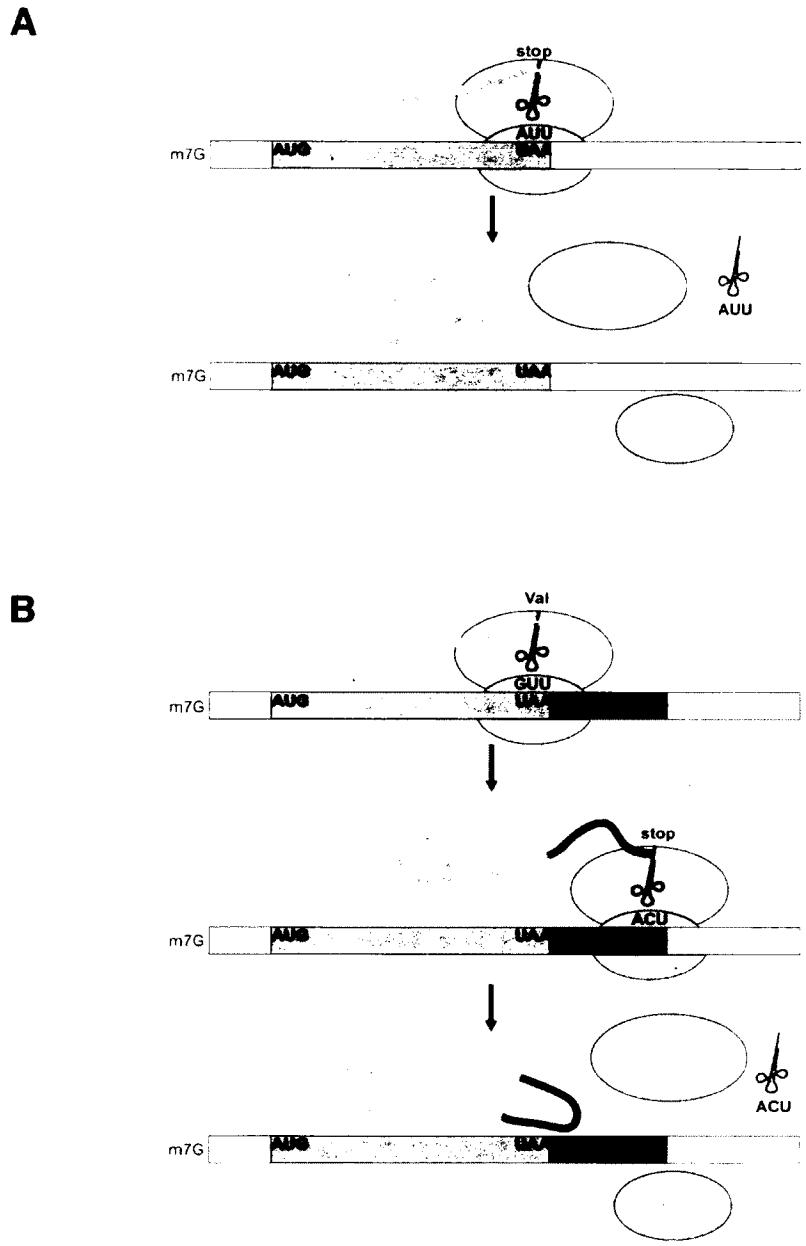


Figure 9. Le *readthrough* traductionnel apporte une variabilité protéique dans les portions C-terminales des protéines. (**A**) Lorsqu'un ribosome rencontre un codon stop, son décodage se fait de manière générale correctement, et la traduction s'arrête. Le ribosome est alors désassemblé, et la protéine qui vient d'être traduite est relâchée. (**B**) Lorsque la fidélité de reconnaissance du codon stop est altérée (contexte nucléotidique autour du codon, modifications dans l'expression des facteurs de terminaison), un ARNt à la séquence anticodon proche de la séquence complémentaire au codon stop peut être utilisé pour continuer l'élongation jusqu'au prochain codon stop reconnu. Une protéine avec une portion C-terminale supplémentaire (en rouge) est alors produite. Les boîtes représentent les différentes régions de l'ARNm (blanc: UTRs ; orange: séquence codante ; rouge: région du 3'UTR traduite additionnellement lors du *readthrough*). Les rubans colorés correspondent à la séquence protéique traduite. Les sous-unités ribosomales sont en gris, les ARNt en bleu.

1.2.5. Modifications post-traductionnelles

Une fois la traduction d'une protéine achevée, celle-ci peut subir des modifications post-traductionnelles, augmentant encore la diversité de protéoformes encodées dans un gène ou un ARNm donné (Seo & Lee, 2004, Smith et al, 2013). Ces modifications peuvent entre autres moduler la structure, la localisation, l'activation, la stabilité, ou le réseau d'intéraction des protéines, agissant directement sur leur fonctionnalité. Bien que beaucoup de modifications post-traductionnelles n'affectent pas la séquence primaire de la protéine (Seo & Lee, 2004), d'autres modifications, en revanche, produisent un remodelage de la séquence primaire, et seules celles-ci seront discutées ici. La plus répandue et la plus connue est le clivage par protéolyse, qui permet soit la maturation nécessaire d'un précurseur en protéine fonctionnelle, soit l'inactivation de l'activité d'une protéine. De plus, certaines protéines peuvent subir l'équivalent d'un épissage, c'est-à-dire l'excision autocatalytique d'un fragment interne (intéine) suivie de la formation d'une liaison peptidique entre les deux fragments externes (extéines) (Perler et al, 1994). Une base de données dédiée aux intéines répertorie plus d'une centaine de cas chez les eucaryotes (InBase (Perler, 2002)). Un trans-épissage protéique peut également avoir lieu chez les eucaryotes entre deux protéines traduites indépendamment, augmentant encore le potentiel de diversité protéique (Wu et al, 1998, Zhu et al, 2010).

2. Potentiel multi-codant des ARN messagers matures

Chez les eucaryotes, la possibilité pour un ARNm mature d'encoder plusieurs groupes de protéoformes différents à la fois comme mécanisme de diversité protéique a longtemps été négligée. Dans cette partie, je montrerai que ce mécanisme est pourtant largement utilisé dans des organismes primitifs (procaryotes) et relativement peu complexes (virus). Ensuite, j'énoncerai les raisons pour lesquelles ce mécanisme a été longtemps ignoré chez les eucaryotes, celles qui suggèrent la possibilité de son existence, et les possibilités offertes par l'utilisation de multiples ORFs dans un ARNm eucaryote unique.

2.1. Chez les procaryotes et les virus

2.1.1. Arrangement structural des ARN messagers matures procaryotes

L'utilisation d'un seul ARNm mature comme substrat pour la traduction de multiples groupes de protéoformes est un mécanisme qui est apparu très tôt dans l'évolution, puisqu'il est très répandu chez les procaryotes. Ceux-ci possèdent des gènes transcrits en ARNm monocistroniques ou polycistroniques dans des proportions variables. Le groupement dans une même unité transcriptionnelle de plusieurs ORFs (arrangement sous-forme d'opéron, Figure 10) offre un avantage au niveau de la régulation temporelle et stœchiométrique de l'expression de protéines impliquées dans une même voie, pour un coût énergétique minimal puisqu'ici un seul évènement transcriptionnel est nécessaire pour l'expression de plusieurs protéines (Ma et al, 2002a). La nature courte des régions non codantes séparant les différents ORFs dans les génomes des bactéries et des archées explique en partie pourquoi une telle organisation d'ARNm est possible (seulement un quart environ de leur génome ne code pas pour des protéines (Mattick, 2004)). Une explication supplémentaire réside dans le mécanisme général d'initiation de la traduction chez ces deux grands groupes du vivant (Figure 10). Brièvement, chaque ORF procaryote est précédé d'une séquence appelée Shine-Dalgarno (SD), qui interagit avec la séquence de l'ARN ribosomal (ARNr) 16S composant la petite sous-unité 30S du ribosome, et permet de recruter ce dernier à proximité du codon d'initiation AUG (Shine & Dalgarno, 1974). Ainsi, pourvu qu'une séquence SD soit présente et disponible pour interagir avec la sous-unité 30S, l'initiation de la traduction peut avoir lieu en de multiples régions d'un ARNm,

autorisant la traduction de plusieurs ORFs. Cette vision de ce que j'appelle le potentiel multi-codant des ARNm (plusieurs ORFs dans un seul ARNm mature) procaryotes est établie depuis longtemps (la caractérisation de l'opéron lactose par Jacob et Monod date de 1960 (Jacob et al, 1960)), et est encore complexifiée par des découvertes plus récentes. Il apparaît maintenant que les archées (et les bactéries dans une moindre mesure) ont développé des mécanismes d'initiation de la traduction SD-indépendants, augmentant le nombre de possibilités. L'interaction de la protéine ribosomale S1 avec une région riche en pyrimidines en amont de l'AUG (Boni et al, 1991), des intéractions SD-*like* entre l'ARNm et l'ARNr (Kozak, 2005), ou encore une initiation 5'UTR-dépendante et SD-indépendante pour l'instant peu caractérisée au niveau moléculaire (Hering et al, 2009) sont autant de mécanismes alternatifs qui peuvent être utilisés, et ce assez fréquemment (Chang et al, 2006). De plus, certains ARNm qui n'ont pas de 5'UTR peuvent tout de même recruter un ribosome 70S (contenant les deux sous-unités 30S et 40S déjà assemblées) directement à l'AUG (Moll et al, 2002).

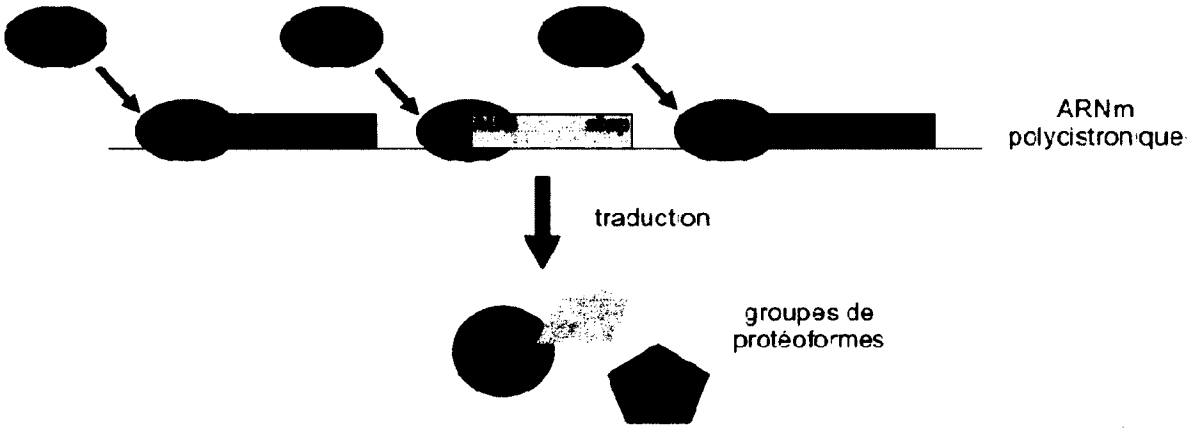


Figure 10. Les ARNm procaryotes sont souvent polycistroniques, organisés sous forme d'opéron. Les bactéries et les archées possèdent des structures génomiques permettant d'inclure plusieurs ORFs dans une unité transcriptionnelle unique. En amont de chaque ORF, une séquence Shine-Dalgarno (SD) permet de recruter directement dans des régions internes de l'ARNm la sous-unité ribosomale 30S, et de la placer à proximité du codon initiateur (AUG ici). Chaque ORF est alors traduit à partir de cet opéron, aboutissant à la production de protéines souvent impliquées dans la même voie métabolique. La ligne horizontale noire représente l'ARNm, les boîtes colorées les ORFs, et les formes colorées les protéines correspondantes. Les sous-unités 30S sont en gris.

2.1.2. Arrangement structural des ARN messagers matures viraux

Les virus, pour effectuer leur cycle viral et se multiplier, doivent produire un certain nombre de protéines encodées dans leur génome. Ils ne possèdent pas leur propre système de traduction, et ils utilisent donc la machinerie de la cellule hôte à cette fin. Ils utilisent donc de manière subversive des mécanismes déjà disponibles et opérationnels dans la cellule infectée. De plus, face à la petite taille de leur génome, les virus ont optimisé leur potentiel codant, et plusieurs groupes de protéoformes sont souvent encodées par ARNm mature. Utilisant plusieurs sites d'initiation de la traduction dans différents cadres de lecture (mais aussi du *frameshift*), ils profitent efficacement des différents mécanismes de traduction à leur disposition, autant classiques que non-canoniques.

Les bactériophages, tout comme leurs hôtes bactériens, possèdent un génome organisé sous forme d'opérons (Figure 11 A), et produisent des ARNm polycistroniques. En plus de permettre la traduction de plusieurs ORFs successifs, les ARNm des bactériophages présentent un niveau de diversification supplémentaire en utilisant des ORFs se chevauchant (Normark et al, 1983). En utilisant plusieurs cadres de lecture, soit par *frameshift*, soit par initiation de la traduction à des sites alternatifs, les ARNm de bactériophages peuvent donc coder pour de nombreuses protéines dans un espace génomique réduit. L'exemple du bactériophage G4 est particulièrement éloquent, puisque certains nucléotides peuvent être traduits dans les trois cadres de lecture différents (Godson et al, 1978, Shaw et al, 1978).

La situation est similaire chez les virus eucaryotes, où l'utilisation de cadres de lecture chevauchants et de sites alternatifs d'initiation de la traduction a été bien caractérisée. C'est le cas chez les virus de plantes, comme le virus de la mosaïque du navet jaune où deux codons d'initiation de la traduction séparés de 7 nucléotides dirigent l'expression de deux protéines distinctes dans des cadres de lecture différents mais qui se chevauchent (Matsuda et al, 2004). Chez les mammifères également, les virus peuvent utiliser plusieurs sites d'initiation de la traduction pour traduire des ORFs distincts à partir d'un seul ARNm. En particulier, chez l'humain, le virus de l'hépatite C produit une polyprotéine par l'usage d'un très long ORF, mais une initiation de la traduction peut avoir lieu en aval de l'AUG de cet ORF principal, permettant la production d'une protéine immunogénique impliquée dans le

développement de complications hépatiques associées à l'infection (Figure 11 B)(Baril & Brakier-Gingras, 2005, Vassilaki & Mavromara, 2009, Walewski et al, 2001). Le virus de l'immunodéficience humaine de type 1 produit également des ARNm avec plusieurs ORFs se chevauchant dans différents cadres de lecture (Karn & Stoltzfus, 2012). D'autres exemples similaires d'organisations d'ARNm existent (Bolinger & Boris-Lawrie, 2009, Kozak, 1986a), ce qui semble être la norme plus que l'exception chez les virus.

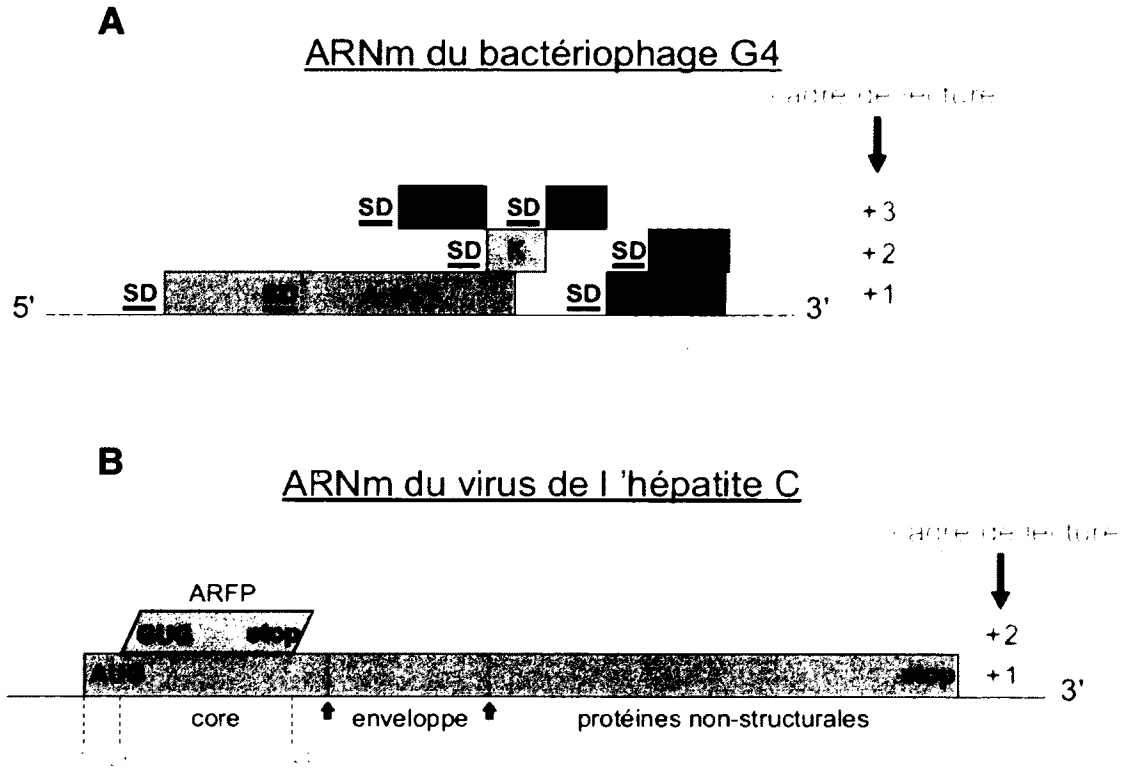


Figure 11. Organisation des ORFs dans les ARNm matures viraux. (**A**) Adapté de (Godson et al, 1978). Pour les virus de procaryotes, l'exemple du bactériophage G4 est présenté. L'un des ARNm (ligne noire horizontale) produit par le phage G4 dirige la synthèse de multiples protéines en utilisant les trois cadres de lectures disponibles. Un site de liaison des ribosomes (SD) précède chaque ORF (boîtes colorées). Seule une portion de l'ARNm complet est présentée. (**B**) Adapté de (Vassilaki & Mavromara, 2009). Pour les virus de cellules eucaryotes, l'ARNm (ligne noire horizontale) du virus de l'hépatite C (HCV) est présenté. La polyprotéine de HCV est encodée par l'ORF principal (boîte grise), et subit des clivages aux sites indiqués par les flèches noires afin de produire les protéines structurales et non-structurales du virus. Une seconde protéine appelée ARFP (pour *alternative reading frame protein* en anglais) est encodée dans un ORF situé dans le cadre de lecture +2 (boîte orange). L'initiation de la traduction de l'ORF d'ARFP a lieu à un codon GUG situé environ 26 codons en aval de l 'AUG de l'ORF principal.

2.2. Chez les eucaryotes

2.2.1. Des ARN messagers matures multi-codants chez les eucaryotes ?

2.2.1.1 Arrangement structural des ARNm matures : vision classique et remise en question du dogme

Un ARNm mature eucaryote, tel que typiquement décrit dans les livres de biologie moléculaire, est monocistronique. Il est constitué de trois régions distinctes : deux régions dites non traduites (UTRs), situées en amont (5'UTR) et en aval (3'UTR) d'une région codante (CDS, pour *coding sequence* en anglais) unique. Le CDS, qui code pour une seule protéine, peut aussi être appelé ORF de référence (RefORF), la protéine qu'il encode étant référencée dans les bases de données nucléotidiques/protéiques (RefSeq ou Ensembl, par exemple) comme l'unique produit protéique issu de cet ARNm (Figure 12). Bien qu'il soit reconnu que les procaryotes et les virus possèdent des ARNm multi-codants, plusieurs caractéristiques du système de traduction des ARNm eucaryotes font qu'historiquement ils ne sont pas reconnus pour partager ce mécanisme de diversité protéique. Les séquences codantes eucaryotes ont été définies selon des règles assez simples (Brent, 2005). Tout d'abord, incluant chez les procaryotes, un *a priori* sur la taille minimale que doit avoir une séquence codante (100 codons) est quasiment toujours utilisé lors de la recherche d'ORFs dans une séquence nucléotidique donnée. Ce seuil minimal, justifié pour limiter les mauvaises prédictions en absence de la puissance de la génomique comparative, devrait être revisité à l'heure où un très grand nombre de génomes sont séquencés et que les ressources bioinformatiques adéquates sont disponibles. Ensuite, pour chaque locus, n'a en général été prédit (manuellement ou de façon automatisée) que le plus grand ORF, commençant si possible par le premier codon AUG rencontré depuis l'extrémité 5' de la région analysée. Ainsi, des ARNm eucaryotes matures encodant plusieurs protéines n'ont été découverts que de façon sporadique et acceptés comme des exceptions (voir paragraphe 3.1).

Pourtant, certaines raisons suggèrent qu'un ARNm eucaryote mature pourrait produire plusieurs groupes de protéoformes à la fois. Le fait que des virus de cellules eucaryotes, qui utilisent la machinerie de traduction de leurs hôtes, possèdent plusieurs ORFs distincts traduits à partir d'un ARNm mature unique indique que ce phénomène pourrait prendre

place lors de la traduction d'ARNm cellulaires. De plus, au vu de l'importance de plus en plus grande que semblent revêtir les protéines de petite taille, il est possible que les méthodes de prédiction d'ORFs utilisées précédemment aient ignorés des ORFs réellement traduits. Les ARNm multi-codants pourraient ainsi ne pas être limités au règne procaryote, mais être également utilisés chez les eucaryotes. Cela permettrait spéculativement une diversité protéique avec un coût énergétique moindre (un ARNm, plusieurs protéines) (Kochetov, 2008).

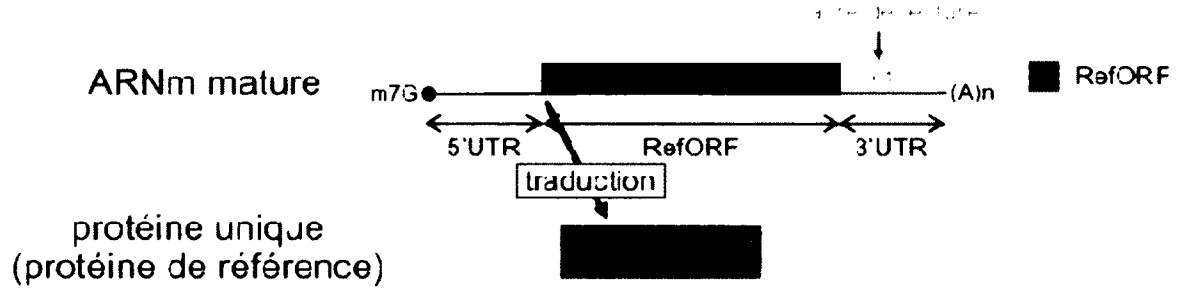


Figure 12. Vision classique de l'arrangement structural des ARNm matures. Dans le modèle classique d'un ARNm mature eucaryote, celui-ci comporte une coiffe à l'extrémité 5', une queue poly-A en 3', et trois régions distinctes (5'UTR, RefORF, 3'UTR). Selon ce modèle, un ARNm eucaryote typique est monocistronique. La seule séquence codante qu'il contient est le RefORF (boîte grise), aussi appelé séquence codante (CDS) associée à l'ARNm dans la base de données RefSeq. Cet ORF code pour la protéine de référence, unique protéine traduite à partir de cet ARNm. Par convention, le cadre de lecture dans lequel est encodé la protéine de référence est le cadre de lecture +1.

2.2.1.2 Diversité protéique procurée par l'initiation alternative de la traduction

Comme mentionné au paragraphe 1.2.4, mis à part le *frameshift* et le *readthrough*, la régulation au niveau de l'initiation de la traduction est un facteur de diversification des produits protéiques issus d'un ARNm mature unique (protéoformes et groupes de protéoformes).

En plus de la protéine codée par le RefORF, d'autres protéoformes de cette même protéine peuvent être formées par l'utilisation sur le même ARNm mature de sites d'initiation alternative de la traduction partageant le même codon stop (et donc le même cadre de lecture) que le RefORF (Figure 13). Ces nouvelles protéoformes partageront la même partie C-terminale que la protéine encodée dans le RefORF (protéine dite de référence), mais elles varieront dans leur portion N-terminale (Kochetov, 2008). Si le codon initiateur alternatif

est en amont de celui du RefORF, des acides aminés sont ajoutés en amont de la séquence correspondant à la protéine de référence. Au contraire, une portion N-terminale est manquante si une initiation de la traduction a lieu en aval du site canonique du RefORF. Les conséquences de la formation d'isoformes allongées ou raccourcies en N-terminal de la protéine de référence peuvent être importantes au niveau fonctionnel, en particulier en donnant naissance à des protéoformes dont la localisation varie (Kazak et al, 2013, Kochetov et al, 2005, Kochetov, 2008). En effet, de nombreux signaux de localisation sont situés dans la partie N-terminale des protéines, tels les peptides signaux pour le ciblage vers l'appareil de sécrétion (Blobel, 1980), ou vers les mitochondries (Chacinska et al, 2009). Par ailleurs, des éléments modulant la fonction ou la liaison à des partenaires d'interaction peuvent également être exclus ou inclus (Calligaris et al, 1995, Lin et al, 1993). Par exemple, le gène suppresseur de tumeur *TP53* code pour le facteur de transcription p53, dont une protéoforme tronquée en N-terminal manque le domaine de transactivation, agissant comme régulateur négatif au cours du cycle cellulaire (Courtois et al, 2002).

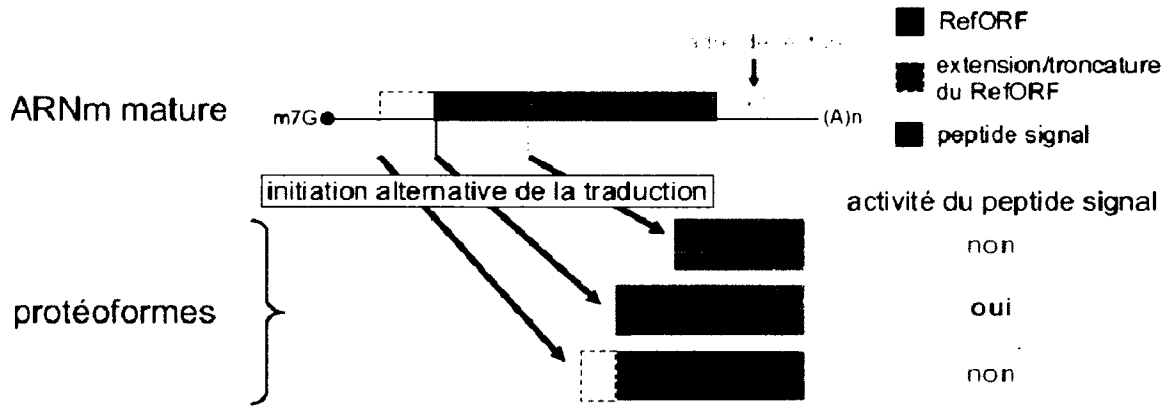


Figure 13. L'initiation alternative de la traduction permet d'augmenter le nombre de protéoformes produites depuis un ARNm mature unique. L'initiation de la traduction peut avoir lieu en amont ou en aval du site d'initiation du RefORF, à des codons initiateurs dits alternatifs (lignes verticales pointillées). Si elle a lieu dans le même cadre de lecture (+1) et que le même codon stop que celui du RefORF est utilisé suite à ces évènements d'initiation alternatifs, alors des isoformes allongées ou raccourcies de la protéine de référence sont traduites. Ces nouvelles protéoformes varient dans leur extrémité N-terminale par rapport à la protéine de référence. Si l'initiation a eu lieu en aval, le peptide signal (en rouge) est absent de la protéoforme, et si elle a eu lieu en amont alors une séquence peptidique précède le peptide signal, l'empêchant d'être reconnu. Dans les deux cas, la localisation finale des protéoformes alternatives va donc différer de celle de la protéoforme de référence. Un seul ARNm mature code donc ici pour plusieurs protéoformes.

De nombreux autres exemples de ce type existent (Kochetov, 2008). Certaines études à large échelle récentes indiquent que la création d'isoformes protéiques étendues ou tronquées par ce mécanisme semble participer de façon non négligeable à la diversité protéique chez les eucaryotes (Fritsch et al, 2012, Ingolia et al, 2011, Kazak et al, 2013). Des prédictions à l'échelle du génome de ce phénomène et de ses conséquences ont été réalisées et une base de données référençant ces informations est déjà disponible (Cai et al, 2005).

L'initiation de la traduction à un site alternatif permet aussi l'expression de groupes de protéoformes issus d'ORFs ne partageant pas le même codon stop que le RefORF. Cela peut être le cas soit si ces ORFs ne chevauchent pas le RefORF, soit s'ils les chevauchent (au moins partiellement) mais sont localisés dans des cadres de lecture différents (Figure 14). Dans les deux cas, nous sommes alors en présence d'ARNm eucaryotes matures capables de diriger la synthèse de protéines avec une séquence en acides aminés distincte de celle de la protéine de référence. **Un terme unificateur pour désigner les ORFs qui ne partagent pas le même codon stop que le RefORF est celui d'ORF alternatifs (AltORFs), qui codent pour des protéines alternatives, distinctes du groupe de protéoformes issues du RefORF.** Dans le cas où il n'y a pas chevauchement entre le ou les AltORFs et le RefORF, la situation est similaire à l'organisation en opéron des ARNm polycistroniques procaryotes. Dans le cas où il y a chevauchement, la situation est plus proche des ORFs chevauchant observés fréquemment chez les virus. Dans le reste du manuscrit, pour des raisons de clarté, nous parlerons donc d'ARNm eucaryote multi-codant pour l'ensemble des deux situations, et d'ARNm eucaryote polycistronique uniquement lorsque celui-ci contient au moins deux ORFs qui ne se chevauchent pas. En résumé, si des AltORFs sont effectivement présents et traduits à partir d'ARNm matures eucaryotes, alors plusieurs groupes de protéoformes sont traduits depuis un ARNm mature unique.

Par la production de multiples protéoformes et groupes de protéoformes à partir d'un seul ARNm mature, l'initiation alternative de la traduction revêt donc un grand potentiel dans sa capacité à générer de la diversité protéique.

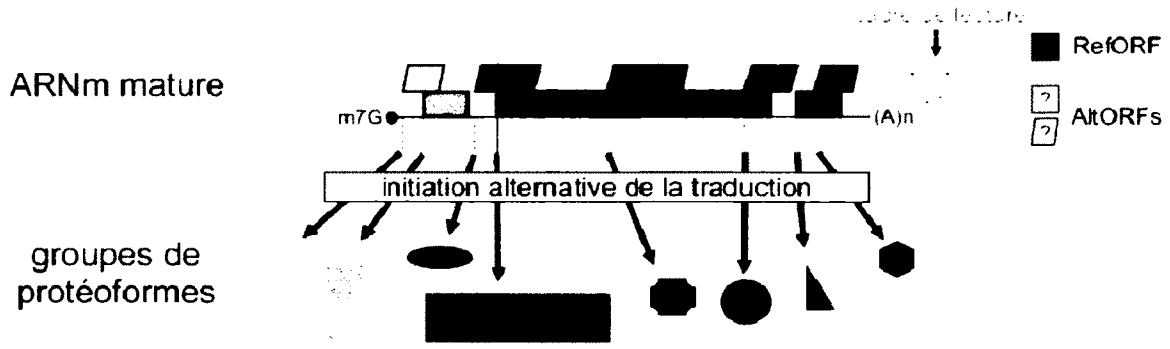


Figure 14. Un ARNm mature unique peut produire plusieurs groupes de protéoformes diffférents par l'utilisation ORF alternatifs (AltORFs). L'initiation alternative de la traduction peut avoir lieu à des codons initiateurs qui ne partagent pas le même codon stop que le RefORF. Cela arrive s'ils se situent dans un cadre de lecture alternatif (+2 ou +3) et/ou si leur codon stop est localisé dans les régions UTRs. Les ORFs définis par ces types de sites d'initiation alternatifs sont appelés ORFs alternatifs (AltORFs, boîtes non grises). Ils codent pour des protéines alternatives (formes non grises) dont la séquence primaire est entièrement distincte de celle de la protéine de référence, portant donc probablement des fonctions qui diffèrent de celle de cette dernière. Ici, à partir d'un ARNm mature unique, plusieurs groupes de protéoformes sont donc possiblement encodés.

2.2.2. Mécanismes d'initiation de la traduction chez les eucaryotes

L'initiation de la traduction peut être résumée comme l'ensemble des étapes qui permettent de placer un ribosome apte à l'élongation au niveau d'un site d'initiation de la traduction. Ici, les mécanismes d'initiation de la traduction eucaryotes seront décrits, ainsi que la manière dont ils permettent l'utilisation de sites d'initiation alternatifs.

2.2.2.1 Mécanisme canonique : le balayage dépendant de la structure coiffe

Contrairement aux ARNm procaryotes, les ARNm eucaryotes portent une modification appelée coiffe, qui consiste en une 7-methylguanosine ajoutée post-transcriptionnellement à l'extrémité 5' de l'ARNm (Shuman, 2002). Il a été démontré que la coiffe stimule la traduction par la machinerie eucaryote (Mitchell et al, 2010, Shatkin, 1976). Elle est reconnue et liée par un facteur d'initiation de la traduction appelé eIF4E, qui interagit directement avec eIF4G (lui-même lié à d'autres facteurs) dans un complexe appelé eIF4F (Figure 15). En parallèle de l'ARNm lié à eIF4F par l'interaction eIF4E-coiffe, les petites sous-unités 40S des ribosomes eucaryotes sont associées à un complexe multifactoriel (MFC), composé entre autres d'eIF3, le facteur d'échafaudage (Hinnebusch, 2006), et d'eIF2, qui amène l'ARNt-Met initiateur (Met-ARNt$_i^{Met}$, permettant la reconnaissance du codon AUG initiateur et l'incorporation du premier résidu méthionine). L'association entre sous-unité 40S et MFC constitue le complexe 43S de pré-initiation (43S PIC). Puisque eIF4G et eIF3 peuvent interagir ensemble, le 43S PIC peut être recruté par eIF4F sur l'ARNm, complétant la première étape de l'initiation de la traduction : le recrutement de la sous-unité 40S sur l'ARNm (Sonenberg & Hinnebusch, 2009). Une fois recruté en 5' de l'ARNm, le 43S PIC va alors procéder à un balayage (*scanning* en anglais), avançant vers l'extrémité 3' de l'ARNm à la recherche d'un codon initiateur AUG. Une fois l'AUG reconnu, la grande sous-unité 60S du ribosome est à son tour recrutée sur l'ARNm, et forme avec la sous-unité 40S le complexe d'initiation 80S final. L'élongation de la traduction peut alors débuter (Sonenberg & Hinnebusch, 2009).

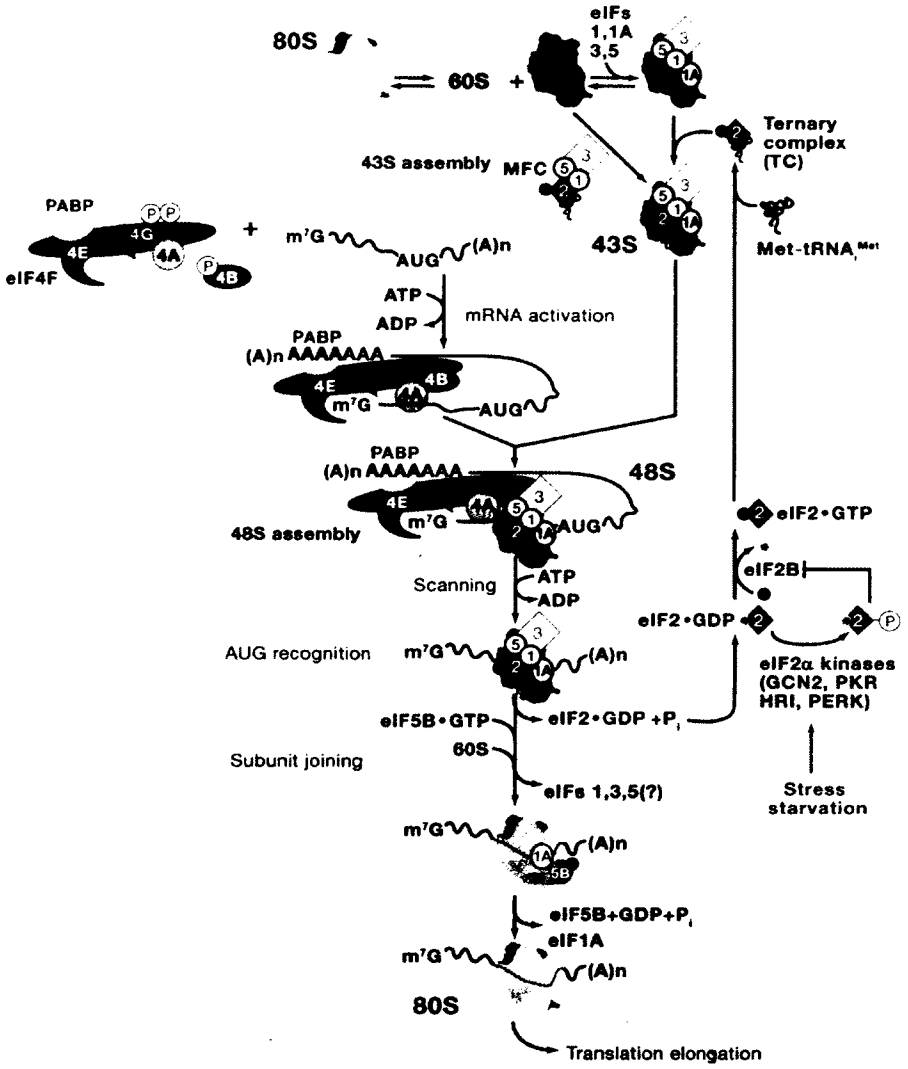


Figure 15. Mécanisme moléculaire de l'initiation de la traduction coiffe-dépendante chez les eucaryotes. Adapté de (Sonenberg & Hinnebusch, 2009). L'initiation de la traduction consiste dans l'ensemble des étapes permettant la formation d'un complexe ribosome/ARNm à un site d'initiation de la traduction apte à procéder à l'élongation de la traduction. Dans le cytoplasme, les petites sous-unités ribosomales 40S libres s'associent à un complexe multifactoriel (MFC) composé entre autres du facteur d'échaffaudage eIF3, et d'eIF2 lié à l'ARNt-Met initiateur (Met-tRNAi^{Met}). Cette association aboutit à la formation du complexe de pré-initiation 43S (43S PIC). D'autre part, les ARNm libres vont voir leur structure coiffe reconnue et liée par le facteur d'initiation eIF4E, qui lie également eIF4G (associé à d'autres facteurs d'initiation). La liaison d'eIF4G à eIF3 permet de recruter sur l'ARNm le 43S PIC (la liaison de eIF4E sur la coiffe favorise donc indirectement le recrutement des ribosomes sur l'ARNm). Dans le complexe 48S ainsi formé, le 43S PIC peut alors procéder au balayage de l'ARNm (*scanning*), de façon à reconnaître le codon d'initiation par appariement entre l'ARNm et le Met-tRNAi^{Met}. La sous-unité 60S rejoint alors la sous-unité 40S pour former un ribosome 80S apte à l'élongation de la traduction.

Le modèle de balayage ribosomal intègre plusieurs paramètres qui semblaient être des normes de l'initiation de la traduction à l'époque : i) la coiffe facilite l'initiation ; ii) l'initiation est restreinte à la région 5' des ARNm ; iii) absence d'une séquence pré-requise fixe (de type SD) ; iv) l'initiation commence au codon AUG le plus proche de l'extrémité 5' (Figure 16 A). La découverte que l'ajout d'un codon AUG en amont du CDS de la préproinsuline diminue l'efficacité d'initiation à l'AUG du RefORF a amené à une complexification du modèle. Suite à l'observation que la suppression du codon stop situé entre les deux ORFs (dans le cadre de lecture du premier AUG) abolit l'expression de la préproinsuline, il a été conclu que les ribosomes reconnaissent le premier AUG, et sont ensuite capables de réinitier au codon AUG du RefORF (Kozak, 1984). La réinitiation semble cependant fonctionner uniquement si certaines règles sont respectées, en particulier que le premier ORF doit être court et que le second AUG ne doit être ni trop proche ni trop éloigné du codon stop de l'ORF précédemment traduit (Kozak, 1987, Kozak, 2001)(Figure 16 B). Cette découverte a fait germer l'idée que l'initiation de la traduction pouvait avoir lieu à partir de plusieurs codons initiateurs sur un même ARNm eucaryote mature.

Par la suite, les facteurs déterminant quel(s) site(s) d'initiation pouvait servir efficacement de site d'initiation de la traduction ont été découverts. En particulier, l'observation des nucléotides flanquant les sites d'initiation du RefORF d'une centaine d'ARNm eucaryotes a mené à la suggestion que le motif [A/G]nnAUGG (A ou G en position -3 par rapport au A de l'AUG, et un G en position +4) semblait être la séquence consensus qui pourrait faciliter la reconnaissance d'un codon AUG par le complexe de balayage (Kozak, 1981) (Figure 16 C), même si d'autres nucléotides peuvent avoir un rôle dans une moindre mesure (Kozak, 1997). Ceci a été confirmé expérimentalement par la suite (Kozak, 1986b), et ce contexte nucléotidique autour d'un codon d'initiation a été appelé « contexte Kozak ». Ainsi, si un AUG se situe dans un contexte suboptimal (i.e. autre que le consensus), il se peut qu'il ne soit pas reconnu comme site d'initiation, et que le ribosome n'initie la traduction qu'au prochain AUG dans un contexte optimal rencontré : il s'agit du balayage en fuite (*leaky scanning* en anglais) (Kozak, 1986a, Kozak, 1986c) (Figure 16 C). Cependant, il a été observé que l'initiation peut avoir lieu à un codon AUG qui n'est pas dans un contexte optimal, et ne pas avoir lieu à un autre qui l'est (Kozak, 1997, Sloan et al, 1999, Stallmeyer et al, 1999). Le contexte Kozak n'est donc pas un déterminant final à la reconnaissance ou

non d'un codon d'initiation. Il a été démontré que l'initiation à un codon situé dans un contexte suboptimal est facilitée par la présence légèrement en aval d'une structure secondaire (type tige-boucle, ou épingle à cheveux) modérément stable. En ralentissant le complexe de balayage dans une position où l'AUG est placé dans ou proche du site de reconnaissance de la sous-unité 40S, cela favoriserait l'appariement avec le Met-ARNt$_i^{Met}$ (Kozak, 1990)(Figure 16 D). Ces deux facteurs (contexte Kozak, structure secondaire en aval) peuvent également moduler la reconnaissance à des codons d'initiation non-AUG, expliquant l'utilisation parfois efficace de ces triplets comme sites d'initiation (Baril & Brakier-Gingras, 2005, Kozak, 1990). Etant donné les multiples facteurs qui le modulent, le balayage est par nature probabilistique : chaque codon initiateur a une probabilité variable d'être reconnu par un ribosome en cours de balayage. Au lieu de ne permettre la reconnaissance que d'un unique site d'initiation à tout coup, le *leaky scanning* permet donc l'utilisation de multiples sites d'initiation de la traduction (Kochetov et al, 2005, Kozak, 1986a).

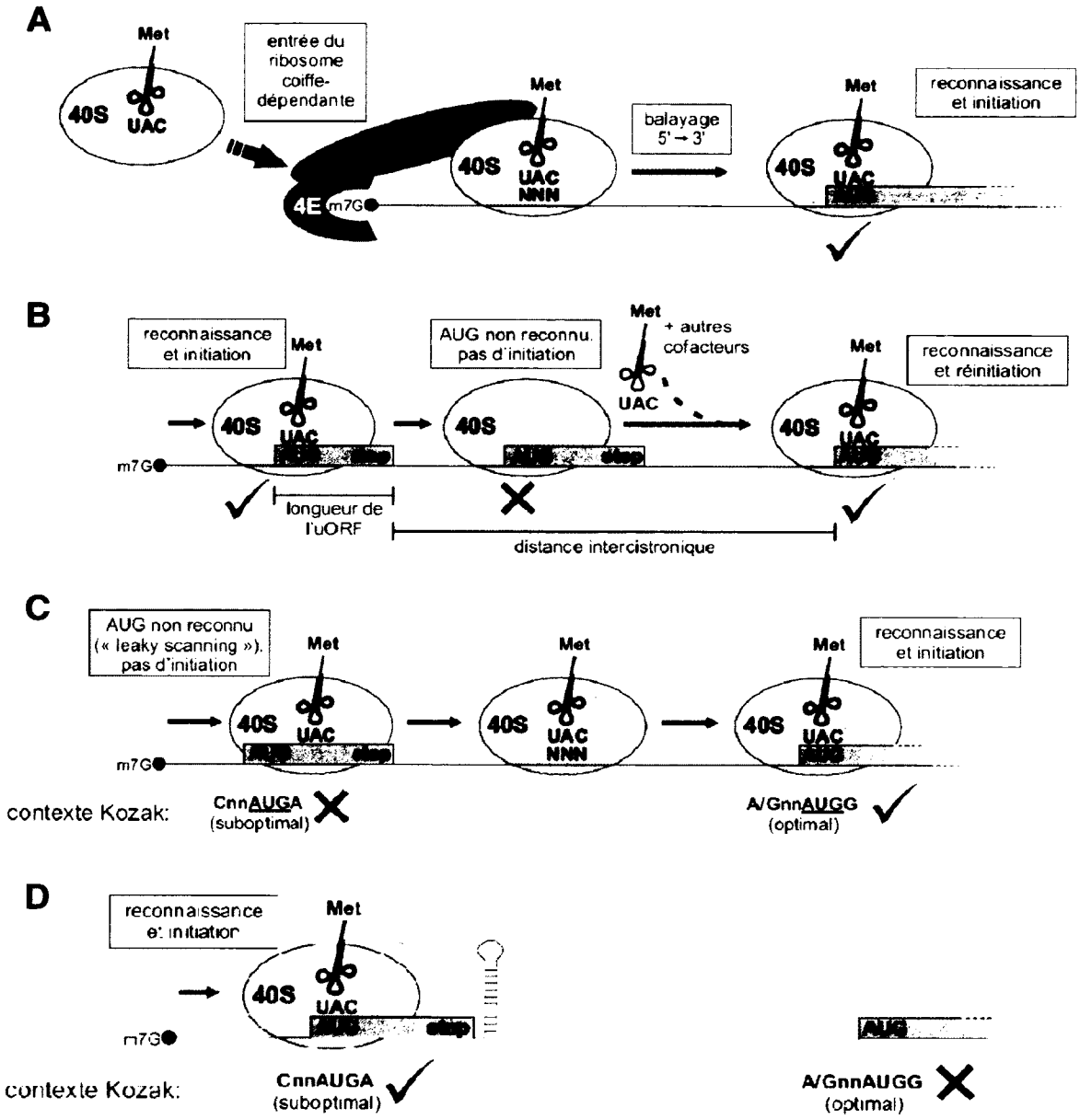


Figure 16. Mécanismes d'initiation de la traduction chez les eucaryotes. (**A**) Dans le mécanisme canonique utilisé par la majorité des ARNm cellulaires, le complexe de pré-initiation 43S (43S PIC, ici représenté par la sous-unité ribosomale 40S liée au Met-ARNt$_i^{Met}$) est recruté à l'extrémité 5' de l'ARNm de façon coiffe dépendante par interaction avec eIF4F (ici eIF4E + eIF4G). S'en suit un balayage vers l'extrémité 3' jusqu'à reconnaissance d'un codon initiateur (ici AUG). A ce point, la sous-unité ribosomale 60S est recrutée et l'élongation peut débuter (non schématisé). (**B**) Après qu'un court uORF ait été traduit, la petite sous-unité 40S peut reprendre le balayage et réinitier la traduction en aval. La distance entre l'uORF et le codon initiateur en aval est importante (ni trop grande ni trop courte), car elle doit permettre la ré-acquisition de certains facteurs nécessaires à l'initiation, en évitant la perte d'autres qui augmente avec la distance à balayer. Ainsi, si l'ORF en aval est trop proche ou trop loin, la traduction de l'uORF a un effet négatif sur la traduction de cet ORF en aval. *(suite en page suivante)*

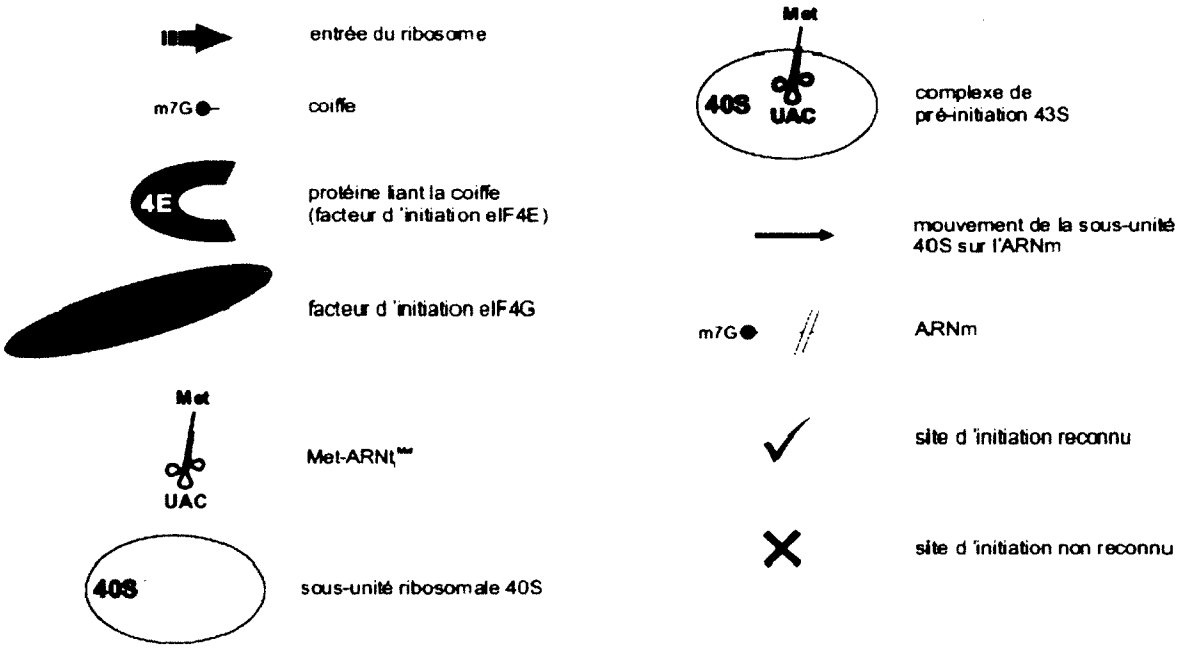


Figure 16. *(suite)* (**C**) Lors du balayage, la séquence nucléotidique autour des codons initiateurs est un facteur important influençant la probabilité de reconnaissance d'un site d'initiation. Cette séquence, appelée contexte Kozak, est optimale lorsqu'elle correspond au consensus (A/GnnAUGG). Si un codon d'initiation n'est pas reconnu lors du balayage, on dit qu'il y a *leaky scanning*, ce qui arrive plus souvent lorsque le contexte Kozak est suboptimal. (**D**) La présence de structures secondaires stables proche d'un site d'initiation et en aval ralentit le balayage, augmentant le temps accordé au 43S PIC pour reconnaître le codon initiateur, et la probabilité que ce site d'initiation soit utilisé. Cela peut compenser pour un contexte Kozak suboptimal, favorisant la traduction du premier ORF mais réduisant celle du second dans le cas présenté.

2.2.2.2 Mécanismes additionnels (non-canoniques) d'initiation de la traduction

Sur certains ARNm, du shuntage ribosomal (*ribosome shunting* en anglais) peut avoir lieu pour éviter de rencontrer certains éléments inhibant le balayage (Figure 17 A). Le shuntage ribosomal permet par exemple, dans le cas de l'ARNm de la protéine Gtx à homéodomaine, à la sous-unité 40S en balayage d'ignorer un codon AUG en amont de celui du RefORF ainsi qu'une structure stable en épingle à cheveux, et de reprendre le balayage en aval de ces éléments. Ils peuvent être ainsi évités, favorisant l'initiation de la traduction au RefORF (Chappell et al, 2006). Dans le cas d'un reovirus aviaire, le shuntage ribosomal permet l'expression de plusieurs ORFs dans un ARNm tricistronique par initiation alternative de la traduction (Racine & Duncan, 2010). La dépendance de ce mécanisme à un appariement entre l'ARNm et l'ARNr 18S indique qu'il pourrait être utilisé de façon assez régulière (Chappell et al, 2006, Yueh & Schneider, 2000). En plus de la réinitiation de la traduction et du *leaky scanning*, le shuntage ribosomal peut donc permettre l'utilisation de plusieurs sites d'initiation sur un ARNm mature unique. Un autre phénomène permet également d'exclure du balayage une région de l'ARNm : l'utilisation d'un site d'entrée interne des ribosomes (IRES, pour *internal ribosome entry site* en anglais). Les IRESs sont des structures secondaires capables de recruter de façon coiffe-indépendante les ribosomes directement à proximité d'un site d'initiation de la traduction, même si celui-ci est situé loin de l'extrémité 5' de l'ARNm (Shatsky et al, 2010, Sonenberg & Hinnebusch, 2009) (Figure 17 B). Bien que les mécanismes de fonctionnement varient d'un IRES à un autre, ils ont tous des points en commun : i) au moins un élément spécifique de leur structure secondaire possède une haute affinité pour un facteur clé du recrutement ribosomal (facteur d'initiation, ou sous-unité ribosomale directement) (Shatsky et al, 2010) ; ii) des éléments structuraux de l'IRES extrêmement spécifiques permettent à la séquence proximale au codon initiateur d'être accommodées dans le canal de liaison à l'ARNm du ribosome (Filbin & Kieft, 2011, Shatsky et al, 2010) ; iii) leur capacité à favoriser l'initiation de la traduction n'est pas dépendante d'une région 5' libre dans l'ARNm, ce qui suggère un caractère facultatif du balayage pour la reconnaissance du site d'initiation (Terenin et al, 2013). Les IRESs sont particulièrement utilisés chez les virus afin de résister à l'inhibition globale de l'initiation de la traduction coiffe-dépendante lors de conditions de stress, comme une infection virale (Shatsky et al, 2010). Malgré de nombreuses études

controversées rapportant la découverte d'un IRES dans des ARNms cellulaires eucaryotes (basée sur l'observation d'ARNms traduits lorsque la traduction coiffe-dépendante est inhibée) (Jang et al, 1988, Pelletier & Sonenberg, 1988, Shatsky et al, 2010), aucun d'entre eux n'a subi ou passé la batterie de tests nécessaires à la conclusion définitive de la présence d'un IRES (Shatsky et al, 2010). Il est ici important de noter qu'initiation coiffe-indépendante n'est pas synonyme d'initiation IRES dépendante (Shatsky et al, 2010), pour les raisons mentionnées ci-dessous.

Un dernier mécanisme d'initiation de traduction, récemment proposé, permet d'apporter une alternative à l'utilisation d'IRES pour qu'un ARNm cellulaire soit efficacement traduit de façon coiffe-indépendante. Ce mécanisme alternatif met en jeu des éléments appelés CITE (pour *cap independent translational enhancer* en anglais) (Figure 17 C-D). Ces éléments sont en particulier utilisés chez les virus de plantes (Dreher & Miller, 2006, Kneller et al, 2006, Miller et al, 2007), et une preuve de concept de leur utilisation dans des cellules humaines a été apportée (Terenin et al, 2013). Ces structures particulières dans les régions 5'UTR (5' CITE) ou 3'UTR (3' CITE) des ARNm sont capables d'interagir spécifiquement avec des facteurs d'initiation de la traduction ou directement avec les sous-unités ribosomales. Dans le cas des 5' CITE, les composants nécessaires à l'initiation de la traduction se lient directement à la région 5' (Figure 17 D), alors que dans le cas des 3' CITE, une liaison coopérative entre les régions 5' et 3' de l'ARNm permet d'amener le complexe d'initiation dans la région 5' (Figure 17 C) (Shatsky et al, 2010). L'initiation médiée par les CITEs est donc coiffe-indépendante, mais est dépendante de la région 5' et invoque ensuite un balayage ribosomal depuis cette extrémité afin de sélectionner le(s) codon(s) initiateur(s) (Terenin et al, 2013). Le potentiel de ce mécanisme coiffe-indépendant pour l'initiation alternative de la traduction existe, de manière similaire au mécanisme coiffe-dépendant traditionnel, vu la spécificité toute relative de la sélection d'un codon initiateur par balayage.

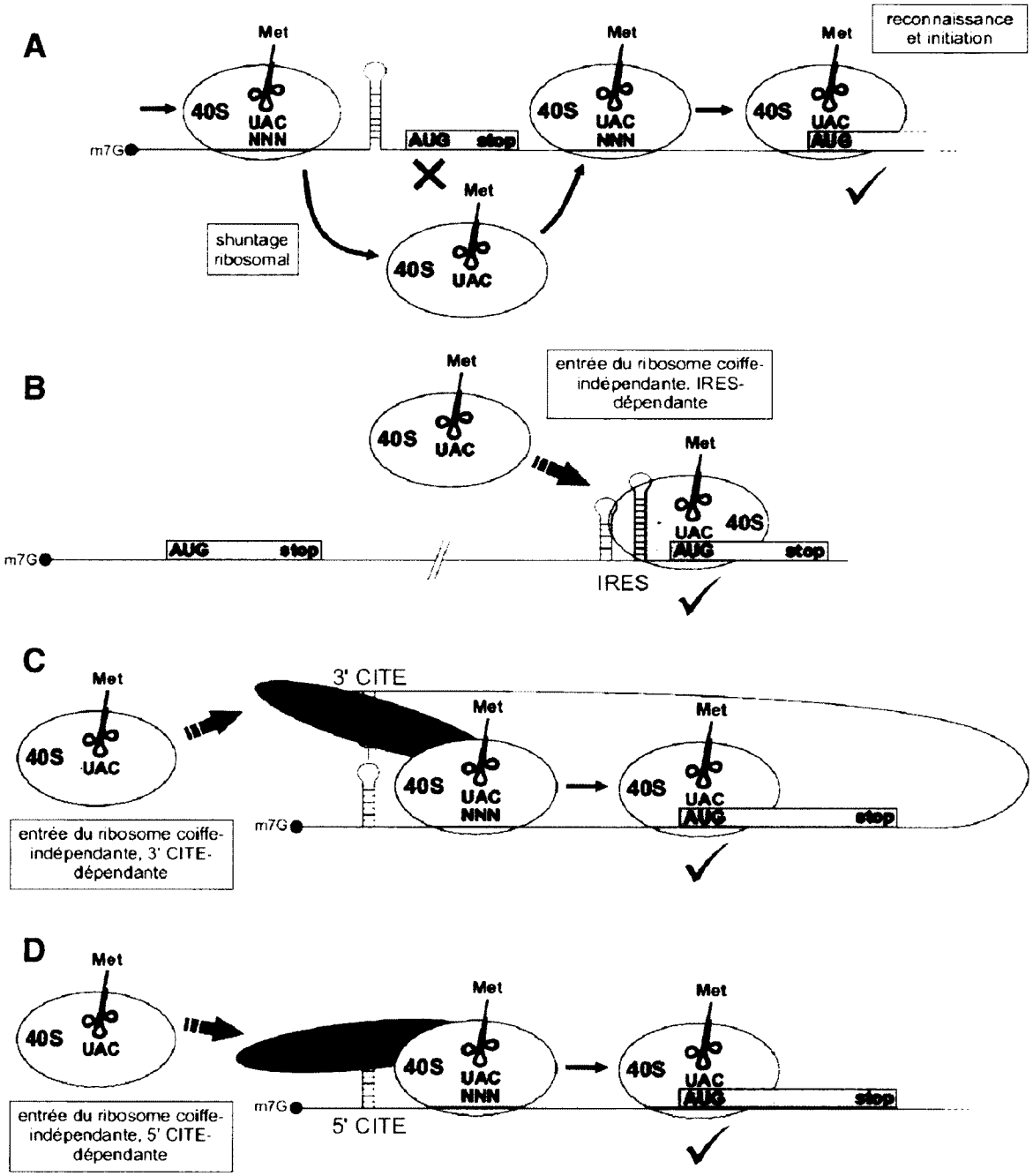


Figure 17. Mécanismes non-canoniques d'initiation de la traduction chez les eucaryotes. (**A**) Le shuntage ribosomal permet à un 43S PIC en cours de balayage de « sauter » par dessus une région contenant des éléments inhibant potentiellement la reconnaissance de l'AUG en aval, tels qu'une structure secondaire stable dans l'ARNm ou un codon AUG en amont, et de reprendre ensuite le balayage. (**B**) Un mécanisme alternatif d'initiation de la traduction est l'utilisation d'IRES (*internal ribosome entry site*), une région structurée capable d'interagir avec des facteurs d'initiation de la traduction ou le ribosome lui-même. Ici le ou les sous-unités ribosomales (variable selon les IRESs ; 43S PIC représenté ici) sont attirées sur l'ARNm à un site interne, à proximité du codon

initiateur. Elles entrent sur l'ARNm à cet endroit, et non nécessairement dans la région 5'. Ce mode d'entrée est indépendant de la coiffe et d'une extrémité 5' libre. Le balayage est ici facultatif, et même rare. (**C**,**D**) Un autre mode alternatif d'entrée des ribosomes est celui mettant en jeu les CITEs (*cap independent translational enhancer*). Ces structures peuvent être présentes dans le 3' UTR (3' CITE, **C**) ou le 5' UTR (5' CITE, **D**) de certains ARNm, et interagir avec des facteurs d'initiation de la traduction comme eIF4G, attirant indirectement vers eux les 43S PIC. L'utilisation de CITEs constitue un mécanisme d'initiation coiffe-indépendant. Mais à la différence du mécanisme IRES-dépendant, si le ribosome est aussi attiré dans une région interne, il va entrer sur l'ARNm dans la région 5', d'où il va débuter le balayage. Les 5' CITEs attirent les ribosomes directement vers la région 5', alors que pour les 3' CITEs ce recrutement dépend d'une interaction entre le 3' CITE et la région 5' de l'ARNm.

3. Evidences expérimentales chez les eucaryotes de l'utilisation d'AltORFs

Comme mentionné précédemment, l'utilisation dans les ARNm cellulaires eucaryotes d'AltORFs permettant la production de multiples protéines à partir d'un seul ARNm mature est un phénomène relativement peu étudié. Cependant, le nombre d'exemples documentés de ce mécanisme est en constante augmentation, et leur découverte s'accélère ces dernières années avec l'avènement de l'ère post-génomique. Dans cette partie, une revue de littérature sera effectuée pour recenser ces découvertes. Tout d'abord, les évènements d'utilisation d'AltORFs découverts de façon isolée seront décrits. Par la suite, j'expliquerai comment les approches à large échelle ont permis de comprendre l'étendue potentielle de ce phénomène.

3.1. Exemples découverts de façon sporadique

Les études décrites ici ont permis la mise en évidence chez les eucaryotes que le RefORF associé à chaque gène ou ARNm n'est pas nécessairement le seul produit protéique qu'ils peuvent générer. Sans définir l'étendue du phénomène, elles amènent à la compréhension de l'importance des implications fonctionnelles qui y sont associées.

3.1.1. Les upstream ORFs *régulateurs*

Les premiers exemples décrits, et les plus nombreux, concernent l'utilisation d'AltORFs dont le site d'initiation est situé en amont du RefORF. Ils sont classiquement appelés *upstream ORFs* ou uORFs (en anglais). Ils ont été découverts par hasard, lors de la quête pour la compréhension du mécanisme permettant la reconnaissance par le ribosome du codon AUG initiateur de la traduction du RefORF. Marilyn Kozak décrivit en 1984 que l'ajout d'un codon AUG en amont du CDS de la préproinsuline avait pour effet de diminuer l'expression de cette protéine (Kozak, 1984). Ce résultat appuya non seulement la théorie naissante du modèle de *scanning* pour l'initiation de la traduction eucaryote (Kozak, 1978), mais démontra également que deux sites d'initiation de la traduction pouvaient être utilisés sur un seul et même ARNm eucaryote. Cette compétition pour la reconnaissance par le ribosome créée par l'AUG en amont et son effet traductionnel sur le RefORF a valu à ces AltORFs le qualificatif d'uORFs régulateurs.

De nombreux cas similaires ont été décrits depuis, marquant une grande variabilité des mécanismes mis en jeu dans l'activité répressive des uORFs. Tout d'abord, l'efficacité de reconnaissance du codon d'initiation de l'uORF (et donc son contexte Kozak) modulent, par compétition, l'utilisation de l'AUG du RefORF (Kozak, 1986b). Dans d'autres cas, c'est la diminution de la capacité du ribosome à réinitier qui réprime la traduction du RefORF. Après avoir traduit une courte séquence, un ribosome est en général capable de reprendre le *scanning* et de réinitier la traduction en aval (Hinnebusch, 2005, Luukkonen et al, 1995, Poyry et al, 2004, Vattem & Wek, 2004). Certains facteurs agissant en *cis* augmentent le temps nécessaire pour traduire un (ou plusieurs) uORF, favorisant la perte de facteurs d'initiation nécessaires pour la réinitiation de la traduction à l'AUG du RefORF (Poyry et al, 2004). Ces facteurs sont la longueur de l'uORF (Luukkonen et al, 1995), la présence de structures secondaires, la longueur de la région intercistronique (Kozak, 1987, Kozak, 2001), l'utilisation de codons rares (Col et al, 2007), ou encore des interactions entre le peptide traduit et la machinerie de traduction (Law et al, 2001, Mize et al, 1998)). De façon similaire, si le codon stop d'un uORF se situe après le codon initiateur du RefORF, la réinitiation est alors impossible, diminuant l'expression du RefORF (Sarrazin et al, 2000, Vattem & Wek, 2004). Parfois, la séquence en acides aminés du peptide codé par l'uORF module la vitesse de traduction par des interactions avec la machinerie de traduction (Law et al, 2001, Mize et al, 1998), ce qui est révélateur que pour un ARNm mature donné, des produits protéiques autres que le RefORF peuvent être non seulement exprimés mais fonctionnels (Figure 18 A). Ceci est encore plus flagrant pour d'autres exemples d'uORFs qui, même exprimés à partir d'un transcrit différent, peuvent réprimer la traduction d'un RefORF en *trans* (Pendleton et al, 2005, Rabadan-Diehl et al, 2007) (Figure 18 B). Ces observations ont été obtenues aussi bien dans des systèmes de traduction *in vitro* que *in cellulo*, bien que le mécanisme moléculaire associé soit méconnu.

Des études computationnelles et génétiques indiquent que 40-50% des ARNm humains et de rongeurs contiennent au moins un uORF (Calvo et al, 2009, Iacono et al, 2005, Matsui et al, 2007), et des approches génétiques et protéomiques indiquent que ces uORFs réduisent l'expression du RefORF de 30 à 80% (Calvo et al, 2009). Les uORFs, bien qu'essentiellement considérés comme répresseurs transcriptionnels en *cis*, constituent une excellente indication que l'utilisation de multiples ORFs à partir d'un seul ARNm mature

pourrait être un mécanisme fréquent. Il est important de noter que des mutations associées à l'apparition, la disparition, ou modifiant la longueur ou la séquence protéique de certains uORFs ont été liées à de nombreuses maladies, comme des formes héréditaires de mélanomes ou de thrombocytémie (Somers et al, 2013).

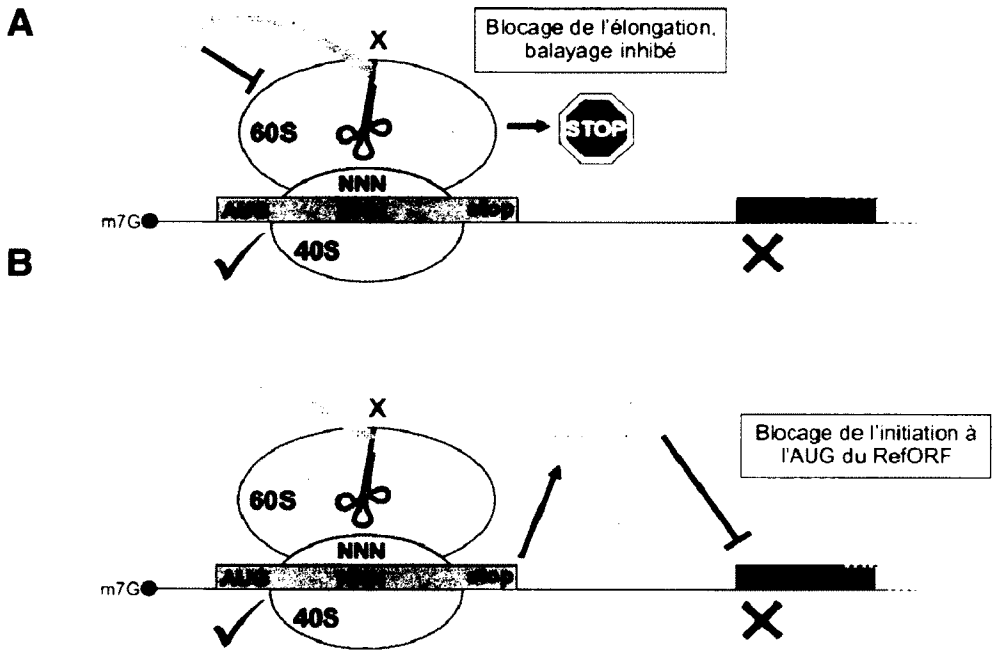


Figure 18. Effet régulateur d'uORFs dépendant de la séquence du peptide encodé. Adapté de (Somers et al, 2013). Dans les deux exemples décrits ici, même courts, les peptides encodés par les uORFs constituent des exemples de protéines alternatives fonctionnelles. (**A**) Certains effets inhibiteurs d'uORFs (boîte orange) sur la traduction du RefORF (boîte bleue) en aval sont médiés par le peptide encodé lui-même, qui interagit en *cis* avec le ribosome en train de le traduire. La séquence particulière du peptide est ici importante, permettant de bloquer l'élongation, et ainsi d'empêcher les ribosomes en cours de balayage situés en amont d'initier la traduction au codon initiateur du RefORF. (**B**) Les séquences primaires particulières de certains peptides leur permettent d'agir en *trans* pour réprimer spécifiquement l'initiation de la traduction à l'AUG du RefORF du gène qui les encode, par un mécanisme restant à caractériser.

3.1.2. Epitopes cryptiques de cellules T

L'utilisation d'AltORFs comme source de diversité protéique a reçu une importante contribution de la part d'études visant à comprendre l'établissement du répertoire peptidique reconnu par les cellules T CD8 cytotoxiques, en particulier dans les cas de cancers. Les cellules T CD8 sont capables d'infiltrer une tumeur, reconnaissant certains antigènes spécifiques aux cellules cancéreuses de cette tumeur. Rosenberg *et al.* ont démontré que l'infusion autologue de telles cellules T CD8, activées par l'ajout concomitant d'interleukine 2, pouvait permettre une régression de la tumeur (Rosenberg et al, 1988). Il a ensuite été démontré que dans certains cas, l'utilisation d'un cadre de lecture alternatif était à l'origine de l'antigène présenté par les cellules tumorales et ciblé par les cellules T CD8 (Ho & Green, 2006, Wang et al, 1996). Indépendamment de l'identité de leur codon initiateur (AUG ou non-AUG), les AltORFs dirigeant l'expression des exemples d'épitopes cryptiques décrits sont issus d'AltORFs inclus dans le RefORF mais traduits à partir du cadre de lecture +2 ou +3. C'est le cas des gènes *TRP-1/gp75* (Wang et al, 1996), *BING-4* (Rosenberg et al, 2002) (mélanomes), *iCE* (Ronsin et al, 1999) (carcinomes rénaux), et *NY-ESO-1* (Wang et al, 1998) (mélanomes et cancer du sein). L'expression d'épitopes cryptiques non désirée (et non remarquée) issus de tels AltORFs peut également poser problème lors d'essais cliniques de thérapie génique, basés sur des vecteurs adénoviraux par exemple. Li *et al.* ont démontré que si un transgène possède un AltORF et que celui-ci est traduit, une réponse cytotoxique peut apparaître, détruisant les cellules transduites et résultant en un échec de la thérapie (Li et al, 2009). Ces exemples démontrent encore que le potentiel multi-codant des ARNm peut avoir une importance particulière dans un contexte physiologique et thérapeutique. L'utilisation d'AltORFs pour élargir le répertoire peptidique possiblement reconnu par le système immunitaire permet à un organisme de se défendre contre un éventail élargi d'agressions (Ho & Green, 2006).

3.1.3. ARN messagers multi-codants

Des AltORFs peuvent-ils avoir des fonctions plus variées que celles décrites précédemment ? Des preuves directes de l'expression endogène à des niveaux détectables de protéines issues d'AltORFs (i.e. autrement que par l'effet traductionnel sur un RefORF, ou par l'activation de cellules immunitaires en contexte pathologique) et de leur importance

physiologique ont permis de répondre à cette question. Les études décrites ici seront présentées en fonction de l'organisation des ORFs dans les ARNm impliqués.

3.1.3.1 ARNm bicistroniques (ORFs non chevauchants)

Le premier ARNm bicistronique stable qui a été décrit chez les eucaryotes est celui exprimant le facteur de croissance/différenciation 1 (GDF-1) ; son caractère bicistronique a été découvert en 1991 chez l'humain et la souris (Lee, 1991) (Figure 19). Le second cistron code pour GDF-1, une protéine sécrétée stimulant la croissance/différentiation cellulaire embryonnaire au niveau du système nerveux. Celui en amont, d'abord baptisé *upstream of* GDF-1 (UOG-1), code pour une protéine de 350 AA chez la souris et a été depuis renommé CERS1, en lien avec son activité céramide synthase au réticulum endoplasmique (Koybasi et al, 2004). Ainsi, ce seul transcrit permet l'expression de deux protéines impliquées dans la croissance cellulaire, la surproduction de céramides issue d'une surexpression de CERS1 diminuant la croissance cellulaire (Figure 19).

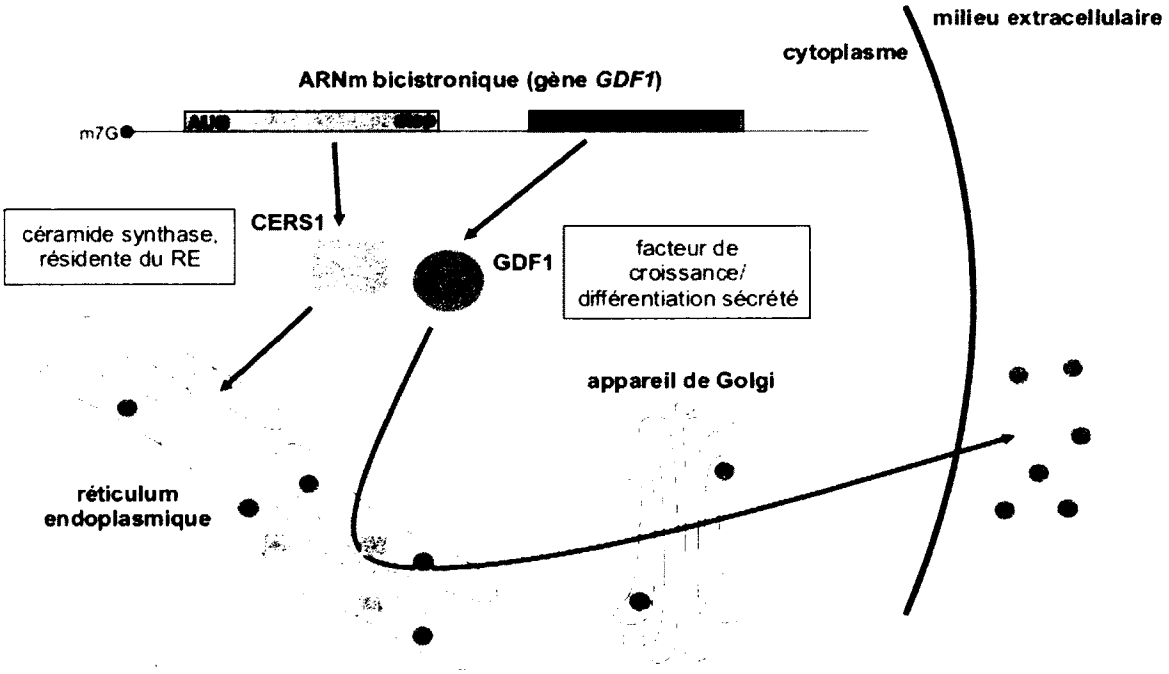


Figure 19. Exemple d'ARNm bicistronique eucaryote. Adapté de (Lee, 1991). L'ARNm codant pour le facteur de croissance/différentiation GDF-1 (en gris) contient un second ORF en amont (en orange), qui permet la production d'une protéine (CERS1) ayant une fonction de céramide synthase. Les deux protéines sont dirigées vers l'appareil de sécrétion. Alors que la première est finalement sécrétée dans le milieu extracellulaire, la seconde réside au réticulum endoplasmique.

D'autres cas d'ARNm bicistroniques ont été découverts par la suite. Chez les mammifères, le locus *Snrpn*, impliqué dans le syndrome de Prader-Willi, est également transcrit en un ARNm bicistronique, l'AltORF identifié codant pour une protéine de 71 AA appelée SNURF (pour SNRPN *upstream reading frame*) localisée au noyau tout comme la protéine de référence SNRPN (Gray et al, 1999). La fonction de cette protéine alternative reste inconnue malgré une implication probable dans la régulation de la prise alimentaire (Naik et al, 2012). Le gène *RPP14*, qui encode une sous-unité de la ribonucléase P humaine, a aussi une structure bicistronique (Autio et al, 2008). L'AltORF en aval a été découvert lors d'un criblage chez la levure d'ADNc humains permettant de complémenter la déficience respiratoire d'une souche mutante pour la 3-hydroxyacyl thioester dehydratase 2 (*htd2*). Tous les ADNc complémentants contenaient deux ORFs : l'un codant pour RPP14, et un autre en aval codant pour une protéine de 168 AA, appelée HsHTD2 chez l'humain. La structure bicistronique de ce gène est conservée chez les vertébrés, du poisson à l'humain, suggérant un lien génétique ancien entre la biogenèse d'acide gras mitochondriaux et la maturation des ARN. Bien qu'une preuve définitive que HsHTD2 est bien traduite depuis un ARNm bicistronique manque, aucun ADNc ne contenant que cet ORF n'a pu être identifié à ce jour (Autio et al, 2008). Des gènes de récepteurs de sucres dans les neurones gustatifs de *D. melanogaster* sont également transcrits en tant qu'ARNm bicistroniques (Slone et al, 2007), tout comme le locus *stoned* important pour la fonction neuronale chez la larve et l'adulte (Andrews et al, 1996), ainsi que l'alcool deshydrogénase (Adh) et une protéine qui lui est homologue (Adhr) (Brogna & Ashburner, 1997). Chez les plantes, ce mécanisme existe également puisque le gène *ENOD40* a été décrit comme codant pour deux peptides de 12 et 24 AA capables de lier la sucrose synthase (Rohrig et al, 2002). L'un de ces deux peptides est même capable de réguler l'activité de clivage du sucrose par cette enzyme (Rohrig et al, 2004). *ENOD40* est le premier représentant découvert d'une nouvelle classe d'ARNs, qui sont à la fois polycistroniques, et codent pour des peptides de taille limitée : les ppcRNAs (polycistronic peptide coding RNAs) (Savard et al, 2006, Tautz, 2009).

3.1.3.2 ARNm polycistroniques avec plus de deux ORFs non chevauchants.

Plusieurs occurrences d'ARNm possédant plus de deux ORFs et dont l'expression des protéines associées a été validée existent. L'exemple le plus édifiant a été découvert

simultanément par trois équipes de recherche, chez le scarabée *Tribolium* et la mouche *D. melanogaster* (Figure 20). Le gène en question, lorsque délété, provoque des défauts développementaux évidents qui lui ont valu les noms de *milles-pattes* (*mlpt*, *Tribolium*) (Savard et al, 2006), *tarsal-less* (*tal*, *D. melanogaster*) (Galindo et al, 2007) et *polished rice* (*pri*, *D. melanogaster*) (Kondo et al, 2007). Conservé au cours de l'évolution chez les insectes, ce gène dirige l'expression d'au moins quatre peptides de 11 à 32 AA à partir d'un ARNm polycistronique (un ppcRNA), probablement par réinitiation de la traduction. Certains de ces peptides régulent le clivage de la protéine Shavenbaby (Svb). Après clivage, Svb passe d'une fonction de répresseur à celle d'activateur transcriptionnel. Les peptides issus de *pri* permettent ainsi le contrôle temporel précis d'une reprogrammation transcriptionnelle primordiale à la morphogenèse épithéliale (Kondo et al, 2010) (Figure 20).

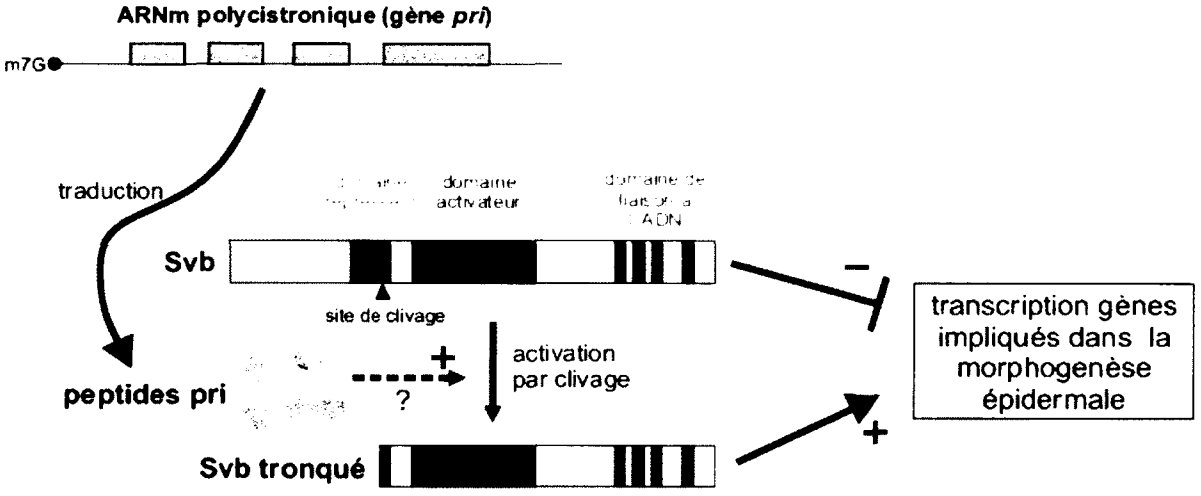


Figure 20. Exemple d'ARNm polycistronique eucaryote. Adapté de (Kondo et al, 2010). Quatre peptides, traduits depuis des ORFs non chevauchants, sont encodés dans le transcrit du gène *pri* chez les insectes. Ils régulent positivement, par un mécanisme moléculaire encore peu caractérisé, le clivage du facteur de transcription Shavenbaby (Svb). Ce clivage aboutit à la perte du domaine répresseur de Svb, qui peut alors activer le programme transcriptionnel nécessaire à la morphogenèse épidermale au cours du développement embryonnaire.

Toujours chez les insectes (ici les mites dont *Bombyx mori*), un ARNm tricistronique a été découvert récemment dans le gène encodant le précurseur du peptide paralytique (PP, famille des peptides ENF) (Kanamori et al, 2010). Il permet l'expression par un mécanisme de *leaky scanning* de trois protéines de 105, 89 et 131 AA (respectivement appelées uENF1, uENF2 et PP/ENF). La taille plus élevée de ces produits de traduction par rapport à ceux des ppcRNAs pourrait définir encore une autre classe d'ARN polycistroniques. Au niveau fonctionnel, PP et uENF1 induisent l'étalement des plasmatocytes, un évènement impliqué dans l'immunité cellulaire chez les insectes. uENF2 régule négativement l'activité des deux premiers, et cette organisation fonctionnelle (peptides impliqués dans la même voie codés par un même transcrit) n'est pas sans rappeler celle des opérons procaryotes. Des ARNm polycistroniques codant pour plus de deux protéines différentes existent aussi chez les mammifères. Il a récemment été montré que le gène *MKKS* produisait deux ARNm par l'utilisation d'un site alternatif de polyadénylation situé en amont du RefORF. La forme longue d'ARNm code pour la protéine MKKS ainsi que pour au moins trois uORFs régulateurs localisés dans le 5'UTR. La forme courte permet l'expression des uORFs seulement. Ces deux ARNm sont donc polycistroniques, avec deux des trois uORFs qui se chevauchent (Akimoto et al, 2013). Les produits de traduction d'au moins deux uORFs (nommés uMKKS1 et uMKKS2) sont exprimés à des niveaux détectables par western blot et conservés chez les mammifères. Bien que les fonctions de uMKKS1 et uMKKS2 soient inconnues, ces deux protéines de 63 et 50 AA sont localisés aux mitochondries, à la différence de MKKS qui fait la navette entre cytosol et centrosome. Contrairement à d'autres exemples cités précédemment où plusieurs ORFs codés par le même transcrit agissaient dans la même voie (ex. gène du PP chez les insectes) (Kanamori et al, 2010), le couplage fonctionnel entre les AltORFs et le RefORF semble ici moins évident.

3.1.3.3 Utilisation de cadres de lecture alternatifs (ORFs chevauchant le RefORF)

Pour les exemples répertoriés ci-après, les AltORFs chevauchent le RefORF, et utilisent donc des cadres de lecture alternatifs +2 ou +3 dans une séquence de l'ARNm qui est doublement codante. Trois cas de figures sont alors possibles : **i)** le codon initiateur alternatif est dans le 5'UTR et le codon stop dans le RefORF (c'est donc un uORF chevauchant) ; **ii)** le codon initiateur et le codon stop sont dans le RefORF (AltORF inclus dans le RefORF) ; **iii)** le codon initiateur est dans le RefORF, et le stop dans le 3'UTR.

i) Pour les uORFs régulateurs chevauchant cités précédemment (Sarrazin et al, 2000, Vattem & Wek, 2004), la preuve de l'expression au niveau protéique est indirecte (modulation de l'expression du RefORF). En revanche, des essais de traduction *in vitro* ont démontré l'expression (par *leaky scanning* pour l'ORF en aval) de deux ORFs à partir d'un même ARNm mature issu du gène *MOCS2* chez les mammifères (Stallmeyer et al, 1999). Les protéines issues de ces deux ORFs, MOCS2A et MOCS2B, constituent les deux sous-unités de l'hétérodimère portant la fonction de molybdopterine synthase. Cette activité, conservée des bactéries à l'humain, est nécessaire à l'activité de toutes les molybdoenzymes (ex : sulphite oxidase). Les deux ORFs se chevauchent sur 77 nucléotides, codant pour des protéines de 88 (MOCS2A) et 188 (MOCS2B) AA respectivement. Dans des études sur deux autres gènes, deux produits protéiques d'uORFs régulateurs chevauchants ont été détectés, sans qu'une fonction physiologique autre qu'une compétition traductionnelle avec le RefORF ne leur soit attribuée (Diba et al, 2001, Hernandez-Sanchez et al, 2003). Inclus dans la région génomique (exons 1 à 3) de l'oncogène *rab34*, un uORF codant pour une protéine de 198 AA a été identifié par analyse par spectrométrie de masse de protéines nucléaires hautement chargées contre une banque d'ESTs (expressed sequence tags) (Zougman et al, 2011). Appelée NARR (pour Nine Amino acid Residue Repeat), cette protéine est très conservée chez les mammifères, exprimée de façon ubiquitaire, et localisée aux nucléoles. Elle y interagit avec d'autres protéines nucléolaires, mais la fonction de ce complexe reste inconnue à ce jour. Au niveau du mécanisme d'expression, il n'a pas été déterminé si Rab34 et NARR étaient traduites à partir du même ARNm, bien qu'aucune donnée ne suggère l'existence de deux ARNm différents.

ii) Comme mentionné précédemment (voir paragraphe 2.1.2), le fait que des ARNms viraux eucaryotes permettent l'expression d'AltORFs inclus dans un cadre de lecture alternatif du RefORF indique que la machinerie de traduction eucaryote est capable de les prendre en compte. Les épitopes cryptiques de cellules T en sont déjà une preuve, mais ce mécanisme est-il utilisé en conditions non-pathologiques ? Une étude précurseur de Klemke *et al.*, en 2001 a démontré que le cadre de lecture +2 de l'exon 1 du gène *GNAS* chez les mammifères sert à la traduction d'une seconde protéine à partir du même ARNm exprimant le RefORF (Klemke et al, 2001) (Figure 21 A). Le RefORF de *GNAS* code pour

la protéine XLαs, une sous-unité α neuroendocrine spécifique d'une protéine G. L'exon 1 de GNAS encode le domaine XL de XLαs dans le cadre de lecture +1, et le cadre de lecture +2 contient un AltORF codant pour une protéine appelée ALEX (alternative gene product encoded by the XL-exon). ALEX est coexprimée avec XLαs à partir d'un transcrit unique lors d'essais de traduction *in vitro*, mais aussi de façon endogène dans des lignées cellulaires et tissus neuroendocriniens. Il est remarquable de noter qu'ALEX interagit avec le domaine XL de XLαs au niveau de la face interne de la membrane cytoplasmique (Klemke et al, 2001), régulant négativement l'activation de XLαs et la formation d'AMPc subséquente (Freson et al, 2003) (Figure 21 B). Des mutations dans la séquence encodant les deux protéines amènent à une diminution de cette interaction, et à une augmentation de la signalisation associée. Les individus touchés par ces mutations présentent des tendances hémophiles, des désordres neurologiques, et de la brachydactylie (Freson et al, 2003). Ceci souligne l'importance fonctionnelle en conditions physiologiques et pathologiques d'un AltORF inclus dans un cadre de lecture alternatif d'un RefORF. Récemment, une correction a été apportée sur la longueur de l'exon 1 du gène *GNAS*, et des formes plus longues de XLαs et ALEX ont été identifiées (700 AA chez l'humain pour ALEX) (Abramowitz et al, 2004).

Au meilleur de ma connaissance, depuis la découverte d'ALEX, en dehors des épitopes crytiques de cellules T et des découvertes accomplies grâce aux travaux présentés dans cette thèse, le seul exemple d'initiation alternative de la traduction dans un AltORF inclus dans un cadre de lecture alternatif du RefORF concerne le gène *RMD1* chez *S. cerevisiae* (Ben-Yehezkel et al, 2013). Sans détecter la protéine correspondante au niveau endogène, cette étude décrit la mise au point d'un élégant système rapporteur pour étudier l'utilisation de multiples sites d'initiation de la traduction dans les régions entourant l'ATG du RefORF (±150 nt). Brièvement, le RefORF ou les AltORFs sont successivement fusionnés à une protéine fluorescente, produisant un système rapporteur quantitatif de l'efficacité d'initiation à un codon initiateur donné. Les auteurs ont ainsi démontré qu'environ 3.3% des évènements d'initiation dans *RMD1* correspondaient à des isoformes d'un même AltORF inclus dans le RefORF. Cela indique pour la première fois que ce phénomène pourrait également être utilisé chez la levure, ce qui soutient son importance fonctionnelle.

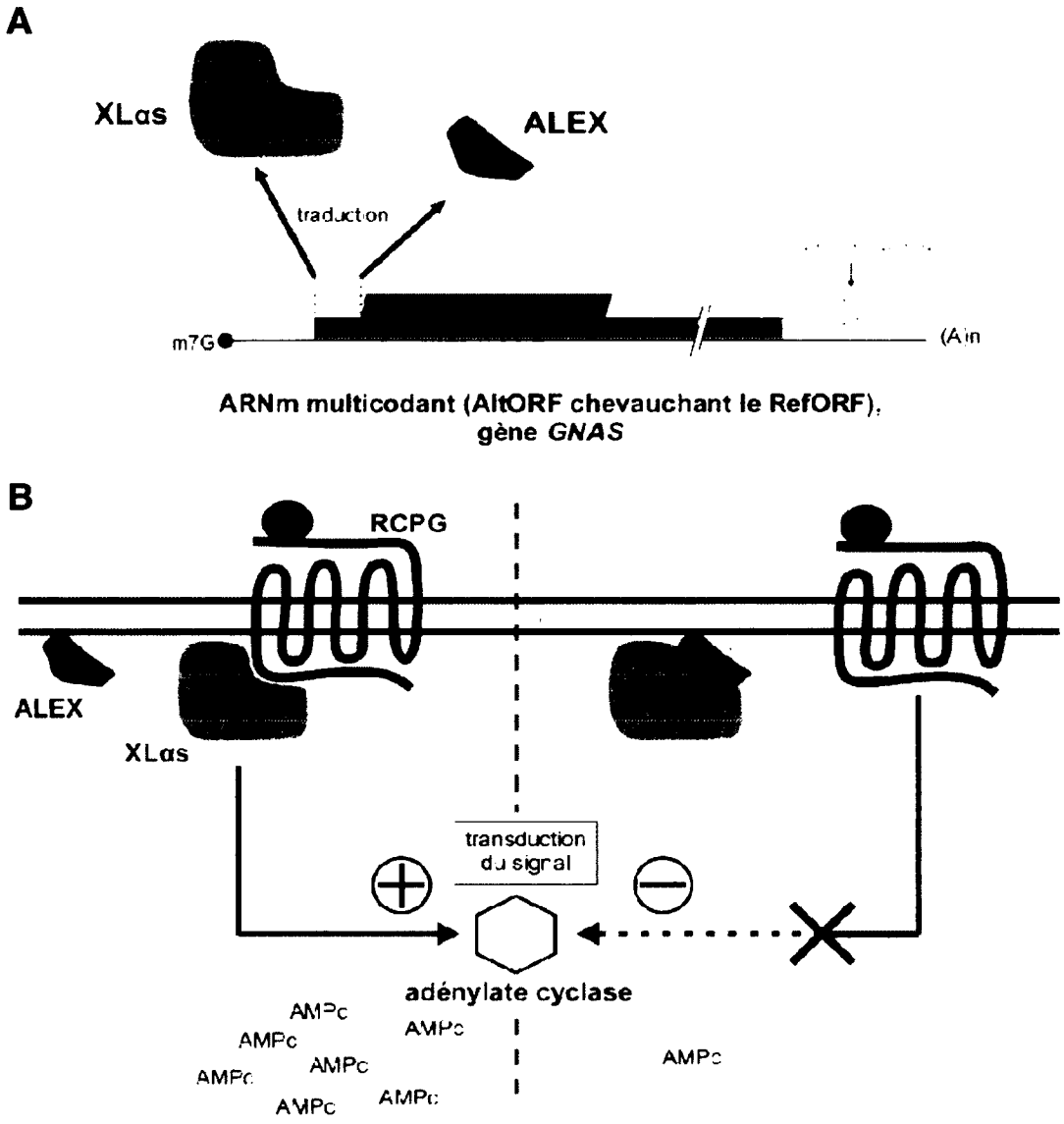


Figure 21. Un AltORF chevauchant le RefORF dans un cadre de lecture alternatif : exemple du gène *GNAS*. (**A**) Adapté de (Klemke et al, 2001). L'exemple le mieux caractérisé d'AltORF chevauchant le RefORF dans un cadre de lecture alternatif chez les eucaryotes est celui du gène *GNAS*. Le RefORF (gris) de ce gène code pour une sous-unité α (Xlαs) d'une protéine G. Dans le cadre de lecture +2, la traduction depuis le même ARNm d'un AltORF chevauchant le RefORF permet la synthèse de la protéine alternative ALEX (vert). (**B**) Adapté de (Freson et al, 2003). En plus d'être encodée dans le même ARNm, Xlαs et ALEX sont couplées fonctionnellement. A gauche: lors de la liaison d'un ligand à un récepteur couplé aux protéines G (RCPG), la protéine XLαs participe à la transduction du signal qui résulte en une activation de l'adénylate cyclase et en une augmentation des niveaux du second messager, l'AMPc. A droite : l'interaction directe entre Xlαs et ALEX empêche la liaison de Xlαs sur le RCPG. Ceci résulte en une diminution de la transduction du signal.

iii) Concernant le troisième type d'AltORFs chevauchant le RefORF dans un cadre de lecture alternatif (site d'initiation inclus dans le RefORF, codon stop dans le 3'UTR), aucun exemple n'a été décrit à ce jour. Cependant, cette configuration n'est pas fondamentalement différente de celle d'un ARNm contenant un uORF chevauchant, et l'on peut considérer par exemple que si l'ORF codant pour NARR (Zougman et al, 2011) avait été découvert avant celui codant pour Rab34 (voir **i)** ci-dessus), alors ce dernier aurait été un exemple de ce cas de figure. Par ailleurs, bien que les codons initiateurs de tels ORFs pourraient être éloignés de l'extrémité 5' du transcrit qui les contient, certains cas d'ARNm bicistroniques discutés précédemment (Gray et al, 1999, Lee, 1991) suggèrent que cela n'empêche pas nécessairement l'expression du cistron distal associé.

3.2. Approches à large échelle

Suite aux découvertes d'exemples isolés d'ARNm à potentiel multicodant chez les eucaryotes s'est posée la question de l'étendue de l'utilisation du phénomène à l'échelle du génome et du transcriptome. Tout d'abord, il a été tenté, par des approches par bioinformatique, de prédire le potentiel d'utilisation de multiples ORFs dans des transcrits matures uniques. Plus récemment, une approche appelée *ribosome profiling* (en anglais) a été mise au point, permettant d'offrir une vision globale de la traduction (et de l'étape d'initiation) de milliers d'ARNm simultanément dans un même échantillon. Enfin, un complément indispensable à ces approches étant la détection directe des produits protéiques des AltORFS, un nombre croissant d'études utilisant des techniques protéomiques (spectrométrie de masse) a vu le jour, commençant à dépeindre la contribution des AltORFs aux protéomes eucaryotes.

3.2.1. Approches in silico

Des études bioinformatiques ont prédit que 40 à 50% des ARNm humains ou de rongeurs possèdent au moins un uORF dans leur 5'UTR. En utilisant une banque d'ARNm suffisamment représentative et complète au niveau des extrémités 5', et avec un codon initiateur du RefORF correctement annoté, Iacono *et al.* ont prédit un uORF ou un AUG en amont du RefORF dans 44% des ARNm humains et 42% des ARNm murins (Iacono et al, 2005). Une autre analyse en parallèle chez l'humain et la souris a aussi permis de prédire la

présence de tels éléments régulateurs dans environ la moitié des transcrits pris en compte (Matsui et al, 2007). Un quart de ces uORFs code pour des peptides d'au moins 20 AA, et plus de 200 d'entre eux sont conservés entre l'humain et la souris, indiquant un potentiel fonctionnel pour ces produits protéiques (Crowe et al, 2006). Ceci est appuyé par une sélection au niveau peptidique sous-tendue par un biais très significatif vers des mutations synonymes. Devant l'augmentation du nombre d'exemples d'utilisation de sites d'initiation de la traduction non-AUG dans les 5'UTR validés expérimentalement, un algorithme a été mis au point permettant de prédire lesquels de ces codons sont effectivement utilisés pour initier la traduction (Wegrzyn et al, 2008). L'implémentation de nombreux critères (contexte Kozak, longueur du 5'UTR, de l'uORF, structures secondaires, AUGs en amont, entre autres) a permis de mettre sur pied une méthode de prédiction robuste des uORFs réellement sujets à l'initiation de la traduction.

Des approches *in silico* par génomique comparative ont été également appliquées à la prédiction d'AltORFs chevauchant (ou entièrement inclus dans) le RefORF dans un cadre de lecture alternatif. Chung *et al.* ont été les premiers à tenter de prédire dans une banque d'ARNm matures l'utilisation de cadres de lectures alternatifs conservés entre la souris et l'humain, plus soit le rat soit le chien. Afin de négliger l'apparition d'une région double-codante par chance, une taille minimale de 500 nt a permis de filtrer les résultats, fournissant une liste de 40 candidats de haute confiance (Chung et al, 2007). Une autre étude a montré, à l'aide d'un système rapporteur contenant un AltORF *out-of-frame* initié en aval de l'AUG du RefORF, des critères de contexte de codon d'initiation alternatif et de référence favorables à la traduction de l'AltORF. En prenant en compte la conservation entre l'humain et la souris ainsi qu'une taille minimale de 500 nt, 138 candidats ont finalement été retenus (Xu et al, 2010). La dernière étude de ce type a identifié 1793 AltORFs d'un minimum de 150 nt conservés entre l'homme, la souris et le rat. Une augmentation de la stringence de prédiction de candidats possiblement exprimés, en ajoutant un filtre ne conservant que les candidats ayant un contexte Kozak optimal, a mené à une liste de 217 AltORFs putatifs (Ribrioux et al, 2008).

Bien que ces études prédictives indiquent que plusieurs centaines d'AltORFs inclus dans le RefORF mais dans un cadre de lecture alternatif puissent être conservés et exprimés chez les mammifères, plusieurs limites y sont associées. Ceci est illustré par l'absence dans les

candidats prédits par l'ensemble de ces études d'exemples validés expérimentalement comme ALEX dans le gène *GNAS*. Plusieurs explications à ce constat peuvent être apportées. Tout d'abord, face à l'importance fonctionnelle possible des protéines de petite taille, et à leur contribution sous-estimée au protéome (Frith et al, 2006), les limites de taille d'AltORFs de 500 nt et 150 nt utilisées semblent trop restrictives. Par ailleurs, la conservation entre espèces de mammifères, bien qu'indicatif de fonctionnalité, ne permet pas d'exclure la présence d'AltORFs à importance biologique dans un groupe plus restreint du vivant, comme les primates par exemple. Il existe d'ailleurs des gènes primates spécifiques (Tay et al, 2009), et ceci pourrait être applicable aux AltORFs. La fonctionnalité supposée par la conservation évolutive ne garantit pas non plus qu'un AltORF soit exprimé. D'ailleurs, les épitopes cryptiques de cellules T n'ont pas été associés à des fonctions biologiques en condition physiologique, mais leur expression n'en est pas moins importante dans un contexte pathologique (Ho & Green, 2006). L'utilisation d'un critère conservant les candidats à contexte Kozak optimal seulement est une autre limitation, puisque l'exemple d'ALEX indique qu'un tel AltORF peut être exprimé malgré un contexte Kozak suboptimal (Klemke et al, 2001). En se limitant aux mammifères, les prédictions effectuées n'analysent pas non plus l'étendue possible du phénomène dans des espèces eucaryotes plus primitives, bien qu'il semble avoir lieu même chez *S. cerevisiae* (Ben-Yehezkel et al, 2013).

3.2.2. Approches par ribosome profiling

Récemment, des avancées dans les techniques de séquençage profond ont rendu possible la lecture en parallèle de gigantesques quantités de séquences courtes d'ADN (Bentley et al, 2008). Ingolia *et al.* en ont tiré profit en développant une nouvelle technique appelée *ribosome profiling* (Ingolia et al, 2009). Brièvement, la méthode repose sur le séquençage à haut débit, après transcription inverse, de fragments d'ARNm protégés par les ribosomes. En normalisant les lectures aux quantités de chaque ARNm dans l'échantillon (obtenu par séquençage d'ARN classique (RNA-SEQ)), il est ainsi possible de connaître les régions du transcriptome qui sont activement traduites et avec quelle efficacité. La résolution au niveau nucléotidique permet de connaître le cadre de lecture traduit, et l'utilisation de drogues figeant les ribosomes aux sites d'initiation amène à l'identification des codons

utilisés pour démarrer la synthèse protéique (Ingolia et al, 2011, Lee et al, 2012). En utilisant les avantages fournis par le *ribosome profiling*, plusieurs études ont mené à l'élaboration d'un profil global de la traduction chez les eucaryotes, donnant des informations de première importance sur l'étendue de l'utilisation d'AltORFs chez les eucaryotes.

Tout d'abord, il a été déterminé que dans au moins la moitié des cas (49.6% dans des cellules humaines HEK 293 ; 65% dans des cellules souches embyronnaires (CSE) de souris), deux sites d'initiation de la traduction différents étaient utilisés au minimum par ARNm (Ingolia et al, 2011, Lee et al, 2012). Dans les CSE murines, 16% des transcrits analysés possédaient même quatre ou plus sites d'initiation utilisés (Figure 22 A). Ces résultats suggèrent une utilisation très fréquente à l'échelle du transcriptome de l'initiation alternative de la traduction, même en conditions physiologiques. L'étude d'Ingolia *et al.* en 2011 (Ingolia et al, 2011) nous apprend également que dans les CSE murines, les RefORFs semblent ne représenter qu'environ 28% du total des ORFs traduits, et les isoformes allongées ou tronquées du RefORF environ 14%. Ceci implique que plus de la moitié des codons initiateurs utilisés correspondent à des AltORFs (~40% d'uORFs chevauchants ou non, 16% d'AltORFs inclus dans le RefORF mais dans un cadre de lecture alternatif, et ~2% d'AltORFs dans les 3'UTR) (Figure 22 B). La fréquence d'utilisation de l'initiation alternative de la traduction dans les 5'UTR prédite par des méthodes computationnelles (~50%) a été confirmée par le *ribosome profiling*, puisque cela a lieu pour 54% des ARNms de cellules humaines HEK 293 (Lee et al, 2012). L'application spécifique du *ribosome profiling* à l'identification de sites d'initiation en amont du RefORF a mené à l'identification de 4400 uORFs dans 5062 transcrits d'une lignée monocytaire humaine (Fritsch et al, 2012). L'utilisation d'uORFs chez *S. cerevisiae* a également été observée pour 1800 gènes, soit environ 30% des gènes codant pour des protéines chez la levure (Gerashchenko et al, 2012). Une autre information importante apportée par le *ribosome profiling* concerne les triplets utilisés pour l'initiation de la traduction. Ainsi, d'un point de vue global, le codon AUG est le plus utilisé (~50%), puis le codon CUG (~16%), et le reste des codons qui diffèrent d'AUG par un seul nucléotide (~24%). Il est intéressant de noter que pour les sites alternatifs d'initiation, ces proportions varient selon que l'on considère les sites d'initiation alternatifs en amont ou à l'intérieur du RefORF. Les AUG sont

favorisés à l'intérieur du RefORF, alors que les CUG sont les plus représentés en amont (bien que les AUG y représentent encore un quart des évènements d'initiation) (Lee et al, 2012). Ainsi, même s'il reste justifié d'utiliser le codon AUG pour prédire les ORFs (au moins chez les mammifères), d'autres codons pourraient aussi être pris en compte afin de compléter les prédictions. Enfin, Michel *et al.* ont utilisé l'avantage de la précision au nucléotide près du *ribosome profiling* pour identifier des régions candidates subissant une traduction dans deux cadres de lectures différents (Michel et al, 2012). Outre des cas de *frameshift* ou de changement de cadre dans des variants de transcrits obtenu par initiation de transcription ou épissage alternatifs, 29 uORFs chevauchants et 15 AltORFs inclus dans un cadre de lecture alternatif du RefORF ont été identifiés.

L'énorme avantage du *ribosome profiling* est sa profondeur d'analyse, indépendamment de la séquence traduite, alors que les techniques de protéomique quantitative sont peu adaptées à la détection de protéines de petite taille ou peu abondantes (Lubec & Afjehi-Sadat, 2007), ou inconnues. De plus, en mesurant l'efficacité traductionnelle, le *ribosome profiling* surmonte une partie du problème du manque de corrélation entre niveau d'expression d'un ARNm et de la protéine associée (Sonenberg & Hinnebusch, 2007) que d'autres techniques (puces à ADN, séquençage d'ARN) ne peuvent prendre en compte. Il met d'ailleurs en évidence l'étendue des régulations traductionnelles auxquelles sont sujets les ARNm eucaryotes. Mais le *ribosome profiling* ne fournit pas une preuve directe de l'expression à des niveaux détectables de protéines issues d'AltORFs.

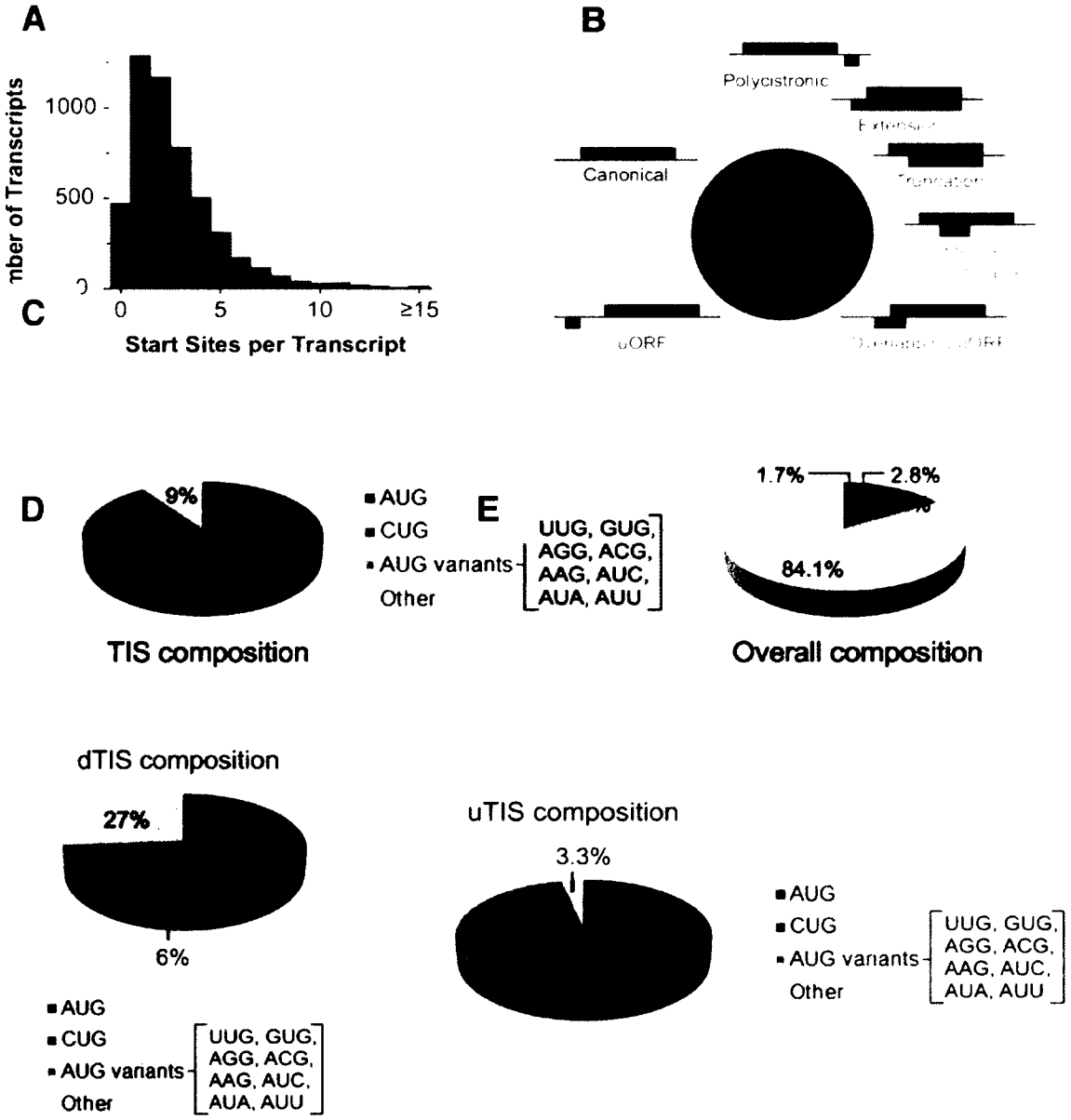


Figure 22. Données sur l'utilisation des sites d'initiation de la traduction (TIS) issues d'études par *ribosome profiling* dans des cellules de mammifères. (**A**,**B**) Adapté de (Ingolia et al, 2011). (**A**) Histogramme du nombre de TIS utilisés par transcrit dans des cellules souches embryonnaires murines. (**B**) Diagramme montrant la distribution des ORFs correspondant aux TIS identifiés dans les mêmes cellules. Noter que les RefORFs (*canonical*) constituent à peine plus d'un quart de l'ensemble des ORFs exprimés. (**C**,**D**,**E**) Adapté de (Lee et al, 2012). (**C**) Utilisation des codons aux TIS identifiés dans des cellules humaines HEK 293, et comparaison à la distribution des codons dans l'ensemble du transcriptome. (**D**) Utilisation des codons aux TIS situés en aval du TIS annoté du RefORF. (**E**) Utilisation des codons aux TIS situés en amont du TIS annoté du RefORF.

3.2.3. Approches protéomiques

En plus de l'intensité de traduction, mesurée par le *ribosome profiling*, la vitesse de dégradation conditionne le niveau final d'expression d'une protéine. Les approches de protéomique quantitative, malgré leurs limites de sensibilité (Lubec & Afjehi-Sadat, 2007) (en constante amélioration par ailleurs), restent donc primordiales pour valider expérimentalement à large échelle l'expression des protéines issues d'AltORFs et pour comprendre leur contribution au protéome. En 2005, Oyama *et al.* ont utilisé la chromatographie liquide couplée à de la spectrométrie de masse en tandem (LC-MS/MS) pour rechercher de nouveaux ORFs courts à partir de séquences complètes de cDNAs de cellules leucémiques humaines K562. 54 protéines de moins de 100 AA ont été détectées, incluant 4 nouvelles protéines toutes traduites à partir d'uORFs. Les auteurs concluent que des ORFs courts peuvent être exprimés *in vivo* même lorsqu'un autre ORF plus long est présent sur le même ARNm (Oyama et al, 2004). La même équipe a identifié dans des cellules humaines (HEK 293 et K562), parmi les protéines de moins de 20 kDa, 8 nouvelles protéines en plus de 197 protéines déjà annotées. Six d'entre elles sont issues d'AltORFs (1 uORF non chevauchant, 3 uORFs chevauchants, un AltORF chevauchant le RefORF et le 3'UTR dans un cadre de lecture alternatif, et un AltORF non chevauchant en aval du RefORF), par initiation à des codons non-AUG dans certains cas (Oyama et al, 2007). Plus récemment, Menschaert *et al.* (Menschaert et al, 2013) ont utilisé des données de *ribosome profiling* publiquement disponibles (Ingolia et al, 2011) pour créer une base de données de protéines putatives traduites spécifiques aux CSE murines. Cela leur a permis de produire une base de données de taille réduite, ce qui diminue le taux de mauvaises identifications tout en en augmentant le nombre total (Nesvizhskii, 2010). En analysant le protéome de ces cellules avec cette base de données personnalisée, seulement 4 protéines alternatives ont été détectées, toutes issues d'uORFs. En 2013, une étude peptidomique de grande stringence a mené à l'identification avec une très bonne confiance de 90 protéines de 18 à 149 AA, dont 86 nouvelles (Slavoff et al, 2013). Environ 80% d'entre elles font moins de 100 AA, une taille classique pour les petites protéines fonctionnelles décrites jusqu'à présent, ce qui renforce l'intérêt de la méthode. Une base de données personnalisée pour l'échantillon de cellules K562 (obtenue sur une base de RNA-SEQ et d'ARNm RefSeq validés) a été utilisée. Parmi les petites protéines identifiées à partir de RefSeq, il est intéressant de noter

que la majorité sont issues d'ARNm multi-codants (et sont donc des AltORFs), bien que 14% d'entre elles semblent être exprimées à partir d'ARNm antisens ou d'ARNs précédemment définis comme non codants. Ici encore, 43% seulement des sites d'initiation les plus probables étaient des codons AUG, bien que d'autres codons (en particulier variant d'un seul nucléotide par rapport à AUG) soient aussi utilisés (21% des cas). Le tour de force réalisé dans cette étude tient dans la quantification exacte du nombre de copies par cellules de produits d'AltORFs : entre 10 et 2000 copies pour 3 candidats étudiés. En résumé, la profondeur d'analyse offerte par les approches protéomiques est, à l'heure actuelle, bien inférieure à celle obtenue par *ribosome profiling* (bien moins d'évènements de traduction d'AltORFs identifiés). Néanmoins, l'utilisation de la spectrométrie de masse permet de fournir l'indication indéniable que les exemples d'AltORFs découverts sporadiquement ne représentent que la pointe de l'iceberg d'un phénomène dont l'étendue exacte reste à définir.

4. Question, hypothèses et objectifs de recherche

4.1. Question de recherche

Nous avons vu qu'il est établi que l'utilisation de l'initiation alternative de la traduction permet de produire plusieurs groupes de protéoformes à partir d'un ARNm mature unique chez les eucaryotes, y compris chez l'humain. Cependant, les preuves au niveau protéique de l'expression d'AltORFs à des niveaux détectables restent minces, seulement quelques dizaines d'exemples ayant été découverts de façon sporadique ou dans des études à large échelle. De plus, de nombreuses protéines alternatives n'ont toujours pas été associées à des fonctions, et leur importance biologique reste donc à établir. Le modèle traditionnel selon lequel un ARNm mature eucaryote est monocistronique reste donc largement dominant face aux rares contre-exemples d'ARNm multi-codants. La question de recherche à laquelle j'ai voulu répondre est donc la suivante :

L'expression de protéines alternatives à partir des gènes eucaryotes est-il un phénomène généralisé ?

4.2. Hypothèse de recherche

En réponse à cette question, j'ai donc émis l'hypothèse suivante :

Les ARNm matures eucaryotes contiennent fréquemment un ou des AltORFs efficacement traduits en plus du RefORF, et les protéines alternatives correspondantes contribuent significativement à l'établissement du protéome.

4.3. Objectifs

Lors du début de mon doctorat, l'intérêt de recherche principal du laboratoire du Dr Roucou était l'étude des fonctions du gène de la protéine prion en conditions physiologiques et pathologiques (maladies neurodégénératives). Mon projet de recherche a donc débuté avec une attention particulière portée à ce gène, dans une approche par gène candidat, avant de s'élargir pour généraliser l'étude des AltORFs en accord avec la question de recherche. Les objectifs ont donc été les suivants :

a) Démontrer qu'un AltORF est utilisé dans le gène de la protéine prion, permettant l'expression endogène d'une protéine distincte de la protéine prion issue du RefORF. Caractériser la protéine alternative encodée dans cet AltORF.

b) A partir du prototype constitué par l'AltORF présent dans le gène de la protéine prion, mettre au point une méthode de prédiction d'AltORFs dans le transcriptome, en particulier chez l'humain, afin d'estimer l'étendue potentielle de l'utilisation des AltORFs chez les eucaryotes.

c) Mettre au point une méthode permettant de valider à large échelle l'expression de protéines alternatives prédites, afin de définir l'étendue de la contribution des protéines alternatives au protéome eucaryote, en particulier humain.

ARTICLE 1

An overlapping reading frame in the *PRNP* gene encodes a novel polypeptide distinct from the prion protein

Auteurs de l'article: Benoît Vanderperre, Antanas B. Staskevicius, Guillaume Tremblay, Marie McCoy, Megan A. O'Neill, Neil R. Cashman, et Xavier Roucou

Statut de l'article: publié dans *The FASEB Journal*, 25(7):2373-2386, 2011.

Avant-propos: Pour cet article, je suis co-premier auteur avec Antanas B. Staskevicius. J'ai participé à 70% de la planification des expériences, et à 60% de la réalisation des expériences présentées. J'ai aussi participé à 50% de l'écriture du manuscrit (résultats et discussion, et une partie du matériel et méthodes).

Résumé : Le gène de la protéine prion *PRNP* dirige la synthèse d'une des protéines de mammifères les plus étudiées, la protéine prion (PrP). Pourtant, la fonction physiologique de PrP est restée évasive et a créé des controverses dans la littérature. Nous avons trouvé dans le cadre de lecture +3 de *PRNP* un codon AUG d'initiation de la traduction en aval de celui de PrP et entouré par une séquence Kozak optimale. Le cadre de lecture ouvert alternatif correspondant code pour un polypeptide appelé protéine prion alternative (AltPrP) avec une séquence en acides aminés totalement différente de celle de PrP. Nous avons introduit une étiquette hémagglutinine (HA) dans le cadre de lecture d'AltPrP dans l'ADNc PrP de différentes espèces afin de tester l'expression de ce nouveau polypeptide en utilisant des anticorps anti-HA. AltPrP est constitutivement co-exprimée avec PrP chez l'humain, le bovin, le mouton, et le cerf. AltPrP est localisée aux mitochondries et son expression est régulée positivement par le stress au réticulum endoplasmique et l'inhibition du protéasome. La synthèse d'anticorps anti-AltPrP nous a permis de tester l'expression endogène d'AltPrP dans les cellules humaines de type sauvage exprimant PrP. En transfectant des cellules avec un siRNA contre l'ARNm de PrP, nous avons réprimé l'expression à la fois de PrP et AltPrP, confirmant expression endogène d'AltPrP à partir de

PRNP. Ces résultats démontrent une fonction inattendue pour *PRNP*, qui en plus de PrP, ancrée à la membrane plasmique, encode également un second polypeptide appelé AltPrP.

An overlapping reading frame in the *PRNP* gene encodes a novel polypeptide distinct from the prion protein

Benoît Vanderperre[*1], Antanas B. Staskevicius[*1], Guillaume Tremblay[1], Marie McCoy[1], Megan O'Neill[2], Neil R. Cashman[2] and Xavier Roucou[1]

[1]Department of Biochemistry, Faculty of Medicine, Université de Sherbrooke, Sherbrooke, Quebec, J1H 5N4, Canada.

[2] Brain Research Centre, University of British Columbia, 2211 Wesbrook Mall, Vancouver, BC, V6T 2B5 Canada.

[*]These authors contributed equally to this work.

Address correspondence to Dr. Xavier Roucou, Department of Biochemistry, Faculty of Medicine, University of Sherbrooke, 3001 12ème avenue nord, Sherbrooke, QC, J1H 5N4, Canada, Tel: (819) 346 1110x12248; Fax: (819) 564 5340; E-mail: xavier.roucou@usherbrooke.ca

Short title: *PRNP*: one gene, two distinct proteins

Abstract

The prion protein gene *PRNP* directs the synthesis of one of the most intensively studied mammalian proteins, the prion protein (PrP). Yet the physiological function of PrP has remained elusive and has created controversies in the literature. We found a downstream alternative translation initiation AUG codon surrounded by an optimal Kozak sequence in the +3 reading frame of *PRNP*. The corresponding alternative open reading frame encodes a polypeptide termed Alternative Prion Protein (AltPrP) with a completely different amino acid sequence from PrP. We introduced a hemagglutinin (HA) tag in-frame with AltPrP in PrP cDNAs from different species to test the expression of this novel polypeptide using anti-HA antibodies. AltPrP is constitutively co-expressed with human, bovine, sheep, and deer PrP. AltPrP is localized at the mitochondria and is upregulated by ER stress and proteasomal inhibition. Generation of anti-AltPrP antibodies allowed us to test for endogenous expression of AltPrP in wild-type human cells expressing PrP. By transfecting cells with siRNA against PrP mRNA, we repressed expression of both PrP and AltPrP, confirming endogenous expression of AltPrP from *PRNP*. These results demonstrate an unexpected function for *PRNP*, which in addition to plasma membrane-anchored PrP also encodes a second polypeptide termed AltPrP.

Key words: alternative translation initiation, PrP, overlapping reading frame, mitochondria

Introduction

The prion protein (PrP) is a glycoprotein anchored to the plasma membrane by virtue of a glycosylphosphatidylinositol (GPI) anchor (1). Transmissible Spongiform Encephalopathies, otherwise known as TSEs, involve the conversion of PrP^{C}, the normally folded conformer of PrP, into PrP^{Sc}, an aggregation-prone isoform of PrP that is resistant to proteinase K (2). PrP^{Sc} is known to be the infectious agent in TSEs. However, the presence of normally folded PrP^{C} is absolutely necessary for the onset of disease, as it acts as a continuous supply for the generation of PrP^{Sc}. Therefore, *PRNP,* the gene encoding PrP, is essential for the development of TSEs, and not surprisingly, PrP knockout animals are resistant to prion infection (3,4,5). Many missense mutations within human *PRNP* are also associated with genetic forms of TSEs (6), providing support for a central role of PrP in TSE pathogenesis.

In contrast to its well-established pathogenic role, the quest for the normal physiological function of PrP has proven very difficult. Several functions have been proposed, and controversies on the role of PrP remain in the literature (7). Some of the explanations put forward include the study of different cultured cell models used to investigate the function of PrP, such as neuronal versus non-neuronal cell lines, immortalized cell lines versus primary cells, and mouse neurons versus human neurons. How one gene and the associated protein can result in such a complexity in terms of physiological function is confounding.

It has recently been hypothesized that alternative translation initiation of eukaryotic mRNAs might be used as a method to expand the proteome (8). Based on the idea that a single mRNA can produce three completely independent amino acid sequences if read in all three possible reading frames, this hypothesis suggests that the complexity of the eukaryotic proteome is largely underestimated. Several examples of out-of-frame alternative translation initiation in eukaryotes exist to support this hypothesis. However, almost all of these examples occur at an upstream AUG codon in relation to the +1 position of the main open reading frame (ORF) (9,10). These alternative AUG codons are usually situated within an optimal Kozak context (11). Although extremely rare, a small number of examples of out-of-frame alternative translation initiation at a downstream AUG codon in an optimal Kozak context exist in mammals (12,13).

Upon re-examination of the PrP coding sequence (CDS), we found a potential ORF whose initiator codon is surrounded by an optimal Kozak context in a number of species. In this study, we show that the protein encoded in this overlapping ORF is co-expressed with PrP from the *PRNP* gene. This finding has direct implications regarding the comprehension of the physiological function of *PRNP*.

Materials and Methods

Cloning of plasmids and transfection- All primer sequences are outlined in Table 1. Cloning of human PrP^{C} in pCEP4β vector (Invitrogen, Carlsbad, CA, USA) has been described previously (14). Human $PrP^{(HA)}$ was produced by inserting an HA tag in the +3 frame of human PrP between bases 308 and 309 of the PrP CDS by PCR overlap extension using the forward primers 1 and 3 and the reverse primers 2 and 4. Human $PrP^{(HA)*}$, in which the alternative AUG at bp 90-92 of $PrP^{(HA)}$ was mutated to CUG, was produced by the Quikchange method (Stratagene, La Jolla, CA, USA) using the forward primer 5 and the reverse primer 6. Human C-terminally tagged $AltPrP^{HA}$ was produced using $huPrP^{(HA)}$ as a template with the forward primer 7 and the reverse primer 9. Untagged human AltPrP was amplified from $huPrP^{C}$ with the forward primer 7 and the reverse primer 8. Human $PrP^{\Delta 1-66(HA)}$, in which the first 66 bps of the $PrP^{(HA)}$ CDS were deleted, was produced using the forward primer 10 and the reverse primer 2. All constructs were inserted in pCEP4β vector using HindIII and BamHI restriction enzymes. Recombinant huAltPrP was produced with the forward primer 11 and the reverse primer 12. The PCR product was then inserted in pET-21b vector (EMD Chemicals, Gibbstown, NJ, USA) using NheI and BamHI restriction enzymes. The recombinant protein was then expressed as previously described (15). Bovine PRNP was reverse-transcribed from a total RNA extract using the forward primer 13 and the reverse primer 14. The bovine PrP CDS was then amplified using the forward primer 15 and the reverse primer 16. Bovine PrP was then inserted in pCEP4β vector using the HindIII and NotI restriction enzymes. Bovine $PrP^{(HA)}$ and $AltPrP^{HA}$ were produced as described above with primers 15, 16, 17, 18, 19 and 20, and inserted in pCEP4β vector using HindII and BamHI restriction enzymes. White-tailed deer (wtd) PrP in pCEP4β was a kind gift from Debbie McKenzie (Alberta Centre for Prion and Protein Folding Diseases, Department of

Biological Sciences, University of Alberta, Edmonton, Alberta, Canada). wtdPrP$^{(HA)}$ and AltPrP^{HA} were produced as described above with primers 15, 16, 19, 20, 21 and 22. PCR products were then inserted in pCEP4β vector using HindIII and BamHI restriction enzymes. Sheep PrP in pCI expression vector was a kind gift from Dr. Michael A. Tranulis (Norwegian School of Veterinary Science, Dept. of Biochemistry and Physiology, Institute of Basic Sciences and Aquatic Medicine, Oslo, Norway). Sheep PrP$^{(HA)}$ and AltPrP^{HA} were produced as described above with primers 23, 24, 25, 26, 27 and 28. PCR products were then reinserted in the pCI expression vector (Promega, Madison, WI, USA) using the XhoI and XbaI restriction enzymes. All restriction enzymes were acquired from New England Biolabs (Ipswitch, MA, USA). Cells were transfected with ExGen 500 transfection reagent (Fermentas, Burlington, ON, Canada) or GeneCellin transfection reagent (BioCellChallenge, Toulon, France) according to the manufacturer's instructions.

Table 1. Primer Sequences

Primer	Sequence, 5'-3'
huPrP F	CCCAAGCTTCGTAATGGCGAACCTTGGCTGCTGG
huPrP R	CGCGGATCCTCATCGACTATGAGGAAGATG
huPrP$^{(HA)}$ F	TATCCGTACGACGTACCAGACTACGCCTAAGCCAAAAACCAAC
huPrP$^{(HA)}$ R	GTCGAACAAGCCGAGTATCCGTACGACGTACCAGACTACGCC
huPrP$^{(HA)}$ F	CGAAGCCTGGAGCCTGGAACACTGGGG
huPrP$^{(HA)}$ R	CCCCAGTGTTCCAGGCTCCAGGCTTCG
huAltPrP F	GGAAGCTTCGCCATGGAACTCTCGGGGCA
huAltPrP R	CCCGGATCCTTACTCGGCTTGTTC
huAltPrP^{HA} R	CGACGTACCAGACTACGCCTAAGGATCCGA
huΔ1-60 F	AGTAAGCTTCGCCGCCATGGGCGCTCTGCAAGAAGCGGCCGAAGC
huAltPrP pET F	AGGCTAGCGAACACTGGGGGCAGCCGATA
huAltPrP pET R	TCTGGATCCTTACTCGGCTTGTTCCACTGACTGT
boPrP F	TTCAACCAAGCCGAAGCATCTGTC
boPrP R	AGCACGAAATGAGACACCACGACT
boPrP$^{(HA)}$ F	TATCCGTACGACGTACCAGACTACGCCTAAGCCAAAAACCAAC
boPrP$^{(HA)}$ R	GGCGTAGTCTGGTACGTCGTACGGATACTCCTGGGTTTGTTCC
boAltPrP F	ATCAAGCTTAGCAGGATGGAACACTCG
boAltPrP^{HA} R	GATGGATCCTTAGGCGTAGTCTGGTAC
wtdPrP F	ACACCCTCTTTATTTTGCAG
wtdPrP R	AGAAGATAATGAAAACAGCAAGG
wtdPrP$^{(HA)}$ F	TATCCGTACGACGTACCAGACTACGCCTAAGCCAAAAACCAAC
wtdPrP$^{(HA)}$ R	GGCGTAGTCTGGTACGTCGTACGGATACTCCTGGGCTTGTTCC
shPrP F	TAGCCTCGAGGTCATCATGGTGAAAAGCCACATAGG
shPrP R	CGACTCTAGAGTACTATCCTACTATGAGAAAAATGAGG
shPrP$^{(HA)}$ F	TATCCGTACGACGTACCAGACTACGCCTAGCCAGATCAGTCAGTGGAACAAA
shPrP$^{(HA)}$ R	GGCGTAGTCTGGTACGTCGTACGGATACTCCCAGCTTGACCCCAG
shAltPrP^{HA} F	TATCTCGAGGCCACCATGGAACACTGGGGCA
shAltPrP^{HA} R	GGCTGGGCTCAAGGTGGGAGTATCCGTACGACGTACCAGACTACGCCTAGTCTAGATAT

F, Forward ; R, Reverse.

Antibodies and reagents- Primary antibodies used were monoclonal anti-Cox IV (ab14744), polyclonal anti-VDAC1 (ab15895), polyclonal anti-GRP78 BiP (ab53068), polyclonal anti-GAPDH (ab9485), polyclonal anti-HA (ab9110), monoclonal anti-phospho-eIF2α (ab32157) (Abcam, Cambridge, UK), monoclonal anti-βactin (clone AC-15, Sigma-Aldrich, St. Louis, MO, USA), monoclonal anti-Hsp70 (SPA-810, Stressgen, Ann Arbor, MI, USA), polyclonal anti-Bax (sc-493, Santa Cruz Biotechnology, Santa Cruz, CA, USA), monoclonal anti-Cytochrome c (clone 6H2.B4, BD Pharmingen, Franklin Lakes, NJ, USA), monoclonal anti-HA (clone C29F4, Cell Signaling Technology, Danvers, MA, USA), monoclonal anti-mitochondrial Hsp70 (clone JG1, Affinity BioReagents, Waltham, MA, USA), monoclonal anti-PrP (clone SAF32, Cayman Chemical, Ann Arbor, MI, USA), monoclonal anti-α-tubulin (clone A11126, Molecular Probes, Eugene, OR, USA). Anti-PrP clone 3F4 was purified from hybridoma cell lines. Rabbit polyclonal antibodies against human AltPrP were raised against residues 59-73 and affinity purified by GenScript (Piscataway, NJ, USA). Secondary antibodies used were horseradish peroxidase (HRP)-conjugated sheep anti-mouse IgG (NA931V), HRP-conjugated donkey anti-rabbit IgG (NA934V) (GE Healthcare, Little Chalfont, Buckinghamshire, UK), Alexa Fluor 488-conjugated goat anti-mouse IgG (A-11701) and Alexa Fluor 568-conjugated goat anti-rabbit IgG (A-21069) (Invitrogen, Carlsbad, CA, USA). All other reagents were obtained from Sigma-Aldrich (St. Louis, MO, USA), unless otherwise stated.

Cell culture and drug treatments- Human epithelial kidney cells (HEK293), murine neuroblastoma (N2a) and human astrocytoma (U-118 and U-87) cells were grown in Dulbecco's Modified Eagle's Medium supplemented with 10% Fetal Bovine Serum (FBS) (Wisent, St-Bruno, QC, Canada). Human neuroblastoma (BE(2)-M17) cells were grown in a 1:1 mixture of Eagle's Minimum Essential Medium and F-12 Medium supplemented with 5% FBS. All culture media were supplemented with amphotericin B as well as penicillin/streptomycin. Human primary neurons (catalog number 1520-5, ScienCell, Carlsbad, CA, USA) were grown in Neuronal Medium (catalog number 1521, ScienCell, Carlsbad, CA, USA) according to the manufacturer's protocol. Peripheral blood mononuclear cells were purified from human peripheral blood using Ficoll-Paque PLUS (GE Healthcare, Little Chalfont, Buckinghamshire, UK) according to the manufacturer's protocol. Drug

treatments were conducted on cells transfected for 24 hours as follows, unless otherwise stated: MG132 and epoxomicin were used for 8 hours, each at a concentration of 10μM. Thapsigargin, tunicamycin and A23187 were used for 24 hours at a concentration of 3.5μM, 1.5μM and 1.3μM, respectively. Cycloheximide was used at a concentration of 107μM.

siRNA treatments- U-118 cells were plated in a 6-well plate at $2x10^5$ per well in fresh medium containing no antibiotics. 24 h later, *PRNP* siRNA (SI03019324) or AllStars Negative Control siRNA (1027281) (Qiagen, Mississauga, ON, Canada) were transfected into the cells at a final concentration of 100nM using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's protocol. 72h later, cells were harvested and lysates were processed for SDS-PAGE and western blot analysis to assess knock-down efficiency.

Sample preparation and Immunoblotting- Cells were grown in 6-well plates for 24 hours and were then transfected as described above. Cells were rinsed and harvested in phosphate-buffered saline (PBS) and centrifuged for 60 seconds at 5,000 rpm. Cells were then lysed in RIPA buffer and samples were quantified using BCA protein assay reagent (Pierce, Waltham, MA, USA). Preparation of 10% human brain homogenate in PBS supplemented with EDTA Complete protease inhibitor cocktail (Roche Applied Science, Laval, Quebec, Canada) was prepared using an Omni TH115 Tissue Homogenizer (Omni International, Kennesaw, GA, USA) according to the manufacturer's protocol. 100μg of protein from each sample were precipitated using the chloroform/methanol technique described by Wessell and Flugge (16) and the resulting pellets were resuspended in 4X SDS-PAGE loading dye. After electrophoresis, proteins were transferred to PVDF membranes according to the manufacturer's protocol. Membranes were then exposed using Luminata Forte Western HRP Substrate (Millipore Corporation, Billerica, MA, USA) or Western Lightning ECL reagent (Perkin Elmer, Waltham, MA, USA) according to the manufacturer's instructions. Films used were Amersham Hyperfilm ECL films (GE Healthcare, Little Chalfont, Buckinghamshire, UK). Membranes were stripped by washing twice in 0.2N NaOH for 20 min, rinsed in PBS, blocked and reprobed as described above. Densitometric analysis was conducted using the

ImageJ software. Densitometric values were corrected for loading (anti-tubulin or anti-actin signal obtained by western blot).

Immunofluorescence- Immunofluorescence was carried out as previously described (17). Confocal analysis was carried out as previously described (18).

Mitochondrial Fractionation- Mitochondria were isolated and treated with sodium carbonate as previously described (19).

PNGase F treatment and solubility assay- PNGase F treatment of transfected cell lysates was carried out according to the manufacturer's instructions (New England Biolabs, Ipswitch, MA, USA). Solubility of PrP was assessed as previously described (20).

Ethics statement- The human brain sample was a kind gift from Dr. Cheryl Wellington (Associate Professor, Department of Pathology, University of British Columbia, Vancouver, British Columbia, Canada). All experiments were performed according to ethics protocol number UBC C04-0595.

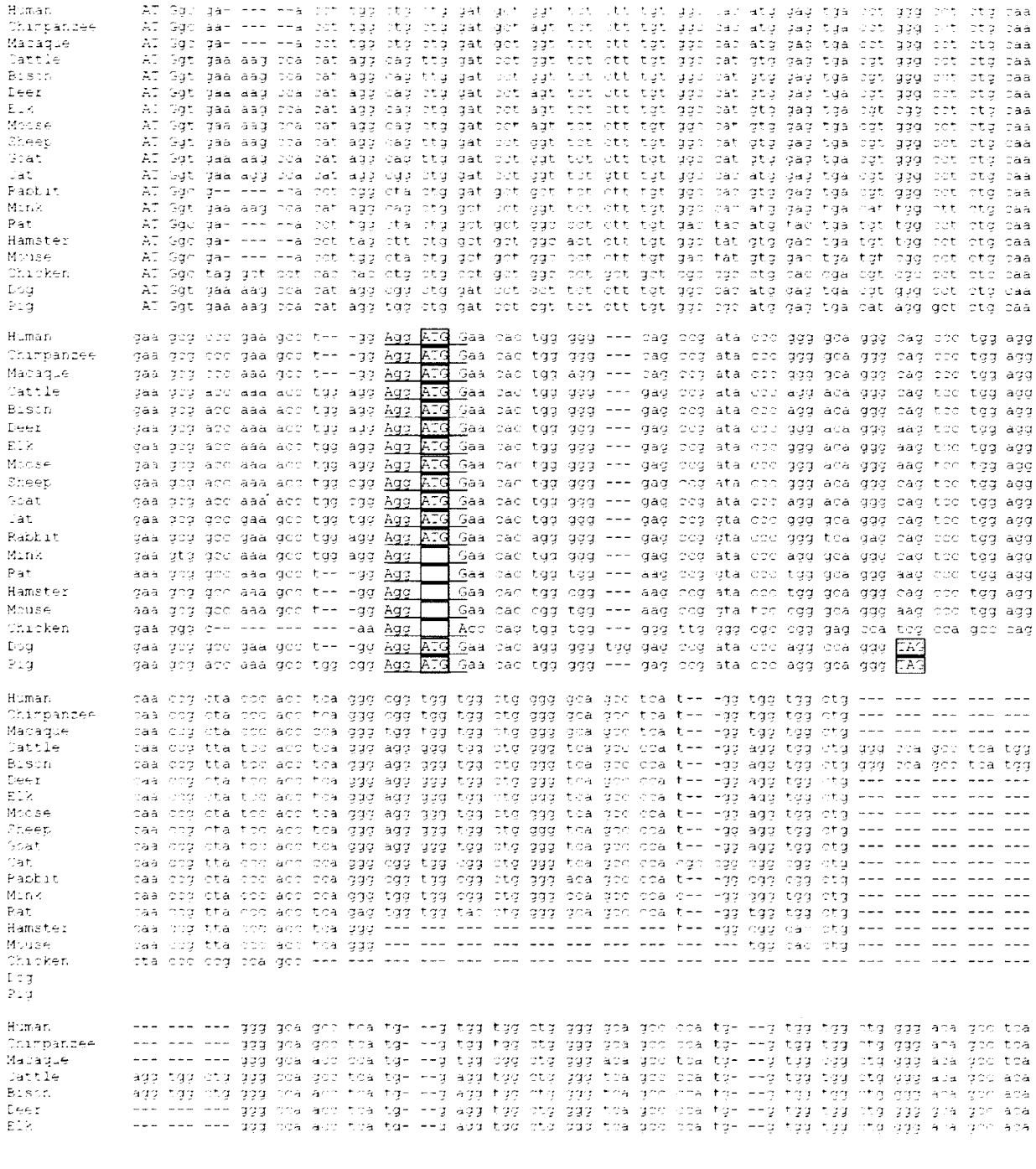
Results

An alternative ORF exists in the coding sequence of PrP

In an attempt to uncover the reason for the numerous functions attributed to PrP, we have re-examined the sequence of *PRNP* from several species. We noticed an ORF in the +3 reading frame in which the AUG initiation codon (bp 90-92 in the human PrP CDS) is positioned within an optimal Kozak context (Fig. 1A). We termed the putative polypeptide encoded by this ORF Alternative PrP (AltPrP). The AltPrP ORF covers the entire octarepeat (OR) region of PrP. AltPrP ranges in length between 64 and 81 amino acids, and contains several tryptophan-rich repeats resulting from translation of the OR region of PrP in the +3 reading frame (Fig. 1B). In other species, such as mouse and hamster, the alternative initiation AUG codon is absent and is replaced by a GUG codon still located within an optimal Kozak sequence (Fig. 1A). Although there is at least one example in

which GUG is an efficient initiation codon to translate a protein (21), we focused on the expression of AltPrP in humans, cattle, sheep, and white-tailed deer in this study.

A



(continued on next page)

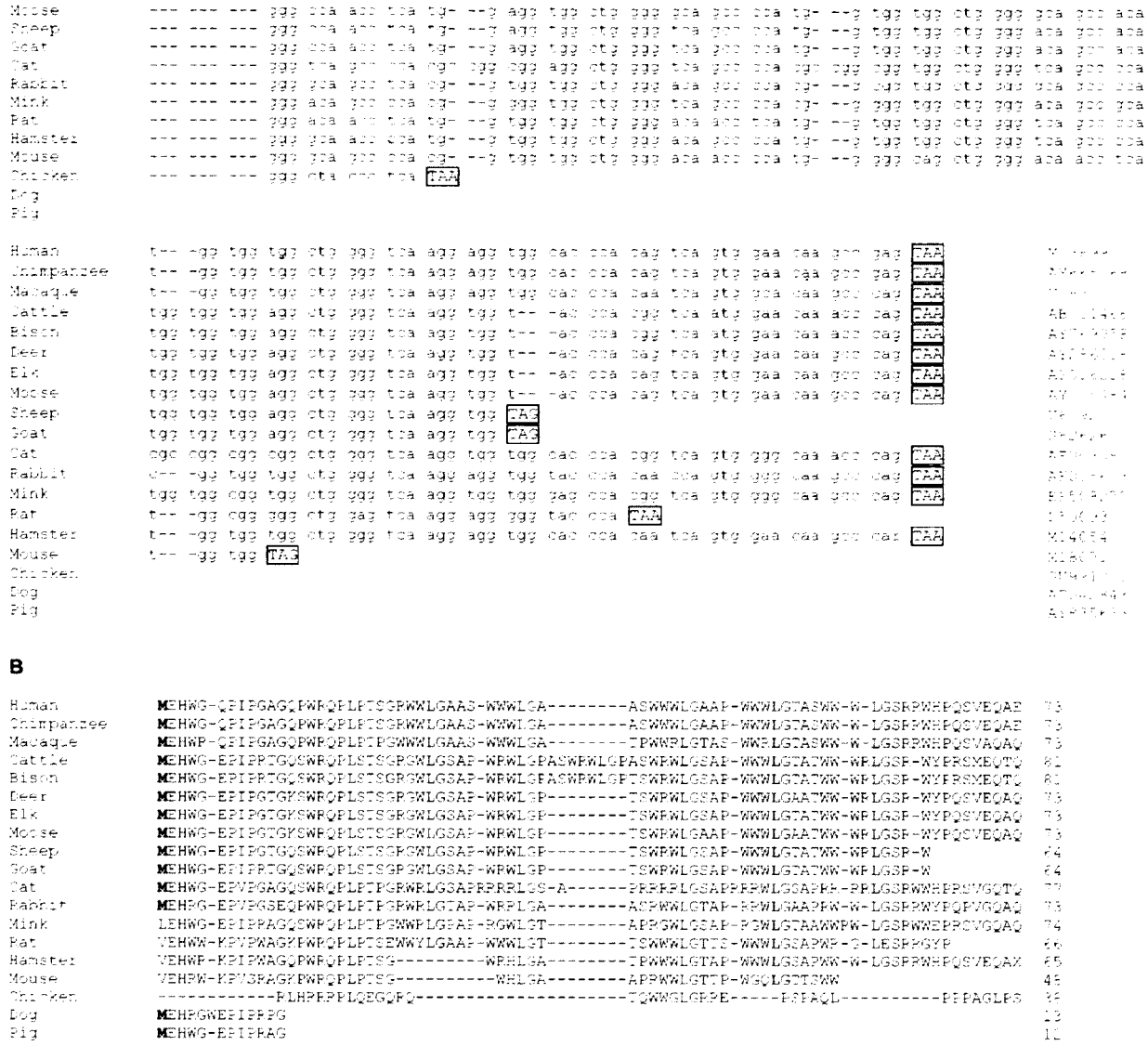


Figure 1. An alternative ORF overlapping with the PrP octarepeat (OR) region exists in the +3 frame of several species. (**A**) DNA sequence alignment of *Prnp* in several species. *Prnp* DNA sequences of several species were aligned starting from the PrP start codon (capitalized letters) until the AltPrP stop codon (boxed and capitalized). All larger mammals analyzed contain a start codon (boxed and capitalized) surrounded by a Kozak consensus sequence (underscored). Most smaller species analyzed do not possess a conventional ATG start codon (shown in gray), though they all possess a Kozak consensus sequence, except chicken. Sequences were obtained from GenBank and aligned using ClustalW software. GenBank accession number of each sequence used is included at the end of the DNA sequence. (**B**) Amino acid sequence alignment of AltPrP in several species. AltPrP is present in several large mammals and has conserved tryptophan-rich repeats. Some smaller mammals contain the AltPrP ORF, yet have interrupted tryptophan-rich repeats replaced by arginine (such as cat and rabbit). Other species do not possess a conventional initiator methionine residue (mink, rat, hamster, and mouse) or contain a stop codon early on in the peptide sequence (dog and pig). Length of the ORF in each species is indicated at the end of the peptide sequence. Sequences were translated from DNA

sequences obtained from GenBank using the ExPaSy Translate tool and then aligned using ClustalW software. Nucleotide sequence data reported are available in the Third Party Annotation Section of the DDBJ/EMBL/GenBank databases under accession numbers TPA: BK007887–BK007890.

Table 2. Amino acid sequences of constructs used to detect AltPrP

Construct	Reading frame	Sequence
huPrP	1	MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGTWGQPHGG WGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVGGLGGY MLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTT TKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLISFLIFLIVG
	3	MEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE
huPrP^{PEX}	1	MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGG GWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPS*IRTTYQTTP*KPKTNMKHMAGAAA AGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILL ISFLIFLIVG
	3	MEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE*YPTDVPDYA*
huPrP^{HA}	1	MANLGCWMLVLFVATWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGG GWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPS*IRTTYQTTP*KPKTNMKHMAGAAA AGAVVGGLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCV NITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILL ISFLIFLIVG
	3	LEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE
huAltPrP	1	MEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE
huAltPrP^{HA}	1	MEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE*YPYDVPDYA*
PrP$^{\Delta 23-88/HA}$	1	MGLGKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGW GQPHGGGWGQGGGTHSQWNKPS*IRTTYQTTP*KPKTNMKHMAGAAAAGAVVGGLGGYMLGSAMS RPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENF TETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLISFLIFLIVG
	3	MEHWGQPIPGAGQPWRQPLPTSGPWWLGTASWWWLGAASWWWLGAAPWWWLGTASWWWLGSRRW HPQSVEQAE*YPYDVPDYA*
boPrP	1	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH GGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGTHGGWNKPSKPKTNMKHVAGAAA AGAVVGGLGGYMLGSAMSRPLIHFGSDYEDRYYRENMHRYPNQVYYRPVDQYSNQNNFVHDCVN ITVKEHTVTTTTKGENFTETDIKMMERVVEQMCITQYERESQAYYQRGASVILFSSPPVILLIS FLIFLIVG
	3	MEHWGEPIPRTGQSWRQPLSTSGRGWLGSAPWRWLGPASWRWLGPASWRWLGSAPWWWLGTATW WWRLGSRWYPRSMEQTQ
boPrP^{HA}	1	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH GGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGTHGGWNKPS*IRTTYQTTL*KPKTN MKHVAGAAAAGAVVGGLGGYMLGSAMSRPLIHFGSDYEDRYYRENMHRYPNQVYYRPVDQYSNQ NNFVHDCVNITVKEHTVTTTTKGENFTETDIKMMERVVEQMCITQYERESQAYYQRGASVILFS SPPVILLISFLIFLIVG
	3	MEHWGEPIPRTGQSWRQPLSTSGRGWLGSAPWRWLGPASWRWLGPASWRWLGSAPWWWLGTATW WWRLGSRWYPRSMEQTQ*YPYDVPDYA*
boAltPrP^{HA}	1	MEHWGEPIPRTGQSWRQPLSTSGRGWLGSAPWRWLGPASWRWLGPASWRWLGSAPWWWLGTATW WWRLGSRWYPRSMEQTQ*YPYDVPDYA*
wtPrP	1	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH GGGWGQPHGGGWGQPHGGGWGQPHGGGGWGQGGTHSQWNKPSKPKTNMKHVAGAAAAGAVVGGL GGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYNNQNTFVHDCVNITVKQHTV TTTTKGENFTETDIKMMERVVEQMCITQYQRESQAYYQRGASVILFSSPPVILLISFLIFLIVG
	3	MEHWGEPIPGTGKSWRQPLSTSGPGWLGSAPWPWLGPTSWRWLGSAPWWWLGAATWWWFLGSRW YPQSVEQAQ
wtPrP^{HA}	1	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH GGGWGQPHGGGWGQPHGGGWGQPHGGGGWGQGGTHSQWNKPS*IRTTYQTTL*KPKTNMKHVAGAA AAGAVVGGLGGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYNNQNTFVHDCV NITVKQHTVTTTTKGENFTETDIKMMERVVEQMCITQYQRESQAYYQRGASVILFSSPPVILLI SFLIFLIVG
	3	MEHWGEPIPGTGKSWRQPLSTSGPGWLGSAPWPWLGPTSWRWLGSAPWWWLGAATWWWFLGSRW YPQSVEQAQ*YPYDVPDYA*
wtAltPrP^{HA}	1	MEHWGEPIPGTGKSWRQPLSTSGPGWLGSAPWPWLGPTSWRWLGSAPWWWLGAATWWWFLGSRW YPQSVEQAQ*YPYDVPDYA*

continued on next page

Table 2. *(continued)*

Construct	Reading frame	Sequence
		MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH-
		GGGWGQPHGGGWGQPHGGGWGQPHGGGGWGQGGSHSQWNKPSKPKTNMKHVAGAAAAGAVVGGL-
		GGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYSNQNNFVHDCVNITVKQHTV-
shPrP	1	TTTTKGENFTETDIKIMERVVEQMCITQYQRESQAYYQRGASVILFSSPPVILLISFLIFLIVG
	3	MEHWGEPIPGTGGSWRQPLSTGRGWLGSAPWRWLGFTSWRWLGSAPWWWLGTATWWWRLGSRW
shPrP^{HA}	1	MVKSHIGSWILVLFVAMWSDVGLCKKRPKPGGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPH-
		GGGWGQPHGGGWGQPHGGGWGQPHGGGGWGQGG*IRTTYQTTP*SHSQWNKPSKPKTNMKHVAGAA-
		AAGAVVGGLGGYMLGSAMSRPLIHFGNDYEDRYYRENMYRYPNQVYYRPVDQYSNQNNFVHDCV-
		NITVKQHTVTTTTKGENFTETDIKIMERVVEQMCITQYQRESQAYYQRGASVILFSSPPVILLI
		SFLIFLIVG
	3	MEHWGEPIPGTGGSWRQPLSTGRGWLGSAPWRWLGFTSWRWLGSAPWWWLGTATWWWRLGSR-
		WYPYDVPDYA
shAltPrP^{HA}	1	MEHWGEPIPGTGGSWRQPLSTGRGWLGSAPWRWLGFTSWRWLGSAPWWWLGTATWWWRLGSR-
		WYPYDVPDYA

Amino acid sequences of all constructs used are included. Note that wherever possible, the amino acid sequence of both the PrP reading frame (+1) and the AltPrP reading frame (+3) are included for each construct. Underscored letters represent the hemagglutinin (HA) tag. When shown in roman type, the HA sequence is in its proper reading frame in order to be detected by an anti-HA antibody. When shown in italics, the HA sequence is in another reading frame, which cannot be detected by an anti-HA antibody. Note that for huPrP(HA)*, the initiator methionine codon (M) is replaced by leucine (L, double underscored).

AltPrP is expressed from PrP cDNA of several species

A detailed list of all constructs used, along with amino acid sequences of each construct, is outlined in Table 2. In order to test if AltPrP is expressed, we introduced a hemagglutinin (HA) tag in-frame with AltPrP to produce carboxy-tagged AltPrP ($AltPrP^{HA}$), within the human PrP cDNA (Fig. 2A). For clarity reasons, this construct is termed $PrP^{(HA)}$, where (HA) indicates that the HA tag is silent within the reading frame of PrP. As a positive control, we created a construct encoding solely $AltPrP^{HA}$. We also engineered a $PrP^{(HA)}$ construct with an inactivated alternative initiation codon. In this construct termed $PrP^{(HA)*}$, the AUG at bp 90 was changed to CUG. Lysates from mammalian cells transfected with PrP, $PrP^{(HA)}$ or $PrP^{(HA)*}$ were probed with both anti-PrP and anti-HA antibodies to test for the expression of PrP, $PrP^{(HA)}$, $PrP^{(HA)*}$ and $AltPrP^{HA}$. As expected, the introduction of the HA tag resulted in a slight increase in the molecular weight of PrP (Fig. 2B). On a SDS-PAGE gel, $PrP^{(HA)}$ migrated as several bands, indicating the presence of glycosylations similar to native PrP. Remarkably, a band corresponding to the expected molecular weight for $AltPrP^{HA}$ was detected with an anti-HA antibody in cells transfected with $PrP^{(HA)}$. The identity of this band was confirmed by probing lysates from cells directly transfected with cDNA encoding $AltPrP^{HA}$. $AltPrP^{HA}$ was not detected in cells transfected with $PrP^{(HA)*}$ (Fig. 2B), clearly showing that translation of AltPrP is indeed initiated at the identified alternative AUG codon (Fig. 1A), most likely by alternative translation initiation. Analogous results were obtained in neuronal as well as in non-neuronal cell lines (Fig. 2B). We then confirmed these results by immunofluorescence on N2a and HEK293 cells expressing each of these constructs (Fig. 2C). Detection of PrP, $PrP^{(HA)}$, $PrP^{(HA)*}$ and $AltPrP^{HA}$ was done using anti-PrP and anti-HA antibodies. In both cell types, expression of $PrP^{(HA)}$ and $AltPrP^{HA}$ resulted in an HA signal with a granular cytoplasmic distribution. Expression of PrP or $PrP^{(HA)*}$ did not provide an HA signal, confirming the results described in figure 2B. The addition of the out-of-frame HA tag in $PrP^{(HA)}$ and $PrP^{(HA)*}$ did not affect the subcellular localization of PrP, which is typically localized at the plasma membrane and the Golgi apparatus.

The ORF encoding AltPrP is also present in *PRNP* from cattle, sheep, and white-tailed deer (Fig. 1A). An HA tag was inserted at the C-terminus of AltPrP in PrP cDNA of each of these species using the same strategy as above for human PrP. Western blot

analysis of HEK293 cells transfected with cattle, sheep, and deer $PrP^{(HA)}$ cDNA demonstrated the presence of AltPrP in these three species in addition to humans (Fig. 2D). This data clearly demonstrates that under normal conditions, the prion protein cDNA of numerous species directs the synthesis of two co-expressed polypeptides: PrP, as well as the newly identified AltPrP.

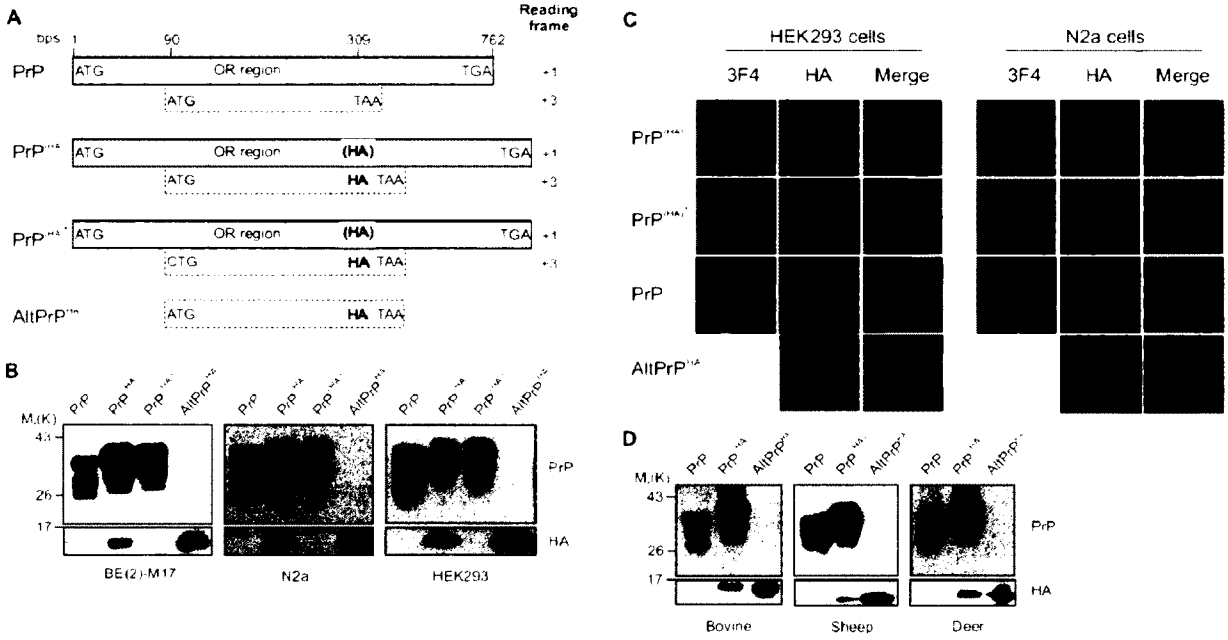


Figure 2: Detection of AltPrP. (**A**) Strategy used to detect AltPrP by introducing an HA tag at the C-terminus of AltPrP. AltPrP is in the +3 reading frame in relation to PrP and spans the entire octapeptide repeat (OR) region in the N-terminal domain of PrP. The PrP reading frame (+1) is represented by the upper boxes in each construct, while the dashed boxes represent the AltPrP reading frame. In $PrP^{(HA)}$, $PrP^{(HA)*}$ and $AltPrP^{HA}$, an HA tag (grey box) was inserted at the C-terminus of AltPrP. The parentheses surrounding the HA in the PrP reading frame represent the fact that the HA epitope sequence is encoded in the AltPrP reading frame, and is therefore undetected if expressed from the ATG codon at bp 1 of the PrP CDS. $PrP^{(HA)*}$ is identical to $PrP^{(HA)}$ except that the ATG codon at bp 90 has been mutated to CUG. (**B**) Western blot against PrP (3F4 epitope) and AltPrP (HA epitope) in BE(2)-M17, N2a, and HEK293 cells transfected with PrP, $PrP^{(HA)}$, $PrP^{(HA)*}$ and $AltPrP^{HA}$ constructs. (**C**) Cells transfected with PrP, $PrP^{(HA)}$, $PrP^{(HA)*}$, or $AltPrP^{HA}$ were immunostained with anti-PrP (3F4; green) or anti-HA (red) antibodies. Merged images also show nuclei stained with Hoechst (blue). (**D**) Expression of $AltPrP^{HA}$ from bovine, sheep, and deer cDNA PrP constructs. Note the slight change in molecular weight between species.

AltPrP is a mitochondrial protein

We then addressed the subcellular localization of AltPrP by immunofluorescence. The precise cytoplasmic localization of AltPrP was determined in both N2a and HEK293 cells expressing human $AltPrP^{HA}$ using antibodies against different cytoplasmic organelles. We observed a clear co-localization of $AltPrP^{HA}$ with mitochondrial Hsp70 (Fig. 3A) and Cytochrome c (data not shown). The presence of AltPrP at the mitochondria was confirmed by subcellular fractionation and differential centrifugation of HEK293 cells stably transfected with $AltPrP^{HA}$. $AltPrP^{HA}$ was detected in the crude cellular fraction as well as in the mitochondrial fraction, but not in the cytosol (Fig. 3B). Bioinformatic analysis predicts a transmembrane domain within AltPrP (Fig. 3C). In order to test this hypothesis, mitochondria isolated from cells expressing $AltPrP^{HA}$ were treated with sodium carbonate. Similar to other mitochondrial membrane-integrated proteins, including VDAC and Cox IV, AltPrP was not extracted after alkali treatment. In contrast, Bax, a peripheral mitochondrial membrane protein in non-apoptotic cells (19) was completely extracted by treatment with sodium carbonate (Fig. 3D). This result demonstrates that AltPrP is a mitochondrial membrane-integrated protein. We next tested if AltPrP is inserted in the mitochondrial inner or outer membrane with digitonin to selectively solubilize the outer membrane (19). In the presence of digitonin, most of the VDAC (an outer membrane component) was extracted from mitochondria while Cox IV (localized at the inner membrane) remained associated with the mitochondria. In these conditions that allow selective extraction of the outer membrane, AltPrP did not remain attached to the mitochondria (Fig. 3E), indicating that AltPrP is inserted in the outer mitochondrial membrane.

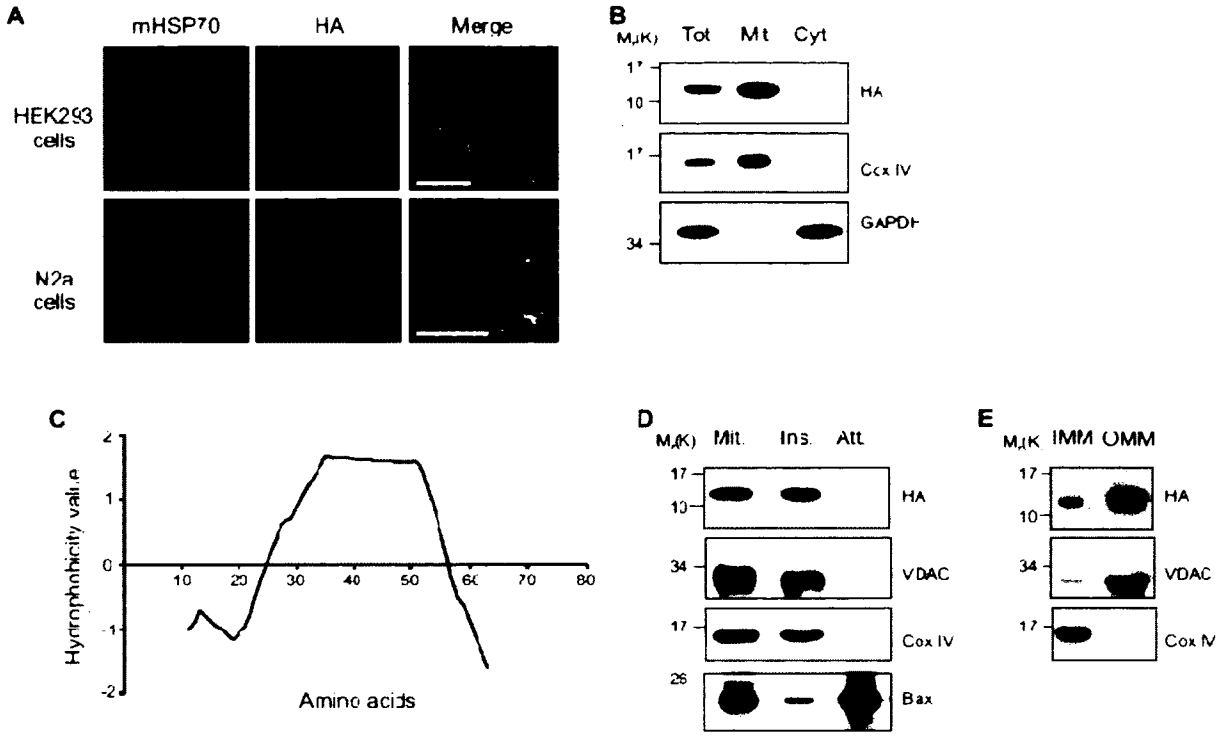


Figure 3. AltPrP^{HA} is localized at the mitochondria and is an integral membrane protein. (**A**) Cells transfected with AltPrP^{HA} were immunostained with anti-HA (red) and anti-mHSP70 (green) antibodies. Scale bar, 10μm. (**B**) Total cell extracts (Tot.), mitochondrial (Mit.) and cytoplasmic (Cyt.) fractions from cells expressing AltPrP^{HA} were immunoblotted for AltPrP, Cox IV, a mitochondrial marker, and GAPDH, a marker of the cytosol. (**C**) Transmembrane domain prediction of AltPrP. The amino acid sequence of human AltPrP was analyzed using TopPred 0.01 software (22). Note that the Goldman Engelman Steitz hydrophobicity profile is computed using a window of 10 amino acids. Human AltPrP contains a predicted transmembrane domain between amino acids 35 and 55. (**D**) Mitochondria isolated from AltPrP^{HA}-expressing cells were treated with 0.1M Na_2CO_3 to produce alkali-resistant (inserted, Ins.) and sensitive (attached, Att.) fractions, and were analyzed by western blot for AltPrP, VDAC, Cox IV, and Bax. (**E**) After treatment with 0.2mg/ml digitonin, the digitonin-sensitive (outer mitochondrial membrane, OMM) and resistant (inner mitochondrial membrane, IMM) fractions were analyzed by western blot for the presence of AltPrP^{HA}, VDAC, and Cox IV.

AltPrP expression is increased by proteasome inhibition and ER stress

The observation that expression levels of $AltPrP^{HA}$ in cells transfected with $PrP^{(HA)}$ are lower than levels in cells directly transfected with $AltPrP^{HA}$ suggests that expression of AltPrP from PrP cDNA is constitutively negatively regulated (Fig. 2). Certain cellular stresses modulate the expression level of a great number of proteins (23,24). Because AltPrP originates from *PRNP*, a locus linked to neurodegeneration, we decided to test if AltPrP expression levels were affected by stresses implicated in neurodegenerative diseases. One such example is proteasomal dysfunction, which is often associated with neurodegeneration (25,26). Treatment of HEK293 cells expressing $PrP^{(HA)}$ or $AltPrP^{HA}$ with the proteasome inhibitors MG132 or epoxomicin for 8 h resulted in a 2 to 3-fold increase of $AltPrP^{HA}$ levels (Fig. 4A,B). Neither drug modified the levels of $PrP^{(HA)}$ or tubulin in the same experimental conditions, showing that this increase is specific to AltPrP. The effect of MG132 was rapid since a significant increase of $AltPrP^{HA}$ levels was observed as soon as 2 h after addition of the drug (Fig. 4C), an indication that AltPrP is a naturally labile protein with a short half-life, probably degraded in a proteasome-dependent manner. The half-life of AltPrP was determined in the presence of cycloheximide, an inhibitor of protein synthesis. In the absence of MG132, the half-life of AltPrP was estimated to be less than 0.5 h (Fig. 4D). The half-life of AltPrP in the presence of MG132 was over 2 h, confirming its stabilization by proteasome inhibition.

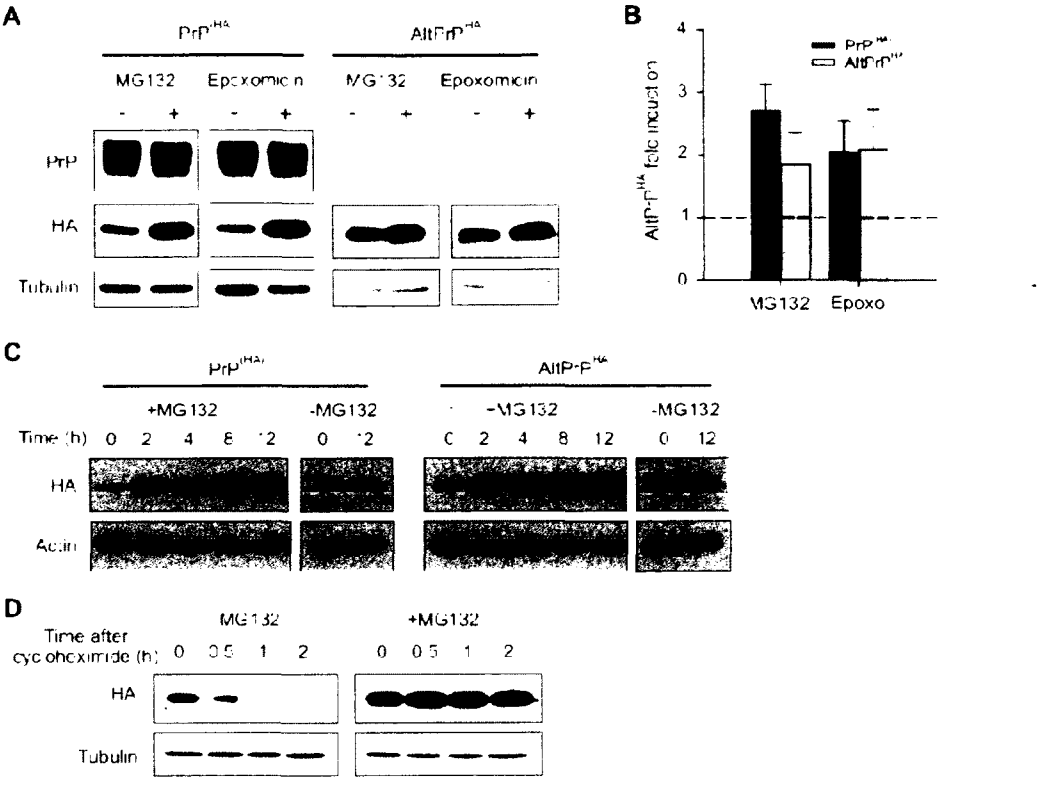


Figure 4: Proteasomal inhibition increases levels of AltPrP^{HA}. (**A**) Western blot of cells expressing PrP$^{(HA)}$ or AltPrP^{HA} treated in the absence or in the presence of MG132 or epoxomicin (10 μM for 8 h) with anti-PrP, -HA, and -tubulin antibodies. (**B**) Densitometric quantification of AltPrP^{HA} induction following proteasome inhibition. Note that both MG132 and epoxomicin (Epoxo.) induce 2-3 fold increase of AltPrP^{HA} whether expressed from PrP$^{(HA)}$ or AltPrP^{HA}. No significant differences (t-test) in induction of AltPrP^{HA} could be observed between the two transfection conditions regardless of the proteasome inhibitor used. Dotted line represents AltPrP^{HA} levels in untreated transfected cells normalized to 1. Values are expressed as the mean value (± s.d.) from at least three independent experiments. (**C**) Time course analysis of MG132-mediated proteasome inhibition on the accumulation of AltPrP in cells expressing PrP$^{(HA)}$ or AltPrP^{HA}. Cell extracts were immunoblotted with anti-HA and -actin antibodies. (**D**) AltPrP^{HA}-expressing cells pre-incubated with MG132 (10 μM for 6 h) were treated with cycloheximide for various times and immunoblotted for HA and tubulin.

Another stress associated with neurodegenerative disorders is endoplasmic reticulum (ER) stress (27,28). It is known that the rate of alternative translation initiation in several mRNAs can be significantly increased during ER stress (29,30). Since AltPrP is most likely produced by alternative translation initiation (Fig. 2B), we tested if ER stress could modulate its expression. HEK293 cells expressing $PrP^{(HA)}$ or $AltPrP^{HA}$ were treated with drugs known to perturb ER homeostasis: thapsigargin (an ER-calcium ATPase inhibitor), tunicamycin (an N-glycosylation inhibitor), and A23187 (a calcium ionophore). The three drugs induced an ER stress as monitored by interfered glycosylation of PrP (demonstrated by the disappearance of higher molecular weight bands detected by the anti-PrP antibody) as well as increased levels of BiP/GRP78 chaperone (Fig. 5A) (31,32). All three drugs also induced at least a 2-fold increase in $AltPrP^{HA}$ levels in cells expressing $PrP^{(HA)}$. However, this effect was not observed in cells directly transfected with $AltPrP^{HA}$, in which $AltPrP^{HA}$ is no longer produced by alternative translation initiation at a downstream AUG codon (Fig. 5A,E). We concluded that ER stress specifically increases the synthesis of AltPrP from PrP cDNA.

We hypothesized that the target of this regulation is located upstream of the AltPrP start codon (bps 1-90 of the human PrP CDS). If so, this region is expected to exert an inhibitory effect on the translation of AltPrP from PrP cDNA that might be abolished during ER stress. In order to test this hypothesis, we engineered a mutant cDNA construct termed $PrP^{\Delta 1\text{-}66(HA)}$ with a deletion of the first 66 bps of the PrP CDS. Since bps 1-66 encode the N-terminal signal peptide, $PrP^{\Delta 1\text{-}66(HA)}$ is expressed as a cytoplasmic protein and does not undergo post-translational modifications, migrating as a single band (Fig. 5B). $AltPrP^{HA}$ was constitutively expressed in cells transfected with $PrP^{\Delta 1\text{-}66(HA)}$ similarly to cells transfected with $PrP^{(HA)}$. More importantly, induction of $AltPrP^{HA}$ by thapsigargin was abolished, confirming that bps 1-66 comprise at least part of the regulation domain for AltPrP expression from PrP cDNA (Fig. 5B,E).

Other stresses, including heat shock and oxidative stress did not induce any change in the levels of AltPrP. In these experiments, cellular stress was monitored by increased levels of phospho-eIF2α (Fig 5 C,D,E). Therefore, the synthesis of AltPrP from PrP cDNA is specifically regulated by ER stress.

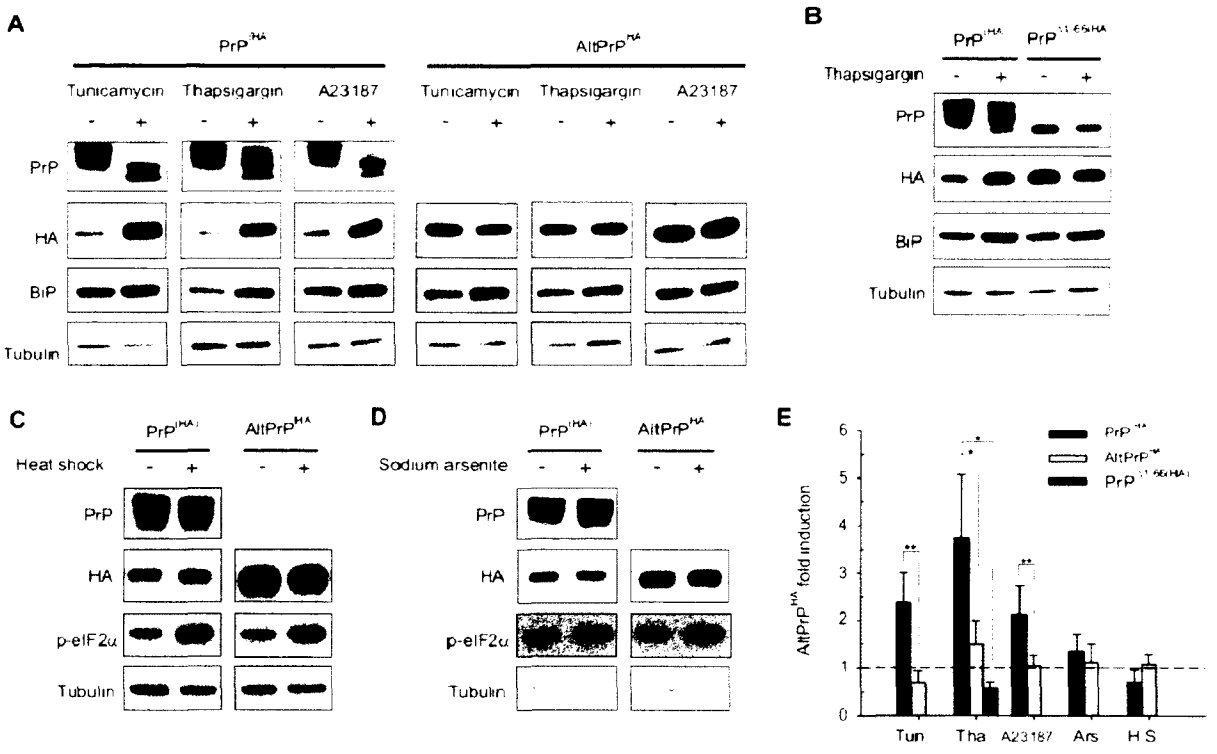


Figure 5. Levels of expression of AltPrP are regulated only by certain conditions of cellular stress. (**A**) Western blot of cells expressing $PrP^{(HA)}$ or $AltPrP^{HA}$ treated in the absence (-) or in the presence (+) of tunicamycin (1.5 μM), thapsigargin (3.77 μM) or A23187 (1.3 μM) for 24 h using PrP, HA, BiP, and tubulin antibodies. (**B**) Lysates of cells expressing $PrP^{(HA)}$ or $PrP^{\Delta 1\text{-}66(HA)}$ and incubated in the absence or in the presence of 3.5 μM thapsigargin for 24 h were immunoblotted for PrP, HA, BiP, and tubulin. (**C-D**) Western blot analysis of cells transfected with $PrP^{(HA)}$ or $AltPrP^{(HA)}$ and untreated or treated for 45 min at 42°C (heat shock) (**C**), or for 30 min with 0.5 mM sodium arsenite (**D**) with anti-PrP, -HA, -phospho-eIF2α (p-eIF2α) and -tubulin antibodies. Cells were allowed to recover for 8 h before harvesting. (**E**) Densitometric quantification of $AltPrP^{HA}$ induction under several cellular stress conditions. Tunicamycin (Tun.), thapsigargin (Tha.) and A23187 each significantly induced $AltPrP^{HA}$ expression from $PrP^{(HA)}$ as compared to the $AltPrP^{HA}$ construct. The stimulation of AltPrP expression in the presence of thapsigargin was abolished by the deletion of bps 1-66 relative to PrP's translation initiation site ($PrP^{\Delta 1\text{-}66(HA)}$). Control experiments using cells transfected with $PrP^{(HA)}$ and treated with heat shock (H.S.) or sodium arsenite (Ars.) did not produce significant changes in $AltPrP^{HA}$ expression levels. Dotted line represents $AltPrP^{HA}$ levels in untreated transfected cells normalized to 1. Values are expressed as the mean value (± s.d.) from at least three independent experiments. All statistical significances were determined by t-test. * $p<0.05$, ** $p<0.01$.

The absence of AltPrP has no obvious effect on PrP biology

Because PrP and AltPrP are produced from the same gene, it was important to test for a direct functional relation between these two proteins. We verified if the absence of AltPrP would alter the biology of PrP, namely its post-translational modifications, localization, and solubility. Since AltPrP expression is abolished when the AUG codon at bp 90 is mutated (Fig. 2B), we compared cells expressing either $PrP^{(HA)*}$ with cells expressing $PrP^{(HA)}$ to answer these questions. As a control, we also included cells transfected with wild-type PrP to ensure that each treatment was performed properly. PNGase F is a specific N-glycosidase commonly used to deglycosylate PrP in order to characterize its post-translational modifications. After PNGase F treatment of lysates of HEK293 cells transfected with human PrP, $PrP^{(HA)}$ and $PrP^{(HA)*}$, only the unglycosylated band was detected by western blot (data not shown). Thus, neither the HA tag nor the co-expression of AltPrP interfered with post-translational modifications of PrP. Similarly, AltPrP did not perturb the trafficking of PrP since both $PrP^{(HA)}$ and $PrP^{(HA)*}$ were mainly localized at the plasma membrane and in the Golgi apparatus (Fig. 2C). Finally, an essential characteristic of PrP is its ability to switch from a soluble into a disease-associated insoluble isoform (20). In order to test if the solubility of PrP was affected by co-expression of AltPrP, we monitored PrP, $PrP^{(HA)}$ and $PrP^{(HA)*}$ solubility in HEK 293 cells by ultracentrifugation. Each construct displayed the same solubility (data not shown). We concluded that the presence of AltPrP has no apparent effect on the biology of PrP. This result was not unexpected since under normal conditions, PrP and AltPrP are not localized in the same cellular compartments and are therefore unlikely to interact.

AltPrP is endogenously expressed from *PRNP*

Next, we raised and affinity-purified a polyclonal antibody against the C-terminus of human AltPrP to detect wild-type AltPrP encoded by the endogenous *PRNP* gene. The anti-AltPrP antibody was validated in western blot experiments using lysates from untransfected (Mock) HEK293 cells and cells transfected with human PrP or AltPrP. A band was detected with the expected molecular weight in lysates from PrP- and AltPrP-expressing cells (Fig. 6A). Addition of the immunogenic peptide to neutralize the antibodies completely prevented the detection of AltPrP. AltPrP also was not detected with pre-

immunized serum. In order to determine the expression of AltPrP from cells expressing endogenous PrP, we used human astrocytoma U-118 cells, which express high levels of PrP. A band was detected at the expected size in U-118 whole cell lysates, suggesting that endogenous AltPrP is indeed detected by the antibody (Fig. 6B). The identity of AltPrP was confirmed by treating the cells with siRNA against *PRNP*. Western blot analysis proves that AltPrP is endogenously expressed in these cells, since *PRNP* knock-down resulted in a 60% decrease of the intensity of the bands corresponding to both PrP and AltPrP (Fig. 6B,C). Furthermore, mitochondrial fractionation shows that, like $AltPrP^{HA}$, the band corresponding to AltPrP is enriched in mitochondria, and is absent in the cytosolic fraction (Fig. 6D). Importantly, expression of endogenous AltPrP was upregulated following treatment with thapsigargin, further indicating that the level of AltPrP expression might be increased during neurodegeneration (Fig. 6E,F). Following this experiment, primary human cells which express PrP were tested for AltPrP expression using the same antibody. In addition to U-118 cells, human primary neurons as well as human peripheral blood mononuclear cells (PBMC) express AltPrP, judging by the prominent band at the same molecular weight as the band present in a control cell lysate transfected with AltPrP (Fig. 6G). Furthermore, a healthy human brain homogenate was tested for AltPrP expression using the same technique. As expected, the homogenate expressed a band at the same molecular weight (Fig. 6H).

Overall, these results clearly establish that, in different tissues, primary cells, and cell lines, two protein products, PrP and AltPrP, are endogenously expressed from two distinct overlapping reading frames present in the *PRNP* gene.

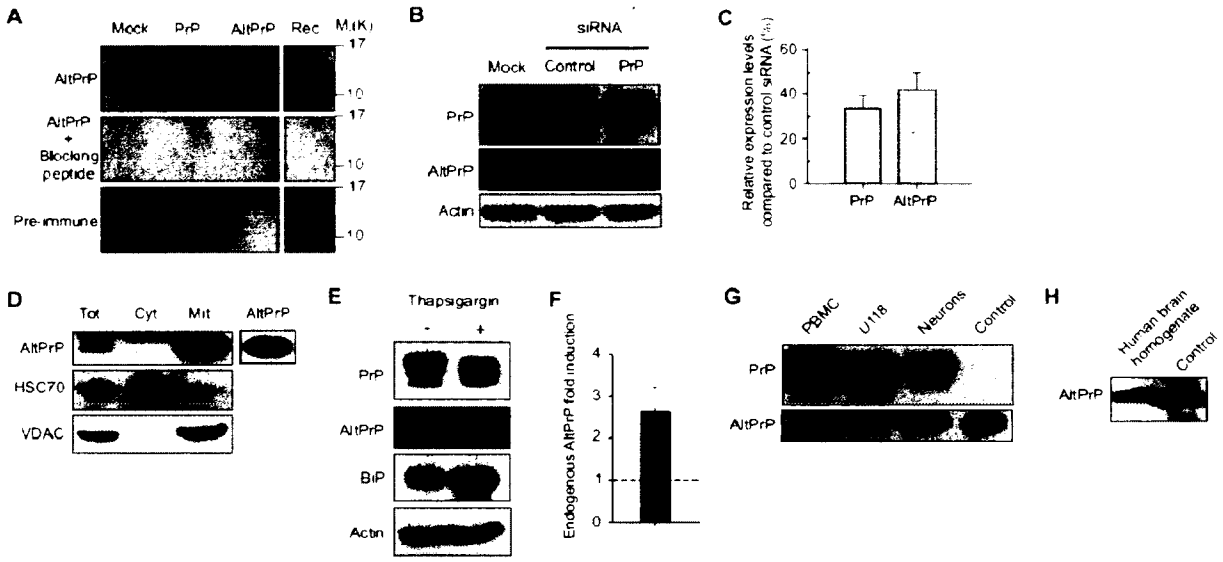


Figure 6. AltPrP is expressed in cells expressing endogenous PrP. (**A**) Characterization of a polyclonal antibody raised against AltPrP. HEK293 cells were either transfected with human PrP, AltPrP, or untransfected (Mock) and were probed for AltPrP using an antibody raised against the C-terminus of AltPrP. A band with the same molecular weight as recombinant AltPrP (Rec.) is detected in cells transfected either with PrP or AltPrP. The same lysates were also probed with the anti-AltPrP antibody blocked with the immunogenic peptide against which the antibody was targeted, or with the animal's pre-immunized serum in order to demonstrate the specificity of the antibody. (**B**) Endogenous AltPrP was detected in U-118 cells, which express high levels of PrP. An siRNA against the 3'UTR of *PRNP* (PrP) notably reduced the expression of PrP as well as the band corresponding to AltPrP at comparable levels, whereas a control siRNA did not change PrP and AltPrP expression. (**C**) Densitometric analysis of siRNA treatment revealed that U-118 cells transfected with an siRNA against PrP showed a decrease (60-70%) in expression of both PrP and AltPrP following treatment. No statistically significant difference was observed between the knock-down levels of PrP and AltPrP by paired t-test (n=2). (**D**) Total cell extracts (Tot.), mitochondrial (Mit.) and cytoplasmic (Cyt.) fractions from U-118 cells were immunoblotted for AltPrP, VDAC, a mitochondrial marker, and HSC70, a marker of the cytosol. The band corresponding to AltPrP was enriched in mitochondria, as expected. AltPrP represents a lysate from cells transfected with AltPrP and is used as a control. (**E**) Western blot of cells human astrocytoma U-87 cells expressing endogenous PrP treated in the absence (-) or in the presence (+) of thapsigargin (3.77 μM) for 24 h using PrP, AltPrP, BiP, and actin antibodies. (**F**) Densitometric quantification of endogenous AltPrP induction in the presence of thapsigargin. Dotted line represents AltPrP levels in untreated U-87 cells normalized to 1. Value is expressed as the mean value (± s.d.) from at least four independent experiments. (**G**) Detection of endogenous AltPrP in several cell types. Human peripheral blood mononuclear cells (PBMC), U-118 cells and human primary neurons (neurons) were immunoblotted for both PrP and AltPrP. All cell types tested showed a prominent band at the same molecular weight as a positive control consisting of cells transfected with human AltPrP (Control). (**H**) A human brain homogenate was probed for AltPrP using the anti-AltPrP antibody. A band at the same

molecular weight as the one seen in cells transfected with human AltPrP was detected (Control).

Discussion

In this study, we provide evidence that out-of-frame alternative translation initiation in the human, sheep, bovine, and deer *PRNP* gene, which encodes the cellular prion protein, results in the synthesis of a novel polypeptide that we termed AltPrP. In-frame alternative translation in PrP mRNA has been reported to produce N-terminally truncated cytoplasmic or nucleocytoplasmic forms of hamster, human and sheep PrP, which might have a different function from that of GPI-anchored PrP (33,34). Here, we propose that co-expression of AltPrP along with PrP represents an additional level of complexity regarding the functions attributed to the *PRNP* locus. The production of several polypeptides from this single gene, with modulation depending on cellular conditions, represents a very plausible explanation for the difficulty in assessing the physiological function of PrP, and more generally of the *PRNP* locus. Knock-down/out of PrP expression by targeted inactivation of PrP mRNA or the *PRNP* gene most likely results in knock-down of all PrP isoforms as well as AltPrP (Fig. 6C). In order to get insights into the molecular mechanisms of PrP function, these considerations should be taken into account in the interpretation of experimental results in the future.

The investigation of AltPrP function(s) is likely to shed some light on the physiological role of PrP, as controversies have emerged in the literature (35,36,37). We propose that part of the difficulties in assessing the physiological relevance of the *PRNP* gene could be attributed to the co-expression of AltPrP together with PrP from the *PRNP* gene. For instance, the neuroprotective role of PrP has been largely debated, and seems to be context-dependent (38,39). Because AltPrP is localized at the mitochondria (Fig. 3), an essential organelle for energy metabolism, stress response and apoptosis (among other functions), it is tempting to think that some toxic/protective functions have been misattributed to PrP. Though these speculations remain to be proven, we have provided evidence that support this hypothesis. The increased level of AltPrP synthesis under conditions of ER stress (Fig. 5) represents a characteristic feature of various proteins

implicated in the unfolded protein response, that participate to either stress recovery, or induction of apoptosis if cellular damage is irreversible (30). It is to be noted that AltPrP does not seem to be pro-apoptotic on its own, since we were able to maintain cell lines stably over-expressing $AltPrP^{HA}$ without any noticeable change in growth rate or cellular morphology (data not shown). However, one can imagine that AltPrP might participate in either a pro- or anti-apoptotic response under specific cellular conditions.

An interesting consideration regarding the discovery of AltPrP is its high degree of conservation and homology in a large number of mammals (Fig. 1). Although numerous studies have demonstrated the conservation of the C-terminal portion of PrP among different species explained by the highly structured nature of this domain, as well as its necessity for forming PrP^{Sc} (40,41), there still lacks a convincing argument explaining why the unstructured N-terminal domain of PrP is conserved as well (40,42,43,44). Despite the fact that the function of AltPrP remains unknown, this newly discovered protein may be the reason for the conservation of the N-terminal portion of PrP. Selective pressure might act on the *Prnp* gene in order to conserve AltPrP rather than the unstructured fragment of PrP. At this point, this is merely a speculation, but it provides another compelling reason to continue studying both the *Prnp* gene as well as AltPrP.

The role of PrP in the pathogenesis of TSEs is very well established, as self-templated transition of the native isoform of PrP (PrP^{C}) to its misfolded, disease-linked isoform (PrP^{Sc}) is at the origin of the neurotoxic mechanisms implicated in these neurodegenerative disorders. As a result, the *PRNP* locus is the main genetic risk factor in these diseases (45). Since this gene also directs the synthesis of AltPrP, the possibility exists that this novel polypeptide might be linked to TSEs. AltPrP could be functionally implicated in TSEs, although our data suggests that basic features of PrP biology, including glycosylation, localization and solubility, do not seem to depend on co-expression of AltPrP. Our results are based on a cell culture system, a model that does not reproduce the events taking place at the level of a whole organism during disease. Moreover, it is plausible that the as-yet unknown function of AltPrP will have no direct effect on PrP biology. Nevertheless, the possibility that AltPrP might participate to the pathogenesis of TSEs deserves further investigation and determining its function will likely answer this question. Whether AltPrP will be proven to have a functional relationship with the

pathogenesis of TSEs or not, the fact that its expression is greatly enhanced following proteasome inhibition and ER stress (Fig. 4 and 5) makes it a potential biomarker for neurodegenerative disorders. Indeed, these two cellular stresses have been shown to be hallmarks not only of TSEs, but also of other neurodegenerative diseases such as Huntington's, Parkinson's and Alzheimer's diseases (27,28,46).

Other examples of polypeptides synthesized from out-of-frame downstream alternative translation initiation sites exist in the literature, although this mechanism occurs most often in viruses (47,48,49). Very few mammalian examples of this phenomenon have so far been described (12,13). Klemke and colleagues have discovered and characterized ALEX, a protein encoded by an alternative ORF in the XLalphas/Galphas gene, and that is expressed in vivo in both rats and humans. Cryptic T-cell epitopes represent other examples of this phenomenon in humans. However, except for their immunogenic potential, these polypeptides have not been further characterized (50). To our knowledge, AltPrP represents the second extensively characterized product of out-of-frame alternative translation initiation in a human gene. The discovery of AltPrP could not only change our understanding of the *PRNP* locus, but the fact that AltPrP is endogenously expressed supports the idea that out-of-frame alternative translation initiation in mammals is very likely to participate in protein diversity. This reveals an additional level of complexity to the proteome and to gene function, which deserves further investigation in order to determine its physiological importance.

Acknowledgements

We thank Debbie McKenzie and Judd Aiken for their gift of the deer PrP cDNA clone, Michael Tranulis for providing the sheep PrP cDNA clone, and David Fortin's kind gift of the U-118 cell line. Human brain tissue was kindly provided by Dr. Cheryl Wellington. We also thank Catherine Grenier for her technical help in the initial stages of this project, and Alireza Roostaee for critical review of the manuscript. This research was funded by PrioNet Canada. XR is a senior research scholar from the Fonds de la Recherche en Santé du Québec.

References

1. Stahl, N., Borchelt, D.R., Hsiao, K., and Prusiner, S.B. (1987) Scrapie prion protein contains a phosphatidylinositol glycolipid. *Cell.* **51**, 229-40
2. McKinley, M.P., Bolton, D.C., and Prusiner, S.B. (1983) A protease-resistant protein is a structural component of the scrapie prion. *Cell.* **35**, 57-62
3. Prusiner, S.B., Groth, D., Serban, A., Koehler, R., Foster, D., Torchia, M., Burton, D., Yang, S.L., and DeArmond, S.J. (1993) Ablation of the prion protein (PrP) gene in mice prevents scrapie and facilitates production of anti-PrP antibodies. *Proc Natl Acad U S A.* **90**, 10608-12
4. Büeler, H.R., Aguzzi, A., Sailer, A., Greiner, R.A., Autenried, P., Aguet, M., and Weissmann, C. (1993) Mice devoid of PrP are resistant to scrapie. *Cell.* **73**, 1339–47
5. Sailer, A., Büeler, H., Fischer, M., Aguzzi, A., and Weissmann, C. (1994) No propagation of prions in mice devoid of PrP. *Cell.* **77**, 967–68
6. Mead, S. (2006) Prion disease genetics. *Eur J Hum Genet.* **14**, 273-81
7. Aguzzi, A. and Calella, A.M. (2009) Prions: protein aggregation and infectious diseases. *Physiol Rev.* **89**, 1105-52
8. Kochetov, A.V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays.* **30**, 683-91
9. Lee, Y.Y., Cevallos, R.C., and Jan, E. (2008) An upstream open reading frame regulates translation of GADD34 during cellular stresses that induce eIF2alpha phosphorylation. *J Biol Chem.* **284**, 6661-73
10. Calvo, S.E., Pagliarini, D.J., and Mootha, V.K. (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad U S A.* **106**, 7507-12
11. Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell.* **44**, 283-92.
12. Klemke, M., Kehlenbach, R.H., and Huttner, W.B. (2001) Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *EMBO J.* **20**, 3849-60

13. Ho, O. and Green, W.R. (2006) Alternative translational products and cryptic T cell epitopes: expecting the unexpected. *J Immunol.* **177**, 8283-9

14. Roucou, X., Guo, Q., Zhang, Y., Goodyer, C.G., LeBlanc, A.C. (2003) Cytosolic prion protein is not toxic and protects against Bax-mediated cell death in human primary neurons. *J Biol Chem.* **278**, 40877-81

15. Roostaee, A., Côté, S., and Roucou, X. (2009) Aggregation and amyloid fibril formation induced by chemical dimerization of recombinant prion protein in physiological-like conditions. *J Biol Chem.* 284, 30907-16

16. Wessel, D. and Flugge, U.I. (1984) A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal Biochem.* **138**, 141-3

17. Roucou, X., Giannopoulos, P.N., Zhang, Y., Jodoin, J., Goodyer, C.G., and LeBlanc, A. (2005) Cellular prion protein inhibits proapoptotic Bax conformational change in human neurons and in breast carcinoma MCF-7 cells. *Cell Death Differ.* **12**, 783-95

18. Beaudoin, S., Vanderperre, B., Grenier, C., Tremblay, I., Leduc, F., and Roucou, X. (2009) A large ribonucleoprotein particle induced by cytoplasmic PrP shares striking similarities with the chromatoid body, an RNA granule predicted to function in posttranscriptional gene regulation. *Biochim Biophys Acta.* **1793**, 335-45

19. Eskes, R., Desagher, S., Antonsson, B., and Martinou, J.C. (2000) Bid induces the oligomerization and insertion of Bax into the outer mitochondrial membrane. *Mol Cell Biol.* **20**, 929-35

20. Daude, N., Lehmann, S., and Harris, D.A. (1997). Identification of intermediate steps in the conversion of a mutant prion protein to a scrapie-like form in cultured cells. *J Biol Chem.* **272**, 11604-12

21. Dubot, A., Godinot, C., Dumur, V., Sablonnière, B., Stojkovic, T., Cuisset, J.M., Vojtiskova, A., Pecina, P., Jesina, P., and Houstek, J. (2004) GUG is an efficient initiation codon to translate the human mitochondrial ATP6 gene. *Biochem Biophys Res Commun.* **313**, 687-693

22. von Heijne, G. (1992) Membrane Protein Structure Prediction: Hydrophobicity

Analysis and the 'Positive Inside' Rule. *J Mol Biol.* **225**, 487-94

23. Zeng, L., Liu, Y.P., Sha, H., Chen, H., Qi, L., and Smith, J.A. (2010) XBP-1 couples endoplasmic reticulum stress to augmented IFN-beta induction via a cis-acting enhancer in macrophages. *J Immunol.* **185**, 2324-30
24. Horowitz, M. (2010) Genomics and proteomics of heat acclimation. *Front Biosci (Schol Ed).* **2**, 1068-80
25. Ding, Q., Keller, J.N. (2003) Does proteasome inhibition play a role in mediating neuropathology and neuron death in Alzheimer's disease? *J Alzheimers Dis.* **5**, 241-5
26. Deriziotis, P. and Tabrizi S.J. (2008) Prions and the proteasome. *Biochim Biophys Acta.* **1782**, 713-22
27. Salminen, A. Kauppinen, A., Suuronen, T., Kaarniranta, K., and Ojala, J. (2009) ER stress in Alzheimer's disease: a novel neuronal trigger for inflammation and Alzheimer's pathology. *J Neuroinflammation.* **6**, 41
28. Yoshida, H. (2007) ER stress and diseases. *FEBS J.* **274**, 630-58
29. Ma, Y. and Hendershot, L.M. (2003). Delineation of a negative feedback regulatory loop that controls protein translation during endoplasmic reticulum stress. *J Biol Chem.* **278**, 34864-73
30. Holcik, M. and Sonenberg, N. (2005). Translational control in stress and apoptosis. *Nat Rev Mol Cell Biol.* **6**, 318-27
31. Kozutsumi, Y., Segal, M., Normington, K., Gething, M.J., and Sambrook, J. (1988) The presence of malfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. *Nature.* **332**, 462-4
32. Orsi, A., Fioriti, L., Chiesa, R., and Sitia, R. (2006) Conditions of endoplasmic reticulum stress favor the accumulation of cytosolic prion protein. *J Biol Chem.* **281**, 30431-8
33. Lund, C., Olsen, C.M., Skogtvedt, S., Tveit, H., Prydz, K., and Tranulis, M.A. (2009) Alternative translation initiation generates cytoplasmic sheep prion protein. *J Biol Chem.* **284**, 19668-78
34. Juanes, M.E., Elvira, G., Garcia-Grande, A., Calero, M., and Gasset, M. (2009) Biosynthesis of prion protein nucleocytoplasmic isoforms by alternative

initiation of translation. *J Biol Chem.* **284**, 2787-94

35. Martins, V.R., Mercadante A.F., Cabral A.L., Freitas, A.R., and Castro, R.M. (2001) Insights into the physiological function of cellular prion protein. *Braz J Med Res.* **34**, 585-95

36. Sakudo, A. and Ikuda, K. (2009) Prion protein functions and dysfunction in prion diseases. *Curr Med Chem*, **16**, 380-9

37. Aguzzi, A., Baumann, F., and Bremer, J. (2008) The prion's elusive reason for being. *Annu Rev Neurosci.* **31**, 439-77

38. Steinacker, P., Hawlik, A., Lehnert, S., Jahn, O., Meier, S., Görz, E., Braunstein, K.E., Krzovska, M., Schwalenstöcker, B., Jesse, S., Pröpper, C., Böckers, T., Ludolph, A., and Otto, M. (2010) Neuroprotective function of cellular prion protein in a mouse model of amyotrophic lateral sclerosis. *Am J Pathol.* **176**, 1409-20

39. Steele, A.D., Zhou, Z., Jackson, W.S., Zhu, C., Auluck, P., Moskowitz, M.A., Chesselet, M.F., and Lindquist, S. (2009) *Prion.* **3**, 240-9

40. Wopfner, F., Weidenhöfer, G., Schneider, R., von Brunn, A., Gilch, S., Schwarcz, T.F., Werner, T., and Schätzl, H.M. (1999). Analysis of 27 mammalian and 9 avian PrPs reveals a high conservation of flexible regions of the prion protein. *J Mol Biol.* **289**, 1163-78

41. Rogers, M., Yehiely, F., Scott, M., and Prusiner, S.B. (1993). Conversion of truncated and elongated prion proteins into the scrapie isoform in cultured cells. *Proc Natl Acad Sci U S A.* **90**, 3182-6

42. Krakauer, D.C., Zanotto, P.M., and Pagel, M. (1998). Prion's progress: patterns and rates of molecular evolution in relation to spongiform disease. *J Mol Evol.* **47**, 133-45

43. Premzl, M. and Gamulin, V. (2009). Positive selection in prion protein. *J Mol Evol.* **68**, 205-7

44. Harrison, P.M., Khachane, A., and Kumar, M. (2010). Genomic assessment of the evolution of the prion protein gene family in vertebrates. *Genomics.* **95**, 268-77

45. Mastrangelo, P. and Westaway, D. (2001). Biology of the prion gene complex.

Biochem Cell Biol. **79**, 613-28

46. Ding, Q., Dimayuga, E., and Keller, J.N. (2006) Proteasome regulation of oxidative stress in aging and age-realted diseases of the CNS. *Antioxid Redox Signal.* **8**, 163-72
47. Kozak, M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene.* **299**, 1-34
48. Yamasaki, K., Weihl, C.C., and Roos, R.P. (1999) Alternative translation initiation of Theiler's murine encephalomyelitis virus. *J Virol.* **73**, 8519-26
49. Branch, A.D., Stump, D.D., Gutierrez, J.A., Eng, F., and Walewski, J.L. (2005) The hepatitis C virus alternate reading frame (ARF) and its family of novel products: the alternate reading frame protein/F-protein, the double-frameshift protein, and others. *Semin Liver Dis.* **25**, 105-17
50. Wang, R.F., Parkhurst, M.R., Kawakami, Y., Robbins, P.F., and Rosenberg, S.A. (1996) Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J Exp Med.* **183**, 1131-40

ARTICLE 2

HAltORF: a database of predicted out-of-frame alternative open reading frames in human

Auteurs de l'article: Benoît Vanderperre, Jean-François Lucier, Xavier Roucou

Statut de l'article: publié dans *Database (Oxford)*, 2012:bas025, Mai 2012.

Avant-propos: J'ai amené l'idée originale à la base de cet article, à savoir la construction d'une base de données de cadres ouverts de lectures *out-of-frame* à partir du transcriptome humain. J'ai rédigé l'ensemble du manuscrit et du texte contenu sur le site web associé à l'exception des sections « Database Generation ». J'ai également créé les figures insérées dans le manuscrit. Sur mes directives et celles de mon directeur le Dr Xavier Roucou, Jean-François Lucier a généré la base de données HAltORF, et créé l'interface web.

Résumé : « Human alternative open reading frames (HAltORF) » est une base de données accessible au public et consultable en ligne qui référence des produits putatifs d'initiation alternative de la traduction (IAT) « out-of-frame » dans les ARNms humains. L'IAT « out-of-frame » est un processus par lequel un seul ARNm peut encoder des protéines indépendantes, lorsque des codons d'initiation distincts situés dans différents cadres de lecture sont reconnus par un ribosome pour initier la traduction. Ce mécanisme est largement utilisé chez les virus, augmentant le potentiel codant des petits génomes viraux. De plus en plus de preuves indiquent que l'IAT « out-of-frame » est également utilisée chez les eucaryotes, y compris l'humain, et pourrait contribuer à la diversité du protéome humain. HAltORF est la première base de données consultable en ligne qui permet une recherche approfondie dans le transcriptome humain de cadres ouverts de lecture « out-of-frame » ayant un codon d'initiation situé dans un contexte Kozak fort, et étant donc plus susceptibles d'être exprimés. C'est également la première étude à large échelle sur le transcriptome humain à prédire avec succès l'expression de produits protéiques issus d'IAT « out-of-frame » précédemment découverts expérimentalement. HAltORF sera un outil

utile pour l'identification de gènes humains avec des séquences codantes multiples, et aidera à mieux définir et comprendre la complexité du protéome humain.

HAltORF: a database of predicted out-of-frame alternative open reading frames in human

Benoît Vanderperre[1], Jean-François Lucier[2], Xavier Roucou[1]

[1]Département de Biochimie, [2]Département de Microbiologie et d'infectiologie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Québec J1E 4K8, Canada

Address correspondence to Dr Xavier Roucou, Département de Biochimie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Québec J1E 4K8, Canada, Tel. (819) 821 8000x72240 ; Fax. (819) 820 6831; E-mail: xavier.roucou@usherbrooke.ca

Abstract

Human alternative open reading frames (HAltORF) is a publicly available and searchable online database referencing putative products of out-of-frame alternative translation initiation (ATI) in human mRNAs. Out-of-frame ATI is a process by which a single mRNA encodes independent proteins, when distinct initiation codons located in different reading frames are recognized by a ribosome to initiate translation. This mechanism is largely used in viruses to increase the coding potential of small viral genomes. There is increasing evidence that out-of-frame ATI is also used in eukaryotes, including human, and may contribute to the diversity of the human proteome. HAltORF is the first web-based searchable database that allows thorough investigation in the human transcriptome of out-of-frame alternative open reading frames with a start codon located in a strong Kozak context, and are thus the more likely to be expressed. It is also the first large scale study on the human transcriptome to successfully predict the expression of out-of-frame ATI protein products that were previously discovered experimentally. HAltORF will be a useful tool for the identification of human genes with multiple coding sequences, and will help to better define and understand the complexity of the human proteome.

Database URL: http://haltorf.roucoulab.com/.

Introduction

Each eukaryotic mRNA encoding a protein is usually associated with only one open reading frame (herein called reference ORF) or coding sequence (CDS) delineated by a start codon (most of the time AUG) and a stop codon, required to initiate and end translation, respectively. This simplistic view is however being challenged by the existence of at least two mechanisms resulting in increased protein diversity. In-frame alternative translation initiation (ATI) at downstream AUG codons allows the production of truncated protein isoforms with new functions or localization and is a well-characterized mechanism in eukaryotes (1,2). Out-of-frame ATI at the start codon of alternative ORFs (AltORFs) in the two other reading frames is a second mechanism producing proteins with an amino acid sequence completely different from the reference protein. The nomenclature regarding reading frames used thereafter is the following (3). The +1 reading frame is determined by the coding sequence of the reference ORF for each transcript (independently of the gene or transcript). Hence, the annotated reference ORF is defined as frame +1, and there are two possible frames for AltORFs: frame +2 and frame +3.

The presence of overlapping ORFs and the use of out-of-frame ATI are well described in viruses (4–6) and provide small viral genomes with an increased coding capacity. In addition, a database referencing putative alternative ORFs in many prokaryotic genomes already exists (7). The role of out-of-frame ATI in eukaryotes has been overlooked. Yet, there is some evidence that proteins derived from AltORFs can affect physiological as well as pathological aspects of gene function. This is the case for the alternative protein ALEX encoded in the *GNAS* gene (8,9). In addition, we recently discovered the endogenous expression in human of an alternative protein product termed AltPrP which ORF (+3 reading frame) partially overlaps with the prion protein CDS (Figure 1) (10). Four other examples exist in human (11–14), which correspond to peptides that are targeted by anti-tumor responses in several types of cancers, and may thus serve as biomarkers or therapeutic targets (15). Interestingly, these AltORFs are all but one included within the reference ORF (11). This observation is critical since the expression of cDNAs composed solely of the CDS in experimental systems such as cultured cells may actually result in the expression of more than one protein (10). Consequently, co-expression of an alternative

protein together with the reference protein in functional studies likely result in unnoticed confounding results. A database containing a list of all human mRNAs containing AltORFs overlapping with the reference ORF is important to identify potential genes with multiple CDS.

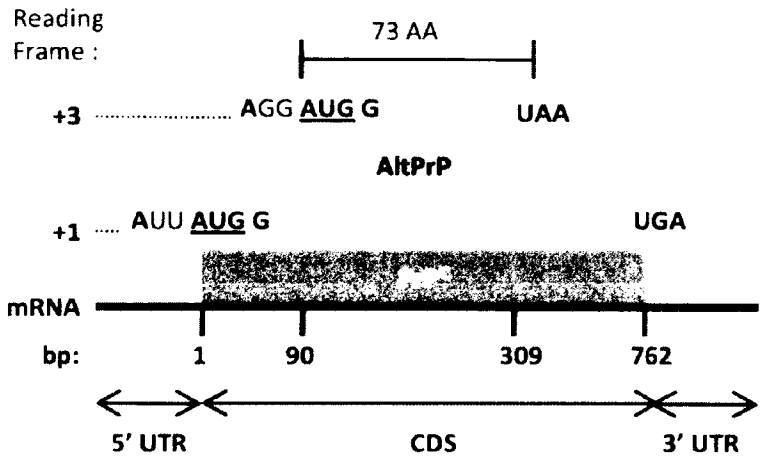


Figure 1. AltPrP, a typical example of AltORFs in the HAltORF database. All mRNAs produced from the *PRNP* gene have the same reference ORF (nt 1–762, gray box) which encodes the prion protein (PrP^C) in the +1 reading frame. An AltORF (white box) is present in the +3 reading frame (nt 90–309). Similar to all AltORFs present in the database, the alternative prion protein (AltPrP) encoding AltORF is entirely included in the CDS of the reference protein, and encodes a protein longer than 24 amino acids (minimum size threshold). Additionally, its AUG codon is in a different reading frame than the reference protein, and is located in an optimal Kozak context (shown in bold; consensus: **A/GNNAUGG**).

To our knowledge, three bioinformatics genome-wide studies aiming at the identification of AltORFs in mammals have been performed previously (16–18). However, none of them provided an online searchable option with links to GenBank and NCBI databases for further investigation. In one study, criteria such as conservation among species and a minimum length of 500 bp for the predicted AltORFs were used and only 40 putatively expressed AltORFs were referenced (16). In a more recent study, 138 potential dual coding transcripts were identified in human (18). In another study, a filter of a minimal length of 150 bp was applied and 1793 AltORFs were found to be conserved among rat, mouse and human (17). When the 1793 human AltORFs were filtered for the presence of an optimal Kozak context around the initiator AUG codon, known to be extremely important for efficient initiation of translation (19), this number dropped to 217 putative AltORFs. One objective of these three studies was to predict high confidence candidate AltORFs, and the highly stringent criteria used were extremely pertinent in this matter. However, they were unsuccessful in predicting the expression of two experimentally proven AltORFs, AltPrP and ALEX. For all these reasons, it is obvious that a less stringent and potentially more comprehensive large scale bioinformatics analysis of AltORFs in the human transcriptome and a publicly available and searchable online database of predicted AltORFs are lacking.
Human alternative open reading frames (HAltORF; http:// haltorf.roucoulab.com/) is the first web-based searchable database that allows thorough investigation in the human transcriptome of AltORFs overlapping with annotated CDS, and putatively expressed by out-of-frame ATI. It is also the first large scale study on the human transcriptome to successfully predict the expression of AltPrP and ALEX, two experimentally discovered out-of-frame ATI protein products. HAltORF will be a useful tool for the identification of genes containing multiple CDS in human, and will help to better define and understand the complexity of the human proteome.

Database generation

The HAltORF database was built using a pipeline of Perl scripts that populate a MySQL database. All GenBank human mRNA and protein entries (release 37) were downloaded from the NCBI website (http://www.ncbi.nlm.nih.gov/), and each mRNA was associated

with its reference protein. For each mRNA, in silico translation of the full sequence was performed using the Transeq software (20), and subsequent comparison of the results with the amino acid sequence of the reference protein allowed to map the translation start and stop sites coordinates of the reference ORF on its corresponding mRNA. The sequence 5' of the translation start site of the reference ORF was then deleted. This action set the reading frame associated with the reference ORF in each mRNA to +1. The remaining sequence was then translated again using the Transeq software. All translation results equal to or above 24 amino acids, regardless of the reading frame, were stored in the database along with their start and stop sites coordinates. The arbitrary threshold of 24 amino acids was selected to reduce the database to an acceptable size, since we (data not shown) and other groups (16,17) noticed that the numbers of predicted AltORFs increases as the size threshold decreases. Additionally, the validation of the expression of smaller peptides by standard techniques, such as SDS–PAGE and western blots, would be technically too challenging. Next, based on a simplified consensus Kozak sequence (**A/GNNATGG**) known to be favorable for efficient translation initiation (19), we determined for each predicted ORF start site if it was located in a strong (perfect fit to the consensus) or weak (any other sequence) Kozak context. The last step was to select, in the CDS of each mRNA, the putative AltORFs that are the most likely to be expressed. To do so, we filtered the database using the following criteria: (i) ORFs had to be in the +2 or +3 reading frames to be selected, thus storing AltORFs, which are currently absent from existing protein databases; (ii) the predicted AltORFs had to possess a strong Kozak context around their AUG codon, to increase the chance of efficient translation initiation; (iii) the stop site of the AltORFs had to be located prior to the stop site of the reference ORFs, thus removing ORFs that are not entirely contained within the CDS of the reference protein. More details on the construction of the database are available on the HAltORF website. For a typical example of AltORFs found in this new database, see Figure 1.

Database content

We identified 17096 distinct predicted AltORFs in the CDS of 31422 mRNAs (41.2% of total human mRNAs) transcribed from 8744 genes (42,5% of total human genes). 14195

(83%) are located in the +2 reading frame, and 2901 (17%) are located in the +3 reading frame.

For each AltORF, the gene name and accession number of the mRNA in which it is encoded are provided. Other information can also be found, including the reference protein produced from the corresponding mRNA, the coordinates of the start and stop codon of both the reference ORF and the alternative ORF in the mRNA, and the predicted length and amino acid sequence of the alternative protein.

Web interface

The HAltORF database (http://haltorf.roucoulab.com/) can be searched by gene name or symbol, by mRNA or protein GenBank accession number, and by protein sequence (with a minimum of 5 amino acids). Detailed explanations on how to perform a search and how results are displayed are available on the website under the Documentation tab. The search results are summarized in a table containing information for each retrieved AltORF, including the gene symbol, mRNA and reference protein accession numbers, reading frames, the location of the reference and alternative ORFs on the mRNA sequence, and the alternative protein length (Figure 2). The nucleotide numbers indicating the location of the ORFs are the first nucleotide of the start codon, and the first nucleotide of the stop codon, respectively. If multiple transcript variants exist for a given gene, all variants containing an alternative ORF are listed. If a search by protein sequence is performed, the table includes a supplementary column displaying part of the alternative protein sequence matching the query sequence. For each retrieved alternative ORF, a detailed result page is accessible through a link and provides the user with basic information concerning the reference mRNA and protein. Links to the NCBI website are also provided to help the user retrieve supplementary information on the gene, mRNA and reference protein associated with the AltORF. The detailed result page also contains an alignment section where the reference and alternative protein sequences are aligned on the reference mRNA sequence (Figure 2). The complete HAltORF database can be freely downloaded in Microsoft Excel or FASTA format under the download tab. The complete MySQL data dump is also available in this

section, thus providing developers with the possibility to predict other AltORFs using different parameters such as the length of AltORFs for example.

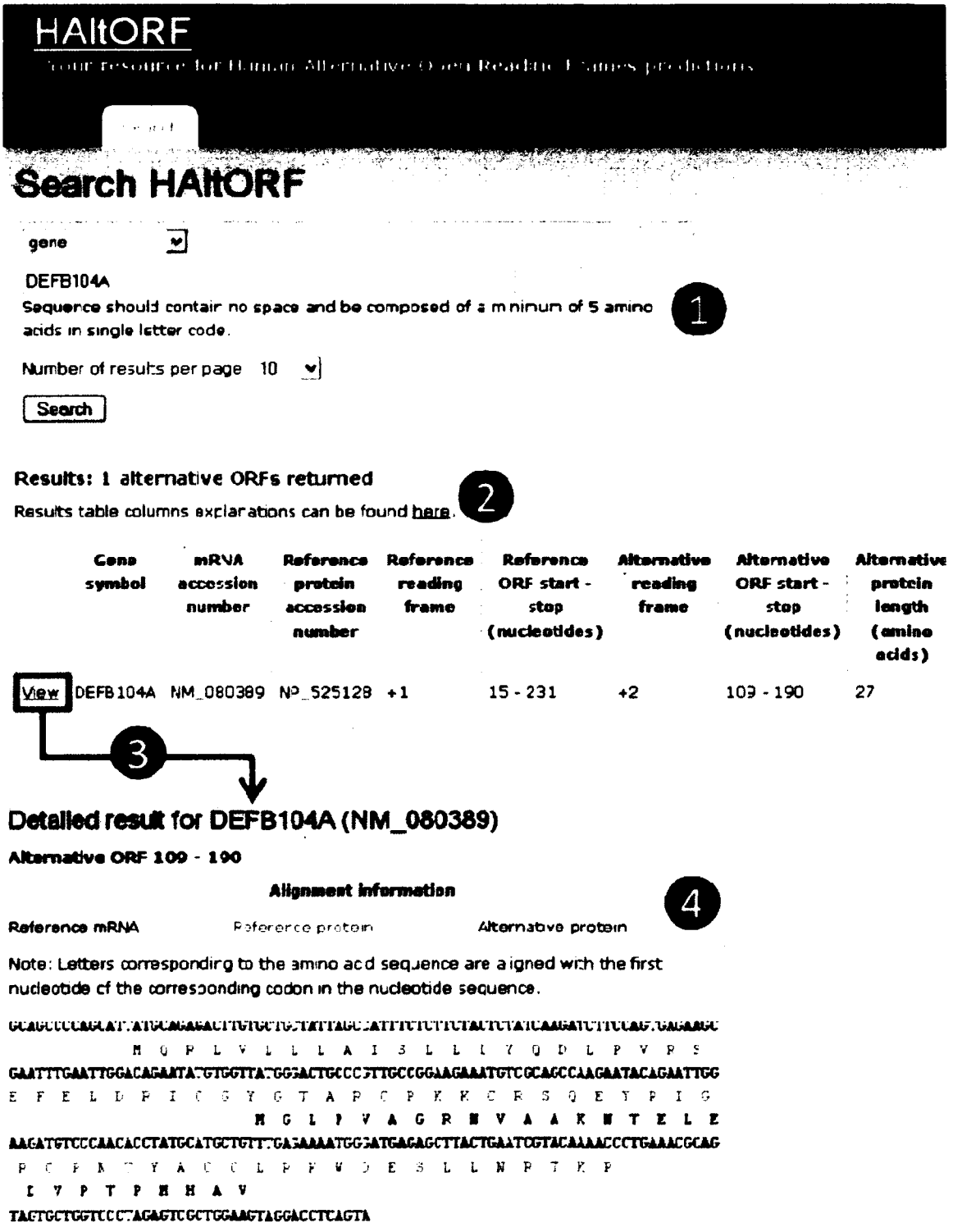


Figure 2. Snapshot of a typical search and associated results pages. (**1**) Search by gene (DEFB104A, which encodes the b-defensin 104 protein). (**2**) The number of corresponding AltORFs is indicated, and details on each AltORF are summarized in a table. Although this is not the case for this particular example, note that for a single gene, all AltORFs present in each transcript variants would be listed. The reference ORF is by definition in the +1 frame, and the alternative ORFs is in the +2 frame in this example. The nucleotide numbers indicating the location of the ORFs are the first nucleotide of the start codon, and the first nucleotide of the stop codon, respectively. (**3**) A detailed result page is available for each AltORF through the 'View' link. (**4**) In the detailed result page, basic information on the gene and mRNA of origin as well as the associated reference protein are displayed along with links to GenBank for each of these items (not shown). An alignment of the reference (blue letters) and alternative (green letters) protein sequences on the reference mRNA sequence (black letters) is provided.

Relevance and research avenues

The number of predicted AltORFs present in HAltORF is much greater when compared to other studies (16–18). This can be explained by different reasons. In particular, we used a lower cut-off for the size of AltORFs, and chose not to consider criteria such as conservation among species and specific codon usage. However, in our approach, we have established several limits, including AUG initiation codons located in an optimal Kozak context. Expression from AUG codons in the absence of an optimal Kozak sequence or from non-traditional CUG sites (21,22) is also possible and may be included in further studies. Nevertheless, the reduced stringency of our approach resulted in the successful prediction of AltPrP and ALEX, two experimentally well-characterized out-of-frame ATI products. It is likely that at least one of the several functions previously attributed to the prion protein is actually catalyzed by AltPrP (10), and we expect that some paradoxical experimental results regarding the function of other genes might be explained by multiple coding as well. This example highlights the fact that conservation along evolution of an alternative ORF is not necessary to be biologically relevant since the initiation codon for AltPrP is present in higher order mammals but not in lower mammals, including rodents (10). In addition, the presence of ALEX in HAltORF, for which polymorphisms have been associated with inherited neurological problems and increased trauma-related bleeding tendency (9), indicates that HAltORF could be valuable for the identification of biologically important AltORFs in human genes with multiple CDS. Last but not least, the

complete database may help mass spectrometry services to identify the great proportion of unknown peptides in their data sets which cannot be currently matched to any protein in existing databases. Altogether, HAltORF will help in the meticulous exploration of this potential alternative proteome which has been largely overlooked to date.

Funding

The Canadian Institutes for Health Research to XR [grant number MOP-89881]. X.R. is a senior research scholar from the Fonds de la Recherche en Santé du Québec. Funding for open access charge: The Canadian Institutes for Health Research [grant number MOP-89881].

Conflict of interest: none declared.

References

1. Kochetov,A.V. (2006) Alternative translation start sites and their significance for eukaryotic proteome. Mol. Biol. (Mosk), 40, 788–795.
2. Kochetov,A.V. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays, 30, 683–691.
3. Veloso,F., Riadi,G., Aliaga,D. et al. (2005) Large-scale, multi-genome analysis of alternate open reading frames in bacteria and archaea. Omics, 9, 91–105.
4. Branch,A.D., Stump,D.D., Gutierrez,J.A. et al. (2005) The hepatitis C virus alternate reading frame (ARF) and its family of novel products: the alternate reading frame protein/F-protein, the doubleframeshift protein, and others. Semin. Liver Dis., 25, 105–117.
5. Kozak,M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. Gene, 299, 1–34.
6. Yamasaki,K., Weihl,C.C. and Roos,R.P. (1999) Alternative translation initiation of Theiler's murine encephalomyelitis virus. J. Virol., 73, 8519–8526.

7. Pedroso,I., Rivera,G., Lazo,F. et al. (2008) AlterORF: a database of alternate open reading frames. Nucleic Acids Res., 36, D517–D518.

8. Klemke,M., Kehlenbach,R.H. and Huttner,W.B. (2001) Two overlapping reading frames in a single exon encode interacting proteins: a novel way of gene usage. EMBO J., 20, 3849–3860.

9. Freson,K., Jaeken,J., Van Helvoirt,M. et al. (2003) Functional polymorphisms in the paternally expressed XLalphas and its cofactor ALEX decrease their mutual interaction and enhance receptormediated cAMP formation. Hum. Mol. Genet., 12, 1121–1130.

10. Vanderperre,B., Staskevicius,A.B., Tremblay,G. et al. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. FASEB J., 25, 2373–2386.

11. Oh,S., Terabe,M., Pendleton,C.D. et al. (2004) Human CTLs to wild-type and enhanced epitopes of a novel prostate and breast tumor-associated protein, TARP, lyse human breast cancer cells. Cancer Res., 64, 2610–2618.

12. Ronsin,C., Chung-Scott,V., Poullion,I. et al. (1999) A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. J. Immunol., 163, 483–490.

13. Rosenberg,S.A., Tong-On,P., Li,Y. et al. (2002) Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. J. Immunol., 168, 2402–2407.

14. Wang,R.F., Parkhurst,M.R., Kawakami,Y. et al. (1996) Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. J. Exp. Med., 183, 1131–1140.

15. Ho,O. and Green,W.R. (2006) Alternative translational products and cryptic T cell epitopes: expecting the unexpected. J. Immunol., 177, 8283–8289.

16. Chung,W.Y., Wadhawan,S., Szklarczyk,R. et al. (2007) A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput. Biol., 3, e91.

17. Ribrioux,S., Brungger,A., Baumgarten,B. et al. (2008) Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. BMC Genomics, 9, 122.

18. Xu,H., Wang,P., Fu,Y. et al. (2010) Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. Cell Res., 20, 445–457.

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

20. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European molecular biology open software suite. Trends Genet., 16, 276–277.

21. Ivanov,I.P., Firth,A.E., Michel,A.M. et al. (2011) Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res., 39, 4220–4234.

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

ARTICLE 3

Direct detection of alternative open reading frames translation products in human significantly expands the proteome

Auteurs de l'article: Benoît Vanderperre, Jean-François Lucier, Cyntia Bissonnette, Julie Motard, Guillaume Tremblay, Solène Vanderperre, Maxence Wisztorski, Michel Salzet, François-Michel Boisvert, Xavier Roucou

Statut de l'article: publié dans *PLoS ONE,* 2013 Aug 12;8(8):e70698. doi:10.1371/journal.pone.0070698

Avant-propos: Dans cet article, j'ai réalisé l'ensemble du travail expérimental à l'exception de la construction des bases de données de protéines alternatives (JF Lucier), du clonage de la construction AltTP53 (figure 4C ; S Vanderperre), des mutagenèses par QuikChange (J Motard), de 4 des 6 panneaux de western blot en figure 4C, des analyses conservatives ainsi que l'analyse des bases de données d'espèces non-humaines associées (figure 6, figure S2, table S5, databases S2-8 ; C Bissonnette), des expériences associées à la LC-MS/MS sur tissus ovariens, de tube de fallope et d'endomètre (M Wisztorski et M Salzet), et de ceux de colon (FM Boisvert). Au niveau de la rédaction, j'ai participé à 50% de l'écriture de l'introduction, des résultats et de la discussion, ainsi qu'à 70% des méthodes et des légendes de figures.

Résumé : Un ARNm pleinement maturé est généralement associé à un cadre de lecture ouvert (ORF, en anglais) de référence codant pour une seule protéine. Cependant, les ARNm matures contiennent des cadres de lecture ouverts alternatifs (AltORFs) non conventionnels situés dans les régions non traduites (UTRs) ou qui chevauchent l'ORF de référence (RefORF) dans les cadres de lecture non-canoniques +2 et +3. Bien que de récentes approches par *ribosome profiling* et *ribosome footprinting* ont suggéré l'utilisation importante des sites d'initiation de la traduction non conventionnels chez les mammifères, une preuve directe de l'expression de protéines alternatives à large échelle au niveau du protéome fait encore défaut. Pour déterminer la contribution des protéines alternatives au protéome humain, nous avons généré une base de données d'AltORFs prédits révélant un nouveau protéome principalement composé de petites protéines avec une longueur médiane

de 57 acides aminés, comparativement à 344 acides aminés pour le protéome de référence. Nous avons détecté expérimentalement un total de 1 259 protéines alternatives grâce à l'analyse par spectrométrie de masse de lignées cellulaires, de tissus et de fluides humains. Dans le plasma et le sérum, les protéines alternatives représentent jusqu'à 55% du protéome et pourraient être une nouvelle source insoupçonnée de biomarqueurs potentiels. Nous avons observé la co-expression constitutive de RefORFs et d'AltORFs à partir de gènes endogènes et d'ADNc transfectés, y compris depuis le suppresseur de tumeur p53, et avons fourni des preuves que des clones *out-of-frame* représentant des AltORFs sont rejetés à tort comme faux positifs lors d'essais par criblages d'ADNc. L'importance fonctionnelle des protéines alternatives est fortement appuyée par une conservation évolutive significative chez les vertébrés, les invertébrés et la levure. Nos résultats impliquent que l'encodage de multiples protéines dans un seul gène par l'utilisation d'AltORFs est une caractéristique commune chez les eucaryotes, et confirment que la traduction des ORFs non conventionnels génère un protéome encore inexploré.

Direct detection of alternative open reading frames translation products in human significantly expands the proteome

Benoît Vanderperre[1], Jean-François Lucier[2], Cyntia Bissonnette[1], Julie Motard[1], Guillaume Tremblay[1], Solène Vanderperre[1], Maxence Wisztorski[3], Michel Salzet[3], François-Michel Boisvert[4], Xavier Roucou[1]

[1]Département de biochimie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada.

[2]Département de microbiologie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada.

[3]PRISM, Laboratoire de Protéomique, Réponse Inflammatoire, Spectrométrie de Masse, EA 4550, SN3, Université Lille 1, Villeneuve d'Ascq, France.

[4]Département d'anatomie et de biologie cellulaire, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada.

Correspondence to Xavier Roucou: Département de biochimie (Z8-2001), Faculté de Médecine et des Sciences de la Santé, 3201 Jean Mignault, Sherbrooke, Québec J1E 4K8, Canada, Tel. (819) 821-8000x72240; Fax. (819) 820 6831; E-mail: xavier.roucou@usherbrooke.ca

Short title: Alternative proteins expand the human proteome

Keywords: Alternative translation initiation, Mass spectrometry, Out-of-frame translation, Polycistronic mRNA, Proteome database

Abstract

A fully mature mRNA is usually associated to a reference open reading frame encoding a single protein. Yet, mature mRNAs contain unconventional alternative open reading frames (AltORFs) located in untranslated regions (UTRs) or overlapping the reference ORFs (RefORFs) in non-canonical +2 and +3 reading frames. Although recent ribosome profiling and footprinting approaches have suggested the significant use of unconventional translation initiation sites in mammals, direct evidence of large-scale alternative protein expression at the proteome level is still lacking. To determine the contribution of alternative proteins to the human proteome, we generated a database of predicted human AltORFs revealing a new proteome mainly composed of small proteins with a median length of 57 amino acids, compared to 344 amino acids for the reference proteome. We experimentally detected a total of 1,259 alternative proteins by mass spectrometry analyses of human cell lines, tissues and fluids. In plasma and serum, alternative proteins represent up to 55% of the proteome and may be a potential unsuspected new source for biomarkers. We observed constitutive co-expression of RefORFs and AltORFs from endogenous genes and from transfected cDNAs, including tumor suppressor p53, and provide evidence that out-of-frame clones representing AltORFs are mistakenly rejected as false positive in cDNAs screening assays. Functional importance of alternative proteins is strongly supported by significant evolutionary conservation in vertebrates, invertebrates, and yeast. Our results imply that coding of multiple proteins in a single gene by the use of AltORFs is a common feature in eukaryotes, and confirm that translation of unconventional ORFs generates an as yet unexplored proteome.

Introduction

The proteome impacts all aspects of health and disease and deciphering the human proteome represents an important challenge in the post-genomic era. A typical fully processed mRNA includes one RefORF and is associated with a reference protein (Fig. 1A). Reference proteins populate current protein databases used to support research in life sciences. For example, protein databases are central to the success of mass spectrometry-based protein identification for the discovery of expression and interaction proteomics, and of biomarkers [1].

Two cellular mechanisms have evolved to increase proteomic diversity by encoding more than one protein per gene, increasing the diversity of the transcriptome or producing more than one protein from a single transcript. Transcriptome diversity [2] is achieved by utilization of alternative promoters [3], reiterative transcription [4], or post-transcriptional processing, including alternative splicing [5], alternative polyadenylation [6] and RNA editing [7]. On the other hand, N-terminal extension [8], ribosomal frameshifting [9,10] and utilization of multiple coding ORFs in one transcript [11,12] can generate functional or disease-related proteins.

Out-of-frame alternative translation initiation in the same transcript is used to encode proteins of totally different amino acid composition and is mainly observed in viruses and bacteriophages [13]. This mechanism provides such small organisms with increased coding capacity. Until recently, very few examples were documented in human [11,12,14-18], and it was assumed such mechanisms were anecdotal in eukaryotes considering the flexibility and coding capacity provided by their large genome. Two recent studies described the discovery of several protein products resulting from translation of non-canonical, alternative open reading frames (AltORFs) [19,20], present in 5'UTRs, overlapping the RefORFs, or in 3'UTRs. These studies demonstrate that the human proteome is more complex than previously appreciated and suggest that there are many more alternative proteins that remain undiscovered.

Paradoxically, several databases predicting AltORFs in eukaryotes are available [21]. Yet, most of these databases did not actually predict the rare examples of AltORFs translation products documented in humans [11,12,14-18]. Most importantly these databases did not

include AltORFs present within UTRs. To address this issue, we generated a database of predicted AltORFs present in mature human mRNAs.

Here, we provide evidence that AltORFs located within UTRs or overlapping the RefORF in a different reading frame are translated and significantly contribute to the human proteome. We demonstrate that basic transfection of cDNAs containing a RefORF and an AltORF results in coexpression of the reference and alternative proteins. In addition, we provide evidence that alternative proteins are conserved with a high sequence identity in vertebrates and invertebrates, suggesting an important function for this unexplored proteome.

Methods

Ethics protocol (human tissues)- Ovarian, fallopian tube and endometrial formalin fixed, paraffin-embedded tissues were obtained from the CHRU de Lille pathology department (institutional review approval from the Ethical Research Committee CPP Nord Ouest IV 12/10). The ethical committee considered contacting the patients, often many years after surgery, to be unnecessary and waived the need for written informed consent from the participants.

AltORFs database- Few databases predicting alternative ORFs (AltORFs) in eukaryotes are available. Yet, most of these databases did not actually predict the very rare examples of out-of-frame alternative translation products documented in humans. Most importantly these databases did not include AltORFs present within UTRs. Other criteria included in previous predictions –conservation among species, presence of an optimal Kozak context around the initiator AUG codon, location within the reference coding sequence– were not taken into account in this study to reduce biases in prediction and generate a comprehensive database.

The AltORF database for each species was built as previously described [21] with minor modifications to the pipeline of Perl scripts that populate a MySQL database, to retain AltORFs regardless of the Kozak context or the location on the mRNA. GenBank entries

(http://www.ncbi.nlm.nih.gov/genbank/) were used to generate the AltORF databases for human (release 37), chimpanzee (release 102), mouse (release 103), cow (version 1) and frog (version 1), whereas Ensembl entries (http://www.ensembl.org/) were used to generate the databases for fly (version 70.546), nematode (version 70.230) and yeast (version 70.4). The AltORF databases are provided as excel files (Supplemental Databases S1-S8).

Antibodies- Primary antibodies used for western blot were monoclonal anti-HA (Covance), monoclonal anti-GFP (SantaCruz), polyclonal anti-NPTII (Millipore), monoclonal anti-BRCA1 (Bethyl). Secondary antibodies used for western blots were horseradish peroxidase (HRP)-conjugated sheep anti-mouse IgG (NA931V), HRP-conjugated donkey anti-rabbit IgG (NA934V) (GE Healthcare). Primary antibody used for immunofluorescence was monoclonal anti-BRCA1 (Novus Biologicals). Secondary antibodies used for immunofluorescence was Alexa Fluor 568-conjugated goat anti-mouse IgG (Invitrogen).

Clones- These clones were generated to verify translation initiation at predicted AUG codons, to confirm the coexpression of reference and alternative proteins from the same mRNA, and to visualize the expression of alternative proteins using a GFP tag.

Predicted AltORFs with initiation and stop codons in 5'UTRs: SLC35A4 and *ZNF83* (GenBank mRNA entries NM_080670 and NM_001105552, respectively): gBlocks (IDT) gene fragments containing the 5'UTR sequence up to the last coding nucleotide of their respective AltORF were designed and inserted into the *BamHI* restriction site of pEGFP-N1 using the Gibson assembly kit (New England Biolabs) according to the manufacturer's protocol.

Predicted AltORF overlapping the 5'UTR and the reference CDS: IDH3B (NM_174856). This construct was synthesized as gBlocks gene fragments (IDT). The construct was assembled and inserted into the *BamHI* site of pcDNA$^{HA\text{-}EGFP+3}$ using the Gibson assembly kit. pcDNA$^{HA\text{-}EGFP+3}$ was designed to express HA-tagged reference proteins and GFP-tagged alternative proteins with the corresponding AltORF in the +3 reading frame.

Predicted AltORFs entirely included in the reference CDS: NIPA1 (NM_001142275), *BDH2* (NM_020139), *SCARB2* (NM_005506), *LGALS3BP*

(NM_005567), *CDC42* (NM_144681), *VEGFC* (NM_005429), *BDKRB2* (NM_000623), *TP53* (NM_000546), and *SRSF1* (NM_006924). For *NIPA1*, *BDH2*, *SCARB2*, *LGALS3BP*, *CDC42* and *VEGFC*, constructs were synthesized as gBlocks gene fragments (IDT), then assembled and inserted into the *BamHI* site of pcDNA$^{HA\text{-}EGFP+2}$ using the Gibson assembly kit. For *TP53* and *SRSF1*, inserts were PCR amplified from cDNA containing plasmids (Addgene). The resulting PCR fragments were inserted into the *BamHI* site of pcDNA$^{HA\text{-}EGFP+2}$. *BDRKB2* was PCR amplified from a cDNA containing plasmid (clone SC119794, OriGene), and the resulting PCR fragments were inserted into the *BamHI* site of pcDNA$^{HA\text{-}EGFP+3}$.

Mutation of the predicted alternative initiation ATG codons to AAG was performed with complementary oligonucleotides containing the mutation, using the QuikChange kit (Stratagene).

Each construct contained the endogenous Kozak sequence (positions -3 to +4) around both the RefORF and AltORF putative initiation codons. 5' UTR regions were present in the constructs only when the corresponding AltORF was at least partially contained within this region.

The AltMRVI1^{EGFP} fusion was obtained by inserting a gBlock gene fragment containing the AltMRVI1 coding sequence (GenBank mRNA entry NM_001100167, nucleotides 5403 to 5688) into the *BamHI* digested pEGFP-N1 plasmid.

Primers and sequence verified gBlocks gene fragments were purchased from IDT. All constructs were sequenced in both orientations.

Cell culture and transfection- Human epithelial kidney (HEK293), human cervical cancer HeLa and human colon CCL227, CCL228, CCL233, CRL1459 and HCT116 cells were grown in Dulbecco's Modified Eagle's Medium supplemented with 10% Fetal Bovine Serum and penicillin/streptomycin. Human colon cell lines CCL227, CCL228, CCL233, CRL1459, and HCT116 were purchased from ATCC. Transfections were carried out using GeneCellin according to the manufacturer's protocol (BioCellChallenge).

Protein sample preparation, immunoprecipitation and western blot- For validation of alternative proteins expression by HA- and GFP-tagging, HeLa cells were grown in 12-well

plates for 24 h and were then transfected as described above. Cells were rinsed with PBS and lysed in SDS-PAGE sample buffer (0.5% SDS (w/v), 1.25% 2-β-mercaptoethanol (v/v), 4% glycerol (v/v), 0.01% bromophenol blue (w/v), 15 mM Tris-HCl, pH 6.8). After electrophoresis, proteins were transferred to PVDF membranes and detected by western blot using anti-HA (Covance, 1/1000), anti-GFP (Santa Cruz, 1/1000), and anti-NPTII (Millipore, 1/1000) antibodies.

For immunoprecipitation of AltMRVI1^{EGFP}, HEK 293 cells were grown in 100 mm plates for 24 h before transfection. After 24 h, cells were rinsed twice with ice-cold PBS and lysed in 1 mL NETN buffer (50 mM Tris-HCL, pH 8.0, 0.15 M NaCl, 1 mM EDTA, 0.5% NP-40, with protease inhibitors (Roche) and protein phosphatase inhibitors (Thermo Scientific)) for 15 min at 4 °C. Nuclei were broken by successive passing in 18G, 20G, 21G and 25G needles, and the lysate was centrifuged at 15,000 × g for 15 min at 4 °C. Protein concentration was quantified using BCA protein assay reagent (Pierce). Ten µL GFP-Trap agarose beads (ChromoTek) were used for GFP immunoprecipitation of 0.75 mg sample (1 mg/mL in NETN buffer) during 1 h at 4 °C. The beads were then washed three times with 1 mL of NETN buffer, and the bound proteins eluted by incubating for 5 min at 95 °C in SDS-PAGE sample buffer. Proteins were detected by western blot using anti-GFP (Santa Cruz, 1/1000), and anti-BRCA1 (Bethyl, 1/1000) antibodies.

Protein samples preparation and in-gel digestion of proteins from normal colon tissue and colon cell lines prior to LC-MS/MS analysis were performed as previously described [24]. For LC-MS/MS analysis of proteins contained between the 4.6 and 10 kDa markers of an 1D SDS-PAGE, HeLa cells were lysed in a buffer containing 4% SDS/100 mM DTT/100 mM Tris-HCl pH 7.6, and proteins alkylated in 50 mM iodoacetamide. Eight hundred µg of sample was migrated in 8 different lanes (100 µg per lane) of a 4-12% Bis-Tris polyacrylamide NuPAGE Novex gel (Invitrogen), and stained with SimplyBlue Safestain (Invitrogen). In each lane, a single gel slice (between the 4.6 and 10 kDa markers) was cut and processed for in-gel digestion by trypsin [48]. Tryptic peptides were extracted by 1% formic acid followed by 100% acetonitrile before lyophilization in a SpeedVac and resuspension in 1% formic acid prior to LC-MS/MS.

Immunofluorescence and confocal microscopy- Immunofluorescence and confocal microscopy analyses were carried out as previously described [49,50].

LC-MS/MS analyses of HeLa cell lysates, colon cell lines and colon tissues- LC-MS/MS analyses of HeLa cell lysates, colon cell lines (CCL227, CCL228, CCL233, CRL1459 and HCT116 cells) and colon tissues were performed as previously described [24].

LC-MS/MS analyses of cancerous ovarian, normal ovarian, cancerous fallopian tube and normal endometrial tissues- Formalin fixed, paraffin-embedded tissues were obtained from the CHRU de Lille pathology department (institutional review approval CPP Nord Ouest IV 12/10). For the histological imaging prior to LC-MS/MS analysis, 4 µm-thick tissue sections were cut from the formalin-fixed, paraffin-embedded (FFPE) whole-mount ovarian tissue blocks. The sections were placed on ITO-coated slides and heated for 60 min at 58 °C [51]. The tissue was counterstained with hematoxylin, eosin and safran (HES), dehydrated using graded ethanol solutions, and air-dried for histological examination by our staff pathologist. The tissues appeared heterogeneous and contained cancerous, hyperplastic, and normal regions, with stromal tissue in each region [51]. The International Federation of Gynecology and Obstetrics (FIGO) stages were determined.

For tissue de-waxing, sections (10 µm) were generated using a microtome and were applied to conductive glass slides that were coated with ITO (indium tin oxide) on one side. The paraffin was removed by submersion in toluene twice for 5 min, followed by a light rehydration in ethanol baths (100%, 96%; 70% and 30%) before the slides were dried in a desiccator at room temperature [51,52].

Citric acid antigen retrieval was performed by immersing the slides in 10 mM of citric acid for 20 min at 90 °C and then drying them in a desiccator for 10 min. Prior to the enzymatic digestion, the slides were incubated in 10 mM NH_4HCO_3 twice to remove the remaining antigen retrieval solution and to condition the tissue for effective enzyme activity.

Ten milliliters of a solution of 40 mM trypsin in 50 mM ammonium bicarbonate was dropped onto each tissue region of interest. The slides were then incubated for 4 hours at 37 °C in a customized humidity chamber (a 10 cm x 15 cm box filled with water to one quarter of the box height and placed in a 37 °C incubator). After trypsin digestion, 10 µL of a 10

mg/mL HCCA solution in aqueous TFA 0.1%/ACN (3:7) was dropped onto each section [53,54].

Trypsin-digested peptides were manually extracted from specific tissue regions. Using a micropipette, specific regions were subjected to 20 successive washes with 100 µL of a solution of 80% ACN in water. The extract solution was then submitted to freeze-dried with a SpeedVac (Savent). The dried peptides were then re-dissolved in 10 µL of 0.1% TFA. Salts were removed from the solutions, and peptides were concentrated using a solid-phase extraction procedure with the Millipore ZipTip device with a final 10 µL 80% ACN elution solution. The solution was then dried again using the SpeedVac. Dried samples were solubilized in water / 5% acetonitrile / 0.1% formic acid. Samples were separated by online reversed-phase chromatography using a Thermo Scientific Proxeon Easy-nLC system equipped with a Proxeon trap column (100 µm ID x 2 cm, Thermo Scientific) and a C18 packed-tip column (100 µm ID x 15 cm, NikkyoTechnos Co. Ltd). Peptides were separated using a 5%-40% gradient of acetonitrile over 110 minutes at a flow rate of 300 nL/min. The LC eluent was electrosprayed directly from the analytical column and a voltage of 1.7 kV was applied via the liquid junction of the nanospray source. The chromatography system was coupled to a Thermo Scientific Orbitrap Elite mass spectrometer. The mass spectrometer was programmed to acquire in a data-dependent mode. The survey scans were acquired in the Orbitrap mass analyzer operated at 120,000 (FWHM) resolving power. A mass range of 400 to 2000 m/z and a target of 1E6 ions were used for the survey scans. Precursors observed with an intensity over 500 counts were selected "on the fly" for ion trap collision-induced dissociation (CID) fragmentation with an isolation window of 2 amu and a normalized collision energy of 35%. A target of 5000 ions and a maximum injection time of 200 ms were used for CID MS^2 spectra. The method was set to analyze the 20 most intense ions from the survey scan and a dynamic exclusion was enabled for 20 s.

LC-MS/MS analyses of lung tissue, cerebrospinal fluid, urine, plasma and serum- Raw data were obtained from the PeptideAtlas online repository [55] under the following accession numbers: lung (PAe001771), cerebrospinal fluid (PAe001777), urine (PAe000761 and PAe000763), plasma (PAe000846), serum (A: PAe000135 ; B: PAe000347 ; C:

PAe000281 ; D: PAe000331, PAe000332, PAe000333, PAe000334, PAe000335, PAe000337, PAe000338).

Bioinformatics analysis- Quantitation was performed using the program MaxQuant version 1.2.2.5 [56,57]. The derived peak list generated by Quant.exe (the first part of MaxQuant) was searched using Andromeda as the database search engine for peptide identifications against the human GenBank protein entries (release 37) containing 37,390 proteins, to which the 83,886 predicted alternative proteins and 175 commonly observed contaminants and all the reversed sequences had been added. The first search mass tolerance was set to 20 p.p.m. and main search mass tolerance was 6 p.p.m. Enzyme was set to trypsin/p with 2 missed cleavages. Carbamidomethylation of cysteine was searched as a fixed modification, whereas N-acetyl protein and oxidation of methionine were searched as variable modifications. For the serum sample PAe000347, asparagine deamidation was added as a fixed modification, and glutamine deamidation as a variable modification. For analysis of fractionated HeLa cells, colon cell lines, colon tissue, <10 kDa HeLa cells proteins, and paraffin-embedded human tissues, identification was set to a false discovery rate of 5%, determined by the use of a reverse database (similar results were obtained by using a randomly generated database). To achieve reliable identifications, only the proteins associated with a PEP value of less than 0.05, and identified with at least one unique peptide were retained. Indeed, the detection of more than one tryptic fragment for small proteins is unlikely [19], though we detected at least two tryptic fragments for 148 alternative proteins (Supplemental Table S4). For analysis of lung tissue and fluids, identification was set to a false discovery rate of 1%, with at least one unique peptide. Protein isoforms and proteins that cannot be distinguished based on the peptides identified are grouped and displayed on a single line, but with a single GenBank accession number (Supplemental Tables S1-S3).

Accession numbers for detected alternative proteins are provided in Supplemental Table S4 and are publicly available on the nucleotide database of the European Bioinformatics Institute website (http://www.ebi.ac.uk/).

The conservation analyses of alternative and reference proteins were carried out with BLASTP [58] (version 2.2.27+). The AltORFs databases and reference databases from

different species were searched against the corresponding human database with an expectation (E) value cutoff of $\leq 10^{-4}$. Only the best matching hit for each predicted protein was considered. The reference protein databases match the GenBank and Ensembl releases used to generate the AltORFs databases.

Results

To generate our database, we defined AltORFs as ORFs located in a non-canonical reading frame of the RefORF, in the 5' and 3'UTR regions of an mRNA, or partially overlapping with both the RefORF and an UTR region (Fig. 1A). We based our prediction algorithm on characteristics of known AltORFs (AUG as TIS, alternative stop codon different from the RefORF stop codon), and added a size cut-off of 40 codons to keep the database to a reasonable list of polypeptides readily analysable by LC-MS/MS or detectable by SDS-PAGE (Supplemental Database S1; Fig. 1B). Criteria included in previous predictions – conservation among species, presence of an optimal Kozak context around the initiator AUG codon, location within the reference coding sequence– were not taken into account in this study because experimental evidence indicate that these criteria are not necessarily required for an AltORF to be expressed [11,12]. Our database predicts 83,886 unique AltORFs with a minimum size of 40 codons (Fig. 1B). Most predicted AltORFs overlap RefORFs (41.09%) or populate 3'UTRs (46.55%) (Fig. 1C). The majority of mRNAs (87.58 %) have at least one predicted AltORF (Fig. 1D), and there is an average of 3.88 AltORFs for each mRNA. These proportions are in agreement with the number of detectable translation initiation sites (TIS) determined by ribosome profiling [22,23]. Most predicted AltORFs have less than 100 codons, and the median alternative protein length is 57 amino acids, compared to 344 for the conventional proteome (Fig. 1E).

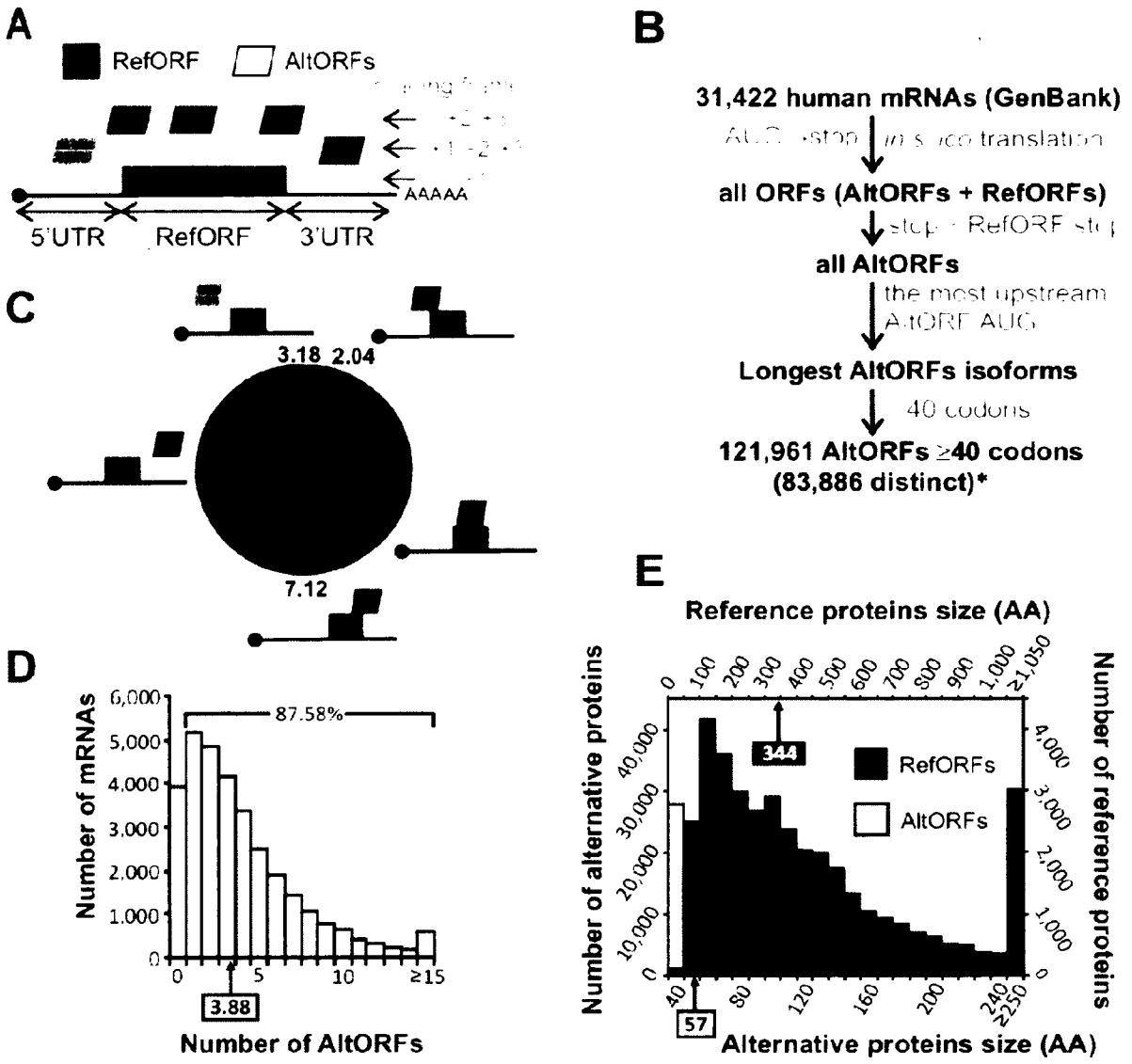


Figure 1. A database to predict AltORFs in human mRNAs. (**A**) A canonical mRNA and its possible AltORFs. The RefORF is the main protein coding ORF annotated in current nucleotide databases. An AltORF is a nucleotide region comprised between an AUG codon and a stop codon distinct from the RefORF and is predicted to encode an alternative protein. AltORFs may be localized in 5' UTRs, overlapping the 5'UTR and the RefORF, overlapping the RefORF, overlapping the RefORF and the 3'UTR, or in the 3'UTR. (**B**) Representation of the database generation process. Distinct AltORFs number indicates the total number of predicted AltORFs that encode alternative proteins with unique amino acid sequences. Since an AltORF may be present in several transcripts, the total number of AltORFs in the transcriptome exceeds the number of distinct AltORFs. (**C**) Distribution in % of AltORFs. (**D,E**) Distribution of the number of predicted AltORFs per mRNA in (**D**) and the size distribution of AltORFs (empty bars, left and bottom scale) compared to RefORFs (grey bars, right and top scale) in (**E**). Boxes and arrows indicate the median.

Using this novel alternative protein database as well as GenBank protein entries, we analyzed a HeLa cells proteomic data set we had previously generated by LC-MS/MS [24]. A total of 68,035 peptides from 5,558 reference proteins and 280 peptides from 129 alternative proteins were identified (Table 1; Supplemental Table S1; Fig. 2A). The mean sequence coverage for reference and alternative proteins was 28.8% and 32.3%, respectively. Overall, alternative proteins represented 2.27% of the total identified proteins. This result clearly shows that the contribution of alternative proteins to the proteome, and thus the number of multiple coding genes, has been overlooked. It is noteworthy that alternative proteins coding sequences are spread across the different regions of mRNAs in agreement with the predicted distribution (compare Fig. 2A and 1C). Co-expression of an alternative protein and its reference protein was observed for 42 genes (Supplemental Table S1). For each of these genes, the average peptide intensity plot of both the reference and alternative proteins revealed large variations in co-expression ratio (Fig. 2B), indicating that a reference protein might not always be the main protein product of a gene. To confirm the expression of alternative proteins in cell lines different from HeLa cells, we performed LC-MS/MS on human colon cell lines and identified 45 alternative proteins (Table 1; Supplemental Table S1; Fig. 2C). AltORFs associated with these 45 proteins were distributed within UTRs and RefORFs with frequencies comparable to those observed in HeLa cells (compare Fig. 2C and 1C). Comparative analysis of alternative proteins detected in both HeLa cells and colon cell lines indicated that 14 are expressed in at least two cell lineages (Fig. 2D). This is more than expected by chance (Fisher's exact test, $p = 3.196e-29$).

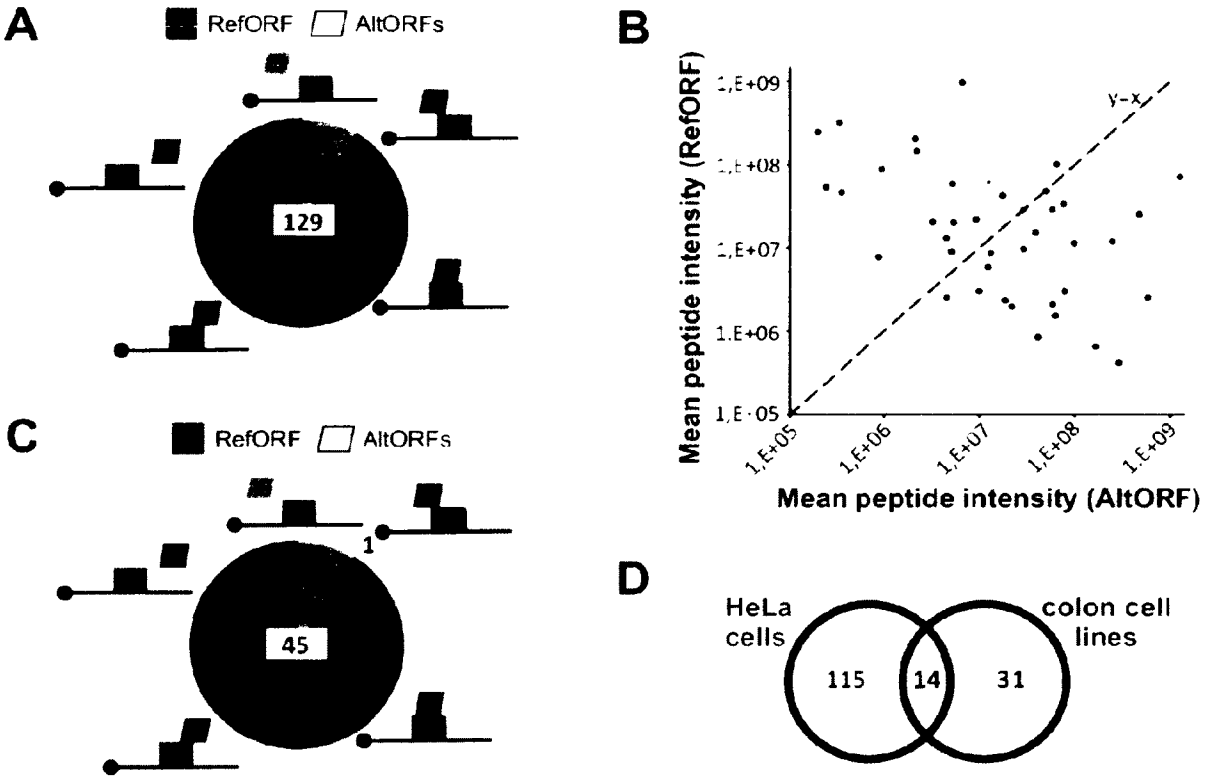


Figure 2. Endogenous expression of alternative proteins in cultured cells. (**A**) Alternative proteins expression was analyzed by LC-MS/MS in HeLa cells, and a schematic distribution of AltORFs in absolute numbers is shown. There were a total of 129 identified alternative proteins (indicated in the center). (**B**) Average peptide intensity plot of both the reference and alternative proteins that were co-expressed from 42 genes. (**C**) Same as (**A**) with colon cell lines. (**D**) Venn diagram showing the number of alternative proteins identified in HeLa cells and colon cell lines. The overlap identifies a common list of 14 alternative proteins.

Table 1. Summary of LC-MS/MS analyses of human samples

Sample	Alternative proteins (peptides)	Reference proteins (peptides)	Alternative proteins (% of total identified proteins)[a]	% sequence coverage (alternative / reference)
Fractionnated HeLa cells	129 (280)	5,558 (68,035)	2.27	28.8 / 32.3
Colon cell lines	45 (63)	3,512 (39,285)	1.27	17.1 / 28.0
Colon tissue	13 (15)	1,985 (16,068)	0.65	19.8 / 23.7
Normal ovary	4 (4)	1,389 (6,224)	0.29	19.5 / 12.7
Serous ovary	6 (7)	1,935 (8,372)	0.31	17.7 / 11.8
Serous fallopian tube	3 (3)	1,379 (5,090)	0.22	8.8 / 11.0
Endometrioid ovary	8 (8)	1,451 (5,762)	0.55	13.0 / 11.7
Normal endometrium	3 (3)	1,240 (4,649)	0.24	9.3 / 10.1
<10kDa HeLa	14 (18)	44 (109)	24.14	25.4 / 35.3
Lung[b]	40 (60)	2,373 (16,987)	1.66	19.3 / 18.4
Cerebrospinal fluid[b]	16 (22)	266 (1,963)	5.67	18.4 / 19.0
Urine[b]	47 (50)	754 (2,898)	5.87	24.9 / 15.5
Plasma[b]	90 (96)	70 (92)	56.25	34.7 / 4.9
Serum A[b]	311 (326)	192 (351)	61.83	23.4 / 3.5
Serum B[b]	269 (293)	339 (847)	44.24	28.4 / 9.7
Serum C[b]	158 (160)	279 (977)	36.16	43.1 / 8.0
Serum D[b]	230 (248)	190 (365)	54.76	N/A
Serum total[b]	928 (1,002)	754 (2,066)	55.17	N/A
TOTAL	1,259 (1,525)	7,341 (85,311)	14.64	N/A

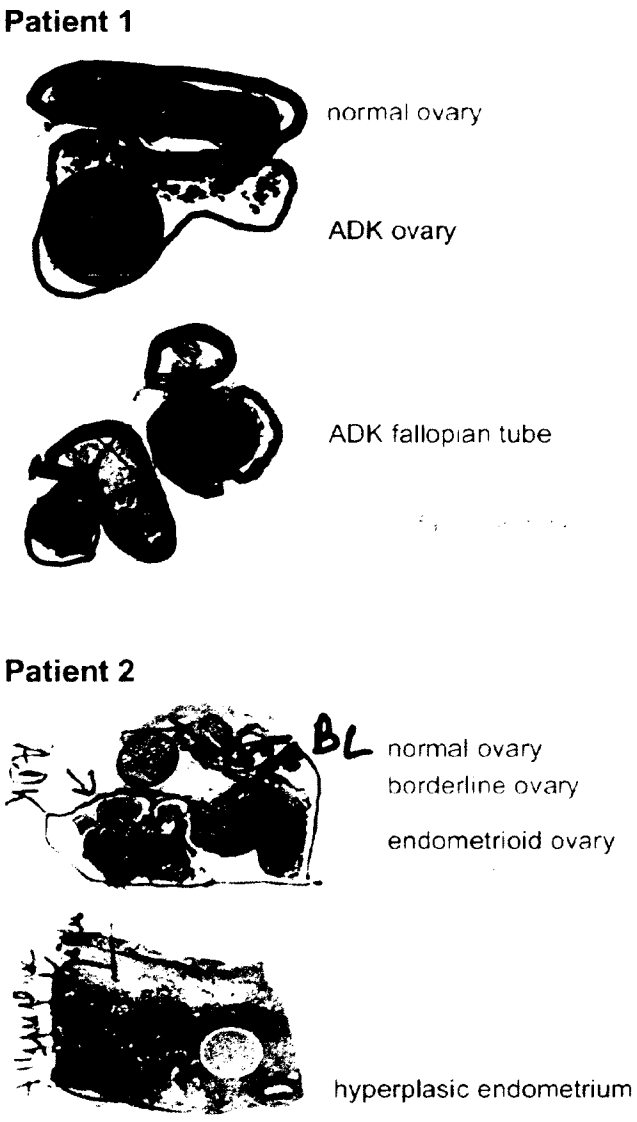
[a] Total identified proteins = alternative + reference proteins.

[b]PeptideAtlas accession numbers associated with lung, cerebrospinal fluid, urine, plasma and serum samples are indicated in the Methods section. The Serum total line indicates the sum of distinct proteins detected in Serum A, B, C and D.

SDS-PAGE in combination with LC-MS/MS is generally limited to the analysis of proteins above 10 kDa, and a low molecular weight is a known limitation in protein identification by MS [25,26]. Since the majority of the predicted alternative proteome is composed of proteins less than 90 amino acids long which have a predicted molecular weight below 10 kDa (Fig. 1E), it is not surprising to have detected much more peptides corresponding to the conventional proteome compared to the alternative proteome. To further assess the abundance of the alternative proteome compared to the conventional proteome, HeLa cells proteins were separated by 1-D SDS-PAGE, and one gel slice between the 4.6 and 10 kDa markers was trypsin digested. The resulting peptides were analyzed by LC-MS/MS. A total of 44 reference and 14 alternative proteins were detected, and alternative proteins

represented 24.14% of the total identified proteins (Table 1; Supplemental Table S1), thus showing that alternative proteins are enriched in the pool of small cellular proteins. The detection of alternative proteins with MW between 4.78 and 9.49 kDa (Supplemental Table S1) is further proof that peptides were not misassigned and that these alternative proteins are actually expressed.

Next, we tested the expression of alternative proteins in a variety of human tissues by LC-MS/MS. First, we analyzed normal colon and lung tissues and detected 13 and 40 alternative proteins respectively (Table 1; Supplemental Table S2). In a second set of experiments, we analyzed ovarian cancer tissue areas and normal areas from the same formalin fixed, paraffin-embedded tissue section of two patients, one presenting endometrioid ovarian cancer and the second presenting a serous ovarian cancer (Supplemental Fig. S1). A total of 19 alternative proteins were identified in the normal endometrium, endometrioid ovary, serous ovary, normal ovary, and serous fallopian tube (Table 1; Supplemental Table S2). We completed these proteomic studies with human fluids, including cerebrospinal fluid, urine, plasma, and serum, identifying 16, 47, 90, and 928 alternative proteins in each fluid respectively (Table 1; Supplemental Table S3). Strikingly, alternative proteins represent approximately 55% of the proteins identified in plasma and serum (Table 1). Overall, we detected a total of 1,259 alternative proteins (Table 1), and 47 were expressed in different cell lines and/or tissues (Supplemental Table S4).



Supplemental Figure S1. Hematoxylin, eosin and saffron-stained sections of normal and cancerous tissues in two patients. *Patient 1*, sections of normal and serous ovarian tissues and normal and serous cancerous fallopian tissue. *Patient 2*, sections of normal, borderline and endometrioid ovarian tissues and normal and hyperplasic endometrial tissues. Annotations of the tissues were performed by a pathologist (Dr. O. Kerdraon, Centre Oscar Lambret, Lille, France).

In accordance with the scanning model of translation initiation, we used the first AUG rule in order to predict the TIS of AltORFs present in our database. Since other non-AUG codons can be used as TIS [22], we tested the reliability of our TIS prediction for the alternative proteins previously detected by two independent methods. First, the detection of N-acetylated peptides, a modification specific to protein N-termini [27], in 889 out of the 1,259 total alternative proteins detected throughout our different LC-MS/MS experiments allowed us to determine that in most cases (886/889), the alternative TIS predicted in our database was correct (Supplemental Table S4). Second, we randomly selected and tested the co-expression of 6 alternative proteins and their corresponding reference proteins from the 129 alternative proteins detected by LC-MS/MS in the fractionated HeLa cells lysate. A strategy based on the transfection of constructs with two tags, an HA tag in frame with the reference protein and a GFP tag in frame with the alternative protein, was used to report the co-expression of both proteins in transfected cells (Fig. 3). The corresponding alternative proteins were all detected by both western blot and GFP fluorescence. Importantly, inactivating mutations (AUG to AAG) of the predicted alternative TIS significantly reduced their expression (Fig. 3).

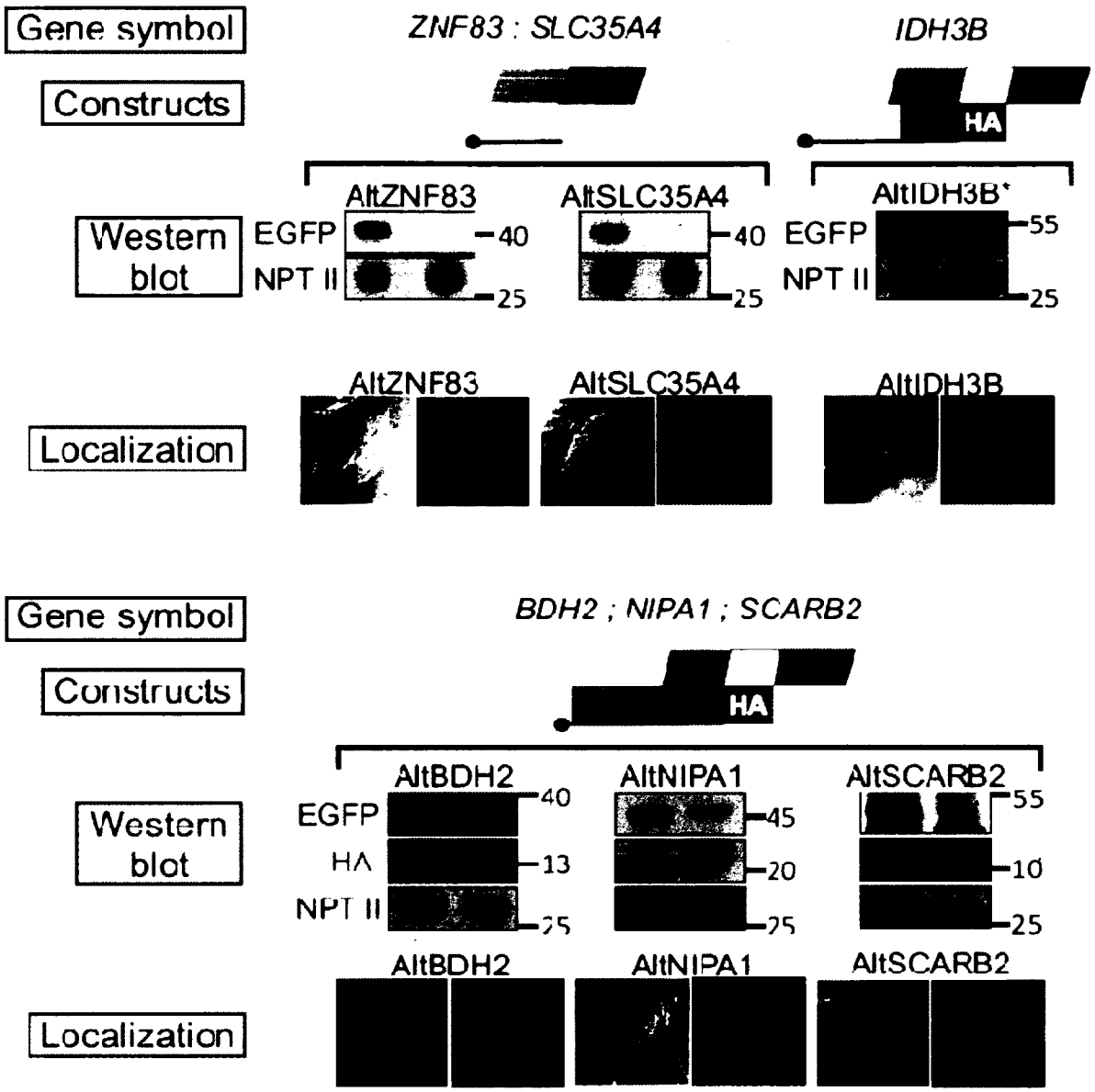


Figure 3. Transfection of tagged constructs validate the expression and translation initiation site prediction of alternative proteins detected by LC-MS/MS. Top diagrams represent the constructs used to detect the co-expression of HA-tagged reference and GFP-tagged alternative proteins by western blot analyses of HeLa cell lysates. GFP is inserted before the alternative stop codon in frame with the AltORF. The black line represents a specific region of the endogenous mRNA. For AltORFs located in 5'UTRs of *ZNF83* and *SLC35A4*, the constructs do not contain the RefORF since the insertion of GFP may prevent the expression of the downstream RefORF. For AltORFs overlapping the 5'UTR and the RefORF (*IDH3B*), and for AltORFs overlapping the RefORF (*BDH2*, *NIPA1*, *SCARB2*), the HA tag was introduced before the GFP tag in frame with the RefORFs. Western blots show the co-expression of reference and corresponding alternative proteins in cell lysates with anti-HA and anti-GFP antibodies, respectively. The left and right lanes

are cell lysates from cells expressing a construct with a normal alternative initiation AUG codon or with an inactivated alternative initiation AAG codon, respectively. NPTII, encoded in the expression plasmid, was used as a transfection control. Molecular weight markers in kDa are indicated on the right. Bottom panels show confocal/DIC images with the various cellular distributions of GFP-tagged alternative proteins. Nuclei were stained with Hoechst. Scale bar: 10 μm. * The reference protein was not detected due to the small size (<3 kDa) of the truncated HA-tagged reference IDH3B protein.

Transfection of cDNAs in cultured cells is a routine technique in most laboratories. The possible unnoticed co-expression of an alternative protein with the reference protein could be a major issue, as 67.36% of human protein coding genes are predicted to have at least one AltORF contained within the RefORF (Fig. 4A). We selected 6 well studied RefORFs from the AltORFs database, including the tumor suppressor p53 (Fig. 4C). The strategy to detect the co-expression of reference and alternative proteins is shown in Fig. 4B. After transfection, we determined that each cDNA led to the constitutive co-expression of the alternative and reference proteins as observed by western blot and fluorescence (Fig. 4C). Diverse subcellular distributions could be observed among the tested constructs (Fig. 4C, see also Fig. 3), suggesting a variety of possible functions associated with alternative proteins.

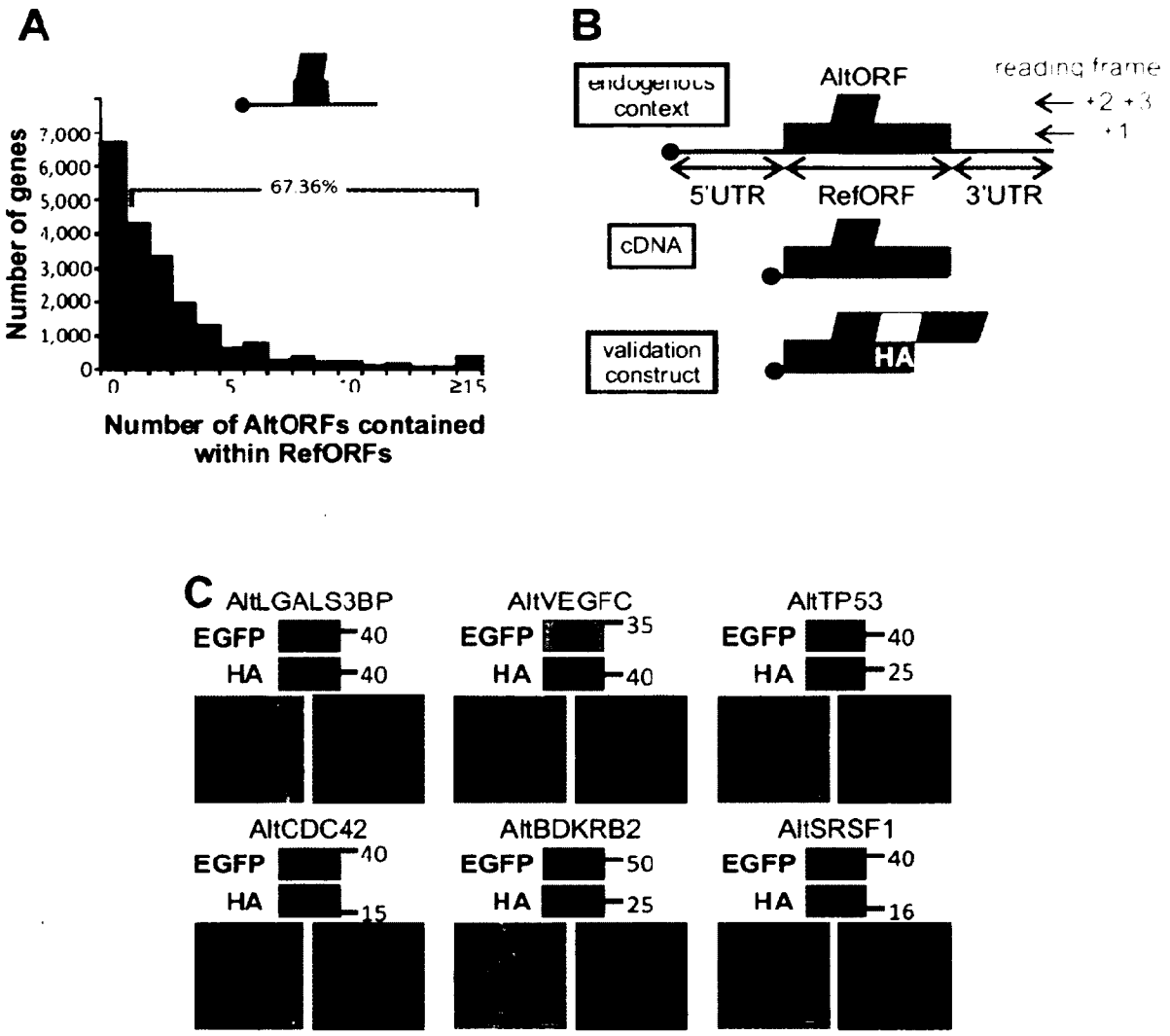


Figure 4. Co-expression of alternative and reference proteins in cDNA transfection experiments is common. (**A**) Distribution of the number of predicted RefORFs-contained AltORFs per gene in the human genome. Top, schematic representation of a mRNA with a RefORF (grey)-containing AltORF (green). By definition, RefORFs are present in the +1 reading frame and AltORFs are present in the non-canonical +2 and +3 reading frames. (**B**) Strategy to detect the co-expression of reference and alternative proteins in cDNA transfection experiments. HA and GFP tags permit the detection of reference and alternative proteins, respectively. Top, graphical representation of a mRNA with a RefORF-contained AltORF. Middle, typical cDNA construct used in transfection experiments. Bottom, representation of constructs used in (**C**). (**C**) Western blot analyses of HA-tagged LGALS3BP (Lectin galactoside-binding soluble 3 binding protein), VEGFC (vascular endothelium growth factor), p53 (cellular tumor antigen p53), CDC42 (cell division cycle 42), BDKRB2 (bradykinin receptor), and SRSF1 (serine/arginine-rich splicing factor 1), and their respective GFP-tagged alternative proteins using anti-HA and anti-GFP antibodies (top panels). Bottom panels show the cellular distribution of alternative proteins by confocal fluorescence microscopy (differential interference contrast and Hoechst, left panels; GFP, right panels). Scale bar: 10 μm.

Many cDNA clones identified in large scale screening assays, including yeast two-hybrid (Y2H) studies do not match any known protein of the conventional proteome because they represent out-of-frame clones [28,29]. In Y2H, these unknown interacting proteins are usually rejected as false positive hits; yet, we reasoned that a proportion of such clones could represent alternative proteins with real affinity for the bait. We found in the literature the partial sequence of five out-of-frame clones from a Y2H experiment performed with the tandem BRCT domain of breast cancer susceptibility protein 1 (BRCA1) [29]. One sequence was 100% identical to an alternative protein from our database whose AltORF is located in the 3'UTR of the mRNA produced from the *MRVI1* gene (Fig. 5A). AltMRVI1^{EGFP} was cloned and transfected into HeLa cells. Similar to BRCA1, AltMRVI1^{EGFP} localized to the nucleus (Fig. 5B). We confirmed the interaction between BRCA1 and AltMRVI1^{EGFP} by co-immunoprecipitation (Fig. 5C). Thus, AltMRVI1 is possibly a novel BRCA1 interacting protein that was already identified by Y2H, but mistakenly rejected as a false positive hit.

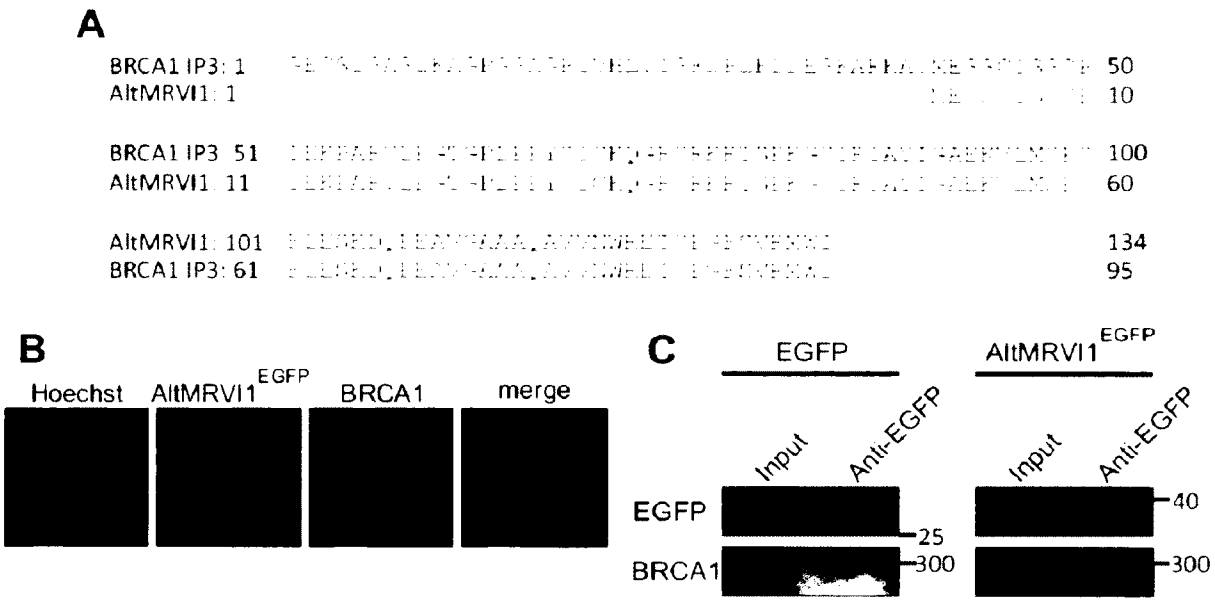
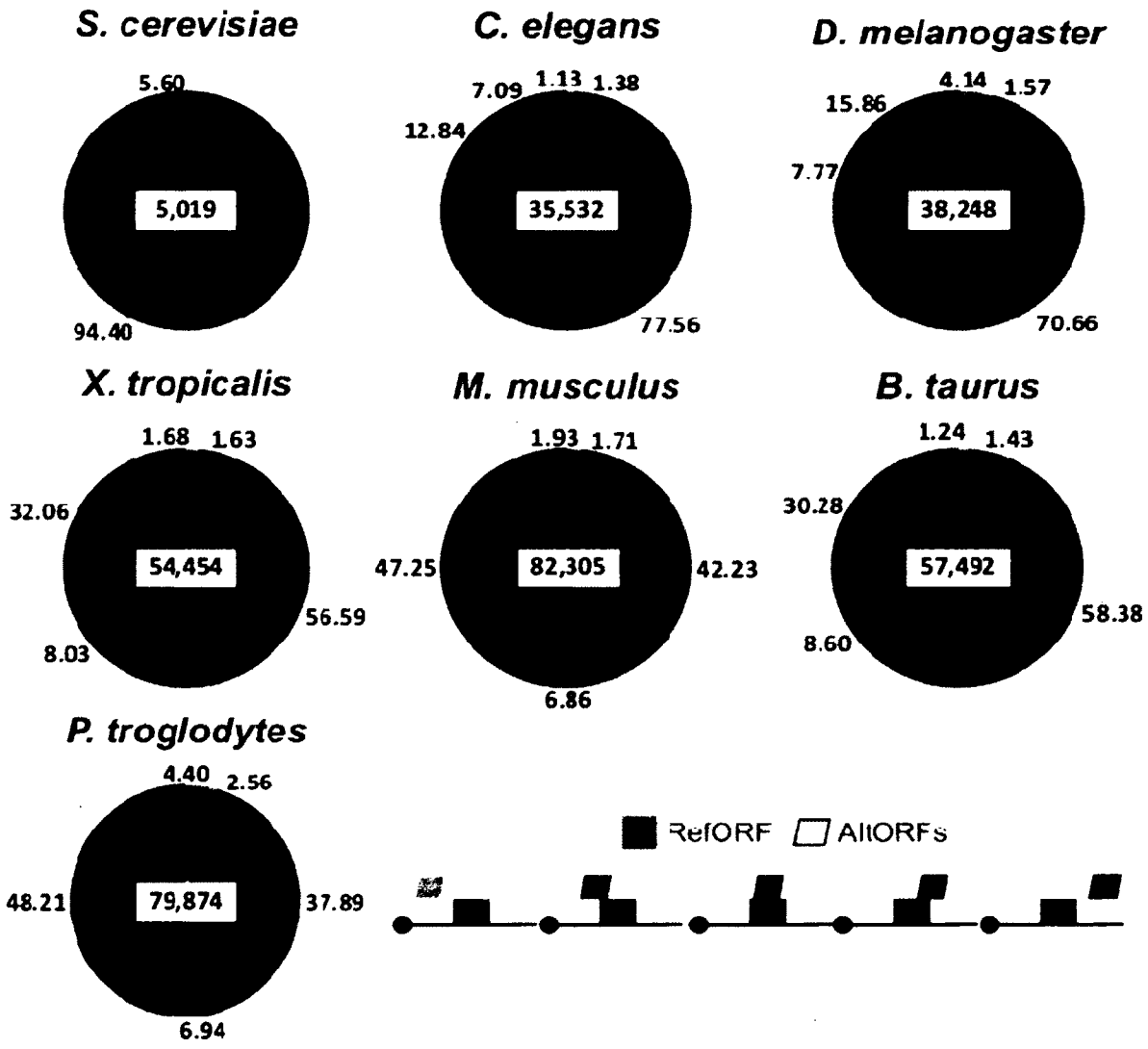


Figure 5. The alternative protein AltMRVI1 interacts with BRCA1. (**A**) Sequence alignment between the third BRCA1 interacting protein 3 (BRCA1 IP3) identified by Liu et al by Y2H [29] and AltMRVI1. (**B**) Localization of BRCA1 and AltMRVI1^{EGFP} in the nucleus. HeLa cells transfected with AltMRVI1^{EGFP} (green channel) were immunostained with anti-BRCA1 (red channel) antibodies. Nuclei were stained with Hoechst (blue channel). Scale bar, 10 μm. (**C**) Co-immunoprecipitation with anti-GFP antibodies was performed using lysates from HeLa cells expressing AltMRVI1^{EGFP}. Molecular weight markers in kDa are indicated on the right.

To assess the evolutionary conservation of predicted alternative proteins, we generated a database of AltORFs present in mature mRNAs from different eukaryote species (Supplemental Databases S2-S8), and predicted 5019 distinct alternative proteins in *Saccharomyces cerevisiae*, 35,532 in *Caenorhabditis elegans*, 38,248 in *Drosophila melanogaster*, 52,454 in *Xenopus tropicalis*, 82,305 in *Mus musculus*, 57,492 in *Bos taurus*, and 79,874 in *Pan troglodytes* (Supplemental Fig. S2 ; Supplemental Table S5). The median AltORF size of these predicted proteins ranges from 50 to 57 codons, showing that as observed for humans, the alternative proteome is composed of small proteins compared to the reference proteome (Supplemental Table S5). Although the homology between reference proteins is greater than that of alternative proteins, we also identified thousands of human alternative proteins conserved with predicted AltORFs in vertebrates, hundreds in

invertebrates, and were even surprised to find that 13 alternative proteins are conserved between human and yeast with a median sequence identity of 47.8% (Fig. 6).



Supplemental Figure S2. AltORFs distribution among eukaryote species. Distribution in % across the different mRNA regions. The number of distinct proteins predicted for each species is displayed in the insert.

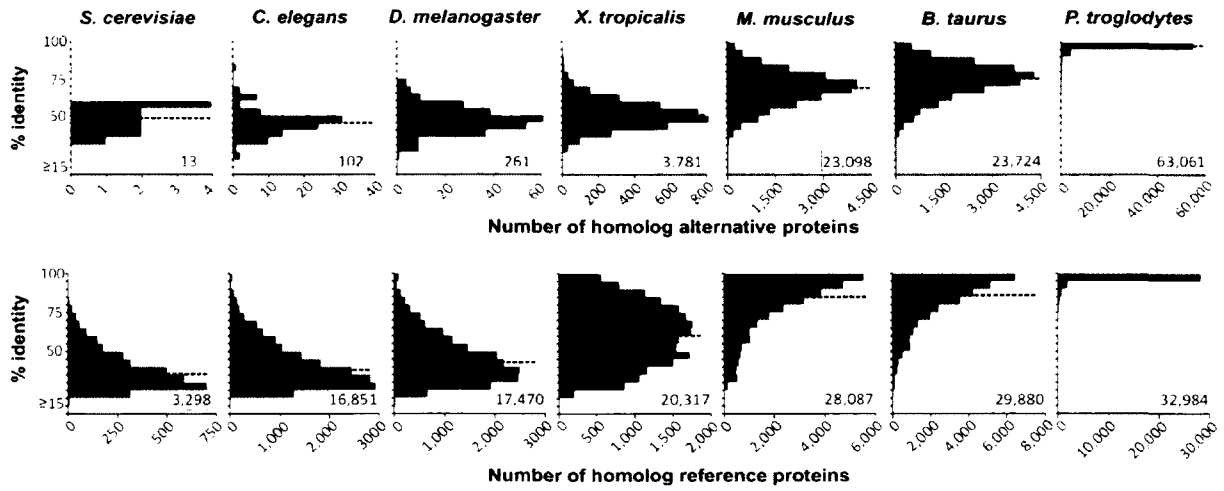


Figure 6. Alternative proteins are conserved from human to yeast. Percent identity of predicted alternative proteins or reference proteins from different eukaryote species analysed with BLASTP (cutoff expectation value of $\leq 10^{-4}$) against human corresponding proteins. The number of homologous proteins displaying each percent identity is shown on the y axis. The dotted black bar indicates the median percent identity for each species. Insets indicate the total number of conserved proteins.

Supplemental Table S5. Summary of AltORFs and alternative proteins characteristics in different eukaryote species

Species	mRNAs[a]	% of mRNAs with AltORF(s)	#AltORFs/ mRNA[b]	Total AltORFs	Distinct alternative proteins[c]	AltORFs size[b] (# codons)	RefORFs size[b] (# codons)
S. cerevisiae	6,692	47.37	0.79	5,288	5,019	50	359
C. elegans	32,693	72.62	1.89	61,922	35,532	56	346
D. melanogaster	24,019	83.06	3.18	76,401	38,248	57	455
X. tropicalis	22,472	78.53	2.43	54,661	52,454	53	386
B. taurus	32,229	62.74	2.39	76,925	57,492	57	360
M. musculus	28,853	85.43	3.80	109,774	82,305	56	397
P. troglodytes	33,850	85.23	3.75	126,875	79,874	55	384

[a]Number of transcripts used to generate the database

[b]Median value

[c]Since an AltORF may be present in several transcripts, the total number of AltORFs in the transcriptome exceeds the number of distinct alternative proteins

Discussion

The potential of eukaryotic genomes for encoding alternative proteins from non-canonical open reading frames is well known and was recently featured in ribosome profiling studies [22,30,31]. Yet, a large scale approach allowing the detection of the corresponding alternative proteins was lacking. In this study, we have generated a database of predicted alternative protein-coding ORFs with a minimum length of 40 codons present in human mRNAs. The data presented here indicate an average of 3.8 AltORFs per human mRNA with a median length of 57 amino acids. Using this database, 1259 human alternative proteins were detected by mass spectrometry in the present study, 3 of which (accession numbers HF548059, HF547970, HF548029) were previously detected [19,32,33]. This result strongly supports the hypothesis that the complexity of the proteome has been underestimated and alternative translation initiation already well characterized in viruses cannot be ignored in humans. Importantly, evolutionary conservation of alternative proteins between vertebrates and invertebrates implies that these proteins have significant biological functions.

It is very likely that similar to proteins translated from canonical ORFs, alternative proteins display a wide variety of biological functions. This is suggested by the great diversity of subcellular localizations that we observed in our fluorescence microscopy experiments (Fig. 3, 4C, 5B), and by the fact that a growing number of important functions are attributed to small proteins and peptides [34-37]. The polycistronic nature of AltORF encoding mRNAs can potentially lead to intriguing functional interplays between reference and alternative proteins, such as direct interaction between the reference and alternative proteins [11,38]. There is also evidence that an upstream ORF not only regulates the expression of a downstream RefORF by interfering in *cis* with canonical AUG recognition by scanning ribosomes, but also reduces the translational efficiency of the RefORF in *trans* [34,39]. AltORFs translation products could also be of particular importance during the thymal selection of T lymphocytes, serving as "cryptic T-cell epitopes". In some cases, this has been shown to lead to the selection of lymphocytes with antiviral or antitumor activities [40].

Our databases of AltORFs will be useful to identify genes containing multiple protein coding ORFs and to unravel their functions. This is particularly important in experimental settings where gene expression studies (cDNA transfection, knock-down, transgenes, and gene therapy) could result in the expression or down regulation of a reference protein and an unnoticed alternative protein, leading to confounding results [12, this study]. Another striking example is the co-expression of therapeutic transgenes and their associated alternative proteins, which elicit a cytotoxic T lymphocyte response [41]. Thus, transgene sequences should be carefully examined for possible AltORFs to decrease potential adverse immune responses during therapeutic gene transfer.

We also propose that proteins recalcitrant to mass spectrometry identification or proteins with no sequence homology with the conventional proteome identified in large-scale cDNAs screens should be revisited with the AltORFs databases. Additionally, large fundamental and clinical proteomic studies using organs and tissues would likely benefit from the AltORFs database to achieve complete catalogs of proteins in different tissues [42,43].

The presence of a large fraction of alternative proteins in plasma and serum is particularly interesting as there is a constant need for biomarkers to identify a variety of disorders at an early stage [44-46]. The reason why so many alternative proteins are secreted is currently unknown. We did not find any enrichment for classical export signal peptides (not shown), and their secretion mechanism remains to be investigated.

As for any databases, the AltORFs database has some limitations. Although AUG remains the main translation initiation site, recent ribosome profiling studies clearly indicate the use of non-AUG start sites [22]. Yet, we did not take into account non-AUG initiation sites as an accurate prediction method for such functional translation start sites is not yet available [47]. Although there is strong evidence that short peptides are also translated [19], we introduced a cut-off of 40 amino acids to discard from our databases polypeptides shorter than 40 residues, which are less readily detected by conventional mass spectrometry approaches, and to keep the database to a reasonable size. Finally, we used the NCBI reference sequence database (RefSeq) as a source for RNA transcripts. This non-redundant and well-annotated database is fairly conservative, and thus is a quality source for identifying candidate AltORFs, but it would be interesting to compare with other databases

(e.g. Gencode) to verify if different mRNA isoforms could also serve as template for the expression of corresponding alternative proteins. Nevertheless, the location of AltORFs is comparable with the distribution obtained in a peptidomic study of small ORFs encoded polypeptides, with the exception of AltORFs located in 5'UTRs [19]. For these AltORFs, the apparent discrepancy probably results from our prediction of AltORFs initiating at AUG sites only [23].

In conclusion, we have provided compelling evidence that alternative proteins significantly contribute to the human proteome by identifying 1,259 new proteins and many more will likely be detected in further MS experiments. A comprehensive knowledge of the proteome is of crucial interest to unravel the cellular mechanisms underlying health and disease. We believe that proteomics approaches supported by ribosome profiling will further benefit the establishment of an exhaustive catalog of proteins to fulfill this goal in the future.

Supplemental Tables

Supplemental Table S1. Alternative and reference proteins list in diverse cell lines (provided as separate Excel file). For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. When co-expression of the reference and alternative proteins is observed for a particular gene, the lane is highlighted in gray. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. For the HeLa cell line whole-protein analysis, a total of 129 alternative proteins identified by 280 peptides, and 5,558 reference proteins identified by 68,035 peptides were detected. For the HeLa cell line analysis of proteins between the 4.6 and 10 kDa marker of a 1-D SDS-PAGE, a total of 14 alternative proteins identified by 18 peptides, and 44 reference proteins identified by 109 peptides were detected. We excluded from the identified proteins those with expected molecular weights above 10 kDa since fragments below 10 kDa likely represent breakdown products. For the colon cell lines, a total of 45 alternative proteins identified by 63 peptides, and 3,512 reference proteins identified by 39,285 peptides were detected.

Supplemental Table S2. Alternative and reference proteins list in human tissues (provided as separate Excel file). For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. For the colon tissue, a total of 13 alternative proteins identified by 17 peptides, and 1,985 reference proteins identified by 16,068 peptides were detected. For the lung tissue, a total of 40 alternative proteins and 2,373 reference proteins were detected. For the normal endometrium, endometrioid ovary, serous ovary, normal ovary, and serous fallopian tube, a total of 19 alternative proteins and 2,748 reference proteins were detected.

Supplemental Table S3. Alternative and reference proteins list in human fluid (provided as separate Excel file). For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. In cerebrospinal fluid, a total of 16 alternative proteins and 266 reference proteins were detected. In urine, a total of 47 alternative proteins and 754 reference proteins were detected. In plasma, a total of 90 alternative proteins and 70 reference proteins were detected. In serum. a total of 928 alternative proteins and 754 reference proteins were detected.

Supplemental Table S4. Combined alternative proteins list (provided as separate Excel file). All alternative proteins identified across all analysed samples are shown. For a given alternative protein, if N-terminal N-acetylated peptides were detected, they are indicated in an additional column and the entire lane is highlighted in gray. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column.

Supplemental Databases

Supplemental Database S1. A database of alternative ORFs in *Homo sapiens* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded are provided. Other information can also be found for both

the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated.

Supplemental Database S2. A database of alternative ORFs in *Pan troglodytes* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for chimpanzee and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S3. A database of alternative ORFs in *Mus musculus* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for mouse and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S4. A database of alternative ORFs in *Bos taurus* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene

name and accession number for cow and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S5. A database of alternative ORFs in *Xenopus tropicalis* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for frog and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S6. A database of alternative ORFs in *Drosophila melanogaster* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for fly and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S7. A database of alternative ORFs in *Caenorhabditis elegans* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human.

The gene name and accession number for nematode and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Supplemental Database S8. A database of alternative ORFs in *Saccharomyces cerevisiae* (provided as separate Excel file). For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for yeast and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Acknowledgements

We thank Dr Michelle Scott (Université de Sherbrooke), Francis Gaudreault (Université de Sherbrooke), Dr Paul Harrison (McGill University) and members of the Roucou lab for helpful comments. The plasmid containing the SRSF1 coding sequence was a kind gift from Dr Benoit Chabot (Université de Sherbrooke).

Author Contributions

Conceived and designed the experiments: XR BV. Performed the experiments: BV JM GT SV MW. Analyzed the data: XR BV CB FMB MS. Contributed reagents/materials/analysis tools: BV JFL CB FMB MS MW SV JM GT. Wrote the paper: XR BV CB MS. Responsible for the project: XR. Contributed to computational analyses: JFL BV CB. Generated the databases: JFL. Performed the conservation analyses: CB.

References

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

2. Licatalosi DD, Darnell RB. (2010) RNA processing and its regulation: Global insights into biological networks. Nat Rev Genet 11: 75-87.

3. Davuluri RV, Suzuki Y, Sugano S, Plass C, Huang TH. (2008) The functional consequences of alternative promoter use in mammalian genomes. Trends Genet 24: 167-177.

4. Anikin M, Molodtsov V, Temiakov D, McAllister WT. (2010) Transcript slippage and recoding. In: Atkins JF GR, editor. RECODING: EXPANSION OF DECODING RULES ENRICHES GENE EXPRESSION. : Springer. pp. 409.

5. Nilsen TW, Graveley BR. (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463: 457-463.

6. Di Giammartino DC, Nishida K, Manley JL. (2011) Mechanisms and consequences of alternative polyadenylation. Mol Cell 43: 853-866.

7. Farajollahi S, Maas S. (2010) Molecular diversity through RNA editing: A balancing act. Trends Genet 26: 221-230.

8. Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV. (2011) Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res 39: 4220-4234.

9. Namy O, Rousset JP, Napthine S, Brierley I. (2004) Reprogrammed genetic decoding in cellular gene expression. Mol Cell 13: 157-168.

10. Wills NM, Atkins JF. (2006) The potential role of ribosomal frameshifting in generating aberrant proteins implicated in neurodegenerative diseases. RNA 12: 1149-1153.

11. Klemke M, Kehlenbach RH, Huttner WB. (2001) Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage. EMBO J 20: 3849-3860.

12. Vanderperre B, Staskevicius AB, Tremblay G, McCoy M, O'Neill MA, et al. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. FASEB J 25: 2373-2386.

13. Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, et al. (1983) Overlapping genes. Annu Rev Genet 17: 499-525.

14. Wang RF, Parkhurst MR, Kawakami Y, Robbins PF, Rosenberg SA. (1996) Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. J Exp Med 183: 1131-1140.

15. Ronsin C, Chung-Scott V, Poullion I, Aknouche N, Gaudin C, et al. (1999) A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. J Immunol 163: 483-490.

16. Rosenberg SA, Tong-On P, Li Y, Riley JP, El-Gamil M, et al. (2002) Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. J Immunol 168: 2402-2407.

17. Poulin F, Brueschke A, Sonenberg N. (2003) Gene fusion and overlapping reading frames in the mammalian genes for 4E-BP3 and MASK. J Biol Chem 278: 52290-52297.

18. Oh S, Terabe M, Pendleton CD, Bhattacharyya A, Bera TK, et al. (2004) Human CTLs to wild-type and enhanced epitopes of a novel prostate and breast tumor-associated protein, TARP, lyse human breast cancer cells. Cancer Res 64: 2610-2618.

19. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, et al. (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 9: 59-64.

20. Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappe J, et al. (2013) Deep proteome coverage based on ribosome profiling aids MS-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. Mol Cell Proteomics .

21. Vanderperre B, Lucier JF, Roucou X. (2012) HAltORF: A database of predicted out-of-frame alternative open reading frames in human. Database (Oxford) 2012: bas025.

22. Ingolia NT, Lareau LF, Weissman JS. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147: 789-802.

23. Lee S, Liu B, Lee S, Huang SX, Shen B, et al. (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A 109: E2424-32.

24. Boisvert FM, Ahmad Y, Gierlinski M, Charriere F, Lamont D, et al. (2012) A quantitative spatial proteomics analysis of proteome turnover in human cells. Mol Cell Proteomics 11: M111.011429.

25. Lubec G, Afjehi-Sadat L. (2007) Limitations and pitfalls in protein identification by mass spectrometry. Chem Rev 107: 3568-3584.

26. Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, et al. (2006) The abundance of short proteins in the mammalian proteome. PLoS Genet 2: e52.

27. Van Damme P, Arnesen T, Gevaert K. (2011) Protein alpha-N-acetylation studied by N-terminomics. FEBS J 278: 3822-3834.

28. Vidal M, Legrain P. (1999) Yeast forward and reverse 'n'-hybrid systems. Nucleic Acids Res 27: 919-929.

29. Liu Y, Woods NT, Kim D, Sweet M, Monteiro AN, et al. (2011) Yeast two-hybrid junk sequences contain selected linear motifs. Nucleic Acids Res 39: e128.

30. Kochetov AV. (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. Bioessays 30: 683-691.

31. Michel AM, Roy Choudhury K, Firth AE, Ingolia NT, Atkins JF, et al. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res .

32. Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, et al. (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. Genome Res 14: 2048-2052.

33. Akimoto C, Sakashita E, Kasashima K, Kuroiwa K, Tominaga K, et al. (2012) Translational repression of the McKusick-kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. Biochim Biophys Acta .

34. Parola AL, Kobilka BK. (1994) The peptide product of a 5' leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. J Biol Chem 269: 4497-4505.

35. Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, et al. (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. Nat Cell Biol 9: 660-665.

36. Gomes I, Grushko JS, Golebiewska U, Hoogendoorn S, Gupta A, et al. (2009) Novel endogenous peptide agonists of cannabinoid receptors. FASEB J 23: 3020-3029.

37. Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, 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: 6336-6341.

38. Freson K, Jaeken J, Van Helvoirt M, de Zegher F, Wittevrongel C, et al. (2003) Functional polymorphisms in the paternally expressed XLalphas and its cofactor ALEX decrease their mutual interaction and enhance receptor-mediated cAMP formation. Hum Mol Genet 12: 1121-1130.

39. Pendleton LC, Goodwin BL, Solomonson LP, Eichler DC. (2005) Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame. J Biol Chem 280: 24252-24260.

40. Ho O, Green WR. (2006) Alternative translational products and cryptic T cell epitopes: Expecting the unexpected. J Immunol 177: 8283-8289.

41. Li C, Goudy K, Hirsch M, Asokan A, Fan Y, et al. (2009) Cellular immune response to cryptic epitopes during therapeutic gene transfer. Proc Natl Acad Sci U S A 106: 10770-10774.

42. Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, et al. (2013) The biology/disease-driven human proteome project (B/D-HPP): Enabling protein research for the life sciences community. J Proteome Res 12: 23-27.

43. Legrain P, Aebersold R, Archakov A, Bairoch A, Bala K, et al. (2011) The human proteome project: Current state and future direction. Mol Cell Proteomics 10: M111.009993.

44. Hanash SM, Baik CS, Kallioniemi O. (2011) Emerging molecular biomarkers--blood-based strategies to detect and monitor cancer. Nat Rev Clin Oncol 8: 142-150.

45. Blennow K, Hampel H, Weiner M, Zetterberg H. (2010) Cerebrospinal fluid and plasma biomarkers in alzheimer disease. Nat Rev Neurol 6: 131-144.

46. Gerszten RE, Wang TJ. (2008) The search for new cardiovascular biomarkers. Nature 451: 949-952.

47. Kochetov AV, Prayaga PD, Volkova OA, Sankararamakrishnan R. (2013) Hidden coding potential of eukaryotic genomes: NonAUG started ORFs. J Biomol Struct Dyn 31: 103-114.

48. Shevchenko A, Wilm M, Vorm O, Mann M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal Chem 68: 850-858.

49. Beaudoin S, Vanderperre B, Grenier C, Tremblay I, Leduc F, et al. (2009) A large ribonucleoprotein particle induced by cytoplasmic PrP shares striking similarities with the chromatoid body, an RNA granule predicted to function in posttranscriptional gene regulation. Biochim Biophys Acta 1793: 335-345.

50. Roucou X, Giannopoulos PN, Zhang Y, Jodoin J, Goodyer CG, et al. (2005) Cellular prion protein inhibits proapoptotic bax conformational change in human neurons and in breast carcinoma MCF-7 cells. Cell Death Differ 12: 783-795.

51. Lemaire R, Menguellet SA, Stauber J, Marchaudon V, Lucot JP, et al. (2007) Specific MALDI imaging and profiling for biomarker hunting and validation: Fragment of the 11S proteasome activator complex, reg alpha fragment, is a new potential ovary cancer biomarker. J Proteome Res 6: 4127-4134.

52. Bonnel D, Longuespee R, Franck J, Roudbaraki M, Gosset P, et al. (2011) Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: Application to prostate cancer. Anal Bioanal Chem 401: 149-165.

53. Franck J, Arafah K, Barnes A, Wisztorski M, Salzet M, et al. (2009) Improving tissue preparation for matrix-assisted laser desorption ionization mass spectrometry imaging. part 1: Using microspotting. Anal Chem 81: 8193-8202.

54. Lemaire R, Desmons A, Tabet JC, Day R, Salzet M, et al. (2007) Direct analysis and MALDI imaging of formalin-fixed, paraffin-embedded tissue sections. J Proteome Res 6: 1295-1305.

55. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, et al. (2006) The PeptideAtlas project. Nucleic Acids Res 34: D655-8.

56. Cox J, Mann M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol 26: 1367-1372.

57. Cox J, Matic I, Hilger M, Nagaraj N, Selbach M, et al. (2009) A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics. Nat Protoc 4: 698-705.

58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

DISCUSSION

Les protéines alternatives : une couche de complexité supplémentaire dans l'établissement du protéome

La combinaison de l'ensemble des mécanismes décrits dans la partie 1.2 de l'introduction offre des possibilités immenses en termes de diversité protéique produite depuis une séquence génomique unique. L'utilisation d'AltORFs pour produire des groupes de protéoformes supplémentaires à partir de gènes uniques vient maintenant s'ajouter à la complexité combinatoire déjà connue. Pour un gène à partir duquel seraient transcrits deux pré ARNm, chacun produisant 3 variants d'ARNm matures par épissage alternatif, un total de 6 protéoformes totales pourraient être produites, en ignorant une combinatoire potentielle de plusieurs modifications post-traductionnelles. En considérant un ORF alternatif différent par pré ARNm, produisant chacun deux protéoformes par épissage alternatif, et chacun porteur d'une seule modification post-traductionnelle optionnelle, il faudrait compter pour ce gène 8 protéoformes supplémentaires, pour un total de 14 protéoformes réparties dans trois groupes de protéoformes (1 RefORF + 2 AltORFs). L'expression validée de plusieurs AltORFs n'est peut-être pas un scénario réaliste pour la majorité des gènes eucaryotes. Mais il est intriguant de noter qu'en compilant l'ensemble des données de spectrométrie de masse analysées dans le manuscrit 3, un total de 49 gènes possèdent deux AltORFs pour lesquels un ou plusieurs peptides ont été détectés, pour 1161 avec un seul AltORF détecté. En admettant que nos critères de sélection ne permettent pas de prédire la totalité des AltORFs (ne serait-ce qu'à cause d'une taille minimale imposée), que la couverture du protéome de chacun des échantillons analysés est incomplète, et que la variété d'échantillons analysés est restreinte, il n'est pas à exclure que des gènes supportent l'expression de plus de 3 groupes de protéoformes chez l'humain. Ceci a d'ailleurs déjà été démontré chez les insectes (Galindo et al, 2007, Kondo et al, 2007, Savard et al, 2006).

Le protéome n'est pas seulement défini comme l'ensemble des protéines encodées par un génome, mais aussi comme l'ensemble de celles exprimées dans un système biologique particulier (tissu, cellule, structure sub-cellulaire) dans un état particulier (pathologie,

réponse à un stimulus, etc) (Bayes & Grant, 2009). Ainsi, de manière additionnelle à la complexité combinatoire décrite ci-dessus, une composante spatio-temporelle inhérente à la complexité du protéome est affectée par la découverte de l'expression des AltORFs. Tout d'abord, pour un ARNm sujet à une expression tissu- ou type cellulaire-spécifique, ou à un transport actif vers un compartiment cellulaire particulier, la présence d'un AltORF en son sein offre la possibilité d'une co-expression spatiale de composants impliqués dans une même voie. C'est le cas, dans le gène *GNAS*, de l'expression d'ALEX et de XLαs, ce qui permet la régulation de la seconde protéine par la première (Freson et al, 2003). Ceci n'est pas sans rappeler les avantages offerts par la structure en opérons chez les procaryotes. Des « opérons post-transcriptionnels » (ARNm dont la traduction est régulée de façon concomitante) étaient déjà connus chez les eucaryotes pour atteindre cet objectif de co-régulation spatio-temporelle de plusieurs ARNm (Keene & Lager, 2005). Par exemple, les ARNm codant pour des protéines liées à l'acquisition (récepteur de la transferrine) et au stockage (ferritine) du fer sont régulés de façon coordonnée au niveau traductionnel par les protéines régulatrices IRP1 et 2 (IRP=iron-regulatory protein) (Ponka et al, 1998). L'expression d'ORFs (RefORF + AltORF) impliqués dans une voie/réponse similaire à partir d'un seul ARNm (tel qu'ALEX et XLαs) offre des possibilités similaires tout en supprimant la dépendance à la coévolution d'un système régulateur complexe. Dépendamment du mécanisme d'initiation de la traduction régissant l'expression d'un AltORF, il est aussi envisageable que son expression soit mutuellement exclusive à celle du RefORF. Ainsi, la traduction du second uORF présent dans l'ARNm du facteur de transcription ATF4 est mutuellement exclusive avec celle du RefORF. L'expression de cet uORF est levée en conditions de stress cellulaire, permettant l'expression d'ATF4 et la transcription de gènes importants pour la réponse au stress (Vattem & Wek, 2004). Ingolia *et al.* ont observé une diminution de l'initiation de la traduction aux codons initiateurs localisés dans les 5'UTRs lors de la différentiation de CSE murines, ce qui pourrait être expliqué par une baisse de l'initiation de la traduction coiffe-dépendante par rapport à l'initiation coiffe-indépendante (Ingolia et al, 2011). Un découplage dans l'expression des ORFs portés par un même ARNm (expression mutuellement exclusive) semble être observable dans mes résultats d'analyses protéomiques, bien que cela doive être confirmé par une autre méthode de détection (western blot par exemple) étant donné qu'en MS,

absence de détection ne signifie pas forcément absence d'expression. De nombreux AltORFs ont été détectés dans des échantillons sans que les RefORFs correspondant ne l'aient été. C'est le cas pour 87 des 129 AltORFs dont la protéine alternative correspondante a été détectée dans l'expérience de cellules HeLa fractionnées. Pour les 42 restants, des ratios variables d'expression RefORF/AltORF ont été observés (manuscrit 3, Figure 2).

Intégration des AltORFs dans l'étude de la structure et de l'expression des gènes

Un des grands objectifs de l'ère post-génomique est d'annoter le(s) génome(s) de façon à ce qu'un ORF soit défini à chaque locus transcrit en ARNm (Brent, 2005). Bien que cela soit une étape nécessaire vers le décryptage des génomes, cet objectif devra être revu avec l'ambition de déterminer l'ensemble des séquences ultimement traduites en protéines que portent les génomes. C'est le but que tente d'atteindre la protéogénomique : l'annotation, guidée par les techniques de protéomique à haut débit (MS), des régions génomiques traduites. De multiples approches existent pour cette identification de nouveaux produits protéiques, et seront discutées dans le paragraphe suivant. La protéogénomique a déjà mis et continue à mettre à jour (manuscrit 3) des structures de gènes toujours plus complexes. La démonstration que les AltORFs sont exprimés de manière plus fréquente que supposée précédemment amène des changements non négligeables au concept de gène, et par là même à la manière dont nous les étudions. Les ressources bioinformatiques, telles que bases de données de gènes, de transcrits, de polymorphismes, ou encore outils d'annotation fonctionnelle devront prendre en compte les multiples groupes de protéoformes encodées par des substrats uniques, et rendre cette complexité facilement accessible et compréhensible à la communauté des sciences de la vie. Au fur et à mesure de la découverte de nouveaux ORFs, ceux-ci devront être intégrés à l'étude de l'expression des gènes, aussi bien qu'à celle de leurs rôles fonctionnels. Ceci sera valable pour les approches ciblées (par gène candidat par exemple), et le développement d'outils tel que la base de données HAltORF facilement consultable en ligne (manuscrit 2) sera là encore nécessaire. Une étude récemment publiée par notre laboratoire a d'ailleurs utilisé cette ressource comme point de départ pour la découverte d'une nouvelle protéine alternative dans le gène

ATXN1 (Bergeron et al, 2013). Les études à large échelle seront évidemment concernées par ces changements apportés au concept de gène. Dans le manuscrit 3, j'ai souligné l'importance que revêt la ré-analyse de données brutes (MS, Y2H et screening d'ADNcs de manière générale) à la lumière du concept de l'expression d'AltORFs depuis les gènes eucaryotes. La remise en question des *a priori* sur la structure des gènes eucaryotes sera certainement cruciale afin d'obtenir une vision compréhensive des résultats expérimentaux et donc des systèmes biologiques à l'étude à travers l'ensemble des domaines des sciences de la vie, du fondamental au clinique.

Étude de la fonction et des intérêts biologiques des protéines alternatives

L'intérêt fonctionnel de certaines protéines alternatives a déjà été établi (voir introduction). Après avoir démontré l'utilisation fréquente d'AltORFs, il est désormais important de découvrir les fonctions associées aux protéines alternatives. Ces fonctions seront fort probablement extrêmement variées (comme le sont celles des protéines de référence). Ceci sera discuté dans ce paragraphe, ainsi que les méthodes utilisées pour l'étude de la fonction des protéines alternatives.

Caractéristiques des séquences de protéines alternatives : Il est connu que la composition nucléotidique (le pourcentage de nucléotides G et C) varie entre les différentes régions des ARNm (Zhang et al, 2004). De plus, les ORFs traduits dans des cadres de lecture alternatifs possèdent un biais dans l'utilisation des codons, au moins chez les virus (Pavesi et al, 2013, Sabath et al, 2012). Puisqu'ils sont traduits depuis des régions UTRs ou dans des cadres de lecture alternatifs, les AltORFs pourraient avoir une composition générale en acides aminés différente des RefORFs. L'utilisation de scripts Perl m'a permis de déterminer, chez l'humain, les compositions en AA des protéines de référence RefSeq, ainsi que de l'ensemble des protéines alternatives prédites ou détectées par MS dans le manuscrit 3 (données non présentées). Les résultats n'ont pas mis à jour de différence pour aucune des deux comparaisons, mais des analyses en fonction de la distribution des AltORFs dans les régions des ARNm, ou encore des analyses de biais d'utilisation des codons, amèneraient peut-être des résultats différents. Pour ce qui a trait à la structure des protéines alternatives, certaines études permettent de présumer de celles des protéines traduites à partir de

séquences double-codantes. En effet, l'étude des séquences traduites dans deux cadres de lecture différents par l'usage d'épissage alternatif indique que les protéines associées ont tendance à être plus désordonnées que le reste des séquences codantes (Kovacs et al, 2010). Les protéines désordonnées peuvent avoir des cinétiques de liaison augmentées, des surfaces de liaison agrandies, et une capacité à s'adapter à la structure de multiples interacteurs (Tompa et al, 2005), et il faudra vérifier si cela est le cas pour les protéines alternatives issues d'AltORFs chevauchant le RefORF. L'étude d'ORFs chevauchant chez des virus eucaryotes a souligné cette caractéristique de désordre intrinsèque, et a également indiqué que l'apparition de ces ORFs est souvent genre ou espèce spécifique (Rancurel et al, 2009). Leur fonctionnalité dans la pathogénicité ou la transmission virale est pourtant établie, ce qui souligne qu'un ORF chevauchant (et par extension une protéine alternative) n'a pas besoin d'être conservée évolutivement pour avoir une importance biologique. D'ailleurs, pour une région non structurée, plus de mutations peuvent apparaître sans modifier la structure que dans une région structurée (Brown et al, 2002, Daughdrill et al, 2007), ce qui autoriserait les AltORFs chevauchant à subir une contrainte évolutive amoindrie. C'est d'ailleurs ce qui est observé pour la protéine alternative ALEX du gène *GNAS*, qui est peu conservée en termes de séquences d'AA, mais qui en contrepartie co-évolue rapidement avec la protéine de référence correspondante pour maintenir son interaction avec la protéine XLαs (Nekrutenko et al, 2005).

Localisation des protéines alternatives et implications fonctionnelles: La variété des fonctions associées aux protéines alternatives est suggérée par la grande diversité dans la localisation subcellulaire observée pour les différents candidats étudiés dans le manuscrit 3 et pour les protéines de petite taille en général (Frith et al, 2006, Slavoff et al, 2013). Il est intriguant de remarquer que la localisation d'une protéine alternative est parfois similaire à celle de la protéine de référence associée. Dans le manuscrit 3 (Figure 3), la protéine alternative du gène *ZNF83*, dont le RefORF encode un facteur de transcription nucléaire probable, est aussi localisée au noyau. Celle du gène *NIPA1* semble située dans l'appareil de sécrétion, la protéine de référence étant localisée entre autres dans les endosomes précoces (Tsang et al, 2009). Un essai de colocalisation entre AltNIPA1 et le marqueur d'endosomes précoce EEA1 a permis de confirmer cela (Figure 23 A). Dans la figure 4 du manuscrit 3, la localisation commune des protéines alternative et de référence dans

l'appareil de sécrétion semble valable aussi pour le gène *VEGFC* (encodant un facteur de croissance sécrété), et a été validée pour le gène *BDKRB2* (encodant un récepteur couplé aux protéines G) (Figure 23 B). Ce phénomène était déjà connu pour le gène *GNAS* (Klemke et al, 2001), et dans ce cas, un lien fonctionnel existe en les deux protéines *via* une interaction directe qui permet de réguler la fonction de la protéine de référence. Il serait intéressant de valider la colocalisation en microscopie à fluorescence d'autres protéines alternatives candidates avec leur protéine de référence respective, et de vérifier ensuite si protéines alternative et de référence sont trouvées dans un même complexe par co-immunoprécipitation. Cela a d'ailleurs été réalisé pour la protéine alternative du gène *ATXN1*, où les deux produits protéiques interagissent au noyau de manière directe (Bergeron et al, 2013). D'autres travaux dans notre laboratoire indiquent qu'une interaction serait aussi observable dans le cas du gène *BDKRB2*, mais au niveau du réticulum endoplasmique (données non présentées). Ces localisations semblables, couplées ou non à des interactions, ne sont cependant pas toujours observées. C'est le cas pour PrP (feuillet externe de la membrane plasmique) et AltPrP (membrane externe mitochondriale) (manuscrit 1, Figure 2), et pour d'autres gènes candidats (manuscrit 3, Figures 3 et 4, gènes *BDH2*, *SCARB2*, *TP53*, *SRSF1*). Ainsi, l'étude au cas par cas de la fonction des protéines alternatives permettra certainement d'identifier des gènes où AltORFs et RefORFs sont impliqués dans des voies biologiques différentes, ou dans la régulation d'un même processus. L'analyse de l'expression spatio-temporelle (couplée ou découplée) des multiples ORFs traduits dans un gène donné pourra certainement aider à cerner ces couplages ou découplages fonctionnels.

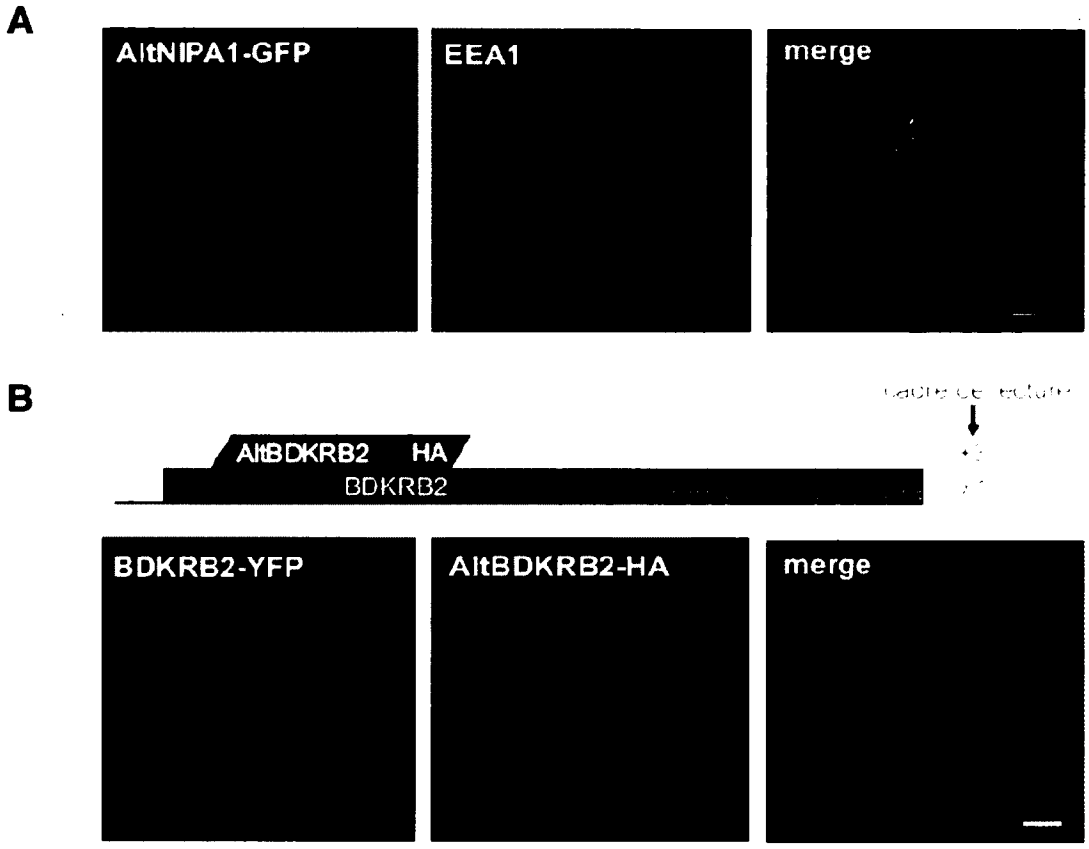


Figure 23. Localisation subcellulaires similaires entre certaines protéines alternatives et leur protéine de référence respective. (**A**) Des cellules HeLa transfectées avec une protéine fusion AltNIPA1-GFP (vert) ont été marquées avec un anticorps anti-EEA1 par immunofluorescence, soulignant les endosomes précoces en rouge. (**B**) Des cellules HeLa ont été transfectées avec une construction contenant l'ADNc correspondant à l'ARNm du gène BDKRB2, contenant l'AltORF encodant AltBDKRB2, et le RefORF encodant BDKRB2. Les protéines ont été étiquetées avec un épitope HA (AltBDKRB2-HA) ou YFP (BDKRB2-YFP, vert) respectivement, en C-terminal. Une immunofluorescence a ensuite été réalisée avec un anticorps anti-HA (rouge). (**A**,**B**) Échelle, 10 µM

Limites des approches par homologie à des protéines connues : Afin d'étudier les fonctions des protéines alternatives, plusieurs méthodes peuvent être utilisées. Tout d'abord, une approche par homologie (alignement de séquences : Blastp ou Pfam par exemple) permet d'identifier des protéines présentant des similarités de séquence ou des domaines fonctionnels conservés, respectivement. Basé sur la fonction des séquences homologues identifiées, des hypothèses peuvent alors être proposées sur la fonction des protéines alternatives. Cette approche, très utilisée et efficace pour les protéines de référence (les RefORFs sont bien souvent annotés dans les génomes grâce à des homologies de séquence), n'a pas donné de résultats significatifs pour la majorité des protéines alternatives. Ceci peut s'expliquer par la taille réduite de nombreuses protéines alternatives, et donc la probabilité diminuée d'y trouver un domaine protéique complet. Pour les régions d'AltORFs chevauchant le RefORF, un autre problème se pose. Les domaines protéiques ont tendance à être portés par des exons uniques (Holland & Blake, 1987), qui n'encodent ces domaines que dans le cadre de lecture canonique du RefORF (en général). Le fait que les AltORFs chevauchant le RefORF soient traduits dans un cadre de lecture non canonique (+2 ou +3) fait en sorte que la séquence en AA et la structure conservée des domaines protéiques ne sont pas incluses dans les protéines alternatives correspondantes. Toute homologie de séquence avec des domaines connus est alors improbable. Ainsi, sur les 1 259 protéines alternatives détectées par LC-MS/MS dans le manuscrit 3, seules 103 d'entre elles trouvent une homologie parmi les protéines présentes dans la base de données de séquences protéiques non redondante de NCBI (Blastp, base de données « nr », données non présentées). Les protéines alternatives des gènes *PRNP*, *ATXN1* et *GNAS* ne trouvent pas non plus d'homologie à des protéines ou domaines connus, ce qui ne les empêche pas d'adopter des localisations subcellulaires précises, et n'empêche pas la protéine alternative ALEX d'être fonctionnelle. Les approches par homologies aux protéines connues peuvent donc présenter des limites importantes pour de nombreuses protéines alternatives. D'autres méthodes doivent donc être développées pour sélectionner des candidats potentiellement fonctionnels parmi les AltORFs prédits et/ou dont l'expression a été validée.

Approches pour la sélection de protéines alternatives candidates et leur étude fonctionnelle : Certains critères permettent de concentrer les efforts sur des candidats d'intérêt supérieur. L'estimation de la conservation évolutive des protéines alternatives est

une possibilité, la méthode la plus simple étant l'alignement de séquences par paires, comme effectué pour les protéines alternatives prédites dans le manuscrit 3 (Figure 6, et Databases S2-S8). Cela a fourni une liste de 13 candidats communs chez l'humain et la levure. Afin d'étudier leur fonction, nous avons établi une collaboration avec le laboratoire du Dr Landry, à l'Université Laval. Tout d'abord, des souches contenant les AltORFs candidats étiquettés seront générées afin de tester l'expression des protéines alternatives associées. Ensuite, des souches mutantes (mais sans étiquette) où l'expression des AltORFs sera inactivée (codon initiateur muté, ou codon stop précoce introduit) seront utilisées afin de comparer leur croissance dans diverses conditions de culture (aérobie/anaérobie, stress, ...) par rapport aux souches sauvages, permettant de montrer l'implication des protéines alternatives candidates dans la croissance dans ces conditions variables. Puis les protéines alternatives d'intérêt ainsi identifiées verront leur interactome défini par des approches à large échelle (tel que le protein-fragment complementation assay (Tarassov et al, 2008)), offrant une piste vers les mécanismes moléculaires régissant leur importance biologique.

La conservation évolutive des protéines peut également être estimée par le rapport de mutations silencieuses sur non-silencieuses dans une séquence nucléotidique codante donnée. Cela pourrait être appliqué aux AltORFs, et pourrait même permettre de discriminer, dans des régions traduites dans plusieurs cadres de lecture, lequel porte la pression sélective. Si la région du RefORF encodée dans la séquence double-codante est moins bien conservée que celle de l'AltORF qui la chevauche, cela pourrait être un indice que la conservation de l'AltORF est fonctionnellement plus importante que celle du RefORF dans cette région. Cependant, nous avons vu que la conservation évolutive d'une protéine alternative n'est pas absolument nécessaire à son expression et à sa fonctionnalité (exemple d'ALEX, ayant peu d'homologie entre humain et souris) (Nekrutenko et al, 2005) et ne peut donc être utilisé comme seul critère pour la sélection de protéines alternatives candidates à des études fonctionnelles.

Un autre critère important est donc l'existence de preuves de l'expression de la protéine alternative en question *in vivo*, condition primordiale pour suggérer un intérêt biologique à celle-ci. Pour les approches par sélection de candidats prédits (AltPrP, AltATXN1), qui ont cependant porté leurs fruits, nous avons généré des anticorps pour produire cette preuve, ce qui s'est avéré fastidieux, coûteux, long, et sans aucune garantie sur l'obtention d'un

anticorps fonctionnant dans les applications voulues. Le fait de se concentrer sur les candidats ayant été détectés dans des expériences de MS, si possible avec une haute confiance (couverture de séquence, qualité des spectres, détection dans plusieurs échantillons), permet ainsi d'avoir une preuve dès le début d'un projet d'une expression *in vivo*, et de s'assurer d'un rapport coût/bénéfices minimal de la synthèse d'anticorps, qui restent des outils indispensables pour l'étude des protéines endogènes. Une liste de plus de 1 250 candidats est ainsi à disposition (manuscrit 3, Table S4), qui pourra être complétée avec l'analyse d'autres données de MS.

Le couplage d'une preuve de l'expression d'une protéine alternative par MS à la conservation évolutive pourrait offrir une confiance meilleure encore dans la sélection des candidats. Aucune des 13 protéines alternatives conservées chez la levure n'a de preuve d'expression à ce jour, mais 2 candidats conservés à partir de *C. elegans* répondent à ces deux critères. En particulier, une protéine alternative encodée dans le 5'UTR de l'ARNm du gène *SMCR7L* a été identifiée chez l'humain (cellules HeLa) par la détection de 4 peptides, avec une couverture de séquence de 38%. Appelée AltSMCR7L, elle a une longueur de 70 AA. En accord avec son poids moléculaire calculé, elle est également détectable dans un échantillon de protéines inférieures à 10 kDa d'un lysat de cellules HeLa (manuscrit 3, Table S1 et S4). L'AltORF correspondant est correctement identifié comme ayant un potentiel codant par l'outil de recherche de petits ORFs sORFfinder (Hanada et al, 2010)(données non présentées). Une recherche par Pfam indique la présence d'un domaine conservé LYR, possiblement impliqué dans la biogenèse des centres Fe-S aux mitochondries (Adam et al, 2006, Shi et al, 2009) et à l'assemblage d'autres complexes impliqués dans les voies métaboliques mitochondriales (Ghezzi et al, 2009) (Figure 24). AltSMCR7L est prédite pour être mitochondriale (WolfPsort, mitoprotII, données non présentées), et cela a été vérifié par microscopie à fluorescence (Figure 25), par fractionnement subcellulaire, ainsi que par détection par spectrométrie de masse dans le protéome de la matrice mitochondriale (ré-analyse de données brutes de MS de Rhee *et al.* (Rhee et al, 2013) par la méthode utilisée dans le manuscrit 3) (données non présentées). La protéine de référence du gène *SMCR7L*, chez les vertébrés, s'appelle MIEF1 (Figure 24). Elle est localisée à la membrane externe mitochondriale, où elle régule positivement la fusion des mitochondries (Palmer et al, 2011, Zhao et al, 2011). AltSMCR7L ayant été

détectée lors de l'analyse des données de MS de Rhee *et al.* (Rhee et al, 2013), elle est en contact avec la matrice mitochondriale, et est donc localisée soit à la matrice soit à la membrane interne, ce qui sera confirmé par fractionnement submitochondrial. Toujours est-il qu'un couplage fonctionnel entre la protéine alternative et de référence de *SMCR7L* n'est pas *a priori* évidente malgré une localisation mitochondriale commune (Figure 25). Un anticorps est en cours de synthèse afin de valider l'expression d'AltSMCR7L *in vivo*, et des études fonctionnelles préliminaires indiquent que sa surexpression semble perturber l'homéostasie du fer dans des cellules HeLa (données non présentées). Il sera intéressant de comprendre les mécanismes moléculaires associés à cette observation, et une liste de protéines interagissant avec AltSMCR7L sera déterminée par co-immunoprécipitation et analyse par LC-MS/MS. Cela pourrait également être obtenu par utilisation d'un crible double hybride chez la levure.

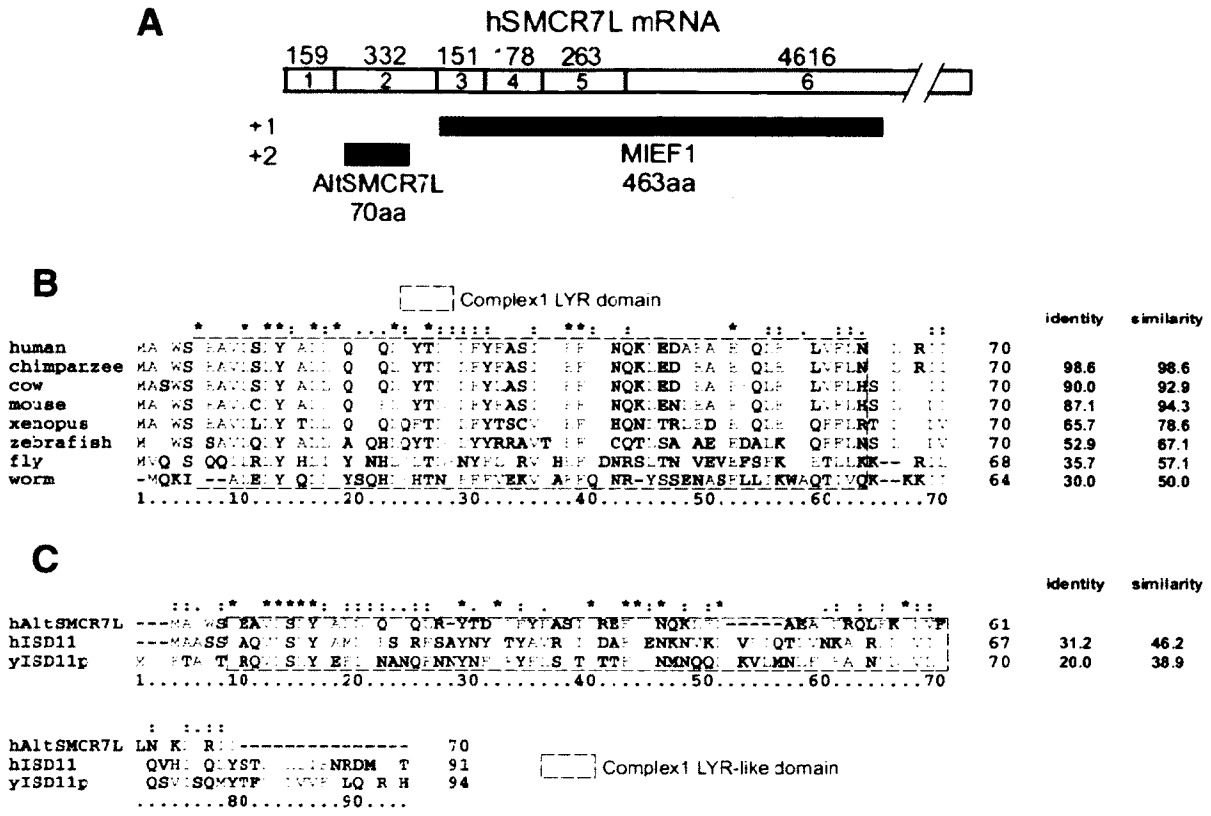


Figure 24. AltSMCR7L, une protéine encodée dans le 5'UTR du gène *SMCR7L* chez les vertébrés, est conservée de l'humain au ver. (**A**) Chez l'humain, le gène *SMCR7L* est composé de 6 exons assemblés en un seul ARNm bicistronique. Le 5'UTR contient un AltORF codant pour AltSMCR7L (cadre de lecture +2), une nouvelle protéine non caractérisée de 70 acides aminés. Le RefORF annoté (cadre de lecture +1) encode MIEF1, qui favorise la fusion mitochondriale (Zhao et al, 2011). (**B**,**C**) Des alignements de séquences multiples ont été réalisés avec ClustalX 2.1, et les pourcentages d'identité et de similarité ont été calculés à l'aide de l'algorithme Needle (suite Jemboss). (**B**) AltSMCR7L humaine présente une conservation significative avec ses orthologues dans d'autres espèces animales. Noter la présence conservée d'un domaine LYR du complexe 1, impliqué dans la biogenèse des centres Fe/S mitochondriaux et dans l'assemblage d'autres complexes impliqués dans des voies métaboliques mitochondriales. (**C**) AltSMCR7L présente de l'homologie avec ISD11, une protéine conservée évolutivement de l'humain à la levure et contenant un domaine LYR-*like* du complexe 1.

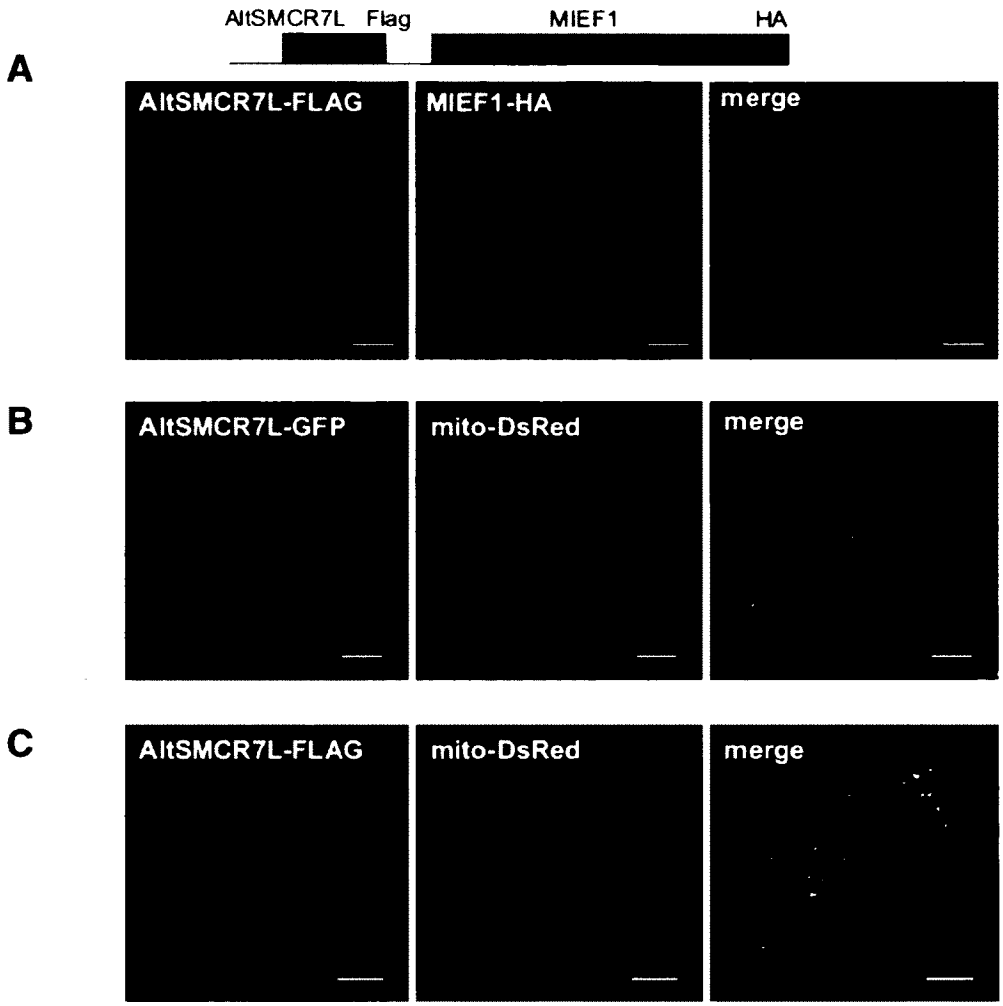


Figure 25. AltSMCR7L est localisée aux mitochondries. (**A**) Des cellules HeLa ont été transfectées avec une construction contenant l'ADNc correspondant à l'ARNm du gène *SMCR7L*, contenant l'AltORF encodant AltSMCR7L, et le RefORF encodant MIEF1. Les protéines ont été étiquetées avec un épitope FLAG (AltSMCR7L-FLAG) ou HA (MIEF1-HA) respectivement, en C-terminal. Une immunofluorescence a ensuite été réalisée avec des anticorps anti-FLAG (vert) et anti-HA (rouge). (**B**) Une construction codant pour une fusion entre AltSCMR7L et la GFP (en C-terminal) (AltSMCR7L-GFP, vert) a été cotransfectée dans des cellules HeLa avec une construction encodant le marqueur mitochondrial mito-DsRed (rouge). (**C**) La colocalisation entre mito-DsRed et AltSMCR7L-FLAG a également été observée dans une autre expérience de contransfection dans des cellules HeLa. (**A**,**B**,**C**) Les noyaux ont été marqués au Hoechst. Échelle, 10 µM

Prise en considération de l'expression d'AltORFs dans l'étude de la fonction des gènes: Les approches de routine actuellement utilisées dans les laboratoires pour évaluer la fonction d'une protéine donnée dans un processus cellulaire particulier devront certainement être revisitées. En effet, j'ai démontré que des ARN interférents (ARNi) utilisés pour diminuer l'expression d'un RefORF (celui codant pour PrP^C) provoquait la diminution à des niveaux similaires d'une protéine alternative (AltPrP) portée par le même gène (*PRNP*) (manuscrit 1, figure 6). Cela a également été observé pour le gène *ATXN1* (Bergeron et al, 2013). Les observations corollaires sont applicables lors d'essais de sur-expression (manuscrit 1, Figure 2 ; manuscrit 3, Figures 3 et 4) (Bergeron et al, 2013, Klemke et al, 2001). Des précautions particulières devraient donc être prises dans les méthodes utilisées pour moduler l'expression d'une protéine donnée. Tout d'abord, l'expression de protéines alternatives devrait être écartée. Si un ou des AltORFs sont présents et exprimés, alors les méthodes utilisées devraient prendre cela en compte. Par exemple, une expérience d'ARNi devrait être validée avec la complémentation dans le système expérimental par l'une ou l'ensemble des protéines dont l'expression est affectée, afin de s'assurer de l'identité de l'ORF portant la fonction affectée par l'ARNi.

Méthodes d'identification de nouveaux ORFs (protéogénomique)

Au cours de mes travaux, j'ai utilisé des méthodes de prédiction d'ORFs à partir de séquences transcrites afin de construire des bases de données utilisées pour l'identification de protéines par MS. Des approches alternatives de détection de nouveaux produits protéiques consistent en la traduction *in silico* de génomes ou transcriptomes entiers dans 6 cadres de lecture (3 sens, 3 antisens) et 3 cadres de lecture, respectivement. L'utilisation de séquences génomiques traduites constitue la méthode la moins biaisée, mais un problème associé est que le caractère fragmenté des gènes eucaryotes (structure intron/exon) complexifie grandement la cartographie des peptides détectés sur le génome. De plus, l'absence possible de preuves de la transcription d'une séquence donnée qui viendrait appuyer le potentiel de cette séquence à être traduite peut jeter des doutes sur la validité de certaines identifications. L'utilisation de bases de données de séquences transcriptomiques traduites écarte ces problèmes. Mais la taille de ces bases de données (comme celle

obtenues depuis le génome) est très grande, ce qui diminue le ratio signal/bruit et la sensibilité dans l'identification des peptides (Nesvizhskii, 2010). L'utilisation de bases de données de taille réduite telles que celles utilisées dans le manuscrit 3, certes moins exhaustives, limite cela. De plus en plus fréquemment, des bases de données personnalisées pour chaque échantillon analysé sont générées à partir de données de séquençage d'ARNm ou de *ribosome profiling* des échantillons avant analyse par MS (Menschaert et al, 2013, Slavoff et al, 2013). Cela permet de créer, en théorie, un espace de recherche idéal, puisque seuls les produits protéiques potentiellement traduits depuis les transcrits exprimés seront recherchés, diminuant la taille de la base de données. Néanmoins, là encore, des problèmes sont associés. Tout d'abord, l'instabilité des ARNm rend impossible cette approche dans certains échantillons biologiques. Par ailleurs, ces approches sont basées sur un préjugé : une protéine et le transcrit qui l'encode sont forcément exprimés (ou au moins tous deux détectables) à un instant donné dans l'échantillon analysé. Mais il ne peut être exclu que la durée de vie de produits protéiques dépasse parfois de loin celle des transcrits qui les encodaient, ou qu'un ARNm très faiblement exprimé, et donc possiblement non détecté, soit très efficacement traduit. Certaines protéines qui pourraient être présentes dans l'échantillon seraient donc exclues à tort de l'analyse. Cela serait aussi le cas pour les protéines présentes, par exemple, dans un type cellulaire analysé A, mais qui ont été synthétisées par un autre type cellulaire B et transporté vers le type cellulaire A. Enfin, l'utilisation d'une base de données protéique issue de données de ribosome profiling n'est pas applicable aux fluides biologiques, puisqu'ils ne synthétisent pas leurs propres constituants protéiques. Malgré les limitations qu'elle comporte (voir paragraphe ci-dessous), notre méthode de découverte de nouvelles protéines par prédiction d'ORFs depuis un transcriptome de référence, avec des critères définis, limite bon nombre des problèmes associés aux autres méthodes de protéogénomique.

Avantages et limites de la méthode de prédiction des AltORFs

Prédiction d'ORFs à partir de séquences transcrites : La prédiction d'ORFs depuis une séquence génomique est un problème non trivial, au regard de la complexité structurale des gènes eucaryotes (multiplicité d'introns/exons, de promoteurs alternatifs, de sites de polyadénylation alternatifs) (Brent, 2005). En comparaison, la prédiction d'ORFs à partir de séquences transcrites est bien plus aisée, puisque les règles de décodage de la traduction seules peuvent être utilisées dans l'algorithme de prédiction. Les multiples protéoformes possibles d'une même protéine alternative peuvent ainsi être prédites à partir des multiples variants d'ARNm issus d'un même gène. De plus, la prédiction d'ORFs à partir d'une molécule validée comme servant de substrat pour la traduction (un ARNm) renforce la confiance que cet ORF peut réellement être utilisé pour la synthèse protéique. L'exhaustivité des prédictions devient certes dépendante de celle de la base de données de transcrits utilisée au départ, mais l'avantage est que la capacité de toute ou partie de la séquence à être transcrite en ARNm n'est pas à mettre en doute. Le choix de la base de données d'ARNm revêt une importance particulière. Dans nos études, nous avons utilisé RefSeq comme base de données initiale d'ARNm afin de guider nos prédictions d'AltORFs. L'avantage d'utiliser cette ressource est qu'elle est sujette à une vérification régulière et stringente, permettant le référencement dans celle-ci d'ARNm dont les séquences sont dignes de confiance (Pruitt et al, 2007). La majorité des ORFs déjà référencés ainsi que ceux prédits à partir de ces séquences sont donc moins sujets à changements et corrections que si une base de données d'ARNm plus libérale (par exemple Gencode, ou UCSC known genes) (Harrow et al, 2012, Karolchik et al, 2007) était utilisée, ce qui rendrait la base de données d'AltORFs régulièrement obsolète. L'utilisation de ces autres bases de données pourrait en revanche présenter l'avantage d'écarter l'hypothèse selon laquelle certains AltORFs pourraient être en réalité des RefORFs dans d'autres transcrits, voire dans d'autres gènes non identifiés dans RefSeq. Ceci est illustré par le fait que suite à une mise à jour de RefSeq postérieure à la génération de la base de données d'AltORFs humains, un nouvel ARNm a été incorporé correspondant au gène *Pet117*, situé en amont du gène *CSRP2PB*, pour lequel un AltORF avait été prédit dans la région 5'UTR. Cet AltORF (dont l'expression a été validée par LC-MS/MS, voir manuscrit 3, Table S4) prédit dans l'ARNm de *CSRP2BP* correspond en fait au RefORF de *Pet117*. La séquence

5'UTR de l'ARNm de *CSRP2BP* contenant cet AltORF a d'ailleurs été mise à jour également, l'AltORF ne s'y trouvant plus. Ainsi, la mise à jour de la base de données RefSeq rendra nécessaire celle de la base de données d'AltORFs. Il sera également important de vérifier que des AltORFs dont l'expression au niveau protéique aura été validée expérimentalement ne soient pas éliminés des prédictions après mise à jour en raison de la suppression d'un transcrit par manque de preuves de l'expression du RefORF associé, par exemple.

Il apparaît donc crucial, pour obtenir les prédictions d'AltORFs les plus justes possibles, que des bases de données exhaustives et de haute qualité soient utilisées (telles que RefSeq). Ces bases de données ne sont pas encore parfaites, mais l'analyse du nombre croissant de séquences générées dans les expériences de séquençage d'ARNm contribue à leur amélioration continue.

Choix des codons initiateurs pris en compte : Il est désormais bien établi que bien d'autres codons que l'AUG (en particulier CUG et les autres codons différant d'AUG à un nucléotide près) sont efficacement utilisés comme sites d'initiation de la traduction (Ingolia et al, 2011). Cependant, AUG reste le codon initiateur majoritaire identifié dans les études de *ribosome profiling* (51% des sites d'initiations sont des AUG, 16% des CUG), bien qu'il ne représente que 1,7% des codons du transcriptome (Lee et al, 2012). AUG reste donc le codon semble-t-il le plus favorable à l'initiation de la traduction, d'où son utilisation pour prédire les AltORFs dans l'approche que nous avons utilisée. Comme mentionné dans l'introduction, la situation diffère selon la région de l'ARNm. Les sites d'initiation en aval de celui du RefORF sont en majorité des AUG (47%), où les CUG n'en constituent que 6%. Dans les régions 5'UTRs, où CUG devient le codon initiateur majoritaire (30,3%), l'AUG n'est plus utilisé qu'à hauteur de 25,6% (Lee et al, 2012). Une implémentation, dans la prédiction d'AltORFs dont le site d'initiation est situé dans le 5'UTR, de critères validés permettant de prédire les sites d'initiation (Wegrzyn et al, 2008) pourrait être réalisée afin d'améliorer notre algorithme. Il faudrait tout de même valider que les résultats obtenus (pourcentage d'utilisation des codons initiateurs) avec ces critères additionnels sont en accord avec les observations expérimentales apportées par les études de *ribosome profiling*. Une meilleure définition de règles de l'initiation de la traduction en aval du codon initiateur de référence serait nécessaire avant d'imaginer faire de même avec la prédiction de ces sites

d'initiation. En l'absence de telles règles régions-spécifiques validées applicables pour la prédiction d'AltORFs, l'utilisation de l'AUG comme codon d'initiation, avec potentiellement l'ajout du CUG dans les régions 5'UTRs, semble être à mes yeux la meilleure solution, sous peine d'augmenter drastiquement le nombre de faux positifs. D'ailleurs, la méthode de prédiction semble déjà assez robuste comme en atteste l'analyse des peptides N-acétylés pour les protéines alternatives détectées à travers l'ensemble de mes analyses par LC-MS/MS (886/889 prédictions correctes), et la validation par mutagenèse de plusieurs sites d'initiation alternatifs (manuscrit 3, Table S4 et Figure 3).

Comparaison aux autres études *in silico* prédisant les AltORFs : L'algorithme de prédiction des AltORFs utilisé dans le manuscrit 2, et modifié dans le manuscrit 3 pour inclure les régions UTRs des ARNm, a été développé de manière à ce qu'il prédise avec succès des exemples connus d'AltORFs générant un produit protéique. La tentation est forte, en générant un tel algorithme, d'appliquer plusieurs critères permettant de produire une liste prédisant les candidats ayant le plus de chance d'être exprimés et fonctionnels, avec un taux de faux positifs le plus bas possible. Pour les raisons expliquées dans l'introduction et dans le manuscrit 2, les critères de longueur minimale des AltORFs, de contexte Kozak, et de conservation évolutive au niveau de la séquence protéique sont à double tranchant puisqu'ils excluent de ces listes soit disant de haute confiance des AltORFs dont l'expression a été validée *in vivo* (Chung et al, 2007, Ribrioux et al, 2008, Xu et al, 2010). Les AltORFs chevauchant les RefORFs des gènes *PRNP* et *GNAS* étaient ainsi absents des études *in silico* réalisées avant celles présentées ici, sans pour autant qu'il ait été vérifié que les candidats soit disant de haute confiance identifiés sont réellement exprimés, mettant en doute l'utilité d'une stringence trop grande dans les prédictions. En particulier, le nombre de plus en plus grand de produits protéiques de petite taille effectivement exprimés et ayant une importance fonctionnelle tend à diminuer l'intérêt de n'inclure que les AltORFs les plus grands dans les prédictions. De même la conservation au cours de l'évolution n'est pas un gage de l'expression ou de la fonctionnalité d'un AltORF. AltPrP est exprimée chez l'humain malgré l'absence d'un AltORF correspondant chez la souris (manuscrit 1). D'autre part, ALEX montre une conservation évolutive faible : la coévolution avec sa protéine de référence pour une interaction intermoléculaire conservée amène sa séquence protéique à subir une évolution rapide (Nekrutenko et al, 2005). Les critères choisis pour

les algorithmes utilisés dans les manuscrits 2 et 3 sont certes bien moins stringents, mais permettent ainsi de prédire des candidats à l'expression validée. Intuitivement, le taux de faux positifs dans les candidats prédits est possiblement augmenté, mais cela restera à vérifier une fois qu'un répertoire plus exhaustif de protéines alternatives exprimées sera disponible.

D'une manière générale, indépendamment des méthodes prédictives utilisées, seule la validation expérimentale de l'expression des AltORFs permettra d'évaluer l'étendue réelle de leur contribution au protéome, et d'améliorer les méthodes de prédictions.

Evaluation *a priori* ou *a posteriori* de la conservation évolutive des AltORFs : En fonction des espèces, les bases de données d'ARNm disponibles varient dans leur qualité (validation des séquences, aux extrémités 5' et 3' en particulier), et leur exhaustivité. Les bases de données de séquences humaines sont ainsi souvent très complètes, tant les efforts pour décrypter notre génome ont été immenses. Cela n'est pas le cas d'autres espèces comme le bovin par exemple, où les séquences UTRs sont moins caractérisées que chez l'humain ou la souris. Ainsi, les différences dans le contenu des bases de données spécifiques à chaque espèce posent problème dans le cadre d'une analyse comparative de séquences. Dans une approche où, pour être considéré comme un candidat prédit de façon valide, un AltORF doit être conservé évolutivement, de nombreux candidats potentiels pourraient ainsi être éliminés, par l'absence imputable à un manque de données expérimentales d'un transcrit donné dans la base de données d'ARNm d'une espèce donnée. Pour cette raison, l'approche utilisée dans le manuscrit 3 d'alignements par paires présente un intérêt évident. Elle permet d'éviter l'élimination de vrais positifs, tout en identifiant certains candidats à potentiel fonctionnel fort. Cela se fait au détriment de la robustesse introduite par les alignements multiples de séquences, qui donnent une meilleure estimation de la conservation évolutive. Cependant, comme je l'ai déjà discuté dans le cas d'ALEX ou d'AltPrP par exemple, la conservation au strict niveau de la séquence en acides aminés n'est gage ni de la capacité à être exprimé, ni de celle à'être fonctionnel. La conservation évolutive estimée après prédiction dans mes travaux est donc particulièrement adaptée à l'analyse exploratoire de l'expression des protéines issues d'AltORFs, tout en permettant d'identifier des candidats d'intérêt particulier pour des études fonctionnelles subséquentes.

Mécanismes d'expression des protéines alternatives

Variétés des mécanismes d'expression possibles : La compréhension du mécanisme d'expression d'une protéine alternative donnée n'est pas triviale. Cela a des conséquences directes sur les possibilités de régulation spatio-temporelle de l'expression de différents groupes de protéoformes à partir d'une région génomique donnée. Deux cas de figures généraux sont envisageables : un AltORF peut être traduit depuis un ARNm multicodant (celui dans lequel il a été prédit et contenant le RefORF) et/ou depuis un autre ARNm séparément du RefORF. Les AltORFs du gène MKKS constituent un très bon exemple de ces deux possibilités (Akimoto et al, 2013). Des indications peuvent être obtenues en consultant des bases de données d'ADNc (analyse dans le UCSC Genome Browser par exemple), mais cela doit ultimement être déterminé expérimentalement (par exemple 5' et 3' RACE (Rapid Amplification of cDNA Ends) ou Northern blot, transfection *in cellulo* d'ARNm synthétisés *in vitro*).

Le fait de prédire des AltORFs depuis les ARNm permet de s'assurer qu'ils sont bien présents dans des molécules capables de subir la traduction, et donc qu'ils sont potentiellement traduits. Cela étant dit, il n'est pas garanti qu'un AltORF prédit soit bien exprimé depuis l'ARNm dans lequel il a été prédit. Entre autres, l'utilisation de promoteurs ou sites de polyadénylation alternatifs, ou des mécanismes de trans-épissage pourraient expliquer cela. Dans une moindre mesure, l'utilisation de cadres de lecture chevauchant par épissage alternatif pourrait expliquer la détection par MS de peptides présents dans des AltORFs prédits (chevauchant le RefORF). Seulement 7% des gènes épissés alternativement semblent cependant avoir des régions multi-codantes de ce type (Liang & Landweber, 2006), ce qui limite la portée de ce phénomène. L'exemple de l'AltORF à cheval entre le RefORF et le 3'UTR du gène *C11Orf48*, dont l'expression a été validée par mes expériences de LC-MS/MS (manuscrit 3, Table S4) et par des expériences de *ribosome profiling* (Michel et al, 2012), est particulièrement éloquent. Son codon initiateur est localisé plus de 600 nt en aval de l'extrémité 5' de l'ARNm de la base de données RefSeq dans lequel il a été prédit. D'après le modèle de l'initiation de la traduction par balayage, majoritairement utilisé chez les eucaryotes, son codon d'initiation AUG (le 8ème de la séquence en ordre d'occurrence, et dans un contexte Kozak sub-optimal) n'a que peu de chances de servir de site d'initiation. Pourtant, la protéine alternative correspondante

semble exprimée à des niveaux relativement abondants, puisqu'elle a été détectée avec 11 peptides distincts pour une couverture de séquence de 40,3% dans les cellules HeLa (manuscrit 3, Table S1). Néanmoins, un ARNm absent dans RefSeq mais présent dans la base de données Ensembl semble pouvoir soutenir l'expression du seul AltORF (Michel et al, 2012), offrant une alternative à la présence d'un ARNm multicodant dans ce cas, bien que cela n'ait pas été vérifié expérimentalement. Conceptuellement, cet ORF pourrait être considéré comme alternatif au niveau du gène, mais si son expression provient bel et bien d'un transcrit alternatif, elle serait considérée comme le RefORF de cet ARNm. Le cas de *C11Orf48* n'est certainement pas isolé, et pourrait bien être observé pour l'expression d'AltORFs ayant été prédits dans les régions 3'UTR de certains ARNm, le mécanisme canonique de balayage expliquant difficilement une initiation de la traduction si loin de l'extrémité 5'. Les essais de validation de l'expression de tels AltORFs détectés en LC-MS/MS par la transfection de vecteurs d'expression contenant des ADNc bicistroniques n'ont d'ailleurs pas abouti à une détection des protéines alternatives correspondantes (données non présentées).

Si la traduction d'un AltORF est réellement soutenue par l'ARNm à partir duquel il a été prédit, qui permet aussi l'expression d'un RefORF, alors il s'agit bel et bien d'un ARNm multicodant. Un tel ARNm présente un avantage énergétique pour la cellule, lui permettant de produire plusieurs groupes de protéoformes à partir d'un évènement transcriptionnel unique. Dans ce cas, l'investigation de leur mécanisme d'expression doit porter sur les évènements amenant à l'initiation de la traduction à leur codon d'initiation. La validation expérimentale de l'identité de leur codon d'initiation est alors d'une grande importance, par la détection de peptides N-acétylés par MS, ou par mutagenèse dirigée. Afin de déterminer si leur expression est dépendante ou indépendante de la coiffe, plusieurs méthodes sont disponibles. Des inhibiteurs de l'interaction eIF4E/eIF4G (comme la petite molécule 4EGI-1, ou des versions constitutivement active des protéines 4EBP) permettent d'inhiber spécifiquement la traduction coiffe-dépendante. L'utilisation, dans des vecteurs d'expression eucaryotes, d'un promoteur permettant une transcription par l'ARN polymérase 1 (qui génère des ARNs non coiffés) permet aussi de répondre à cette question. Si l'expression d'un AltORF est maintenue dans de telles conditions, cela permettrait de conclure qu'un mécanisme indépendant de la coiffe (IRES ou CITE) est à l'origine de leur

expression. Afin de discriminer entre ces deux mécanismes, l'utilisation d'un système rapporteur bicistronique est un moyen efficace. Si la région en amont du codon initiateur de l'AltORF, placée dans la région intercistronique du système rapporteur, permet d'induire la traduction du second cistron, la présence d'un IRES est envisageable (bien que cela nécessite des contrôles additionnels, pour écarter l'hypothèse d'un promoteur alternatif par exemple). Il faut noter que l'utilisation d'IRES pourrait bien expliquer l'expression d'AltORFs traduits dans des régions éloignées de l'extrémité 5' des ARNm (telles que les 3'UTR). Si la séquence introduite dans la région intercistronique n'induit pas la traduction du second cistron, un CITE (qui est dépendant d'une extrémité 5' libre pour permettre l'entrée du ribosome) est probablement présent. Cela peut être vérifié en introduisant les séquences putatives du CITE dans un ARNm hétérologue, et en vérifiant si elles permettent la traduction de l'ORF de l'ARNm hétérologue dans des conditions ou la traduction coiffe dépendante est inhibée. Une fois le mécanisme d'entrée du ribosome établi, il faudra ensuite vérifier par des expériences de mutagenèse dirigée si les mécanismes de balayage, de shuntage ribosomal ou de réinitiation de la traduction sont nécessaires à l'expression de l'AltORF étudié.

En résumé, les mécanismes d'expression des AltORFs devront probablement être étudiés au cas par cas, en déterminant d'abord quel est le ou quels sont les ARNm à partir du ou desquels il est exprimé, avant de décortiquer les mécanismes moléculaires autorisant l'initiation de la traduction à leur codon d'initiation. Les mécanismes régissant l'expression d'AltORFs en comparaison aux RefORFs correspondants permettront d'obtenir des indications quant à la régulation de leur expression, couplée ou découplée, en fonction de conditions cellulaires particulières. En condition de stress cellulaire, par exemple, la traduction coiffe-dépendante est inhibée (Ron, 2002, Sonenberg & Hinnebusch, 2009). La majorité de la traduction est inhibée (du moins cela est-il démontré pour les RefORFs), mais la traduction de certains ORFs peut être spécifiquement maintenue ou augmentée dans de telles conditions (Sonenberg & Hinnebusch, 2009). Cela semble être le cas pour AltPrP en condition de stress au réticulum endoplasmique (manuscrit 1, Figure 5).

Conciliation entre la présence d'AltORFs et la dégradation des ARNm non-sens (NMD) :

Le mécanisme de NMD (*nonsense mRNA decay*, en anglais) permet aux cellules eucaryotes d'éliminer les ARNm contenant des codons stop prématurés, afin d'éviter la synthèse de

protéines ayant des effets possiblement délétères. Lorsqu'un ARNm est traduit pour la première fois, le complexe d'élongation de la traduction élimine les complexes de jonction des exons (CJE) localisés à proximité des jonctions exon-exon. La règle de la NMD est que si le codon stop reconnu lors de ce premier tour de traduction est localisé en amont d'un point limite (50-55 nucléotides en amont du dernier CJE), il est considéré comme un codon de terminaison prématuré. L'ARNm est alors ciblé vers la machinerie de dégradation. Un codon stop prématuré provoquant la NMD peut survenir dans plusieurs cas : en conséquence à un épissage alternatif (rétention d'un intron par exemple) ; dans un ARNm bicistronique, contenant un AltORF inclus dans le RefORF ; dans des transcrits de pseudogènes, d'éléments transposables ou de gènes sujets à réarrangement programmé (Kervestin & Jacobson, 2012). Ainsi, selon la localisation du codon stop d'un AltORF subissant la traduction, il est possible que l'ARNm qui l'encode soit sujet à la NMD. Après la traduction d'un uORF, la NMD peut avoir lieu, comme cela a été démontré chez les plantes, d'une manière dépendante de la taille de l'uORF (Nyiko et al, 2009). Cependant, la réinitiation de la traduction après un codon stop prématuré permet à un ARNm d'échapper à la NMD (Neu-Yilik et al, 2011), suggérant qu'un uORF par exemple pourrait être traduit dès le premier tour de traduction. Un codon stop prématuré situé dans le dernier exon peut également rendre un ARNm insensible à la NMD, ce qui permettrait la traduction de certains AltORFs (Perrin-Vidoz et al, 2002). De plus, l'efficacité de la NMD apparaît tissu-spécifique (Zetoune et al, 2008). La NMD ne toucherait que 1 à 10 % des ARNm (Kervestin & Jacobson, 2012), alors que plus de la moitié des ARNm de mammifères subissent l'utilisation de multiples sites d'initiation (et donc surement de sites de terminaison précoce) de la traduction (Ingolia et al, 2011, Lee et al, 2012). Enfin, il faut noter que si le RefORF d'un ARNm est traduit en premier lors du premier tour de traduction, la NMD provoqué par la traduction d'un AltORF n'est plus à considérer. Toutes ces indications sur l'efficacité toute relative des mécanismes de NMD impliquent que l'existence de ce mécanisme est loin d'être incompatible avec une utilisation répandue des AltORFs comme source de diversité protéique.

Régulation de la synthèse d'AltPrP : Dans le manuscrit 1, j'ai analysé la régulation du niveau d'expression d'AltPrP en réponse à des stress variés (Figures 4 et 5). Lors d'un stress au réticulum endoplasmique (RE), mais pas lors d'un stress thermique ou oxidatif,

l'expression d'AltPrP se trouvait augmentée lorsque celle-ci était exprimée depuis une construction codant pour PrP et AltPrP ($PrP^{(HA)}$), mais pas depuis une construction codant juste pour AltPrP (condition $AltPrP^{HA}$). Ces résultats excluent une régulation au niveau post-traductionnel. De plus, cette augmentation d'expression est spécifique à AltPrP, l'expression de PrP ne semblant pas augmentée, et touche aussi AltPrP produite de façon endogène. Un effet au niveau transcriptionnel n'est que peu probable. Un effet traductionnel spécifique à l'expression d'AltPrP par initiation alternative de la traduction est donc envisageable. Il est pour l'instant inconnu si cette utilisation accrue d'un AUG en aval de celui du RefORF est spécifique au locus *PRNP* ou si ce mécanisme pourrait être généralisé. En plus du lancement d'un programme transcriptionnel bien défini lors du stress au RE, il est intéressant de noter qu'une réponse traductionnelle est mise en place (Lewis et al, 2008, Ventoso et al, 2012). Ainsi, la traduction de certains ARNm est maintenue voire augmentée, alors que le taux de traduction global est diminué de manière précoce suite à l'induction du stress au RE (Ron, 2002). Ceci est obtenu par une diminution de l'initiation de la traduction coiffe-dépendante (via la phosphorylation d'eIF2α) (Ron, 2002), et une augmentation de mécanismes coiffe-indépendants (IRES ou CITE, ce qui reste à définir clairement) (Lewis et al, 2008, Sonenberg & Hinnebusch, 2009). Il peut donc être supposé qu'AltPrP serait traduite de manière coiffe-indépendante (au moins en condition de stress au RE). Plusieurs protéines dont la traduction est maintenue lors du stress au RE permettent de favoriser la survie cellulaire. Cela ne semble pas être le cas pour AltPrP, puisque sa surexpression n'affecte pas la viabilité cellulaire estimée par essai XTT en réponse à un stress au RE (données non présentées). Si d'autres protéines produites par utilisation de codons d'initiation en aval de celui du RefORF sont également induites lors d'un stress au RE, il serait important de valider les implications biologiques associées en rapport à la réponse à ce stress.

Un lien entre AltPrP et les maladies à prion ?

Outre la régulation de la synthèse d'AltPrP, j'ai également étudié l'expression d'AltPrP au niveau de son mécanisme de dégradation. J'ai montré qu'AltPrP était dégradée rapidement (durée de vie <30 min) de manière protéasome-dépendante (manuscrit 1, Figure 4). Il a été

observé qu'une inhibition du protéasome ainsi qu'un stress au RE semblaient être impliqués dans le processus de pathogenèse des maladies à prion. Puisque ces deux stress augmentent l'expression d'AltPrP, la question se pose de savoir si AltPrP pourrait être impliquée dans le développement des maladies à prion. La surexpression d'AltPrP au cours de la pathogenèse des maladies à prion devrait être testée. Les anticorps dirigés contre AltPrP bovine, ovine ou de cervidé que nous avons fait synthétiser ne nous ont malheureusement pas permis de détecter AltPrP endogène chez ces espèces (données non présentées).

Comme mentionné dans le paragraphe précédent, la surexpression d'AltPrP ne semble pas moduler la réponse au stress au RE. En revanche, il a été démontré qu'un stress au RE ainsi qu'une inhibition du protéasome induisaient une augmentation d'agrégats insolubles de PrP^C dans le cytosol (Dron et al, 2009, Kristiansen et al, 2005, Ma et al, 2002b). J'ai donc étudié l'impact de ces deux stress sur la localisation cellulaire d'AltPrP. Le stress au RE n'a pas eu d'effet, AltPrP (étiquetée HA ou eGFP en C-terminal) restant localisée aux mitochondries (données non présentées). Mais lors d'une inhibition du protéasome par traitement à l'epoxomycine de cellules exprimant AltPrP fusionnée à eGFP, cette dernière subit une délocalisation des mitochondries. Elle est alors localisée majoritairement dans un agrésome caractéristique, entouré d'une cage de vimentine et situé dans une région péri-nucléaire (Figure 26) (Johnston et al, 1998). Il serait important de valider l'existence de telles structures *in vivo*, dans des tissus d'animaux ou de patients atteint de maladies à prion. Un anticorps qui fonctionne en immunohistochimie devra cependant être obtenu au préalable.

L'infection de souris par des prions constitue un moyen répandu d'étudier dans un temps raisonnable la contribution de différents facteurs dans la susceptibilité au développement de ces maladies (pénétrance, temps d'incubation). Nous avons donc généré une lignée transgénique de souris avec de multiples insertions aléatoires de transgènes dirigeant la transcription d'un ARNm codant pour AltPrP humaine. Les tests d'expression du transgène dans le cerveau des souris ont cependant indiqué, au niveau de l'ARNm comme de la protéine, un niveau d'expression trop faible dans cette lignée pour que le projet soit poursuivi (données non présentées).

Un point supplémentaire de discussion concernant l'implication potentielle d'AltPrP dans le développement des maladies à prion est disponible en annexe, sous forme d'un échange de lettres avec le Dr Peter Wills (Auckland University, Nouvelle-Zélande).

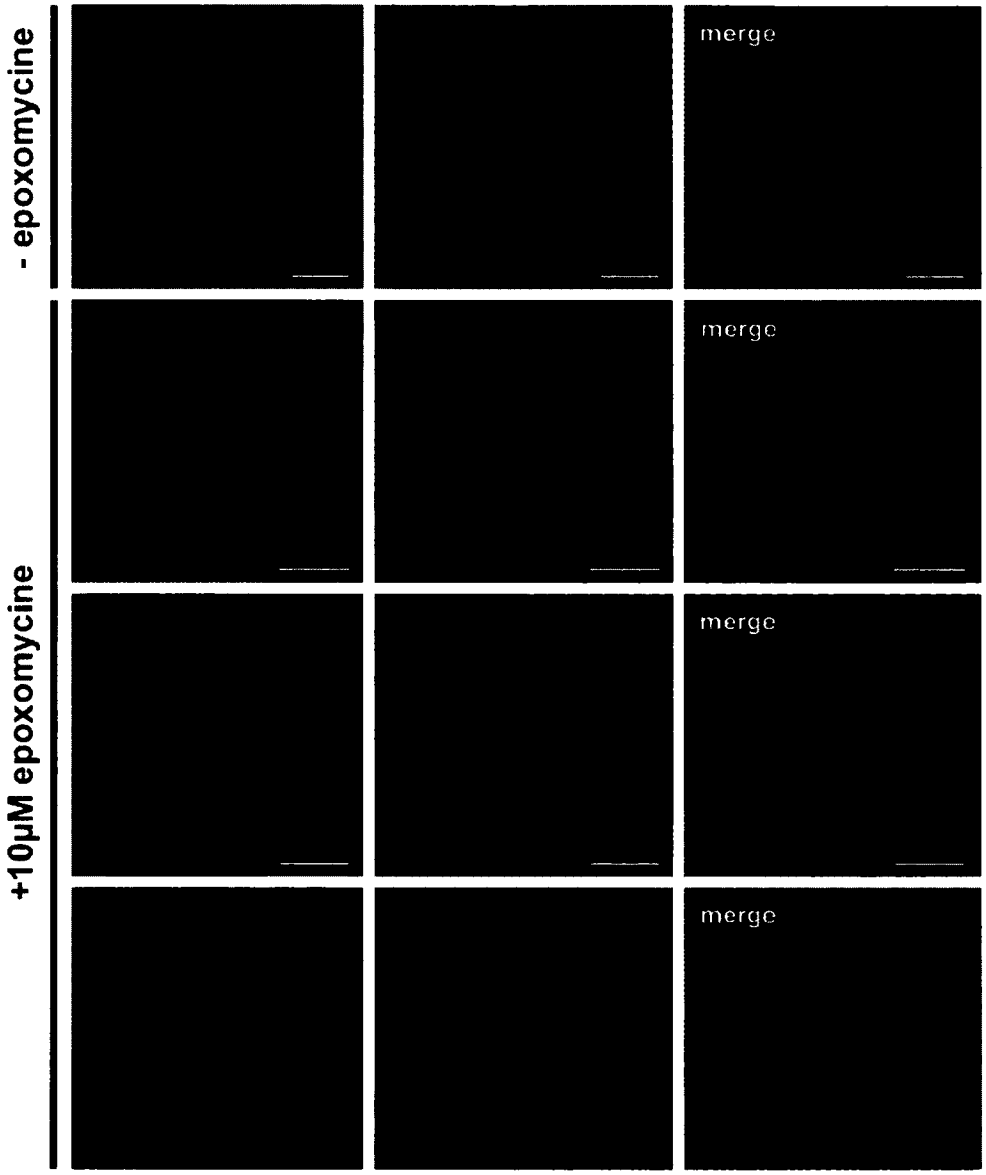


Figure 26. AltPrP s'accumule sous forme d'agrésomes lorsque le protéasome est inhibé. Des cellules HeLa ont été transfectées avec une construction codant pour AltPrP avec une étiquette eGFP en C-terminal (vert), puis cultivées en absence ou en présence d'epoxomycine (un inhibiteur du protéasome). Puis les cellules ont été immunomarquées avec des anticorps anti-mHSP70 (mitochondries), anti-vimentine (cytosquelette) ou anti-γ-tubulin (centrosome) (rouge). Les images ont été obtenues par microscopie confocale. Échelle : 10µm.

Détection des protéines alternatives

Bien que les résultats présentés dans le manuscrit 3 indiquent une contribution sans précédent des AltORFs à l'établissement du protéome, le nombre absolu et relatif par rapport aux protéines de référence de protéines alternatives détectées peut paraître faible. Excepté pour les échantillons de sérum/plasma, où les protéines alternatives sont largement représentées (~50% du nombre total de protéines détectées), ces dernières représentent moins de 2,5% des protéines détectées dans la plupart des échantillons analysés. Cependant, le nombre de protéines alternatives détectées est fort probablement une sous-estimation du nombre total de protéines alternatives contribuant réellement au protéome, pour plusieurs raisons. Tout d'abord, des protéines alternatives dont l'expression a été validée dans la littérature (AltPrP, AltATXN1, ALEX pour ne citer qu'elles) n'ont pas été détectées dans les expériences de MS analysées dans le manuscrit 3, indiquant le caractère non exhaustif du répertoire généré dans cette étude. Ensuite, un nombre limité de tissus et types cellulaires a été analysé. L'analyse d'un plus grand nombre d'échantillons, avec des profils transcriptomiques variés, amènera à la découverte de nouveaux AltORFs effectivement traduits en produits protéiques. Par exemple, il a été déterminé que le protéome des cellules HeLa était composé d'au moins 10255 protéines (de référence) issues de 9207 gènes (Nagaraj et al, 2011). Cela signifie que probablement moins de la moitié des 20 300 gènes codant pour des protéines chez l'humain sont exprimés dans ces cellules. Par ailleurs, dans l'étude présentée dans le manuscrit 3, la profondeur de couverture protéomique était inférieure à celle de l'étude de Nagaraj *et al.*, puisque 5 558 protéines de référence seulement ont été détectées. La base de données utilisée pour l'identification protéique était différente entre les deux études (IPI pour Nagaraj *et al.*, Genbank dans notre cas), et surtout les techniques utilisées pour le fractionnement des protéines et peptides différaient, ce qui peut avoir une grande influence sur le nombre de protéines identifiées. L'application de méthodes de fractionnement supplémentaires pourrait donc amener à une plus grande sensibilité dans l'identification des protéines alternatives. Une autre explication au nombre inférieur de protéines alternatives détectées, comparé aux protéines de référence, tient dans la taille réduite des AltORFs comparé aux RefORFs. La détection de protéines de petite taille par spectrométrie de masse est un problème récurrent. En effet, en protéomique *shotgun* (identification des peptides issus d'un clivage protéolytique), la probabilité de

détecter une protéine donnée est reliée au potentiel de cette protéine à produire une quantité abondante de peptides. Ainsi, plus une protéine est de taille et d'abondance réduite, moins elle génère de peptides. L'étude de Slavoff *et al.* suggère que les protéines alternatives sont exprimées à des abondances comparables aux protéines de référence (10 à 2 000 copies par cellules, pour trois candidats testés) (Slavoff et al, 2013). Leur taille réduite (manuscrit 3, figure 1E) explique donc certainement en grande partie la proportion réduite de protéines alternatives détectées, ce qui pourrait être amélioré en utilisant des méthodes de préparation d'échantillons contenant des étapes de fractionnement supplémentaires (séparation des peptides sur gradient isoélectrique par exemple), diminuant la complexité du mélange de peptides à analyser. L'analyse de protéines de taille réduite et les méthodes de peptidomique offrent également une avenue intéressante en complément aux analyses indépendantes du poids moléculaire, comme en atteste la proportion élevée (~24% du total d'identifications) de protéines alternatives identifiées parmi les protéines de moins de 10 kDa (cellules HeLa, manuscrit 3, Table 1). De manière générale, l'amélioration continuelle des technologies de spectrométrie de masse vers une sensibilité toujours accrue (Mann et al, 2013) permettra également de contribuer à l'établissement d'un répertoire exhaustif des protéines alternatives dans plusieurs espèces. Il est important de noter que même sans générer de nouvelles données brutes de spectrométrie de masse, et comme réalisé en partie dans le manuscrit 3, la simple ré-analyse de données brutes déjà générées avec une base de données à laquelle sont ajoutées les séquences de protéines alternatives prédites permettra de contribuer significativement à ce répertoire. L'amélioration de la détection des AltORFs permettra également d'étudier plus en détail les protéines alternatives, en obtenant des informations sur les protéoformes produites par épissage alternatif ou modifications post-traductionnelles par exemple. Des profondeurs d'analyse protéomique plus poussées autoriseront aussi la détection facilitée et reproductible de produits protéiques difficilement détectables (comme les protéines de petite taille). Ceci mènera à la possibilité de leur incorporation dans des analyses protéomiques comparatives, et à identifier leurs différentes protéoformes possibles.

Expression d'AltORFs chez d'autres espèces que l'humain

Dans les travaux présentés ici, un accent particulier a été mis sur la validation de l'expression de protéines alternatives chez l'humain. Qu'en est-il de leur expression dans d'autres espèces ?

Concernant AltPrP, son codon initiateur (celui identifié chez l'humain) est présent chez les mammifères supérieurs, mais absent chez les rongeurs, où aucun autre codon AUG (en amont ou en aval) présent dans ce cadre de lecture +3 ne pourrait être utilisé pour soutenir l'expression d'AltPrP (manuscrit 1, Figure 1). Il est envisageable que le codon GTG qui le remplace chez le rat, le hamster et la souris, ou le CTG chez le vison, puisse être utilisé comme codon initiateur, même si cela semble peu probable, étant donné leur fréquence réduite d'utilisation en aval de l'AUG de l'ORF de référence (Lee et al, 2012). En revanche, bien que la preuve de l'expression *in vivo* d'AltPrP n'ait été obtenue que chez l'humain (manuscrit 1, Figure 6), le haut niveau de conservation d'AltPrP chez les mammifères supérieurs et l'expression d'une version étiquetée HA à partir de plasmides permettant la surexpression de PrP^C (bovine, ovine, de cervidés) indique que cette protéine alternative est certainement exprimée de façon endogène chez les mammifères supérieurs (manuscrit 1, Figures 1 et 2).

Tout comme chez l'humain, et grâce aux mécanismes conservés d'initiation de la traduction, il est possible de prédire de nombreux AltORFs chez d'autres eucaryotes, jusqu'à la levure, en utilisant l'algorithme présenté dans le manuscrit 3. Toutefois, la prédiction seule d'un AltORF ne présage en rien de son potentiel à être traduit. En effet, il serait par exemple possible d'identifier des ORFs dans des introns qui ne seront jamais inclus dans des ARNm subissant la traduction. La liste d'AltORFs prédits par l'utilisation de notre algorithme est certainement composée d'un nombre important de faux positifs, c'est-à-dire d'AltORFs qui ne sont pas réellement utilisés pour traduire des protéines. Une méthode qui permet d'identifier des AltORFs candidats potentiellement exprimés et fonctionnels est l'utilisation d'analyse comparative de séquences, afin de retenir ceux conservés au cours de l'évolution, donc possiblement sous pression de sélection positive. Bien que cela ne soit pas une garantie de l'expression des candidats identifiés, cela nous a permis d'identifier des dizaines (chez la levure), centaines (chez les invertébrés), ou milliers (vertébrés) de protéines alternatives candidates qui pourraient être exprimées et

fonctionnelles à travers plusieurs espèces eucaryotes (manuscrit 3, figure 6). Néanmoins, seules des approches expérimentales permettent de valider l'expression des AltORFs prédits, qu'ils soient conservés ou non.

Au cours de mon doctorat, j'ai analysé des données brutes de spectrométrie de masse provenant d'échantillons d'espèces pour lesquelles les bases de données de protéines alternatives avaient été générées. J'ai ainsi identifié 24 protéines alternatives chez la levure *S. cerevisiae* à partir des données brutes de l'étude de de Godoy *et al.* (de Godoy et al, 2008) (données non présentées), dont aucune n'était conservée chez l'humain. Dans le sécrétome de cellules β-pancréatiques murines, j'ai détecté 8 protéines alternatives, dont 3 sont conservées chez l'humain (Tattikota et al, 2013) (données non présentées).

Ces résultats, conjugués aux exemples décrits dans la partie 3.1 de l'introduction du présent manuscrit, indiquent que des protéines alternatives sont exprimées dans d'autres espèces eucaryotes que l'humain, comme pouvait le laisser penser la conservation des mécanismes de traduction.

Des AltORFs chez les procaryotes ?

Chez les procaryotes, plusieurs mécanismes d'initiation de la traduction autorisent l'utilisation d'AltORFs dans les ARNm matures. Comme indiqué dans l'introduction (paragraphe 2.1), le mécanisme canonique d'initiation de la traduction chez les procaryotes dépend de la présence d'une séquence Shine-Dalgarno en amont d'un codon initiateur (dans une région non traduite), ce qui permet déjà la traduction de plusieurs ORFs par ARNm. Des séquences SD peuvent également être présentes à l'intérieur des régions codantes bactériennes (Ponnala, 2010). Bien qu'il ait été suggéré que la présence de ces séquences SD internes permet de faciliter la traduction de régions contenant des codons rares, leur association à la présence d'AltORFs n'a pas été vérifiée. L'existence de mécanismes d'initiation ne nécessitant pas la présence d'une séquence SD (Boni et al, 1991, Hering et al, 2009, Kozak, 2005, Scharff et al, 2011) laisse aussi ouverte la possibilité que des ORFs autres que ceux référencés dans les ARNm procaryotes (qu'ils soient déjà définis comme polycistroniques ou non) soient utilisés pour la production de nouvelles protéines. Une base de données référençant des AltORFs prédits de plus de 100 codons dans le génome de 481

organismes procaryotes existe déjà (incluant des AltORFs présents sur les brins complémentaires des gènes inclus dans les prédictions) (Pedroso et al, 2008). La traduction des protéines alternatives associées n'a en revanche pas été vérifiée de façon approfondie.

CONCLUSION

En accord avec les objectifs présentés dans la partie 4 de l'introduction, les conclusions majeures de mes travaux de recherche de doctorat sont les suivantes :

- dans une approche par gène candidat, j'ai prouvé que chez les mammifères supérieurs, le gène *PRNP* dirige l'expression d'une seconde protéine appelée AltPrP, traduite depuis un AltORF situé dans un cadre de lecture alternatif et chevauchant l'ORF codant pour la protéine prion. La contribution d'AltPrP à la pathogenèse des maladies à prion est suggérée par la régulation de son expression lors de conditions de stress cellulaire qui participent à ces maladies.

- une approche computationnelle indique que des milliers d'AltORFs peuvent être prédits dans le transcriptome de plusieurs organismes eucaryotes. Environ 90 % des ARNm humains possèdent au moins un AltORF prédit.

- la validation par une méthode protéomique à large échelle de l'expression endogène d'AltORFs prédits permet de conclure que les protéines alternatives contribuent significativement au protéome chez les eucaryotes.

Analysés dans leur globalité, mes travaux de doctorat mettent en relief l'étendue de l'utilisation de sites d'initiation alternative de la traduction chez les eucaryotes, permettant l'expression de nouvelles protéines depuis des ORFs alternatifs. Les outils développés constituent une base solide afin de mener à un recensement le plus exhaustif possible des protéines alternatives exprimées chez plusieurs organismes eucaryotes. Ils s'intègrent dans l'effort vers l'objectif majeur de l'ère post-génomique qu'est l'établissement du protéome complet de multiples organismes.

La tâche principale restant à accomplir, afin de déterminer l'importance biologique du protéome alternatif, est de caractériser les fonctions portées par les produits protéiques d'AltORFs. Les résultats et outils présentés ici stimuleront la découverte et la caractérisation de protéines alternatives de plus en plus nombreuses. L'exemple du gène codant pour l'ATXN1 et la protéine alternative AltATXN1, étudié dans notre laboratoire,

en est une illustration (Bergeron et al, 2013). D'autres protéines alternatives candidates sont actuellement à l'étude dans notre laboratoire et dans d'autres, par le biais de collaborations.

La découverte que les AltORFs sont utilisés plus fréquemment qu'estimé précédemment amène à repenser la conception de la structure d'un gène eucaryote classique. L'expression de plusieurs groupes de protéoformes depuis un locus génomique unique, et possiblement depuis des transcrits uniques, a des implications directes sur l'étude de la fonction des gènes. L'étude fonctionnelle des protéines alternatives amènera à la compréhension de leur rôle dans les mécanismes cellulaires en conditions physiologiques et pathologiques.

REMERCIEMENTS

Je tiens à adresser un immense merci à mon directeur de recherche, le Dr Xavier Roucou, pour l'accueil qu'il m'a accordé au sein de son laboratoire. Il fut un excellent mentor, me poussant à donner le meilleur de moi-même, et participant à promouvoir dans le laboratoire de la collaboration, ainsi qu'une ambiance de travail saine et productive. Il m'a aussi donné les libertés nécessaires dans la gestion de mes projets de recherche, ce qui a été un facteur majeur dans mon éclosion scientifique. D'un point de vue plus personnel, je me souviendrai toujours de l'aide qu'il m'a apporté pour faciliter mon arrivée au Québec. Mes remerciements vont ensuite au reste des membres de mon laboratoire, passés et présents, pour l'ambiance de travail mémorable à laquelle ils ont participé. Un merci spécial à Guillaume Tremblay, assistant de recherche extraordinaire et enseignant de québécois personnel, à Alireza Roostae, compagnon sportif au labo, à Antanas Staskevicius, excellent collaborateur et boute-en-train, à Simon Beaudoin, Maxime Béland, Cyntia Bissonnette, Julie Motard pour les discussions, conseils, et aides en tout genre. Merci à Jean-François Lucier, acteur bioinformatique déterminant dans la réussite de mes projets. Merci aux professeur(e)s François Bachand, Eric Massé, Léa Brakier-Gingras, François-Michel Boisvert, Michel Salzet, et Michelle Scott pour leurs conseils et leur œil critique sur mes résultats. Merci aux membres du département de Biochimie d'avoir créé un environnement propice à l'excellence académique. Une grande reconnaissance est adressée aux membres de mon jury de thèse de doctorat, les professeurs Benoit Coulombe, Benoit Chabot, François Bachand et Xavier Roucou, pour avoir pris le temps de me lire et de m'écouter de façon critique et juste. Je remercie également les organismes subventionnaires (FQRNT, PrioNet Canada, IRSC) pour leur aide financière. Enfin, mes remerciements infinis vont à tous ceux qui m'ont soutenu moralement, et qui ont participé à mon bien-être quotidien au cours de ces quatre années : mes parents, mes frères et sœurs et leurs conjoints, leurs enfants, mes grands-parents, qui m'emplissent de bonheur à chaque retrouvailles. Un tendre merci à Mathilde Monnier, qui partage ma vie et mes escapades au laboratoire depuis trois ans. Une grande pensée à tous mes amis, qui ont joué à merveille le rôle de soupape de décompression et de partenaires d'évasion quand le besoin s'en ressentait.

LISTE DES RÉFÉRENCES

Abramowitz J, Grenet D, Birnbaumer M, Torres HN, & Birnbaumer L (2004) XLalphas, the extra-long form of the alpha-subunit of the Gs G protein, is significantly longer than suspected, and so is its companion Alex. *Proc Natl Acad Sci U S A* **101**: 8366-8371

Adam AC, Bornhovd C, Prokisch H, Neupert W, & Hell K (2006) The Nfs1 interacting protein Isd11 has an essential role in Fe/S cluster biogenesis in mitochondria. *EMBO J* **25**: 174-183

Akimoto C, Sakashita E, Kasashima K, Kuroiwa K, Tominaga K, Hamamoto T, & Endo H (2013) Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim Biophys Acta* **1830**: 2728-2738

Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, & Sorek R (2006) Transcription-mediated gene fusion in the human genome. *Genome Res* **16**: 30-36

Andrews J, Smith M, Merakovsky J, Coulson M, Hannan F, & Kelly LE (1996) The stoned locus of Drosophila melanogaster produces a dicistronic transcript and encodes two distinct polypeptides. *Genetics* **143**: 1699-1711

Ara T, Lopez F, Ritchie W, Benech P, & Gautheret D (2006) Conservation of alternative polyadenylation patterns in mammalian genes. *BMC Genomics* **7**: 189

Atkins,J.F., Gesteland,R.F. (2010) *Recoding: Expansion of Decoding Rules Enriches Gene Expression.* Springer: New York

Autio KJ, Kastaniotis AJ, Pospiech H, Miinalainen IJ, Schonauer MS, Dieckmann CL, & Hiltunen JK (2008) An ancient genetic link between vertebrate mitochondrial fatty acid synthesis and RNA processing. *FASEB J* **22**: 569-578

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, & Eichler EE (2002) Recent segmental duplications in the human genome. *Science* **297**: 1003-1007

Baril M & Brakier-Gingras L (2005) Translation of the F protein of hepatitis C virus is initiated at a non-AUG codon in a +1 reading frame relative to the polyprotein. *Nucleic Acids Res* **33**: 1474-1486

Basrai MA, Hieter P, & Boeke JD (1997) Small open reading frames: beautiful needles in the haystack. *Genome Res* **7**: 768-771

Bayes A & Grant SG (2009) Neuroproteomics: understanding the molecular organization and complexity of the brain. *Nat Rev Neurosci* **10**: 635-646

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, & Tromp MC (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**: 819-826

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59

Ben-Yehezkel T, Zur H, Marx T, Shapiro E, & Tuller T (2013) Mapping the translation initiation landscape of an S. cerevisiae gene using fluorescent proteins. *Genomics*

Bergeron D, Lapointe C, Bissonnette C, Tremblay G, Motard J, & Roucou X (2013) An Out-of-frame Overlapping Reading Frame in the Ataxin-1 Coding Sequence Encodes a Novel Ataxin-1 Interacting Protein. *J Biol Chem* **288**: 21824-21835

Berry MJ, Banu L, Harney JW, & Larsen PR (1993) Functional characterization of the eukaryotic SECIS elements which direct selenocysteine insertion at UGA codons. *EMBO J* **12**: 3315-3322

Blaschke RJ, Topfer C, Marchini A, Steinbeisser H, Janssen JW, & Rappold GA (2003) Transcriptional and translational regulation of the Leri-Weill and Turner syndrome homeobox gene SHOX. *J Biol Chem* **278**: 47820-47826

Blobel G (1980) Intracellular protein topogenesis. *Proc Natl Acad Sci U S A* **77**: 1496-1500

Blumenthal T (2004) Operons in eukaryotes. *Brief Funct Genomic Proteomic* **3**: 199-211

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, & Kim SK (2002) A global analysis of Caenorhabditis elegans operons. *Nature* **417**: 851-854

Bolinger C & Boris-Lawrie K (2009) Mechanisms employed by retroviruses to exploit host factors for translational control of a complicated proteome. *Retrovirology* **6**: 8-4690-6-8

Bonetti B, Fu L, Moon J, & Bedwell DM (1995) The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. *J Mol Biol* **251**: 334-345

Boni IV, Isaeva DM, Musychenko ML, & Tzareva NV (1991) Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. *Nucleic Acids Res* **19**: 155-162

Brent MR (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res* **15**: 1777-1786

Brett D, Pospisil H, Valcarcel J, Reich J, & Bork P (2002) Alternative splicing and genome complexity. *Nat Genet* **30**: 29-30

Brogna S & Ashburner M (1997) The Adh-related gene of Drosophila melanogaster is expressed as a functional dicistronic messenger RNA: multigenic transcription in higher organisms. *EMBO J* **16**: 2023-2031

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, & Dunker AK (2002) Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55**: 104-110

Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* **325**: 728-730

Cai J, Zhang J, Huang Y, & Li Y (2005) ATID: a web-oriented database for collection of publicly available alternative translational initiation events. *Bioinformatics* **21**: 4312-4314

Calligaris R, Bottardi S, Cogoi S, Apezteguia I, & Santoro C (1995) Alternative translation initiation site usage results in two functionally distinct forms of the GATA-1 transcription factor. *Proc Natl Acad Sci U S A* **92**: 11598-11602

Calvo SE, Pagliarini DJ, & Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* **106**: 7507-7512

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V et al (2005) The transcriptional landscape of the mammalian genome. *Science* **309**: 1559-1563

Caudevilla C, Serra D, Miliar A, Codony C, Asins G, Bach M, & Hegardt FG (1998) Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci U S A* **95**: 12185-12190

Chacinska A, Koehler CM, Milenkovic D, Lithgow T, & Pfanner N (2009) Importing mitochondrial proteins: machineries and mechanisms. *Cell* **138**: 628-644

Chang B, Halgamuge S, & Tang SL (2006) Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene* **373**: 90-99

Chappell SA, Dresios J, Edelman GM, & Mauro VP (2006) Ribosomal shunting mediated by a translational enhancer element that base pairs to 18S rRNA. *Proc Natl Acad Sci U S A* **103**: 9488-9493

Chung WY, Wadhawan S, Szklarczyk R, Pond SK, & Nekrutenko A (2007) A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput Biol* **3**: e91

Col B, Oltean S, & Banerjee R (2007) Translational regulation of human methionine synthase by upstream open reading frames. *Biochim Biophys Acta* **1769**: 532-540

Conant GC & Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**: 938-950

Courtois S, Verhaegh G, North S, Luciani MG, Lassus P, Hibner U, Oren M, & Hainaut P (2002) DeltaN-p53, a natural isoform of p53 lacking the first transactivation domain, counteracts growth suppression by wild-type p53. *Oncogene* **21**: 6722-6728

Crowe ML, Wang XQ, & Rothnagel JA (2006) Evidence for conservation and selection of upstream open reading frames suggests probable encoding of bioactive peptides. *BMC Genomics* **7**: 16

Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, & Brown CJ (2007) Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol Evol* **65**: 277-288

Davuluri RV, Suzuki Y, Sugano S, Plass C, & Huang TH (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**: 167-177

de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, & Mann M (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **455**: 1251-1254

Di Giammartino DC, Nishida K, & Manley JL (2011) Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853-866

Diba F, Watson CS, & Gametchu B (2001) 5'UTR sequences of the glucocorticoid receptor 1A transcript encode a peptide associated with translational regulation of the glucocorticoid receptor. *J Cell Biochem* **81**: 149-161

Doronina VA & Brown JD (2006) Non-canonical decoding events at stop codons in eukaryotes. *Mol Biol (Mosk)* **40**: 731-741

Dreher TW & Miller WA (2006) Translational control in positive strand RNA plant viruses. *Virology* **344**: 185-197

Dron M, Dandoy-Dron F, Farooq Salamat MK, & Laude H (2009) Proteasome inhibitors promote the sequestration of PrPSc into aggresomes within the cytosol of prion-infected CAD neuronal cells. *J Gen Virol* **90**: 2050-2060

Farabaugh PJ (1996) Programmed translational frameshifting. *Annu Rev Genet* **30**: 507-528

Filbin ME & Kieft JS (2011) HCV IRES domain IIb affects the configuration of coding RNA in the 40S subunit's decoding groove. *RNA* **17**: 1258-1273

Fixsen SM & Howard MT (2010) Processive selenocysteine incorporation during synthesis of eukaryotic selenoproteins. *J Mol Biol* **399**: 385-396

Flouriot G, Brand H, Seraphin B, & Gannon F (2002) Natural trans-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. *J Biol Chem* **277**: 26244-26251

Freson K, Jaeken J, Van Helvoirt M, de Zegher F, Wittevrongel C, Thys C, Hoylaerts MF, Vermylen J, & Van Geet C (2003) Functional polymorphisms in the paternally expressed XLalphas and its cofactor ALEX decrease their mutual interaction and enhance receptor-mediated cAMP formation. *Hum Mol Genet* **12**: 1121-1130

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, & Grimmond SM (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet* **2**: e52

Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, Schreiber S, Platzer M, Krawczak M, Hampe J, & Brosch M (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res* **22**: 2208-2218

Galindo MI, Pueyo JI, Fouix S, Bishop SA, & Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* **5**: e106

Gendron K, Charbonneau J, Dulude D, Heveker N, Ferbeyre G, & Brakier-Gingras L (2008) The presence of the TAR RNA structure alters the programmed -1 ribosomal frameshift efficiency of the human immunodeficiency virus type 1 (HIV-1) by modifying the rate of translation initiation. *Nucleic Acids Res* **36**: 30-40

Gerashchenko MV, Lobanov AV, & Gladyshev VN (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc Natl Acad Sci U S A* **109**: 17394-17399

Ghezzi D, Goffrini P, Uziel G, Horvath R, Klopstock T, Lochmuller H, D'Adamo P, Gasparini P, Strom TM, Prokisch H, Invernizzi F, Ferrero I, & Zeviani M (2009) SDHAF1, encoding a LYR complex-II specific assembly factor, is mutated in SDH-defective infantile leukoencephalopathy. *Nat Genet* **41**: 654-656

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, & Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245

Godson GN, Barrell BG, Staden R, & Fiddes JC (1978) Nucleotide sequence of bacteriophage G4 DNA. *Nature* **276**: 236-247

Goossens S, Janssens B, Vanpoucke G, De Rycke R, van Hengel J, & van Roy F (2007) Truncated isoform of mouse alphaT-catenin is testis-restricted in expression and function. *FASEB J* **21**: 647-655

Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet* **17**: 100-107

Gray TA, Saitoh S, & Nicholls RD (1999) An imprinted, mammalian bicistronic transcript encodes two independent proteins. *Proc Natl Acad Sci U S A* **96**: 5616-5621

Gregory TR (2002) Genome size and developmental complexity. *Genetica* **115**: 131-146

Guo B, Zhai D, Cabezas E, Welsh K, Nouraini S, Satterthwait AC, & Reed JC (2003) Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature* **423**: 456-461

Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, & Shiu SH (2010) sORF finder: a program package to identify small open reading frames with high coding potential. *Bioinformatics* **26**: 399-400

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

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C et al (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760-1774

Hedges SB, Blair JE, Venturi ML, & Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol Biol* **4**: 2

Heo HS, Lee S, Kim JM, Choi YJ, Chung HY, & Oh SJ (2010) tsORFdb: theoretical small open reading frames (ORFs) database and massProphet: peptide mass fingerprinting (PMF) tool for unknown small functional ORFs. *Biochem Biophys Res Commun* **397**: 120-126

Hering O, Brenneis M, Beer J, Suess B, & Soppa J (2009) A novel mechanism for translation initiation operates in haloarchaea. *Mol Microbiol* **71**: 1451-1463

Hernandez-Sanchez C, Mansilla A, de la Rosa EJ, Pollerberg GE, Martinez-Salas E, & de Pablo F (2003) Upstream AUGs in embryonic proinsulin mRNA control its low translation level. *EMBO J* **22**: 5582-5592

Hinnebusch AG (2006) eIF3: a versatile scaffold for translation initiation complexes. *Trends Biochem Sci* **31**: 553-562

Hinnebusch AG (2005) Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* **59**: 407-450

Ho O & Green WR (2006) Alternative translational products and cryptic T cell epitopes: expecting the unexpected. *J Immunol* **177**: 8283-8289

Hobbs EC, Fontaine F, Yin X, & Storz G (2011) An expanding universe of small proteins. *Curr Opin Microbiol* **14**: 167-173

Holland SK & Blake CC (1987) Proteins, exons and molecular evolution. *BioSystems* **20**: 181-206

Iacono M, Mignone F, & Pesole G (2005) uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349**: 97-105

Ingolia NT, Ghaemmaghami S, Newman JR, & Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223

Ingolia NT, Lareau LF, & Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802

Jacks T, Madhani HD, Masiarz FR, & Varmus HE (1988) Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. *Cell* **55**: 447-458

Jacob F (1977) Evolution and tinkering. *Science* **196**: 1161-1166

JACOB F, PERRIN D, SANCHEZ C, & MONOD J (1960) Operon: a group of genes with the expression coordinated by an operator. *C R Hebd Seances Acad Sci* **250**: 1727-1729

Jang SK, Krausslich HG, Nicklin MJ, Duke GM, Palmenberg AC, & Wimmer E (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* **62**: 2636-2643

Johnston JA, Ward CL, & Kopito RR (1998) Aggresomes: a cellular response to misfolded proteins. *J Cell Biol* **143**: 1883-1898

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, & Kellis M (2011) Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res* **21**: 2096-2113

Kaessmann H (2010) Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313-1326

Kaessmann H, Vinckenbosch N, & Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19-31

Kanamori Y, Hayakawa Y, Matsumoto H, Yasukochi Y, Shimura S, Nakahara Y, Kiuchi M, & Kamimura M (2010) A eukaryotic (insect) tricistronic mRNA encodes three proteins selected by context-dependent scanning. *J Biol Chem* **285**: 36933-36944

Karn J & Stoltzfus CM (2012) Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med* **2**: a006916

Karolchik D, Hinrichs AS, & Kent WJ (2007) The UCSC Genome Browser. *Curr Protoc Bioinformatics* **Chapter 1**: Unit 1.4

Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, & Basrai MA (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res* **16**: 365-373

Kazak L, Reyes A, Duncan AL, Rorbach J, Wood SR, Brea-Calvo G, Gammage PA, Robinson AJ, Minczuk M, & Holt IJ (2013) Alternative translation initiation augments the human mitochondrial proteome. *Nucleic Acids Res* **41**: 2354-2369

Keegan LP, Gallo A, & O'Connell MA (2001) The many roles of an RNA editor. *Nat Rev Genet* **2**: 869-878

Keene JD & Lager PJ (2005) Post-transcriptional operons and regulons co-ordinating gene expression. *Chromosome Res* **13**: 327-337

Kervestin S & Jacobson A (2012) NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* **13**: 700-712

Klemke M, Kehlenbach RH, & Huttner WB (2001) Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage. *EMBO J* **20**: 3849-3860

Kneller EL, Rakotondrafara AM, & Miller WA (2006) Cap-independent translation of plant viral RNAs. *Virus Res* **119**: 63-75

Knowles DG & McLysaght A (2009) Recent de novo origin of human protein-coding genes. *Genome Res* **19**: 1752-1759

Kochetov AV (2008) Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* **30**: 683-91

Kochetov AV (2006) [Alternative translation start sites and their significance for eukaryotic proteome]. *Mol Biol (Mosk)* **40**: 788-95

Kochetov AV, Sarai A, Rogozin IB, Shumny VK, & Kolchanov NA (2005) The role of alternative translation start sites in the generation of human protein diversity. *Mol Genet Genomics* **273**: 491-496

Konarska MM, Padgett RA, & Sharp PA (1985) Trans splicing of mRNA precursors in vitro. *Cell* **42**: 165-171

Kondo T, Hashimoto Y, Kato K, Inagaki S, Hayashi S, & Kageyama Y (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* **9**: 660-665

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

Koslowsky DJ, Bhat GJ, Perrollaz AL, Feagin JE, & Stuart K (1990) The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. *Cell* **62**: 901-911

Kovacs E, Tompa P, Liliom K, & Kalmar L (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proc Natl Acad Sci U S A* **107**: 5429-5434

Koybasi S, Senkal CE, Sundararaj K, Spassieva S, Bielawski J, Osta W, Day TA, Jiang JC, Jazwinski SM, Hannun YA, Obeid LM, & Ogretmen B (2004) Defects in cell growth regulation by C18:0-ceramide and longevity assurance gene 1 in human head and neck squamous cell carcinomas. *J Biol Chem* **279**: 44311-44319

Kozak M (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**: 13-37

Kozak M (2001) Constraints on reinitiation of translation in mammals. *Nucleic Acids Res* **29**: 5226-5232

Kozak M (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J* **16**: 2482-2492

Kozak M (1990) Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci U S A* **87**: 8301-8305

Kozak M (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Mol Cell Biol* **7**: 3438-3445

Kozak M (1986a) Bifunctional messenger RNAs in eukaryotes. *Cell* **47**: 481-483

Kozak M (1986b) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44**: 283-292

Kozak M (1986c) Regulation of protein synthesis in virus-infected animal cells. *Adv Virus Res* **31**: 229-292

Kozak M (1984) Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. *Nucleic Acids Res* **12**: 3873-3893

Kozak M (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucleic Acids Res* **9**: 5233-5252

Kozak M (1978) How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell* **15**: 1109-1123

Kristiansen M, Messenger MJ, Klohn PC, Brandner S, Wadsworth JD, Collinge J, & Tabrizi SJ (2005) Disease-related prion protein forms aggresomes in neuronal cells leading to caspase activation and apoptosis. *J Biol Chem* **280**: 38851-38861

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, & Couso JP (2011) Hundreds of putatively functional small open reading frames in Drosophila. *Genome Biol* **12**: R118-2011-12-11-r118

Law GL, Raney A, Heusner C, & Morris DR (2001) Polyamine regulation of ribosome pausing at the upstream open reading frame of S-adenosylmethionine decarboxylase. *J Biol Chem* **276**: 38036-38043

Le Y, Zhou Y, Iribarren P, & Wang J (2004) Chemokines and chemokine receptors: their manifold roles in homeostasis and disease. *Cell Mol Immunol* **1**: 95-104

Lee S, Liu B, Lee S, Huang SX, Shen B, & Qian SB (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A* **109**: E2424-32

Lee SJ (1991) Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc Natl Acad Sci U S A* **88**: 4250-4254

Lespinet O, Wolf YI, Koonin EV, & Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**: 1048-1059

Lewis SM, Cerquozzi S, Graber TE, Ungureanu NH, Andrews M, & Holcik M (2008) The eIF4G homolog DAP5/p97 supports the translation of select mRNAs during endoplasmic reticulum stress. *Nucleic Acids Res* **36**: 168-178

Li C, Goudy K, Hirsch M, Asokan A, Fan Y, Alexander J, Sun J, Monahan P, Seiber D, Sidney J, Sette A, Tisch R, Frelinger J, & Samulski RJ (2009) Cellular immune response to cryptic epitopes during therapeutic gene transfer. *Proc Natl Acad Sci U S A* **106**: 10770-10774

Liang H & Landweber LF (2006) A genome-wide study of dual coding regions in human alternatively spliced genes. *Genome Res* **16**: 190-196

Lin FT, MacDougald OA, Diehl AM, & Lane MD (1993) A 30-kDa alternative translation product of the CCAAT/enhancer binding protein alpha message: transcriptional activator lacking antimitotic activity. *Proc Natl Acad Sci U S A* **90**: 9606-9610

Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky L, Darnell J (eds) (2005) *Biologie moléculaire de la cellule.* De Boeck Université: Bruxelles

Lubec G & Afjehi-Sadat L (2007) Limitations and pitfalls in protein identification by mass spectrometry. *Chem Rev* **107**: 3568-3584

Luukkonen BG, Tan W, & Schwartz S (1995) 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**: 4086-4094

Lynch M & Conery JS (2003) The origins of genome complexity. *Science* **302**: 1401-1404

Ma J, Campbell A, & Karlin S (2002a) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184**: 5733-5745

Ma J, Wollmann R, & Lindquist S (2002b) Neurotoxicity and neurodegeneration when PrP accumulates in the cytosol. *Science* **298**: 1781-1785

Magny EG, Pueyo JI, Pearl FM, Cespedes MA, Niven JE, Bishop SA, & Couso JP (2013) Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* **341**: 1116-1120

Malys N & McCarthy JE (2011) Translation initiation: variations in the mechanism can be anticipated. *Cell Mol Life Sci* **68**: 991-1003

Mann M, Kulak NA, Nagaraj N, & Cox J (2013) The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell* **49**: 583-590

Mathe C, Sagot MF, Schiex T, & Rouze P (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res* **30**: 4103-4117

Matsuda D, Bauer L, Tinnesand K, & Dreher TW (2004) Expression of the two nested overlapping reading frames of turnip yellow mosaic virus RNA is enhanced by a 5' cap and by 5' and 3' viral sequences. *J Virol* **78**: 9325-9335

Matsui M, Yachie N, Okada Y, Saito R, & Tomita M (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS Lett* **581**: 4184-4188

Matsuoka M (2011) Humanin signal for Alzheimer's disease. *J Alzheimers Dis* **24 Suppl 2**: 27-32

Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet* **5**: 316-323

Menschaert G, Van Criekinge W, Notelaers T, Koch A, Crappe J, Gevaert K, & 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**: 1780-1790

Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, & Baranov PV (2012) Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res* **22**: 2219-2229

Miller WA, Wang Z, & Treder K (2007) The amazing diversity of cap-independent translation elements in the 3'-untranslated regions of plant viral RNAs. *Biochem Soc Trans* **35**: 1629-1633

Mira A, Ochman H, & Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**: 589-596

Mitchell SF, Walker SE, Algire MA, Park EH, Hinnebusch AG, & Lorsch JR (2010) The 5'-7-methylguanosine cap on eukaryotic mRNAs serves both to stimulate canonical translation initiation and to block an alternative pathway. *Mol Cell* **39**: 950-962

Mize GJ, Ruan H, Low JJ, & Morris DR (1998) The inhibitory upstream open reading frame from mammalian S-adenosylmethionine decarboxylase mRNA has a strict sequence specificity in critical positions. *J Biol Chem* **273**: 32500-32505

Moll I, Grill S, Gualerzi CO, & Blasi U (2002) Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. *Mol Microbiol* **43**: 239-246

Morse DP & Bass BL (1999) Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. *Proc Natl Acad Sci U S A* **96**: 6048-6053

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, & Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Paabo S, & Mann M (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* **7**: 548

Naik S, Thomas NS, Davies JH, Lever M, Raponi M, Baralle D, Temple IK, & Caliebe A (2012) Novel Tandem Duplication in Exon 1 of the SNURF/SNRPN Gene in a Child with Transient Excessive Eating Behaviour and Weight Gain. *Mol Syndromol* **2**: 76-80

Namy O, Hatin I, & Rousset JP (2001) Impact of the six nucleotides downstream of the stop codon on translation termination. *EMBO Rep* **2**: 787-793

Nekrutenko A, Wadhawan S, Goetting-Minesky P, & Makova KD (2005) Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. *PLoS Genet* **1**: e18

Nesvizhskii AI (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**: 2092-2123

Neu-Yilik G, Amthor B, Gehring NH, Bahri S, Paidassi H, Hentze MW, & Kulozik AE (2011) Mechanism of escape from nonsense-mediated mRNA decay of human beta-globin transcripts with nonsense mutations in the first exon. *RNA* **17**: 843-854

Normark S, Bergstrom S, Edlund T, Grundstrom T, Jaurin B, Lindberg FP, & Olsson O (1983) Overlapping genes. *Annu Rev Genet* **17**: 499-525

Nyiko T, Sonkoly B, Merai Z, Benkovics AH, & Silhavy D (2009) Plant upstream ORFs can trigger nonsense-mediated mRNA decay in a size-dependent manner. *Plant Mol Biol* **71**: 367-378

Oyama M, Itagaki C, Hata H, Suzuki Y, Izumi T, Natsume T, Isobe T, & Sugano S (2004) Analysis of small human proteins reveals the translation of upstream open reading frames of mRNAs. *Genome Res* **14**: 2048-2052

Oyama M, Kozuka-Hata H, Suzuki Y, Semba K, Yamamoto T, & Sugano S (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* **6**: 1000-1006

Pajares MJ, Ezponda T, Catena R, Calvo A, Pio R, & Montuenga LM (2007) Alternative splicing: an emerging topic in molecular and clinical oncology. *Lancet Oncol* **8**: 349-357

Pal S, Gupta R, Kim H, Wickramasinghe P, Baubet V, Showe LC, Dahmane N, & Davuluri RV (2011) Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development. *Genome Res* **21**: 1260-1272

Palaniswamy R, Teglund S, Lauth M, Zaphiropoulos PG, & Shimokawa T (2010) Genetic variations regulate alternative splicing in the 5' untranslated regions of the mouse glioma-associated oncogene 1, Gli1. *BMC Mol Biol* **11**: 32-2199-11-32

Palmer CS, Osellame LD, Laine D, Koutsopoulos OS, Frazier AE, & Ryan MT (2011) MiD49 and MiD51, new components of the mitochondrial fission machinery. *EMBO Rep* **12**: 565-573

Pan Q, Shai O, Lee LJ, Frey BJ, & Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413-1415

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, & Guigo R (2006) Tandem chimerism as a means to increase protein complexity in the human genome. *Genome Res* **16**: 37-44

Paulding CA, Ruvolo M, & Haber DA (2003) The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc Natl Acad Sci U S A* **100**: 2507-2511

Pavesi A, Magiorkinis G, & Karlin DG (2013) Viral proteins originated de novo by overprinting can be identified by codon usage: application to the "gene nursery" of deltaretroviruses. *PLoS Comput Biol* **9**: e1003162

Pedroso I, Rivera G, Lazo F, Chacon M, Ossandon F, Veloso FA, & Holmes DS (2008) AlterORF: a database of alternate open reading frames. *Nucleic Acids Res* **36**: D517-8

Pelletier J & Sonenberg N (1988) Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature* **334**: 320-325

Pendleton LC, Goodwin BL, Solomonson LP, & Eichler DC (2005) Regulation of endothelial argininosuccinate synthase expression and NO production by an upstream open reading frame. *J Biol Chem* **280**: 24252-24260

Perler FB (2002) InBase: the Intein Database. *Nucleic Acids Res* **30**: 383-384

Perler FB, Davis EO, Dean GE, Gimble FS, Jack WE, Neff N, Noren CJ, Thorner J, & Belfort M (1994) Protein splicing elements: inteins and exteins--a definition of terms and recommended nomenclature. *Nucleic Acids Res* **22**: 1125-1127

Perrin-Vidoz L, Sinilnikova OM, Stoppa-Lyonnet D, Lenoir GM, & Mazoyer S (2002) The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. *Hum Mol Genet* **11**: 2805-2814

Pertea M & Salzberg SL (2010) Between a chicken and a grape: estimating the number of human genes. *Genome Biol* **11**: 206-2010-11-5-206. Epub 2010 May 5

Ponka P, Beaumont C, & Richardson DR (1998) Function and regulation of transferrin and ferritin. *Semin Hematol* **35**: 35-54

Ponnala L (2010) A plausible role for the presence of internal shine-dalgarno sites. *Bioinform Biol Insights* **4**: 55-60

Poyry TA, Kaminski A, & Jackson RJ (2004) What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? *Genes Dev* **18**: 62-75

Pozner A, Goldenberg D, Negreanu V, Le SY, Elroy-Stein O, Levanon D, & Groner Y (2000) Transcription-coupled translation control of AML1/RUNX1 is mediated by cap- and internal ribosome entry site-dependent mechanisms. *Mol Cell Biol* **20**: 2297-2307

Pruitt KD, Tatusova T, & Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61-5

Puttaraju M, Jamison SF, Mansfield SG, Garcia-Blanco MA, & Mitchell LG (1999) Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. *Nat Biotechnol* **17**: 246-252

Rabadan-Diehl C, Martinez A, Volpi S, Subburaju S, & Aguilera G (2007) Inhibition of vasopressin V1b receptor translation by upstream open reading frames in the 5'-untranslated region. *J Neuroendocrinol* **19**: 309-319

Racine T & Duncan R (2010) Facilitated leaky scanning and atypical ribosome shunting direct downstream translation initiation on the tricistronic S1 mRNA of avian reovirus. *Nucleic Acids Res* **38**: 7260-7272

Rancurel C, Khosravi M, Dunker AK, Romero PR, & Karlin D (2009) Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. *J Virol* **83**: 10719-10736

Rhee HW, Zou P, Udeshi ND, Martell JD, Mootha VK, Carr SA, & Ting AY (2013) Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**: 1328-1331

Ribrioux S, Brungger A, Baumgarten B, Seuwen K, & John MR (2008) Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* **9**: 122

Rohrig H, John M, & Schmidt J (2004) Modification of soybean sucrose synthase by S-thiolation with ENOD40 peptide A. *Biochem Biophys Res Commun* **325**: 864-870

Rohrig H, Schmidt J, Miklashevichs E, Schell J, & John M (2002) Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci U S A* **99**: 1915-1920

Ron D (2002) Translational control in the endoplasmic reticulum stress response. *J Clin Invest* **110**: 1383-1388

Ronsin C, Chung-Scott V, Poullion I, Aknouche N, Gaudin C, & Triebel F (1999) A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. *J Immunol* **163**: 483-490

Rosenberg SA, Packard BS, Aebersold PM, Solomon D, Topalian SL, Toy ST, Simon P, Lotze MT, Yang JC, & Seipp CA (1988) Use of tumor-infiltrating lymphocytes and interleukin-2 in the immunotherapy of patients with metastatic melanoma. A preliminary report. *N Engl J Med* **319**: 1676-1680

Rosenberg SA, Tong-On P, Li Y, Riley JP, El-Gamil M, Parkhurst MR, & Robbins PF (2002) Identification of BING-4 cancer antigen translated from an alternative open reading frame of a gene in the extended MHC class II region using lymphocytes from a patient with a durable complete regression following immunotherapy. *J Immunol* **168**: 2402-2407

Rueter SM, Dawson TR, & Emeson RB (1999) Regulation of alternative splicing by RNA editing. *Nature* **399**: 75-80

Sabath N, Wagner A, & Karlin D (2012) Evolution of viral proteins originated de novo by overprinting. *Mol Biol Evol* **29**: 3767-3780

Samayoa J, Yildiz FH, & Karplus K (2011) Identification of prokaryotic small proteins using a comparative genomic approach. *Bioinformatics* **27**: 1765-1771

Sarrazin S, Starck J, Gonnet C, Doubeikovski A, Melet F, & Morle F (2000) Negative and translation termination-dependent positive control of FLI-1 protein synthesis by conserved overlapping 5' upstream open reading frames in Fli-1 mRNA. *Mol Cell Biol* **20**: 2959-2969

Savard J, Marques-Souza H, Aranda M, & Tautz D (2006) A segmentation gene in tribolium produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* **126**: 559-569

Scharff LB, Childs L, Walther D, & Bock R (2011) Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet* **7**: e1002155

Schluter H, Apweiler R, Holzhutter HG, & Jungblut PR (2009) Finding one's way in proteomics: a protein species nomenclature. *Chem Cent J* **3**: 11-153X-3-11

Seeburg PH (1996) The role of RNA editing in controlling glutamate receptor channel properties. *J Neurochem* **66**: 1-5

Seo J & Lee KJ (2004) Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. *J Biochem Mol Biol* **37**: 35-44

Shatkin AJ (1976) Capping of eucaryotic mRNAs. *Cell* **9**: 645-653

Shatsky IN, Dmitriev SE, Terenin IM, & Andreev DE (2010) Cap- and IRES-independent scanning mechanism of translation initiation as an alternative to the concept of cellular IRESs. *Mol Cells* **30**: 285-293

Shaw DC, Walker JE, Northrop FD, Barrell BG, Godson GN, & Fiddes JC (1978) Gene K, a new overlapping gene in bacteriophage G4. *Nature* **272**: 510-515

Shi Y, Ghosh MC, Tong WH, & Rouault TA (2009) Human ISD11 is essential for both iron-sulfur cluster assembly and maintenance of normal cellular iron homeostasis. *Hum Mol Genet* **18**: 3014-3025

Shine J & Dalgarno L (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71**: 1342-1346

Shuman S (2002) What messenger RNA capping tells us about eukaryotic evolution. *Nat Rev Mol Cell Biol* **3**: 619-625

Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, & Saghatelian A (2013) Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat Chem Biol* **9**: 59-64

Sloan J, Kinghorn JR, & Unkles SE (1999) The two subunits of human molybdopterin synthase: evidence for a bicistronic messenger RNA with overlapping reading frames. *Nucleic Acids Res* **27**: 854-858

Slone J, Daniels J, & Amrein H (2007) Sugar receptors in Drosophila. *Curr Biol* **17**: 1809-1816

Smith LM, Kelleher NL, & Consortium for Top Down Proteomics (2013) Proteoform: a single term describing protein complexity. *Nat Methods* **10**: 186-187

Solnick D (1985) Trans splicing of mRNA precursors. *Cell* **42**: 157-164

Somers J, Poyry T, & Willis AE (2013) A perspective on mammalian upstream open reading frame function. *Int J Biochem Cell Biol* **45**: 1690-1700

Sonenberg N & Hinnebusch AG (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**: 731-745

Sonenberg N & Hinnebusch AG (2007) New modes of translational control in development, behavior, and disease. *Mol Cell* **28**: 721-729

Spieth J, Brooke G, Kuersten S, Lea K, & Blumenthal T (1993) Operons in C. elegans: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell* **73**: 521-532

Stallmeyer B, Drugeon G, Reiss J, Haenni AL, & Mendel RR (1999) Human molybdopterin synthase gene: identification of a bicistronic transcript with overlapping reading frames. *Am J Hum Genet* **64**: 698-705

Tappe A & Kuner R (2006) Regulation of motor performance and striatal function by synaptic scaffolding proteins of the Homer1 family. *Proc Natl Acad Sci U S A* **103**: 774-779

Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, & Michnick SW (2008) An in vivo map of the yeast protein interactome. *Science* **320**: 1465-1470

Tattikota SG, Sury MD, Rathjen T, Wessels HH, Pandey AK, You X, Becker C, Chen W, Selbach M, & Poy MN (2013) Argonaute2 regulates the pancreatic beta-cell secretome. *Mol Cell Proteomics* **12**: 1214-1225

Tautz D (2009) Polycistronic peptide coding genes in eukaryotes--how widespread are they? *Brief Funct Genomic Proteomic* **8**: 68-74

Tay SK, Blythe J, & Lipovich L (2009) Global discovery of primate-specific genes in the human genome. *Proc Natl Acad Sci U S A* **106**: 12019-12024

Terenin IM, Andreev DE, Dmitriev SE, & Shatsky IN (2013) A novel mechanism of eukaryotic translation initiation that is neither m7G-cap-, nor IRES-dependent. *Nucleic Acids Res* **41**: 1807-1816

Tian B, Hu J, Zhang H, & Lutz CS (2005) A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**: 201-212

Tompa P, Szasz C, & Buday L (2005) Structural disorder throws new light on moonlighting. *Trends Biochem Sci* **30**: 484-489

Torabi N & Kruglyak L (2012) Genetic basis of hidden phenotypic variation revealed by increased translational readthrough in yeast. *PLoS Genet* **8**: e1002546

True HL & Lindquist SL (2000) A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**: 477-483

Tsang HT, Edwards TL, Wang X, Connell JW, Davies RJ, Durrington HJ, O'Kane CJ, Luzio JP, & Reid E (2009) The hereditary spastic paraplegia proteins NIPA1, spastin and spartin are inhibitors of mammalian BMP signalling. *Hum Mol Genet* **18**: 3805-3821

Tu C, Tzeng TH, & Bruenn JA (1992) Ribosomal movement impeded at a pseudoknot required for frameshifting. *Proc Natl Acad Sci U S A* **89**: 8636-8640

Van de Peer Y, Maere S, & Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* **10**: 725-732

Vassilaki N & Mavromara P (2009) The HCV ARFP/F/core+1 protein: production and functional analysis of an unconventional viral product. *IUBMB Life* **61**: 739-752

Vattem KM & Wek RC (2004) Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proc Natl Acad Sci U S A* **101**: 11269-11274

Ventoso I, Kochetov A, Montaner D, Dopazo J, & Santoyo J (2012) Extensive translatome remodeling during ER stress response in mammalian cells. *PLoS One* **7**: e35915

Vogel C & Chothia C (2006) Protein family expansions and biological complexity. *PLoS Comput Biol* **2**: e48

Vogel C, Teichmann SA, & Chothia C (2003) The immunoglobulin superfamily in Drosophila melanogaster and Caenorhabditis elegans and the evolution of complexity. *Development* **130**: 6317-6328

von der Haar T & Tuite MF (2007) Regulated translational bypass of stop codons in yeast. *Trends Microbiol* **15**: 78-86

Walewski JL, Keller TR, Stump DD, & Branch AD (2001) Evidence for a new hepatitis C virus antigen encoded in an overlapping reading frame. *RNA* **7**: 710-721

Wang RF, Johnston SL, Zeng G, Topalian SL, Schwartzentruber DJ, & Rosenberg SA (1998) A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames. *J Immunol* **161**: 3598-3606

Wang RF, Parkhurst MR, Kawakami Y, Robbins PF, & Rosenberg SA (1996) Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. *J Exp Med* **183**: 1131-1140

Wegrzyn JL, Drudge TM, Valafar F, & Hook V (2008) Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. *BMC Bioinformatics* **9**: 232-2105-9-232

Wiedemann N, Urzica E, Guiard B, Muller H, Lohaus C, Meyer HE, Ryan MT, Meisinger C, Muhlenhoff U, Lill R, & Pfanner N (2006) Essential role of Isd11 in mitochondrial iron-sulfur cluster synthesis on Isu scaffold proteins. *EMBO J* **25**: 184-195

Wu CT, Chiou CY, Chiu HC, & Yang UC (2013) Fine-tuning of microRNA-mediated repression of mRNA by splicing-regulated and highly repressive microRNA recognition element. *BMC Genomics* **14**: 438-2164-14-438

Wu H, Xu MQ, & Liu XQ (1998) Protein trans-splicing and functional mini-inteins of a cyanobacterial dnaB intein. *Biochim Biophys Acta* **1387**: 422-432

Xu H, Wang P, Fu Y, Zheng Y, Tang Q, Si L, You J, Zhang Z, Zhu Y, Zhou L, Wei Z, Lin B, Hu L, & Kong X (2010) Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell Res* **20**: 445-57

Yang Y & Walsh CE (2005) Spliceosome-mediated RNA trans-splicing. *Mol Ther* **12**: 1006-1012

Yueh A & Schneider RJ (2000) Translation by ribosome shunting on adenovirus and hsp70 mRNAs facilitated by complementarity to 18S rRNA. *Genes Dev* **14**: 414-421

Zetoune AB, Fontaniere S, Magnin D, Anczukow O, Buisson M, Zhang CX, & Mazoyer S (2008) Comparison of nonsense-mediated mRNA decay efficiency in various murine tissues. *BMC Genet* **9**: 83-2156-9-83

Zhang L, Kasif S, Cantor CR, & Broude NE (2004) GC/AT-content spikes as genomic punctuation marks. *Proc Natl Acad Sci U S A* **101**: 16855-16860

Zhang X, Virtanen A, & Kleiman FE (2010) To polyadenylate or to deadenylate: that is the question. *Cell Cycle* **9**: 4437-4449

Zhao J, Liu T, Jin S, Wang X, Qu M, Uhlen P, Tomilin N, Shupliakov O, Lendahl U, & Nister M (2011) Human MIEF1 recruits Drp1 to mitochondrial outer membranes and promotes mitochondrial fusion rather than fission. *EMBO J* **30**: 2762-2778

Zhu F, Liu Z, Chi X, & Qu H (2010) Protein trans-splicing based dual-vector delivery of the coagulation factor VIII gene. *Sci China Life Sci* **53**: 683-689

Zougman A, Mann M, & Wisniewski JR (2011) Identification and characterization of a novel ubiquitous nucleolar protein 'NARR' encoded by a gene overlapping the rab34 oncogene. *Nucleic Acids Res* **39**: 7103-7113

ANNEXES

Échanges de lettres avec le Dr Peter Wills (Auckland University, Nouvelle-Zélande)

Alternative Prion Proteins

Auteur de la lettre à l'éditeur : Peter R. Wills

Affiliations : Integrative Transcriptomics, Center for Bioinformatics Tübingen, University of Tübingen, Tübingen, Allemagne. Adresse de correspondance : Department of Physics, The University of Auckland, PB 92019, Auckland 1142, New Zealand. E-mail: p.wills@auckland.ac.nz

Statut de la lettre à l'éditeur : publié dans *FASEB Journal* ; 26(8):3100-3101, 2012

Avant-propos: Cette lettre à l'éditeur à été écrite par le Dr Peter R. Wills en réaction à la publication de l'article 1 du présent manuscrit.

I read with interest the article entitled, "An overlapping reading frame in the *PRNP* gene encodes a novel polypeptide distinct from the prion protein," by Vanderperre et al. (1). The paper describes the discovery of a polypeptide that is synthesized when translation is initiated at an AUG codon in the +3 reading frame downstream from the normal +1 open-reading frame (ORF) initiation site of the gene, *PRNP*, which encodes the prion protein (PrP), the main component of the etiological agent of transmissible spongiform encephalopathy (TSE). This alternative PrP (AltPrP) spans a region of *PRNP* that encodes a series of octapeptide repeats in the normal ORF, and it is comprised of 60–80 aa in most species known to be susceptible to TSE, the exact length depending on the number of encoded octa-repeats and the location of the next downstream stop codon in the +3 frame. AltPrP is coexpressed with the normal PrP, at least in humans, cattle, sheep, and deer, and is localized in the outer mitochondrial membrane. The possibility that AltPrP might participate in the pathogenesis of TSEs is discussed in the paper (1), but there is no mention of prior consideration given to the role of PrP molecules containing AltPrP sequence

elements as the main etiological agents of these diseases (2; 3). What may be the most significant feature of the new findings has been overlooked.

Although still the subject of occasional dispute, the infectious agent active in TSEs appears to be a variant of PrP. The difference between the normal cellular form of the PrP (PrP^{C}) and the "scrapie associated" form (PrP^{Sc}) is widely thought to come down to nothing more than differences in the way the amino acid chain is folded—PrP^{Sc} replicates by inducing molecules of PrP^{C} to refold into the PrP^{Sc} form. However, there has never been an investigation precise enough to exclude the possibility that a small minority of prion molecules containing frameshifted segments is an obligatory component of the etiological agents of most TSEs, serving as nuclei for the aggregation of normal-sequence, misfolded PrP that merely obscures the presence of full-length PrPs containing AltPrP elements.

Stochastic variation occurs in all molecular biological processes, and ribosomal framekeeping is no exception. Frameshifting errors occur at an average rate of $\sim 10^{-5}$/codon assignment. Thus, it is to be expected that during synthesis of AltPrP, the normal reading frame will occasionally be restored, before the stop codon in the +3 frame is encountered, resulting in the production of hybrid AltPrP–PrP molecules. The C-terminal bulk of these molecules will be identical to the corresponding part of normal PrP, and the hybrids will be only 8 aa shorter than normal PrP, which has a 22-aa N-terminal signal peptide removed. These hybrid molecules would be expected to copurify with PrP^{Sc} when it is extracted from cells, and they would form an occult component of infectious preparations, including those derived from recombinant microorganisms through "protein misfolding cyclic amplification" (PMCA). This procedure is generally very efficient as far as the refolding of PrP^{C} into PrP^{Sc} is concerned but often fails to produce high specific infectivity (4), suggesting that the production of infectivity may arise from the selective concentration of small numbers of lower solubility species of AltPrP sequence-containing prion molecules with highly hydrophobic N-termini.

The orthodox "induced misfolding model" of prion replication is encountering difficulties, as it cannot define the TSE disease threshold in terms of the amount and form of PrP^{Sc} needed for infectivity. There are cases in which misfolding of PrP is transmissible but

innocuous and others in which disease is transmitted, but misfolded protein cannot be detected. On the other hand, hybrid AltPrP–PrP molecules could only account for infectivity if they were capable of stimulating, above some critical threshold, the abnormal processes that led to their production during cellular synthesis of PrP. The primary process to be stimulated would be translation in the +3 frame of the N-terminal region of the *PRNP* gene, either through initiation at the appropriate AUG codon (1) or −1 ribosomal frameshifting at the nearby site just upstream from a conserved pseudoknot in the mRNA (3; 5). The extremely rare, condensed tryptophan triplet elements in the resulting AltPrP peptides could easily be effectors of either ribosomal process. Once amplified translation in the +3 frame was established, a compensating downstream +1 frameshift could occur simply as a result of ribosome slippage at a UGG "hungry codon" (6). Local depletion of Trp-tRNA is very likely during amplified synthesis in the +3 frame of the octa-repeat region of *PRNP*, as it contains 32.5% of UGG codons compared with a proteome-wide average of only 1.3%. Thus, hybrid AltPrP–PrP could be self-replicating through a combination of −1 and +1 frameshifting processes.

Minimal units of the most infectious preparations of the scrapie agent, in the form of brain homogenate, contain some 10^6 molecules of protease-resistant PrP^{Sc}, dispersed as soluble aggregates consisting of 14–28 molecules of PrP^{Sc} (7). It is not clear why enormous numbers of soluble aggregates are needed to infect a brain when aggregates of this size represent the most efficient means of propagating infectivity, unless of course most aggregates do not contain the etiological agent. This would be the case if infectious preparations contained a small proportion, perhaps up to 100 ppm, of truly active AltPrP sequence-containing molecules of PrP. Normal brain material and recombinant microorganisms that synthesize PrP would be expected to contain only a tiny proportion of the active species (~0.01 ppm; one molecule/10 cells of healthy brain material), produced as a result of normal but infrequent errors in ribosomal framekeeping and retrievable through selective precipitation during the PMCA process.

The hypothesis that the etiological agents active in TSEs are sequence variants of the PrP^{C} was first proposed one-quarter century ago (8). Now that AltPrP has been discovered (1), and the orthodox model of prion replication cannot produce a quantitatively consistent

interpretation of the available data, searching for odd protein molecules with AltPrP sequence elements in infectious preparations has renewed purpose.

Footnotes

The opinions expressed in editorials, essays, letters to the editor, and other articles comprising the Up Front section are those of the authors and do not necessarily reflect the opinions of FASEB or its constituent societies. *The FASEB Journal* welcomes all points of view and many voices. We look forward to hearing these in the form of op-ed pieces and/or letters from its readers addressed to journals@faseb.org.

References

1. Vanderperre, B., Staskevicius, A. B., Tremblay, G., McCoy, M., O'Neill, M. A., Cashman, N. R., and Roucou, X. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*25, 2373-2386

2. Wills, P. R. (1989) Induced frameshifting mechanism of replication for an information-carrying scrapie prion. *Microbial Pathogenesis*6, 235-249

3. Wills, P. R. (1992) Potential pseudoknots in the PrP-encoding mRNA. *Journal of Theoretical Biology*159, 523-527

4. Klingeborn, M., Race, B., Meade-White, K. D., and Chesebro, B. (2011) Lower specific infectivity of protease-resistant prion protein generated in cell-free reactions. *Proceedings of the National Academy of Sciences of the United States of America*108, E1244-53

5. Barrette, I., Poisson, G., Gendron, P., and Major, F. (2001) Pseudoknots in prion protein mRNAs confirmed by comparative sequence analysis and pattern searching. *Nucleic Acids Research*29, 753-758

6. Gallant, J. A., and Lindsley, D. (1998) Ribosomes can slide over and beyond "hungry" codons, resuming protein chain elongation many nucleotides downstream. *Proceedings of the National Academy of Sciences of the United States of America*95, 13771-13776

7. Silveira, J. R., Raymond, G. J., Hughson, A. G., Race, R. E., Sim, V. L., Hayes, S. F., and Caughey, B. (2005) The most infectious prion protein particles. *Nature*437, 257-261

8. Wills, P. R. (1986) Scrapie, ribosomal proteins and biological information. *Journal of Theoretical Biology*122, 157-178

Response

Auteur de la lettre à l'éditeur : Benoît Vanderperre, Xavier Roucou

Affiliations : Département de biochimie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada. Adresse de correspondance : Dr Xavier Roucou, Département de Biochimie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, 3201 Jean Mignault, Sherbrooke, Québec J1E 4K8, Canada, Tel. (819) 821 8000x72240 ; Fax. (819) 820 6831; E-mail: xavier.roucou@usherbrooke.ca

Statut de la lettre à l'éditeur : publié dans *FASEB Journal* ; 26(8):3100-3101, 2012

Avant-propos: Cette lettre à l'éditeur à été écrite en réponse à celle du Dr P. Wills.

Two observations led us to test the expression of alternative prion protein (AltPrP) (1). First, many functions were attributed to PrP, a protein encoded in the *PRNP* gene (2; 3). How one gene could encode one protein with such a variety of function is puzzling. Second, we noticed a putative overlapping reading frame in the +3 reading frame of *PRNP*. This coding sequence overlaps the octapeptide repeat region, a domain well conserved across species. Based on these two considerations only, a strategy was used that undoubtedly established that *PRNP* encodes two proteins: PrP and AltPrP (1). AltPrP is a tryptophan-rich mitochondrial protein, and its physiological function or its role in prion diseases is still unknown.

Previously, a hypothetical-induced frameshifting mechanism was proposed to explain the replication of prions, and the following model was suggested (4). First, a frameshift followed by a compensating frameshift during the translation of PrP results in the production of hybrid PrP molecules with sequence elements of AltPrP. Second, hybrid

AltPrP–PrP molecules are capable of stimulating these frameshifting errors. Third, hybrid AltPrP–PrP molecules are the infectious agent and provide nuclei for PrP aggregation.

In our experiments, we have had no evidence that AltPrP synthesis involves ribosomal frameshifting. Furthermore, we could not detect hybrid AltPrP–PrP molecules by Western blot, and no experimental data indicated the possibility of frameshifting errors during PrP translation. However, levels of hybrid molecules may be too low to be detected by Western blot. Alternatively, the expression of hybrid molecules may require specific experimental conditions.

As noted by Dr Wills, some unresolved gaps persist in the current model of prion replication, despite major breakthroughs in recent years. Nonetheless, the existence and the role for hybrid AltPrP–PrP molecules in prion disease remain speculative. Sensitive proteomic techniques on infectious prion particles may help address this issue. Another strategy to challenge the frameshifting model would be to test if expression of artificial hybrid AltPrP–PrP molecules in cultured cells or in transgenic animals results in spontaneous production of infectious prions.

References

1. Vanderperre, B., Staskevicius, A. B., Tremblay, G., McCoy, M., O'Neill, M. A., Cashman, N. R., and Roucou, X. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*25, 2373-2386

2. Linden, R., Martins, V. R., Prado, M. A., Cammarota, M., Izquierdo, I., and Brentani, R. R. (2008) Physiology of the prion protein. *Physiological Reviews*88, 673-728

3. Westergard, L., Christensen, H. M., and Harris, D. A. (2007) The cellular prion protein (PrP(C)): Its physiological function and role in disease. *Biochimica Et Biophysica Acta*1772, 629-644

4. Wills, P. R. (1989) Induced frameshifting mechanism of replication for an information-carrying scrapie prion. *Microbial Pathogenesis*6, 235-249