Étude d'éléments cis-régulateurs de l'initiation de la traduction eucaryote par des méthodes in vitro et par des méthodes de criblage à haut-débit

Antonin Tidu

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE Sciences de la Vie et de la Santé

Architecture et Réactivité de l'ARN : UPR – 9002 CNRS

Institut de Biologie Moléculaire et Cellulaire (IBMC), Strasbourg

THÈSE

présentée par :

Antonin TIDU

Soutenue le : **20 janvier 2023**

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

Spécialité : Aspects moléculaires et cellulaires de la biologie

Etude d'éléments *cis*-régulateurs de l'initiation de la traduction eucaryote par des approches *in vitro* et par des méthodes de criblage à haut-débit

THÈSE dirigée par :

Dr. MARTIN Franck — Directeur de Recherche CNRS

RAPPORTEURS :

Dr. MILLEVOI Stefania — Directrice de Recherche INSERM

Dr. PRATS Anne-Catherine — Directrice de Recherche INSERM

EXAMINATEURS :

Dr. MARQUET Roland — Directeur de Recherche CNRS

Pr. Dr. LEIDEL Sebastian — Professeur à l'Université de Berne (Suisse)

REMERCIEMENTS

Je voudrais d'abord remercier les membres du jury, Dr. Stefania Millevoi, Dr. Anne-Catherine Prats, Dr. Roland Marquet et Pr. Dr. Sebastian Leidel, d'avoir accepté et pris le temps d'évaluer mon travail de thèse.

Ma profonde gratitude s'adresse naturellement aux personnes que j'ai rencontrées au cours de ce doctorat. Leurs conseils, leur expertise, leur soutien, leur enthousiasme et leur maîtrise du tire-bouchon ont façonné ces années de thèse aussi bien au laboratoire qu'en dehors :

Franck, Gilbert, Christine et Laure, qui m'ont introduit aux us et coutumes du 333, que ce soit par la pipette ou par la boisson,

Aurélie, avec qui j'espère toujours détenir le prix de la plus belle addition à East Canteen, puissent mes dires te hanter pour l'éternité,

Fatima, Hassan, Aurélie et Justine, grâce à qui mon arrivée au 333 a été des plus agréables,

Piotr, Paulo, Claire, Antoine, Yann et Alexandre, aux premières loges du spectacle de la fissure, et toujours à l'affût d'un traquenard,

Sarah, Manon, Onaïs, Léa, Zina, Solène, Myriam, et Jeanne, dont les passages respectifs au 333, bien que furtifs, sont empreints de bons souvenirs,

Natacha, Michaël, Dominique et Karen, pour la confiance qu'ils m'ont accordée dans le cadre de nos collaborations respectives,

Laurence, Sandrine, Philippe, Lauriane et Johana pour leur aide et leur disponibilité,

Dr. Sébastien Pfeffer et Pr. Bruno Chatton, qui ont accepté de faire partie de mon comité de suivi de thèse,

Dr. Yves Nominé, qui a eu la gentillesse de me recevoir pour me conseiller sur l'analyse de données,

Toutes les personnes de l'institut avec qui j'ai pu échanger, même très brièvement.

Un dernier mot pour Franck, mon directeur de thèse, à qui je dois ces quatre années et demie passées dans l'équipe du 333. Tu as toujours su me guider de l'aval vers l'amont de la rivière doctorale, tout en étant suffisamment brave pour laisser libre cours à mes idées. Merci également de m'avoir fait confiance pour m'impliquer dans tes autres projets de recherche, qui, je l'espère, conduiront à de belles découvertes.

Table des matières

LISTE DES ABBREVIATIONS

aa-ARNt	ARN de transfert aminoacylé	Aminoacylated transfer RNA
ADN	Acide désoxyribonucléique	Desoxyribonucleic acid
ADP	Adénosine diphosphate	Adenosine diphosphate
ARN	Acide ribonucléique	Ribonucleic acid
ARNm	ARN messager	Messenger RNA
ARNpm	ARN pré-messager	Pre-messenger RNA
ARNr	ARN ribosomique	Ribosomal RNA
ARNt	ARN de transfert	Transfer RNA
ATP	Adénosine triphosphate	Adenosine triphosphate
BPS	Séquence de branchement	Branch point sequence
CoV	Coronavirus	Coronavirus
CrPV	Virus de la paralysie du cricket	Cricket paralysis virus
CTP	Cytidine triphosphate	Cytidine triphosphate
eEF	Facteur d'élongation eucaryotique	Eucaryotic elongation factor
eIF	Facteur d'initiation eucaryotique	Eucaryotic initiation factor
EMCV	Virus de l'encéphalo-myocardite	Encephalomyocarditis virus
eRF	Facteur de terminaison eucaryotique	Eucaryotic release factor
FACS	Tri de cellules activé par fluorescence	Fluorescence activated cell sorting
FMDV	Virus de la fièvre aphteuse	Foot and mouth disease virus
GAP	Protéine activatrice de GTPase	GTPase-activation protein
GCN2	Kinase de contrôle général non-dérépressible	General control nonderepressible 2
GDP	Guasosine diphosphate	Guasosine diphosphate
GEF	Facteur échangeur de guanine	Guanine exchange factor
GFP	Protéine fluorescence verte	Green fluorescent protein
GTP	Guasosine triphosphate	Guasosine triphosphate
HCV	Virus de l'hépatite C	Hepatitis C Virus
HIV	Virus de l'immunodéficience humaine	Human immunodeficiency virus
HRI	Inhibiteur régulé par l'hème	Heme-regulated inhibitor
IGR	Région intergénique	Intergenic region
IRE	Elément de réponse au fer	Iron responsive element
IRES	Site d'entrée interne du ribosome	Internal ribosome entry site
ISR	Réponse intégrée au stress	Integrated stress response
ITAF	Facteurs *trans* associés aux IRES	IRES trans-acting factor
miARN	microARN	microRNA
mTOR	Cible de la rapamycine chez les mammifères	Mammalian target of rapamycin
NSP	Protéine non structurale	Non-structural protein
ORF	Séquence ou phase codante ; cadre de lecture ouvert	Open Reading Frame
uORF	ORF située en amont de la phase codante principale	Upstream open reading frame
dORF	ORF située en aval de la phase codante principale	Downstream open reading frame
PAPB	Protéine de fixation à la queue poly-A	Poly-A binding protein
PERK	Protéine kinase du RE similaire à la protéine kinase R	Protein kinase R-like endoplasmic reticulum kinase
pH	Potentiel hydrogène	Hydrogen potential
PKR	Protéine kinase R	Protein kinase R
PV	Virus de la polio	Poliovirus
RE	Réticulum endoplasmique	Endoplasmic reticulum
RISC	Complexe induisant l'invalidation des ARN	RNA induced silencing complex
RRL	Lysats de réticulocytes de lapin	Rabbit reticulocytes lysates
SARS	Syndrome de l'insuffisance respiratoire aigüe sévère	Severe acute respiratory syndrome
SL	Tige-boucle	Stem loop
SRP	Particule de reconnaissance du signal	Signal recognition particle
START	Traduction des ARN assistée par une structure	Structure-assisted RNA translation
TAS	Récepteur du goût	Taste receptor
TIRF	Fluorescence par réflexion totale interne	Total internal reflection fluorescence
UPR	Réponse aux protéines mal repliées	Unfolded protein response
UTP	Uridine triphosphate	Uridine triphosphate
UTR	Région non-traduite	Untranslated region
VCE	Enzyme de capping du virus de la vaccine	Vaccinia capping enzyme

INTRODUCTION GENERALE

1. Introduction générale

1.1. Vue d'ensemble des rôles biologiques des protéines

Les protéines remplissent l'ensemble des fonctions biologiques des organismes uni- et pluri-cellulaires. Tous les mécanismes de régulation de ces fonctions qui ont été sélectionnés au cours de l'évolution permettent d'adapter le taux de synthèse des protéines et par conséquent leur(s) activité(s) biologiques en fonction de l'environnement.

1.1.1. A l'échelle d'un organisme

Les exemples ci-dessous illustrent certains rôles biologiques des protéines à l'échelle des organismes, principalement pluricellulaires.

1.1.1.1. Certaines protéines sont des senseurs de l'environnement d'un organisme

Chez les animaux, ces protéines sont situées à la surface de cellules spécifiques qui sont en relation directe avec le milieu extérieur, et sont le plus souvent transmembranaires.

La perception des sens chez l'homme requiert la coopération de protéines impliquées dans les voies de signalisation cellulaires pour assurer la transmission d'un signal sensoriel jusqu'au cerveau. Par exemple, les opsines sont les protéines photoréceptrices essentielles aux cellules de la vision, les cônes et les bâtonnets, qui sont situées dans la rétine (Hussey *et al.* 2022). La perception du goût implique les nombreuses protéines TAS (Taste Receptors) et la transduction du signal se fait par des protéines couplées aux protéines G et par des canaux ioniques (Roper and Chaudhari 2017). La perception de l'odorat fait appel à des protéines réceptrices qui sont toutes couplées aux protéines G (Lee *et al.* 2019), tandis que celle du toucher implique principalement des canaux ioniques (thermorécepteurs, mécanorécepteurs, etc.) composés de protéines situées à la surface des membranes des cellules spécialisées. Dans tous les cas, ces protéines réceptrices convertissent un stimulus extérieur en un signal moléculaire transmis vers le cerveau. Cette conversion passe par l'activation de voies de signalisation cellulaires qui mène à l'activation de canaux ioniques permettant de générer un potentiel d'action qui sera ensuite véhiculé jusqu'au cerveau où il sera interprété par l'intermédiaire d'autres protéines effectrices.

La réponse de l'organisme à une infection par un pathogène dépend des immunoglobulines présentes dans les membranes des cellules immunitaires. La reconnaissance d'un antigène induit une réponse précoce qui se poursuit par une cascade d'évènements constituant la réponse immunitaire adaptative, les deux reposant sur une signalisation moléculaire largement coordonnée par des protéines.

1.1.1.2. Les protéines circulantes assurent le maintien et la coordination des fonctions biologiques chez les organismes pluricellulaires

Les protéines circulantes assurent le transport d'informations, de métabolites et de nutriments entre les organes. L'hémoglobine assure par exemple le transport du dioxygène dans l'organisme : l'acheminement du dioxygène aux cellules est essentiel à la production d'ATP par la chaine respiratoire mitochondriale et donc à leur potentiel énergétique. Les anticorps circulants et les cytokines sont des acteurs centraux de la réponse immunitaire adaptative. L'hydrosolubilité des lipoprotéines, qui assurent le transport de lipides dans l'organisme et notamment le cholestérol, est due aux protéines qui les constituent. Certaines hormones,

produites par le système endocrinien, sont de nature protéique et sont essentielles à la coordination des fonctions biologiques des organismes en agissant sur l'activation des cellules cibles. On peut citer l'insuline, sécrétée par le pancréas, qui est essentielle pour la modulation du métabolisme du glucose des cellules musculaires, hépatiques et adipeuses.

1.1.1.3. Les protéines des matrices extracellulaires sont essentielles au maintien de la structure globale d'un organisme

L'organisation interne des organismes pluricellulaires est assurée par les matrices extracellulaires qui sont principalement constituées de protéines comme des collagènes, des protéoglycanes, des glycoprotéines ou des élastines.

La lame basale est une matrice extracellulaire synthétisée par les cellules épithéliales et assure leur organisation en épithélium. Les cellules évoluent dans la matrice extracellulaire qu'elles fabriquent. Les intégrines sont des protéines transmembranaires localisées à l'interface entre la matrice extracellulaire et la cellule épithéliale et sont essentielles à l'adhésion, à la mobilité, à la différentiation et à la prolifération cellulaires.

L'exemple le plus remarquable de matrice extracellulaire est incarné par la structure osseuse qui est synthétisée dans un premier temps par les ostéoblastes puis par les ostéocytes. La partie organique, interne, des os est synthétisée par les ostéoblastes et est principalement constituée de collagène. La matrice extracellulaire ainsi formée servira de matrice pour la fixation du calcium et du phosphate inorganique pour former l'hydroxyapatite, principal constituant minéral des os (Schlesinger *et al.* 2020).

1.1.2. A l'échelle cellulaire

Les cellules sont les unités fonctionnelles du vivant. Chaque cellule produit un ensemble de protéines (protéome) qui diffère d'un type cellulaire à l'autre selon le profil d'expression de l'ensemble de ses gènes (génome). En plus de ses fonctions, le protéome d'une cellule détermine aussi son état physiologique dans la mesure où il est le reflet de mécanismes moléculaires régulant la différenciation et le cycle cellulaires, ou encore l'apoptose.

1.1.2.1. Les protéines transmembranaires participent au fonctionnement cellulaire

Les protéines transmembranaires constituent le lien entre la cellule et son milieu extérieur et permettent l'échange d'informations codées de manière moléculaire, soit par contact direct avec une cellule voisine, soit par molécules effectrices sécrétées. Le même principe s'applique pour les organites au sein d'une cellule, leur milieu extérieur étant dans ce cas le cytosol.

<u>Rôles des protéines transmembranaires dans le maintien du potentiel de membrane</u>

Les protéines de la membrane plasmique comme les canaux ioniques et les pompes à ions sont impliquées dans la perméabilité membranaire et de ce fait dans le maintien du potentiel de membrane. Ce potentiel est crucial pour le transport de molécules intra- et extra- cellulaires. A ce titre, on peut citer les pompes à Na^+/K^+ ATP-dépendantes (Skou 1957; Nguyen *et al.* 2022), qui, en maintenant simultanément un flux global entrant de potassium et un flux global sortant de sodium, sont responsables de concentrations cytosoliques de 5-25 mM sodium et de 100-160 mM de potassium (Sejersted and Sjøgaard 2000; McKenna *et al.* 2008; Gast *et al.* 2021). Le maintien du potentiel de membrane est un paramètre critique pour, entre autres, la transmission de potentiels d'action et donc pour la communication nerveuse, ainsi que pour la

contraction musculaire. Dans les expériences de traduction *in vitro* réalisées dans le cadre de ce travail de thèse, nous avons choisi d'utiliser des concentrations en ions K^+ voisines de ces concentrations cellulaires.

Rôles des protéines transmembranaires dans la signalisation cellulaire

Les protéines transmembranaires sont impliquées dans les voies de signalisation cellulaires en réponse à un stimulus extracellulaire (hormone(s), ligand(s), métabolite(s), photons, etc.) et permettent la transduction d'un signal qui régule les fonctions de base de la cellule.

Dans le cas particulier d'une infection virale, certaines protéines transmembranaires sont nécessaires pour l'invagination de la particule virale dans la membrane plasmique, ce qui aboutit à l'injection du génome viral dans la cellule désormais infectée.

Rôles des protéines transmembranaires dans la maturation des protéines

Les protéines transmembranaires du réticulum endoplasmique et de l'appareil de Golgi sont impliquées dans la maturation des protéines membranaires et des protéines sécrétées, en étant notamment essentielles à la formation de complexes moléculaires nommés cargos qui permettent le transport de ces protéines vers leur localisation fonctionnelle, voire à la formation de vésicules d'exocytose dans le cas des protéines sécrétées.

Rôles des protéines transmembranaires dans le métabolisme cellulaire

L'énergie apportée par l'hydrolyse de l'ATP en ADP est exploitée par la cellule dans un grand nombre de réactions chimiques, ce qui fait de l'ATP une molécule essentielle au fonctionnement cellulaire. La principale source de synthèse de l'ATP dans la cellule est l'ATP synthase de la mitochondrie. La chaine respiratoire mitochondriale est principalement constituée de protéines transmembranaires situées dans la membrane interne de la mitochondrie. Ces protéines assurent le transport d'électrons jusqu'à la réduction du dioxygène en eau et c'est le pompage de protons vers l'espace intermembranaire qui permet le maintien du gradient de pH essentiel à la synthèse d'ATP par l'ATP synthase.

1.1.2.2. Les protéines du cytosquelette maintiennent l'architecture cellulaire

Les protéines du cytosquelette déterminent l'organisation tridimensionnelle de la cellule et sont particulièrement importantes pour les cellules polarisées comme les cellules épithéliales ou les neurones. Le cytosquelette est constitué de trois types de polymères de protéines qui sont constamment remodelés. Les filaments d'actine sont des polymères d'actines et sont plutôt flexibles. Situés sous la membrane plasmique, ils participent à la forme globale de la cellule ainsi qu'à l'élasticité cellulaire. Ces propriétés sont essentielles lors de la mitose, la méiose ou encore lors de la migration cellulaire car ils sont les constituants des lamellipodes. Les filaments intermédiaires constituent la charpente de la cellule et servent notamment d'ancrage aux organites et au maintien de la structure du noyau. Enfin, les microtubules, d'une vingtaine de nanomètres de diamètre, sont des polymères de tubulines α et β. Polarisés, leur pôle négatif est situé dans le centrosome et leur pôle positif peut être localisé dans toute la cellule. L'assemblage et le désassemblage des microtubules est extrêmement dynamique et régule de la sorte le réseau de transport des molécules dans la cellule.

Les protéines motrices associées aux microtubules permettent le transport actif de protéines, de molécules d'ARN et d'ADN, ou même d'organites, de manière ATP-dépendante. Elles se distinguent selon qu'elles se déplacent vers le pôle négatif ou positif des microtubules.

Les protéines motrices sont en particulier impliquées dans les mécanismes moléculaires mis en œuvre lors de la contraction musculaire, de la migration cellulaire, de l'exocytose, du transport intracellulaire d'ARN messagers, ou encore des mouvements chromosomiques lors de la division cellulaire.

1.1.2.3. Les enzymes permettent la majorité des réactions chimiques du vivant

Les enzymes sont des protéines qui, en rapprochant spatialement les réactants, en les plaçant dans une conformation stéréochimique idéale et/ou en rendant certains groupements chimiques davantage réactifs, catalysent les réactions chimiques de la cellule dont la cinétique est trop lente pour se produire sans l'assistance des enzymes.

Les enzymes ont été classées selon le type de réaction chimique qu'elles catalysent. Elles interviennent dans l'ensemble des processus biologiques.

La localisation des enzymes dans la cellule est indissociable de leur fonction : certaines sont trouvées dans les membranes, les organites, le cytosol ou le milieu extracellulaire (on parle dans ce dernier cas d'exo-enzymes).

1.1.2.4. Certaines protéines participent à l'assemblage de complexes moléculaires

La cellule contient des complexes macromoléculaires qui catalysent certaines réactions chimiques fondamentales de la biologie. C'est par exemple le cas du ribosome qui catalyse la polymérisation des acides aminés en protéines lors de la traduction des ARN messagers. L'assemblage du ribosome est abordé dans le paragraphe 1.2.1.2. Au cours de la synthèse des protéines, certaines protéines nommées facteurs de traduction permettent l'assemblage et la progression du ribosome sur l'ARN messager, et ainsi la synthèse d'une protéine. Ces mécanismes seront abordés dans les parties à suivre.

Plus généralement, l'assemblage de complexes moléculaires concerne l'ensemble des protéines qui, en ayant une affinité pour une molécule donnée, permettent le recrutement d'autres protéines vers cette molécule sans qu'elles aient nécessairement d'affinité l'une pour l'autre : c'est un recrutement indirect.

1.1.2.5. Des dysfonctionnements de la synthèse des protéines sont à l'origine de maladies

De nombreuses maladies sont liées au dysfonctionnement d'une ou plusieurs protéine(s), à leur sur- ou sous- expression, ainsi qu'à leur taux de synthèse. La liste étant très longue, citons ici quelques exemples de maladies où le mécanisme de synthèse d'une protéine est directement impliqué.

La synthèse de poly-dipeptides toxiques dans les motoneurones est une des causes de la sclérose latérale amyotrophique (maladie de Charcot). Ces dipeptides proviennent de la traduction de séquences répétées apparaissant dans les transcrits du gène C9ORF72 chez les patients atteints de sclérose latérale amyotrophique. Ces transcrits sont pris en charge par le ribosome de la cellule, ce qui induit la synthèse de poly-dipeptides toxiques qui provoquent la mort des motoneurones (Tabet *et al.* 2018).

La présence d'un codon stop prématuré dans l'ARN messager codant pour la dystrophine, une protéine transmembranaire indispensable au maintien de l'architecture des cellules

musculaires, stoppe la synthèse de la dystrophine, ce qui cause la myopathie de Duchenne (Hoffman 2020).

Dans le cas d'une infection virale, le détournement de la machinerie cellulaire de synthèse des protéines est essentiel à la production des protéines virales et donc à l'assemblage de nouvelles particules virales fonctionnelles et infectieuses (Walsh and Mohr 2011; Jaafar and Kieft 2019).

Enfin, un facteur d'initiation de la synthèse des protéines, eIF4E, est surexprimé dans de nombreux cancers (Smith *et al.* 2021). Ce facteur guide l'assemblage de la machinerie traductionnelle sur l'extrémité 5' des ARN messagers. Sa surexpression dérègle les taux de synthèse des protéines de la cellule et participe dès lors au phénomène de cancérisation.

1.1.3. A l'échelle moléculaire

Les fonctions des protéines sont strictement dépendantes de leur conformation tri-dimensionnelle, elle-même dictée par la séquence d'acides aminés qui les constituent, d'éventuelles modifications post-traductionnelles, ainsi que par la vitesse de leur synthèse.

Les protéines sont des polymères d'acides aminés qui adoptent un repliement tridimensionnel leur conférant leur(s) fonction(s). Localement, les séquences d'acides aminés se replient en deux motifs structuraux élémentaires : les hélices α et les feuillets β. Les acides aminés hydrophiles sont généralement localisés à la surface des protéines qui fait face au solvant et permettent souvent des interactions avec diverses molécules. Les acides aminés hydrophobes permettent l'ancrage dans les membranes dans le cas de protéines membranaires, sinon ils forment généralement le cœur hydrophobe de la protéine. Certaines protéines se décomposent en plusieurs domaines fonctionnels, comme les domaines catalytiques ou les domaines d'interaction avec d'autres molécules. Ces domaines sont souvent séparés par une chaîne peptidique non structurée qui leur confère de ce fait une certaine mobilité.

Les protéines sont parfois constituées de plusieurs chaînes polypeptidiques qui peuvent s'associer pour former des homo- (protéines identiques) ou hétéro- (protéines différentes) oligomères. C'est par exemple le cas de l'hémoglobine qui est un hétérotétramère formé de deux chaînes α (α-globine) et de deux chaînes β (β-globine). C'est aussi le cas des anticorps qui sont constitués de deux chaînes lourdes et de deux chaînes légères.

Enfin, certaines protéines contiennent des modifications post-traductionnelles sur des acides aminés précis qui peuvent affecter leur repliement et/ou leur(s) spécificité(s) d'interaction avec des substrats. On peut citer la phosphorylation des résidus sérines, thréonines et tyrosines par les protéines kinases qui interviennent dans la totalité des voies de signalisation cellulaires.

Conclusion

Ce tour d'horizon des rôles biologiques des protéines montre que leur synthèse cellulaire est un mécanisme biologique fondamental qui est au cœur du fonctionnement du vivant. Des études ont montré que la cellule dédie près d'un tiers, voire la moitié pour certaines bactéries, de ses ressources énergétiques à la mise en œuvre de mécanismes permettant la traduction fidèle et efficace de ses ARN messagers en protéine(s) (Buttgereit and Brand 1995; Russell and Cook 1995; Li *et al.* 2014).

1.2. Mécanismes assurant la biosynthèse des protéines et sa fidélité

La biosynthèse des protéines, ou traduction des ARN messagers, est un mécanisme conservé dans les trois domaines du vivant. Elle fait intervenir une machinerie moléculaire dédiée au décodage de l'information génétique permettant la synthèse des protéines. Cette information est codée dans la phase codante des ARN messagers sous la forme de triplets de nucléotides appelés codons selon les règles du code génétique. De fortes homologies de fonction et de séquences sont trouvées parmi les constituants de la machinerie de synthèse de protéines entre les Eucaryotes, les Bactéries et les Archées (Tableau 1). Cette forte conservation est révélatrice du caractère fondamental et ancestral de la traduction des ARN messagers en protéines chez tous les êtres vivants. Les mécanismes régulateurs de la traduction seront abordés dans les parties 1.3 et 1.4.

Le travail présenté dans ce manuscrit se focalise sur les mécanismes de la traduction des ARN messagers chez les eucaryotes (Figure 1).

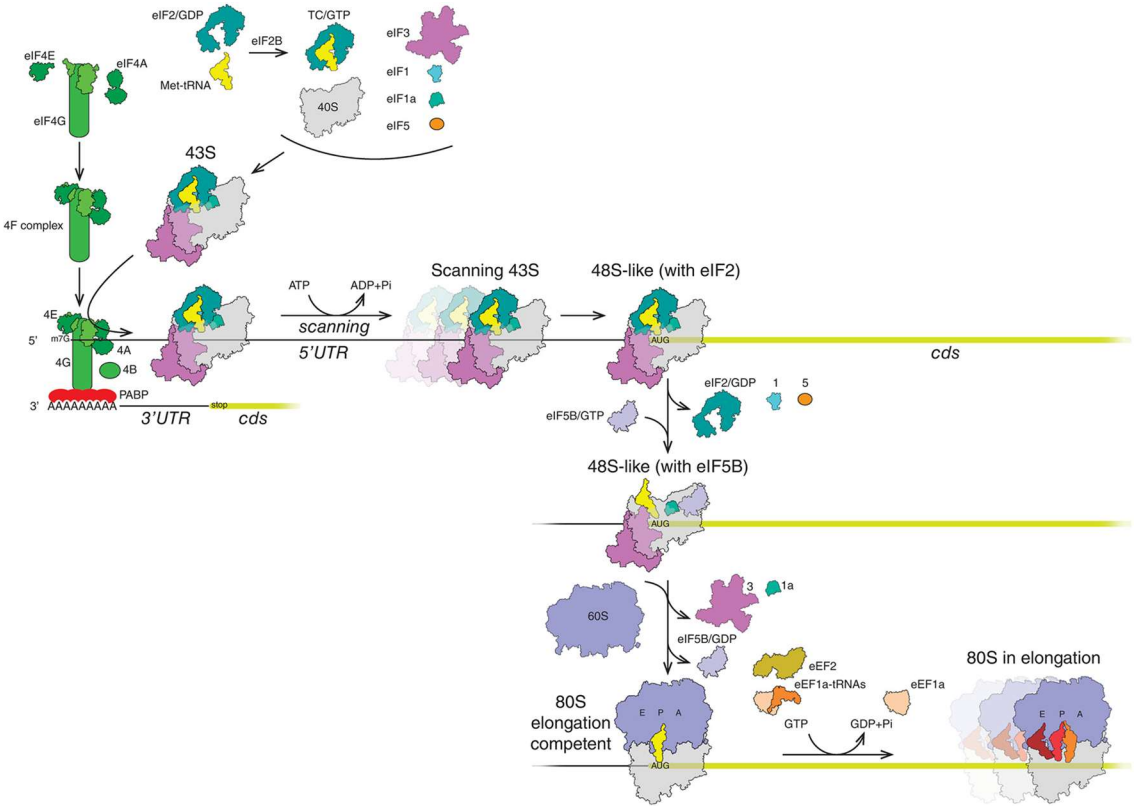


Figure 1 : vue d'ensemble des mécanismes de traduction canoniques, ou coiffe-dépendants, chez les eucaryotes. eIF : facteur d'initiation eucaryotique, eEF : facteur d'élongation eucaryotique, eRF : facteur de terminaison eucaryotique, UTR : région non traduite, cds : phase codante. D'après (Mailliot and Martin 2018).

1.2.1. Les principaux acteurs moléculaires de la traduction chez les eucaryotes

La synthèse des protéines nécessite la coopération ordonnée et localisée de molécules d'ARN et de protéines.

Les ARN sont le résultat de l'expression d'un gène, c'est-à-dire de sa transcription, qui est soumise à de nombreux mécanismes de régulation qui ne seront pas détaillés dans ce rapport.

Les ARN sont des polymères de ribonucléotides (adénosine, cytidine, guanosine et uridine) qui peuvent être modifiés ou non. Chez les eucaryotes, ils sont synthétisés par trois ARN polymérases et se distinguent par leur caractère codant ou non-codant. Les ARN codants, ou ARN messagers (ARNm), portent l'information génétique qui est traduite en protéines par les ribosomes. Les ARN non-codants interviennent dans les mécanismes de régulation de la totalité des processus biologiques, même si l'étendue et parfois la nature de leurs implications restent à déterminer précisément. Parmi les ARN non-codants figurent des acteurs majeurs de la traduction : les ARN de transfert (ARNt) délivrent les acides aminés aux ribosomes qui eux-mêmes sont principalement constitués d'ARN ribosomiques (ARNr). Par conséquent, même s'il ne fait pas l'objet de cette étude, il est clair qu'un premier niveau de régulation de la traduction a lieu au niveau transcriptionnel par un étroit contrôle des taux de synthèse des molécules d'ARN mises en jeu, qu'elles soient codantes ou non. A l'instar des protéines, le repliement tridimensionnel des ARN est fondamental pour leur fonction. Il fait intervenir d'une part l'établissement de liaisons hydrogènes intramoléculaires entre les bases des nucléotides, les groupements phosphates étant la plupart du temps côté solvant et les plateaux de bases très souvent à l'intérieur de l'hélice ainsi formée, et d'autre part des interactions avec diverses molécules. L'ensemble de ces interactions détermine la topologie de repliement des différentes régions de l'ARN concerné, qui peut être une simple tige-boucle, un pseudo-nœud (interaction longue distance entre deux tiges-boucles) ou des structures bien plus stables comme les G-quartets (Dumas *et al.* 2021). Les structures ou domaines qui en résultent peuvent être modulés par les conditions de pH, les concentrations en ions, ainsi que par leurs interactions avec différentes protéines et petites molécules.

Les protéines directement impliquées dans la traduction peuvent se distinguer selon le type d'ARN avec lequel elles interagissent. Leurs interactions avec les ARNm, les ARNr ou les ARNt sont fondamentales et permettent l'assemblage de complexes de traduction fonctionnels.

1.2.1.1. Les ARN messagers

Les ARN messagers (ARNm) eucaryotes sont le résultat de multiples modifications des ARN issus de la transcription des gènes correspondants par l'ARN polymérase II dans le noyau des cellules. Au contraire des ARNm procaryotes, ils sont essentiellement monocistroniques, c'est-à-dire qu'ils ne codent en principe pour qu'une seule protéine.

Le produit de la transcription par l'ARN polymérase II, l'ARN pré-messager (ARNpm), est sujet à de nombreuses modifications co- et post- transcriptionnelles interconnectées. Parmi les plus importantes figurent l'ajout d'une coiffe 7-méthyl-guasosine-triphosphate (m^7Gppp) en 5', l'ajout d'une queue poly-adénosines (poly-A) en 3' ainsi que l'épissage des introns, auxquels peuvent s'ajouter d'autres modifications de bases. Le déroulement de ces mécanismes qui ont lieu pendant la transcription peut influencer l'efficacité de la traduction des ARNm.

<u>Mécanisme de synthèse de la coiffe et de modifications de bases</u>

Chez les eucaryotes, l'ajout de la coiffe m^7G (Figure 2) fait intervenir trois activités enzymatiques portées par trois enzymes nucléaires distinctes et a lieu de manière co-transcriptionnelle après la polymérisation d'une vingtaine de nucléotides (Cowling 2010; Ramanathan *et al.* 2016). D'abord, une ARN triphosphatase clive l'extrémité 5'-triphosphate de l'ARN néosynthétisé, ce qui donne une extrémité diphosphate et libère un phosphate inorganique (Pi) :

$$pppN_1pN_2 \rightarrow ppN_1pN_2 + Pi$$

Ensuite, une ARN guanylyl-transférase catalyse la formation d'une liaison 5'-5'-triphosphate en utilisant une molécule de guanosine triphosphate (GTP), ce qui libère un pyrophosphate (PPi) :

$$ppN_1pN_2 + GTP \rightarrow GpppN_1pN_2 + PPi$$

Enfin, une guanine-7-méthyltransférase catalyse la méthylation de l'azote 7 de la guanosine précédemment ajoutée en 5' en utilisant la S-adénosyl-méthionine (SAM) comme donneur de groupement méthyl :

$$GpppN_1pN_2 + \text{S-adénosyl-méthionine} \rightarrow \mathbf{m^7GpppN_1pN_2} + \text{S-adénosyl-L-homocystéine}$$

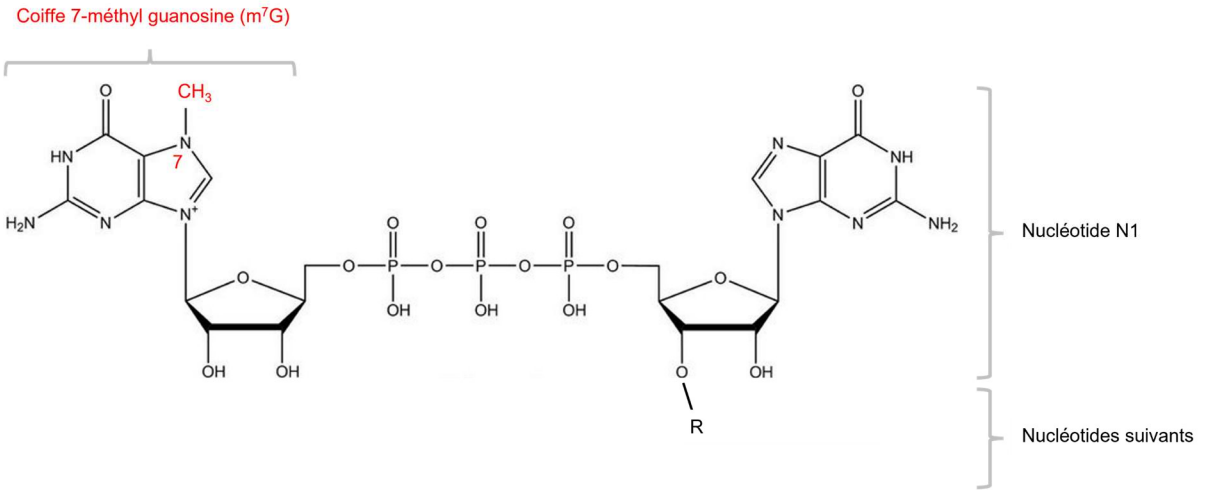


Figure 2 : la coiffe 7-méthyl guanosine. R correspond aux nucléotides ($N_{i>1}$) qui composent l'ARNm. Adapté de (Ramanathan *et al.* 2016).

La synthèse de la coiffe m^7G à l'extrémité 5' de l'ARN en cours de synthèse est fondamentale pour sa stabilité ainsi que pour nombre de mécanismes moléculaires comme l'épissage de ses introns, son export nucléaire ou sa traduction.

En plus de la coiffe m^7G, d'autres modifications de base comme la N_6-méthyl-adénosine ou les coiffes tri-méthylées sont possibles et peuvent influencer la traduction des ARNm. Ces mécanismes seront abordés dans le paragraphe 1.4.1.1. Par ailleurs, certains virus ajoutent une coiffe m^7G en 5' des ARN viraux au moyen d'une seule protéine qui possède les trois activités enzymatiques précédemment décrites : c'est notamment le cas de la Vaccinia Capping Enzyme (VCE), une protéine du virus de la vaccine (Martin and Moss 1975; Shuman and Hurwitz 1981).

<u>Mécanisme d'épissage des introns</u>

L'ARN pré-messager en cours de synthèse est également soumis à un large remodelage nommé épissage des introns (Shi 2017) dont les aspects mécanistiques et régulateurs ne seront pas développés ici. Sur l'ARN pré-messager, l'information génétique codant pour la protéine est fragmentée : les séquences codantes, les exons, sont séparées par des régions non-codantes, les introns. L'épissage des introns correspond au clivage, par un complexe ribonucléoprotéique nommé spliceosome, des séquences comprises entre les extrémités

respectivement 5'-P et 2'-OH de deux introns qui ne sont pas nécessairement consécutifs (Figure 3).

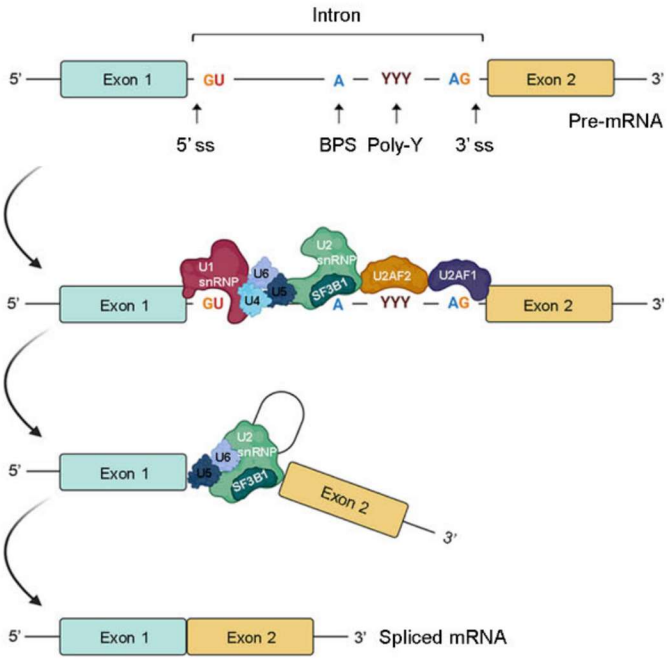


Figure 3 : mécanisme simplifié de l'épissage des introns. BPS : séquence de branchement ; ss : site d'épissage ; poly-Y : poly-pyrimidines. D'après (Bessa *et al*. 2020).

La jointure des exons qui respectivement précèdent et suivent le ou les intron(s) concerné(s) fait que l'ARNm sera composé d'une combinaison d'exons. Tous les exons présents dans la séquence de l'ARN pré-messager ne sont pas nécessairement présents sur l'ARNm, et plusieurs combinaisons d'exons sont possibles pour un même transcrit : l'épissage est alternatif. La traduction des ARNm issus d'un épissage alternatif du même ARN pré-messager, et donc de l'expression du même gène, donne des protéines distinctes dites isoformes.

Mécanisme d'ajout de la queue poly-adénosines à l'extrémité 3'

A l'exception notable des ARN messagers codant pour les histones dites « réplication-dépendantes » qui ont une structure en tige boucle à leur extrémité 3' (Lyons *et al.* 2016), les ARN messagers eucaryotes sont polyadénylés à leur extrémité 3' par un mécanisme co-transcriptionnel en deux étapes. D'abord, la fixation des facteurs de polyadénylation sur des motifs spécifiques localisés de part et d'autre du site de polyadénylation guide leur clivage par des endonucléases. Intervient ensuite la polymérisation d'une succession de 150 à 250 adénosines (Tian 2005) par la poly-adénosines polymérase (PAP).

Transport des ARNm du noyau au cytoplasme

Toujours de manière co-transcriptionnelle et en étroite coordination avec les mécanismes précédents, l'ARN messager s'associe avec de nombreuses protéines pour préparer son export du noyau vers le cytoplasme par les pores nucléaires (Stewart 2019). Cet assemblage requiert l'accomplissement des mécanismes d'ajout de la coiffe, d'épissage et de polyadénylation car ces modifications interviennent dans la reconnaissance des protéines

formant le complexe d'exportation. Ce complexe ribonucléoprotéique, ou cargo, présente une affinité pour les nucléoporines, et peut alors se rapprocher du pore nucléaire et sortir du noyau. La diffusion du cargo étant bidirectionnelle, divers réarrangements moléculaires, dont l'intervention d'hélicases, permettent de décrocher les protéines du cargo de l'ARNm, assurant ainsi son transport unidirectionnel du noyau vers le cytoplasme (Figure 4).

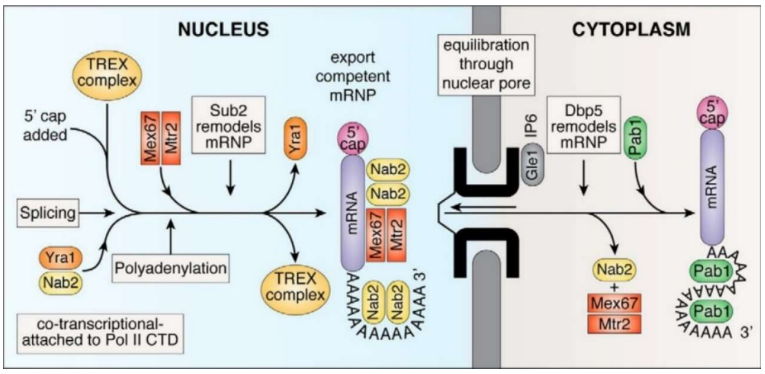


Figure 4 : schéma récapitulatif des différentes étapes de la maturation des ARNm dans le noyau et de leur export nucléaire. D'après (Stewart 2019).

Eléments fonctionnels des ARNm

La séquence des ARNm se divise en trois régions toutes impliquées dans la régulation de leur traduction en protéine (Figure 5), à savoir les régions non traduites 5' et 3' (5'UTR et 3'UTR) ainsi que la séquence codante qui contient l'information génétique pour la synthèse de la protéine. L'impact de ces éléments sur la traduction des ARNm sera abordé dans le paragraphe 1.4 dédié à la régulation de la traduction par la séquence de l'ARNm.

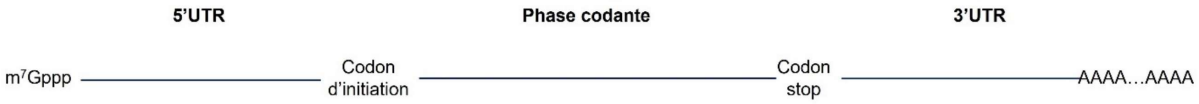


Figure 5 : représentation simplifiée d'un ARN messager eucaryote.

1.2.1.2. Les ribosomes : assemblages d'ARN et de protéines ribosomiques

Les ribosomes désignent un assemblage ribonucléoprotéique constitué d'ARN ribosomiques (ARNr) entourés de protéines ribosomiques.

Les ARN ribosomiques

Ce sont les principaux constituants des ribosomes. Les ARNr sont issus de la maturation d'ARN précurseurs issus de la transcription des gènes correspondants au sein du nucléole par les ARN polymérases I et III. La première synthétise l'ARN précurseur des ARNr 18S, 5.8S et 28S et la seconde le précurseur de l'ARNr 5S. Les ARN précurseurs sont ensuite clivés en ARNr dont le repliement fonctionnel s'effectue de manière co-transcriptionnelle et est assuré par de multiples protéines, dont les principaux mécanismes sont expliqués dans la revue (Henras *et al.* 2015).

Les protéines ribosomiques

Le nombre et la diversité des protéines ribosomiques dans les trois domaines du vivant ont poussé les chercheurs à définir une nomenclature unifiée et universelle des protéines ribosomiques (Ban *et al.* 2014) qui sera utilisée dans ce travail.

Assemblage des ribosomes

La biogenèse des ribosomes a lieu dans le noyau. Les protéines ribosomiques, nécessaires à la biogénèse et produites dans le cytoplasme, y sont donc exportées. Les ARNr et les protéines ribosomiques s'assemblent dans le nucléole pour former les sous-unités 40S et 60S du ribosome. Cet assemblage nécessite plus de 200 protéines, dont la plupart sont des chaperonnes (Klinge and Woolford 2019). Ces nombreuses interactions moléculaires sont nécessaires pour assurer le repliement fonctionnel des ARN ribosomiques des deux sous-unités du ribosome. Une fois formées à la suite d'un processus complexe qui ne sera pas détaillé ici, les sous-unités ribosomiques sont exportées du noyau vers le cytoplasme pour les dernières étapes de leur maturation.

Principaux éléments fonctionnels du ribosome

D'une masse d'environ 4.5 x10^6 daltons (Yusupova and Yusupov 2015) et d'une envergure d'une vingtaine de nanomètres, les ribosomes eucaryotes sont constitués des deux sous-unités 40S et 60S. La petite sous-unité 40S est constituée de l'ARNr 18S ainsi que de 33 protéines ribosomiques. Associée aux facteurs d'initiation, elle forme la particule 43S qui intervient dans l'ensemble du processus d'initiation de la traduction. La grande sous-unité 60S, constituée des ARNr 28S, 5.8S et 5S ainsi que de 46 protéines ribosomiques, se joint à la petite sous-unité après qu'elle a reconnu le codon d'initiation. Le ribosome 80S ainsi formé démarre ensuite la synthèse de la protéine. Le ribosome 80S contient trois sites fonctionnels (Figure 6, Figure 7) nommés par rapport à l'état des ARN de transfert au cours de la synthèse de la protéine : le site E (Exit), le site P (Peptidyl-ARNt) et le site A (Amino-acyl ARNt).

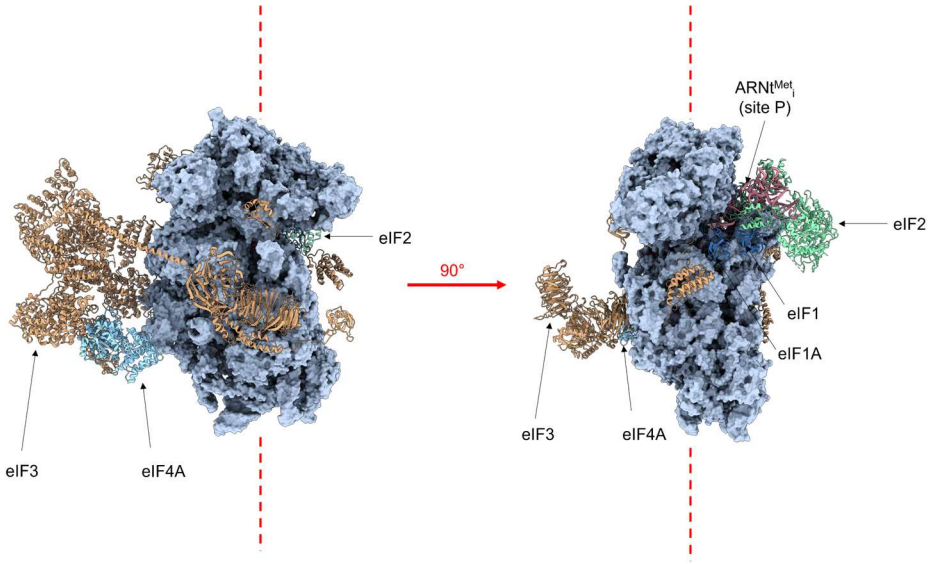


Figure 6 : structure du complexe de pré-initiation 48S humain. Adapté de (Querido *et al.* 2020), PDB : 6ZMW

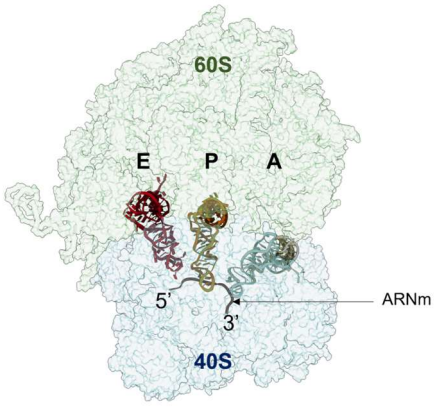


Figure 7 : structure du ribosome 80S humain. Les positions des ARNt ont été modélisées dans les sites E, P et A du ribosome humain (PDB : 6QZP) à l'aide des structures de ribosomes bactériens (PDB : 4V5G). Adapté de (Schmeing *et al.* 2009; Natchiar *et al.* 2017)

Le site A correspond au site de décodage de l'information génétique en acide aminé par l'interaction entre le codon de l'ARN messager et l'anticodon de l'ARN de transfert lié à l'acide aminé (amino-acyl ARNt). Le site P contient le site peptidyl-transférase qui catalyse la formation de la liaison peptidique entre le peptide naissant et l'acide aminé amené dans le site A. Le produit final de la réaction est le peptidyl-ARNt. Enfin, le site E est le site de décrochage de l'ARNt déacylé à la suite de cette réaction. Le détail de ces mécanismes sera abordé dans le paragraphe 1.2.3.

La catalyse de la liaison peptidique ainsi que la dynamique traductionnelle sont des mécanismes extrêmement conservés dans les trois domaines du vivant.

Domaine	**Ribosome** (Svedberg)	**Petite et grande sous-unités du ribosome** (Svedberg)	**ARNr de la petite et de la grande sous-unité du ribosome** (Svedberg)	**Protéines ribosomiques**
Bactéries	**70S**	**30S**	**16S** (1493 nt)	**21** (15u / 6b)
		50S	**23S** (2891 nt) **5S** (117 nt)	**33** (18u / 15b)
Archées	**70S**	**30S**	**16S** (1512 nt)	**25** (15u / 9e / 1a)
		50S	**23S** (2967 nt) **5S** (122 nt)	**38** (18u / 20e)
Eucaryotes	**80S**	**40S**	**18S** (1860 nt)	**33** (15u / 18e)
		60S	**28S** (4039 nt) **5,8S** (158 nt) **5S** (120 nt)	**46** (18u / 28e)

Tableau 1 : les ribosomes dans les trois domaines du vivant. S est le coefficient de sédimentation exprimé en Svedberg. La longueur des ARN ribosomiques (ARNr) est indiquée en nucléotides (nt). Les protéines ribosomiques sont désignées par une nomenclature officielle comprenant un préfixe selon qu'elles sont conservées dans les trois domaines du vivant (u pour universelle), ou conservées chez les bactéries (b), chez les archées (a) ou chez les eucaryotes (e). Ce préfixe est suivi de la lettre S pour les protéines de la petite sous-unité ou L pour les protéines de la grande sous-unité du ribosome, suivi d'un numéro (Ban *et al.* 2014).

1.2.1.3. Les ARN de transfert et les amino-acyl-ARNt synthétases

Les ARN de transfert (ARNt) sont issus de la maturation des ARN précurseurs synthétisés par la transcription des gènes correspondants par l'ARN polymérase III dans le noyau. La maturation nucléaire des ARN précurseurs d'ARNt implique notamment la dégradation contrôlée par des exo- et endo- nucléases des extrémités 5' et 3', couplée ou non à de l'épissage d'introns, ainsi que l'ajout du triplet de nucléotides CCA à l'extrémité 3' par la nucléotidyl-ARNt-transférase (Betat *et al.* 2014). A cela s'ajoutent de multiples modifications de bases qui précèdent ou suivent l'export des ARNt du noyau vers le cytoplasme. Les ARNt sont des ARN non codants strictement essentiels au décodage de l'information génétique et à la traduction des ARN en protéines, mais ils sont aussi impliqués dans d'autres mécanismes moléculaires de la cellule (Schimmel 2018). Le nombre de gènes codant pour les ARNt est très variable selon les organismes. Il en existe néanmoins au moins un par acide aminé protéinogène. L'abondance des ARNt d'une cellule détermine la vitesse de décodage des codons correspondants lors de l'étape d'élongation, et est ainsi reliée à l'efficacité globale de la traduction cellulaire. D'une longueur comprise entre 76 et 93 nucléotides, les ARNt adoptent un repliement fonctionnel se caractérisant par une structure en « L » qui contient plusieurs éléments conservés dans les trois domaines du vivant (Figure 8).

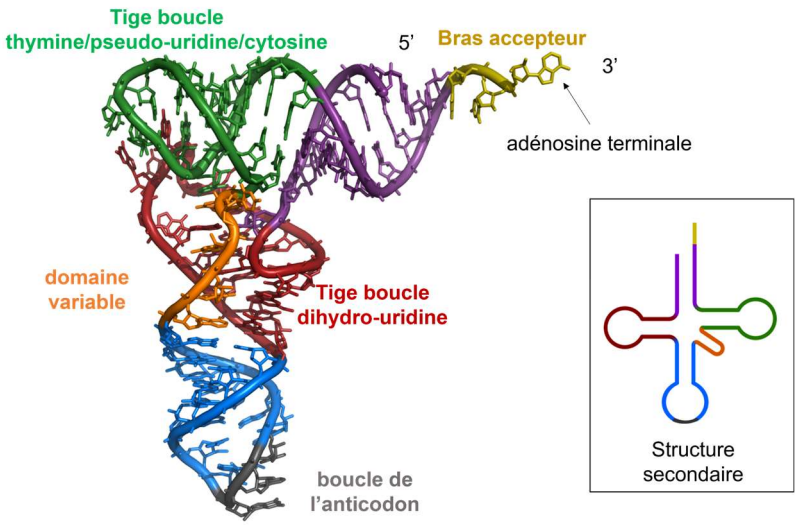


Figure 8 : structures secondaire et tridimensionnelle de l'ARNt-Phénylalanine de la levure (PDB : 1ehz).

De l'extrémité 5' vers 3' se trouvent la tige-boucle dihydro-uridine (D), la tige-boucle de l'anticodon, un domaine variable et la tige-boucle thymine/pseudo-uridine/cytosine. Enfin, le triplet CCA est porté par le bras accepteur de l'ARNt sur lequel sont fixés les acides aminés par les aminoacyl ARNt synthétases. L'élément central du décodage du code génétique est la boucle anticodon. Deux ARNt sont dits isoaccepteurs s'ils portent deux anticodons distincts mais portent le même acide aminé, illustrant ainsi la redondance du code génétique. Deux ARNt sont dits isodécodeurs s'ils portent le même anticodon mais ont par ailleurs une séquence nucléotidique et/ou des modifications de bases distinctes. Les rôles et la diversité des modifications des bases des ARNt sont nombreux (Grosjean and Westhof 2016; Schaffrath and Leidel 2017). Les modifications de la base 34 située dans la boucle anticodon sont en particulier impliquées dans la fidélité de décodage par l'interaction avec le troisième

nucléotide du codon (Pernod *et al.* 2020) mais également dans la vitesse de synthèse des protéines à l'échelle cellulaire (Nedialkova and Leidel 2015).

Les aminoacyl ARNt synthétases (aaRS) sont les enzymes permettant le chargement des acides aminés à l'extrémité 3' des ARNt. Il y en a une par acide aminé protéinogène. Le chargement de ces acides aminés sur le bras accepteur des ARNt constitue la première étape du décodage du code génétique. La fidélité du décodage passe par la reconnaissance spécifique des deux substrats des aaRS : d'une part l'ARNt porteur de l'anticodon correspondant à l'acide aminé chargé par l'aaRS et d'autre part l'acide aminé en question. La reconnaissance de l'ARNt est dépendante d'un jeu de nucléotides spécifiques appelés déterminants qui peut être couplé à des modifications de base. La reconnaissance de l'acide aminé a lieu dans une poche spécifique de l'aaRS et implique principalement les discriminations de taille et de charges de surface, assistée dans certains cas par des ions métalliques (Rubio Gomez and Ibba 2020).

Le chargement de l'acide aminé sur l'extrémité 3' de l'ARNt s'effectue en deux étapes. D'abord, le groupement acide carboxylique C-terminal est rendu davantage réactif par une réaction d'aminoacylation qui utilise l'ATP, formant ainsi un aminoacyl-adénylate. Les aaRS catalysent ensuite l'estérification du groupe carbonyl de l'adénylate par l'extrémité hydroxyle 2' ou 3' du ribose de l'adénosine terminale, conduisant ainsi au chargement de l'acide aminé sur le bras accepteur de l'ARNt. Les aaRS se distinguent en deux classes selon que la réaction d'estérification implique l'hydroxyle 2' (classe I) ou 3' (classe II) de l'adénosine terminale de l'ARNt (Eriani *et al.* 1990). Selon la classe d'aaRS, l'interaction avec l'ARNt se fait par l'une des deux faces de la structure en L des ARNt.

1.2.1.4. Les facteurs de traduction

De nombreux facteurs protéiques interviennent lors des étapes d'initiation (eIF pour « eukaryotic initiation factor »), d'élongation (eEF pour « eukaryotic elongation factor ») et de la terminaison (eRF pour « eukaryotic release factor ») de la traduction. Certains remplissent des fonctions très similaires entre les trois domaines du vivant. Le Tableau 2 récapitule les différents facteurs eucaryotiques mis en jeu au cours de la traduction chez les eucaryotes. Ils seront évoqués au fur et à mesure dans le texte à suivre.

eIF	Nombre de sous-unités	Masse (kDa)
eIF1	1	12,7
eIF1A	1	16,5
eIF2	3	36,1 / 38,4 / 51,1
eIF2B	5	33,7 / 39 / 50,2 / 59,7 / 80,3
eIF3	13	800 (total)
eIF4A	1	46,1
eIF4B	1	69,3
eIF4E	1	24,5
eIF4G	1	175,5
eIF4F	3	eIF4E + eIF4G + eIF4A
eIF4H	1	27,4
eIF5	1	49,2
eIF5B	1	138,9

eEF	Nombre de sous-unités	Masse (kDa)
eEF1A	1	50,1
eEF1B	3	24,8 / 31,1 / 50,1
eEF2	1	95,3
eIF5A	1	16,8

eRF	Nombre de sous-unités	Masse (kDa)
eRF1	1	49
eRF3	1	72,1

Tableau 2 : les principaux facteurs de traduction eucaryotiques. eIF : facteur d'initiation, eEF : facteur d'élongation, eRF : facteur de terminaison.

1.2.2. L'initiation de la traduction chez les eucaryotes

L'initiation de la traduction chez les eucaryotes constitue l'étape limitante de la synthèse des protéines. C'est une étape fondamentale dans la mesure où elle détermine si une protéine est synthétisée. L'étape d'initiation est régulée par la majorité des mécanismes de régulation de la traduction connus à ce jour. Malgré la diversité des acteurs protéiques et ribonucléiques mis en jeu, les mécanismes d'initiation de la traduction se distinguent selon qu'ils nécessitent ou non la présence de la coiffe m^7G à l'extrémité 5' des ARNm.

1.2.2.1. L'initiation coiffe-dépendante

Le mécanisme canonique de l'initiation de la traduction (Figure 9) réfère à la stricte nécessité de la présence d'une coiffe 5'-m^7G pour recruter la petite sous-unité du ribosome et ses facteurs associés à l'extrémité 5' de l'ARNm qui sera traduit, déclenchant la série d'évènements décrite ci-après (Jackson *et al.* 2010; Hinnebusch and Lorsch 2012; Aitken and Lorsch 2012; Merrick and Pavitt 2018).

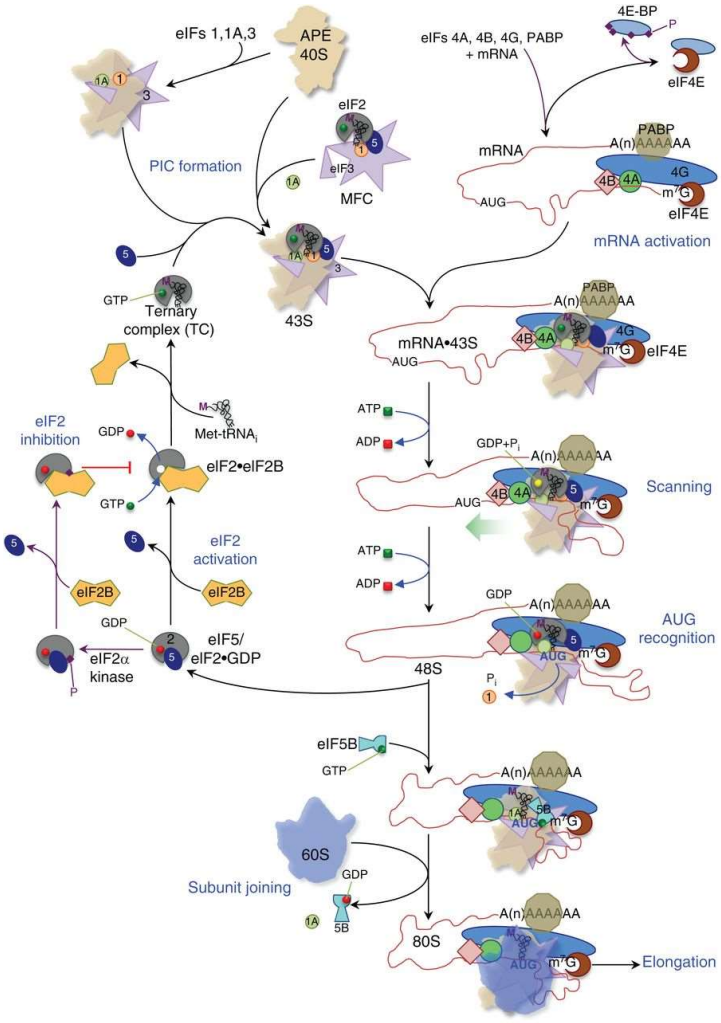


Figure 9 : vue d'ensemble de l'initiation de la traduction chez les eucaryotes. TC : complexe ternaire. D'après (Merrick and Pavitt 2018).

Fixation du complexe 43S à l'extrémité 5' de l'ARNm

Le complexe protéique eIF4F, composé des facteurs d'initiation eIF4G, eIF4A et eIF4E, se fixe sur la coiffe m^7G par l'intermédiaire d'eIF4E et l'assistance d'autres eIF. L'affinité de eIF4E

pour la coiffe est augmentée par l'interaction de son domaine N-terminal avec eIF4G (Gross *et al.* 2003). Le facteur eIF4G joue le rôle de plateforme en interagissant avec divers facteurs du complexe de pré-initiation. Son interaction avec la protéine de fixation à la queue poly-A (PABP) permet la circularisation de certains ARNm (Kahvejian *et al.* 2005) et de ce fait augmente la concentration locale des facteurs de traduction ainsi que la stabilisation de l'ARNm. Ce phénomène n'est toutefois pas indispensable à l'initiation de la traduction des ARNm, comme suggéré par le fait que l'invalidation de l'interaction eIF4G-PABP n'est pas léthale chez la levure (Kessler and Sachs 1998; Park *et al.* 2011). La proximité spatiale des extrémités 5' et 3' des ARNm qui permet leurs interactions avec de nombreuses protéines pourrait être simplement due au repliement tridimensionnel de ces ARNm (Vicens *et al.* 2018). L'interaction de eIF4G avec le facteur eIF3 (Villa *et al.* 2013) permet le recrutement de la petite sous-unité du ribosome associée à d'autres facteurs d'initiation. Parmi ces facteurs figurent notamment le complexe ternaire eIF2/GTP/ARNt-méthionine-initiateur ($ARNt^{Met}_i$), le facteur eIF5 qui est impliqué dans l'hydrolyse du GTP lié à eIF2 en GDP après la reconnaissance du codon initiateur, ainsi que les facteurs eIF1 et eIF1A impliqués dans le scanning et la fidélité de décodage (Asano *et al.* 2000 ; Sokabe *et al.* 2012). Le complexe ribosomique ainsi recruté sur la coiffe m^7G est nommé complexe 43S ou complexe de pré-initiation. eIF3 est notamment impliqué dans la stabilisation des composants du complexe de pré-initiation par ses multiples interactions avec les facteurs d'initiation (Zhou *et al.* 2008; des Georges *et al.* 2015; Valášek *et al.* 2017).

Scanning directionnel de la région 5'UTR à la recherche du codon d'initiation

Muni d'une activité hélicase par le facteur eIF4A qui est exacerbée par ses interactions avec eIF4G (Schütz *et al.* 2008), eIF4B (Rogers *et al.* 2001) ou eIF4H (Sun *et al.* 2012), la particule 43S scanne de manière unidirectionnelle la région 5'UTR de l'ARNm de manière ATP-dépendante, nucléotide par nucléotide, jusqu'à la reconnaissance du codon initiateur par l'$ARNt^{Met}_i$ dans le site P (Kozak 1978). Cette étape de scanning peut également faire intervenir d'autres hélicases selon la longueur de la région 5'UTR, la présence de structures secondaires, ou l'état physiologique de la cellule (Guenther *et al.* 2018; Sen *et al.* 2019).

Reconnaissance du codon initiateur

C'est une étape cruciale dans la mesure où elle détermine l'exactitude de la traduction du code génétique et par conséquent la nature de la protéine synthétisée. Le principal moteur de la reconnaissance du codon initiateur est la stabilité de l'interaction entre l'anticodon porté par l'$ARNt^{Met}_i$ lié à la particule 43S par l'intermédiaire d'eIF2 et le codon d'initiation de la traduction situé dans le site P de la particule 43S (Lomakin and Steitz 2013). La fidélité de reconnaissance du codon initiateur est principalement assurée par la coopération des facteurs eIF1 et eIF1A qui sont localisés de part et d'autre du site P (Pestova *et al.* 1998; Thakur and Hinnebusch 2018; Zhou *et al.* 2020). En induisant une gêne stérique au niveau du site P, ils assurent le maintien d'une conformation du 43S en cours de scanning telle que seule une interaction codon-anticodon suffisamment stable permet de basculer de cette conformation « ouverte » à une conformation « fermée », permettant à l'$ARNt^{Met}_i$ de s'accommoder dans le site P tout en bloquant l'avancement du complexe de scanning. Dans la conformation « ouverte », eIF1 agit comme un inhibiteur de l'hydrolyse spontanée du GTP lié à eIF2 en GDP (Algire *et al.* 2005) en interagissant avec eIF5, et inhibe ainsi l'assemblage des sous-unités du ribosome et le démarrage de la traduction.

Assemblage du ribosome 80S et relargage différentiel des facteurs d'initiation

La reconnaissance du codon d'initiation dans le site P entraine une cascade d'évènements moléculaires menant à l'assemblage des deux sous-unités du ribosome et au recyclage différentiel des facteurs d'initiation. Le complexe moléculaire assemblé sur le codon d'initiation est maintenant nommé complexe 48S. En tant que protéine activatrice de GTPase (GAP) spécifique du complexe ternaire eIF2-GTP- $ARNt^{Met}_i$, eIF5 stimule l'activité GTPase de la sous-unité eIF2-γ en se fixant à la sous-unité β de eIF2 (eIF2-β), ce qui a pour conséquence l'hydrolyse des GTP liés à eIF2-γ en GDP (Paulin *et al.* 2001), elle-même facilitée par le déplacement d'eIF1 après l'accommodation de $ARNt^{Met}_i$ dans le site P. L'avancement de l'hydrolyse du GTP entraine une perte progressive d'affinité d'eIF2 pour l'$ARNt^{Met}_i$ (Kapp and Lorsch 2004), ce qui conduit à son relargage du complexe 48S. L'assemblage des deux sous-unités est grandement assuré par le facteur eIF5B (Pestova *et al.* 2000) et en particulier par l'interaction de son domaine C-terminal avec celui de eIF1A (Nag *et al.* 2016), qui est rendue possible par le déplacement d'eIF1 après l'accommodation de l'$ARNt^{Met}_i$ dans le site P. Cette interaction facilite l'assemblage des deux sous-unités du ribosome et permet le relargage conjoint de eIF2, eIF1A et eIF5B par l'hydrolyse du GTP lié à eIF5B. Très récemment, il a été montré que eIF5B guide le positionnement du bras accepteur de $ARNt^{Met}_i$ de telle sorte qu'il soit compatible avec l'assemblage de la grande sous-unité du ribosome (Lapointe *et al.* 2022). Quant aux facteurs d'initiation eIF3 et eIF4A/G/E, l'existence de mécanismes de ré-initiation (Kozak 2001; Pöyry *et al.* 2004) suggèrent qu'ils pourraient, à ce stade, demeurer associés au ribosome 80S ainsi assemblé durant les premiers temps de l'élongation.

L'interaction codon-anticodon, qui permet de décrocher le facteur eIF1 du site P, est donc le principal moteur de l'assemblage des sous-unités du ribosome et par conséquent du démarrage de la synthèse des protéines.

Recyclage des facteurs d'initiation

Les facteurs d'initiation sont soumis à des mécanismes de recyclage avant d'être à nouveau engagés dans un processus d'initiation.

Le recyclage du facteur eIF2 nécessite l'échange de la molécule de GDP liée à eIF2 en GTP pour la formation de nouveaux complexes de pré-initiation 43S. Après son relargage du 48S, eIF2-GDP interagit avec eIF2B, une protéine échangeuse de guanosine (GEF) qui catalyse l'échange du GDP en GTP (Sidrauski *et al.* 2015; Kashiwagi *et al.* 2019). Seul le complexe eIF2-GTP a l'affinité requise pour l'$ARNt^{Met}_i$, formant ainsi le complexe ternaire essentiel à la formation du complexe de pré-initiation. Le taux de recyclage de eIF2 est donc directement relié au taux de traduction des ARNm cellulaires par le mécanisme d'initiation canonique. Le recyclage du facteur eIF2 est au cœur de mécanismes de régulation qui seront abordés dans la partie 1.3.1.2 et est illustré dans la Figure 19.

Coiffes 5' alternatives

Certains ARNm possèdent des coiffes 5' alternatives qui modulent leur traduction. C'est par exemple le cas des ARNm codant pour les séléno-protéines qui ont une coiffe 2,2,7- triméthyl-guanosine ($m^{2,2,7}G$) (Wurth *et al.* 2014). L'ajout des groupements méthyls supplémentaires sur les azotes 2 est réalisé à partir de la coiffe m^7G dans le cytoplasme ou dans le noyau selon la localisation cellulaire de l'enzyme triméthyl-guanosine synthétase 1 (Tgs1).

1.2.2.2. L'initiation coiffe-indépendante

L'initiation coiffe-indépendante ne fait pas intervenir la coiffe m^7G dans le mécanisme de recrutement du ribosome sur l'ARNm, mais plutôt des structures particulières de l'ARNm en association avec des facteurs protéiques. De telles structures sont appelées sites d'entrée interne du ribosome (IRES). Selon la nature du mécanisme de recrutement du ribosome, certains des eIF requis pour l'initiation canonique peuvent être impliqués. Historiquement identifiés et considérés comme une des stratégies mises en œuvre lors d'une infection virale pour favoriser la traduction de l'ARN du virus (Lee *et al.* 2017), un nombre grandissant de travaux suggère que des mécanismes d'initiation par l'intermédiaire d'IRES sont également utilisés pour la traduction de certains ARNm cellulaires (Godet *et al.* 2019; Kwan and Thompson 2019).

Les IRES viraux

Les IRES viraux sont des éléments structurés retrouvés principalement dans les génomes de virus à ARN qui sont capables de détourner la machinerie de traduction de la cellule infectée vers la traduction des ARN viraux, au détriment des ARNm cellulaires. Assistés ou non par des facteurs, ils permettent de recruter la particule 43S ou un ribosome 80S. Les IRES viraux ont été classés en fonction du nombre et du type de facteurs protéiques nécessaires à leur fonctionnement et de la complexité de leur structure tridimensionnelle (Mailliot and Martin 2018).

Les IRES de classe I nécessitent l'ensemble des facteurs d'initiation de la traduction à l'exception d'eIF4E, le facteur qui fixe la coiffe m^7G dans le mécanisme canonique. L'IRES de la région 5'UTR du Poliovirus (PV) est l'IRES de classe I le mieux étudié (Figure 10).

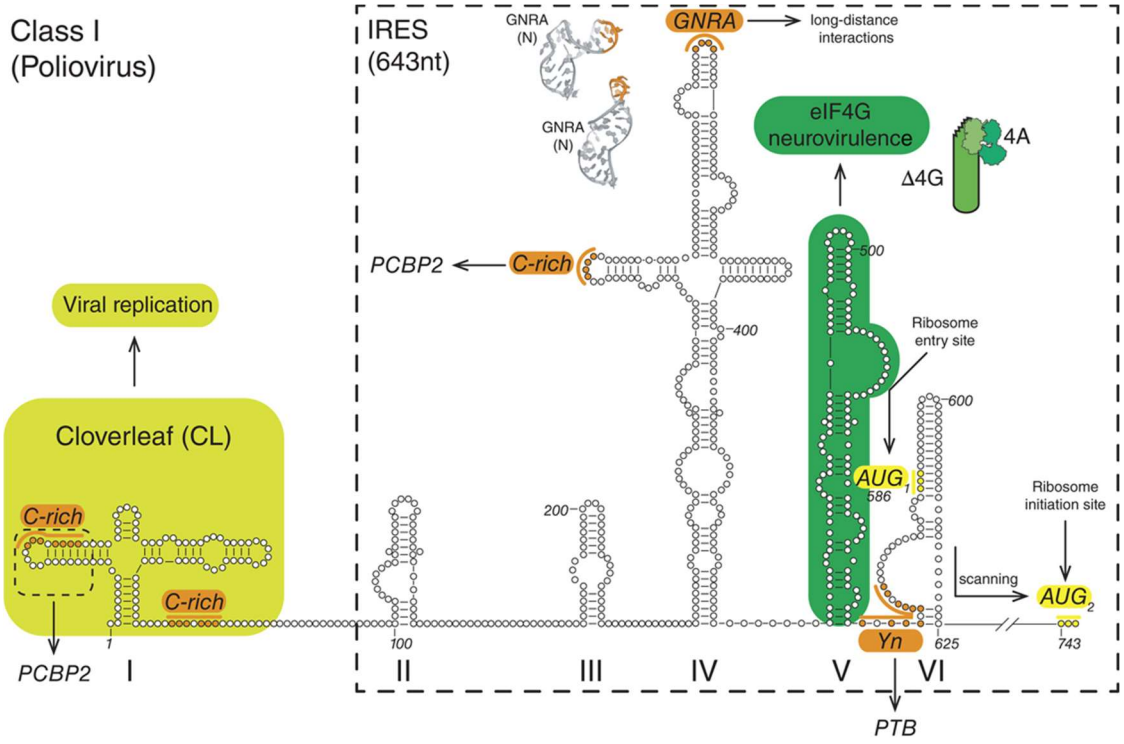


Figure 10 : structure secondaire de la région 5'UTR de l'ARN du Poliovirus. L'IRES minimal est encadré en pointillés. D'après (Mailliot and Martin 2018).

Le repliement fonctionnel de cet IRES a été caractérisé par sondage enzymatique (Pilipenko *et al.* 1989; Skinner *et al.* 1989; Burrill *et al.* 2013) ainsi que par des études de covariations dans les séquences d'autres virus de la famille des poliovirus (Rivera *et al.* 1988). D'une longueur de 743 nucléotides, la région non traduite du poliovirus contient divers éléments structuraux impliqués dans la réplication de l'ARN viral et dans sa traduction (Andino *et al.* 1990, 1993). Le domaine V de l'IRES permet le recrutement de formes complètes et tronquées de eIF4G, qui résultent de son clivage par la protéine 2A, une protéase virale. La protéine résultante conserve son affinité pour eIF4A et eIF3 mais n'est plus incapable d'interagir avec eIF4E, inhibant ainsi la traduction cellulaire coiffe-dépendante (Gradi *et al.* 1998). Etant donné que la protéine eIF4G tronquée est encore capable de recruter les autres facteurs d'initiation, la particule 43S est recrutée sur l'IRES par l'intermédiaire de l'interaction eIF4G-eIF3. Cette interaction amène la particule 43S à être déposée sur un AUG localisé aux positions 586-588 au niveau du domaine VI. S'ensuit le scanning des 160 nucléotides de la région 5' UTR du virus à la recherche du codon de démarrage de la traduction (Lozano and Martínez-Salas 2015). L'assemblage du ribosome sur le codon d'initiation s'effectue de manière similaire au mécanisme canonique. A noter que d'autres IRES de classe I ne font pas nécessairement scanner le complexe 43S recruté et le déposent directement sur un AUG, comme c'est le cas pour l'IRES du Human Rhino Virus 2 (Kaminski *et al.* 2010). Les autres domaines de l'IRES interagissent avec d'autres protéines essentielles au fonctionnement de l'IRES ainsi qu'à la réplication de l'ARN viral : ce sont des ITAF, pour IRES *trans*-acting factors. Elles peuvent être cellulaires ou virales.

Les IRES de classe II nécessitent également l'ensemble des facteurs d'initiation de la traduction à l'exception de celui fixant la coiffe m^7G (eIF4E). Les IRES des régions 5'UTR de l'Encephalomyocarditis Virus (EMCV) et du Foot and Mouth Disease Virus (FMDV) sont les IRES de classe II les plus étudiés (Figure 11).

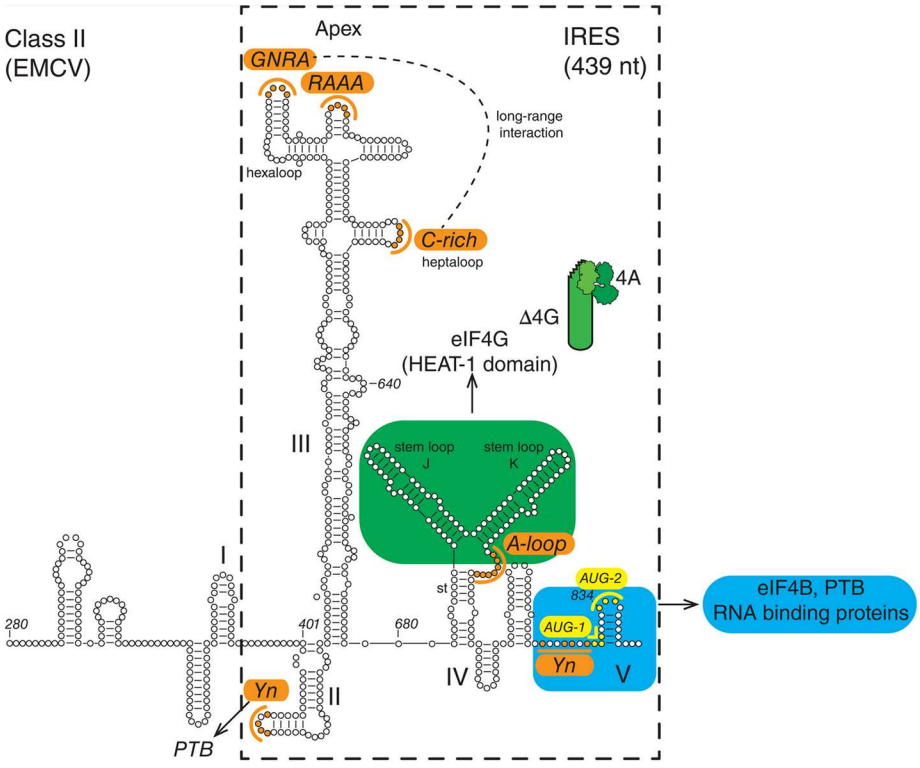


Figure 11 : structure secondaire de l'IRES de EMCV. L'IRES minimal est encadré en pointillés. D'après (Mailliot and Martin 2018).

Au contraire des IRES de la classe I, ces IRES recrutent la particule 43S à proximité du codon d'initiation et ne font donc pas intervenir de scanning à proprement parler. C'est un mouvement oscillant de type brownien limité à quelques nucléotides qui permet de positionner le codon de démarrage dans le site P du 43S recruté sur l'IRES. Les structures secondaires des IRES EMCV et FMDV sont bien caractérisées et sont très similaires. Le domaine J-K des deux IRES est essentiel au recrutement du facteur eIF4G, par l'intermédiaire duquel est recrutée la particule 43S de manière analogue au mécanisme d'initiation canonique (Jang and Wimmer 1990; Hoffman and Palmenberg 1995; Kolupaeva *et al.* 1998; Lozano and Martínez-Salas 2015; Imai *et al.* 2016). A noter que le facteur eIF4G est également clivé lors de l'infection par ces virus (voir paragraphe précédent). Dans l'IRES d'EMCV, le recrutement d'eIF4G est largement influencé par une boucle de six adénosines consécutives au sein du domaine J-K (Hoffman and Palmenberg 1995; Imai *et al.* 2016). Deux AUG initiateurs en phase peuvent servir de codon d'initiation dans les deux IRES. Ils sont séparés de 9 nucléotides dans EMCV, et de 84 nucléotides dans FMDV. En ce sens, le recrutement au niveau du domaine V du facteur eIF4B permettrait d'assister leur reconnaissance, dans ce cas-là, par un mécanisme de scanning. L'assemblage du ribosome qui suit la reconnaissance de l'AUG s'effectue de manière similaire au mécanisme canonique. Le domaine III, et notamment le motif GNRA de sa partie apicale, est essentiel à la fonctionnalité des deux IRES, bien que les mécanismes impliqués demeurent mal compris (Robertson *et al.* 1999; López de Quinto and Martínez-Salas 1997). Les autres domaines de l'IRES fixent des facteurs protéiques, cellulaires ou viraux, essentiels au maintien de la conformation tridimensionnelle de l'IRES (ITAFs chaperonnes) ou plus généralement essentiels au cycle viral.

Les IRES de classe III nécessitent un nombre plus restreint de facteurs d'initiation, parmi lesquels le facteur eIF3 joue un rôle majeur. L'IRES du virus de l'hépatite C (HCV) en est l'exemple le plus connu (Figure 12).

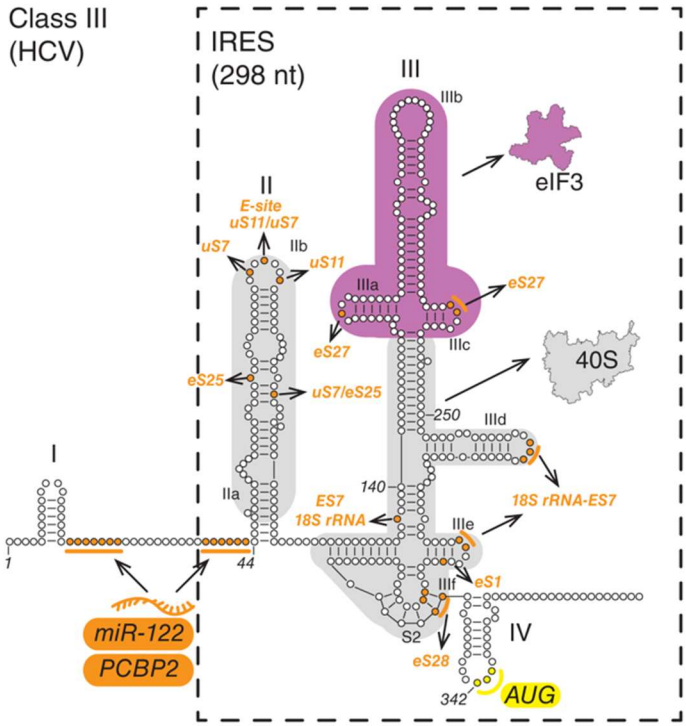


Figure 12 : structure secondaire de l'IRES de HCV. L'IRES minimum est encadré en pointillés. D'après (Mailliot and Martin 2018).

En plus du mécanisme principal qui sera expliqué ci-après, d'autres mécanismes d'initiation de la traduction guidée par l'IRES de HCV coexistent selon l'état d'avancement des mécanismes moléculaires mis en œuvre dans la cellule en réponse à l'infection. L'emploi de l'un ou l'autre mécanisme est notamment dépendant de la concentration intracellulaire en Mg^{2+}, du taux de phosphorylation de eIF2 et donc de la disponibilité des complexes ternaires eIF2-GTP-ARNt$^{Met}_i$ (voir paragraphe 1.3.1), et peut nécessiter des facteurs d'initiation alternatifs comme eIF2A ou eIF2D en coopération avec eIF3 et eIF1A (Lancaster *et al.* 2006; Pestova *et al.* 2008; Terenin *et al.* 2008; Dmitriev *et al.* 2010; Kim *et al.* 2011; Jaafar *et al.* 2016). Modérément compacte (environ 400 nucléotides) et constitué de quatre domaines structuraux, cet IRES a la particularité de fixer directement la petite sous-unité du ribosome par un réseau d'interactions avec les domaines II et III qui est responsable d'une constante d'affinité plutôt élevée : $K_a = 10^9$ (Fuchs *et al.* 2015). Le complexe de pré-initiation 48S se forme par l'intermédiaire du facteur eIF3, dont l'interaction avec la particule 40S recrutée est stabilisée par les interactions des sous-unités eIF3a et eIF3c du facteur eIF3 avec le domaine III de l'IRES (Hashem *et al.* 2013; Sun *et al.* 2013; Erzberger *et al.* 2014). Le déplacement de la sous-unité eIF3j du tunnel par lequel passe l'ARN viral est essentiel au positionnement de l'AUG de l'IRES dans le site P du ribosome. Les changements conformationnels du 48S nécessaires à l'accommodation de l'IRES sont médiés par la formation du complexe ternaire eIF2-GTP-ARNt$^{Met}_i$, couplée à l'action des facteurs eIF3, eIF1A et eIF1 par un mécanisme encore mal compris (Fraser *et al.* 2009). A ce titre, le domaine II serait impliqué dans le positionnement optimal de l'ARNt$^{Met}_i$ dans le site P. Par ailleurs, le domaine II de l'IRES est essentiel à l'assemblage du ribosome 80S, et serait ainsi impliqué dans l'hydrolyse du GTP fixé sur eIF2 et dans les mécanismes de relargage des facteurs d'initiation (Otto and Puglisi 2004; Locker *et al.* 2007).

Les IRES de classe IV assemblent directement un ribosome 80S sur le site d'initiation de la traduction sans l'assistance d'aucun facteur d'initiation, ni même de l'ARNt$^{Met}_i$ (Wilson *et al.* 2000; Pestova and Hellen 2003). Très compacts (environ 200 nucléotides), ils adoptent un repliement dont l'un des domaines se structure en un pseudo-nœud qui mime l'interaction entre la boucle anticodon d'un ARNt et un codon. Ce leurre moléculaire permet le positionnement de ce pseudo-nœud directement dans le site A du ribosome avant sa translocation dans le site P. Démarre alors la traduction de la phase codante par l'incorporation d'un ARNt dans le site A. L'IRES de la région intergénique (IGR) du virus de la paralysie du cricket (CrPV) constitue le modèle d'étude des IRES de classe IV (Figure 13), bien que de tels IRES soient trouvés dans d'autres virus de la famille des dicistrovirus.

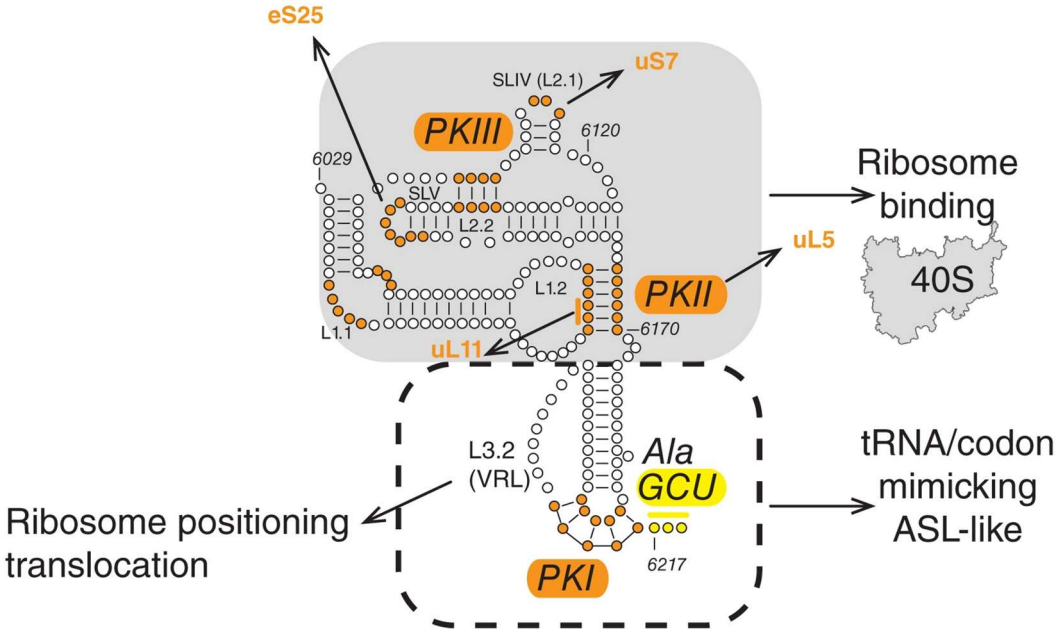


Figure 13 : structure secondaire de l'IGR du CrPV. PK : pseudo-nœud ; ASL : tige-boucle codon-anticodon ; SL : tige-boucle. D'après (Mailliot and Martin 2018).

Ces IRES se structurent généralement en trois pseudo-nœuds nommés PKI, PKII et PKIII. PKI adopte une topologie qui mime l'interaction codon-anticodon, et, avec PKII, est essentiel au recrutement des deux sous-unités du ribosome (Schüler *et al.* 2006). PKIII est quant à lui dispensable au recrutement des sous-unités du ribosome, mais est indispensable au démarrage de la traduction une fois le 80S assemblé (Jan and Sarnow 2002). L'assemblage du ribosome sur cet IRES implique soit les recrutements successifs des sous-unités 40S puis 60S, soit le recrutement direct d'un ribosome 80S (Petrov *et al.* 2016). Dans tous les cas, la présence du PKI dans le site A nécessite sa translocation dans le site P, sans quoi aucun aminoacyl-ARNt ne peut être amené au site A lors de la phase d'élongation. La translocation de PKI du site A vers le site P est médiée par le recrutement du facteur d'élongation eEF2 (voir 1.2.3.3.) au niveau du site A (Yamamoto *et al.* 2007). Le site A désormais vacant, la phase d'élongation peut avoir lieu dans la phase +0 ou +1 (Ren *et al.* 2012). Cette propriété peut être imputée à la relative flexibilité du PKI. La sélection du cadre de lecture est dépendante de la conformation du PKIII pour le positionnement du PKI dans le site A du ribosome (Au *et al.* 2015).

Les propriétés de l'IRES de EMCV et de l'IGR du CrPV seront particulièrement utilisées dans ce travail. Par ailleurs, une partie du travail de thèse est dédiée à la mise au point d'une méthode d'identification systématique d'IRES dans un génome viral (2.2).

Les IRES cellulaires

D'abord cantonnés aux ARN viraux, les mécanismes d'initiation coiffe-indépendants sont aussi impliqués dans la traduction de certains ARNm cellulaires et font partie des stratégies sélectionnées au cours de l'évolution permettant d'adapter rapidement le protéome d'une cellule aux changements environnementaux qui génèrent un stress cellulaire, dont une des conséquences est en général l'inhibition du mécanisme d'initiation canonique de la traduction.

A l'instar de certains IRES viraux, leur traduction nécessite des facteurs protéiques auxiliaires nommés ITAF (IRES *trans*-acting factors). Les mécanismes relatifs à la traduction des ARNm cellulaires en condition de stress seront abordés dans la partie 1.4, et permettront d'évoquer le concept de traduction sélective des ARN messagers (1.5).

1.2.3. La phase d'élongation des protéines

La phase d'élongation (Figure 14) démarre lorsque le ribosome 80S est assemblé sur le codon d'initiation avec l'ARNt$^{Met}_{i}$ situé dans le site P par l'intermédiaire de l'interaction codon-anticodon, les sites E et A étant vacants. L'avancement du ribosome sur l'ARN messager de l'extrémité 5' vers l'extrémité 3' consiste en la répétition successive, codon par codon, des trois étapes décrites dans les paragraphes suivants (Dever *et al.* 2018).

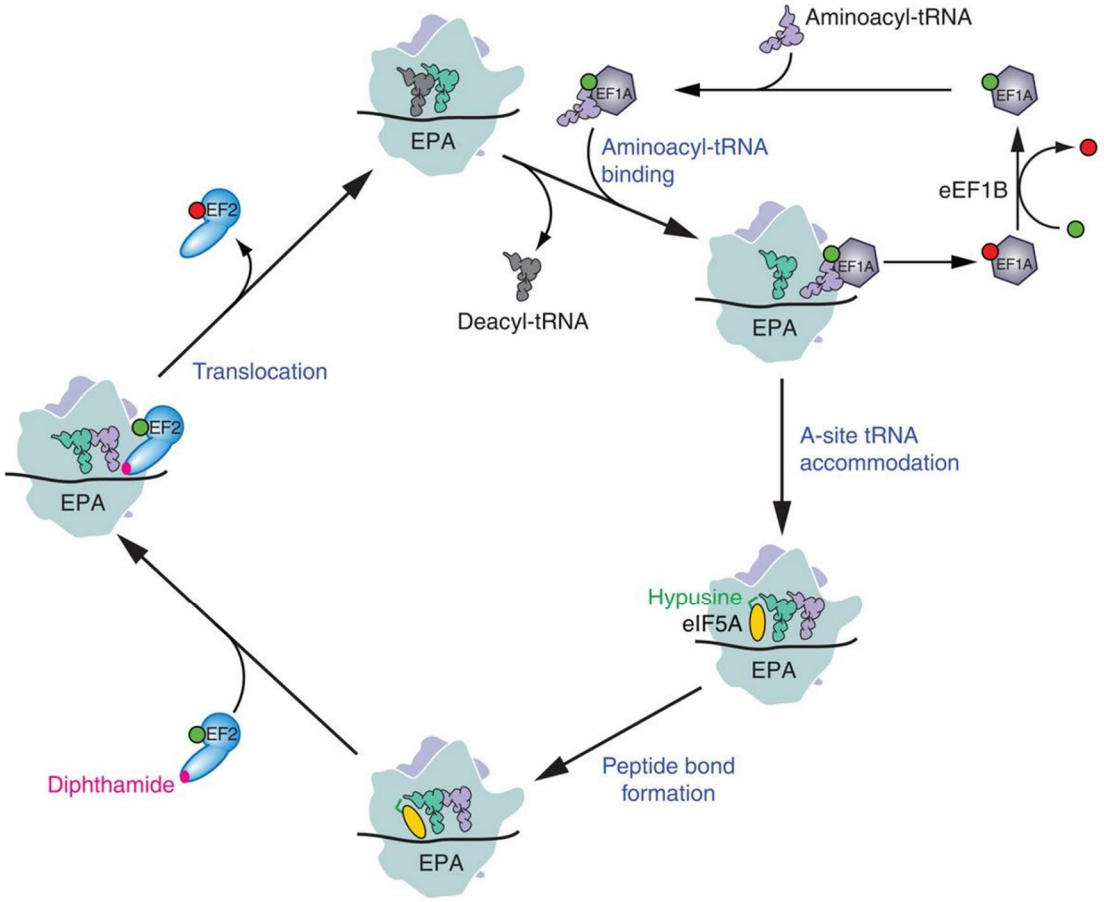


Figure 14 : vue de l'ensemble des mécanismes impliqués dans l'étape d'élongation de la traduction. Le GTP est représenté par un point vert et le GDP par un point rouge. D'après (Dever *et al.* 2018).

1.2.3.1. Accommodation de l'amino-acyl-ARNt dans le site A du ribosome

La diffusion moléculaire fait successivement entrer les complexes ternaires eEF1A/GTP/aa-ARNt au niveau du codon localisé dans le site A du ribosome jusqu'à l'accommodation d'un aa-ARNt. Seule une interaction codon-anticodon suffisamment stable permet l'accommodation de l'aa-ARNt dans le site A et l'activation subséquente de l'hydrolyse du GTP de eIF1A qui entraine son relargage.

La reconnaissance du codon par l'aa-ARNt repose sur les liaisons hydrogènes engagées dans l'interaction codon-anticodon, mais également sur celles engagées avec trois nucléotides clés du ribosome dans l'hélice h44 de l'ARNr 18S (A1824, A1825, G626 du ribosome de lapin) (Ogle *et al.* 2001). Ces trois nucléotides modulent la géométrie de l'hélice formée par les liaisons hydrogènes entre les deux premières bases du codon et les deux dernières de l'anticodon (Loveland *et al.* 2017). Ce faisant, ils permettent de discriminer les interactions codon-anticodon invalides (pas d'appariements Watson-Crick) des interactions valides (appariements Watson-Crick parfaits) ou partiellement valides (appariements Watson-Crick et mésappariements tolérés à ce stade), et constituent ainsi un premier garant de la fidélité de décodage sur les deux premières bases du codon. La tolérance est moins stricte au niveau de la troisième position du codon et des appariements non Watson-Crick y sont possibles : c'est la position 'Wobble' dont une conséquence est la dégénérescence du code génétique. A ce stade, les aa-ARNt invalides ne sont pas stabilisés dans le site de décodage et par conséquent n'y résident pas, permettant ainsi l'entrée d'autres aa-ARNt qui seront testés de la même manière. La stabilisation de l'interaction codon-anticodon par le nucléotide G626 (Loveland *et al.* 2017) entraine des changements conformationnels majeurs au sein du ribosome et aboutissent à l'activation de l'hydrolyse du GTP lié à eEF1A par l'interaction entre d'une part la boucle sarcine-ricine de la grande sous-unité du ribosome et d'autre part la Switch 2 loop du domaine GTPase de eIF1A, ce qui permet le positionnement optimal d'un résidu histidine nécessaire pour la catalyse (Shao *et al.* 2016). La cinétique d'hydrolyse du GTP de eEF1A est dépendante des changements conformationnels induits par l'interaction codon-anticodon et est dès lors impliquée dans la discrimination des interactions codon-anticodon valides des partiellement valides. La vitesse de l'hydrolyse détermine en effet si l'aa-ARNt s'accommode définitivement dans le site A tout en se dissociant de eIF1A-GDP ou s'il est éjecté du ribosome conjointement avec eEF1A-GDP. L'hydrolyse différentielle du GTP constitue ainsi un second garant de la fidélité de décodage. Enfin, des modifications de bases de l'anticodon des ARNt permettent d'augmenter la stabilité de l'interaction avec le codon, ce qui constitue un troisième garant de la fidélité de décodage. Ceci est particulièrement important pour les combinaisons codon-anticodon qui sont riches en nucléotides A et U (Pernod *et al.* 2020).

A l'instar de eIF2-GDP avec eIF2B, le facteur eEF1A-GDP est recyclé en eEF1A-GTP par l'intermédiaire d'une protéine à activité GEF : eEF1B (Gromadski *et al.* 2007).

1.2.3.2. Formation de la liaison peptidique

La catalyse de la liaison peptidique est un mécanisme strictement conservé dans les trois domaines du vivant. Suite au relargage de eEF1A-GDP et à l'accommodation de l'aa-ARNt dans le site A, le domaine peptidyl-transférase de la grande sous-unité du ribosome assure le positionnement optimal du peptidyl ARNt et de l'amino-acyl ARNt situés respectivement dans les sites P et A. Dans le cas particulier du décodage du deuxième codon, l'ARNt présent dans le site P est l'ARNt$^{Met}_{i}$ et non un peptidyl-ARNt. Le facteur eIF5A, en se fixant dans le site E, favorise le bon positionnement du bras accepteur de l'ARNt du site P par rapport à celui du site A, et de ce fait accélère la formation de la liaison peptidique (Gutierrez *et al.* 2013 ; Melnikov *et al.* 2016 ; Schmidt *et al.* 2016). Cette liaison est le résultat de l'amidation du carbone du groupement acide carboxylique de l'extrémité C-terminale du peptide naissant par le groupement α-amine de l'acide aminé lié à l'aa-ARNt du site A (Figure 15). Ensuite, le pivotement de la tête de la petite sous-unité du ribosome repositionne l'ARNt du site P désormais déacylé à cheval sur les sites E et P et le nouvellement formé peptidyl-ARNt à

cheval sur les sites A et P, ce qui amorce la translocation (Ratje *et al.* 2010; Budkevich *et al.* 2011; Behrmann *et al.* 2015).

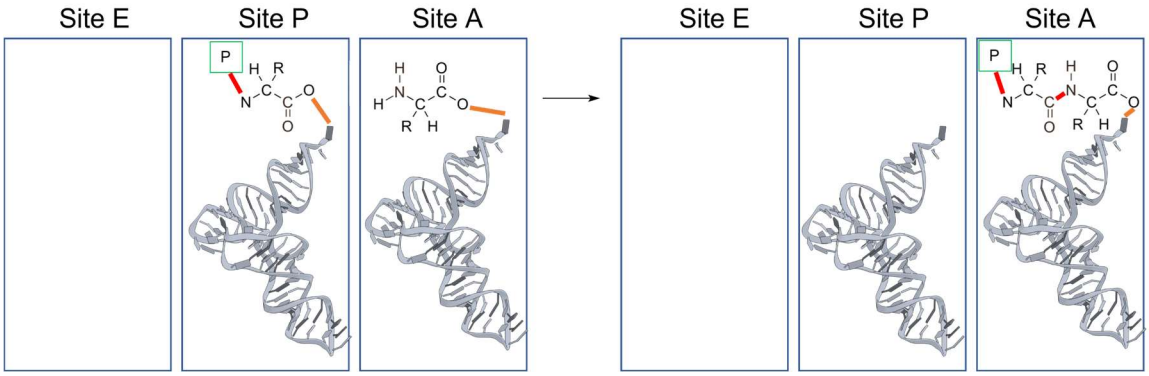


Figure 15 : schéma représentant la formation de la liaison peptidique. R correspond au radical des acides aminés. P correspond au peptide naissant. Les liaisons peptidiques sont indiquées en rouge, les liaisons esters reliant les acides-aminés à l'extrémité 3'-CCA des ARNt sont indiquées en orange.

1.2.3.3. Translocation du ribosome et relargage de l'ARNt déacylé

La translocation de l'ARNt déacylé dans le site E et du peptidyl-ARNt dans le site P est dépendante de l'énergie apportée par l'hydrolyse du GTP lié au facteur eEF2 qui se fixe dans le site A vacant. Bien que son mode d'action ne soit pas entièrement élucidé, l'hypothèse actuelle est que cette hydrolyse finalise les mouvements de rotations de la tête de la petite sous-unité du ribosome précédemment amorcés en levant les interactions entre les nucléotides clés de l'hélice h44 de l'ARNr 18S et l'interaction codon-anticodon impliquant le nouveau peptidyl-ARNt, permettant ainsi leur progression au sein du ribosome (Spahn *et al.* 2004; Taylor *et al.* 2007; Abeyrathne *et al.* 2016; Murray *et al.* 2016). Ces changements conformationnels aboutissent au relargage de eEF2-GDP et au positionnement précis de l'ARNt déacylé et du peptidyl-ARNt dans les sites E et P respectivement. Il est essentiel que les interactions codon-anticodon restent stables pendant ce mouvement, faute de quoi il y aurait un risque de changement de cadre de lecture, et donc d'une protéine au moins tronquée du côté C-terminal, sinon non fonctionnelle. En ce sens, des études suggèrent qu'une modification post-traductionnelle d'une histidine de eEF2 en diphthamide est essentielle pour la fidélité du décodage chez les eucaryotes car son absence augmenterait la probabilité de décalage de cadre de lecture au moment de la translocation (Abeyrathne *et al.* 2016; Murray *et al.* 2016). Enfin, le relargage de l'ARNt déacylé du site E est essentiel pour l'arrivée de eIF5A. Il a également été proposé que l'arrivée de eIF5A dans le site E accélère la sortie de l'ARNt déacylé qui s'y trouve lors de l'accommodation d'un nouvel aa-ARNt dans le site A.

1.2.4. La terminaison de la traduction

Les trois étapes précédentes se répètent successivement, codon par codon, jusqu'à la présence d'un des trois codons stop canoniques (UAA, UAG, UGA) dans le site A. Etant donné qu'aucun ARNt de la cellule ne peut s'apparier avec les codons stop, le site A reste vacant, ce qui enclenche une série d'évènements menant à la dissociation des sous-unités du ribosome ainsi qu'au relargage de la protéine synthétisée (Hellen 2018). Les trois codons stop canoniques (à quelques organismes près) sont reconnus dans le site A vacant par le domaine

N-terminal de eRF1 (Bertram *et al.* 2000) dont le repliement tridimensionnel s'apparente à celui des ARNt (Song *et al.* 2000). La fidélité de reconnaissance du codon stop par eRF1 est dépendante d'une part du codon stop et des nucléotides +4 et +5 en aval, en considérant le U du codon stop comme le +1 (McCaughan *et al.* 1995), et d'autre part des motifs GTS et YxCxxxF de eRF1 (Brown *et al.* 2015). Le relargage de la protéine néosynthétisée est dépendant de l'activité GTPase de eRF3-GTP, un facteur qui interagit avec le domaine C-terminal de eRF1. L'interaction de eRF3 avec la boucle sarcine-ricine du ribosome entraîne l'hydrolyse de sa molécule de GTP et permet l'accommodation de eRF1 dans le site A en positionnant son motif conservé GGQ en face du bras accepteur du peptidyl-ARNt dans le site P (Shao *et al.* 2016). Le motif GGQ catalyse l'hydrolyse de la liaison ester reliant l'extrémité 3'CCA de l'ARNt à la protéine synthétisée qui est ainsi relarguée dans le tunnel du ribosome. Une étude a démontré l'implication du facteur eIF5A qui augmente la vitesse de cette hydrolyse (Schuller *et al.* 2017). Après la terminaison, le recyclage des sous-unités ribosomiques et des facteurs eRF1 et eRF3 repose sur l'interaction entre eRF1 et ABCE1 (Pisarev *et al.* 2010). Après la dissociation, l'ARNt du site P et l'ARNm restent fixés à la petite sous-unité du ribosome et sont décrochés par la fixation des facteurs d'initiation eIF3, eIF1A et eIF1 pour la formation de nouveaux complexes d'initiation (Pisarev *et al.* 2010).

1.2.5. Localisation cellulaire de la traduction

Chez les eucaryotes, la traduction a exclusivement lieu dans le cytosol, et peut être associée dans certains cas aux membranes de certains organites comme le réticulum endoplasmique ou le noyau, ou à certains organelles dépourvus de membranes comme les granules de stress. Pour certains types cellulaires comme les neurones, la spatialisation de la traduction des ARNm est essentielle à leur fonction.

1.2.5.1. La spatialisation de la traduction est dépendante du type cellulaire

C'est dans les cellules spécialisées que la spatialisation et la répartition de la traduction des ARNm sont les plus marquées (Martin and Ephrussi 2009). Dans les fibroblastes, la traduction des ARNm de la β-actine est localisée au niveau des lamellipodes et est essentielle au remodelage du cytosquelette qui dans ce cas permet la migration cellulaire. Dans les neurones matures, la traduction des ARNm localisés dans la partie terminale des axones participe largement à la plasticité synaptique (Figure 16). La traduction localisée de l'ARNm codant pour le facteur de transcription VegT est déterminante dans l'induction du mésoderme lors du développement embryonnaire du Xénope, qui est l'organisme modèle pour ce type d'études.

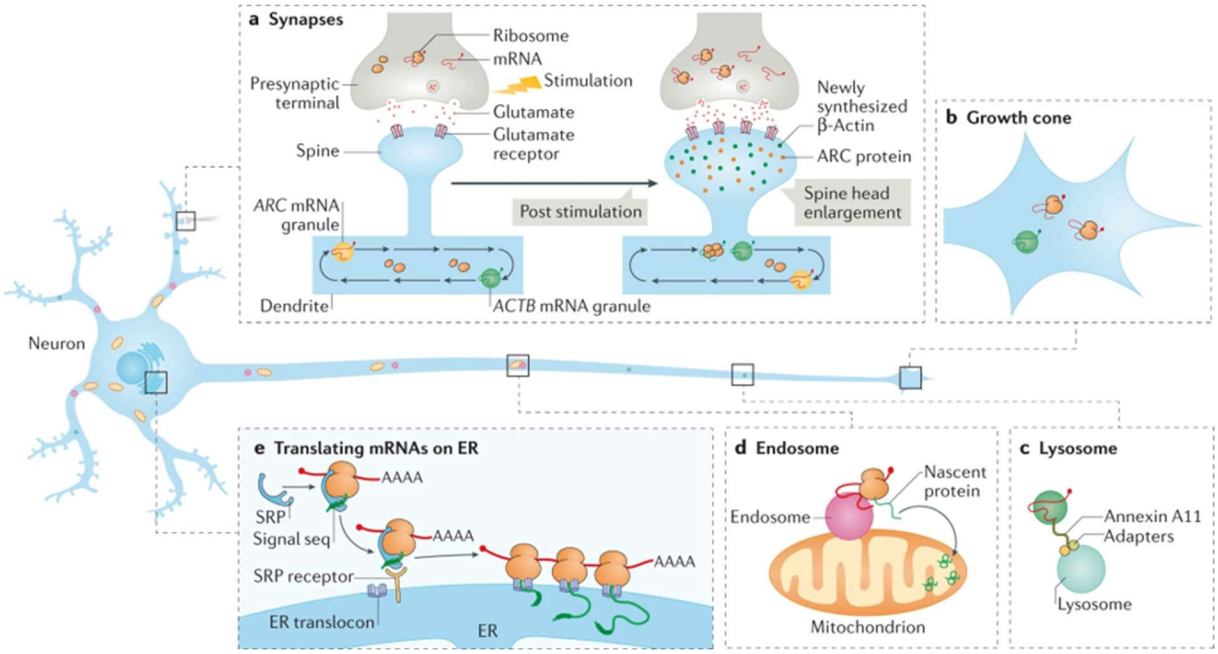


Figure 16 : illustration de l'importance de la localisation cellulaire de la traduction des ARNm pour la fonctionnalité des neurones. Le transport de certains ARNm et leur traduction localisée est essentielle pour la réception de signaux au niveau des parties post-synaptiques comme les épines dendritiques (**a**) et au développement des parties pré-synaptiques à l'extrémité des axones (**b**). Le transport des ARNm dans les neurones peut faire intervenir leur association aux lysosomes (**c**). Certaines protéines mitochondriales codées par les gènes nucléaires sont adressées aux mitochondries par leur traduction associée aux endosomes (**d**). La fonctionnalité des synapses est largement dépendante des protéines transmembranaires et excrétées dont la maturation dépend de la traduction localisée de leurs ARNm au niveau du réticulum endoplasmique (**e**). D'après (Das *et al.* 2021)**.**

Le transport des ARNm vers une localisation cellulaire essentielle à la fonction de la protéine codée implique la formation de granules d'ARNm transportés sur le cytosquelette par des protéines motrices. Les motifs nucléotidiques qui permettent l'assemblage de ces granules sont généralement localisés dans les régions 3'UTR des ARNm.

1.2.5.2. La traduction associée à la membrane du réticulum endoplasmique est essentielle à la maturation des protéines membranaires et sécrétées

La traduction des ARNm associée à la membrane du réticulum endoplasmique est essentielle à la maturation de la protéine synthétisée et concerne en particulier la totalité des protéines membranaires et des protéines sécrétées (Reid and Nicchitta 2015). Dans les deux cas, ce sont des protéines fondamentales pour la communication intra- et inter- cellulaires et donc pour le fonctionnement global de l'organisme, comme cela a été précédemment illustré avec les neurones (Figure 16). Près d'un tiers des protéines synthétisées est concerné (Pechmann *et al.* 2014). L'association de la traduction d'une protéine en cours de synthèse à la membrane du réticulum endoplasmique dépend du repliement co-traductionnel du peptide naissant en un motif N-terminal hydrophobe qui est reconnu par la particule de reconnaissance du signal, abrégée SRP (Walter *et al.* 1981; Walter and Blobel 1981). L'interaction entre la SRP et un récepteur SRP situé dans la membrane du réticulum endoplasmique y associe le complexe traductionnel et libère le motif hydrophobe initialement reconnu par la SRP. S'ensuit alors

l'invagination co-traductionnelle de la protéine en cours de synthèse dans la lumière ou la membrane (selon la protéine) du réticulum endoplasmique après l'ancrage du motif hydrophobe dans le translocon, une protéine transmembranaire reliant le cytoplasme à la lumière du réticulum endoplasmique. Commence ainsi la maturation de la protéine dans le réticulum endoplasmique puis dans l'appareil de Golgi, où elle fera l'objet de nombreuses modifications post-traductionnelles nécessaires à leur(s) fonction(s) dans la cellule.

1.2.5.3. La traduction associée aux granules de stress

La traduction des ARNm associée aux granules de stress (Mateju *et al.* 2020) peut être interprétée comme une stratégie pour concentrer localement les différents facteurs protéiques nécessaires à la traduction des ARNm traduits spécifiquement dans ces conditions.

Ces quelques exemples montrent que l'intégration fonctionnelle d'une cellule au sein d'un organisme passe également par la localisation cellulaire précise de la traduction de l'ensemble de ses ARN messagers.

1.3. Mécanismes *trans*-régulateurs de la traduction

Les trois étapes de la traduction sont sujettes à des mécanismes de régulation faisant intervenir de manière plus ou moins concertée divers facteurs agissant en *trans* qui peuvent être des ARN non-codants, des facteurs protéiques, certaines petites molécules, ainsi que la séquence de la protéine en cours de synthèse. Ces mécanismes sont largement dépendants de l'environnement et de la signalisation cellulaires.

Dans cette partie, l'accent sera mis sur la régulation de l'initiation de la traduction. Par souci d'exhaustivité, les régulations liées aux phases d'élongation et de terminaison seront brièvement évoquées.

Bien qu'indissociables des aspects *trans*-régulateurs de la traduction qui font l'objet de ce paragraphe, les éléments *cis*-régulateurs localisés dans la séquence des ARNm seront discutés dans le paragraphe 1.4.

1.3.1. Certains facteurs d'initiation, d'élongation ou de terminaison sont des cibles d'effecteurs de voies de signalisation

La plupart des mécanismes moléculaires mis en jeu dans la traduction est reliée à des voies de signalisation dont certains facteurs d'initiation, d'élongation ou de terminaison de la traduction sont des cibles. La traduction est ainsi directement connectée aux changements environnementaux captés par la cellule. L'intégration directe de ces signaux au niveau traductionnel permet la synthèse rapide d'un nouvel ensemble de protéines qui permet d'orchestrer la réponse adaptative de la cellule en réponse à ces changements.

1.3.1.1. La voie mTOR (mammalian target of rapamycin)

La voie mTOR (Figure 17) est liée à l'état énergétique et à l'environnement nutritionnel de la cellule et contrôle la synthèse des biomolécules selon les ressources cellulaires disponibles. Parmi ses nombreuses cibles figurent notamment des protéines impliquées dans la traduction. Le terme mTOR réfère au domaine sérine/thréonine kinase de deux complexes protéiques cytosoliques nommés complexes mTOR-1 (mTORC1) et mTOR-2 (mTORC2). Le premier est en lien avec la régulation de la biosynthèse des protéines, des lipides ou des nucléotides, tandis que le second est plutôt lié à la régulation de l'activité de protéines kinases impliquées dans le maintien de l'intégrité cellulaire et dans le remodelage du cytosquelette (Saxton and Sabatini 2017).

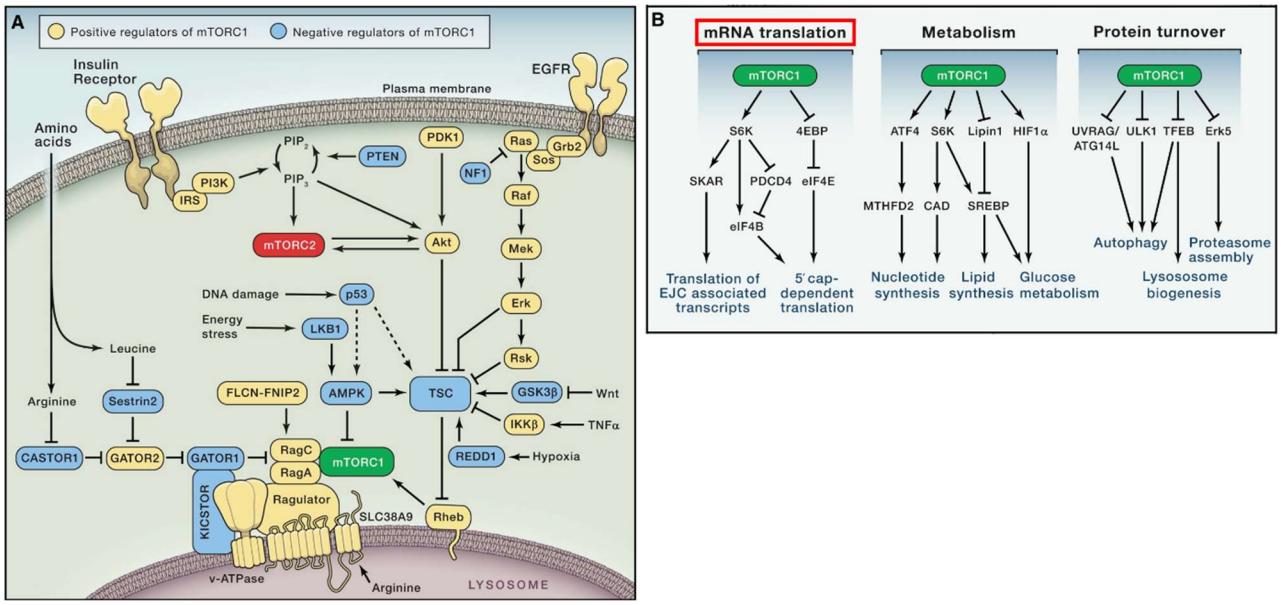


Figure 17 : vue d'ensemble de la voie mTOR. A. Les voies de signalisation en amont de mTOR qui mènent à leur activation ou à leur inactivation**. B.** Les voies de signalisation en aval de mTORC1, qui découlent de son activation ou de son inhibition. D'après (Saxton and Sabatini 2017).

mTORC1

Le complexe mTORC1 est inactivé dans des situations de déficit en acides aminés ou dans des conditions d'hypoxie qui entrainent la baisse du rendement de production d'ATP par la chaine respiratoire mitochondriale, comme c'est notamment le cas dans les cellules tumorales peu voire pas vascularisées. L'inactivation de mTORC1 entraine l'inhibition de la traduction coiffe-dépendante cellulaire et un ralentissement du métabolisme, en réponse à ce déficit énergétique.

Activé, le complexe mTORC1 phosphoryle la protéine de fixation à eIF4E (4E-BP), ce qui réduit son affinité pour eIF4E et favorise ainsi la traduction coiffe-dépendante (Brunn *et al.* 1997). L'inactivation de mTORC1 entraine une baisse du taux de phosphorylation de 4E-BP, et donc l'augmentation de son affinité pour eIF4E. Ainsi séquestré, le facteur eIF4E ne peut plus former le complexe d'initiation eIF4F, inhibant ainsi la traduction coiffe-dépendante.

Une seconde catégorie de substrat de mTORC1 est constituée par les protéines S6 kinases 1 (S6K1). Phosphorylées, les S6K1 phosphorylent à leur tour des facteurs influençant l'initiation de la traduction coiffe-dépendante. En particulier, la phosphorylation de eIF4B favorise l'initiation de la traduction en augmentant l'activité hélicase de eIF4A (Holz *et al.* 2005). Par ailleurs, les S6K1 phosphorylent et dans ce cas inhibent les protéines PDCD4 qui, activées, inhibent eIF4B (Dorrello *et al.* 2006) et eEF2K, la kinase spécifique du facteur d'élongation eEF2 (Wang 2001). La phosphorylation de eEF2 par eEF2K sur la thréonine 56 inhibe la translocation des ARNt et bloque ainsi la traduction.

En revanche, des facteurs de croissance et diverses hormones permettent l'activation de mTORC1 par l'intermédiaire des voies sous contrôle des kinases phosphoinositide 3-kinase (PI3K)-AKT et Ras-ERK (extracellular signal-regulated kinase). Ainsi, mTORC1 assure le

maintien de la synthèse des protéines lors de la croissance et de la division cellulaire (Ma and Blenis 2009).

mTORC2

Enfin, plusieurs travaux suggèrent l'implication de mTORC2 dans la maturation du peptide naissant lors de la phase d'élongation (Oh *et al.* 2010; Zinzalla *et al.* 2011).

1.3.1.2. La voie ISR (Integrated Stress Response)

En réponse aux nombreux changements environnementaux dont certains peuvent être considérés comme des stress, la cellule active une série de cascades de signalisation appelée voie ISR (Intergrated Stress Response) qui convergent toutes vers un effecteur commun : le facteur d'initiation de la traduction eIF2.

La phosphorylation de la sous-unité α de eIF2 (eIF2α) sur la sérine 51 est au cœur de la régulation traductionnelle dirigée par l'ISR. Quatre voies utilisent des sérine/thréonine kinases susceptibles de réaliser cette phosphorylation selon le type de stress : PKR, PERK, HRI, GCN2 (Figure 18).

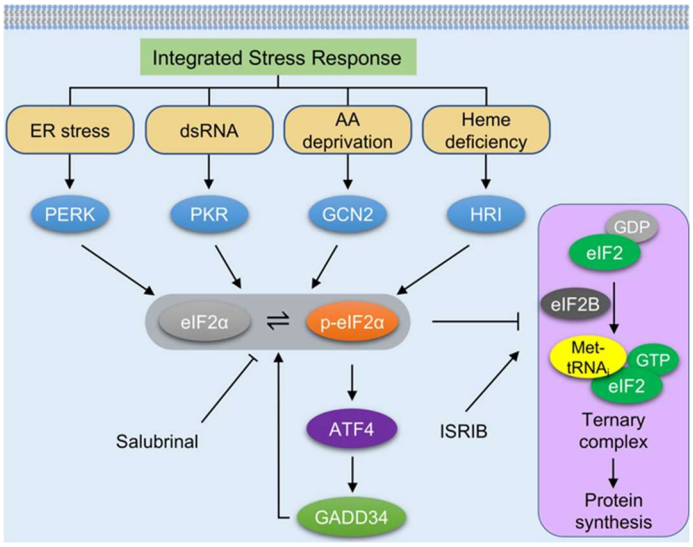


Figure 18 : schéma illustrant les quatre cascades de signalisation menant à phosphorylation de la sous-unité eIF2α lors de l'ISR. ISRIB est une molécule qui améliore l'activité GEF de eIF2B et de ce fait favorise le recyclage de eIF2 (Zyryanova *et al.* 2018). Le salubrinal est une molécule qui inhibe la déphosphorylation de eIF2 (Boyce *et al.* 2005). ER : réticulum endoplasmique ; AA : acide-aminé ; dsRNA : ARN double-brin. D'après (Zhang *et al.* 2022).

Le facteur de transcription ATF4, dont l'ARNm est spécifiquement traduit en présence de eIF2 phosphorylé (voir Figure 28), est l'un des principaux orchestrateurs de l'ISR car il remodèle le transcriptome de la cellule en réponse au stress.

La phosphorylation de eIF2α provoque l'inhibition de la traduction coiffe-dépendante par un mécanisme qui implique eIF2B, le facteur permettant le recyclage du complexe ternaire eIF2-GTP-ARNt$^{Met}_{i}$ (Figure 19). La phosphorylation de eIF2 est possible lorsqu'il est engagé dans le complexe ternaire, qu'il soit lié au GTP ou au GDP, mais aussi lorsqu'il est sous sa forme libre. Lorsque eIF2 est phosphorylé dans le complexe ternaire, il peut toujours former un

complexe 43S d'initiation de la traduction, à la suite de laquelle il sera relargué du complexe d'initiation sous sa forme libre liée au GDP. La forme phosphorylée de eIF2-GDP a une meilleure affinité pour eIF2B que la forme non-phosphorylée, et inhibe son activité GEF (Sudhakar *et al.* 2000; Krishnamoorthy *et al.* 2001; Kashiwagi *et al.* 2019). En séquestrant eIF2B de la sorte, la phosphorylation de eIF2α inhibe la formation de nouveaux complexes ternaires eIF2-GTP-ARNt$^{Met}_{i}$, essentiels à l'initiation de la traduction coiffe dépendante.

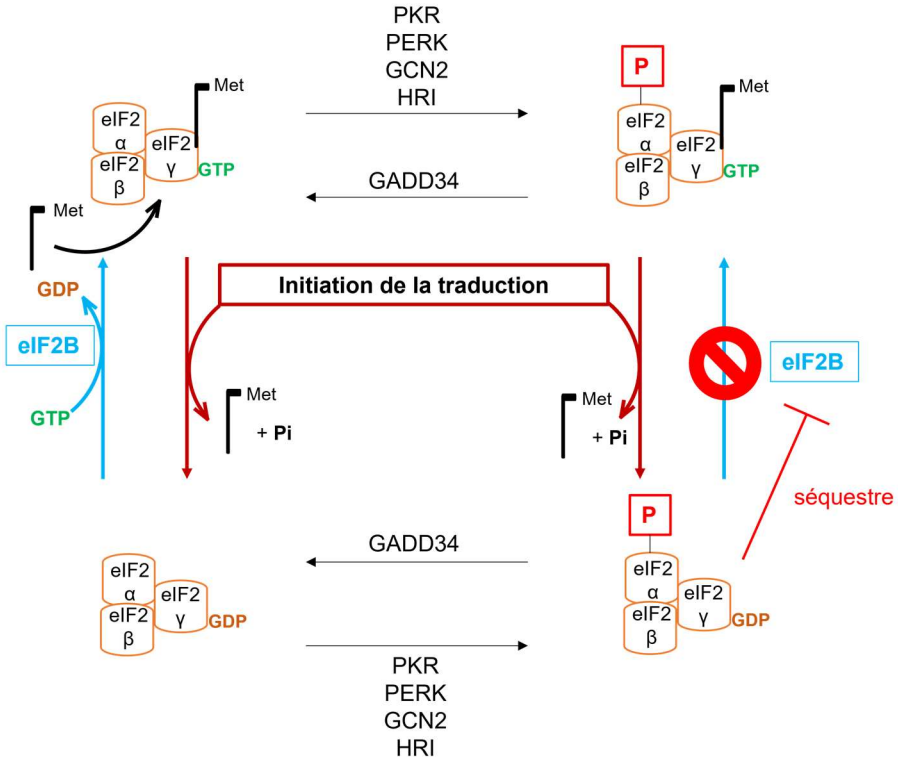


Figure 19 : mécanisme d'inhibition de la traduction coiffe-dépendante cellulaire par la phosphorylation de la sous-unité eIF2α lors de l'ISR.

A ce jour, quatre sérine/thréonine kinases de eIF2α ont été identifiées et caractérisées.

HRI

Son rôle a initialement été caractérisé dans les érythrocytes comme étant le principal coordinateur des taux de α- et β- globines, d'hème et de fer, qui, en s'assemblant, forment l'hémoglobine. L'activation de la HRI passe par une succession d'autophosphorylations qui sont proportionnellement dépendantes de la concentration en hème (Igarashi *et al.* 2011). Ainsi, une faible concentration cellulaire en hème mène à la phosphorylation de eIF2α par la HRI et au ralentissement de la synthèse des constituants de l'hémoglobine.

Le rôle de la HRI n'est cependant pas limité aux globules rouges et son implication a aussi été démontrée dans la maturation des macrophages (Liu *et al.* 2007).

GCN2

A l'instar de mTORC1, la protéine GCN2 est un senseur de la concentration en acides aminés dans la cellule, mais, contrairement à mTORC1, elle est activée par leur faible abondance. GCN2 est donc une protéine essentielle à la réponse au stress métabolique en condition de

carence en acides aminés. La fixation de GCN2 à des ARNt non chargés provoque sa dimérisation et son autophosphorylation (Narasimhan *et al.* 2004). Ainsi activée, GNC2 phosphoryle eIF2α. Par ailleurs, l'activation de GCN2 contribue indirectement au maintien de l'inactivation de mTORC1 par l'intermédiaire de la protéine Sestrin2, dont la transcription du gène correspondant nécessite le facteur de transcription ATF4 dont l'ARNm est spécifiquement traduit lorsque eIF2 est phosphorylé (Ye *et al.* 2015).

PKR

D'abord caractérisée comme un senseur d'ARN double-brin d'origine virale, la protéine PKR est activée par une série d'autophosphorylations suite à la reconnaissance de portions d'ARN double-brin, qu'ils soient d'origine exogènes ou endogènes. Cette reconnaissance est dépendante de la taille et de la conformation des structures adoptées par les ARN. Il a également été démontré que l'ARN simple-brin peut interagir avec la PKR et l'activer (Mayo and Cole 2017). L'activation de la PKR est contrôlée par de multiples acteurs ribonucléiques et protéiques. Le fait que des ARN exogènes et endogènes puissent mener à l'activation de la PKR et à l'inhibition de la traduction cellulaire coiffe-dépendante par la phosphorylation subséquente de eIF2α pose la question de la sélectivité d'interaction entre la PKR et ses interacteurs, et en particulier entre les molécules du soi et du non-soi. D'autres études structurales sont nécessaires pour élucider le « code PKR » (Bou-Nader *et al.* 2019).

PERK

La protéine PERK, avec les protéines IRE1α et ATF6, fait partie des trois protéines transmembranaires du réticulum endoplasmique impliquées dans la réponse au stress provoqué par la présence de protéines mal repliées dans la lumière du réticulum (UPR pour Unfolded Protein Response). Cette voie de signalisation est fondamentale pour le maintien de l'homéostasie cellulaire, dont le dérèglement peut induire l'apoptose (Hetz *et al.* 2020). L'activation de ces trois protéines senseurs passe par la détection d'aberrations conformationnelles de protéines présentes dans la lumière du réticulum endoplasmique par un jeu d'interactions complexes avec des protéines chaperonnes ainsi que les protéines mal repliées. eIF2 étant situé dans le cytosol, l'activation de la PERK dans la lumière du réticulum endoplasmique mène à la phosphorylation de eIF2α par son domaine cytosolique.

1.3.1.3. Les facteurs de traduction non-canoniques impliqués dans la traduction des ARNm cellulaires en conditions de stress

Les voies mTOR et ISR ont en commun de mener à l'inhibition de la traduction coiffe-dépendante cellulaire, soit en inactivant eIF4E, soit en inactivant le complexe ternaire eIF2-GTP-ARNt$^{Met}_{i}$. Cela ne signifie pas pour autant l'arrêt total de la traduction, et certains ARNm dont la protéine résultante est impliquée dans la réponse au stress sont spécifiquement traduits malgré l'inactivation de ces facteurs. Interviennent alors des facteurs de traduction dits non-canoniques, impliqués dans l'initiation de la traduction de ces ARNm.

Un exemple est la traduction de l'ARNm codant pour le facteur de transcription ATF4 qui est responsable de la transcription de gènes impliqués dans la régulation du métabolisme des acides aminés, la résistance au stress oxydant ou qui sont essentiels dans la coordination de la réponse aux protéines mal repliées. L'ensemble des gènes activés par ATF4 constitue la réponse intégrée au stress (ISR). Il a été démontré que la traduction de l'ARNm d'ATF4 est dépendante des facteurs d'initiation non-canoniques tels que eIF2D et DENR (Bohlen *et al.*

2020; Vasudevan *et al.* 2020), qui seraient impliqués dans l'acheminement de l'ARNt initiateur vers le complexe de pré-initiation et dans des mécanismes de ré-initiation.

D'autres facteurs d'initiation non-canoniques seraient impliqués dans l'acheminement de l'ARNt initiateur au complexe de pré-initiation lorsque eIF2 est phosphorylé. Le facteur eIF5B a été caractérisé comme essentiel dans la traduction non-canonique de certains ARNm codant pour des protéines anti-apoptotiques, comme XIAP, Bcl-xL, cIAP1, et c-FLIP$_S$ (Thakor and Holcik 2012; Ross *et al.* 2019). Le facteur eIF2A est également impliqué dans des mécanismes de traduction non-canoniques en réponse à la phosphorylation de eIF2α. Il est en particulier essentiel à la traduction de l'IRES de HCV en conditions de stress induit par l'infection virale (Kim *et al.* 2011). Bien que leur implication dans la réponse au stress cellulaire et plus particulièrement dans les mécanismes de traduction non-canoniques soit clairement démontrées, le mode d'action précis de ces facteurs d'initiation non-canoniques n'est pas encore totalement élucidé (Komar and Merrick 2020).

Un autre exemple est la phosphorylation, induite par un stress métabolique, de la sous-unité eIF3d de eIF3 qui est impliquée dans la traduction non-canonique de l'ARNm codant pour le facteur de transcription *c-Jun* (Lamper *et al.* 2020). Etant responsable de l'activation de gènes impliqués dans la prolifération et la différenciation cellulaires, ainsi que dans l'apoptose, c'est un facteur déterminant pour la fonctionnalité d'une cellule. Un tel stress induit l'inactivation de eIF4E par 4E-BP via mTORC1 et la phosphorylation de eIF3d augmente son affinité de reconnaissance pour la coiffe m^{7}G. La traduction de *c-Jun* est strictement dépendante de la forme phosphorylée de eIF3d, dans la mesure où des structures secondaires de son extrémité 5' bloquent la fixation du complexe eIF4F et de ce fait inhibent sa traduction par un mécanisme canonique (Lee *et al.* 2016).

Enfin, les voies mTOR et ISR sont simultanément mises à contribution dans la réponse de la cellule aux changements environnementaux et ne sont donc pas mutuellement exclusives, ce qui illustre la complexité des mécanismes de réponse aux stress. Par exemple, l'activation de 4E-BP par l'intermédiaire d'ATF4, spécifiquement traduit après la phosphorylation de eIF2 par GCN2, permet la traduction sélective d'ARNm codant pour des peptides antimicrobiens lors d'une infection bactérienne chez la drosophile (Vasudevan *et al.* 2017).

1.3.2. Certains pathogènes produisent des protéines qui manipulent la machinerie traductionnelle de la cellule

Dans cette partie, l'accent sera mis sur la traduction cellulaire lors d'une infection virale.

1.3.2.1. Détournement de la machinerie traductionnelle cellulaire par des protéines virales

Des protéines virales peuvent influencer plusieurs étapes de la traduction des ARNm cellulaires selon le stade de l'infection virale (Walsh and Mohr 2011). La régulation spatiale et temporelle de l'activité des différents facteurs de traduction est critique pour l'aboutissement de l'assemblage de particules virales infectieuses et fait souvent intervenir la modulation des voies mTOR et ISR précédemment évoquées. Les paragraphes à suivre illustrent quelques stratégies mises en place lors de l'infection par certains virus (Figure 20). Ces stratégies ne sont pas mutuellement exclusives.

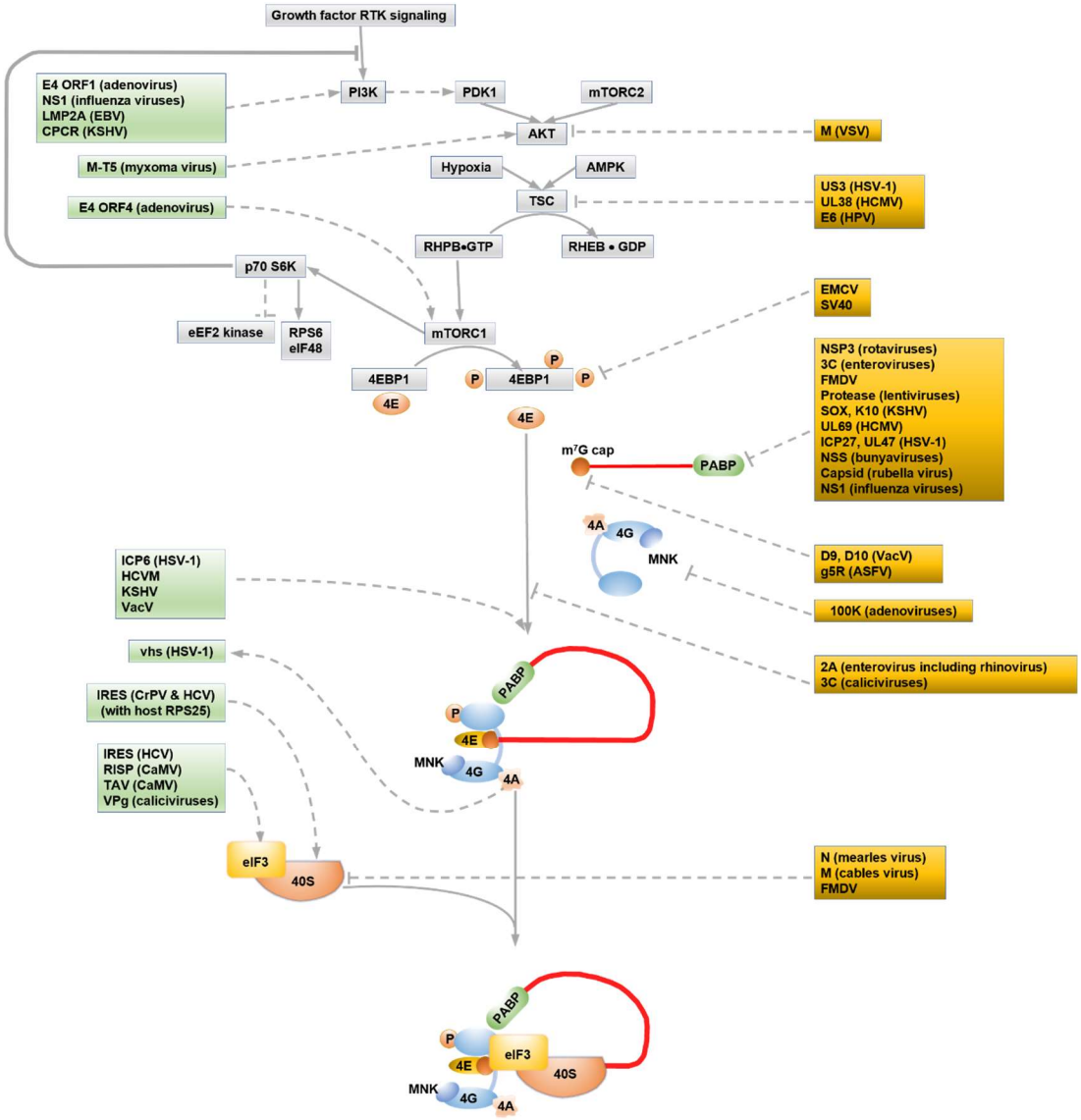


Figure 20 : vue d'ensemble des stratégies virales permettant de détourner la machinerie de traduction cellulaire. Les encadrés verts et orangés recensent respectivement les éléments *cis*- et *trans*- régulateurs utilisés pour contrôler de la traduction. Les pointillés gris indiquent les molécules et les complexes moléculaires ciblés par ces stratégies virales. Adapté de (Walsh and Mohr 2011).

Un certain nombre de facteurs de traduction sont la cible de protéases virales. Le clivage du facteur eIF4G par la protéase 2A des pircornavirus (Gradi *et al.* 1998) ou la protéase du HIV-1 (Ventoso *et al.* 2001) donne une protéine qui conserve son affinité pour eIF4A et eIF3 mais qui devient incapable d'interagir avec eIF4E, inhibant ainsi la traduction coiffe-dépendante cellulaire. La protéase 3C des pircornavirus clive le facteur eIF5B, empêchant ainsi l'assemblage des deux sous-unités du ribosome (de Breyne *et al.* 2008). La protéase 3C de FMDV clive également eIF3 (Belsham *et al.* 2000).

Le facteur eIF4E, essentiel à la traduction coiffe-dépendante, peut également être au cœur de certaines stratégies virales. L'inhibition de la traduction coiffe-dépendante cellulaire est surtout observée lors de l'infection par des virus dont la traduction des ARN n'est pas coiffe-dépendante, que ce soit par l'intermédiaire d'IRES ou par l'intermédiaire de la protéine VpG

(voir ci-après), et peut être fonction du stade de l'infection. La protéine 4E-BP ainsi que la voie mTOR plus en amont sont les principales cibles de ces mécanismes, dont la finalité est la modulation du taux de phosphorylation de 4E-BP et par conséquent du taux de traduction coiffe-dépendante. De manière non exhaustive, on peut citer (Figure 20) :

- parmi les virus à ARN de polarité positive : l'infection par le virus EMCV, dont l'ARN génomique contient un IRES, induit la déphosphorylation de 4E-BP et donc l'inhibition de la traduction coiffe-dépendante (Gingras *et al.* 1996). De manière assez paradoxale aux premiers abords, la protéine NS5A du virus de l'hépatite C, qui contient aussi un IRES, permet l'activation de la voie mTOR et promeut ainsi la traduction coiffe-dépendante (George *et al.* 2012). Le maintien de la traduction cellulaire par la stimulation de eIF4E lors de l'infection par HCV pourrait être impliqué dans la cancérogénèse hépatique, une des conséquences possibles de l'infection par HCV.
- parmi les à ARN de polarité négative : la protéine M du VSV (Vesicular Stomatitis Virus) inhibe la kinase AKT, ce qui résulte en l'inhibition de l'activité de mTORC1 et mène ainsi à la déphosphorylation de 4E-BP en phase tardive d'infection (Dunn and Connor 2011). La traduction des ARN viraux, alors nécessairement coiffe-indépendante, est assurée par un mécanisme encore inconnu (Neidermyer and Whelan 2019), bien que des études aient suggéré le rôle de la protéine 100K dans un mécanisme impliquant un saut du ribosome sur l'ARN « ribosome shunting » (Xi *et al.* 2004) ;
- parmi les virus à ADN : la protéine E6 du Papillomavirus humain (HPV-1), active la voie mTORC1 et par conséquent permet le maintien de la traduction coiffe-dépendante (Spangle and Münger 2010). A l'inverse, le petit antigène de SV40 la désactive et inhibe la traduction coiffe-dépendante (Yu *et al.* 2005). Par ailleurs, SV40 figure parmi les rares virus à ADN dont les ARNm possèdent des IRES (Yu and Alwine 2006).

D'autres stratégies, complémentaires des précédentes, consistent à détourner l'assemblage du complexe de pré-initiation canonique sur l'ARN viral par la présence de la protéine VpG à son extrémité 5'. La protéine VpG est un mime fonctionnel de la coiffe m^7G et permet dès lors le recrutement des facteurs eIF3 et eIF4E à l'extrémité 5' de l'ARN viral, nécessaires à l'assemblage d'un complexe de pré-initiation.

La régulation de la disponibilité du complexe ternaire eIF2-GTP-ARNt^{Met}_i est au cœur de l'ISR activée par la PKR, et est une cible privilégiée lors de l'infection virale. Selon la nature coiffe-dépendante ou coiffe-indépendante du mécanisme de traduction des ARN viraux, les stratégies virales se distinguent selon qu'elles exacerbent ou atténuent la traduction canonique en fonction du stade de l'infection. Par exemple, la protéine K3L du virus de la vaccine promeut la déphosphorylation de eIF2α alors que la protéine NS5A de HCV favorise la phosphorylation de eIF2α, maintenant ainsi le stress cellulaire.

Les facteurs d'élongation et de terminaison sont également impliqués dans ces stratégies virales. Les protéines Gag et reverse transcriptase des rétrovirus interagissent avec eEF1A (Cimarelli and Luban 1999) et eRF1 (Orlova *et al.* 2003) respectivement. Remarquablement, la traduction d'une uORF de l'ARN du cytomégalovirus humain inhibe la terminaison de la traduction en interagissant avec eRF1 (Janzen *et al.* 2002), soulignant le caractère indissociable des mécanismes *cis*- et *trans*- régulateurs de la traduction mis en œuvre lors d'une infection virale.

Enfin, la protéine NSP1 du SARS-CoV-2 affecte à la fois l'initiation (partie 2.3) et la terminaison (Shuvalov et al. 2021) de la traduction. Son rôle sur l'initiation fera l'objet de la

partie 2.3 de ce travail et illustre aussi la coopération de mécanismes cis- et trans- régulateurs permettant la modulation de la traduction lors d'une infection virale.

1.3.2.2. Exemple d'une infection par un champignon

La restrictocine est un peptide fongique synthétisé par le champignon *Aspergillus fumigatus* lors de l'infection de la drosophile. La restrictocine entraine le clivage de la boucle sarcine-ricine du ribosome, ce qui empêche l'accommodation des aa-ARNt dans le site A au cours de l'élongation et inhibe ainsi la traduction (Xu *et al.* 2022).

1.3.2.3. Exemple d'une infection bactérienne

Certaines bactéries produisent des toxines comme la toxine diphtérique ou l'exotoxin A qui provoquent l'ADP-ribosylation du résidu diphthamide de eEF2 et bloquent l'étape de translocation, inhibant ainsi la traduction (Schaffrath *et al.* 2014).

Ces quelques exemples montrent que l'inhibition et/ou le détournement de la traduction cellulaire au cours d'une infection par un pathogène fait partie intégrante des stratégies moléculaires sélectionnées au cours de l'évolution pendant le processus d'infection.

1.3.3. Certaines petites molécules sont des inhibiteurs de la traduction

Initialement caractérisées par leur activité antibiotique ou antifongique, ces molécules ont un intérêt expérimental indéniable dans le cadre d'études en lien avec le ribosome car elles permettent d'isoler et de caractériser des complexes ribosomiques à des stades intermédiaires de la traduction. Leurs modes d'action sur le ribosome ont été caractérisés par des études structurales. Les principaux inhibiteurs qui ont été utilisés dans le cadre de ce travail ou dans d'autres études seront détaillés ci-dessous.

<u>Puromycine</u>

La puromycine est un antifongique et un antibiotique produit par la bactérie *Streptomyces alboniger* et cible spécifiquement les ribosomes procaryotes et eucaryotes. Le groupement lié au carbone 1 de son ribose ressemble structuralement à l'adénosine terminale du brin accepteur des ARNt, ce qui lui confère une affinité pour le site A du ribosome. La puromycine peut entrer dans le site A d'un ribosome pendant la phase d'élongation. Son positionnement dans le site A permet la puromycylation du peptide naissant au niveau du peptidyl-ARNt, ce qui entraine sa dissociation prématurée du ribosome et le recyclage des deux sous-unités ribosomiques. Le mode d'action de la puromycine est donc spécifique des ribosomes engagés dans l'étape d'élongation.

<u>Homoharringtonine</u>

L'homoharringtonine est un alcaloïde produit par les plantes qui se fixe spécifiquement à cheval sur les sites P et A de la grande sous-unité du ribosome (Garreau de Loubresse *et al.* 2014). L'assemblage des deux sous-unités après la reconnaissance du codon initiateur mène alors à un ribosome 80S dont le site A n'est pas vacant, empêchant ainsi le démarrage de la traduction. Le mode d'action de l'homoharringtonine est donc spécifique des ribosomes en initiation, les bloquant au stade 80S sur le site d'initiation.

Cycloheximide

La cycloheximide est un antifongique produit par la bactérie *Streptomyces griseus* et est spécifique du ribosome eucaryote. C'est un inhibiteur de la translocation des ARNt durant l'étape d'élongation, bloquant ainsi les ribosomes au stade 80S sur l'ARNm. Des études structurales ont montré que la cycloheximide se loge dans le site E du ribosome au niveau de l'emplacement du bras accepteur de l'ARNt déacylé, empêchant ainsi l'ARNt provenant du site P de s'y positionner lors de la translocation (Garreau de Loubresse *et al.* 2014). Son mode d'action permet donc de bloquer la translocation et est ainsi spécifique des ribosomes 80S en fin d'initiation et/ou engagés dans l'étape d'élongation.

GMP-PNP (5'-guanylyl imidodiphosphate)

C'est un analogue non-hydrolysable du GTP, et est donc un inhibiteur de toutes les étapes de la traduction nécessitant l'hydrolyse du GTP. Le GMP-PNP inhibe d'une part l'assemblage de la grande sous-unité du ribosome puisqu'il dépend de l'hydrolyse du GTP lié à eIF5B, et d'autre part l'étape d'élongation qui fait intervenir l'hydrolyse du GTP lié à eEF1A. Son mode d'action permet d'isoler des complexes de traduction principalement bloqués au stade 48S sur le codon initiateur.

1.3.4. Les ARN interférents inhibent la traduction des ARNm qu'ils ciblent

La région 3'UTR des ARNm peut contenir des séquences complémentaires à des ARN interférents comme les micro-ARN ou les petits ARN interférents sur lesquelles ces derniers s'hybrident. Ces ARN interférents peuvent être d'origine endogène ou exogène. Leur appariement au niveau de l'ARNm cible permet le recrutement de protéines associées aux ARN interférents et entrainent l'inhibition de la traduction de l'ARNm, soit en induisant la dégradation de l'ARNm, soit en interférant avec le mécanisme d'initiation de la traduction (Iwakawa and Tomari 2022).

Ces mécanismes dépendent de l'assemblage des ARN interférents avec de multiples protéines pour former un « RNA-induced silencing complex » (RISC) en association avec des protéines Argonaute (Ago). Le recrutement de ce complexe est guidé par l'hybridation de l'ARN interférant sur une séquence cible de la région 3' UTR de certains ARNm et induit l'inhibition de leur traduction (Figure 21). Le complexe RISC peut en effet provoquer la coupure de la coiffe en 5' de l'ARNm et sa dégradation subséquente par la nucléase XRN1 et/ou interférer avec des facteurs de traduction.

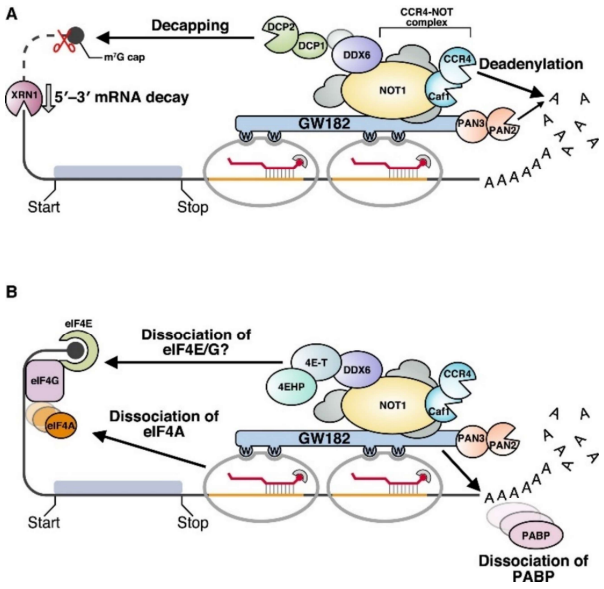


Figure 21 : vue d'ensemble des mécanismes d'inhibition de la traduction par le recrutement du complexe RISC guidé par les ARN interférents sur l'ARNm. A. Le recrutement du complexe RISC induit la coupure de la coiffe en 5' de l'ARNm et sa dégradation. **B**. Le complexe RISC interfère avec la mécanique d'initiation de la traduction. D'après (Iwakawa and Tomari 2022).

1.3.5. Le peptide naissant agit comme un régulateur de sa propre synthèse

Le repliement d'une protéine commence dès la sortie du peptide naissant du tunnel du ribosome, et peut constituer un signal de localisation cellulaire de la traduction, comme cela a été vu avec la traduction associée au réticulum endoplasmique (voir 1.2.5.2). Le repliement co-traductionnel du peptide naissant est déterminé par la nature des acides aminés qui le constituent, mais également par la vitesse d'élongation du ribosome. Réciproquement, le repliement co-traductionnel d'une protéine influence aussi la vitesse d'élongation. Enfin, le décodage des codons, qui dépend de la disponibilité des aa-ARNt correspondants, est un modulateur essentiel de la vitesse d'élongation. Ce dernier point sera abordé dans le paragraphe 1.4.2.2 en tant qu'élément *cis*-régulateur de la traduction. Ce paragraphe se focalise uniquement sur l'influence qu'exerce le peptide naissant sur sa propre synthèse.

La manifestation la plus évidente d'une telle régulation est la traduction de clusters de codons codant pour des acides aminés qui induisent une forte gêne stérique dans le ribosome, dont les plus connus sont les triplets de prolines (PPP). Une conséquence de cette gêne stérique est le ralentissement drastique de la cinétique de formation de la liaison peptidique par le site peptidyl-transférase du ribosome. Lorsque ce ralentissement conduit à une pause complète du ribosome, on parle de « ribosome stalling ». La traduction des poly-prolines induit un ralentissement du ribosome tel que la reprise normale de l'élongation de la traduction nécessite l'intervention du facteur eIF5A (Gutierrez *et al.* 2013). La durée du blocage du ribosome au niveau des motifs PPP est aussi modulée par la nature des acides aminés qui les entourent (Starosta *et al.* 2014). Par ailleurs, des expériences d'invalidation du facteur eIF5A ont permis d'identifier d'autres triplets d'acides aminés qui induisent des pauses du ribosome, dont les deux plus fréquents sont PDP et DPG (Schuller *et al.* 2017). Une autre étude suggère que le repliement en feuillets β du peptide naissant entraine un ralentissement de la vitesse d'élongation du ribosome pour les mêmes raisons (Burke *et al.* 2022).

Certains peptides issus de la traduction de phases codantes situées dans la région 5'UTR en amont de la phase codante principale (uORF) présentent parfois de très fortes conservations de séquences parmi différents organismes, suggérant un rôle potentiel dans la régulation de la traduction de l'ARNm correspondant (Dever *et al.* 2020).

Enfin, des phases codantes peuvent coder pour des peptides auto-clivants comme le peptide 2A, dont la séquence peptidique est telle que le l'activité peptidyl-transférase du ribosome échoue à former la liaison peptidique entre une glycine et une proline (Donnelly *et al.* 2001; Sharma *et al.* 2012). Ces mécanismes sont notamment employés par les picornavirus comme moyen de produire plusieurs protéines à partir d'une seule phase codante (Luke *et al.* 2008).

1.3.6. La vitesse d'élongation est liée au niveau de stress cellulaire

Un ralentissement trop prononcé des ribosomes en cours d'élongation sur un même ARNm mène à leurs collisions. La collision entre deux ribosomes sur un ARNm survient lorsqu'un ribosome ralentit ou s'arrête à un endroit précis de la phase codante qui impacte sa vitesse d'élongation. Il est alors rattrapé par le ribosome en cours de traduction qui le suit à une vitesse normale. Les causes de ce ralentissement peuvent être des modifications de bases dans la phase codante (paragraphe 1.4.1.1), la nature des acides aminés décodés (paragraphe précédent) ou la faible disponibilité des ARNt lorsque le ribosome décode les codons correspondants (paragraphe 1.4.2.2). Un déficit d'acides aminés ralentit la synthèse des aa-ARNt et réduit indirectement la vitesse d'incorporation des aa-ARNt dans le site A, ce qui diminue la vitesse d'élongation. C'est une cause majeure de la collision de ribosomes.

Un taux basal de collisions de ribosomes dans la cellule est résolu par des mécanismes de contrôle qualité des ribosomes et des ARNm nommés « No-go decay » (Yan *et al.* 2019). Les ribosomes arrêtés dans la phase codante (sans codon stop dans le site A) demeurent ainsi car leur mécanisme de recyclage canonique ne peut s'enclencher puisqu'il dépend de la présence d'un codon stop dans le site A. La détection de deux ribosomes en collision est assurée par la protéine GCN1 qui se fixe spécifiquement sur deux ribosomes adjacents, qui en retour active la kinase GCN2 impliquée dans la voie ISR (Pochopien *et al.* 2021). Une des conséquences est l'ubiquitination des protéines ribosomiques eS10 et uS10 par la protéine ZNF598 (Juszkiewicz *et al.* 2018). L'ubiquitination entrainerait ensuite le recrutement direct ou indirect de nucléases menant à la dégradation des ARNm sur lesquels les ribosomes sont à l'arrêt. Le mécanisme par lequel les ribosomes sont recyclés à la suite de l'ubiquitination est encore mal connu.

Lorsque le taux de collisions des ribosomes est très élevé, cela induit une série d'évènements moléculaires reliés à des voies de stress (Wu *et al.* 2020a). D'abord, l'activation prolongée par GCN1 de la kinase GCN2, impliquée dans la phosphorylation de eIF2α dans le cadre de l'ISR, permet l'inhibition de la traduction coiffe-dépendante. Ce mécanisme peut être interprété comme une réponse de la cellule au déficit d'acides aminés en dirigeant les ressources vers la synthèse des protéines indispensables pour la réponse au stress. Si ce mécanisme ne s'avère pas suffisant pour revenir à un état cellulaire basal, les collisions excessives des ribosomes qui résultent de cet échec induisent le recrutement de la protéine ZAKα, qui, en se fixant sur les ribosomes juxtaposés par leur collision, active la signalisation moléculaire menant à l'apoptose.

La vitesse d'élongation est donc un paramètre déterminant de l'état physiologique de la cellule, et apparait comme un facteur décisionnel dans le déclenchement des voies d'apoptose.

1.4. Mécanismes *cis*-régulateurs de la traduction des ARNm cellulaires

L'ARNm code pour une protéine mais également pour la régulation de sa traduction. Ces informations de régulation sont contenues à la fois dans les régions UTR et dans les phases codantes sous la forme de motifs de quelques nucléotides et/ou de structures secondaires plus ou moins complexes (Figure 22). La *cis*-régulation de la traduction est indissociable des aspects *trans*-régulateurs précédemment évoqués.

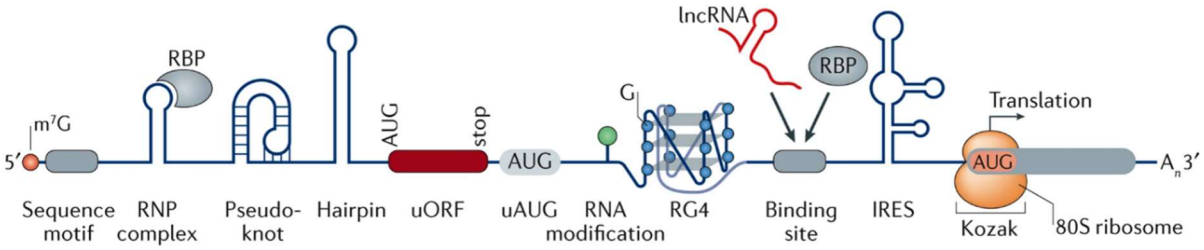


Figure 22 : vue d'ensemble des éléments *cis*-régulateurs de l'initiation de la traduction chez les eucaryotes. Hairpin : structure en épingle à cheveux ; RG4 : structure en G quadruplex ; lncRNA : long ARN non-codant ; Pseudo-knot : structure en pseudo-nœud ; RBP : protéines pouvant se lier à l'ARN ; RNP complex : complexe ribonucléoprotéique. D'après (Leppek *et al.* 2018).

L'accent sera mis sur la régulation de l'initiation de la traduction chez les eucaryotes. Par souci d'exhaustivité, les régulations liées aux phases d'élongation et de terminaison seront évoquées mais ne seront pas nécessairement approfondies.

1.4.1. Eléments de la région 5'UTR

La région 5'UTR des ARNm concentre la majeure partie des éléments *cis*-régulateurs de l'initiation de leur traduction, qu'elle soit canonique ou non-canonique.

1.4.1.1. Les modifications de base

Au-delà de la coiffe m^7G présente à l'extrémité 5' des ARNm eucaryotes sur laquelle repose le mécanisme d'initiation canonique, d'autres modifications de bases sont essentielles pour la traduction de certains ARNm. Elles ont lieu au moment de la synthèse et/ou de la maturation des ARN pré-messagers dans le noyau.

N_6-méthyl-adénosine (m^6A)

La plus importante est certainement la N_6-méthyl-adénosine (m^6A) (Dominissini *et al.* 2012), qui correspond à la méthylation d'une adénosine sur l'azote 6 (N_6). Son impact sur la traduction se manifeste par deux principaux mécanismes qui impliquent soit le facteur d'initiation eIF3, soit la déstabilisation de structures secondaires (Figure 23). La modification m^6A est particulièrement impliquée dans la traduction des ARNm en condition de stress.

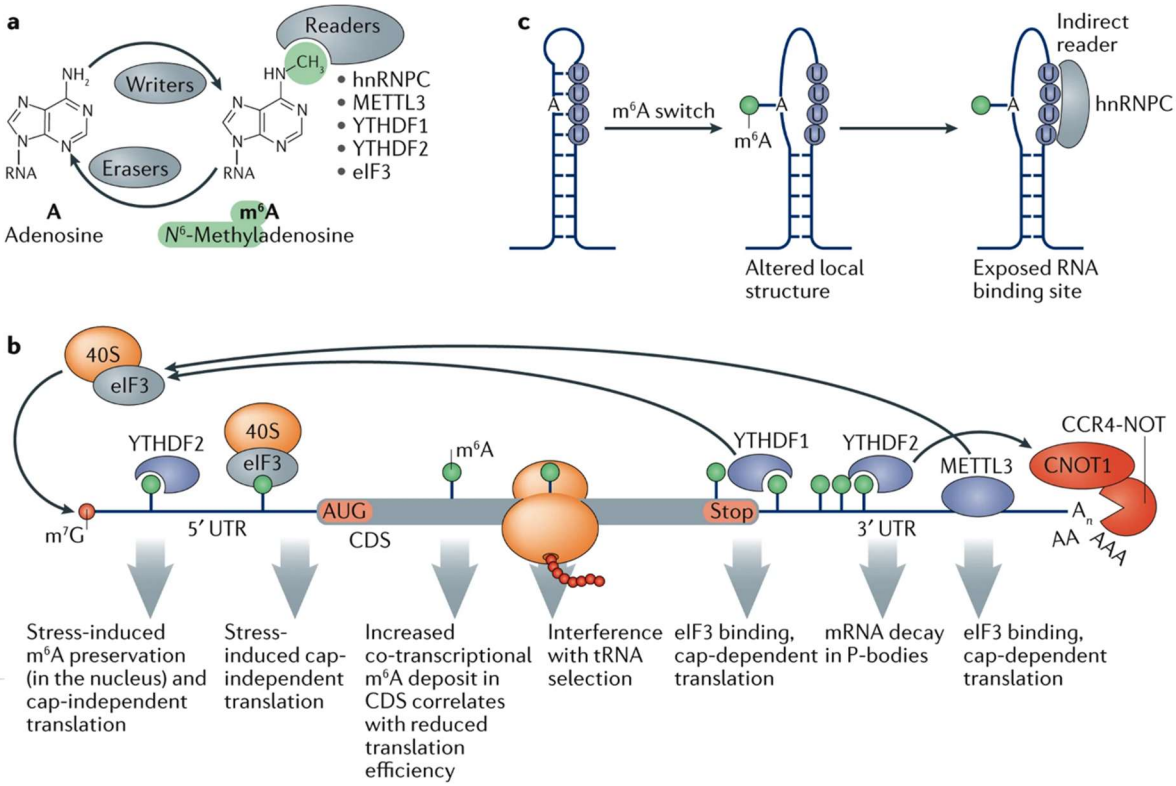


Figure 23 : vue d'ensemble des rôles des m⁶A sur la traduction des ARNm. a. Les enzymes « writers » sont des méthyltransférases tandis que les « erasers » sont des déméthylases. Les protéines « readers » sont celles capables d'interagir avec la base modifiée, en l'occurrence ici la m⁶A. **b.** Les nucléotides modifiés m⁶A interagissent avec diverses protéines impliquées dans la stabilité et/ou la traduction de l'ARNm modifié. **c.** Les modifications m⁶A peuvent induire des changements conformationnels permettant le recrutement de protéines sur l'ARNm. D'après (Leppek *et al.* 2018).

Le premier mécanisme fait intervenir le facteur d'initiation eIF3 qui reconnait une ou plusieurs modifications m⁶A de manière coiffe-indépendante, recrutant ainsi le complexe 43S directement au niveau des adénosines méthylées sur l'ARNm (Meyer *et al.* 2015). Selon le positionnement des m⁶A par rapport au codon d'initiation, le recrutement des facteurs de scanning par l'intermédiaire de l'interaction entre eIF4G et eIF3 peut s'avérer indispensable. Ce type de mécanisme peut être interprété comme une forme d'IRES cellulaire, nommée m⁶A Internal Ribosome Entry Site (MIRES). Le profil de méthylation des adénosines de la région 5'UTR est modulé par l'activité des méthyltransférases nucléaires qui catalysent la méthylation des adénosines sur l'azote 6 durant la maturation des ARN pré-messagers, ainsi que par l'import nucléaire de protéines qui fixent la modification m⁶A et inhibent ainsi l'activité des enzymes de déméthylation nucléaires. Un exemple de ce dernier mécanisme est l'import nucléaire de YTHDF2, qui, en réponse à un choc thermique, migre dans le noyau et stabilise les modifications m⁶A d'ARNm codant pour des protéines de réponse au choc thermique, comme les chaperonnes HSP70 (Zhou *et al.* 2015). En particulier, la méthylation des adénosines intervient dans la traduction nécessairement coiffe-indépendante des ARN circulaires (Legnini *et al.* 2017; Yang *et al.* 2017) puisqu'ils n'ont pas d'extrémité 5'. Enfin, les modifications m⁶A assurent également la distinction par la cellule entre les ARN circulaires

endogènes et exogènes, les premiers n'étant pas pris pour cible par les mécanismes cellulaires d'immunité innée (Chen *et al.* 2019).

Une autre conséquence de la présence de m^6A dans les ARNm est le dépliement de structures secondaires présentes dans la région 5'UTR qui pourraient bloquer la particule 43S en cours de scanning et/ou empêcher le recrutement de certains facteurs essentiels à l'initiation de la traduction (Liu *et al.* 2015; Spitale *et al.* 2015). A cela peut s'ajouter le recrutement interne du facteur eIF3, et avec lui celui de la particule 43S comme précédemment évoqué.

Si la méthylation des adénosines situées dans la région 5'UTR a plutôt un effet stimulant sur la traduction, la méthylation de celles situées dans la phase codante ou dans la région 3'UTR a plutôt un effet inhibiteur. La méthylation des adénosines situées dans la phase codante déstabilise les interactions codon-anticodon et ne permet donc pas l'accommodation des ARNt lors de la phase d'élongation (Choi *et al.* 2016). La méthylation des adénosines situées dans la région 3'UTR peut quant à elle favoriser la dégradation des ARNm en recrutant des protéines impliquées dans le NMD (Non-sense Mediated Decay) (Wang *et al.* 2014). Le mécanisme de NMD entraine alors la dégradation des ARNm concernés (Hug *et al.* 2016).

Autres modifications

D'autres modifications de bases ont été caractérisées, comme la N_1-méthyl-adénosine, trouvées principalement dans les régions 5'UTR, ou l'hydroxyméthyl-cytosine, principalement trouvée dans les phases codantes. Les deux améliorent respectivement l'efficacité de l'initiation (Dominissini *et al.* 2016) et de l'élongation de la traduction (Delatte *et al.* 2016).

Initiation sur une coiffe hyperméthylée

La traduction des ARNm qui codent pour les séléno-protéines est particulière dans la mesure où ces ARNm possèdent une coiffe tri-méthylée $m^{2,2,7}G$ dont l'interaction avec le facteur eIF3 par les sous-unités c, d et e permet le recrutement de la particule 43S (Hayek *et al.* 2022). Toutefois, ce mécanisme est encore mal compris.

Les rôles de eIF3 dans ces mécanismes d'initiation non-canoniques soulignent l'importance de ce facteur qui apparait très impliqué dans les mécanismes d'initiation impliquant des interactions avec les bases modifiées des ARNm.

1.4.1.2. La longueur et la séquence de la région 5'UTR

Chez les eucaryotes, la longueur des régions 5'UTR suit une distribution asymétrique étirée vers les grandes valeurs (Leppek *et al.* 2018), avec des valeurs médianes comprises entre 50 et 200 nucléotides. Chez l'Homme, la valeur médiane est voisine de 200 nucléotides, et environ 25% des régions 5'UTR ont une longueur inférieure à 150 nucléotides ou supérieure à 400 nucléotides (Figure 24).

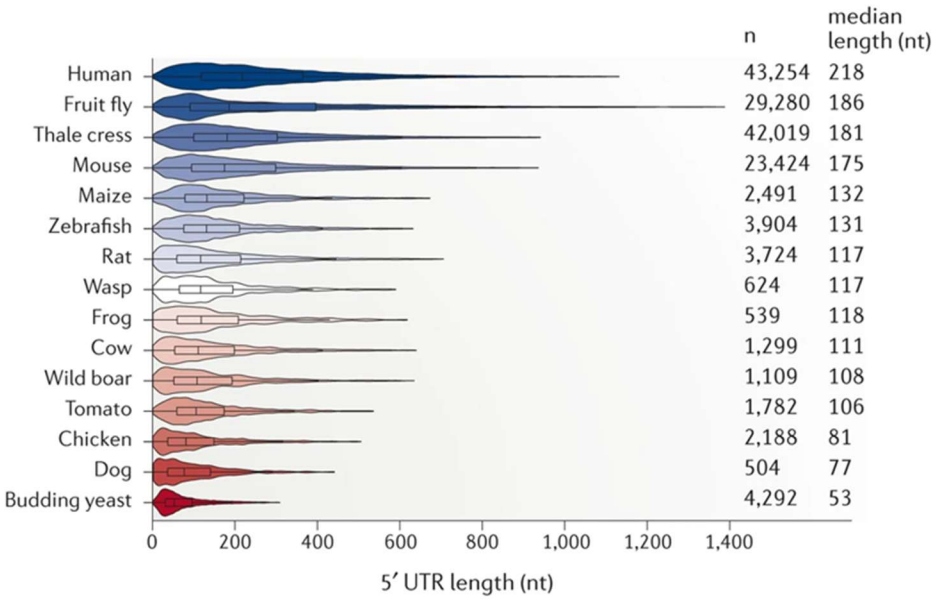


Figure 24 : distribution des longueurs des régions 5'UTR chez différents organismes eucaryotes. La longueur des régions 5'UTR ainsi que leur valeur médiane sont données en nucléotides (nt). n correspond au nombre de régions 5'UTR de transcrits annotés dans la banque RefSeq. D'après (Leppek *et al.* 2018).

Traduction coiffe- et scanning- dépendante des ARNm avec de longues régions 5'UTR

En plus de eIF4A, d'autres ARN hélicases comme Ded1/DDX3 ou DHX29 (Sen *et al.* 2015) peuvent s'avérer nécessaires pour le scanning des longues régions 5'UTR puisqu'elles participent au dépliement des structures secondaires que l'activité hélicase modérée de eIF4A ne peut déplier.

Traduction coiffe-dépendante mais scanning-indépendante des ARNm avec de courtes régions 5'UTR

La traduction des ARNm avec de très courtes régions 5'UTR fait intervenir des mécanismes coiffe-dépendants mais scanning-indépendants.

La traduction des ARNm dont le codon d'initiation est entouré par le motif SAASAUGGCGGC (où S est C ou G) est dépendante d'un mécanisme nommé TISU, pour « Translation Initiation on Short UTR » (Haimov *et al.* 2015). Dans ces ARNm, ce motif est localisé à une distance comprise entre 5 et 30 nucléotides du site de démarrage de la transcription, ce qui résulte en des régions 5'UTR très courtes dont la longueur est comprise entre 5 et 30 nucléotides, avec une longueur médiane de 12 nucléotides (Elfakess and Dikstein 2008). Différentes études ont montré que le mécanisme de traduction des ARNm TISU nécessite la présence de la coiffe m^7G mais n'utilise pas de scanning. L'initiation de la traduction des ARNm TISU repose largement sur le contexte nucléotidique du codon initiateur qui empêche les phénomènes de « leaky scanning » (Haimov *et al.* 2015).

D'autres ARNm ayant de courtes régions 5'UTR comme celui codant pour l'histone H4 font intervenir des structures secondaires pour la reconnaissance du codon d'initiation de manière scanning-indépendante (Martin *et al.* 2011). L'ARNm codant pour l'histone H4 a une région 5'UTR coiffée longue de seulement 9 nucléotides, et l'extrémité 3' n'est pas polyadénylée. La

phase codante et la région 3' UTR comprennent respectivement 312 et 54 nucléotides et sont aussi relativement courtes. L'initiation de la traduction de l'ARNm de l'histone H4 fait intervenir des éléments structuraux présents dans la phase codante. Certains de ces éléments interagissent avec les facteurs d'initiation eIF4E (Martin *et al.* 2011), eIF3 (Hayek *et al.* 2021) ou avec le ribosome par l'interaction avec l'ARNr 18S (Martin *et al.* 2016). Le recrutement interne de eIF4F permet de déloger la coiffe m^7G initialement enfouie dans une structure compacte. Dans ce mécanisme, la particule 43S est alors directement déposée au niveau du codon initiateur sans étape de scanning.

1.4.1.3. Le codon d'initiation et son contexte nucléotidique

Le contexte nucléotidique du codon initiateur est un élément essentiel pour la modulation de l'efficacité de l'initiation de la traduction. Les travaux pionniers de Marylin Kozak ont montré que la présence de purines aux positions -3 et +4 (+1 étant le A de l'AUG initiateur par convention) est impliquée dans la reconnaissance efficace de l'AUG initiateur (Kozak 1986). Les autres positions proximales ont ensuite rapidement été analysées et s'avèrent également importantes, dans une moindre mesure toutefois que les -3 et +4 (Kozak 1987). En moyennant les séquences des phases codantes de génomes eucaryotes, la séquence « consensus » la plus fréquente autour du codon initiateur a été déterminée comme étant GCCRCCAUGR, où R est une purine (A ou G). Ce motif est communément appelé la séquence de Kozak.

Il convient de préciser à ce stade qu'il s'agit là d'une séquence moyenne, établie en calculant le nucléotide le plus abondant sur chaque position, qui ne constitue en aucun cas un motif majoritairement trouvé parmi les séquences eucaryotes. En effet, seules quatre séquences peuvent être dérivées de GCCRCCAUGR (GCCACCAUGA, GCCACCAUGG, GCCGCCAUGA, GCCGCCAUGG), et elles ne sont donc absolument pas représentatives de l'ensemble des quelques cent mille contextes nucléotidiques des AUG des phases codantes annotées dans le génome humain, dont environ vingt-huit mille sont distincts (voir 2.1.1.1). La séquence consensus de Kozak est indicatrice des nucléotides les plus préférentiellement trouvés sur chaque position autour du codon AUG, indépendamment des nucléotides voisins, et ne devrait pas être interprétée en dehors de ce cadre très restreint.

De nombreuses expériences de mutations ponctuelles réalisées par Marylin Kozak ont démontré l'existence de contextes nucléotidiques favorables et défavorables pour l'initiation de la traduction et ont davantage souligné le caractère essentiel des positions -3 et +4. En effet, muter les positions -3 et +4 en pyrimidines (C ou U) est fortement défavorable pour la traduction (Kozak 1986, 1987). Néanmoins, ces observations ne sont valables que dans le cadre des contextes nucléotidiques testés car rien ne permet de les généraliser à l'ensemble des contextes nucléotidiques des codons d'initiation des phases codantes trouvées dans les génomes eucaryotes.

La séquence moyenne autour du codon initiateur a ensuite été confirmée par des analyses bio-informatiques généralisées sur les séquences codantes connues du génome humain et plus généralement sur celles d'autres génomes eucaryotes (Nakagawa *et al.* 2007; Hernández *et al.* 2019). Il ressort de ces analyses une séquence moyenne légèrement distincte de celle déterminée par Marylin Kozak : GCMRNCAUGG, où R est A ou G et M est A ou C. Ces études révèlent néanmoins que seules les positions -3 et -2 sont universellement conservées chez les eucaryotes, les autres étant soumises à une certaine variabilité en fonction de l'organisme étudié (Tableau 3).

Groupe	Contexte nucléotidique du codon initiateur (AUG)							
	-6	-5	-4	-3	-2	-1	AUG	+4
Vertébrés	G	C		R	M	C	-	G
Tuniciers			M	A	M		-	G/U
Invertébrés			M	R	M		-	G/A/U
Plantes vertes				R	M		-	G
Algues rouges				A	M		-	A/U
Mycètes			M	R	M		-	U
Amœbozoaires				R	M		-	R
Alvéolés				R	M		-	R
Excavés				R	M		-	C/A/U
Universellement conservé				**R**	**M**			

Tableau 3 : nucléotides les plus conservés du contexte nucléotidique du codon d'initiation des positions -6 à +4 chez différents eucaryotes. R : A ou G ; M : A ou C. Adapté de (Hernández *et al.* 2019).

Des études structurales ont montré que la quasi-totalité des nucléotides au voisinage immédiat du codon initiateur établissent des interactions (Figure 25) avec des facteurs d'initiation, des protéines ribosomiques ou des portions d'ARN ribosomiques dans le complexe de pré-initiation assemblé sur le codon de démarrage (Pisarev *et al.* 2006; Martin *et al.* 2016; Hussain *et al.* 2014; Babaylova *et al.* 2020; Bhaskar *et al.* 2020; Simonetti *et al.* 2020; Thakur *et al.* 2020).

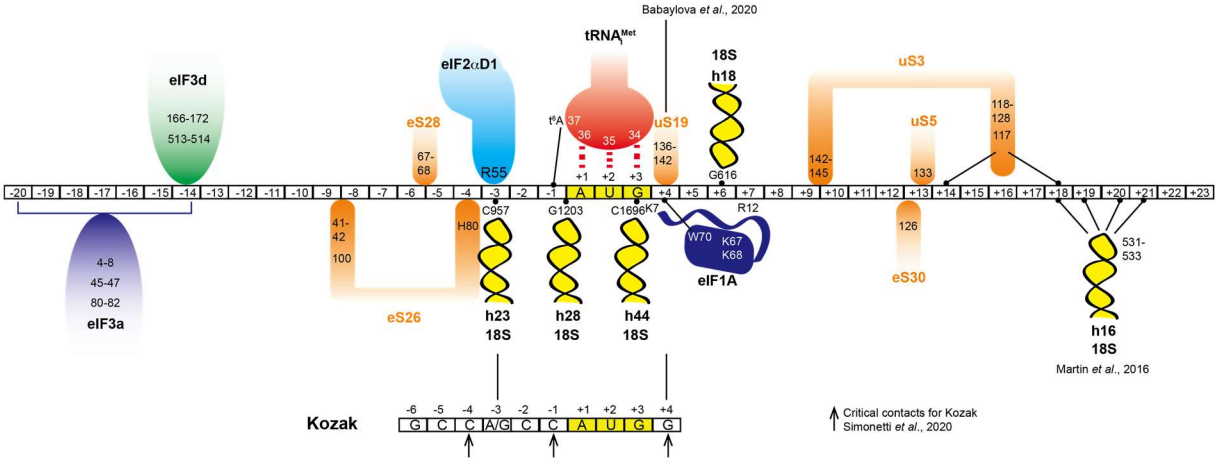


Figure 25 : vue d'ensemble des interactions entre les nucléotides proximaux du codon d'initiation avec les facteurs d'initiation, les ARN et protéines ribosomiques, et l'ARNt$^{Met}_{i}$

Une partie du travail de thèse est dédiée à l'identification des contextes nucléotidiques optimaux pour l'initiation sur un codon AUG, et à la recherche de potentiels motifs minimaux permettant d'expliquer leur efficacité (partie 2.1.1.1).

Enfin, le codon d'initiation AUG, bien que largement préférentiel, n'est pas le seul à pouvoir générer l'assemblage du ribosome et le démarrage de la traduction (Figure 26). Des expériences de ribosome profiling montrent que d'autres codons, qu'ils soient AUG-like (un nucléotide de différence avec l'AUG) ou non-AUG, peuvent être reconnus comme codon d'initiation (Ingolia *et al.* 2011; Chothani *et al.* 2022).

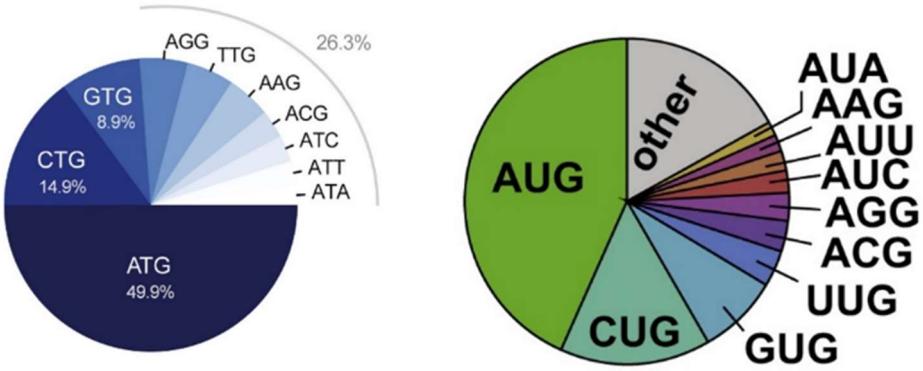


Figure 26 : taux d'utilisation des codons d'initiation dans la cellule déterminés par des études de ribosome profiling. D'après (Chothani *et al.* 2022; Ingolia *et al.* 2011).

Une partie du travail de thèse aborde l'implication de structures secondaires situées en aval des codons dans l'initiation de la traduction sur des codons non-AUG (voir 2.1.1.2).

1.4.1.4. Les phases codantes situées dans la région 5'UTR en amont du codon d'initiation principal et les mécanismes de ré-initiation

Les phases codantes situées en amont de la phase codante considérée comme principale (uORF pour « upstream Open Reading Frame ») sont des éléments majeurs de la *cis*-régulation de la traduction. Généralement, la traduction des uORFs inhibe celle de la phase codante principale en agissant comme un piège à particules 43S. Dans certains cas, lorsque le nombre de nucléotides séparant l'uORF de la phase codante principale n'est pas trop élevé, cette dernière peut être traduite par un mécanisme de ré-initiation. La traduction de l'ARNm d'ATF4 sera prise comme exemple pour illustrer ces mécanismes. A noter que le peptide résultant de la traduction des uORFs peut également agir comme un facteur *trans*-régulateur à part entière. Les uORFs sont trouvées dans près de la moitié des phases codantes annotées du génome humain (Calvo *et al.* 2009) et la plupart est largement utilisée dans la cellule (Ingolia *et al.* 2011; Chothani *et al.* 2022).

Eléments *cis*-régulateurs modulant l'initiation de la traduction des uORFs

Le contexte nucléotidique du codon initiateur de l'uORF module évidemment l'initiation de sa traduction comme expliqué dans le paragraphe précédent.

Parfois, l'inhibition de l'ORF principale par l'initiation sur des uORFs fait intervenir la coopération de facteurs *cis*- et *trans*- régulateurs. Des éléments structuraux de la région 5'UTR peuvent moduler la reconnaissance du codon initiateur des uORFs par un mécanisme d'initiation de la traduction appelé « STructure Assisted RNA Translation » ou mécanisme START (Eriani and Martin 2018; Despons and Martin 2020). Ce mécanisme repose sur la

présence d'une structure secondaire localisée en aval du codon initiateur de l'uORF dont la stabilité est suffisante pour stopper ou du moins suffisamment ralentir la particule 43S en cours de scanning. Selon la nature du triplet de nucléotides se trouvant dans le site P à ce moment-là, ce ralentissement peut favoriser la stabilisation de l'interaction codon-anticodon et mener à l'initiation de la traduction sur ce codon, même si ce n'est pas un AUG. Ce blocage comblerait le déficit énergétique inhérent aux interactions codon-anticodons les moins favorables qui normalement fait que la particule 43S poursuit le scanning, comme c'est le cas pour les codons non-AUG ou les codons se trouvant dans un contexte nucléotidique défavorable à leur reconnaissance par la particule 43S. Le taux de traduction des uORFs par le mécanisme START et le taux de traduction de l'ORF principale sont alors modulés par le recrutement d'hélicases au niveau du complexe de scanning, comme l'ARN hélicase Ded1p/DDX3 (Guenther *et al.* 2018). Son activité ARN hélicase permet le dépliement des structures secondaires en aval de l'uORF et favorise ainsi l'initiation sur l'AUG de l'ORF principale, qui autrement est inhibée par la traduction des uORFs par le mécanisme START (Figure 27).

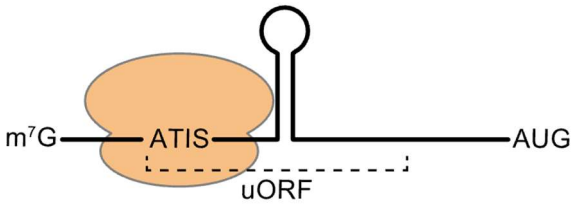


Figure 27 : illustration du mécanisme START pour la traduction d'uORFs. ATIS : site alternatif d'initiation de la traduction, AUG : codon d'initiation de l'ORF principale. D'après (Eriani and Martin 2018).

Traduction de l'ARNm codant pour le facteur de transcription ATF4

Déjà mentionnée à plusieurs reprises comme étant centrale dans la réponse au stress, la traduction de l'ARNm codant pour le facteur de transcription ATF4 fait intervenir un jeu d'uORFs couplé à des mécanismes de ré-initiation (Vattem and Wek 2004) qui sont assistés par des facteurs de traduction non canoniques (Bohlen *et al.* 2020; Vasudevan *et al.* 2020). Ces mécanismes permettent sa traduction sélective en conditions de stress (Figure 28).

Le mécanisme proposé pour expliquer la sélectivité traductionnelle de l'ARNm d'ATF4 est le suivant (Vattem and Wek 2004). La région 5'UTR de l'ARNm d'ATF4, longue de 264 nucléotides, contient deux uORFs. La première (uORF1), dite activatrice, contient 12 nucléotides à partir du nucléotide 68 dans la région 5'UTR, et est essentielle à la traduction de la phase codante principale. Cette uORF est séparée par 87 nucléotides de la seconde uORF (uORF2), dite inhibitrice, qui contient 180 nucléotides et dont la phase de lecture est chevauchante avec la phase codante principale : son codon stop comprend en effet les nucléotides 81, 82 et 83 de la phase codante principale. Il est clair que la traduction de l'uORF2 inhibe celle de la phase codante principale, qui dépend alors de la vitesse de ré-initiation du complexe d'initiation après la traduction de l'uORF1. La concentration cellulaire du complexe ternaire eIF2-GTP-ARNt$^{Met}_{i}$ ainsi que les concentrations des facteurs de traduction non-canoniques DENR et eIF2D modulent la vitesse de ré-initiation. Après la traduction de l'uORF1, le ribosome se dissocie et reforme une particule 43S compétente pour le scanning

grâce au maintien des facteurs essentiels pour le scanning, eIF3, eIF4G et eIF4A, sur le complexe de traduction 80S durant l'élongation sur l'uORF1 (Pöyry *et al.* 2004). L'implication du mécanisme de scanning dans la ré-initiation est suggérée par le fait que l'insertion d'une tige-boucle qui bloque ou du moins ralentit le scanning entre l'uORF1 et l'uORF2 diminue la traduction de l'ORF principale. C'est la vitesse de chargement du l'ARNt$^{Met}_{i}$ dans la particule 43S en cours de scanning qui détermine si la ré-initiation de la traduction aura lieu sur l'AUG de l'uORF2 ou sur l'AUG de l'ORF principale. Une concentration élevée de complexe ternaire, et donc de eIF2α non phosphorylé, permet la ré-initiation rapide du complexe d'initiation sur l'uORF2, dont la traduction empêche celle de l'ORF principale. En revanche, une faible concentration de complexes ternaires, en relation avec le taux de phosphorylation de eIF2α, ralentit sensiblement la vitesse de ré-initiation et favorise alors la ré-initiation sur l'ORF principale plutôt que sur l'uORF2, étant donné que la particule 43S au cours de scanning aura alors dépassé l'AUG de l'uORF2 au moment du chargement de l'ARNt$^{Met}_{i}$. Par conséquent, la distance séparant l'uORF1 de l'uORF2 et de l'ORF principale est un élément *cis*-régulateur critique pour la synthèse d'ATF4. Le mode d'acheminement de l'ARNt$^{Met}_{i}$ dans la particule 43S après la traduction de l'uORF1 ferait intervenir le facteur eIF2D plutôt que le peu de complexes ternaires encore fonctionnels après la phosphorylation de eIF2α (Vasudevan *et al.* 2020). Dans tous les cas, la cinétique de recrutement de l'ARNt$^{Met}_{i}$ dans la particule 43S doit être suffisamment lente pour permettre la ré-initiation sur l'ORF principale et non pas sur l'uORF2 et ainsi produire le facteur de transcription ATF4 qui orchestre ensuite la réponse au stress.

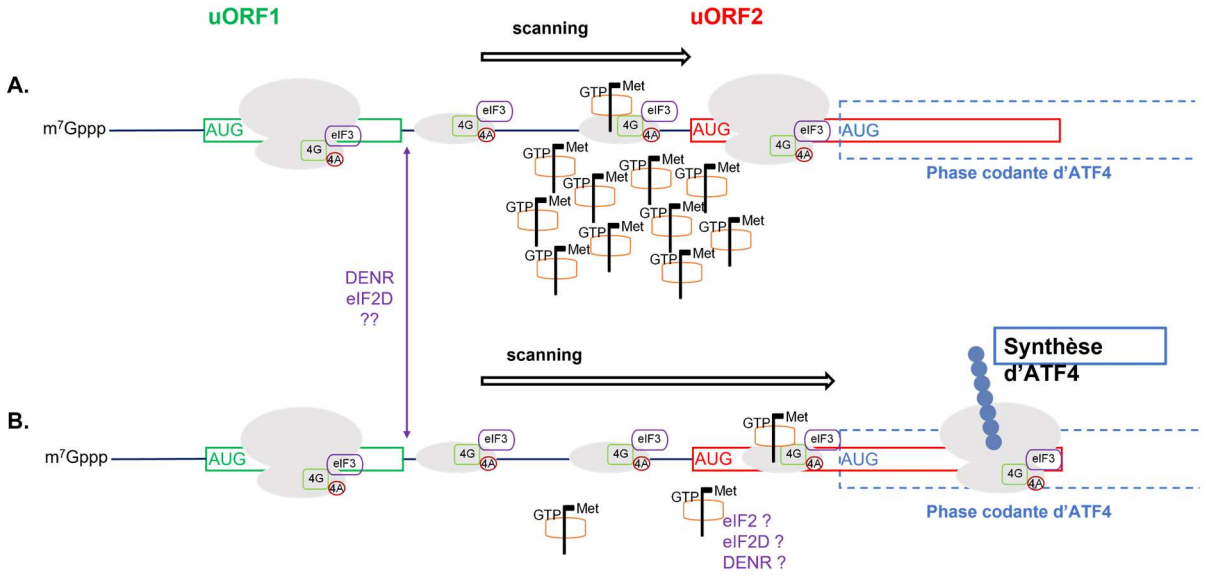


Figure 28 : modèle de la traduction de l'ARNm d'ATF4. L'uORF1 est toujours traduite par le mécanisme d'initiation canonique, à la suite de quoi le ribosome se dissocie et entame un nouveau scanning grâce à la persistance de certains facteurs d'initiation sur le ribosome 80S. **A.** Dans des conditions physiologiques normales, la concentration cellulaire en complexes ternaires est élevée favorisant ainsi l'initiation de la traduction de l'uORF2, ce qui a pour conséquence d'empêcher la traduction de la phase principale codante d'ATF4 située en aval. **B.** Lorsque la concentration du complexe ternaire est faible en raison de la phosphorylation de eIF2α dans le cadre de l'ISR, un phénomène de leaky-scanning de l'uORF2 favorise la traduction de la protéine ATF4. Ce mécanisme utilise des facteurs d'initiation non-canoniques comme eIF2D ou DENR.

1.4.1.5. Les motifs d'ARN inhibiteurs des ribosomes

Un exemple de motif d'ARN inhibiteur de la traduction a été mis en évidence dans la région 5'UTR des ARNm du gène homéotique Hox a11 par notre équipe (Alghoul *et al.* 2021). Il repose sur l'action synergique de deux éléments localisés dans la région 5'UTR : un motif start-stop (un codon AUG suivi d'un codon stop) séparé par 19 nucléotides d'une structure en tige-boucle de 50 nucléotides dont les 16 paires de bases G-C qui la constituent lui confèrent une stabilité assez élevée (enthalpie libre de -21,6 kcal/mol). L'AUG étant reconnu par la particule 43S en cours de scanning, un ribosome 80S est assemblé sur l'AUG, et le codon stop se retrouve alors dans le site A, ce qui doit normalement constituer un signal de recrutement des eRF pour le recyclage du ribosome. Néanmoins, la structure secondaire est positionnée à une distance telle du codon AUG qu'elle bloque l'entrée du site A et empêche le recrutement des eRF et le recyclage du ribosome, le piégeant ainsi au niveau du start-stop.

Par ailleurs, cette structure peut bloquer la particule 43S en cours de scanning. En effet, muter le codon AUG du start-stop ne permet pas d'initier la traduction sur l'AUG principal de a11. Ainsi, en plus d'empêcher l'arrivée des eRF au niveau du site A, la structure bloque aussi les particules 43S qui n'auraient pas reconnu efficacement l'AUG du start-stop en raison d'un « leaky-scanning » et qui n'auraient donc pas été affectées par le mécanisme d'inhibition lié au motif start-stop.

La faculté de la structure a11 à bloquer une particule 43S en cours de scanning en fait un outil de choix pour l'étude de l'initiation de la traduction par un mécanisme START qui sera présentée dans la partie 2.1.1.2.

1.4.1.6. Les structures secondaires situées dans la région 5'UTR qui inhibent la traduction canonique

La région 5'UTR des ARNm peut contenir des structures suffisamment stables qui peuvent soit empêcher la reconnaissance de la coiffe 5'-m^7G, soit bloquer la particule 43S en cours de scanning, soit défavoriser la reconnaissance du codon d'initiation. Dans le premier cas, l'accès à la coiffe m^7G nécessite l'activité hélicase des facteurs eIF4H, eIF4B et eIF4A qui déplient les régions structurées proximales de l'extrémité 5' de l'ARNm et/ou diverses protéines *trans*-régulatrices. Dans les deux autres cas, le niveau d'inhibition dépend de l'activité hélicase de la particule 43S, mais également de protéines remodelant ces structures secondaires, que l'on peut qualifier de chaperonnes d'ARN. Ces trois mécanismes seront illustrés dans les paragraphes qui suivent à partir d'exemples caractérisés dans la littérature.

<u>Structures qui empêchent la reconnaissance de la coiffe 5'-m^7G</u>

Des structures localisées à l'extrémité 5' des ARNm, couplées ou non à des protéines, peuvent induire une gêne stérique au niveau de la coiffe m^7G qui empêche sa reconnaissance par le complexe eIF4F.

Le premier exemple découvert se trouve à l'extrémité 5' des ARNm codant pour la ferritine : c'est l'élément de réponse au fer (IRE) (Hentze *et al.* 1987). L'IRE est formé par une tige-boucle de 55 nucléotides et constitue le site de fixation des protéines IRP-1 et IRP-2 (iron-regulated protein 1/2). La structure IRE étant située à seulement 10 nucléotides de la coiffe m^7G, la fixation de ces protéines sur l'IRE entraine une gêne stérique à l'extrémité 5' de l'ARN qui n'empêche pas le recrutement de eIF4F mais bloque le recrutement de la particule 43S en

interférant avec l'interaction eIF3-eIF4G (Muckenthaler *et al.* 1998). En l'absence des protéines IRP-1 et IRP-2, l'assemblage du complexe de pré-initiation sur la coiffe s'effectue normalement. La fixation de ces protéines à l'IRE est régulée par la faible concentration cellulaire de fer, la fixation de l'ion fer aux IRP réduisant leur affinité pour l'IRE. Par conséquent, ce mécanisme moléculaire participe activement au maintien de l'homéostasie de la concentration cellulaire en fer.

Un autre exemple a été mis en évidence dans la levure concernant la traduction de l'ARNm codant pour HAC1, un facteur de transcription impliqué dans l'activation de gènes impliqués dans la réponse aux protéines mal repliées. Ces gènes codent principalement pour des enzymes ou des chaperonnes résidentes du réticulum endoplasmique (Patil *et al.* 2004). La formation d'un complexe de pré-initiation de la traduction sur la coiffe m^7G de l'ARNm de HAC1 est inhibée par la présence d'une structure proximale (Uppala *et al.* 2022). Remarquablement, cette structure est formée par les interactions entre 11 nucléotides d'un intron, qui n'a pas été retiré lors de l'épissage de l'ARN pré-messager, et 11 nucléotides de la région 5'UTR (du premier exon) (Uppala *et al.* 2022). En conditions de stress générées par la présence de protéines mal repliées dans le réticulum endoplasmique, cet ARNm se rapproche d'une RNase transmembranaire du réticulum endoplasmique qui dégrade spécifiquement cet intron (Sidrauski and Walter 1997) de telle sorte que la jonction des exons 1 et 2 soit effectuée (Sidrauski *et al.* 1996). L'ARNm désormais dépourvu de la structure inhibitrice peut alors être traduit par un mécanisme canonique.

Enfin, la structure TAR (*trans*-activation response element) présente à l'extrémité de la région 5' UTR de HIV-2 serait responsable de l'inhibition du recrutement d'un complexe de pré-initiation au niveau de la coiffe de l'ARN viral (Soto-Rifo *et al.* 2012). La modulation temporelle de la structure du TAR par des protéines virales et/ou cellulaires au cours de l'infection serait un facteur déterminant pour la production des protéines virales et l'assemblage de nouvelles particules virales.

<u>Structures qui bloquent la particule 43S en cours de scanning</u>

Il existe peu d'exemples à l'heure actuelle.

On peut citer l'ARNm codant pour l'ornithine décarboxylase (ODC) chez le rat (Manzella and Blackshear 1990) dont la traduction est inhibée par une structure secondaire très stable de sa région 5'UTR. Cette structure est probablement impliquée dans le blocage de la particule 43S en cours de scanning, empêchant ainsi la reconnaissance du codon initiateur.

Un autre exemple est la structure a11 précédemment évoquée qui, en plus d'empêcher l'arrivée des eRF dans le site A du ribosome 80S piégé sur le codon AUG du motif start-stop, bloque les éventuelles particules 43S qui n'auraient pas reconnu l'AUG du start-stop en raison d'un « leaky scanning ».

Le dépliement de telles structures fait intervenir le recrutement d'hélicases additionnelles au niveau du complexe de scanning comme DDX3 ou DHX29. Etant donné que la suppression de l'activité hélicase de Ded1p (levure) ou de DDX3 (mammifères) est très délétère pour la traduction générale (Sen *et al.* 2015; Guenther *et al.* 2018; Sen *et al.* 2019), le dépliement des régions 5'UTR structurées semble nécessaire pour la traduction de nombreux ARN messagers. La modulation du niveau d'expression et d'activité des hélicases cellulaires constitue ainsi un autre paramètre essentiel de la régulation de la traduction.

Structures qui modulent la reconnaissance du codon initiateur

Certaines structures peuvent être considérées comme les équivalents eucaryotes des riboswitchs bactériens, qui, après la fixation d'un ligand, modulent l'accessibilité du codon AUG. De tels mécanismes ont été décrits chez les eucaryotes et concernent principalement la régulation d'IRES cellulaires. La fixation de protéines spécifiques, qualifiées d'IRES *trans*-acting factors (ITAFs), permet le remodelage de la structure et module ainsi l'accessibilité du codon initiateur.

La traduction de l'ARNm codant pour APAF1, une protéine impliquée dans le déclenchement de l'apoptose, serait régulée par un tel mécanisme (Coldwell *et al.* 2000), bien que des mécanismes coiffe-dépendants aient également été décrits pour cet ARNm, remettant ainsi en cause l'existence de cet IRES (Andreev *et al.* 2012). Un élément de réponse qui permettrait de lever cette ambiguïté serait l'importance du type cellulaire, et donc du protéome, sur la fonctionnalité de l'IRES de l'ARNm APAF1. Il est en effet davantage traduit dans les cellules neuronales (Mitchell *et al.* 2003). Le mécanisme d'initiation par l'IRES consiste en l'ouverture progressive de la structure secondaire de la région 5'UTR par les protéines UNR d'abord et PTB ensuite, ce qui permet l'assemblage du complexe 43S au niveau du site d'initiation de la traduction.

1.4.1.7. Les éléments permettant le recrutement d'un complexe de pré-initiation dans les ARNm cellulaires

L'initiation de la traduction par des IRES cellulaires apparait essentielle pour la traduction sélective de certains ARNm en conditions de stress, en relation avec les facteurs *trans*-régulateurs présentés dans la partie précédente. Les IRES cellulaires sont traduits par des mécanismes de traduction coiffe-indépendants apparus au cours de l'évolution dans les cellules eucaryotes. Leur fonctionnement est surtout lié à l'inhibition de la traduction coiffe-dépendante, soit par l'inactivation de eIF2 et de eIF4E dans le cadre des voies ISR et mTORC, soit par des éléments *cis*-régulateurs agissant de la sorte, comme les uORFs. De nombreux ARNm impliqués dans la réponse au stress ou dans le développement sont traduits par des IRES cellulaires (Godet *et al.* 2019). Globalement moins structurés que les IRES viraux, les IRES cellulaires font plutôt intervenir des facteurs protéiques nommés IRES *trans*-acting factors (ITAFs) pour favoriser le recrutement d'un complexe de pré-initiation de manière coiffe-indépendante et permettre ainsi la traduction sélective d'ARNm.

Certains mécanismes d'IRES cellulaires ont déjà été abordés. Les modifications m^6A dans la région 5'UTR des ARNm permettent le recrutement interne de la particule 43S par l'intermédiaire de l'interaction entre eIF3 et les m^6A, et sont particulièrement impliquées dans la traduction des ARN circulaires (Prats *et al.* 2020). Certains ITAFs permettent le remodelage de la structure des ARNm en jouant le rôle de chaperonnes d'ARN : un exemple a été abordé dans le paragraphe précédent avec la traduction de l'ARNm codant pour APAF1. D'autres mécanismes ont également été suggérés, parmi lesquels figurent la modulation de la localisation cellulaire de certains ITAFs et/ou du complexe de traduction, ou encore l'interaction de certains ITAFs avec certaines protéines ribosomiques qui guident le recrutement du ribosome.

L'implication prononcée du protéome dans les mécanismes de traduction coiffe-indépendants serait indicatrice d'une forte dépendance de la traduction de ces IRES au type cellulaire, dont le protéome est caractéristique.

1.4.2. Eléments de la phase codante

1.4.2.1. La présence d'un codon stop prématuré

La présence d'un codon stop prématuré dans une phase codante, résultat d'un intron non excisé, d'une mutation, d'une erreur de polymérisation de l'ARN polymérase II, ou parfois d'un décalage de cadre de lecture, déclenche une série de mécanismes connue sous le nom de « non-sense mediated decay » (NMD) qui mène à la dégradation de l'ARNm et au retrait de la coiffe m^7G (Lykke-Andersen and Jensen 2015). Le NMD fait intervenir en premier lieu l'interaction du facteur eRF3 avec la protéine UPF1 qui recrute ensuite des protéines responsables de la dégradation de l'ARNm et du clivage de la coiffe. La distinction entre un codon stop canonique et un codon stop prématuré repose sur deux caractéristiques. La première fait intervenir la présence du complexe EJC (Exon Junction Complex) qui est composé de protéines déposées par la réaction d'épissage des ARN pré-messagers sur les jonctions exon-exon. Lorsque le complexe EJC n'est pas présent à proximité du codon stop, il est reconnu comme prématuré, ce qui enclenche le NMD. Le second fait intervenir des interactions entre la PABP (Poly-A Binding Protein) accrochée à l'extrémité 3' des ARNm, le facteur eRF3 et dans certains cas le facteur eIF4G qui empêchent le recrutement des protéines du NMD. Un codon stop suffisamment proche de la queue poly-A est ainsi interprété comme canonique, alors qu'un codon stop trop distant est reconnu comme prématuré. De manière générale, les mécanismes de discrimination des codons stop canoniques des prématurés sont encore mal compris (Lykke-Andersen and Jensen 2015; Hug *et al.* 2016).

1.4.2.2. Les séquences modulant la vitesse d'élongation

En plus des aspects *trans*-régulateurs déjà évoqués dans les paragraphes 1.3.5 et 1.3.6, la nature des codons joue un rôle majeur dans la vitesse d'élongation, en relation avec la concentration cellulaire en aa-ARNt cellulaires correspondants. En effet, l'abondance des aa-ARNt d'une cellule eucaryote ne corrèle pas nécessairement avec l'abondance des codons qui sont utilisés dans le protéome. Un codon qui doit être décodé par un aa-ARNt rare, c'est-à-dire peu abondant dans la cellule, entrainera un ralentissement de la vitesse d'élongation de la traduction.

Dans une situation physiologique, la modulation de la vitesse d'élongation par l'usage des codons permet, en plus de l'acide aminé incorporé dans le peptide naissant, de moduler le repliement de la protéine synthétisée en sortie du ribosome. La position des zones de ralentissement voire de pause du ribosome lors de la synthèse n'est pas aléatoire, et peut par exemple être localisée entre deux domaines protéiques. Dans ce cas, la pause du ribosome permet de réguler le repliement d'un domaine de la protéine synthétisée en sortie du ribosome. Ainsi, la phase codante des ARNm code pour une protéine mais comporte également des informations de repliement par l'intermédiaire de l'usage de codons décodés par des ARNt rares qui sont localisés de telle sorte que leur lente cinétique de décodage assure le repliement fonctionnel de la protéine en cours de synthèse.

La synthèse des protéines destinées à être maturées dans le réticulum endoplasmique permet d'illustrer le principe selon lequel l'usage des codons est essentiel au repliement co-traductionnel des protéines. En effet, l'existence d'un cluster de codons décodés par des ARNt peu abondants situé 35 à 40 codons en aval du domaine dont le repliement co-traductionnel forme le motif hydrophobe reconnu par la SRP à la sortie du ribosome induit le ralentissement de la vitesse d'élongation au niveau de ces codons (Pechmann *et al.* 2014). Il s'avère que ces

35 à 40 codons correspondent exactement à la distance séparant le site P de l'extrémité du tunnel de sortie du peptide naissant (Figure 29). Ainsi, le ralentissement local de la vitesse d'élongation généré par ce cluster de codons est essentiel au repliement fonctionnel du motif hydrophobe afin qu'il puisse être reconnu par la SRP, menant ensuite à l'ancrage du peptide naissant dans la membrane du réticulum endoplasmique et à sa maturation.

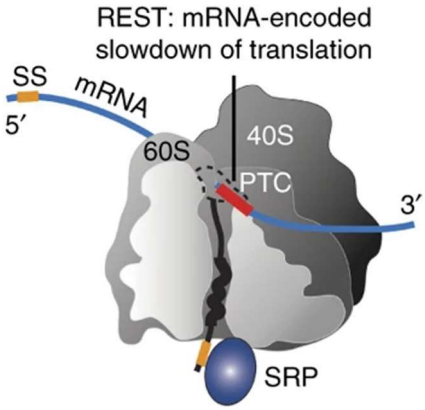


Figure 29 : schéma illustrant le ralentissement programmé du ribosome pour la reconnaissance du motif hydrophobe par la SRP, permettant ensuite l'ancrage du complexe de traduction dans la membrane du réticulum endoplasmique. PTC : site peptidyl-transférase ; SRP : particule de reconnaissance du signal ; SS : séquence du signal. D'après (Pechmann *et al.* 2014).

Plus généralement, la nature des premiers (resp. derniers) codons de la phase codante influence également la vitesse de traduction en tout début (resp. toute fin) d'élongation (Tuller *et al.* 2010). Cette étude a montré que, chez les eucaryotes, les 30 à 50 premiers acides aminés sont traduits à une vitesse relativement faible, suggérant que l'élongation nécessite le décodage d'un certain nombre de codons avant d'atteindre ce qu'on pourrait nommer le régime permanent. Il s'avère que ces premiers codons sont décodés par des aa-ARNt rares, et qu'une telle caractéristique est conservée chez les organismes testés. La faible vitesse de traduction des premiers codons serait indicatrice d'un couplage initiation-élongation qui pourrait favoriser le positionnement ou la stabilisation du ribosome dans la bonne phase de lecture.

Par ailleurs, certaines séquences ou structures secondaires qui provoquent le ralentissement du ribosome en cours d'élongation peuvent être impliquées dans des décalages du cadre de lecture qui sont essentiels à la traduction de certaines protéines virales (paragraphe suivant).

1.4.2.3. Les séquences de décalage de cadre de lecture

Certaines séquences des ARNm peuvent induire un décalage du cadre de lecture des ribosomes lors de la phase d'élongation, en interférant avec le positionnement des ARNt dans les sites P et A du ribosome. Le taux de décalage de cadre de lecture est étroitement lié à la vitesse d'élongation, qui elle-même dépend des facteurs *cis*- et/ou *trans*- régulateurs déjà évoqués.

On distingue les décalages de cadre spontanés qui mènent très probablement à la rencontre d'un codon stop prématuré et donc à une protéine non-fonctionnelle, des décalages dits programmés qui permettent la synthèse d'une protéine fonctionnelle.

Ces décalages peuvent être induits par de courts motifs nucléotidiques et/ou par des structures secondaires. Presque tous les virus à ARN ont évolué vers ce mécanisme pour la synthèse de plusieurs protéines à partir du seul ARN génomique.

Décalage induit par des motifs nucléotidiques

Certains motifs nucléotidiques contiennent des codons décodés par des ARNt rares, ce qui provoque un ralentissement accru du ribosome à cet endroit. Un tel mécanisme a été décrit dans le cadre de la traduction du rétrotransposon Ty1 chez la levure qui contient un motif responsable du décalage de la phase de lecture (Belcourt and Farabaugh 1990). La séquence de décalage, CUU AGG C, contient en particulier le codon AGG décodé dans le site A par un ARNt-Arg très peu abondant dans la levure. En résulte une pause du ribosome à cet endroit, puis un décalage de la phase +0 (CCU) vers la phase +1 (UUA).

On peut noter qu'un déficit d'acides aminés dans la cellule résulterait en l'appauvrissement généralisé de tous les aa-ARNt et donc à une probabilité accrue de décalage de cadre de lecture lors de la traduction des ARNm. Cela mènerait à la synthèse abondante de protéines au repliement aberrant, qui activerait ensuite la réponse cellulaire aux protéines mal repliées. Ce phénomène est atténué par la cellule qui active des voies de stress en réponse à une carence d'acides aminés par les voies ISR (PERK et GCN2) et mTORC1, ce qui conduit à l'inhibition générale de la traduction cellulaire et à la traduction d'ARNm spécifiques. Si la carence est trop prolongée, la cellule enclenche les voies d'apoptose.

Décalage induit par des structures secondaires

Le décalage programmé du cadre de lecture est essentiel pour la synthèse des protéines virales. Par exemple, la synthèse de l'ORF *pol* du HIV-1, dont découlent les enzymes du HIV, est dépendante d'un décalage de cadre de lecture programmé au cours de la traduction de l'ORF *gag,* dont découlent certaines protéines structurales (Jacks *et al.* 1988). Ce décalage fait intervenir l'action synergique de deux éléments *cis*-régulateurs : une séquence de glissement U UUU UUA séparée d'une structure en tige-boucle de 26 nucléotides par un « spacer » de 8 nucléotides. Le rôle de la structure en tige-boucle est de retenir le ribosome sur la séquence de glissement. L'efficacité de cette pause dépend de la stabilité de la tige-boucle et de la distance la séparant de la séquence de glissement, si bien que cette dernière se retrouve dans le site A (Mouzakis *et al.* 2013). La distance séparant les deux éléments *cis*-régulateurs est donc critique pour l'induction du décalage de cadre de lecture. L'arrêt du ribosome induit par la tige-boucle est suffisamment long pour permettre, dans environ 10% des cas (Plant and Dinman 2006; Mathew *et al.* 2015), des réarrangements des appariements codons-anticodons localisés dans les sites P (codon UUU) et A (codon UUA) vers le cadre de lecture -1 (site P : UUU, site A : UUU). Un mécanisme similaire est impliqué lors de la traduction de l'ORF1a du SARS-CoV-2, au cours de laquelle le décalage de cadre permet la traduction de l'ORF1b, à la différence près que la structure située en aval de la séquence de glissement se replie en un pseudo-nœud (Bhatt *et al.* 2021).

1.4.2.4. Les structures secondaires proximales situées en aval du codon initiateur

Le mécanisme START, déjà évoqué pour l'initiation sur les uAUG, peut aussi intervenir dans l'initiation de la traduction de la phase codante principale.

Dans le cas de la sclérose latérale amyotrophique, des séquences répétées GGGGCC dans un intron du gène C90RF72 provoquent un défaut d'épissage qui entraîne l'accumulation d'un ARNm aberrant à partir du gène C90RF72 (Tabet *et al.* 2018). Ce transcrit étant coiffé du côté 5', un complexe de pré-initiation de la traduction peut s'assembler et entame le scanning. Les séquences GGGGCC formant des structures très stables contenant des G-quartets, la progression du complexe de scanning est stoppée si bien qu'un codon CUG se retrouve dans le site P du ribosome et peut de ce fait être reconnu comme un codon d'initiation. L'initiation sur ce codon CUG dépend donc de la présence des G-quartets en aval qui stabilisent le complexe de scanning sur ce codon (Figure 30). La traduction initiée à partir de ce codon entraine la synthèse de poly-dipeptides hautement toxiques pour les motoneurones.

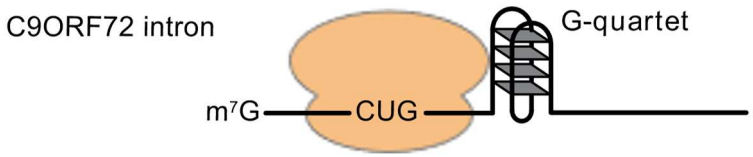


Figure 30 : initiation de la traduction des ARNm aberrants C9ORF72 sur un codon CUG par un mécanisme de type START D'après (Eriani and Martin 2018).

Plus généralement, des prédictions de structures secondaires réalisées sur les 65 premiers nucléotides situés en aval du codon d'initiation AUG des ORFs annotées de divers génomes eucaryotes suggèrent que le mécanisme START pourrait être impliqué dans l'initiation de la traduction de 0.5 à 2% des ORFs eucaryotes (Despons and Martin 2020). Par ailleurs, des expériences de ribosome profiling montrent que l'utilisation de codons non-AUG est très répandue dans la traduction d'ORFs, d'uORFs et de dORFs (downstream ORFs) chez les eucaryotes (Ingolia *et al.* 2011; Chen *et al.* 2020; Chothani *et al.* 2022). Le degré d'implication du mécanisme START dans l'initiation de leur traduction n'est en revanche pas connu. Les travaux de Marylin Kozak avaient déjà suggéré le rôle de telles structures pour l'initiation de la traduction sur des codons AUG ou non-AUG situés dans des contextes nucléotidiques plus ou moins favorables (Kozak 1990).

Une partie de ce travail de thèse est dédiée à l'identification des paramètres cis-régulateurs requis pour initier la traduction par le mécanisme START (partie 2.1.1.2).

1.4.3. Eléments de la région 3'UTR

1.4.3.1. Les phases codantes situées en aval du codon stop principal (dORF)

La traduction de phases codantes situées dans la région 3'UTR, en aval du codon stop de la phase principale (nommées dORF pour « downstream ORF »), a été mise en évidence par des expériences de ribosome profiling (Bazzini *et al.* 2014; Mackowiak *et al.* 2015; Chen *et al.* 2020; Chothani *et al.* 2022). Les dORF, à l'instar des uORF, sont essentielles à la régulation de la traduction de la phase codante principale, et le peptide résultant de leur traduction peut également être un facteur *trans*-régulateur (Chen *et al.* 2020). Là où le nombre d'uORF corrèle généralement avec l'inhibition accrue de la traduction de la phase codante principale, le nombre de dORF corrèle quant à lui avec la traduction accrue de l'ORF principale (Wu *et al.* 2020b). Le mécanisme d'initiation de la traduction sur les dORFs n'est pas le résultat d'un

défaut de reconnaissance du codon stop de la phase principale (Chen *et al.* 2020; Wu *et al.* 2020b), et peut impliquer l'utilisation de codons non-AUG (Chen *et al.* 2020; Chothani *et al.* 2022). L'initiation de la traduction des dORF serait alors nécessairement coiffe-indépendante et pourrait faire intervenir les mécanismes de type IRES précédemment évoqués.

1.4.3.2. La PABP et la queue poly-adénosines

Leurs principaux rôles ont déjà été évoqués. Ils interviennent dans la circularisation des ARNm lors de l'initiation de la traduction. La PABP (Poly-A Binding Protein) est en particulier une cible de certains facteurs *trans*-régulateurs, qu'ils soient endogènes ou exogènes, notamment lors d'une infection virale (Figure 20). Enfin, la PABP est certainement impliquée dans l'inhibition du NMD lors de la reconnaissance des codons stops canoniques (Hug *et al.* 2016).

1.4.3.3. Les séquences complémentaires d'ARN interférents

Les séquences cibles des ARN interférents sont en général localisées dans les régions 3'UTR des ARNm. L'hybridation des ARN interférents permet le recrutement de protéines impliquées dans l'inhibition de la traduction de l'ARNm, soit en induisant sa dégradation, soit en interférant avec des protéines essentielles pour sa traduction (voir paragraphe 1.3.4 et Figure 21).

1.4.3.4. Les motifs de localisation cellulaire des ARNm

L'importance de la localisation cellulaire de la traduction, qui passe par le transport des ARNm vers la localisation en question, a déjà été abordée dans le paragraphe 1.2.5. L'adressage des ARNm nécessite l'assemblage de protéines qui reconnaissent des motifs nucléotidiques et structuraux précis de la région 3'UTR des ARNm concernés. La localisation cellulaire des ARNm et de la traduction est primordiale pour certaines cellules spécialisées comme les neurones, ou pour la régulation spatio-temporelle de l'expression génétique au cours du développement embryonnaire.

1.5. La coordination des mécanismes de régulation permet la traduction sélective des ARNm

Une conséquence de l'action conjointe des mécanismes *cis*- et *trans*- régulateurs est la traduction sélective d'un ensemble d'ARN messagers en fonction de l'état physiologique de la cellule et est ainsi déterminante pour le maintien de son intégration fonctionnelle dans un organisme. Selon les informations environnementales captées par certaines protéines transmembranaires et/ou par diverses protéines intracellulaires, la cellule modifie son protéome en conséquence. L'adaptation de ce schéma traductionnel passe notamment par l'inactivation de certains facteurs de traduction impliqués dans le mécanisme d'initiation canonique, à savoir eIF2 par le biais de la voie ISR, et eIF4E par le biais de la voie mTORC1. En éteignant la traduction d'une majorité de ses ARN messagers de la sorte, la cellule redirige les ressources traductionnelles vers la synthèse de protéines essentielles à la prise en compte des signaux qu'elle a initialement intégrés, parmi lesquelles figurent des facteurs de transcription qui permettent de moduler le transcriptome et ainsi d'orchestrer la suite de la réponse adaptative. C'est pourquoi la synthèse des facteurs de transcription est, la plupart du temps, finement régulée au niveau traductionnel. La modulation du protéome cellulaire peut donc être vue comme le premier maillon de la réponse adaptative de la cellule à des changements environnementaux.

La traduction des ARN messagers dans ces conditions fait nécessairement appel à des mécanismes d'initiation coiffe-indépendants, parmi lesquels les IRES cellulaires occupent une place prépondérante.

Dans le cas particulier d'une infection virale, la traduction cellulaire est détournée vers la traduction des ARN viraux par l'action conjointe de facteurs *trans*-régulateurs, qu'ils soient d'origine cellulaire ou virale, et/ou *cis*-régulateurs.

Il a été proposé que des ribosomes spécialisés puissent constituer une autre facette de la sélectivité traductionnelle (Shi *et al.* 2017). La spécialisation des ribosomes dépend de l'ensemble des protéines ribosomiques qui les constitue et déterminerait ainsi l'ensemble des ARNm qu'ils sont capables de traduire, et notamment ceux arborant des IRES.

Enfin, un ralentissement global de la traduction enclenche une réaction de la cellule, qui, selon la gravité du déficit énergétique qui en est généralement la cause, peut mener à l'activation des voies d'apoptose. Par conséquent, le taux de traduction d'une cellule peut être vu comme un baromètre de son état physiologique.

1.6. Objectifs de la thèse

Les travaux réalisés dans le cadre de ce travail de thèse ont consisté à rechercher et à caractériser des éléments *cis*-régulateurs de la traduction eucaryote à l'aide d'approches *in vitro* et de méthodes de criblage à haut-débit.

Le premier volet de ce travail est dédié à la recherche de séquences qui modulent la reconnaissance du codon initiateur par le ribosome (partie 2.1).

Le contexte nucléotidique du codon initiateur est un élément *cis*-régulateur essentiel pour sa reconnaissance par la particule 43S en cours de scanning. L'analyse des contextes des AUG des phases codantes annotées dans les génomes eucaryotes, bien qu'informative, ne donne pas d'information sur l'efficacité traductionnelle de ces contextes, car ils peuvent très bien avoir été sélectionnés au cours de l'évolution pour d'autres raisons que leur efficacité. Ainsi, les contextes nucléotidiques les plus favorables ou défavorables à l'initiation de la traduction sur un codon AUG ne sont pas vraiment connus. Pour les identifier, nous avons réalisé un criblage d'une banque d'ARN rapporteurs contenant l'ensemble des 262144 combinaisons NNNNNNAUGNNN, avec N = {A ; U ; G ; C}, à l'aide d'un dispositif micro-fluidique de manière à déterminer quelles sont les combinaisons qui sont acceptées et validées par le ribosome, ou au contraire celles qui sont rejetées. Ce dispositif permet de miniaturiser dans des microgouttes des réactions de traductions *in vitro* réalisées à l'aide d'extraits acellulaires de traduction préparés à partir de lysats de cellules HEK293FT. Le séquençage des variants sélectionnés à l'issue des criblages permet d'identifier les contextes les plus favorables.

Par ailleurs, des études de ribosome profiling ont démontré que près de la moitié des initiations de la traduction dans la cellule a lieu sur un codon AUG, démontrant ainsi que des codons non-AUG sont aussi largement utilisés par la cellule, notamment pour la traduction d'uORFs et de dORFS (Ingolia *et al.* 2011; Chen *et al.* 2020; Chothani *et al.* 2022). Dès lors, nous avons cherché à déterminer dans quelle mesure le mécanisme START (Eriani and Martin 2018; Despons and Martin 2020) qui permet la traduction de codons sous-optimaux, est mis en œuvre pour l'initiation de la traduction sur un codon non-AUG.

Enfin, nous avons initié une collaboration avec l'équipe de Karen Perronet à Paris avec pour objectif d'étudier la vitesse de scanning du ribosome par des approches d'études en molécules uniques couplées à de la microscopie TIRF (Total Internal Reflection Fluorescence). Ce projet qui en est encore à ses débuts est présenté brièvement dans la partie 2.4.1.

Le second volet de ce travail est dédié à la mise au point d'une méthode de criblage permettant d'identifier de manière systématique des IRES dans un génome viral (partie 2.2).

Les techniques actuelles d'identification d'IRES viraux s'appuient sur le criblage par FACS (Fluorescence-Activated Cell Sorting) de cellules transfectées par des plasmides contenant des gènes rapporteurs, couplé au séquençage à haut-débit des variants sélectionnés. Le criblage par FACS de cellules transfectées avec une banque de plasmides codant pour des rapporteurs bicistroniques a été employé pour identifier de nombreuses séquences de traduction coiffe-indépendantes cellulaires ou virales (Weingarten-Gabbay *et al.* 2016). Néanmoins, un problème de cette approche est la potentielle présence de promoteurs cryptiques et/ou de sites de clivage dans les régions candidates qui génèrent des ARNm

monocistroniques à l'origine de nombreux faux-positifs. Par ailleurs, la taille des banques criblées est limitée à l'efficacité de transfection des cellules, ce qui force leur conception rationnelle et ne permet pas de cribler des génomes viraux de manière systématique. Enfin, la transfection de cellules ne donne pas beaucoup de marge de manœuvre dans la conception de stratégies expérimentales permettant d'évaluer l'impact de facteurs *trans*-régulateurs sur la traduction, contrairement aux approches *in vitro*.

Une nouvelle méthode de criblage pour l'identification d'IRES viraux a été développée dans le cadre de ce travail : elle repose sur le fractionnement de complexes de traduction sur gradient de sucrose. La mise au point de cette méthode a été réalisée à partir du génome du virus de la paralysie du cricket (CrPV) qui contient deux IRES bien caractérisés dans la littérature. Le criblage de la banque de rapporteurs synthétisée à partir de ce génome modèle a été réalisé avec des extraits de traduction *in vitro* préparés à partir de lysats de réticulocytes de lapin et de lysats de cellules embryonnaires de drosophile (S2), plus pertinents pour le criblage du génome du CrPV puisqu'il infecte la drosophile. A terme, ces méthodes seront appliquées à l'étude d'autres génomes viraux.

Le troisième volet de ce travail est dédié à l'étude de la traduction de l'ARN du SARS-CoV-2 et de modèles d'ARN cellulaires en présence de la protéine virale NSP1, caractérisée comme étant un inhibiteur de la traduction (partie 2.3).

Les résultats de ces trois volets seront présentés sous la forme de manuscrits en vue d'une soumission prochaine (parties 2.1 et 2.2) ou d'articles déjà publiés (partie 2.3). Une dernière partie est dédiée aux autres projets pour lesquels ce travail de thèse a contribué (partie 2.4).

Le chapitre suivant s'organise en quatre parties indépendantes :

- Caractérisation d'éléments *cis*-régulateurs modulant la reconnaissance du codon initiateur par le ribosome
- Recherche systématique d'IRES dans un génome viral
- Etude de l'impact de la protéine NSP1 sur la traduction des ARNm cellulaires et sur la traduction de l'ARN génomique du SARS-CoV-2
- Contributions à d'autres projets de recherche

RESULTATS
DISCUSSIONS
PERSPECTIVES

2. Résultats, discussions et perspectives

2.1. Caractérisation d'éléments *cis*-régulateurs modulant la reconnaissance du codon initiateur par le ribosome

2.1.1. Introduction du projet

2.1.1.1. Etude du contexte nucléotidique du codon AUG

Le contexte nucléotidique du codon initiateur est un élément *cis*-régulateur essentiel pour sa reconnaissance par la particule 43S en cours de scanning. L'analyse des contextes des AUG des phases codantes annotées dans les génomes eucaryotes, bien qu'informative, ne donne pas d'information sur l'efficacité traductionnelle de ces contextes, car ils peuvent très bien avoir été sélectionnés au cours de l'évolution pour d'autres raisons que leur efficacité. Ainsi, les contextes nucléotidiques les plus favorables ou défavorables à l'initiation de la traduction sur un codon AUG ne sont pas vraiment connus à ce jour.

Afin de caractériser des contextes nucléotidiques favorables ou défavorables à l'initiation de la traduction, nous avons réalisé un criblage d'une banque d'ARN rapporteurs contenant l'ensemble des 262 144 combinaisons NNNNNNAUGNNN, avec N = {A ; U ; G ; C}, à l'aide d'un dispositif micro-fluidique. Le criblage s'effectue par la miniaturisation dans des microgouttes de réactions de traductions *in vitro* réalisées à l'aide d'extraits préparés à partir de lysats de cellules HEK293FT. Le séquençage des variants sélectionnés à l'issue du criblage permet d'identifier les contextes les plus favorables.

Les résultats ont confirmé le caractère essentiel du contexte nucléotidique du codon d'initiation pour le démarrage de la traduction car seulement 10% des combinaisons NNNNNNAUGNNN de la banque de départ ont été sélectionnées après trois tours de criblage. Ensuite, la diversité des séquences sélectionnées suggère l'absence de séquence consensus, ou de motif minimal, qui permettrait d'expliquer l'efficacité traductionnelle de ces séquences. En d'autres termes, chaque arrangement de nucléotides aux positions -6 à +6 exerce un effet unique sur l'initiation de la traduction. L'analyse des phases codantes annotées du génome humain a montré que seules 11% des combinaisons NNNNNNAUGNNN possibles sont utilisées. Parmi ces 11%, le contexte le plus retrouvé n'est pas une séquence ayant le consensus Kozak. Dans l'ensemble, la diversité des 28454 contextes nucléotidiques distincts utilisés pour initier la traduction des phases codantes humaines suggère aussi l'absence de contexte nucléotidique consensus pour le démarrage de la traduction, même si une certaine préférence a été observée sur les C entourant la position -3, en particulier aux positions -2 et -1.

2.1.1.2. Implication du mécanisme START dans l'initiation de la traduction sur un codon non-AUG

Des études de ribosome profiling ont démontré que près de la moitié des initiations de la traduction dans la cellule a lieu sur un codon non-AUG, démontrant ainsi que ces codons sont largement utilisés dans la cellule, notamment dans le cadre de la traduction d'uORFs et de dORFS (Ingolia *et al.* 2011; Chen *et al.* 2020; Chothani *et al.* 2022). Dès lors, nous avons cherché à déterminer dans quelle mesure le mécanisme START (Eriani and Martin 2018; Despons and Martin 2020), qui permet la traduction à partir de codons sous-optimaux, est mis en œuvre pour l'initiation de la traduction sur un codon non-AUG.

En utilisant la structure a11 pour bloquer une particule 43S en cours de scanning, nous avons déterminé les paramètres critiques permettant l'initiation de la traduction par un mécanisme START avec des expériences de traduction *in vitro* en utilisant différents extraits acellulaires.

Les résultats montrent que la présence d'une structure secondaire suffisamment stable, située à une distance adéquate du codon d'initiation, permet d'initier la traduction sur certains codons non-AUG comme les codons CUG ou ACG. De plus, la stabilité minimale de la structure permettant d'initier la traduction sur un codon non-AUG par un mécanisme START est dépendante de l'extrait acellulaire utilisé pour la traduction *in vitro*, suggérant le rôle du contenu cellulaire, et notamment celui en ARN hélicases, dans la modulation des paramètres *cis*-régulateurs du mécanisme START.

2.1.2. Etude d'éléments *cis*-régulateurs critiques pour la reconnaissance du codon initiateur chez les eucaryotes

Ces résultats sont présentés sous la forme d'un manuscrit en vue d'une soumission d'ici peu.

MANUSCRIT 1

Etude d'éléments *cis*-régulateurs critiques pour la reconnaissance du codon initiateur chez les eucaryotes

Investigation of critical *cis* regulators of start codon recognition in eukaryotes

Antonin Tidu, Natacha Dentz, Fatima Alghoul, Laurence Despons, Michael Ryckelynck, Gilbert Eriani, Franck Martin*

Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, Architecture et Réactivité de l'ARN, CNRS UPR9002, 2 allée Konrad Roentgen, F-67084 Strasbourg (France)

*Corresponding authors

Email for correspondence:

f.martin@ibmc-cnrs.unistra.fr

Running title: *Cis*-acting elements in eukaryotic translation initiation

Keywords: Ribosome, Translation initiation, Kozak consensus, AUG/non-AUG codons

Abstract

In eukaryotes, translation initiation is a highly regulated process, which combines *cis*-regulatory sequences located on the messenger RNA along with *trans*-acting factors like eucaryotic initiation factors (eIF). One critical step of the process of translation initiation is the start codon recognition by the scanning 43S particle, which leads to ribosome assembly and protein synthesis. Start-codon selection stringency involves *trans*- and *cis*- regulators which both modulate the stability of the codon-anticodon interaction in the P-site and therefore the assembly of the ribosome. In this work, we investigated the influence of the start codon proximal sequences on translation initiation. First, we studied the AUG nucleotide context using a microfluidic-based *in vitro* screening method of an RNA reporter library featuring a randomized AUG context (NNNNNNAUGNNN), where N is any of the four ribonucleotides. We coupled this study with bioinformatic analysis of AUG contexts of the annotated ORFs in the human genome. Our results suggest that although the AUG nucleotide context does play a critical role, it does not feature a minimal sequence pattern that can explain translation initiation efficiency on its own. Next, we investigated the involvement of secondary structures downstream the initiation codon in the so-called the START (Structure-Assisted RNA translation) mechanism on non-AUG translation initiation. The results demonstrate that downstream secondary structures can efficiently promote non-AUG translation initiation provided they are stable-enough to stall a scanning 43S particle and located at an optimal distance from this non-AUG codon to trigger and stabilise the codon-anticodon base-pairing in the P site.

1. Introduction

In eukaryotes, translation initiation is a highly regulated process, which combines *cis*-regulatory sequences located on the messenger RNA along with *trans*-acting factors like eukaryotic initiation factors (eIF). The canonical translation initiation mechanism starts with the recognition of the 5' m^7G cap by eIF4E, which is associated with eIF4G and eIF4A to form the eIF4F complex. The interaction between eIF4G and 43S-bound eIF3, enables the recruitment of the 43S particle on the m^7G cap. Then, the 43S particle scans the 5'UTR thanks to the ATP-dependent RNA helicase activity of eIF4A, which is enhanced through its interactions with eIF4G, eIF4B and eIF3 (Rogers *et al.* 2001; Schütz *et al.* 2008). During scanning, the 43S particle is maintained in a so-called open conformation by the combined actions of eIF1A and eIF1 which destabilize the codon-anticodon interaction in the P-site, inhibit spontaneous hydrolysis of eIF2-bound GTP (Algire *et al.* 2005) and therefore prevent ribosomal subunits joining (Pestova *et al.* 1998; Thakur and Hinnebusch 2018; Zhou *et al.* 2020). The codon-anticodon interaction involves the nucleotide triplet from the codon being analysed in the P-site and the anticodon of the initiator $tRNA^{Met}_i$ which is bound to eIF2-GTP in the 43S particle (Lomakin and Steitz 2013). When a stable-enough codon-anticodon base-pairing is established during scanning, it displaces eIF1 from the P-site and triggers the switch from the open to the close conformation of the 43S. This results in the activation of the GTPase activity of eIF2 by eIF5 and in the subsequent hydrolysis of eIF2-bound GTP into GDP (Paulin *et al.* 2001). As GTP hydrolysis goes on, eIF2 loses its affinity for the Met-$tRNA^{Met}_i$ and is released from the 48S particle (Kapp and Lorsch 2004). Subunits joining is mediated by eIF5B (Pestova *et al.* 2000) and its interaction with eIF1A (Nag *et al.* 2016), which contributes to release eIF2, eIF1, eIF1A and eIF5B from the initiation complex through the hydrolysis of eIF5B-bound GTP. The assembled 80S ribosome then starts protein synthesis by incorporating the first amino-acyl-tRNA in its vacant A-site.

Start-codon selection stringency is mediated by eIF1 and eIF1A which prevent unstable codon-anticodon interactions like those involving non-cognate codons (Pestova *et al.* 1998; Thakur and Hinnebusch 2018; Zhou *et al.* 2020). However, even AUG codons which establish a perfect Watson-Crick base-pairing with the anticodon loop of the Met-$tRNA^{Met}_i$ are sometimes not recognized as the initiation codon. Moreover, several studies (Tang *et al.* 2017; Tabet *et al.* 2018) along with ribosome profiling data (Ingolia *et al.* 2011; Chen *et al.* 2020; Chothani *et al.* 2022) suggest that non-AUG codons can be used for translation initiation in eukaryotes. Therefore start-codon selection is influenced by additional *trans* and *cis* factors. Several reports, starting with the initial studies from Kozak (Kozak 1986, 1987), suggested the start codon nucleotide context plays a major role on start codon recognition efficiency. The established human consensus nucleotide context GCCRCCAUGR (R is A or G) was obtained by averaging the nucleotide contexts of all ORFs. Particularly, -3 and +4 positions (the A of the AUG is +1 by convention) were reported critical for efficient start codon recognition (Kozak 1986, 1987), and the purine at -3 is conserved in eukaryotes (Hernández *et al.* 2019). In addition, structural reports demonstrated in particular interactions between the -3 position and eIF2α subunit (Simonetti *et al.* 2020; Thakur *et al.* 2020). Distal *cis*-regulatory elements may also influence start codon selection stringency. For example, translation of C9ORF72 transcripts into toxic poly-dipeptides is initiated on a CUG codon that is upstream of very stable RNA structures that contain G-quartets (Tabet *et al.* 2018). To that respect, we recently described a mechanism of translation initiation called STructure Assisted RNA Translation or START (Eriani and Martin 2018; Despons and Martin 2020). START initiation relies on the

presence of a downstream secondary structure that enhances initiation and can improve initiation on non-AUG codons by stabilizing the 43S particle with those codons in the P-site.

In this work, we investigated the influence of the start codon proximal sequences on translation initiation. First, we studied the AUG nucleotide context using a microfluidic-based *in vitro* screening of an RNA reporter library featuring a randomized AUG context (NNNNNNAUGNNN) coupled with bioinformatic analysis of the annotated ORFs in the human genome. Our results suggest that although the AUG nucleotide context does play a critical role, it does not feature a minimal sequence pattern that can explain translation initiation efficiency on its own. Moreover, the analysis of the AUG nucleotide context of annotated human ORFs revealed a strong selection pressure towards cytidines flanking the -3 position, especially those at positions -2 and -1. Additionally, the most represented AUG nucleotide context in the human ORFs is not a Kozak sequence, and overall, the nucleotide contexts of the AUG contexts in the human ORFs are very diverse, further supporting the absence of a broad consensus sequence. Next, we investigated the involvement of the START mechanism on non-AUG translation initiation. The results demonstrate that downstream secondary structures can efficiently promote non-AUG translation initiation provided that they are stable-enough to stall a scanning 43S particle and located at an optimal distance from this non-AUG codon to trigger and stabilise the codon-anticodon base-pairing in the P site.

2. Material and methods

2.1. Oligonucleotides information

All the oligonucleotides' sequences used in this study are provided in supplementary material (Table S1). Oligonucleotides were purchased by Integrated DNA Technologies company. In the text, oligonucleotides are referred to by their red numbers in the Table S1.

2.2. Cell-free translation extracts preparation

2.2.1. Rabbit reticulocytes lysates

Untreated rabbit reticulocytes lysates (RRL) were prepared as previously described (Pelham and Jackson 1976).

2.2.2. HEK293FT and SH-SY5Y cell lysates from cells cultured in physiological conditions

HEK293FT cells were seeded at 0.03 x10^6 cells/cm² and grown for 2 days at 37°C in a 5% CO_2 humidified atmosphere. The culture medium is Gibco Dulbecco's Modified Eagle Medium + GlutaMAX (Life Technologies) that is supplemented with 10% of inactivated fetal bovine serum (Life Technologies). Neuroblastoma SH-SY5Y cells were cultured in the same conditions and same medium but supplemented with 1 mM of non-essential amino acids (Gibco). The total culture surface was 4500 cm², leading approx. to 200 to 400 x10^6 cells before harvesting. Cells were harvested at room temperature in their culture medium by centrifugation (300g) at 4°C and washed two times with a cold buffer containing 20 mM HEPES–KOH pH 7.5, 100 mM potassium acetate, 2 mM magnesium acetate, 1 mM DTT. After washing, the cell pellets were resuspended in the same buffer supplemented or not (if TEV protease is be used, see later) with 1X Halt™ Protease Inhibitor Cocktail EDTA-free (Thermo Scientific™) to reach a concentration of 50-100 x10^6 cells/mL. Cells were lysed by nitrogen cavitation with a Cell Disruption Bomb (Parr Instrument Company) after a one-hour incubation under a pressure of 30 bars at 4°C. The lysate was cleared by centrifugations at 10,000g at 4°C, aliquoted, flash-frozen in liquid nitrogen and stored at −80°C. Total protein concentration was determined with Bradford assay (Biorad).

2.2.3. HEK293FT cell lysates from cells cultured in stress conditions

To prepare cell-free translation extracts from HEK293FT cells cultured in stress conditions, the same procedure was used except that the culture medium was supplemented with 5mM DTT 3 hours before harvesting to induce endoplasmic reticulum (ER) stress, which triggers the unfolded protein response (UPR) (Kozutsumi *et al.* 1988; Hetz *et al.* 2020). This pathway leads in particular to the phosphorylation of the eIF2 α-subunit by the protein kinase R-like endoplasmic reticulum kinase (PERK) (Harding *et al.* 2000; Scheuner *et al.* 2001).

2.2.4. Measurements of eIF2 α-subunit phosphorylation by Western Blot

Aliquots of *in vitro* translation reactions were run on 12% SDS-PAGE. Proteins were transferred on PVDF membranes (Immobilon®-P Transfer Membrane) for 1h at 10V. Membranes were saturated with [PBS 1X, Tween20 0.5%, BSA 50mg/mL]. The primary antibodies (@eIF2α: #9722 and @eIF2α-phosphorylated: #3597, Cell Signalling) were diluted ten thousand times in PBS 1X, Tween 20 0.5%, BSA 50 mg/mL and hybridized overnight at 4°C. Membranes were then washed three times in [PBS 1X, Tween20 0.5%] before the 2-hours hybridization of the secondary antibody (@rabbit A120-101-P, Bethyl Laboratories)

which was diluted ten thousand times in PBS 1X, Tween20 0.1%. Membranes were finally washed three times in PBS 1X, Tween 20 0.1% before the addition of the ECL substrate (Clarity Western ECL Substrate #1705061). Chemiluminescence signals were revealed with Biorad chemiDoc (MP) apparatus and resulting TIFF images were analyzed with ImageJ software.

2.3. Screening of the NNNNNNAUGNNN reporter library

2.3.1. Reporter library synthesis

The starting library was synthetized in three steps. First, the insert EMCVscan-NNNNNNATGNNN-eGFP(20nt) used for Gibson assembly was PCR-amplified using primers n° 1 and 2 (Table S1) and the pUC19-EMCVscan plasmid as a template. Then Gibson assembly of the PCR-amplified insert into the pUC19-NcoI-eGFP plasmid linearized by NcoI was performed using the NEBuilder® HiFi DNA Assembly Master Mix (#E2621). Finally, the reporter library was PCR amplified using the Gibson assembly reaction as a template with primers n°3 and 4 (Table S1).

The same strategy was used to construct the reporters with non-randomized AUG contexts for the *in vitro* translation assays performed in 3.2.3. First, the insert scan-NNNNNNATGNNN-Renilla(20nt) used for Gibson assembly was PCR-amplified using forward primers n°5 to 21, reverse primer n°22 (Table S1) and the pUC19-XbaI-Renilla plasmid as a template. Then Gibson assembly of the PCR-amplified insert into the pUC19-EMCVscan-SpeI plasmid linearized by SpeI was performed using the NEBuilder® HiFi DNA Assembly Master Mix (#E2621). Finally, the construct was PCR amplified using the Gibson assembly reaction as a template with primers n°25 and 22.

2.3.2. Microfluidics-based screening of the reporter library

Screening was performed as previously described (Pernod *et al.* 2020). Briefly, each molecule of the DNA reporter library is individualized into 2 pL droplets. Upon thermocycling, these droplets are injected into a fusion device, together with 18 pL droplets containing a mixture for *in vitro* coupled transcription and translation reaction with HEK293FT cell-free translation extracts prepared from cells cultured in physiological conditions. After droplets fusion, the final concentrations are 15% HEK293FT cell-free extract (1.8 µg/µL total protein), 100 mM potassium acetate, 5 mM magnesium acetate, 1 mM of each NTP, 1.5 mM of each amino acid, 20 mM HEPES-KOH pH 7.5, 0.5 mM spermidine, 1 mM DTT, 0.8 mM ATP, 0.1 mM GTP, 8 mM phospho-creatine, 0.1 µg/µL creatine phospho-kinase, 0.5 U/µL RNase inhibitor (Promega) and 0.125 mg/mL of home-made recombinant T7 RNA polymerase. The emulsion is incubated 1h at 30°C and further injected into a fluorescence-activated sorting device. Three rounds of selection were conducted by PCR-amplifying the DNA contained in the sorted droplets, as described in the next paragraph with primers n°3 and 4 (Table S1).

2.3.3. Libraries preparation for sequencing

DNA molecules contained in the sorted droplets were amplified by PCR in a 50 µL reaction using 0.25 mM dNTP, 0.75X GC buffer (NEB), 0.04 µM primers n°23 and 24 (Table S1) and 0.01 µg/µL home-made Phusion DNA polymerase. PCR products were purified by Solid Phase Reverse Immobilization (SPRI) beads (Beckman) and used as a template (0.5 ng/µL) for a 100 µL PCR using 10 µL each Nextera Index primers (Illumina), 0.75X GC buffer (NEB), 0.25mM dNTP, 0.01 µg/µL recombinant Phusion DNA polymerase. Indexed libraries were SPRI-

purified, checked on BioAnalyser and analysed on a MiSeq Instrument (Illumina) with a MiSeq V3 150 cycles reagent kit with a single-end protocol (Illumina).

2.3.4. Data analysis strategy

The whole data analysis strategy has been conducted using a self-made Python algorithm.

FastQ quality scores filtering

Since this work focuses on the region that is covered by the 48S complex when the AUG start-codon is in the P-site of the initiating ribosome (the A of the AUG is the +1 position), we decided to filter out sequences if at least one nucleotide within the -16 to +16 mRNA positions has a Q-score that is inferior to 30. The Q-score is a function of the probability of incorrect base calling. A base with a Q-score of 30 has 1 chance over 1000 to be incorrect. Then, we kept the sequences that have both the AUG start-codon located at the expected position (nucleotides 689, 690, 691 in the reporter sequence) and a perfect match with no mutation in the -16 to +16 window, apart from the randomized sequences flanking the AUG start site.

Background sequences filtering in the libraries

To calculate a threshold proportion below which the sequences are considered background, the algorithm draws the scatterplot $cumulated\ p = f\left(\frac{1}{p}\right) = f(x)$, where p is the observed proportions of each sequence variant. For filtering the data, the following functions $y = {y_{lim}}/{(1 + a.e^{(b.x)^n})}$ for R0 and R1 or $y = y_{lim}.(1 + a.e^{b.x})$ for R2 and R3 are fitted into the data points. The use of two distinct functions was motivated by the different shapes of the distributions of R0-R1 and R2-R3 respectively, which influence the goodness of fit. In both functions, b is called the “characteristic proportion” and is strictly dependent on the shape of the distribution. The algorithm eventually determines a threshold proportion corresponding respectively to 6, 5, 15 and 15 (chosen empirically) times the characteristic proportion calculated for R0, R1, R2 and R3 respectively. Finally, sequences that are found in the selected libraries but not in the starting library are also filtered out.

Analysis of unique sequences in non-overlapping subgroups

Since each selected libraries (R1 to R3) at a round of selection n is contained into the previous one n-1, it is necessary to generate non-overlapping datasets to further highlight sequence patterns that are specific to each selected library. It is done by determining the last library R_n in which a sequence is significantly found (*i.e.* whose count is superior to the background threshold previously determined) and attribute it to the corresponding subgroup X_n, provided the sequence is also present in the previous rounds $R_{i<n}$. This presence-absence approach is preferred over an “enrichment” approach because of the following statement: if the experimental set-up were free of PCR-induced biases (or PCR-free), the enrichment of a given sequence would only depend on the number of unique sequences that have been lost during the screening process**.** This implies that each sequence within a selected library should have the same enrichment, regardless of its proportion. The only translation-related feature is the significant presence of a sequence in a specific library, which has been dealt with in the previous paragraph. This implies that each sequence within a selected library should have the same enrichment. Finally, the sequences that have been filtered out are attributed to the Non-Attributed (NA) subgroup.

Calculation of the importance of a position

To determine which positions i are the most critical among the randomized ones, a convergence parameter is calculated using the chi-square distance between the observed frequency of each nucleotide (nt) and the theoretical frequency that would be observed under the assumption of equiprobability. This convergence parameter is calculated as:

$$convergence(i) = \sum_{nt=A,U,G,C} \frac{(proportion(nt) - 0.25)^2}{0.25^2}$$

It can be interpreted as a convergence indicator in a sense that it measures how much the nucleotide frequencies at a given position diverged from the initial repartition and therefore converged toward new frequencies through the screening.

Motif calculations

A direct consequence of the principle that translation-related enrichment is equivalent for all sequences of the same subgroup is that they cannot be ranked. Then, a way to conduct quantitative approaches within a subgroup is to perform motif calculations from the sequences in the same subgroup. These calculations consist in counting each k-motif that can be generated from the sequences in each subgroup, excluding the +1, +2 and +3 positions that are constant (AUG start-codon). A k-motif refers to any combination of k (k < 9) nucleotides in the randomized region, the other 9 - k nucleotides left are N. The number of k-motifs that can be generated from one sequence is $\binom{9}{k} = \frac{9!}{(9-k)!k!}$, where k is the number of fixed nucleotides {A, U, G, C}. For instance, the NNNACCAUGNNN motif is one out of the 84 possible (k=3)-motifs that can be generated from the sequence GCCACCAUGGCG. Similarly, the number of sequences that feature a k-motif is 4^{9-k}. The total number of distinct k-motifs is $\binom{9}{k} \times 4^k$. Each motif is attributed a score that measures its coverage in the starting library: if all the sequences that feature a k-motif are found in the starting library, their number is 4^{9-k}, then the score of the k-motif equals 1, otherwise $score(k\_motif) = \frac{n_{R0}}{4^{9-k}}$, where n_{R0} is the number of sequences featuring the k-motif in the starting library.

Determination and ranking of non-overlapping motifs

Motifs that do not share the k fixed nucleotides at the same positions are considered overlapping, because one can be contained into the other. For example, GCCNNNAUGNNN and GNNACNAUGNNN are contained into each other. To generate sets of non-overlapping motifs, the following iterative pipeline is used for each value of k = {2…8}. The maximum number of iterations i is $\binom{9}{k} = \frac{9!}{k!(9-k)!}$.

i=1: the k-motifs are ranked according to their abundance, in descending order. The most abundant k-motif, along with all the others sharing the fixed nucleotides at the same positions and weighted according to their abundance. That makes a first list of non-overlapping motifs that have been generated at the first iteration (L1).

i=2: the previously identified motifs are removed and the k-motifs left are ranked according to their abundance, in descending order. The most abundant k-motif, along with all the others sharing the fixed nucleotides at the same positions and weighted according to their abundance. That makes a second list of non-overlapping motifs (L2), that may overlap however with the previous ones (L1).

i=n: the previously identified motifs are removed and the k-motifs left are ranked according to their abundance, in descending order. The most abundant k-motif, along with all the others sharing the fixed nucleotides at the same positions and weighted according to their abundance. That makes a n^{th} list of non-overlapping motifs (Ln).

Because each list contains non-overlapping motifs, the sum of the proportions of each motif contained in each list equals 1.

2.4. Motif research in the human genome

We used a self-made Python algorithm to count the occurrence of a given k-motif in the annotated ORFs starting with an AUG codon in the human genome. The sequences were downloaded from the NCBI site ftp.ncbi.nlm.nih.gov/genomes/ and further processed to remove ORFs which a) are annotated as pseudogenes, b) whose length is not a multiple of 3, c) have a premature stop codon d) don't have a stop codon, e) don't start with an AUG, resulting in 118,629 ORFs. As the frequency of a NNNNNNAUGNNN motif is related to the number of fixed bases (a motif with k=3 fixed nucleotides such as NNNGGGAUGNNN will be more frequent than one with k=5 fixed nucleotides such as NGGGGGAUGNNN), the observed occurrences cannot be used to rank motifs with different numbers of fixed nucleotides. To do so, the following ratio is calculated: observed number of motif M / expected number of motif M, abbreviated from now on Obs/Exp. The expected number of a motif corresponds to its expected occurrence if it is randomly present in human sequences among all the others. For instance, if the motif NNNRNNAUGNNN (R = A or G) is randomly present among all the NNNNNNAUGNNN motifs, it would represent half of them in the absence of bias (*i.e.*, without selection pressure).

2.5. Reporter synthesis for *in vitro* translation

All plasmids containing the 5'UTR and protein reporter sequences of interest were prepared using the NEBuilder® HiFi DNA Assembly Cloning Kit (#E5520S). Site-directed mutagenesis was performed with NEB Q5® Site Directed Mutagenesis Kit (#E0554S). Then, the corresponding T7-5'UTR-reporter DNA construct is PCR-amplified using forward primers n°25 to 27 and reverse primer n°22 (Table S1). Corresponding RNA reporters are synthetized in a 100 µL *in vitro* transcription reaction using 0.1 µM of T7-DNA template, 5 mM Tris-HCl pH 8, 30 mM $MgCl_2$, 1 mM spermidine, 5 mM DTT, 0.01% Triton X-100, 5 mM of each ribonucleotide (ATP, CTP, GTP, UTP pH 7.5), 0.5 U/µL RNase inhibitor (Promega) and 0.125 mg/mL of home-made recombinant T7 RNA polymerase. The reaction is incubated 1h at 37°C. For co-transcriptional capping, the reaction is started in the absence of GTP and in the presence of 0.5 mM of anti-reverse m^7G-cap analog (NEB #S1411) for 10 min. GTP is then gradually added up to 5 mM at 1h incubation. After 1h incubation, 0.02 mg/mL pyrophosphatase (Merck) is added and after 30 min, the DNA template is removed by 1h incubation with 0.2U/µL DNaseI (Roche). The transcription products are loaded on a 1 mL G-25 Superfine Sephadex column (Cytiva). The resulting eluate is phenol-extracted and ethanol-precipitated. Resulting RNA pellets are resuspended in 30 µL milli-Q water, frozen, and quantified by absorbance measurement.

2.6. *In vitro* translation assays

RRL: the final concentrations in the reaction are 50% RRL, 100 mM potassium acetate, 1 mM magnesium acetate, 1.5 mM of all amino acids but methionine, 0.5 U/µL RNase inhibitor (Promega), 0.125 µCi/µL ^{35}S-methionine (Perkin-Elmer) and 0.2 µM reporter RNA in a total

volume of 20 µL. The reaction is incubated 1h at 30°C or 2h at 30°C if TEV protease is used for co-translational cleavage.

HEK293FT extracts: the final concentrations in the reaction are 15% extract (1.8 µg/µL total protein), 100 mM potassium acetate, 1 mM magnesium acetate, 1.5 mM of all amino acids but methionine, 20 mM HEPES-KOH pH 7.5, 0.5 mM spermidine, 1 mM DTT, 0.8 mM ATP, 0.1 mM GTP, 8 mM phospho-creatine, 0.1 µg/µL creatine phospho-kinase, 0.5 U/µL RNase inhibitor (Promega), 0.125 µCi/µL ^{35}S-methionine (Perkin-Elmer) and 0.2 µM reporter RNA in a total volume of 20 µL. The reaction is incubated 1h at 30°C or 2h at 30°C if TEV protease is used for co-translational cleavage.

Co-translational cleavage of *in vitro* synthesized reporter proteins with TEV protease: cleavage occurs by adding home-made recombinant TEV protease into *in vitro* translation reactions at a final concentration of 0.035 ug/µL.

2.7. Assessment of *in vitro* translation products levels

^{35}S-signal quantification: 5 µL of *in vitro* translation reactions are analyzed by 12% SDS-PAGE. ^{35}S- translation products bands are quantified with a phosphorimager (Typhoon FLA 7000). Image analysis is conducted with ImageJ software using the resulting TIFF files.

Luciferase assay: the amount of *in vitro* synthetized Renilla luciferase was assessed upon injection of 100 µL coelenterazine 0.25 µmol/mL (Synchem) into 10 µL of *in vitro* translation reaction using a luminometer (Varioskan LUX, Thermo Scientific™). Subsequent photon emission was measured for 10 seconds.

Real-time measurement of eGFP synthesis: *in vitro* synthesis of eGFP was assessed by measuring the emitted fluorescence (λ_{ex} = 485nm and λ_{em} = 520nm) every minute with a Mx3005P QPCR Instrument (Agilent).

2.8. Predictions and free energy calculations of RNA secondary structures

These calculations were conducted using the Mfold V2.3 for RNA (Zuker 2003) on the various a11 structure mutants used in this work. Constraints were implemented according to our probing data that enabled to draw the 2D-model of the wild type structure (Alghoul *et al.* 2021). No constraints were implemented for predictions of secondary structures in the eGFP coding sequence.

3. Results

3.1. Functional characterization of the cell-free translation extracts prepared from HEK293FT cells cultured in physiological or stressed conditions

We prepared cell-free translation extracts from HEK293FT cells that were grown in optimal conditions and in stress conditions. These extracts were first validated with *in vitro* translation assays using two reporters: a luciferase reporter containing in its 5'UTR the IRES of the intergenic region (IGR) from the Cricket Paralysis Virus (CrPV) genome and a regular cap-dependent reporters shown in Figure 1A. While cap-dependent translation is strongly inhibited in the stressed extracts due to higher rates of eIF2 phosphorylation (see below), IGR-driven translation is enhanced (Figure 1A). This is probably because IGR does not require initiation translation factors. In that sense, since the ribosomes contained in the stressed-extracts fully retained their translational function, the IGR-mediated initiation does not rely on eIF2-dependent mechanisms.

eIF2 α-subunit phosphorylation is a key parameter to assess the stress intensity of the cell free-extracts. We determined by Western Blot the initial rate of eIF2α phosphorylation being 33 +/- 3% in the "stressed" extracts and 15 +/- 2% in the non-stressed ones. (Figure 1B).

These extracts have been used for characterizing *cis*-regulatory elements involved in both AUG and non-AUG translation.

3.2. Identification of optimal and suboptimal AUG-flanking sequences for translation initiation

We used a microfluidic-based high-throughput screening pipeline (Pernod *et al.* 2020) with cell-free translation extracts prepared from HEK293FT cells cultured in physiological conditions to search for optimal and suboptimal AUG-flanking sequences for translation initiation (Figure 2A-B). The sequences and motifs highlighted by the data analysis were further validated and characterized by *in vitro* translation assays. In addition, we also analyzed the motifs that are found in the human genome with bioinformatic analysis.

3.2.1. Microfluidic-based high-throughput screening of NNNNNN<u>AUG</u>NNN reporter library

To mimic cap-dependent translation in a coupled *in vitro* transcription and translation reaction (IVTT), we used a reporter that contains in its 5'UTR a modified version of the EMCV IRES followed by a 128 nucleotides-long flexible linker (Figure 3A). To minimize the potential inhibiting effects of secondary structures in the 5'UTR, the linker is mostly made of CAA repeats that do not fold into secondary structures. Its length was chosen in accordance with the average 5'UTR length of human transcripts (Leppek *et al.* 2018). The coding sequence of the enhanced Green Fluorescent Protein (eGFP) has been inserted downstream the 5'UTR and its AUG nucleotide context was randomized as followed: NNNNNN<u>AUG</u>NNN, where N is any of the four nucleotides (Figure 3A). After IVTT, synthesized eGFP will allow fluorescence activated sorting during the screening process. We used a modified EMCV IRES to recruit a scanning-competent 43S particle at the 5' end of the RNA reporter library (Figure 3B). To fulfill this objective, we depleted the native AUG of the IRES by site-directed mutagenesis to prevent residual translation initiation on the IRES start codon and to force the recruited 43S particle to scan the flexible linker until it reaches the AUG start codon of the eGFP. We found that the eGFP coding sequence has a strong inhibitory effect on translation initiation, possibly because

it folds into secondary structures that interfere with AUG recognition. Therefore, we inserted an unfolded linker GAA(CAA)$_7$ between the AUG start codon and eGFP coding sequence that efficiently rescued translation (Figure S1).

In vitro translation assays conducted with our model reporters confirmed that the modified version of the EMCV IRES can promote efficient scanning until the ribosome reaches the AUG of interest, which makes this IRES suitable for recruiting the translation machinery and mimic cap-dependent initiation (Figure 3C).

3.2.1.1. The sequence variants' representation in the starting library is not uniform

Counts of unique sequences in the starting library R_0 follow a right-skewed distribution, meaning there are over-represented sequences. 75% of sequence variants are present less than 21 times in the starting library (Figure 4A, left). The analysis of the nucleotide composition of the sequences reveals a strong bias towards A-rich sequences, most probably due to the successive PCR amplification steps during library construction because no selection has yet taken place (Figure 2B). The count of sequences is an exponential function of the number of adenosines in the randomized region, regardless of their position (Figure 4A, right). Such a behavior is not observed with the three other nucleotides thereby confirming a specific bias for adenosines (Figure S2). The log-distribution (Figure 4B, right) of sequence counts in the starting library R0 and in the R1 library features a loss of monotonicity in its left tail, suggesting rare observations. The bimodal aspect of the count log-distribution of each unique sequences in the R2 and R3 libraries allows to easily identify rare observations that are considered as background. Graphical representations of the calculated threshold proportions (Figure 4B, left) efficiently discriminate 1) the non-monotonic from the monotonic part of the distribution for R0 and R1, 2) the two modes of R2 and R3, which validates the method for threshold determination explained in the material and method section. Finally, sequences that are found in the selected libraries but not in the starting library are also filtered out.

3.2.1.2. Ninety percent of the sequences are sorted out after two rounds of selection

Sequencing of the starting library reveals that we screened a total of 221,197 different AUG contexts, which is slightly lower than the expected total number of possibilities (4^9 = 262,144). The starting library then contains 84% of the possible combinations. Without applying any of the filters described in the material and methods section or classifying the sequences into non-overlapping subgroups, we recovered a total of 147,633 distinct sequences in R1, 8200 in R2 and 6730 in R3. One third of the sequences were eliminated by the first round of selection and more than 90% were lost after the second round of selection, further suggesting that the start codon nucleotide context is an essential *cis* regulator of translation initiation. Since each selected libraries (R1 to R3) at a round of selection n is contained into the previous one n-1, it is necessary to generate non-overlapping datasets to further highlight sequence patterns that are specific to each selected library. Statistics and numbers for each non-overlapping subgroup after filters application (see Material and Methods section) are depicted in Figure 2C.

3.2.1.3. The combinations with in-frame stop codons and upstream out-of-frame AUGs are lost

To validate our approach, we first checked the most enriched 3-nucleotides combinations in the 'not selected' subgroup X_0. Among the lost sequences, we found the three combinations with in-frame downstream stop codons along with the two combinations containing upstream

and out-of-frame AUG codons (Figure 5A). Since these five combinations are known to inhibit translation initiation, these observations demonstrate that the experimental set-up and analysis strategies allows to functionally characterize translation relevant motifs, even in the presence of strongly PCR-biased sequences.

3.2.1.4. UAGNNNAUGNNN motifs are inefficient for translation

The search for the most enriched 3-nucleotides motifs in the "unselected" X_0 subgroup also leads to the identification of the UAGNNNAUGNNN motif (Figure 5B). Surprisingly, this motif contains an in-frame upstream stop codon UAG. To make sure this motif does not correspond to a stop codon of an upstream ORF whose translation would inhibit the initiation on the main AUG, we calculated the enrichment of the equivalent motif containing the two others stops codons, namely the motifs UAANNNAUGNNN and UGANNNAUGNNN motifs in all the subgroups (Figure 5B). Only the UAGNNNAUGNNN motif was lost indicating that this observation is not due to any upstream translation event, but to an inhibiting effect on translation initiation by UAG at positions -6, -5 and -4 of the AUG context. However, the UAGNNNAUGNNN motif is the least represented 3-uplet in the starting library, which questions its significance (Figure 5B). At this stage, we could not exclude that the loss of this motif is also due to its low abundance in the starting library. The suspected negative effect of UAG at positions -6, -5 and -4 was then further investigated with *in vitro* translation assays (see below).

3.2.1.5. Adenosines are beneficial for translation at all positions

Calculations of the frequencies of each base at each position between the selected $X_{n>0}$ and unselected X_0 sequences reveal that adenosines are strongly over-represented at all the randomized positions in the selected sequences (Figure 6A). This indicates that adenosines have a beneficial effect on translation initiation especially for positions -6 to +4. Calculation of the convergence of each position for each subgroup measures how fast these positions converge to specific nucleotide(s), and therefore diverge from their initial (approx. equiprobable) state. First, the well-characterized -3 and +4 positions, along with the less-characterized -6 position, converge faster toward specific nucleotides (mainly A). Another remarkable feature is that positions +5 and +6, the last two nucleotides of the second codon, are those subjected to the lowest convergence, suggesting that the second codon downstream the AUG start codon is less important for translation efficiency (Figure 6B). The remaining positions -1, -2, -4 and -5 are also influential although to a clearly lesser extent than -6, -3 and +4. Furthermore, except for A-rich sequences, those displaying strong nucleotide homogeneity, with more than four times the same nucleotide in the randomized region, are gradually depleted through the screening (Figure 6C). Even if it may be partly due to the PCR-bias already mentioned, the fact that these sequences emerge in the last rounds of sequences strongly suggest that this is also linked to their high translatability. A strong argument in favor of this statement is that the A-rich contexts with an in-frame stop codons (AAAAAAAUGUAA for instance) or out-of-frame upstream AUGs are not selected (Figure S3).

3.2.1.6. Influence of the second encoded amino acid

We checked the influence of the second encoded amino acid by calculating their frequencies in all subgroups (Figure 6D). First, the three stop codons are largely counter selected. Because each amino acid can be encoded by several codons, those frequencies were divided by the number of codons coding for the corresponding amino acid. A few amino acids stand out in

the efficient X_3 subgroup, mainly Lysine (K) and Glutamic Acid (E) which are enriched after each round of selection. Lysine and Glutamic acid are encoded by codons that have at least two adenosines. On the other hand, Proline (P), Phenylalanine (F), and to a lesser extent, Cysteine (C) and Valine (V), are gradually lost after each selection round most probably because these codons do not contain A residues. Therefore, we cannot conclude that these differences are due to the nature of the amino acid itself or to the nucleotide composition of their codons.

3.2.1.7. Kozak and non-Kozak sequences

The fact that -3 and +4 positions are the first ones that discriminate the first selected sequences X_1 from the unselected ones X_0 leads to question the Kozak motif. The Kozak consensus sequence has been defined since decades (Kozak 1986, 1987). In parallel, we calculated the average sequence context from 118,629 annotated human ORFs that are translated from an AUG start codon. The result is consistent with the previously described Kozak consensus (Figure 7A), which corresponds to the average motif GCCACCAUGGCG. It is cytidine-rich except for the -3 and +4 positions which mostly have a purine (A or G) These two positions have been shown to be critical for translation initiation efficiency (Kozak 1986, 1987). In our selections, motifs calculations reveal that the proportion of the four so-called minimal Kozak motifs, with a purine at -3 and +4 positions, gradually increases after each round of selection, eventually representing 42% of the sequences in the efficient group X_3, thereby confirming that Kozak consensus is efficiently translated in our selections (Figure 7B, left). In contrast, cytidines that are found at positions -5, -4, -2 and -1 in the Kozak consensus, are mostly depleted in the efficient group X_3 (Figure 7B, right), since only 21% of the selected Kozak sequences has at least one cytidine at positions -5 or -4 or -2 or -1. The few motifs with one C found at positions -5 or -4 or -2 or -1 are present in X_3 but do not show a clear preference for a specific pattern at the other positions, with only 42% of them having a minimal Kozak motif (Figure 7C). Moreover, many Kozak sequences are lost after the first round of selection and represent one quarter of the unselected sequences X_0 (Figure 7B), suggesting that -3 and +4 positions are not sufficient for optimal translational efficiency. The -3A and +4A are most of the time associated and represent 17% of the last subgroup X_3 (Figure 8A). Non-Kozak sequences with pyrimidines at -3 and +4 are gradually depleted through the rounds of selection (Figures 7B and 8B). They represent 12% of the final selected sequences X_3, without featuring a strong requirement for a precise -3/+4 combination, although -3U / +4C is slightly more represented in the efficient subgroup X_3 (Figure 8B). Altogether, these data suggest that translation initiation efficiency does not rely so much on the -3/+4 nucleotide combination.

3.2.1.8. Efficient AUG-contexts do not feature any minimal sequence patterns

Data analysis showed that all positions are important to discriminate sequences in each round of selection, with adenosines being more preferred on average than the other three nucleotides. That makes calculation of representative sequence(s) that would synthetize the selected subgroups X_n challenging, as none of the randomized positions can be left behind in the calculation. Average sequence determination is irrelevant anyhow unless assuming all positions are strictly independent from each other indicating the absence of linkage information between nucleotides in the same sequence. Indeed, n abundant nucleotides at n distinct positions are not necessarily found in the same sequences and could belong to at most n separate subclasses of sequences within the data.

In the following paragraphs, a k-motif refers to any combination of k (k < 9) nucleotides in the randomized region, the other 9 - k nucleotides left are N (any of the four ribonucleotides). It is trivial that the higher k, the less probable the k-motifs are and thus the lower their proportion, but at the same time the higher k, the more impactful the motif becomes. Indeed, a motif such as ACNUGNAUGCNN would have more impact than, for example, ACNNNNAUGNNN as its sequences have more nucleotides in common and are consequently more homogeneous. Similarly, the lower k, the broader the motif as it accounts for more sequences, but the less impactful it becomes because it has few determinant nucleotides (k). Therefore, a good consensus sequence is a trade-off between a high value of k and the number of individual sequences it accounts for. The following points further support the absence of clear preferential AUG-contexts in the selected sequences X_3 from positions -6 to +6, and that each individual sequence can be considered as a *cis*-regulatory pattern on its own.

First, there are no non-overlapping k-motifs (Figure 9A) that can efficiently summarize the sequences of the selected subgroups $X_{n>0}$. Figure 9B displays the distribution of the proportions of each k-motif in the subgroups. Using the pipeline described in the material and methods section, we showed that the most influent 2- and 3- non-overlapping motifs contain specific nucleotides at positions (-3 +4) and (-6, -3, +4) with the most abundant ones only representing 17% and 7% of the selected sequences X_3 (Figure 9C). Figure 9D is an extension of Figure 9C to all iterations and shows the X_3 subgroup coverage as a function of the cumulated proportions of each possible set of non-overlapping motifs that constitute the X_3 subgroup. The fact that 50% of 2- and 3- motifs of distinct non-overlapping sets summarize at most 75% of the efficient subgroup X_3 for both 2- and 3- combinations is in favor of the absence of minimal sequence patterns within the selected sequences.

Second, there are very few k-motifs that are exclusive to one subgroup, and when they are, they account for very few sequences because k is equal to 7 (max. 16 distinct sequences) or 8 (max. 4 sequences). To that respect, Figure 10 shows the repartition of each sequence that feature each k-motifs in each subgroup. It reveals there are no k-motifs with k < 7 that are exclusively found in only one subgroup. Among the 7- and 8- motifs, almost 90% and 50% of them respectively are found in distinct subgroups, meaning that even a difference of 1 to 2 nucleotides between sequences can modulate their translation efficiency.

In that sense, the variants of these 7- and 8- motifs found in more than one subgroup do not feature any precise position regarding their variable nucleotide(s). Figure 11A-B show 7- and 8- motifs with respectively at least one quarter or one half of their variants found in the last subgroup X_3 and the rest found in at least one of the other subgroups $X_{n<3}$ except the NA subgroup. Because they are found in distinct subgroups, these are motifs whose variants have distinct translational efficiency, depending on the remaining variable nucleotides N in the motifs (two variable nucleotides in 7-motifs, only one in 8-motifs). The position of these variable nucleotides is strictly dependent on each motif as shown on the middle colormaps of Figure 11. In other words, there are no positions that are specifically targeted by the N (variable) nucleotides, which further supports that every nucleotide combination between the nine (and not less) randomized positions can control the efficiency of translation initiation.

In conclusion, only 10% of the sequences from the starting library are efficiently translated (belonging to the X_3 subgroup). Among these, at least 8 positions are interconnected in a precise pattern that cannot be modified without loss in translation efficiency, which implies the absence of minimal sequence patterns that can explain translation efficiency on their own.

3.2.2. Motifs research in the human genome

To strengthen the conclusions from our selections, we analyzed the previously identified motifs in the annotated ORFs of the human genome to verify that they are indeed used in living cells. For this purpose, we determined the proportion of each motif the human ORFeome and calculated the ratio observed frequency over expected frequency (Obs/Exp ratio). It is worth mentioning here that a high Obs/Exp ratio means that the motif is used by the ribosome but does not necessarily correlate with a high translation efficiency. It is indeed possible to consider that motifs that are poor substrates for the ribosome could have been selected during the evolution as *cis* regulators of translation initiation.

3.2.2.1. A-rich nucleotide contexts

A striking feature of the selected sequences in the screening experiment is that adenosines are on average enriched at all positions. To further investigate this feature, we measured the representativity of A-rich AUG nucleotide contexts in the human ORFs by calculating their Obs/Exp ratios. Figure 12 AC show that the Obs/Exp increases as the number of adenosines in the context increases. Strikingly, adenosines are particularly absent at positions -4 and +6, but well featured at position -2 (Figure 12B). The most represented A-motifs in the genomes are AANAAAAUGAAA and AANAAAAUGAAN with an Obs/Exp ratio of 24 and 20 respectively, in the same range as the previously described Kozak consensus NCCRCCAUGRNN (ratio of 18) and GCCRCCAUGRNN (ratio of 32) contexts.

3.2.2.2. Kozak and C-rich motifs

The screening results suggested that the cytidines flanking the -3 position that are present in the Kozak consensus are not essential for optimal translation efficiency. To evaluate the importance of these cytidines in the human ORFeome, the Obs/Exp ratios were calculated for NNNNNNAUGNNN motifs having either two, three or four C or D (A, U or G) at positions -5, -4, -2, -1, in the presence or absence of a purine (R) at positions -3, +4 (Figure 13). Compared to the minimal Kozak motif NNNRNNAUGRNN, the Obs/Exp ratios of C-rich motifs are higher (Figure 13A-B) and keep increasing as the number of cytidines at positions -5, -4, -2, and -1 increases, suggesting that cytidines at these positions were positively selected in the human genome through the evolution (Figure 13B, left). The presence of a purine (R) at positions -3 and +4 further increases this ratio. However, purines at -3 and +4 alone are not especially enriched. In contrast, the same calculations with D do not lead to the same observation confirming a position selection of C during evolution (Figure 13A, right). Altogether, these observations suggest that cytidines flanking the -3 position were selected through the evolution in the human genome and that strong positive selection of purines at -3 and +4 is only observed in such C-rich nucleotide contexts.

3.2.2.3. The most abundant AUG nucleotide context among the annotated human ORFs is not the Kozak context

By counting each individual nucleotide context among the 118,629 in human ORFeome, we found that there are 28,454 distinct contexts, meaning that only 11% of the 4^9 possible NNNNNNAUGNNN contexts are present in the human ORFs and therefore used for translation initiation (Figure 14A). This observation further supports the fact that the start codon nucleotide context is an essential *cis* regulator of translation initiation. The representativity of these contexts is not uniform and follows a right-skewed distribution (Figure 14B): 75% are found less than 5 times, suggesting a broad diversity of nucleotide contexts that are used by the

human ribosome. The most abundant one, AGCACCAUGUCG, is found 411 times and does not even feature a Kozak context, indeed this abundant motif contains the unfavorable +4U according to the Kozak context (Figure 14C, top line). The second most abundant one, GGGAAGAUGGCG, which is closer to the Kozak consensus is found 149 times, almost 3 times less. The Kozak consensus GCCACCAUGGCG is found 76 times, almost 6 times less than the most abundant one though. However, the broad consensus GCCRCCAUGRNN is found 920 times (not shown), thus representing only 0.8% of the 118,629 analyzed human ORFs, which is too low to be considered as a genuine consensus.

Considering the respective abundance of each motif, only 2-motifs (and 3-motifs to way lesser extents) which represent at best 20% (10%) of the AUG contexts of the human ORFs might be considered as a consensus (Figure 14D). Indeed, non-overlapping motifs calculations using our iterative method show that the most representative motif is the 2-motif NNNNCCAUGNNN that accounts for 20% of the human AUG contexts. It is closely followed by other 2-motifs at distinct positions as shown in Figure 14F.

Altogether, these analyses suggest that there is a great diversity of AUG nucleotide contexts in the human ORFeome that cannot be summarized with one minimal consensus pattern (Figure 14B). At best, 50% of the possible 2-motifs contain 80% of all the human AUG contexts (Figure 14E), which further supports their diversity. Consequently, 2-motifs like NNNNCCAUGNNN (20%), NNNNNCAUGGNN (20%), NNNANNAUGGNN (18%), NNNNCCAUGGNN (10%) can be considered as consensus because they are highly abundant (Figure 14D-E-F), however the fact they only feature two or three fixed nucleotides makes them highly susceptible to different *cis*-regulatory properties because of their link with the other nucleotides left.

3.2.3. Validation of motifs with *in vitro* translation assays

The conclusions drew from the results of the AUG-context screening and the genomic analysis were further investigated with *in vitro* translation assays using untreated rabbit reticulocytes lysates. We used *in vitro* transcribed m^7G-capped reporter mRNAs. To validate the results with m^7G-capped reporters. The eGFP coding sequence was replaced by the Renilla luciferase coding sequence for quantification purposes (Figure 15A).

The reporters having the UAGNNNAUGNNN motif are twice less efficiently translated than a minimal Kozak motif NNNRNNAUGRNN and the UAGUNNAUGNNN is almost inactive (Figure 15B, right). It is however rescued when an extended Kozak context (ACC) is put at -3, -2 -1 positions or when A residues are located at all the other positions, which is in good agreement with the observation made in our selections on the UAGNNNAUGNNN motif and the A-rich motifs.

According to our selection, A-rich sequences are efficiently translated. The reporter having A residues at the nine positions is at least as efficiently translated as sequences having a minimal Kozak motif but is almost half as efficiently translated as the 5'UTR β-globin reporter, a well-characterized and highly efficient template for the ribosome (Figure 15B, right).

Purines at -3 and +4 positions improve translation efficiency of any motif independently or their positions -5, -4, -2 and -1. Concerning motifs with cytidines at positions -5, -4, -2, -1, they are more efficiently translated than the average NNNNNNAUGNNN motif, regardless of the -3 nucleotide. The presence of a purine at the -3 position in a motif when C are present at -5, -4, -2, -1 results in a boost in their translation efficiency, which does not happen with any other

nucleotide combination lacking cytidines at these positions (D at -5 -4 -2 -1) (Figure 15B, left). This result further supports the importance of the -3 position when it is flanked by cytidines as observed in our human ORFeome analysis.

The 40% difference in translational efficiency between the GCCACCAUGGCG and NCCRCCAUGRNN reporters further supports functional importance and the interplay of all positions between -6 and +6, thereby validating the absence of consensus k-motifs (with k < 9) for translation initiation**.**

3.3. Translation initiation on AUG-like codons requires a stable downstream secondary structure

Recently, we described a novel mechanism of translation initiation called STructure Assisted RNA Translation or START mechanism (Eriani and Martin 2018; Despons and Martin 2020). Such an initiation event relies on the presence of a downstream secondary structure that enhances initiation and can improve initiation on suboptimal codons like AUG-like codons. This implies that the structure is stable-enough to efficiently stall a scanning 43S complex on the suboptimal codons, while being located at an appropriate distance downstream the start codon (Despons and Martin 2020). Indeed, a structure located too close to the start codon would rather decrease translation initiation efficiency due to steric hindrance. Similarly, a structure located too far downstream the start codon would not promote its recognition and lead to leaky scanning. Genome-wide analysis performed on the beginning of annotated ORF sequences in various genomes suggested that START mechanisms may be used to regulate translation initiation in the three kingdoms of life (Despons and Martin 2020). The goal of this work is to investigate the potential role of the START mechanism in AUG-like translation.

We performed *in vitro* translation assays using m^7G-capped mRNA reporters encoding the Renilla luciferase. These reporter mRNAs feature a linear 5'UTR followed by a start codon of interest (Figure 16). We inserted a 50-nucleotides GC-rich a11 secondary structure downstream the start codon that we have previously characterized in the 5'UTR of Hox a11 mRNA (Alghoul *et al.* 2021). Since these reporters will produce distinct N-terminal fusion domains that might modulate the luciferase activity, we introduced a TEV protease cleavage site immediately upstream the Renilla coding sequence to generate the same reporter protein with all mRNA reporters. The a11 structure has been shown to block a scanning complex due its high stability (ΔG = -21,6 kcal/mol) since it contains 16 G-C base pairs (Alghoul *et al.* 2021). With these reporters, we investigated the influence of both the location and the structure stability on translation initiation efficiency, as well as the nature of the start codon.

3.3.1. The optimal position of the secondary structure for translation on a CUG codon ranges from +23 to +26

To determine the optimal distance between the a11 secondary structure and the start codon, *in vitro* translation assays using untreated rabbit reticulocytes lysates were conducted with reporter mRNAs containing an AUG or a CUG start codon with the a11 secondary structure located at positions +11 to +35. The results show that the optimal structure positions for CUG-translation ranges from +23 to +26 (Figure 17A). Translation efficiency decreases as the structure is located closer than +17 or further away than +32 from the CUG codon. Interestingly, when the a11 structure is located at +26, translation initiation from the CUG codon is even more efficient than on an AUG codon. When an AUG codon is used with the same reporters in place of the CUG, their translation is insensitive to the distance between a11

structure and the initiation codon, suggesting that only initiation on a CUG requires the presence of a structure at an optimal distance to the start codon for efficient translation (Figure 17B).

3.3.2. The optimal stability of the secondary structure is influenced by the cell type

Next, we examined the required stability of the secondary structure for efficient translation initiation on a CUG codon. We used reporters with AUG or CUG upstream the a11 secondary structure that is localized at the rather optimal position +20 that was determined in the previous section. To modulate the stability of the a11 structure, we introduced silent mutations that progressively decrease the number of G-C base pairs, resulting in stabilities ranging from -21.6 to -4 kcal/mol (Figure 18). Silent mutations were used to avoid any putative side effects due to amino acid substitutions, except for the -4 kcal/mol mutant that contains mutations leading to 3 amino acids changes in the N-terminal fusion of the translated reporter protein. However, since TEV cleavages were performed co-translationally, all the translated mRNAs produced the same luciferase.

We first confirmed with rabbit reticulocyte lysates that AUG-reporters translation is not influenced by downstream structures (Figure 18A). On the contrary, CUG translation requires a structure with a stability that is higher than -21.6 kcal/mol when it is situated at +20. When the stability is gradually diminished, translation efficiency decreases stepwise, and it is almost 70% reduced when the structure has a stability that is lower than -15 kcal/mol. This shows the importance of a downstream secondary structure with a stability of at least -15 kcal/mol for the initiation of translation on a CUG codon.

Using HEK293FT or SH-SY5Y cell-free translation extracts, we observe the same trend as with RRL (Figure 18B-C). However, the stability requirement for efficient translation initiation on a CUG codon is different with these extracts. Initiation on a CUG codon indeed requires an a11 secondary structure with a stability of -21.6 kcal/mol, which suggests that these human extracts do not contain the same set of RNA helicases that could be responsible for distinct optimal stabilities of downstream structures for CUG initiation. When the a11 structure stability is lower, translation from the CUG is exponentially reduced compared to the stepwise reduction observed with RRL.

3.3.3. The START mechanism does not rescue translation initiation for all AUG-like codons

Next, we extended these experiments to other AUG-like codons, CUG, GUG, UUG, ACG and AUC with the -15 kcal/mol a11 structure situated at position +20 in RRL (Figure 19A). Overall, the AUG-like codons are less efficiently recognized by the ribosome. In addition, all AUG-like codons are not equivalent, and their hierarchy is as follows regarding translation efficiency: CUG>ACG>UUG>AUC>GUG. We repeated the same experiment with HEK293FT cell extracts using the -21.6 kcal/mol a11 structure. The previously established hierarchy is identical with HEK293FT extracts using this structure. Finally, with stressed HEK293FT cells extracts, minor differences in initiation patterns are observed, even though the UUG codon is significantly less efficiently used (Figure 19B).

3.3.4. Downstream secondary structures are found in human smORF that are translated from AUG-like codons

This part is still ongoing. To evaluate the involvement of the START mechanism in non-AUG translation in the cell, we plan to run the secondary structure prediction algorithm used in our previous study (Despons and Martin 2020) on a dataset containing the human small ORFs (uORFs, dORFs and ncORFs) identified by Chothani et al. Secondary structure predictions within the +16/+65 window downstream the initiation codons of these ORFs will provide information on the involvement of putative secondary structures in non-AUG translation in human cells.

4. Discussion

4.1. Influence of the AUG context on translation initiation

The screening for the optimal AUG nucleotide contexts highlighted the importance of adenosines at any position between -6 and +6 for efficient translation initiation. The more A residues in the sequence, the more efficient translation initiation is likely to be. Intriguingly, such sequences are not especially enriched in the human ORFeome, but still efficiently translated *in vitro*. Previous *in silico* studies classified the eucaryotic AUG contexts into two independent categories: GCCGCCAUGNNN and AAAAAAAUGNNN (Nakagawa *et al.* 2007). Our screening experiment has functionally corroborated for the first time the importance of stretches of A upstream of the AUG start codon. The fact that we have mostly selected this second A-rich category preferentially might be due at least partly to the PCR bias that we observed during library construction. Nevertheless, our selection and functional validation by cell-free translation assays confirmed that the AAAAAAAUGNNN context co-exist next to the classical Kozak consensus as an efficient nucleotide combination for translation initiation. In the human ORFeome, an increased number of adenosines in the AUG context correlates with an increased observed/expected ratio, further supporting that such contexts are not rejected by initiating ribosomes.

Apart from these A-rich contexts, no precise sequence pattern emerged from the selected sequences by our screening experiment and from the AUG contexts of the human ORFeome. Both analyses showed a high diversity of AUG nucleotide contexts, suggesting that at least every position from -6 to +6 are determinant for translation initiation efficiency and are totally interdependent. Considering that all these positions within and beyond the -6 to +6 window have been shown to interact with either initiation factors and/or ribosomal RNA and proteins (Simonetti *et al.* 2020), such interdependence is explained by the multiple molecular contacts with these nucleotides of the mRNA. Each proximal AUG sequence creates its own chemical environment and local spatial constraints that modulate the various interactions with the 48S complex and therefore influence translation initiation efficiency. In summary, translation initiation is a function of several interdependent variables, including the energetic contributions brought by each nucleotide of the start codon context.

Concerning Kozak contexts, our experiment showed that -3 and +4 positions, although more influential than the others, do not stand alone to explain the translational efficiency of these sequences. Moreover, it did not highlight any requirements in -3 flanking cytidines in either non-Kozak sequences or Kozak sequences. However, ORFeome analysis showed that these cytidines were positively selected through the evolution and do not seem to have such an impact on translation initiation efficiency *in vitro*, except when an A is present at -3. On the contrary, the -3 position is a highly influential position for translation efficiency, as already suggested in the literature (Kozak 1987; Hernández *et al.* 2019; Simonetti *et al.* 2020; Thakur *et al.* 2020) and further supported by our experiments. From these observations, we propose that the -3-flanking cytidines are more likely to be involve in *trans*- than *cis*-regulation and may rather interact with various factors. In fact, the nucleotide in -3 position interacts directly with the eIF2 alpha subunit (Thakur *et al.* 2020). For technical reasons, we used the EMCV IRES in place of the m^7G-cap in the reporter library used for the screening. Although this IRES can promote ribosomal scanning and is expected to recruit a 43S complex with most of the initiation factors except eIF4E, the recruited 43S complex might be slightly different to the one recruited on a regular m^7G-cap. Such subtle differences would likely result in a change regarding the *trans*-regulation mechanisms involved in AUG recognition.

Similar studies have already been conducted *in vivo* using FACS to screen cells transfected with plasmid libraries of AUG-contexts (Ambrosini *et al.* 2022; Diaz de Arce *et al.* 2018; Noderer *et al.* 2014). They always find the average sequence of the selected variants, whether they initiate from an AUG or not, being close to the Kozak motif, and consequently very distant to the one determined by our screening which is A-rich. Although we demonstrated that averaging sequences is not informative, several points might explain these differences. The first one is the 5'UTR used in the RNA reporter upstream of the randomized region. Differences in the length and the sequence of the UTR being scanned to reach the AUG start codon might change the composition of the 48S particle that is assembled on this AUG. To that respect, the fact we used a 5'UTR strictly linear, CAA-rich without any secondary structure could lead to the assembly of a different 48S complex than with a structured 5'UTR. A second point, closely related to the first, is the size of the screened libraries and therefore the analyzed positions. In the mentioned studies, they range from 4^5 to 4^8 distinct variants at positions -6 to +5 (Noderer *et al.* 2014), -4 to +7 (Ambrosini *et al.* 2022) or -4 to +1 (Diaz de Arce *et al.* 2018). In our selection, we used a larger library (4^9 variants) with randomized nucleotides from -6 to +6. The other fixed nucleotides outside this region can influence the selected AUG-contexts. The last point is the coding sequence itself. As already mentioned, we have found that the beginning of the eGFP coding sequence strongly inhibits translation initiation *in vitro* using various cell-free extracts. We believe that it blocks start-codon accessibility by folding into a stable secondary structure (Figure S4A). We alleviated this bias by introducing a linear 24-nucleotides spacer between the second codon and the eGFP coding sequence to improve start-codon accessibility without impairing eGFP fluorescence (Figure S1). Therefore, we studied the influence of the AUG context in a predicted structure-free environment (Figure S4B). If such inhibition also occurs *in vivo,* it will raise another way to interpret those FACS-seq data also obtained with the eGFP coding sequence. Indeed, in these experiments, the randomized AUG-context could be efficiently used by the 43S either because it disrupts the potential secondary structures induced by the eGFP coding sequence that impair AUG recognition, thereby rescuing translation initiation, or because it is actually an optimal initiation context (or both). This issue can be extended to any study dealing with optimal start-codon contexts and their associated protein coding sequences by integrating local secondary structures that modulate start codon accessibility. To our knowledge, the extent of that matter is not known for the annotated ORFs in the human genome, from which Kozak consensus has been determined.

4.2. Translation initiation on non-AUG codons

Our data demonstrates that the START mechanism requires a stable-enough RNA structure to promote translation initiation on AUG-like codons and further supports the initial work on that topic (Kozak 1990). Considering the length of ribosome footprints provided by ribosome profiling data (approx. 30 nucleotides around the AUG start codon), we expected the optimal distance between the structure and the start-codon to be close or slightly superior to half the number of nucleotides protected by initiating ribosomes (approx. superior to 15-16 nucleotides), because a structure too close to the start-codon would induce the destabilization of a scanning complex before reaching the AUG start codon, while a structure too far away from the CUG codon would lead to leaky-scanning. We found that this distance ranges between 20 and 23 nucleotides, which is slightly longer than the 14nt distance found by Kozak (Kozak 1990). It cannot be excluded however that the optimal distance may also be dependent on the tri-dimensional shape of the structure itself. A small structure with few base pairs might require a closer distance to the start-codon than one with many base-pairs that might contain

kinks that could lead to a structure pointing toward the ribosome. This potential relationship remains to be investigated. Finally, the sequence of the structure loop itself may also have its own importance as it may interact with either the ribosome and/or factors. Indeed, such interactions between helix h16 of the rRNA and nucleotides at positions +17 to +19 have been reported in the case of histone H4 mRNA in our lab (Martin *et al.* 2016).

The structure must be stable-enough to block a scanning ribosome and therefore we chose the a11 GC-rich structure (-21 kcal/mol) for this study as we previously demonstrated that it does stall a scanning pre-initiation complex (Alghoul *et al.* 2021). We determined the threshold stability being close to -15 kcal/mol for initiation on CUG and ACG codons. Other non-AUG codons may require an even more stable secondary structure. Other parameters such as enhanced nucleotide context, specific trans-acting factors or base-pairing with the initiator $tRNA^{Met}$ remain to be investigated.

We have not investigated the influence of the start-codon context, but it can reasonably be assumed that it is the same as with an AUG codon. Previous reports indeed suggest that initiation on non-AUG codons is influenced by their nucleotide context in the same way as with an AUG codon, with the average preferred nucleotide context being the Kozak context (Ivanov *et al.* 2011; Diaz de Arce *et al.* 2018; Hernández *et al.* 2019).

Overall, being able to initiate translation on an AUG-like codon regardless the type of cell extracts suggest that a stable-enough downstream secondary structure is the main factor influencing the initiation on a non-AUG codon. This suggests that the *trans*-acting factors contributions seem to be less critical when a stable structure is present at the appropriate distance from the start codon. Because the optimal structure stability for efficient initiation on AUG-like codons is dependent on the type of extract used for the *in vitro* translation assays, scanning-associated factors might play a major role in those regulation mechanisms. For instance, factors such as eIF1 and eIF1A which control the codon-anticodon interactions (Zhou *et al.* 2020; Pestova *et al.* 1998) and RNA helicases (Guenther *et al.* 2018; Sen *et al.* 2019). The helicase content of the cell would indeed determine which secondary structures can be unwound, and therefore in which stability range the 43S ribosome can scan through to prevent aberrant initiation by a START mechanism.

The established AUG-like codon hierarchy is rather consistent with ribosome profiling data that quantified the use of various start-codon by initiating ribosomes (Ingolia *et al.* 2011; Chen *et al.* 2020; Chothani *et al.* 2022). One striking exception is the GUG codon that is efficiently used in those experiments but not in ours. Thermodynamic studies of the codon-anticodon interaction with a GUG codon demonstrated an increased so-called ‘free-energy penalty’ due to eIF1 and eIF1A interactions that further impairs the initiation on GUG codons but not on CUG codons (Lind and Åqvist 2016). Therefore, initiation on GUG codons is more stringent than on a CUG codon and requires more free energy compensation. Such compensation might come from an even more stable downstream secondary structure and/or from various *trans*-acting factors that modulate start codon selection stringency. One example is the balance between eIF5-mimic protein 5MP and eIF5 that have been characterized to respectively increase or decrease that stringency (Singh *et al.* 2011; Tang *et al.* 2017).

Secondary structures predictions may provide a tool to predict ORF translation from non-predicted non-AUG codons and unveil potential new ORFs and uORFs in genomes.

5. Conclusion

This work provides further insights into *cis*-regulatory parameters that influence start codon recognition. First, its proximal nucleotide context plays a critical role, although it does not feature a minimal sequence pattern that can explain translation initiation on its own. Moreover, the analysis of the AUG nucleotide context of annotated human ORFs revealed a strong selection pressure towards cytidines flanking the -3 position, especially those located at -2 and -1 positions. Additionally, the most represented AUG nucleotide context in the human ORFs is not a Kozak sequence, and overall, the nucleotide contexts of the AUG contexts in the human ORFs are very diverse, further supporting the absence of a broad consensus sequence. Finally, downstream secondary structures can efficiently promote non-AUG translation initiation provided that they are stable-enough to stall a scanning 43S particle and located at a favourable distance from this non-AUG codon to trigger codon-anticodon base-pairing in the P site.

6. References

Alghoul F, Laure S, Eriani G, Martin F. 2021. Translation inhibitory elements from Hoxa3 and Hoxa11 mRNAs use uORFs for translation inhibition. *eLife* **10**: e66369.

Algire MA, Maag D, Lorsch JR. 2005. Pi Release from eIF2, Not GTP Hydrolysis, Is the Step Controlled by Start-Site Selection during Eukaryotic Translation Initiation. *Mol Cell* **20**: 251–262.

Ambrosini C, Destefanis E, Kheir E, Broso F, Alessandrini F, Longhi S, Battisti N, Pesce I, Dassi E, Petris G, *et al.* 2022. Translational enhancement by base editing of the Kozak sequence rescues haploinsufficiency. *Nucleic Acids Res* **50**: 10756–10771.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, *et al.* 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**: 1140–1146.

Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, Zhihao NT, Harmston N, D'Agostino G, Whiffin N, *et al.* 2022. A high-resolution map of human RNA translation. *Mol Cell* **82**: 2885-2899.e8.

Despons L, Martin F. 2020. How Many Messenger RNAs Can Be Translated by the START Mechanism? *Int J Mol Sci* **21**: 8373.

Diaz de Arce AJ, Noderer WL, Wang CL. 2018. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res* **46**: 985–994.

Eriani G, Martin F. 2018. START: STructure-Assisted RNA Translation. *RNA Biol* **15**: 1250–1253.

Guenther U-P, Weinberg DE, Zubradt MM, Tedeschi FA, Stawicki BN, Zagore LL, Brar GA, Licatalosi DD, Bartel DP, Weissman JS, *et al.* 2018. The helicase Ded1p controls use of near-cognate translation initiation codons in 5′ UTRs. *Nature* **559**: 130–134.

Harding HP, Zhang Y, Bertolotti A, Zeng H, Ron D. 2000. Perk Is Essential for Translational Regulation and Cell Survival during the Unfolded Protein Response. *Mol Cell* **5**: 897–904.

Hernández G, Osnaya VG, Pérez-Martínez X. 2019. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem Sci* **44**: 1009–1021.

Hetz C, Zhang K, Kaufman RJ. 2020. Mechanisms, regulation and functions of the unfolded protein response. *Nat Rev Mol Cell Biol* **21**: 421–438.

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.

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.

Kapp LD, Lorsch JR. 2004. GTP-dependent Recognition of the Methionine Moiety on Initiator tRNA by Translation Factor eIF2. *J Mol Biol* **335**: 923–936.

Kozak M. 1987. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* **196**: 947–950.

Kozak M. 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci* **87**: 8301–8305.

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

Kozutsumi Y, Segal M, Normington K, Gething M-J, Sambrook J. 1988. The presence of malfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. *Nature* **332**: 462–464.

Leppek K, Das R, Barna M. 2018. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol* **19**: 158–174.

Lind C, Åqvist J. 2016. Principles of start codon recognition in eukaryotic translation initiation. *Nucleic Acids Res* **44**: 8425–8432.

Lomakin IB, Steitz TA. 2013. The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* **500**: 307–311.

Martin F, Ménétret J-F, Simonetti A, Myasnikov AG, Vicens Q, Prongidi-Fix L, Natchiar SK, Klaholz BP, Eriani G. 2016. Ribosomal 18S rRNA base pairs with mRNA during eukaryotic translation initiation. *Nat Commun* **7**: 12622.

Nag N, Lin KY, Edmonds KA, Yu J, Nadkarni D, Marintcheva B, Marintchev A. 2016. eIF1A/eIF5B interaction network and its functions in translation initiation complex assembly and remodeling. *Nucleic Acids Res* gkw552.

Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K -i. 2007. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* **36**: 861–871.

Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by FACS -seq. *Mol Syst Biol* **10**: 748.

Paulin FEM, Campbell LE, O'Brien K, Loughlin J, Proud CG. 2001. Eukaryotic translation initiation factor 5 (eIF5) acts as a classical GTPase-activator protein. *Curr Biol* **11**: 55–59.

Pelham HRB, Jackson RJ. 1976. An Efficient mRNA-Dependent Translation System from Reticulocyte Lysates. *Eur J Biochem* **67**: 247–256.

Pernod K, Schaeffer L, Chicher J, Hok E, Rick C, Geslain R, Eriani G, Westhof E, Ryckelynck M, Martin F. 2020. The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic Acids Res* **48**: 6170–6183.

Pestova TV, Borukhov SI, Hellen CUT. 1998. Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. *Nature* **394**: 854–859.

Pestova TV, Lomakin IB, Lee JH, Choi SK, Dever TE, Hellen CUT. 2000. The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature* **403**: 332–335.

Rogers GW, Richter NJ, Lima WF, Merrick WC. 2001. Modulation of the Helicase Activity of eIF4A by eIF4B, eIF4H, and eIF4F. *J Biol Chem* **276**: 30914–30922.

Scheuner D, Song B, McEwen E, Liu C, Laybutt R, Gillespie P, Saunders T, Bonner-Weir S, Kaufman RJ. 2001. Translational Control Is Required for the Unfolded Protein Response and In Vivo Glucose Homeostasis. *Mol Cell* **7**: 1165–1176.

Schütz P, Bumann M, Oberholzer AE, Bieniossek C, Trachsel H, Altmann M, Baumann U. 2008. Crystal structure of the yeast eIF4A-eIF4G complex: An RNA-helicase controlled by protein–protein interactions. *Proc Natl Acad Sci* **105**: 9564–9569.

Sen ND, Gupta N, K. Archer S, Preiss T, Lorsch JR, Hinnebusch AG. 2019. Functional interplay between DEAD-box RNA helicases Ded1 and Dbp1 in preinitiation complex attachment and scanning on structured mRNAs in vivo. *Nucleic Acids Res* gkz595.

Simonetti A, Guca E, Bochler A, Kuhn L, Hashem Y. 2020. Structural Insights into the Mammalian Late-Stage Initiation Complexes. *Cell Rep* **31**: 107497.

Singh CR, Watanabe R, Zhou D, Jennings MD, Fukao A, Lee B, Ikeda Y, Chiorini JA, Campbell SG, Ashe MP, *et al.* 2011. Mechanisms of translational regulation by a human eIF5-mimic protein. *Nucleic Acids Res* **39**: 8314–8328.

Tabet R, Schaeffer L, Freyermuth F, Jambeau M, Workman M, Lee C-Z, Lin C-C, Jiang J, Jansen-West K, Abou-Hamdan H, *et al.* 2018. CUG initiation and frameshifting enable production of dipeptide repeat proteins from ALS/FTD C9ORF72 transcripts. *Nat Commun* **9**: 152.

Tang L, Morris J, Wan J, Moore C, Fujita Y, Gillaspie S, Aube E, Nanda J, Marques M, Jangal M, *et al.* 2017. Competition between translation initiation factor eIF5 and its mimic protein 5MP determines non-AUG initiation rate genome-wide. *Nucleic Acids Res* **45**: 11941–11953.

Thakur A, Gaikwad S, Vijjamarri AK, Hinnebusch AG. 2020. eIF2α interactions with mRNA control accurate start codon selection by the translation preinitiation complex. *Nucleic Acids Res* **48**: 10280–10296.

Thakur A, Hinnebusch AG. 2018. eIF1 Loop 2 interactions with Met-tRNA$_i$ control the accuracy of start codon selection by the scanning preinitiation complex. *Proc Natl Acad Sci* **115**: E4159–E4168.

Zhou F, Zhang H, Kulkarni SD, Lorsch JR, Hinnebusch AG. 2020. eIF1 discriminates against suboptimal initiation sites to prevent excessive uORF translation genome-wide. *RNA* **26**: 419–438.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

7. Acknowledgments

This work is funded by *Agence Nationale pour la Recherche* (ANR-17-CE12-0025-01, ANR-17-CE11-0024, ANR-20-COVI-0078), by *Fondation pour la Recherche Médicale* (project CoronaIRES), by *Fondation Bettencourt Schueller*, by University of Strasbourg and by the *Centre National de la Recherche Scientifique*. We would like to thank the IBMP AEG sequencing plateform. We would also like to thank Dr. Yves Nominé for useful advice for data analysis.

8. Figure legends

Figure 1: functional validation of the HEK293FT cell-free translation extracts

A. m^7G and IGR translation efficiencies rely on the stress level of the HEK extracts.

Histogram showing ^{35}S signal quantification of *in vitro* translated Renilla luciferase with HEK293FT (blue) or stressed HEK293FT (green) cell-free translation extracts from m^7G- or IGR- driven translation of the reporter RNAs shown on the left. Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution. p-values were calculated using a student t-test for independent samples. cds: coding sequence

*: $0.01 < p < 0.05$

**: $0.001 < p < 0.01$

***: $0.0001 < p < 0.001$

****: $p < 0.0001$

B. Calculation of the stress level of HEK293T cell-free translation extracts by Western Blot measurements of eIF2α phosphorylation. To determine the initial rate (at t = 0 min) of phosphorylated eIF2α (eIF2α-P) in both extracts, we hypothesized that the phosphorylation level of eIF2α is 100% after 75 min of *in vitro* translation of a m^7G-5'UTR β-globin-Renilla reporter RNA. Consequently, the level of the eIF2α-P is considered equal to total eIF2α after 75 min. Then, the initial rate of eIF2α-P is calculated using the ratio eIF2α-P (t = 0 min) / eIF2α-P (t = 75 min).

Left: Western Blots images (left) with antibodies specific for non-phosphorylated eIF2α (bottom) or phosphorylated eIF2α (top) using cell-free translation extracts prepared from HEK293FT cells cultured in physiological conditions (HEK-R, blue) or in stressed conditions (HEK-S, green)

Right: quantification of eIF2α phosphorylation using the Western Blot images

The dashed lines indicate the images (each in one black rectangle) were cropped (*i.e.* the lanes are not adjacent).

Figure 2: overview of the experimental strategy

A. Microfluidic-based screening set-up.

Each molecule of the starting library featuring the randomized AUG context is individualized into 2 pL droplets, which are further fused to 18 pL droplets containing an In Vitro Transcription-Translation (IVTT) reaction mixture. Upon incubation and eGFP synthesis, droplets are sorted depending on their fluorescence intensities. Active variants are recovered and processed for Illumina sequencing or for another round of selection.

B. The experimental set-up requires many PCR amplification steps which can be at the origin of the appearance of biased variants in each library.

C. Non-overlapping subgroups definition based on the expected translational efficiencies of the sequenced libraries. Numbers represent the number of distinct sequence variants in each subgroup after background filtering. See Figure 4 for background filtering.

Figure 3: overview of the EMCV-based RNA reporter library

A. Schematic representation of the RNA reporter library.

B. Representation of the modified EMCV IRES.

The EMCV IRES native AUGs have been removed to promote ribosomal scanning upon 43S assembly on the IRES through eIF4G interaction with the J-K domain. The A6-bifurcation loop, which contains 6 adenosines, is shown in red. The 43S particle composed of the ternary complex (TC = eIF2/GTP/Met-tRNA$^{Met}_{i}$) along with eIF3, eIF1, eIF1A and eIF5 is assembled on eIF4G which is recruited on the J-K domain of the IRES. The absence of the native AUG in the IRES allows the recruitment of a scanning pre-initiation complex that must find the further dowstream AUG whose nucleotide context has been randomized from -6 to +6 positions.

C. The modified EMCV IRES can promote ribosomal scanning.

SDS-PAGE analysis of ^{35}S-labelled translation proteins produced in a HEK293FT cell-free coupled transcription/translation reaction with and without reporter RNAs containing the modified EMCV IRES direclty followed by the eGFP coding sequence (-) or followed by a flexible 128 nucleotides-long 5'UTR and the eGFP coding sequence (scan). The dashed line indicates that the images were cropped.

Figure 4: analysis of the starting-library and background threshold determination

A. A-rich contexts are significantly biased in the starting library R0.

Left: Boxplot showing the distribution of the counts of the sequence variants found in the starting library. Whiskers are calculated using Q1 – 1.5 x IQR or Q3 + 1.5 x IQR where Q1 is the 25% percentile, Q3 the 75% percentile and IQR is the interquartile range Q3 – Q1. The median value is highlighted by the horizontal black line. The red line corresponds to Q3, meaning that 75% of the sequence variants are found less than 21 times.

Right: boxplot showing the distribution of the sequence variants in the starting library as a function of the number of adenosines in their sequence.

B. Background threshold determination for each library.

Left: scatterplot (blue) showing the cumulated proportions p as a function of 1/p ; the dashed black line corresponds to the fitted filtering function and the red dashed line corresponds to the calculated threshold.

Right: log-distribution of the sequence variants proportions (%obs) plotted as kernel density estimation for each library.

R0 (starting library), Rn (library selected after n round(s) of selection), NA (non-attributed sequences)

Figure 5: characterization of inefficient AUG nucleotide contexts for translation initiation

A. In-frame stop codons and out-of-frame upstream AUGs are depleted through the screening process.

From left to right,

- The left heatmap shows the motifs' scores, which correspond to their coverage in the starting library. A score that is equal to 1 (green) means that every possible sequence that feature the motif were found in the starting library.
- The middle part displays the motifs' sequences.
- The right heatmap displays the distribution of the sequences featuring the analysed motif in each previously defined subgroup, with NA being the non-attributed ones. Therefore, the sum of each row equals 1 to the nearest rounding. For example, 87.8% of the sequences that feature the motif NNNNNNAUGUAG (top line) were found in the starting library. Among these sequences, 64% are found in the X0 subgroup, 4% are found in X1, they were lost in X2 and X3 and 32% were non-attributed (*i.e.* poorly represented in the starting library).

B. The motif UAGNNNAUGNNN stands out among the inefficient ones, although it is poorly covered in the starting library.

From left to right,

- The left heatmap shows the motifs' score which correspond to their coverage in the starting library. A score that is equal to 1 (green) means that every sequence that feature the motif were found in the starting library.
- The middle part displays the motifs' sequences.
- The last heatmap displays the distribution of the sequences featuring the analysed motif in each previously defined subgroup, with NA being the non-attributed ones. Therefore, the sum of each row equals 1 to the nearest rounding. For example, 38.7% of the sequences that feature the motif UAGNNNAUGNNN (top line) were found in the starting library. Among these sequences, 17% are found in the X0 subgroup, 7% are found in X1, they were lost in X2 and X3 and 76% were non-attributed (*i.e.* poorly represented in the starting library).

Figure 6: average sequence calculations in the selected libraries

A. The average AUG context of the selected variants is A-rich.

Heatmap representing the observed frequency of nucleotide at each randomized position in subgroups R0, X1, X2, X3 and NA. The sum of frequencies for each position is 1.

The colormap is centered on 0.25, which is the equi-reprensation of each nucleotide.

NA: non-attributed sequences (*i.e.* poorly represented in the starting library).

B. -3 and +4 positions converge to specific nucleotide combinations.

Heatmap representing the calculated convergence factor for each randomized position in each subgroup.

NA: non-attributed sequences (*i.e.* poorly represented in the starting library).

C. A-rich combinations are gradually enriched.

Heatmap representing the calculated ratio (fold change) of the proportion of unique sequences having 0 to 9 times the same nucleotide in the randomized region between each non-overlapping subgroup and the starting library (Xn/R0)

NA: non-attributed sequences (*i.e*. poorly represented in the starting library).

D. The nature of the second codon has weak impact on translation initiation.

Scatterplot showing the normalized proportions of the amino acids encoded by the second codon.

The normalized proportion is calculated by dividing the observed proportion of the amino acid encoded by the second codon by the number of codons encoding this amino acid.

Figure 7: analysis of Kozak and non-Kozak sequences

A. Representation of the nucleotide proportion for each position in the 11,8724 ORFs from the human ORFeome.

B. Kozak and non-Kozak contexts are not equally represented in the non-overlapping subgroups.

Stacked histogram displaying the proportion of Kozak (NNNRNNAUGRNN, R = {A,G}) and non-Kozak (NNNYNNAUGYNN, Y = {C,U}) contexts in each non-overlapping subgroup. The plot on the right shows a magnification of the Kozak sequence having at least one C at positions -5 or -4 or -2 or -1 in X3 subgroup.

NA: non-attributed sequences (*i.e.* poorly represented in the starting library).

C. Cytidines are poorly represented and do not associate with Kozak sequences in the X3 subgroup.

Pie chart representing the proportion of sequence variants having at least one C at positions -5 or -4 or -2 or -1in the X3 subgroup. The plot on the right shows the proportion of sequences having a Kozak sequence (NNNRNNAUGRNN, R = {A,G}).

Figure 8: influence of the nucleotides at positions -3 and +4 on the frequency of Kozak and non-Kozak contexts in each library

A. stacked histogram displaying the proportion of each possible Kozak sequence (NNNRNNAUGRNN, R = {A,G}) in each non-overlapping subgroup.

B. stacked histogram displaying the proportion of each possible non-Kozak (NNNYNNAUGYNN, Y = {C,U}) in each non-overlapping subgroup.

NA: non-attributed sequences (*i.e.* poorly represented in the starting library).

Figure 9: the selected sequences do not converge towards a specific motif (1)

A. Schematic representation of non-overlapping and overlapping motifs within a subgroup.

B. Distribution of the observed proportions of each k-motif in the X3 subgroup.

For example, the most abundant 2-motif in X3 represents 17.5% of X3, and the least abundant represents 2.3%.

C. Analysis of the most abundant 2- and 3- non-overlapping motifs of X3.

Left: pie chart representing the proportions of every 2-motifs of X3 that are part of the -3/+4 combinations. These motifs are non-overlapping.

Right: pie chart representing the proportions of the first 3-motifs of X3 that are part of the -6/-3/+4 combinations. These motifs are non-overlapping. The plot only shows the cumulated proportion until 55% to improve clarity.

D. Plot representing cumulative proportions of sequences of X3 that are contained into each possible 2-motifs. These curves show the X3 subgroup coverage as a function of the cumulated proportions of each possible set of non-overlapping motifs. Each curve represents the increasing proportions from one pie chart. For example, 50% of the non-overlapping 2(3)-motifs of the first iteration summarize 71% (75%) of the efficient subgroup X3.

n/o: non-overlapping

Figure 10: the selected sequences do not converge towards a specific motif (2)

A. scheme representing the general repartition of the sequences sharing the same k-motif.

B. stacked histogram showing the proportions of k-motifs that are found in one, two, three subgroup(s) simultaneously or found in all subgroups. For example, all the analysed 2-motifs are found in all subgroups, and 52% of the analysed 5-motifs are found in three subgroups simultaneously.

Figure 11: the selected sequences do not converge towards a specific motif (3)

These graphs show 7- (**A**) and 8- (**B**) motifs with respectively at least one quarter or one half of their variants found in the X3 subgroup and the rest found in at least one of the other subgroups except the NA subgroup.

From left to right,

- The left heatmap shows the motifs' scores, which correspond to their coverage in the starting library. A score that is equal to 1 (green) means that every possible sequence that feature the motif were found in the starting library.
- The middle part displays the colour representation of the motifs' sequences.
- The right heatmap displays the distribution of the sequences featuring the analysed motif in each previously defined subgroup, with NA being the non-attributed ones. Therefore, the sum of each row equals 1 to the nearest rounding. For example, 100% of the sequences that feature the motif NAAAAAAUGAAN (A, top line) were found in the starting library. Among these sequences, 62% are found in the X0 subgroup and 38% are found in the X3 subgroup.

Figure 12: bias calculations in the representation of A-rich contexts present in the annotated ORFs of the human genome

A. Distribution of the observed/expected (obs/exp) ratios of the 511 possible A-containing k-motifs (k = {2, …, 8}).

B. Colour representation of the outliers A-motifs represented in a red-dashed rectangle in A.

C. Distribution of the observed/expected (obs/exp) ratios as a function of the number of A in the A-rich AUG contexts.

Figure 13: the representation of C-rich AUG contexts is biased in the human genome

A. Colour representation of the analysed C- and D- rich AUG contexts.

B. Distribution of the observed/expected ratios as a function of the number of C or D in the AUG contexts.

R = {A, G}

D = {A, U, G}

Figure 14: the annotated human ORFs do not feature a minimal AUG context

A. Pie chart showing the proportion of the NNNNNNAUGNNN combinations that are present in the analysed human ORFs sequences. Of note, in-frame stop codons and out-of-frame AUGs combinations represent 7.8% of the total number of distinct NNNNNNAUGNNN contexts and therefore cannot account alone for the 89% of the AUG contexts that are not found in the human ORFs.

B. Boxplot showing the distribution of the counts of each individual AUG nucleotide contexts in the human ORFs. The number 5 in red is the 75th percentile, meaning 75% of the AUG nucleotide contexts are found less than 5 times. The dashed cyan rectangle highlights the sequences displayed in C.

C. Color representation of the outliers highlighted with the cyan rectangle in B.

D. Distribution of the observed proportions of each k-motif in the human ORFs. For exemple, the most abundant 2-motif in the human ORFs accounts for 20% of total sequences, and the least abundant represents approx. 2%.

E. Plots showing the human ORFs coverage as a function of the cumulated proportions of each set of non-overlapping motifs displayed in the pie charts (**F**). 50% of the non-overlapping 2-motifs at -3 and -1 positions summarize at best 79% of the AUG nucleotide contexts of the human ORFs.

n/o: non-overlapping

F. Analysis of the first six most abundant sets of non-overlapping 2-motifs in the human ORFs, whose cumulative proportions are shown in E.

Figure 15: *in vitro* translation assays corroborate the results of the screening experiment and the genome analysis

A. schematic representation of the model RNA reporters used for *in vitro* translation assays.

B: histograms showing luminescence quantification of *in vitro* translated Renilla luciferase from various RNA reporters with RRL. Relative luminescence units (R.L.U.) were normalized to those obtained with the GCCACCAUGGCG reporter for both graphs (even though it is not shown the left graph).

cds = coding sequence

UTR: untranslated region

Figure 16: schematic representation of the model RNA reporters used for investigating the involvement of the START mechanism in non-AUG translation initiation with *in vitro* translation assays

The three *cis*-acting parameters that will be investigated are displayed in blue: 1) the distance between the a11wt structure and the start-codon, 2) the stability of the a11 structure and 3) the nature of the start-codon.

TEVp: Tobacco Etch Virus protease

Rluc cds: Renilla luciferase (Rluc) coding sequence (cds)

Figure 17: the optimal position of the secondary structure for translation on a CUG codon ranges from +23 to +26

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with RRL from various a11 RNA reporters initiating with a CUG codon (**A** and **B**, orange) or with an AUG codon (**B**, in blue). Relative luminescence units (R.L.U.) were normalized to those obtained with the +20 CUG reporter (**A**) or AUG reporter (**B**). Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution. p-values were calculated using a student t-test for independent samples.

*: $0.01 < p < 0.05$

**: $0.001 < p < 0.01$

***: $0.0001 < p < 0.001$

****: $p < 0.0001$

Figure 18: the optimal stability of the secondary structure is influenced by the cell type

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with RRL (**A**), HEK293FT (**B**) or ^{35}S signal quantification of *in vitro* translated Renilla luciferase with SH-SY5Y (**C**) cell-free translation extracts from various a11 RNA reporters initiating with a CUG codon (orange) or with an AUG codon (blue). Relative luminescence units (R.L.U.) (A, B) or pixel intensities (C) were normalized to those obtained with the a11wt AUG reporter. Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution. p-values were calculated using a student t-test for independent samples.

*: $0.01 < p < 0.05$

**: $0.001 < p < 0.01$

***: $0.0001 < p < 0.001$

****: $p < 0.0001$

D. Schematic representation of the a11 structures. The silent mutations realized to lower the a11 structure stability are shown in blue circles. They result in either G-U or C-A mismatches. For a11mut2 structure, silent mutations are shown in blue circles and missense mutations in red circles.

Figure 19: the START mechanism does not rescue translation initiation for all AUG-like codons

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with RRL (**A**) or ^{35}S signal quantification of *in vitro* translated Renilla luciferase with regular HEK293FT (**B**, HEK-R in blue) or stressed HEK293FT (**B**, HEK-S in green) cell-free translation extracts from various a11 RNA reporters initiating with AUG, CUG, GUG, UUG, ACG or AUC codons. Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution. p-values were calculated using a student t-test for independent samples.

*: $0.01 < p < 0.05$

**: $0.001 < p < 0.01$

***: $0.0001 < p < 0.001$

****: $p < 0.0001$

A. Relative luminescence units (R.L.U.) were normalized to those obtained with the AUG reporter using the regular HEK293FT cell-free translation extracts.

B. Relative pixel intensities were normalized to those obtained with the AUG reporter for both extracts independently.

C. Relative pixel intensities were normalized to those obtained with the AUG reporter using the regular HEK293FT cell-free translation extracts.

D. Same as B but with the IGR reporter showed as a positive control.

9. Supplementary Figures and Table

Supplementary Figure S1: the eGFP coding sequence is a strong inhibitory *cis*-element of translation initiation that is not specific to the cell-free translation extracts and 5'UTR

A & **B** (left): schematic representation of the reporter RNA.

A-right: fluorescence analysis of *in vitro* translated eGFP from IGR(CrPV)-driven RNA reporters using cell-free translation extracts prepared from drosophila S2 cells. x-axis represents minutes and y-axis represent relative fluorescence units. cds: coding sequence.

B-right: SDS-PAGE analysis of ^{35}S-labelled translation proteins produced in a HEK293FT (left gel) or RRL (right gel) cell-free IVTT reaction in the presence of model reporter RNAs containing the modified EMCV IRES direclty followed by the eGFP coding sequence (3 and 4) or followed by a flexible 30 nucleotides-long 5'UTR and the eGFP coding sequence (5 and 6).

Supplementary Figure S2: only A-rich contexts are biased in the starting library

Boxplots showing the distribution of the sequence variants in the starting library as a function of the number of A, C, G, U in their sequence.

Supplementary Figure S3: an A-rich context does not rescue in frame-stop codons or out-of-frame upstream AUGs

From left to right,

- The left heatmap shows the motifs' scores, which correspond to their coverage in the starting library. A score that is equal to 1 (green) means that every possible sequence that feature the motif were found in the starting library.
- The middle part displays the motifs' sequences.
- The right heatmap displays the distribution of the sequences featuring the analysed motif in each previously defined subgroup, with NA being the non-attributed ones. Therefore, the sum of each row equals 1 to the nearest rounding. For example, 100% of the sequences that feature the motif AAAAANAUGUAG (top line) were found in the starting library. Among these sequences, 100% are found in the X0 subgroup

Supplementary Figure S4: the AUG context of the NNNNNNAUGNNN reporter library is located in a predicted structure-free environment

Mfold modelisations of the reporter RNA without providing any constraints. The initiator AUG is squared in black.

The eGFP coding sequence (without the AUG) starts from the green arrow.

A. Without the GAA(CAA)$_7$ linker.

B. With the GAA(CAA)$_7$ linker.

Supplementary Table S1: oligonucleotides' sequences

Figure 1: functional validation of the HEK293FT cell-free translation extracts

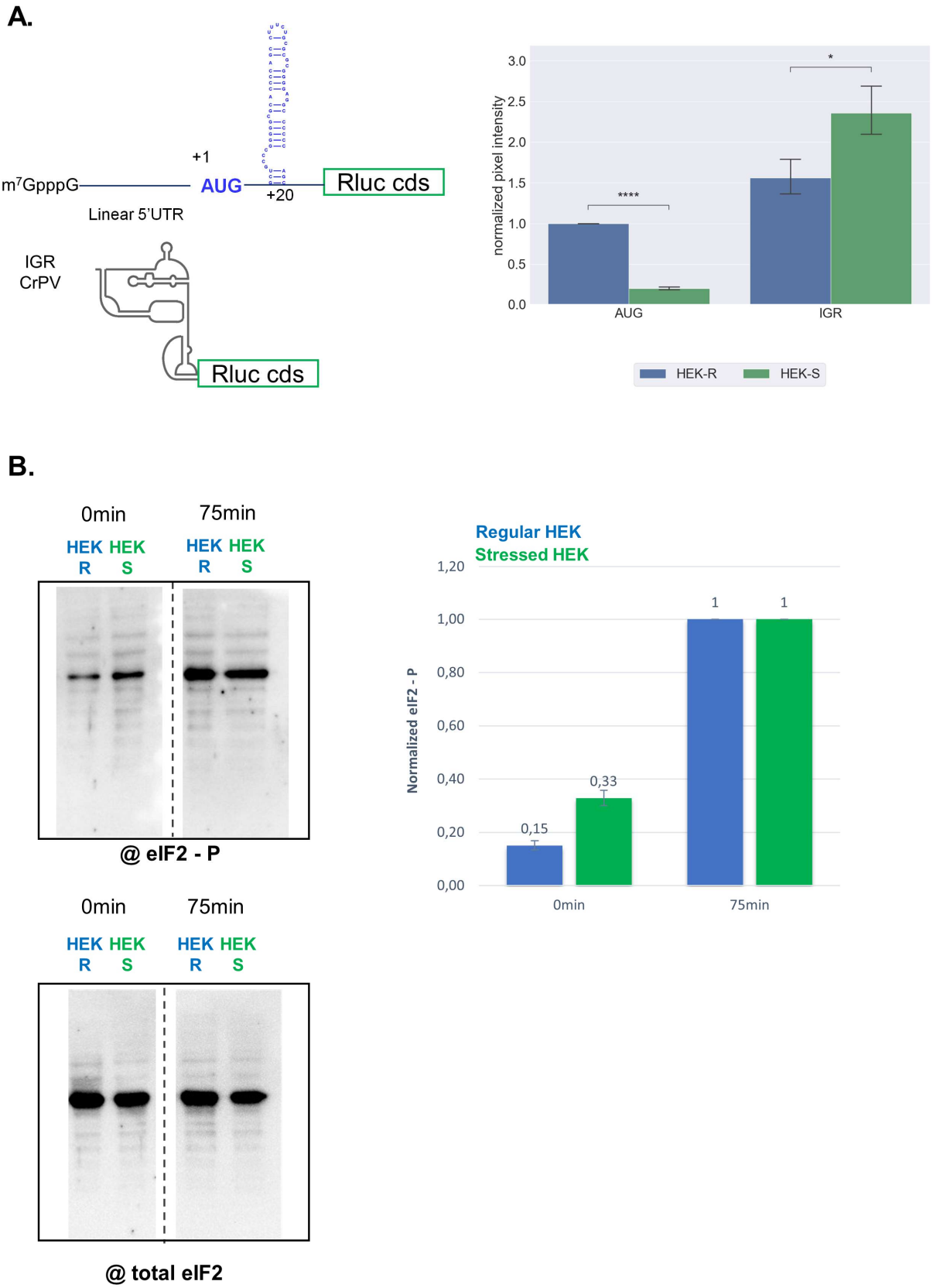


Figure 2: overview of the experimental strategy

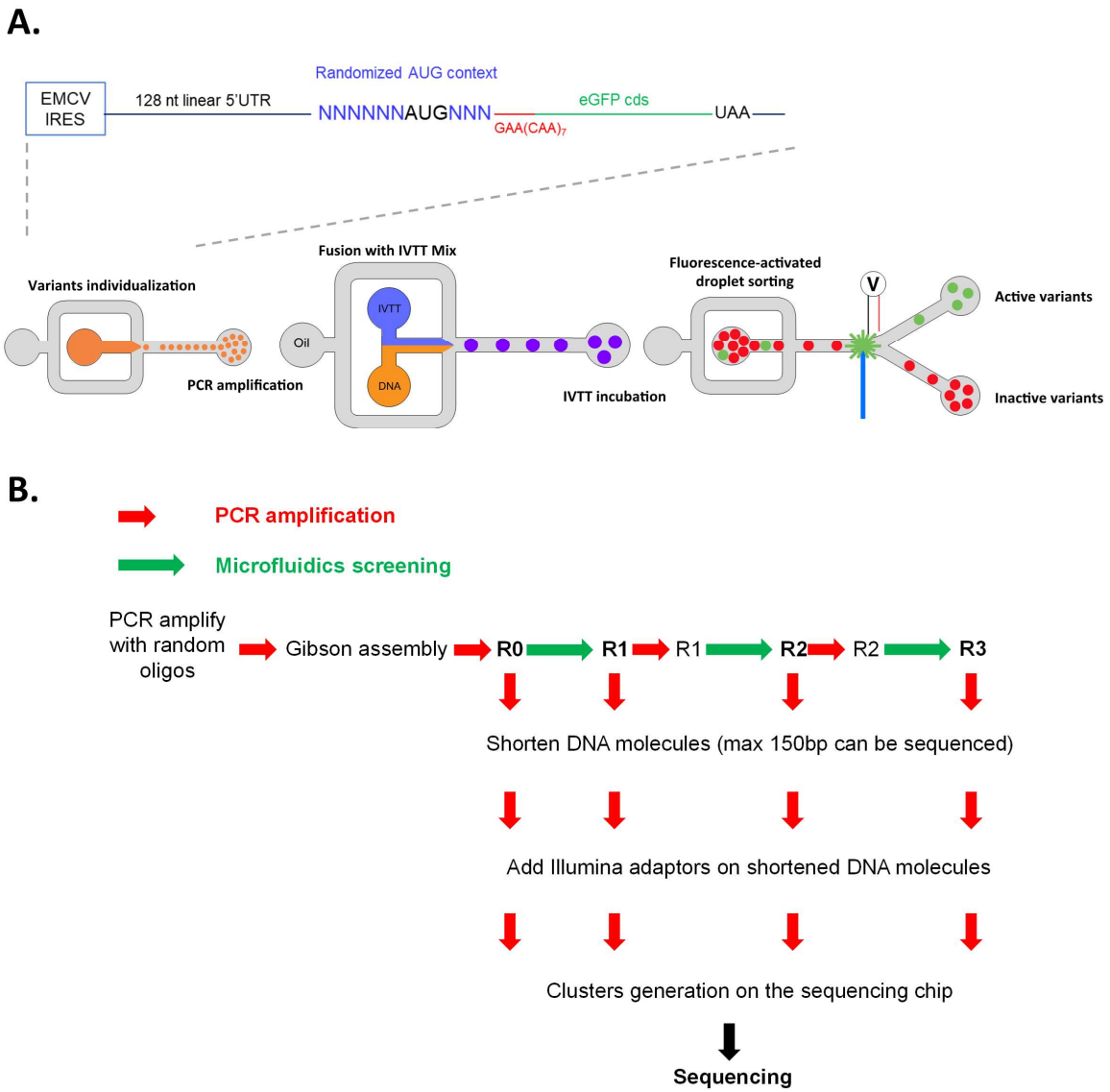


Figure 3: overview of the EMCV-based RNA reporter library

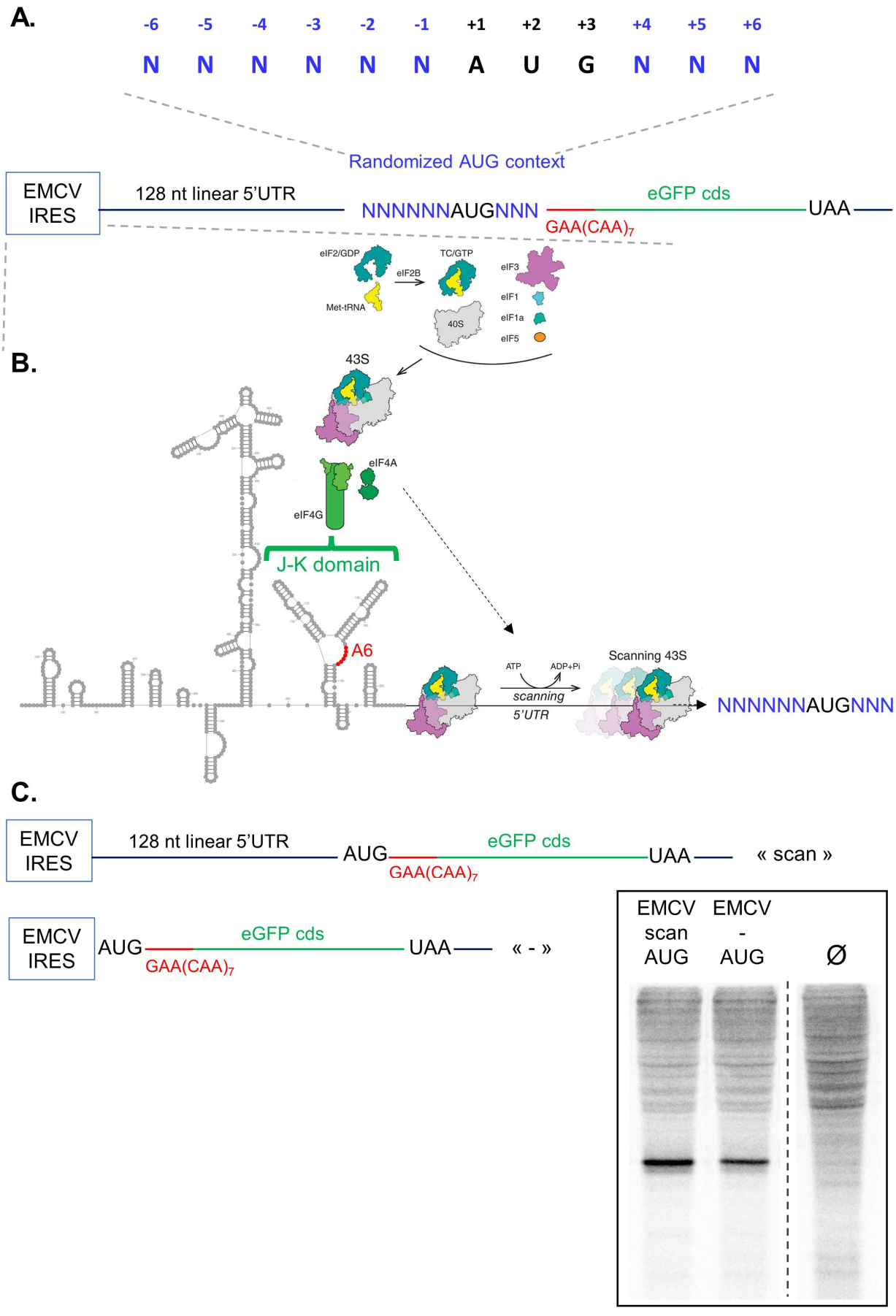


Figure 4: analysis of the starting library and background threshold determination

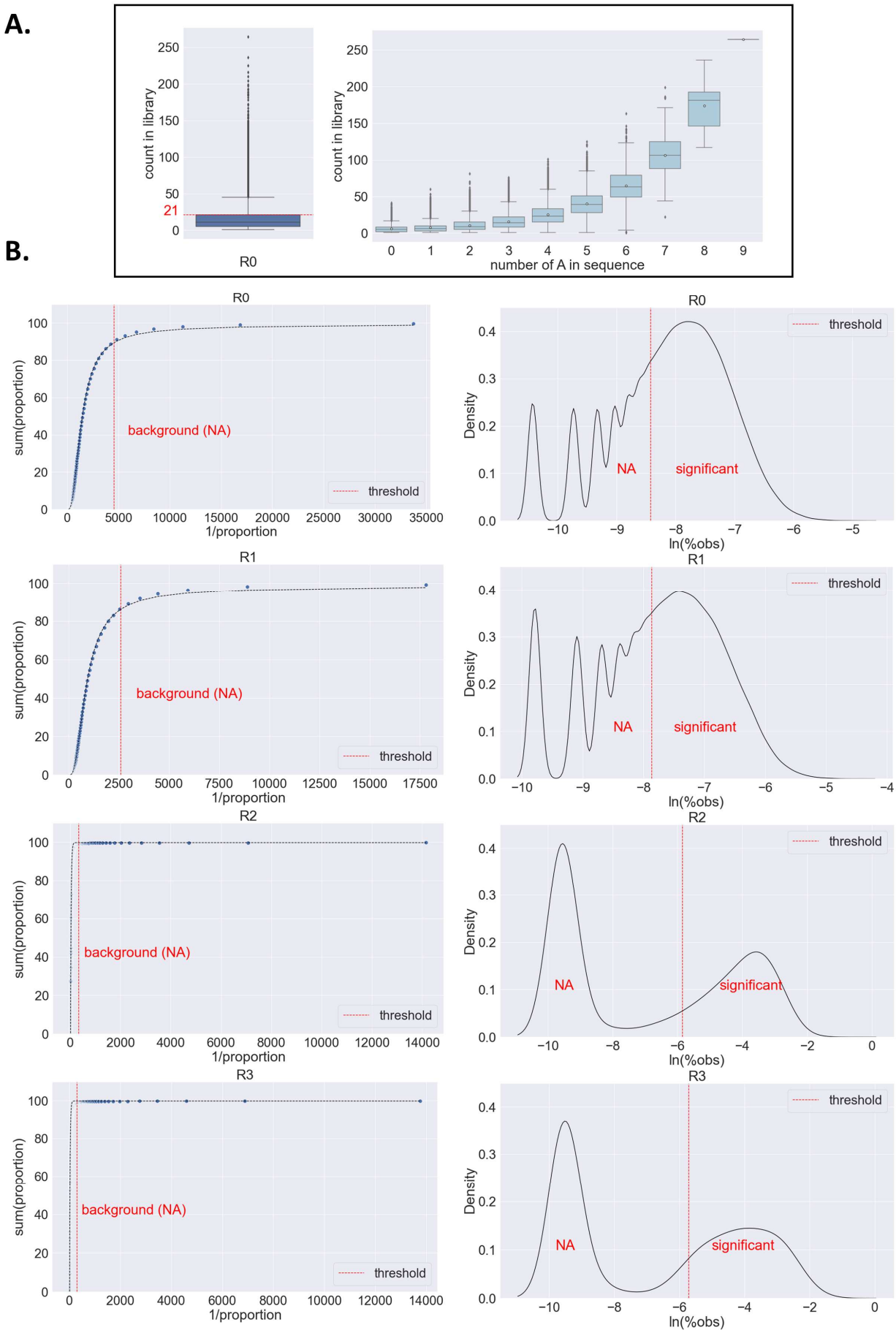


Figure 5: characterization of inefficient AUG nucleotide contexts for translation initiation

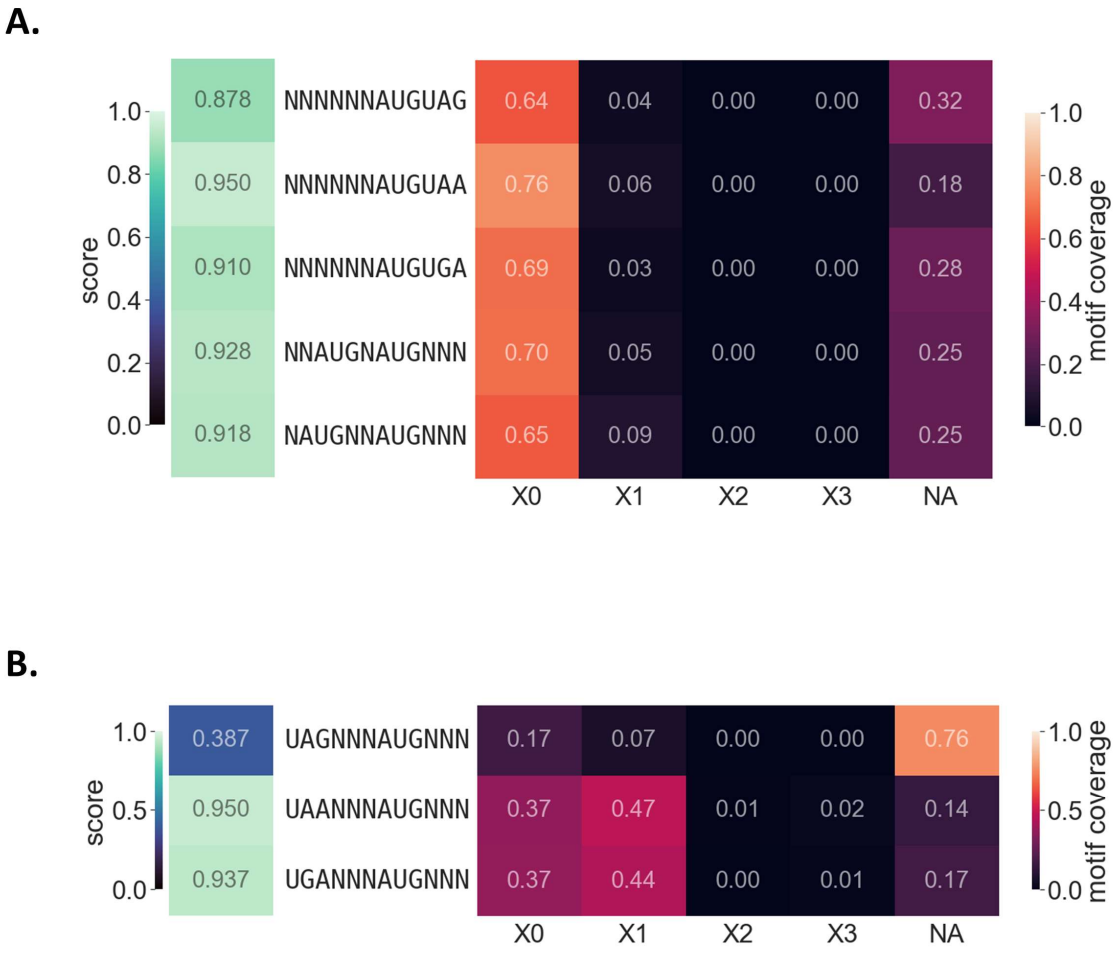


Figure 6: average sequence calculations in the selected libraries

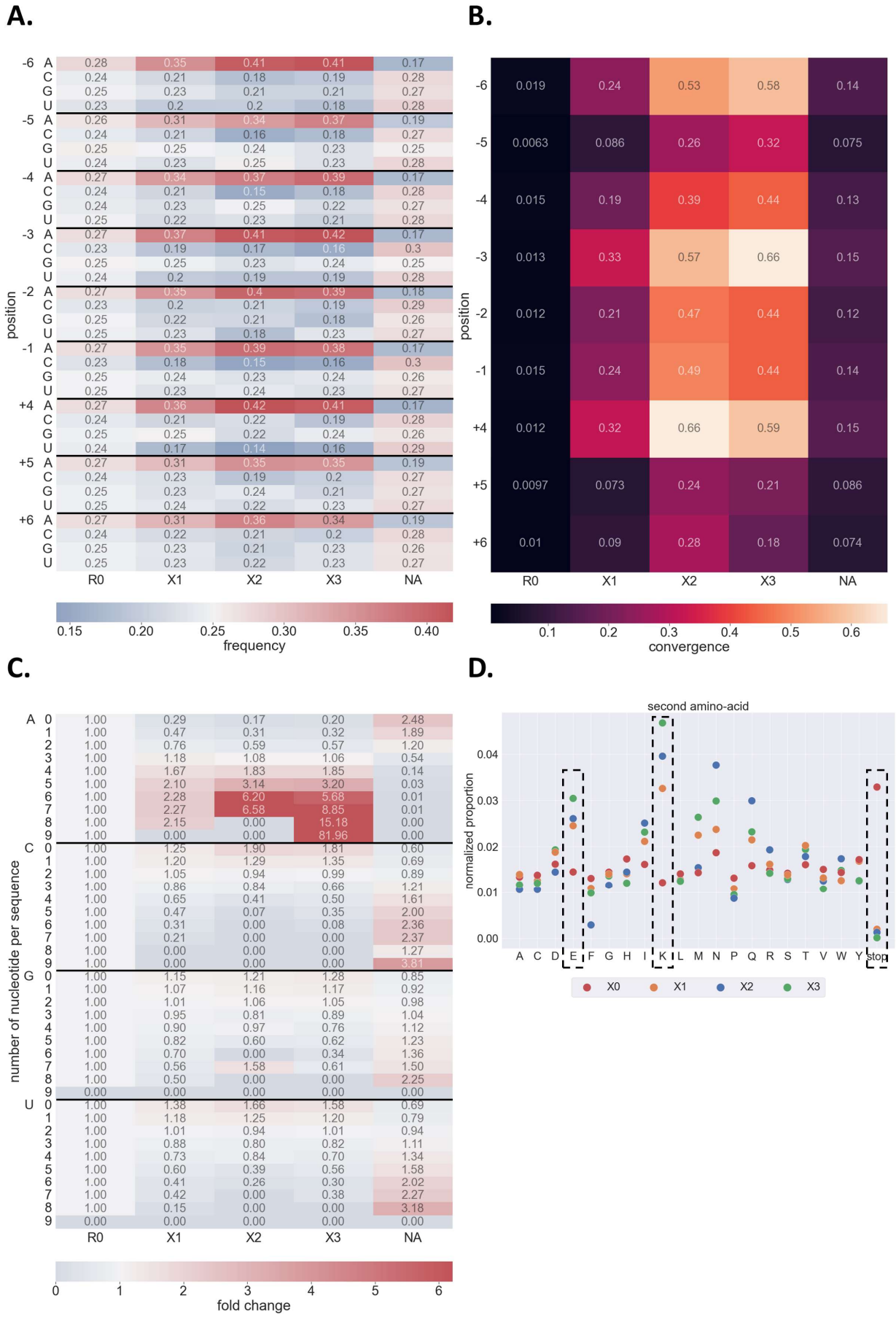


Figure 7: analysis of Kozak and non-Kozak sequences

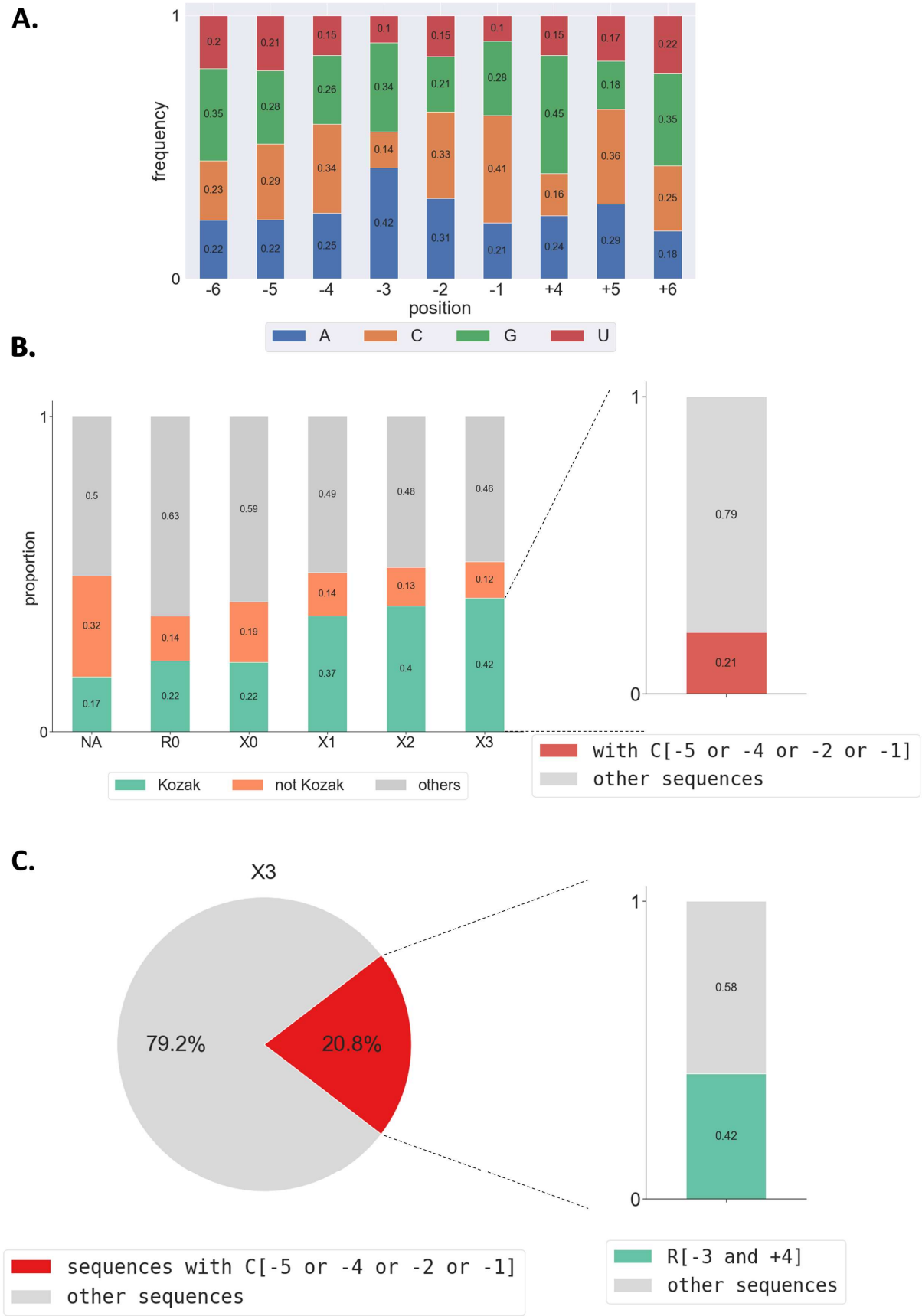
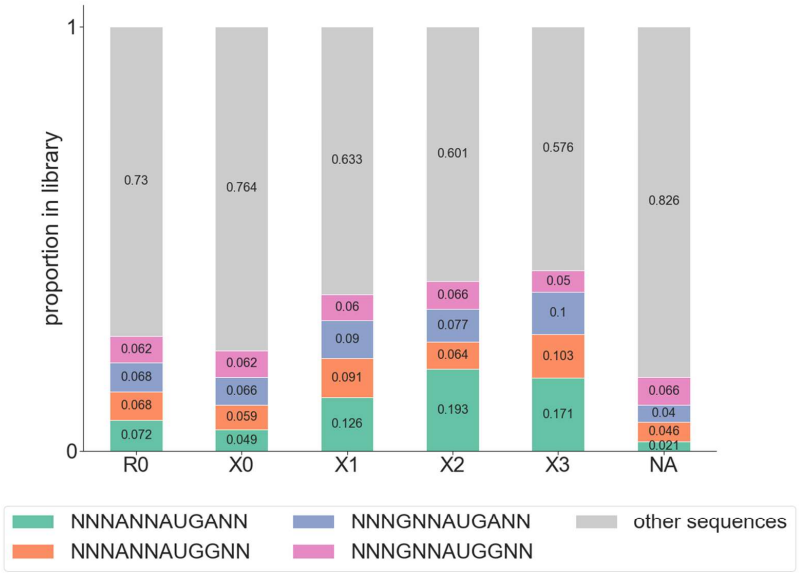


Figure 8: influence of the nucleotides at positions -3 and +4 on the frequency of Kozak and non-Kozak contexts in each library

A.



B.

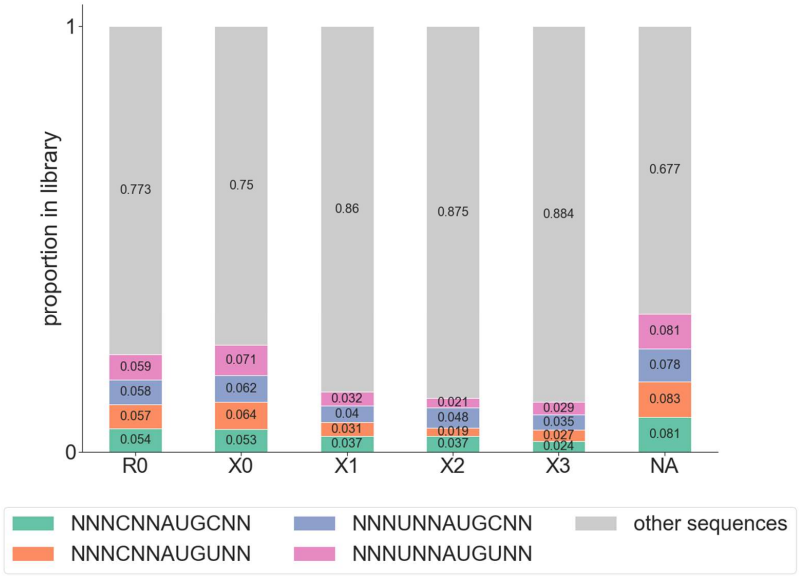


Figure 9: the selected sequences do not converge towards a specific motif (1)

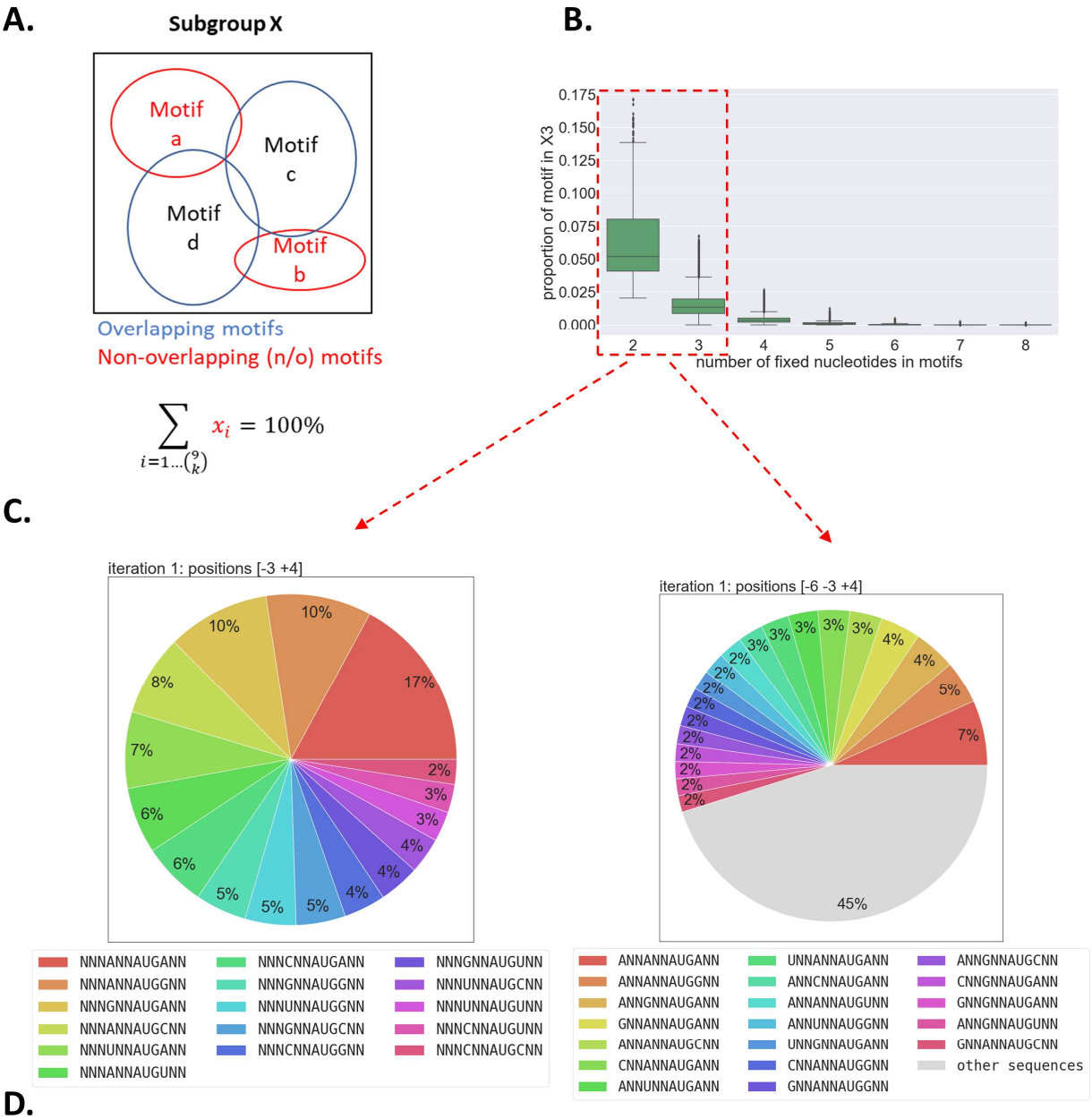
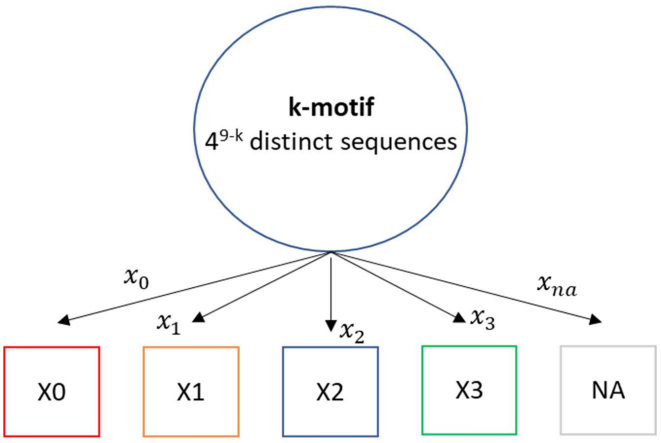


Figure 10: the selected sequences do not converge towards a specific motif (2)

A.



$$\sum_{i=\{0,1,2,3,na\}} x_i = 100\%$$

B.

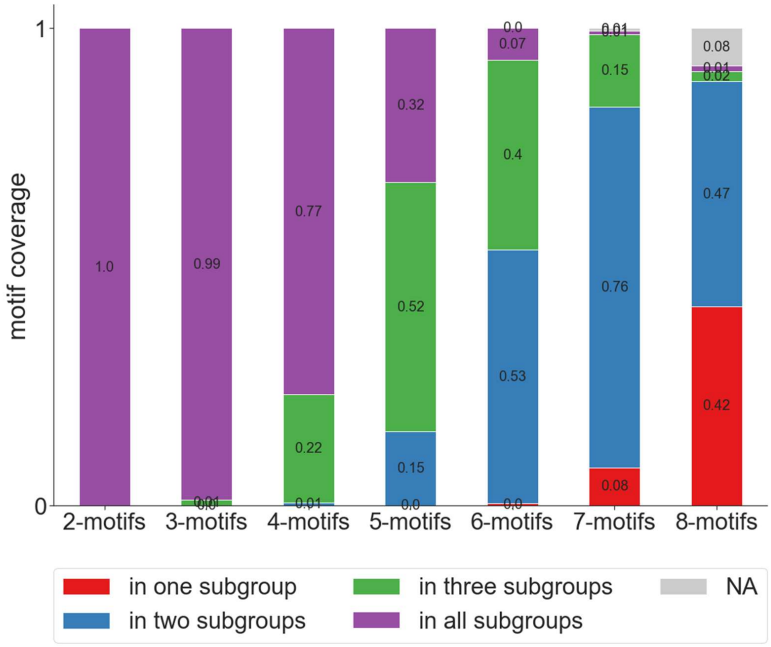


Figure 11: the selected sequences do not converge towards a specific motif (3)

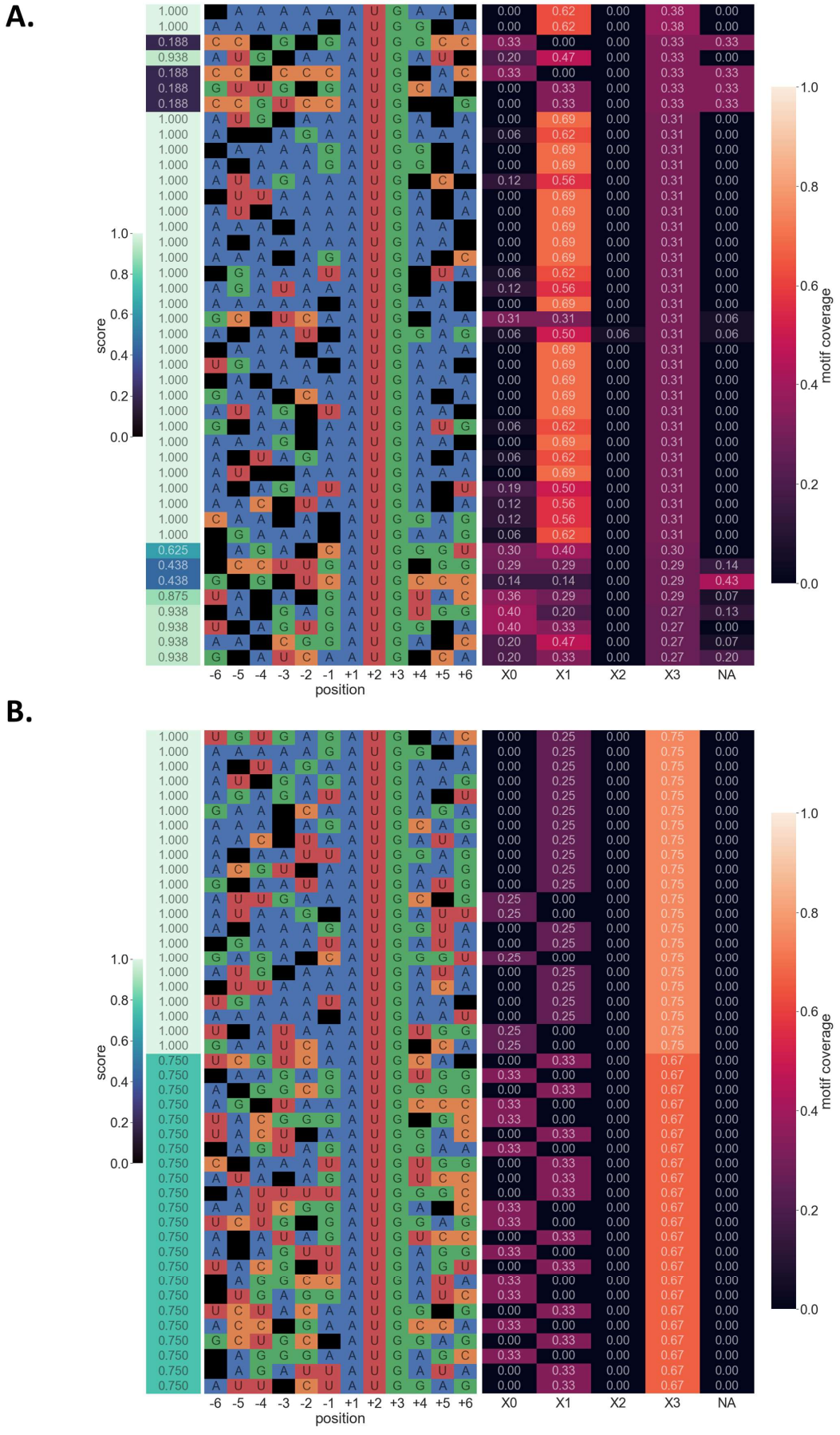


Figure 12: bias calculations in the representation of A-rich contexts present in the annotated ORFs of the human genome

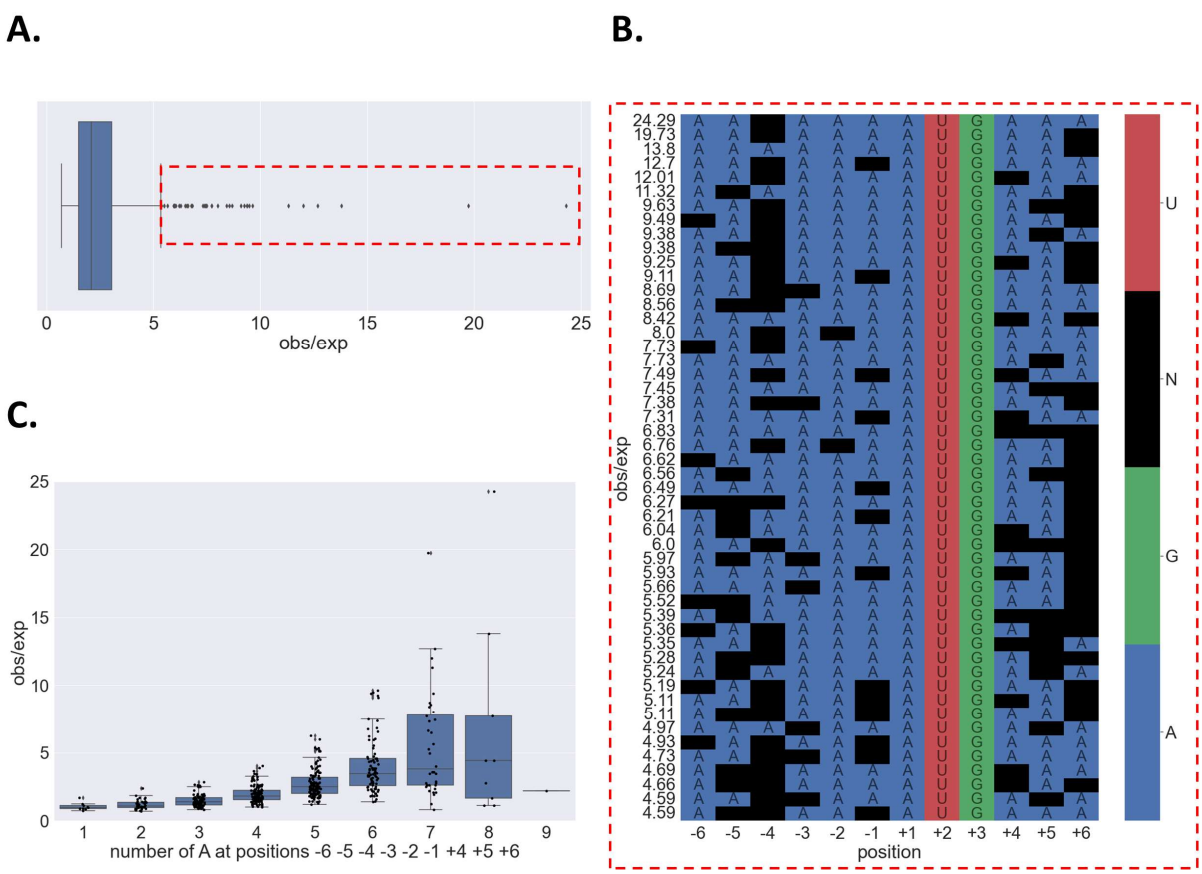


Figure 13: the representation of C-rich AUG contexts is biased in the human genome

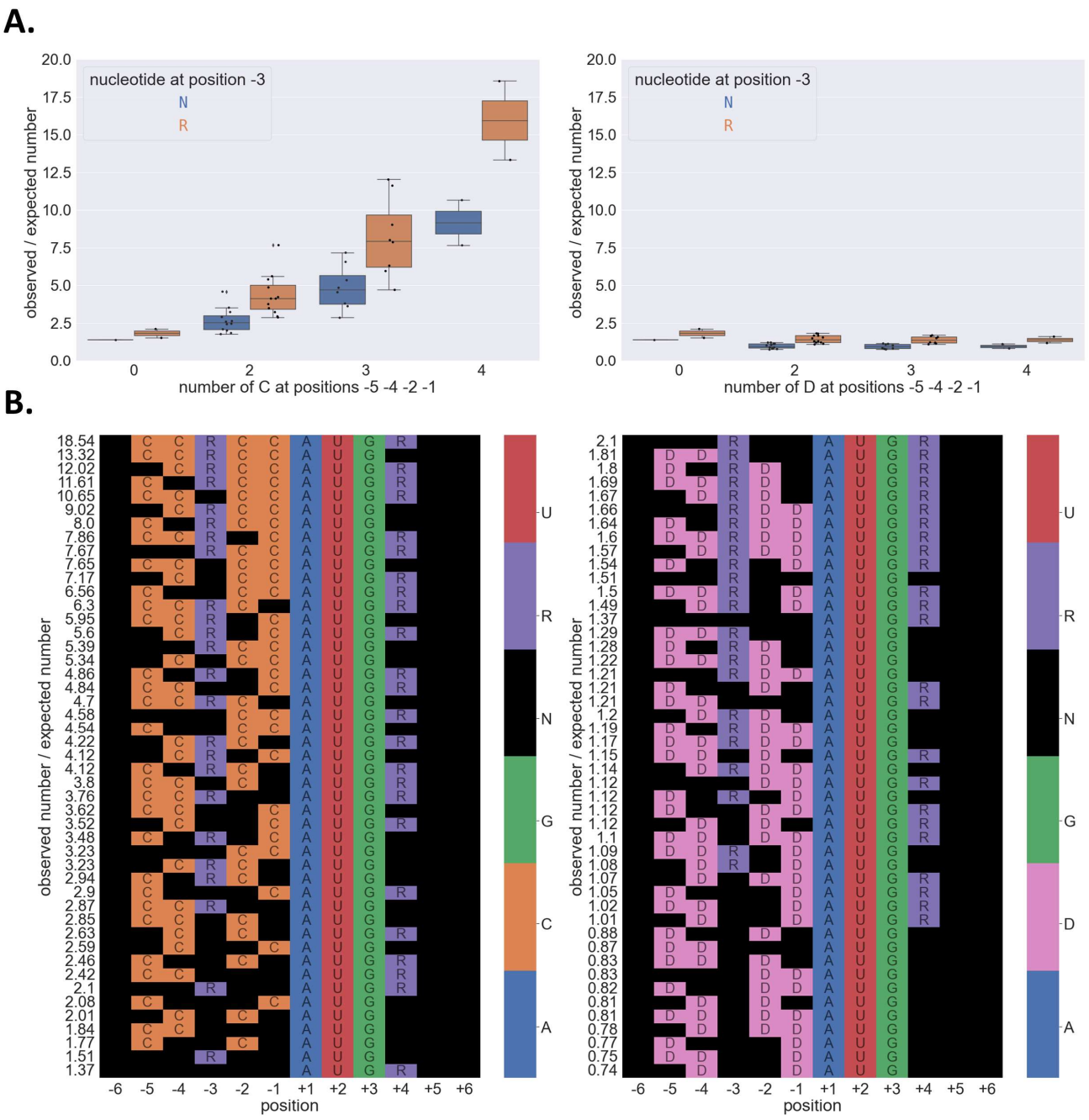


Figure 14: the annotated human ORFs do not feature a minimal AUG context (part 1)

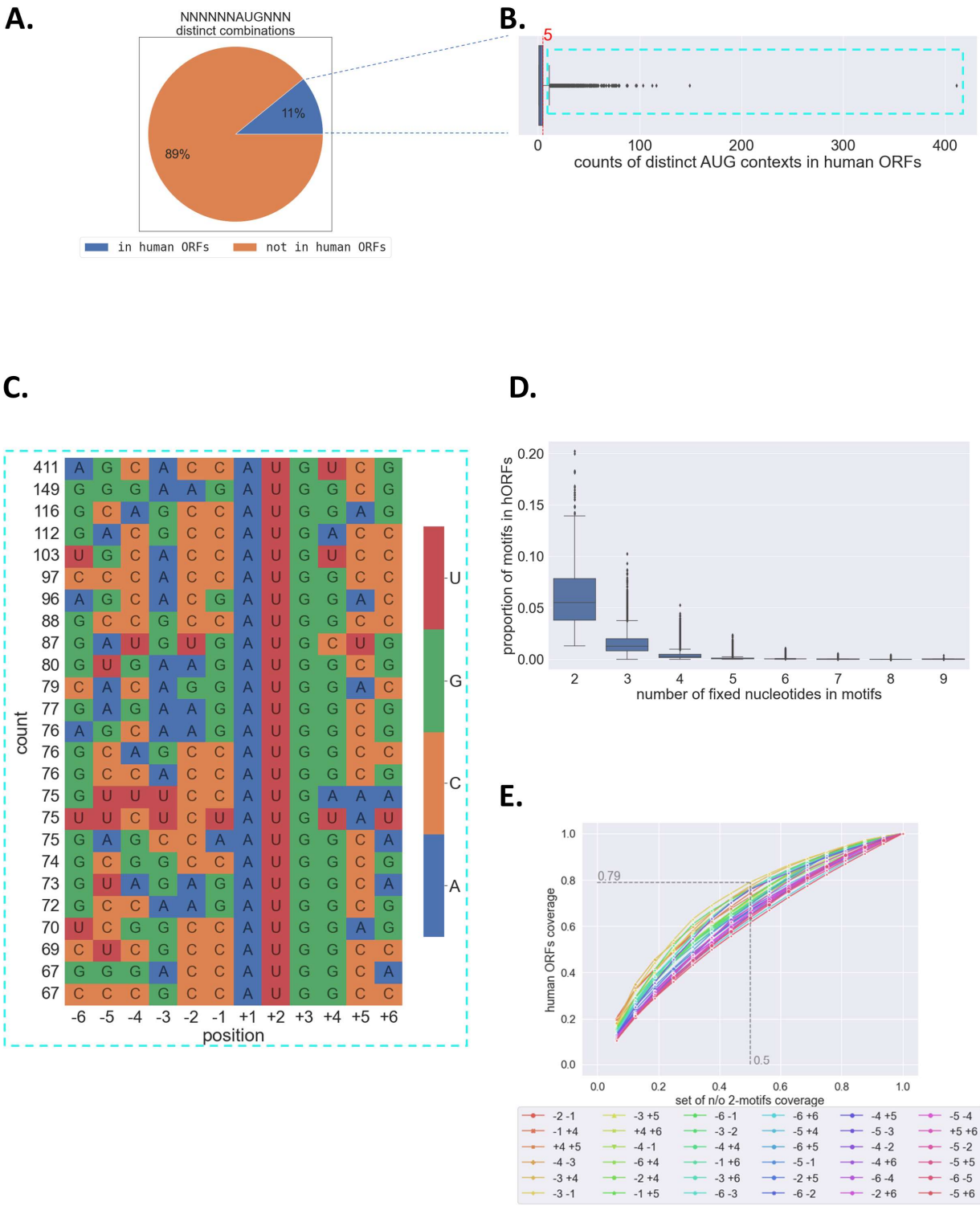


Figure 14: the annotated human ORFs do not feature a minimal AUG context (part 2)

F.

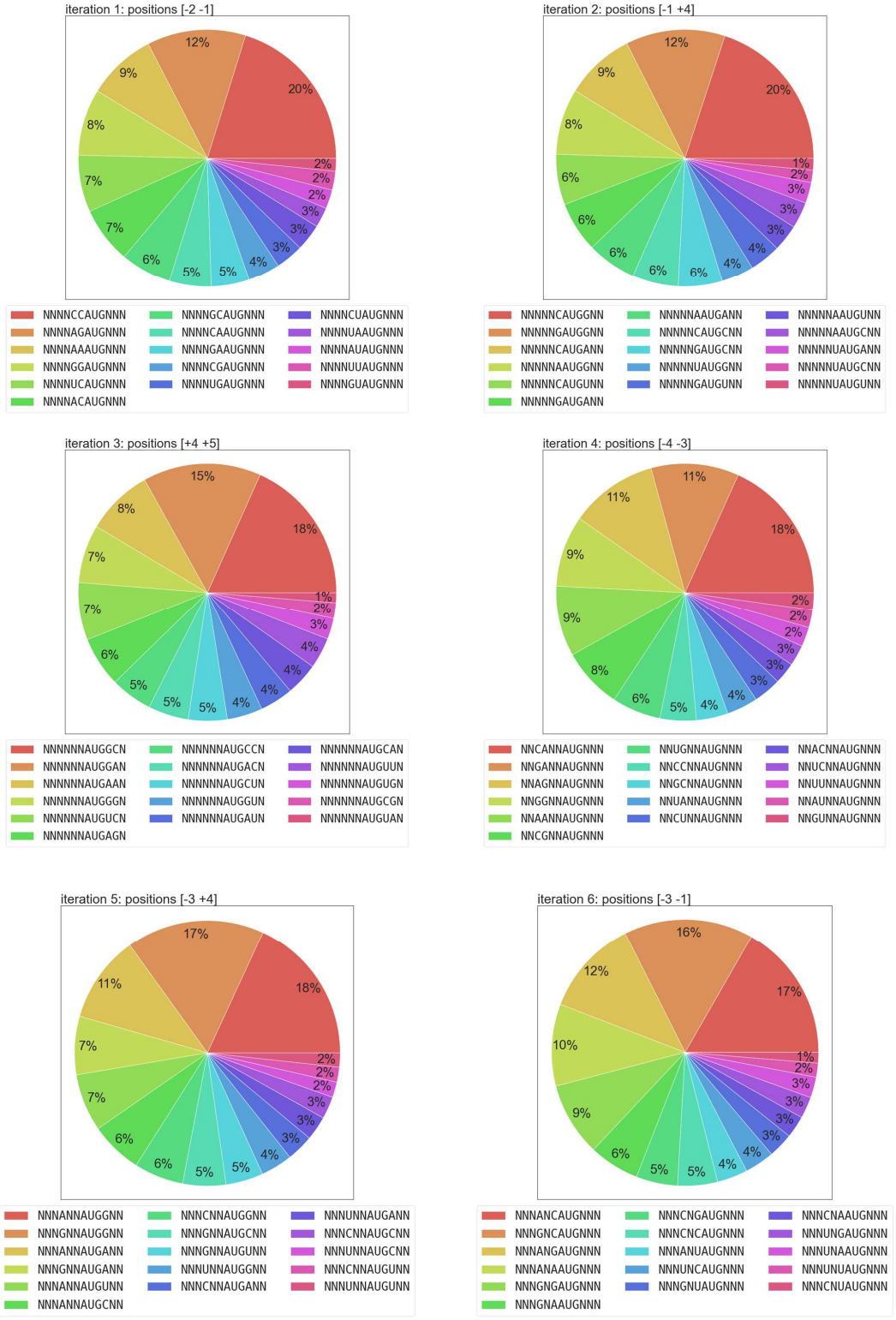


Figure 15: *in vitro* translation assays corroborate the results of the screening experiment and the genome analysis

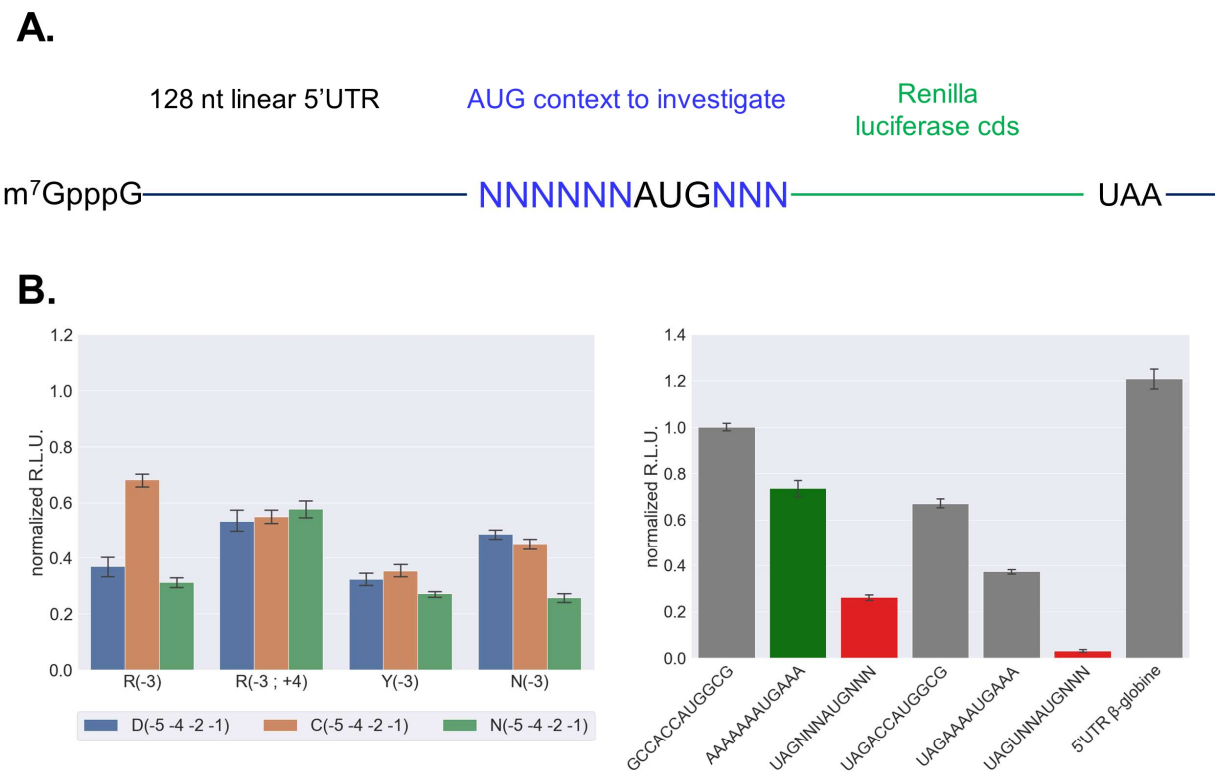


Figure 16: schematic representation of the model RNA reporters used for investigating the involvement of the START mechanism in non-AUG translation initiation with *in vitro* translation assays

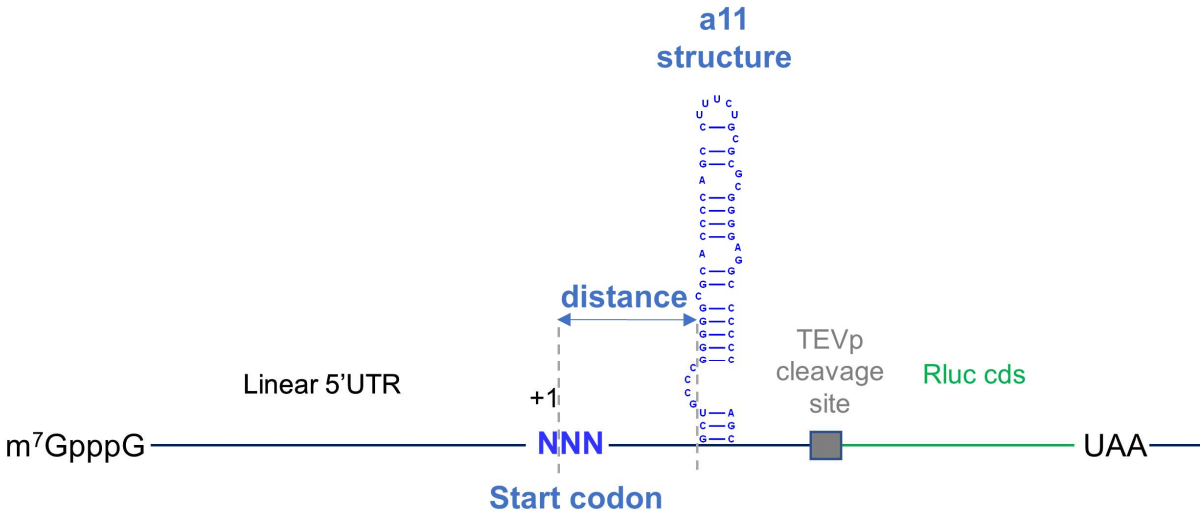


Figure 17: the optimal position of the secondary structure for translation on a CUG codon ranges from +23 to +26

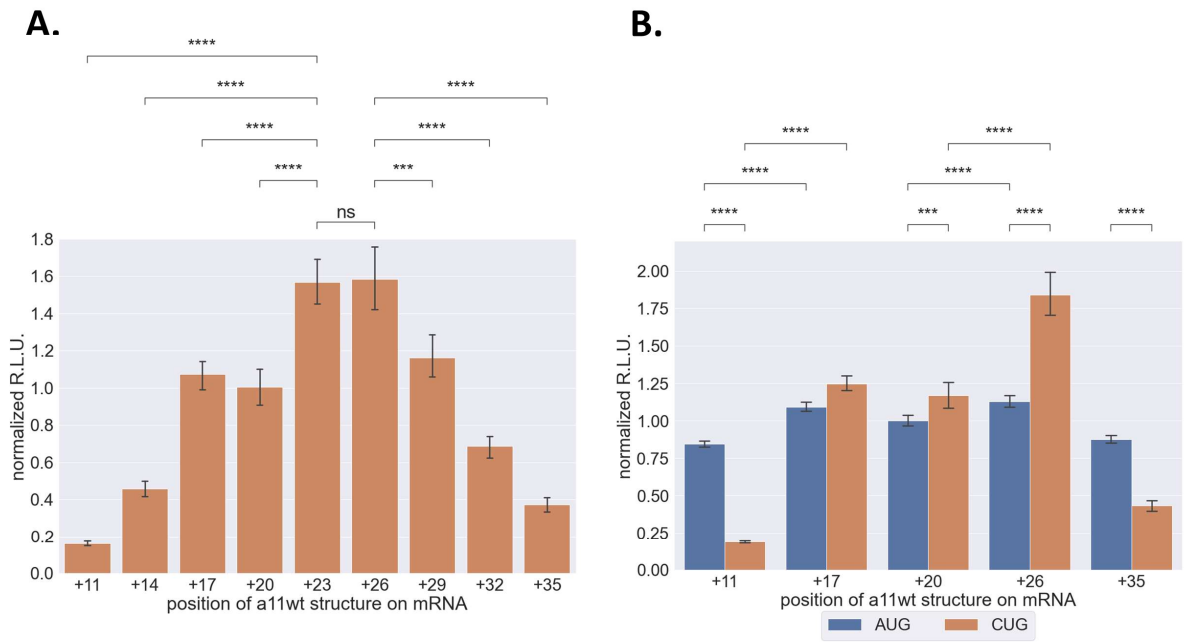
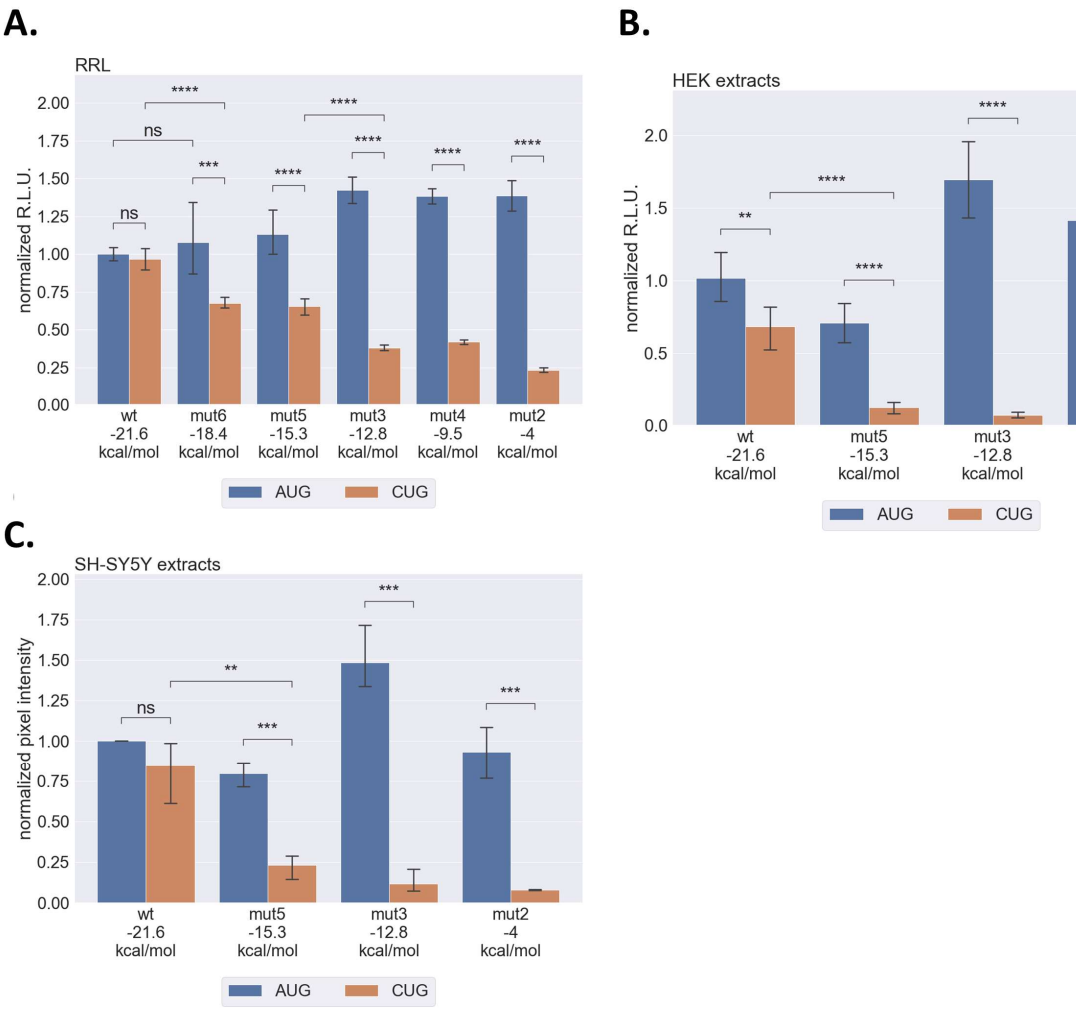


Figure 18: the optimal stability of the secondary structure is influenced by the cell type



D.

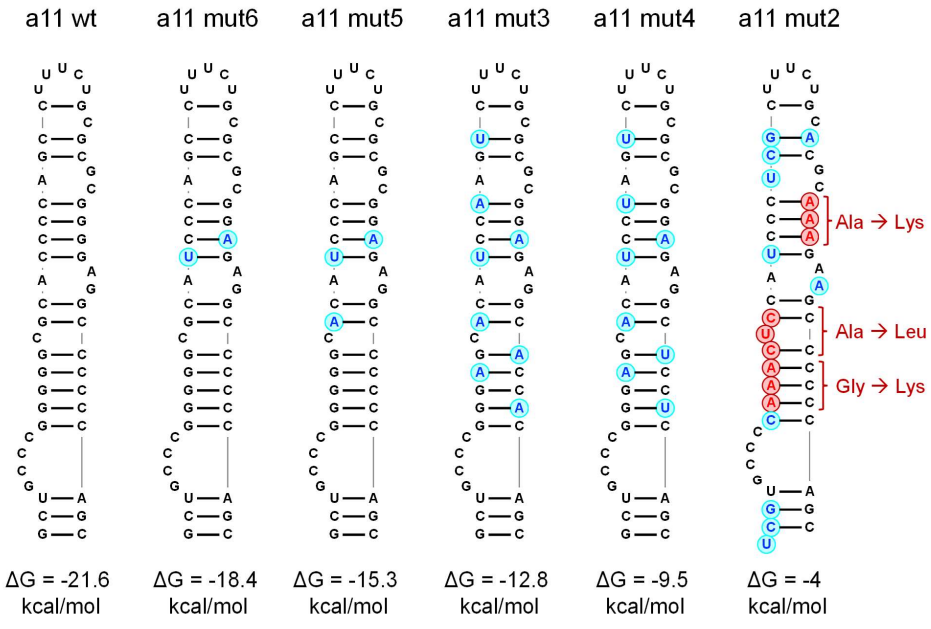
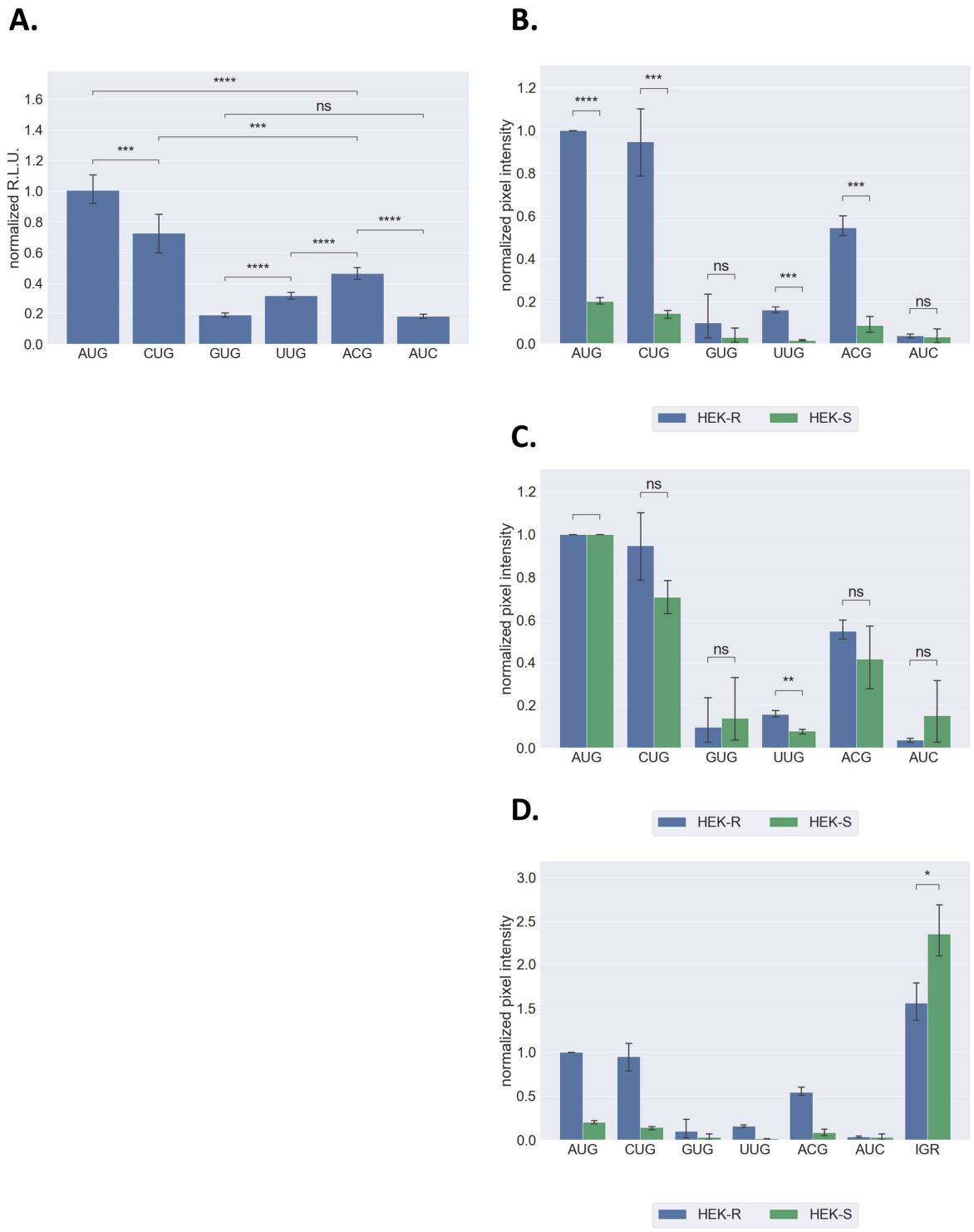
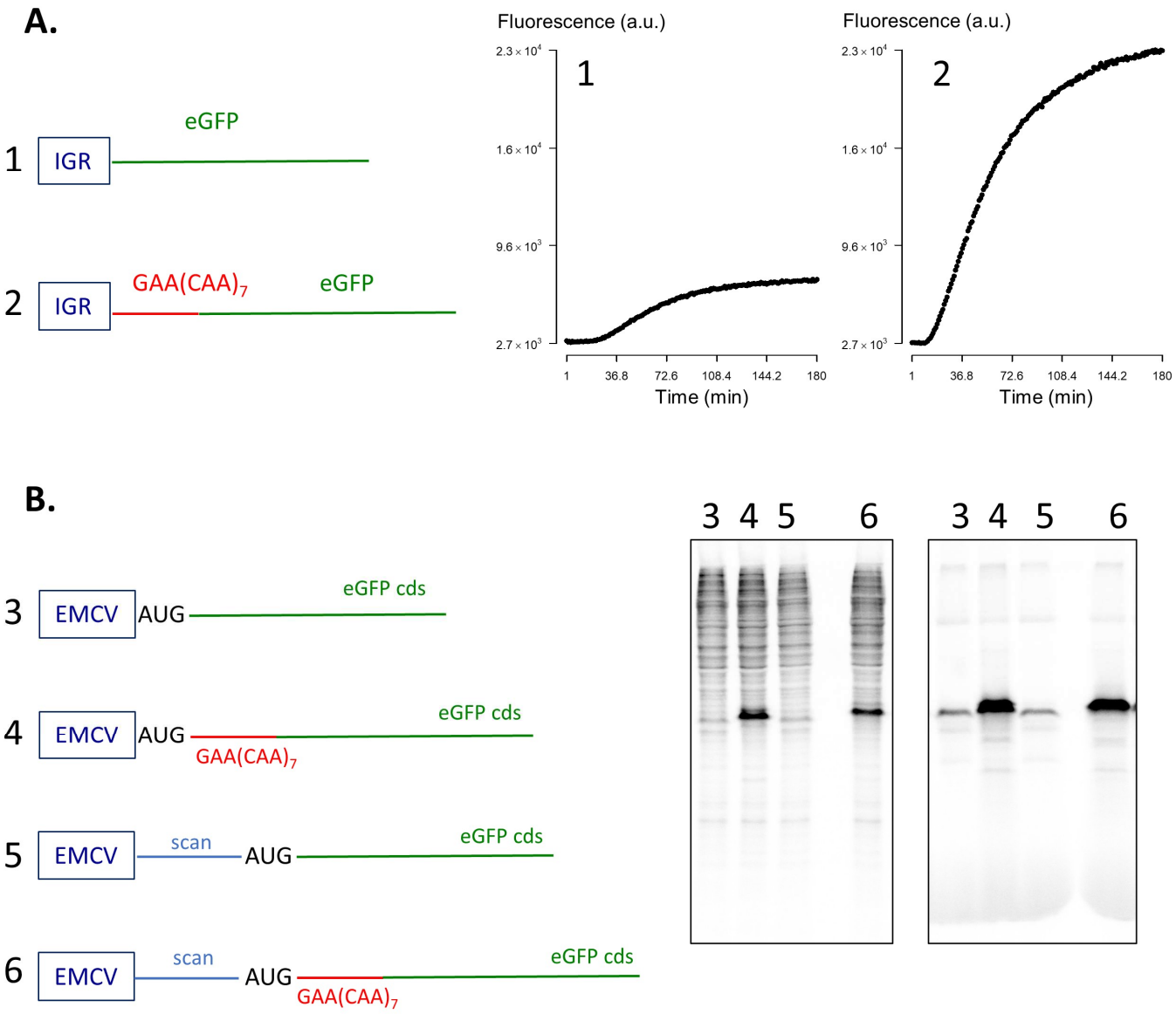


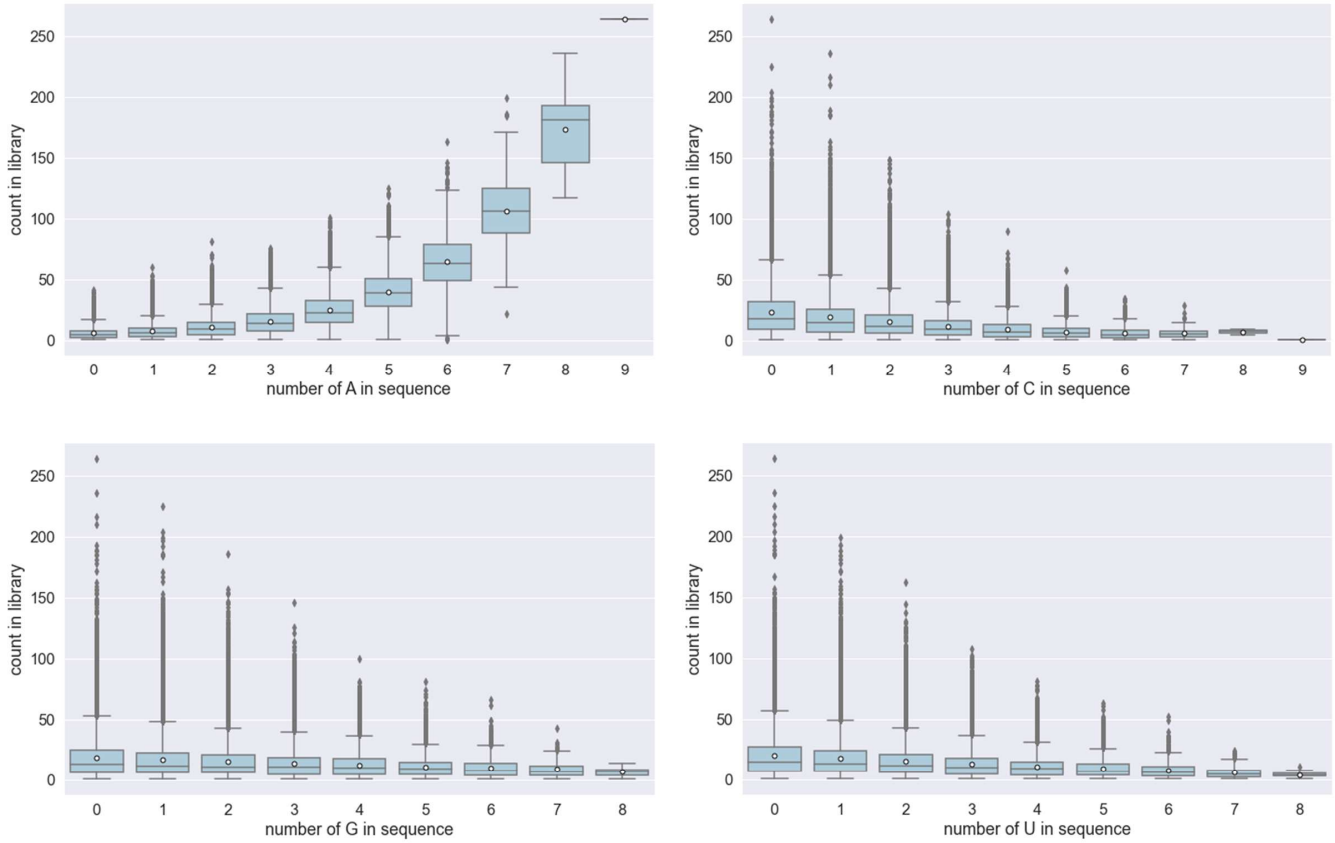
Figure 19: the START mechanism does not rescue translation initiation for all AUG-like codons



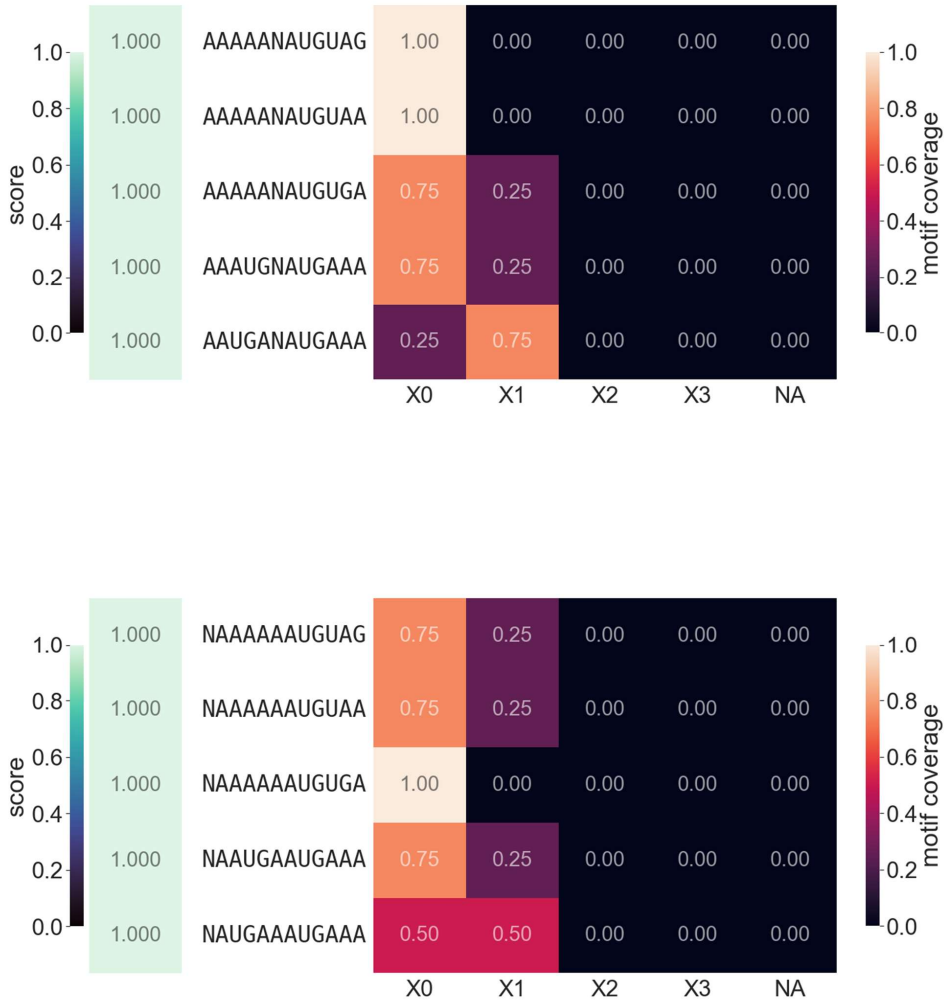
Supplementary Figure S1: the eGFP coding sequence is a strong inhibitory *cis*-element of translation initiation that is not specific to the cell-free translation extracts and 5'UTR



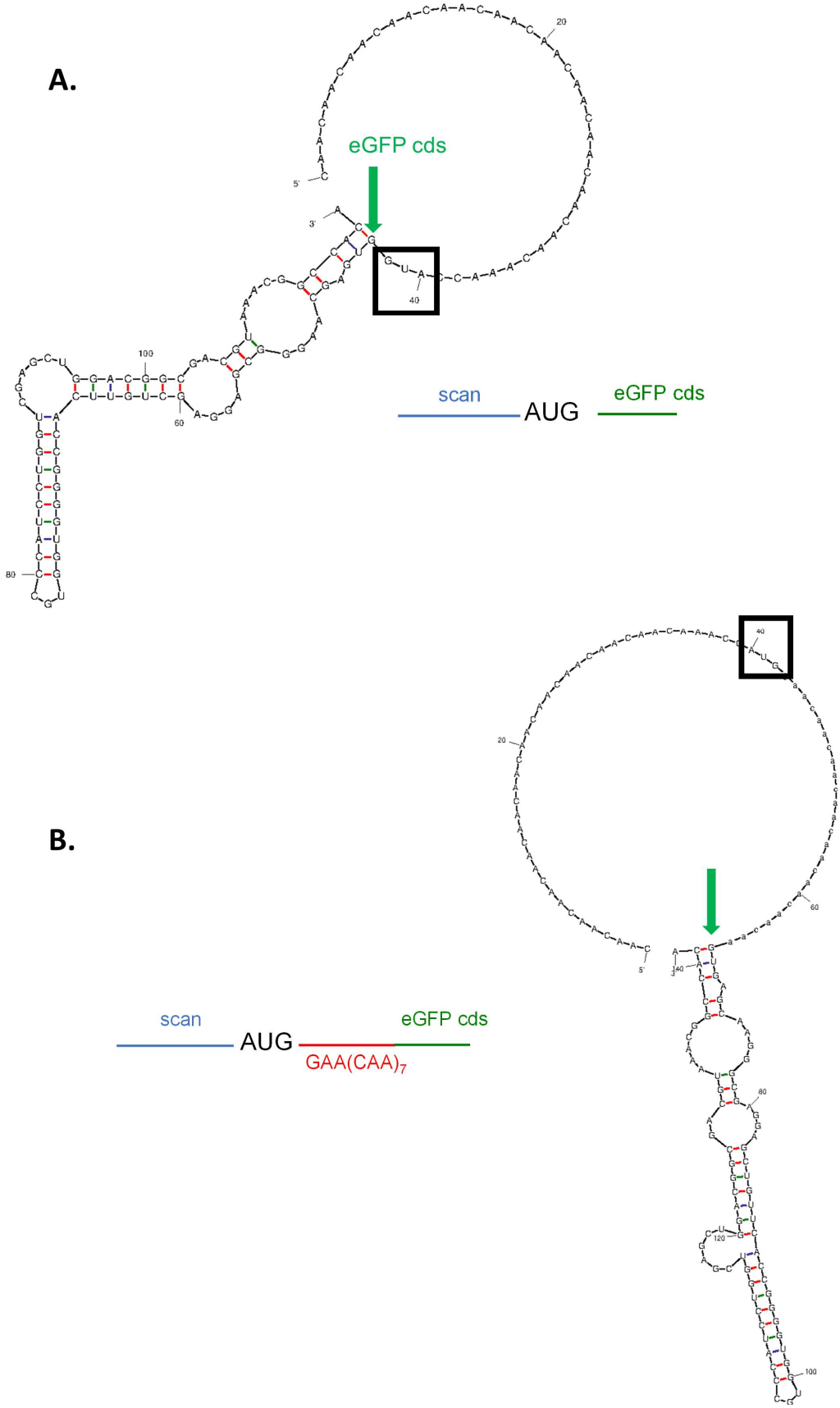
Supplementary Figure S2: only A-rich contexts are biased in the starting library



Supplementary Figure S3: an A-rich context does not rescue in frame-stop codons or out-of-frame upstream AUGs



Supplementary Figure S4: the AUG context of the NNNNNN<u>AUG</u>NNN reporter library is in a predicted structure-free environment



Supplementary Table S1: oligonucleotides' sequences

n° in text	Name	Sequence 5'-3'	Remarks
1	rev_EMCVscan-N6-ATG-N3-eGFP	CAGCTCCTCGCCCTTGCTCACTTGTTGTTGTTGTTCT(N1:25252525)(N1)(N1)CAT(N1)(N1)(N1)(N1)(N1)CCAACTAGTTGGTAGTTG	hand-made mix
2	fw_pUC19-EMCV	GTCGACCTGCAGGCATGCAACTAACGTTACTGGCCGAAGCCGCTTGG	
3	fw_T7-EMCV	CAACAAATATTAATACGACTCACTATAGGTAACGTTACTGGCCGAAGCCGCTTG	
4	rev_TagMut2_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTAATTTAAATCTCTTCTGATAATAATTTTTGTTCTAATGC	
5	fw_TAGNNNATGNNN_Rluc	AACAACAACTACCAACTAGTTGGTAG(N:25252525)(N)(N)ATG(N)(N)(N)ACTTCGAAAGTTTATGATCC	hand-made mix
6	fw_TAGAAAATGAAA_Rluc	AACAACAACTACCAACTAGTTGGTAGAAAATGAAAACTTCGAAAGTTTATGATCC	
7	fw_TAGACCATGGCG_Rluc	AACAACAACTACCAACTAGTTGGTAGACCATGGCGACTTCGAAAGTTTATGATCC	
8	fw_AAAAAATGAAA_Rluc	AACAACAACTACCAACTAGTTGGAAAAAAATGAAAACTTCGAAAGTTTATGATCC	
9	fw-TAGTNNATGNNN	AACAACAACTACCAACTAGTTGGTAGT(N:25252525)(N)ATG(N)(N)(N)ACTTCGAAAGTTTATGATCC	hand-made mix
10	fw-NDDRDDATGNNN	AACAACAACTACCAACTAGTTGGNDDRDDATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
11	fw-NDDRDDATGRNN	AACAACAACTACCAACTAGTTGGNDDRDDATGRNNACTTCGAAAGTTTATGATCC	machine-made mix
12	fw-NDDYDDATGNNN	AACAACAACTACCAACTAGTTGGNDDYDDATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
13	fw-NDDNDDATGNNN	AACAACAACTACCAACTAGTTGGNDDNDDATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
14	fw-NCCRCCATGNNN	AACAACAACTACCAACTAGTTGGNCCRCCATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
15	fw-NCCRCCATGRNN	AACAACAACTACCAACTAGTTGGNCCRCCATGRNNACTTCGAAAGTTTATGATCC	machine-made mix
16	fw-NCCYCCATGNNN	AACAACAACTACCAACTAGTTGGNCCYCCATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
17	fw-NCCNCCATGNNN	AACAACAACTACCAACTAGTTGGNCCNCCATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
18	fw-NNNRNNATGNNN	AACAACAACTACCAACTAGTTGGNNNRNNATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
19	fw-NNNRNNATGRNN	AACAACAACTACCAACTAGTTGGNNNRNNATGRNNACTTCGAAAGTTTATGATCC	machine-made mix
20	fw-NNNYNNATGNNN	AACAACAACTACCAACTAGTTGGNNNYNNATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
21	fw-NNNNNNATGNNN	AACAACAACTACCAACTAGTTGGNNNNNNATGNNNACTTCGAAAGTTTATGATCC	machine-made mix
22	rev_Renilla_pUC19	GCATGCCTGCAGGTCGACTACTAGTTTATTGTTCATTTTTGAGAAC	
23	fw_Ilu2_START_3	CTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAATTATTTTGTAGGGTCAACAACTACCAACTAGTTGG	
24	rev_Ilu1-5'eGFP_74	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCGCCCTTGCTCACTTG	
25	fw_T7-RNA-158_+16	CAACAAATATTAATACGACTCACTATAGGGAAAAACACGACCATAATC	
26	fw_T7-IGR	CAACAAATATTAATACGACTCACTATAGGCAAAAATGTGATCTTGCTT	
27	fw_T7_5'UTR-beta-globin	CAACAAATATTAATACGACTCACTATAGGACATTTGCTTCTGACACAACTGTGTTCACTAGC	

2.1.3. Perspectives

La suite de ce travail consistera dans un premier temps à finaliser l'étude de l'influence du mécanisme START dans la traduction des petites ORFs (uORFs et dORFs) identifiées par l'étude de Chothani *et al.*, dont près de la moitié est traduite à partir de codons non-AUG. Si des structures secondaires suffisamment stables sont effectivement prédites en aval des codons d'initiation de ces ORFs, nous chercherons à confirmer l'effet de ces nouvelles structures sur l'initiation de la traduction par des expériences de traduction *in vitro* à l'aide d'extraits acellulaires. Ces études permettront d'affiner notre outil bio-informatique pour la prédiction de nouvelles phases codantes dans les génomes eucaryotes qui ne démarrent pas par un codon AUG. Ensuite, l'étude des implications biologiques des ARNm qui contiennent ces ORFs par les « GO terms » associés permettra peut-être de définir des conditions physiologiques particulières qui sont favorables à l'activation ou à l'inhibition de la traduction de ces petites ORFs.

Concernant l'étude du contexte nucléotidique de l'AUG, il s'agira d'approfondir les résultats de l'étude de l'ORFeome humain. Il serait intéressant d'étudier si la présence d'un motif particulier corrèle avec une famille d'ARNm particulière qui peut être associée à un ou plusieurs processus biologique(s). Cela concerne par exemple les 411 transcrits portant le contexte le plus représenté, AGCACCAUGUCG. La même opération pourra être réalisée sur les autres contextes relativement abondants que nous avons caractérisés. Par ailleurs, près de 80% des contextes ne présentant pas de codons stops en phase ou d'uAUG hors phase dans la région NNNNNNAUGNNN sont absents de l'ORFeome humain. Il serait intéressant de déterminer si ces séquences partagent des motifs minimaux, ou si, à l'instar des contextes sélectionnés lors des expériences de criblages, chaque contexte nucléotidique trouvé parmi ces 80% exerce en effet unique sur l'initiation de la traduction.

2.2. Recherche systématique d'IRES dans un génome viral

2.2.1. Introduction du projet

Cette partie est dédiée à la mise au point d'une méthode de criblage permettant d'identifier de manière systématique des IRES dans un génome viral.

Cette méthode est basée sur le fractionnement à l'aide de gradients de sucrose de complexes de traduction assemblés sur une banque d'ARN rapporteurs contenant des fragments d'ARN génomique viral fusionnés à la phase codante d'un gène rapporteur. Les fractions contenant les ribosomes en traduction sont collectées et le fragment d'ARN viral est identifié par séquençage Illumina.

Dans ce travail, une expérience de preuve de concept a été réalisée en utilisant le génome du virus de la paralysie du cricket (CrPV), qui contient deux IRES bien caractérisés dans la littérature. Le premier se trouve dans la région 5'UTR de l'ARN viral et a été caractérisé par notre équipe comme un IRES de la classe III (Gross *et al.* 2017). Le second se situe dans la région intergénique de l'ARN viral et a été caractérisé par de nombreux laboratoires comme un IRES de la classe IV. Les criblages réalisés à partir de ce génome viral devraient permettre de détecter ces deux IRES et de ce fait valider la stratégie expérimentale.

Les résultats montrent qu'il est possible d'identifier les deux IRES du CrPV avec cette méthode. Cependant, d'autres régions candidates ont été sélectionnées à l'issue des criblages. Par manque de temps, nous n'avons pas encore déterminé la pertinence de ces régions pour l'initiation de la traduction mais il n'est pas à exclure qu'il puisse s'agir d'artefacts. Par conséquent, des mises au point supplémentaires sont nécessaires pour améliorer la sélectivité de la méthode de criblage afin de réduire la sélection de faux positifs.

2.2.2. Identification d'IRES viraux à l'aide d'une méthode de criblage basée sur le fractionnement de complexes de traduction sur gradient de saccharose

Les résultats sont présentés sous la forme d'un manuscrit qui rend compte de l'état actuel des travaux et des pistes d'amélioration de la méthode de criblage qui seront explorées à l'avenir.

MANUSCRIT 2

Identification d'IRES viraux à l'aide d'une méthode de criblage basée sur le fractionnement de complexes de traduction sur gradient de saccharose

Systematic identification of viral IRES with a sucrose gradient-based screening strategy

Antonin TIDU, Gilbert ERIANI, Franck MARTIN*

Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, *Architecture et Réactivité de l'ARN*, CNRS UPR9002, 2, allée Konrad Roentgen, F-67084 Strasbourg (France)

*Corresponding authors

Email for correspondence:

f.martin@ibmc-cnrs.unistra.fr

Running title: Genome-wide IRES screening

Keywords: Internal Ribosome Entry Site, Ribosomes, Translation

Abstract

Viruses have evolved various strategies to hijack the host cellular translation machinery to produce their essential viral proteins required for functional viral particles synthesis. In positive RNA viruses, the major mechanism used is based on viral RNA structures that can directly recruit cellular ribosomes in a cap-independent manner. Such a translation initiation mechanism is called Internal Ribosome Entry Site (IRES). Here we propose a new method to systematically screen viral genomes for novel uncharacterized IRES elements. This strategy uses a sucrose gradient-based purification method of *in vitro* assembled translation complexes programmed with a viral genome RNA reporter library. The viral library is generated by random fragmentation of a plasmid containing a viral genome. The selection is performed with cell-free translation extracts prepared from various cell types. As a proof-of-concept experiment, we screened the Cricket Paralysis Virus (CrPV) genome, which contains two well-characterized IRES elements. The results enabled the identify the two CrPV IRES along with other CrPV regions that are potentially involved in translation regulation.

1. Introduction

Viruses have evolved various strategies to hijack the host cell translation machinery for the synthesis of functional viral proteins. These strategies rely on the interactions of viral proteins with either translation factors to force viral RNA translation or stressed-induced effectors to inhibit the cellular response to the infection (Walsh and Mohr 2011). Another set of strategies is based on viral RNA structures that can recruit the translational machinery by interacting directly with the ribosomal subunits or indirectly by recruiting translation factors (Jaafar and Kieft 2019). These viral structures are called Internal Ribosome Entry Site (IRES). Viral IRES have been classified into four categories depending on the IRES structural complexity and the translation initiation factors required to efficiently initiate translation (Lee *et al.* 2017; Mailliot and Martin 2018; Jaafar and Kieft 2019). While class I and class II IRES, like those found in Polioviruses (PV) and Encephalomyocarditis viruses (EMCV), require almost all the eucaryotic initiation factors (eIF) except the cap binding protein eIF4E, class III IRES, like the Hepatitis C virus (HCV) IRES, require a limited set of eIF. Finally, class IV IRES, like the Cricket Paralysis Virus (CrPV) IGR (Intergenic Region) don't require any eIF as they directly recruit an elongation 80S ribosome. IRES are often found in positive single stranded RNA viruses such as flaviviruses, coronaviruses and many others. Most of these viruses are emerging threats for human health. Identification of novel uncharacterized IRES in emerging viral genomes would provide critical knowledge to better understand the viral cycles. Moreover, IRES elements are targets of choice to develop antiviral drugs.

Dedicated methods to systematically identify novel IRES in viral genomes have been developed by FACS (Fluorescence Activated Cell Sorting)-based screenings of cells transfected with bicistronic reporters libraries (Weingarten-Gabbay *et al.* 2016). In this strategy, the screening was limited to 200-nt long fragments. The fragments are chosen based on educated guesses. Finally, the fact that the selections were made in cells didn't allow to screen for translation inhibitors that are potentially toxic for living cells. Here we developed a new unbiased approach to systematically screen for viral IRES starting from a complete viral genome. This strategy is based on sucrose gradient-based purification of *in vitro* translation complexes assembled with an RNA reporter library. We generated the viral library by random fragmentation of a plasmid containing a full-length viral genome. We used cell-free translation extracts prepared from various cell types to assemble translation complexes on the viral RNA fragments. The RNA fragments that are contained into 80S ribosomes are purified, reverse-transcribed, and sequenced. As a proof-of-concept, we used the Cricket Paralysis Virus (CrPV) genome that contains two previously characterized IRES. The IRES1 is located in the 5'UTR of CrPV (Gross *et al.*, 2017) and the IRES2 (also called IGR for Intergenic Region) is located in the middle of the genome. Our method enabled to identify the two CrPV IRES along with other CrPV regions that are putatively functionally relevant for the virus.

2. Material and methods

2.1. Oligonucleotides information

All the oligonucleotides' sequences used in this study are provided in Table S1. In the text, oligonucleotides are referred to by their red numbers in the Table S1.

2.2. Reporter library preparation from a fragmented viral genome

TA-plasmid construction

We synthetized a custom TA-cloning plasmid using the NEBuilder® HiFi DNA Assembly Cloning Kit (New England Biolabs #E5520S). It is a pUC19-based plasmid in which we cloned the following cassette into the multiple cloning site: T7 promoter – hairpin sequence – flexible linker – AhdI restriction site – NotI restriction site - AhdI restriction site – Renilla luciferase coding sequence. The variable parts of the AhdI restriction sites were designed in such way that they generate 5'-T overhangs upon restriction digest, and thus compatible for TA-ligation with 3'-A-tailed DNA fragments. In order to have a unique AhdI restriction site, the one that is present in the AmpR gene was inactivated by introducing a silent mutation with the NEB Q5® Site Directed Mutagenesis Kit (New England Biolabs #E0554S). To generate 5'-T overhangs, the plasmid was digested for 3h at 37°C in the following conditions: 50 ng/µL plasmid, 1X Cutsmart buffer (New England Biolabs), 0.5 U/µL AhdI (New England Biolabs). After the reaction, the digested plasmid was dephosphorylated two times 1h at 65°C upon addition of 150 units of Bacterial Alkaline Phosphatase (Invitrogen).

Fragmentation of the CrPV viral genome

3 µg of pCrPV-3 plasmid (Khong *et al.* 2016) containing the Cricket Paralysis Virus (CrPV) genome was fragmented for 12.5 min at 37°C using the NEBNext dsDNA Fragmentase v2 kit (New England Biolabs #M0348) in the recommended conditions without addition of $MgCl_2$. The fragmentation reaction was stopped by addition of 0.1 mM EDTA. 2.5 µg of resulting fragments were purified using Solid Phase Reverse Immobilization beads (SPRI, Beckman) with a beads/DNA ratio of 0.55 in a left-side selection protocol (see Beckman documentation). This procedure allows to concentrate fragments whose size is higher than 400 bp. The selected fragments were ethanol precipitated with 250 mM NaCl and resulting DNA pellets were resuspended in 20 µL milli-Q water. We estimated the DNA fragment concentration at this step being close to 100 ng/µL with 260 nm absorbance measurement with a Nanodrop. Finally, we used the NEBNext® Ultra End Repair/dA-tailing kit (#E7442) to generate blunt- and 3'-A-tailed-fragments for subsequent TA ligation into the custom TA cloning plasmid.

TA ligation and PCR amplification

TA ligation was performed using the NEB Blunt/TA Ligase Master Mix (New England Biolabs #M0367) in the conditions recommended by the manufacturer, except that the reaction was incubated 30 min at 25°C. 2 µL of the ligation mixture was then used as a template for PCR amplification of the reporter library using 0.25 mM dNTP, 3% DMSO, 1X GC-buffer (NEB), 0.2 µM primers n°1 and 2 or n°1 and 3 (Table S1) and 0.01 µg/µL recombinant Phusion DNA polymerase in a 100 µL reaction. The PCR products were SPRI purified using a left-side selection protocol with a beads/DNA volume ratio of 0.75 that allows the removing of fragments whose size is smaller than 300 bp. The PCR products were eluted in 35 µL milli-Q water.

In vitro transcription

Messenger RNA reporters are synthetized in a 100 µL *in vitro* transcription reaction using 30 µL of the SPRI-purified PCR products, 5 mM Tris-HCl pH 8, 30 mM $MgCl_2$, 1 mM spermidine, 5 mM DTT, 0.01% Triton X-100, 5 mM of each ribonucleotide (ATP, CTP, GTP, UTP pH 7.5), 0.5 U/µL RNase inhibitor (Promega) and 0.125 mg/mL of recombinant T7 RNA polymerase in a 100 µL-reaction. The reaction is incubated 1h at 37°C. After 1h incubation, 0.02 mg/mL pyrophosphatase (Merck) is added and after 30 min and the DNA template is removed by the DNA template is removed by 1h incubation with 0.2U/µL DNaseI (Roche). The reaction products are passed through a 1mL-G-25 Superfine Sephadex column (Cytiva), phenol-extracted and ethanol-precipitated. The resulting RNA pellets are resuspended in 30 µL milli-Q water and quantified by absorbance measurement at 260 nm with a Nanodrop.

To synthetize the RNA reporters used in 3.2 and 3.3, a T7-UTR-Renilla construct was PCR-amplified using forward primers n°25 to 29 with reverse primers n°2 (Table S1) for *in vitro* translation assays (see 2.4) or n°3 (Table S1) for radiolabeling (see 2.9). Then, 20µL of the PCR product are used for *in vitro* transcription.

2.3. *In vitro* translation extracts preparation

2.3.1. Rabbit reticulocytes lysates

Untreated rabbit reticulocytes lysates (RRL) were prepared as previously described (Pelham and Jackson 1976).

2.3.2. S2 and Aag2 cells

S2 (embryonic *Drosophila melanogaster*) and Aag2 (embryonic *Aedes aegypti*) cells were seeded at a concentration of 10^6 cells/mL and grown for 4 days at 25°C in 30 mL of Schneider medium (Biowest) supplemented with 10% of inactivated fetal bovine serum (Biowest), 1% GlutaMAX (Life Technologies) and 1% penicillin/streptomycin (Life Technologies). On the fourth day, 20 mL of medium is added to each flask and cells were harvested on the fifth day by centrifugation at 300 g for 5 min at 4°C. The total volume of culture for S2 cells is 500 mL, the total culture surface for Aag2 cells is 1500 cm^2. For both cell types, we obtained a total number of cells of approx. 3 $x10^9$. After harvesting, cells were washed two times with a cold buffer containing 40 mM HEPES–KOH pH 8, 100 mM potassium acetate, 1 mM magnesium acetate, 1 mM DTT and resuspended at a concentration of 10^9 cells/mL in the same buffer supplemented with 1X Halt™ Protease Inhibitor Cocktail EDTA-free (Thermo Scientific™). Cell lysis was performed by nitrogen cavitation with a Cell Disruption Bomb (Parr Instrument Company) after a one-hour incubation under a pressure of 15 bars at 4°C. The lysate was cleared by centrifugation at 10,000 g at 4°C, aliquoted, flash-frozen in liquid nitrogen and stored at −80°C.

2.4. *In vitro* assembly of translation complexes

RRL: the final concentrations in the reaction are 50% RRL, 100 mM potassium acetate, 1 mM magnesium acetate, 1.5 mM of all amino acids, 0.5 U/µL RNase inhibitor (Promega) and 0.2 µM eporter RNA in a total volume of 75 µL. For the purification of 80S complexes, the reaction mixture is incubated 7 min at 30°C in the presence of 1 mg/mL cycloheximide. For the purification of elongating 80S, the reaction is first incubated 10 min at 30°C and then supplemented with either 1 mg/mL cycloheximide or 5 mM puromycin and finally incubated 5 min at 30°C.

S2- and Aag2- cell-free extracts: the final concentrations in the reaction are 20% extract (1.7 µg/µL total protein), 150 mM potassium acetate, 2 mM magnesium acetate, 1.5 mM of all amino acids, 20 mM HEPES-KOH pH 7.5, 0.5 mM spermidine, 1 mM DTT, 1 mM ATP, 0.2 mM GTP, 8 mM phospho-creatine, 0.1 µg/µL creatine phospho-kinase, 0.5 U/µL RNase inhibitor (Promega) and 0.2 µM reporter RNA in a total volume of 75µL. For purification of initiating 80S complexes, the reaction mixture is incubated 15 min at 25°C in the presence of 1 mg/mL cycloheximide. For purification of elongating 80S complexes, the reaction is first incubated 25 min at 25°C and then supplemented with either 1 mg/mL cycloheximide or 5 mM puromycin and finally incubated 5 min at 25°C.

The molar concentration of the RNA reporter library was calculated using its average size.

2.5. Translation complexes analysis on sucrose gradients

In vitro translation reactions were loaded on a linear 7-47% sucrose gradient containing 25 mM Tris-HCl pH 7.5, 100 mM or 700 mM potassium acetate, 5 mM magnesium acetate and 1 mM DTT. Gradients were centrifuged for 2.5h at 37,000 rpm with a SW41-Ti rotor (Beckman) at 4°C. 400 µL fractions were collected from the bottom to the top of the gradients. Fractions containing ribosome complexes were identified by monitoring 260 nm absorbance during fraction collection.

2.6. RNA extraction from 80S-containing fractions and library amplification for sequencing

The 80S-containing fractions were phenol extracted and ethanol precipitated in the presence of 250 mM NaCl for 100 mM potassium-gradients or without additional salts for 700 mM potassium-gradients. RNA pellets were washed with 80% ethanol, dried and resuspended in 20 µL mQ water. cDNA corresponding to the RNA of interest were generated by reverse transcription using 0.1 µM of primer n°4 (Table S1) with 1 µg of extracted RNA, 1X SSIV-RT buffer (Thermo Scientific™), 0.5 mM dNTP Mix, 5 mM DTT, 2 U/µL RNase inhibitor (Promega) and 10 U/µL Superscript® IV reverse transcriptase (Thermo Scientific™) in a 25µL-reaction. First, the amount of RNA and reverse primer are mixed, incubated for 5 min at 65°C and for 5 min on ice. Then the other compounds are added, and the reaction is incubated 15 min at 65°C. 2 µL to 10 µL of cDNA was PCR-amplified in a 100 µL reaction using 0.25 mM dNTP, 1X GC buffer (New England Biolabs), 0.2 µM of each forward primer n°5 to 14 (Table S1), 0.2 µM of each reverse primer n°15 to 24 (Table S1) and 0.01 µg/µL recombinant Phusion DNA polymerase. The amount of cDNA and PCR cycles were adjusted empirically. PCR products were purified by Solid Phase Reverse Immobilization (SPRI) beads (Beckman) and used as a template (0.4 ng/µL) for a 50 µL-PCR amplification using 5 µL each Nextera Index primers (Illumina), 1X GC buffer (New England Biolabs), 0.25 mM dNTP, 0.01 µg/µL recombinant Phusion DNA polymerase. Indexed libraries were SPRI-purified, checked on BioAnalyser and analysed on a MiSeq Instrument (Illumina) with a MiSeq V3 150 cycles reagent kit using a paired-end program (Illumina).

2.7. Data analysis

The whole data analysis strategy has been conducted using a self-made Python algorithm.

2.7.1. Fragments reconstitution from paired-end sequencing data

First, the algorithm checks if the read contains the expected 5' and 3' sequences that correspond to the first 10 nucleotides bordering the TA-cloning site in the plasmid. Only those having both a perfect match with those two sequences and an average Q-score superior to 30 are kept in the analysis. Q-score is a function of the probability of incorrect base calling. A base with a Q-score of 30 has 1 chance over 1000 to be incorrect. The recovered reads from the two paired-end sequencing files are grouped by their index and mapped on the pCrPV-3 reference sequence (Khong *et al.* 2016). The corresponding pCrPV-3 fragment sequence is reconstituted simply by retrieving the sequence between the first mapped nucleotide of read1 and the last one of read2. Finally, each fragment is attributed a weight corresponding to its observed proportion in the library.

2.7.2. Calculation of differential fragments repartition

The screening procedure is expected to make the selected fragments converge toward (respectively diverge from) specific regions of the reference genome that positively (respectively negatively) impact translation. Therefore, comparing the repartition of the sequenced fragments between the starting and the selected libraries by calculating local signal convergence enable to identify translation-impacting sequences in the reference genome. To do so, the algorithm does the following calculations.

2.7.3. Fragment Repartition Matrix

The components of this matrix are calculated from the observed proportions of all the aligned fragments in each library what we call a Fragment Repartition Matrix containing $f_{i,j}$ values: i (rows) refers to each possible fragment length, and j (columns) to each position on the reference sequence. Then, both i and j range from 1 to 12500 (length of pCrPV-3). The genome of CrPV is between 711 and 9895. $f_{i,j}$ is defined as the sum of the observed proportions of fragments of size i that cover position j on the reference genome.

2.7.4. Signal Convergence Matrix

Signal convergence is calculated between two rounds of selection simply by subtracting the Fragment Repartition Matrices corresponding to those two rounds. For simplifying visualization, only positive values will be shown in the corresponding Figures as they correspond to regions where fragments concentration has increased (*i.e.* where fragments have converged) between the selected and the starting libraries.

2.7.5. Graphical cross-correlation

Each Fragment Repartition Matrix and/or Signal Convergence Matrix can be considered as an image. Following linear scaling of the two matrices, graphical "cross-correlation" is performed by overlaying two image-matrices with two different colour themes (shades of red and shades of blue) on the same image. The resulting image contains shades of blue, red, or purple (red + blue) regions. These purple regions are cross-correlating regions between the two matrices. It has been used to cross-correlate background with the sum of background and signal to identify specific signals.

2.8. Calculation of *in vitro* translation product levels

^{35}S-signal quantification: 5 µL of *in vitro* translation reactions are analyzed by 12% SDS-PAGE. ^{35}S- translation products bands are quantified with a phosphorimager (Typhoon FLA 7000). Image analysis is conducted with ImageJ software using the resulting TIFF files.

Luciferase assay: the amount of *in vitro* synthetized Renilla luciferase was measured by injecting of 100 µL coelenterazine 0.25 µmol/mL (Synchem) onto 10 µL of *in vitro* translation reaction using a luminometer (Varioskan LUX, Thermo Scientific™). The resulting photon emission was measured for 10 seconds.

Real-time measurement of eGFP synthesis: *in vitro* synthesis of eGFP was determined by measuring fluorescence (λ_{ex} = 485nm, λ_{em} = 520nm) every minute during translation with a Mx3005P QPCR Instrument (Agilent).

2.9. Radiolabelling of *in vitro*-synthetized mRNA

In vitro synthetized RNAs were separated by 4% PAGE, electro-eluted from the acrylamide slice and ethanol precipitated. m^{7}G-capped reporters were synthesized using the ScriptCap m^{7}G Capping System (CellScript) kit with 0.25 µCi/µL of ^{32}P-α-GTP and further purified by PAGE followed by passive elution and ethanol precipitation. RNA pellets were resuspended in milli-Q water and adjusted to 50,000 cpm/µL after Cerenkov counting. For ribosome salt wash assays using radio-labelled mRNA, we used 250,000 cpm per gradient.

2.10. Prediction and free energy calculations of secondary structures

These calculations of the free energy of specific RNA sequences used in this work were conducted using the Mfold V2.3 for RNA (Zuker 2003).

3. Results

The aim of this work is to design an experimental strategy that leads to the identification of novel IRES element(s) in viral genomes. As a proof-of-concept, we used the CrPV genome that contains two IRES that have been previously characterized (Wilson *et al.* 2000; Gross *et al.* 2017). The IRES1 is located in the 5'UTR of CrPV and the IGR is located in the intergenic region.

A detailed overview of the experimental strategy is depicted in Figure 1. Our approach is dedicated to the identification of *cis*-acting elements such as an IRES that allows cap-independent translation initiation. Briefly, a plasmid containing the whole CrPV genome is randomly fragmented and each resulting dsDNA fragment is inserted upstream the Renilla luciferase coding sequence. In order to avoid leaderless mRNA translation that would lead to false-positive selections, we inserted a stable hairpin at the 5' end of the reporter mRNA library followed by a flexible linker. This will also prevent uncapped-translation events (see 3.2.). The generated DNA reporter library is *in vitro* transcribed from a T7 promoter by recombinant T7 RNA polymerase. Next, translation complexes are assembled on the resulting mRNA reporter library using either RRL or S2 cell-free extracts and separated on a sucrose density gradient. RNA extraction is performed on 80S-containing fractions and the CrPV fragments thus selected are reverse-transcribed and sequenced by Illumina sequencing.

We used two variations of this approach, with the aim of purifiying either initiating ribosomes that have been blocked with cycloheximide on the initiation site, or elongating ribosomes.

3.1. Validation of the insect cell-free extracts used for the screening

First, we investigated the optimal salt conditions for both CrPV IRES *in vitro* translation with the cell-free translation extracts prepared from S2 (Figure 2B and Figure S1) and Aag2 cells (Figure S1). Both extracts allow the efficient translation of mRNA reporters containing the two IRES of CrPV. Interestingly, IGR translation efficiency is highly influenced by the magnesium concentration (Figure 2B).

3.2. Reduction of cap-independent translation with a stable 5' hairpin

Although translation of capped 5'UTR β-globin reporter mRNA is significantly higher, *in vitro* translation with RRL showed that even an uncapped mRNA can be translated efficiently most probably because it is unstructured (Figure 3B, β-globin reporter). This is problematic in a screen for novel IRES because all the mRNA tested are uncapped and might lead to false positive selections. Therefore, we inserted a stable secondary structure (predicted stability of -22.5 kcal/mol) at the 5' extremity of the reporter mRNA library to alleviate residual cap-independent translation (Figure 3A). This structure is followed by a flexible linker to prevent interference of the inserted hairpin with IRES-candidates folding. When this construction is inserted upstream the 5'UTR of β-globin the translation is indeed drastically impaired as expected. In contrast, the same sequence upstream the IGR of CrPV does not affect translation efficiency (Figure 3B). Altogether, the insertion of this hairpin prevents efficiently residual cap-independent translation without affecting IRES-mediated translation. This indicates that the used flexible linker does not interfere with the IRES folding and is therefore suitable for the library construction.

3.3. Determination of the optimal potassium concentration for ribosome salt-wash

The first tests of our strategy suggested a need for reduction of false-positive selection due to unspecific binding to the ribosome (Figure S2). Indeed, another source of false-positive selection may be due to unspecific interactions with ribosomal components (ribosomal proteins or rRNA) in a translation independent manner. To reduce as much as possible these unspecific interactions, we investigated high potassium concentrations that still maintain the integrity of translation-related 80S complexes while interfering with unspecific interactions. We assembled translation initiation complexes with RRL in the presence of cycloheximide or puromycin and monitored 80S formation on sucrose gradients with ^{32}P-m^{7}G-radiolabelled 5'UTR β-globin mRNA in the presence of increasing potassium acetate concentrations (Figure 4). The peak of 80S in the presence of cycloheximide contains translation-competent ribosome along with ribosomal complexes that interact unspecifically with any RNA. In the presence of puromycin, the peak of 80S only contains 80S-mRNA complexes formed by unspecific interactions. The amount of 80S containing the radiolabelled reporter mRNA is identical from 100 mM to 700 mM potassium acetate in the presence of cycloheximide, indicating that translation is still equally efficient with high potassium acetate concentrations. 80S dissociation into 60S and 40S subunits is however observed at 1M potassium acetate. In the presence of puromycin, the unspecific ribosomal complexes are significantly decreased with higher potassium acetate concentrations. Therefore, we chose 700 mM potassium for screening in ribosome salt wash conditions.

3.4. Sequencing of the starting library reveals that it does not follow a uniform distribution

Our experimental strategy was to select either initiating ribosomal complexes or elongating ribosomal complexes. Therefore, we designed two specific libraries. For initiating ribosomal complexes, we used a short reporter mRNA that contains a minimal coding sequence that contains only the 12 N-terminal codons of Renilla coding sequence followed by a stop codon (Figures 5-6A). The length of this minimal coding sequence (36 nt) is compatible with the assembly of an initiating ribosome on the AUG start codon. For the assembly of elongating ribosomal complexes, we used the whole coding sequence of Renilla (933 nt) (Figure 6A). Prior screening, we sequenced the two starting libraries to assess the coverage of the viral genome. Both libraries cover the whole viral genome, however we observed a non-uniform distribution of the fragments (Figures 5-6B). The non-uniformity of the distribution is even more pronounced with the library featuring the whole Renilla coding sequence, thus highlighting a link between the fragment distribution bias with the length of the reporter used for each library construction. Since all the steps for library constructions are identical, we suspected that this difference is due to PCR amplification efficiencies related to the length of the reporter mRNAs. Consequently, there are regions of the CrPV genome that are over- and under-represented in the starting libraries. IRES1 is located in an over-represented region whereas IGR is located in under-represented region (Figures 5-6A). The lengths of the fragments are from the expected size (200 to 400 nt) and there is no strong PCR-amplification bias for smaller fragments (Figures 5-6C). We found that 75% of the fragments of both libraries have a length shorter than 800-900bp.

3.5. Sequencing of CrPV fragments contained in initiating 80S

To isolate CrPV fragments contained in initiating 80S ribosomes, we assembled translation complexes on the CrPV library containing the short coding sequence of 36 nucleotides using

either RRL or S2 cell-free extracts in the presence of 1 mg/mL cycloheximide to block the translocation and immobilize the initiating ribosome on the AUG start codon. The assembled 80S complexes were purified by sucrose gradients in the presence of 700 mM potassium acetate and further analysed by RNA extraction. Upon cDNA synthesis, the selected libraries were either processed for another round of selection and for sequencing (Figure 1-FG).

Raw alignments of the recovered fragments from both RRL and S2 screenings are shown in Figure S3. Convergence of fragments towards IRES1 region is observed after one round of selection (R1) for both extracts. We then calculated the Fragments Repartition Matrix after each round of selection from the reconstituted CrPV fragments and calculated signal convergence between the selected and starting libraries (Material and Methods). After one round of selection with both types of extracts, signals do not specifically converge towards the two IRES regions. Many other regions are also subjected to signal convergence (Figure 7-A-C). The following rounds of selection did not make the signals further converge towards the IRES regions, but instead towards smaller fragments (Figure 7-B-D). In RRL however, IGR fragments are among the most enriched (Figure 7B). This preference for smaller fragments suggests that they have a selection advantage. This could be due to two reasons. First, small fragments are preferred substrates for PCR amplification and *in vitro* transcription between each round of selection. Second, the propensity of those small fragments to efficiently bind non-specifically to the 80S ribosome is higher. That could especially be the case for fragments containing truncated inactive IRES that kept enough residues that are still able to establish specific contacts with the ribosome. Indeed, such fragments are strikingly featured in the R2 selected libraries for each cell extract. (Figure 7-B-D between dashed lines). Overall, CrPV fragments contained in initiating 80S ribosomes feature the two IRES regions, however along with many other fragments covering almost all positions of the genome. At this stage, our screening procedure is not selective enough for applying this method to *de novo* IRES discovery from an unknown viral genome. Therefore, it is necessary to adapt the experimental strategy to improve the selectivity of the screening procedure.

3.6. Selection of fragments contained in elongating 80S

So far, our results clearly demonstrated the need to distinguish translation-related and from translation-unrelated fragments. Using high 700 mM potassium acetate was not sufficient to eliminate completely unspecific or partially unspecific RNA fragments like inactive truncated fragments of IRES from our selections. To further improve the selections of translationally competent complexes, we used the same approach but this time focusing on RNA fragments contained in both 80S and polysomes fractions that have been selected from *in vitro* translation reactions performed in the absence of cycloheximide. In order to identify RNA fragments that bind non-specifically to the ribosomes, we used 5 mM puromycin in a control selection. Fragments that are still present in the presence of puromycin are necessarily non-specific binders while fragments that are depleted by puromycin are mRNA that are engaged in elongating ribosomes. Therefore, the fragments that are still present after puromycin treatment are considered as background fragments for data analysis. In another set of selections, we added 1 mg/mL of cycloheximide after 10 minutes of incubation to immobilize 80S ribosomes on the translated mRNAs. Altogether, we performed three parallel selections of the CrPV library, without inhibitor, with 1 mg/mL cycloheximide after 10 min and with addition of 5 mM puromycin. As previously, the ribosomal complexes were purified by sucrose gradient fractionation.

We conducted two consecutive rounds of selection termed R1 and R2 with RRL and S2 cell-free extracts on the CrPV starting library, which contains the whole Renilla luciferase coding sequence. The fact that the whole coding sequence for Renilla luciferase is present in our reporter mRNA enabled us to monitor global translation by measuring luciferase activity after each round of selection (Figure 8). While the starting library does not promote efficient Renilla luciferase synthesis, global translation from the RNA after selection produce Renilla luciferase indicating an enrichment for translation-prone RNA fragments. This assay was used to assess the translation level of different fractions of the gradients. Polysome fractions are poorly enriched in such fragments, mainly because polysomes are far less abundant than monosomes as indicated by absorbance measurements at 260 nm during fractions collection (Figure S4). On the contrary, monosomes fractions are highly enriched and do promote efficient translation. Both protocols led to an increase of translationally active fragments after one selection round (R1) however addition of cycloheximide is slightly less efficient than without inhibitor. We also observed that 80S fractions contain more active RNA fragments than polysomal fractions. Altogether, we used the most efficient protocol without any inhibitor and focused our analysis on the fragments that have been extracted from the monosomes fractions. Fragments selected after two rounds of selection are also globally active, although less than those selected after one round (Figure S5). This might be due to a loss of fragments during sample processing between two consecutive rounds (Figure 1).

Raw alignments of the recovered fragments are shown in Figure S6. Compared to the previous approach using initiating 80S, the signal convergence analysis of the fragments extracted from monosomes points more efficiently to both IRES simultaneously with fewer other candidates with both extracts after two rounds of selection (Figures 9-10 A-B). The RNA fragments obtained in the presence of puromycin correspond to unspecific binders that are considered as background (Figures 9-10 A-B). Combining on one hand signal convergence between the selected fragments and the starting library and on the other hand cross correlation between the selected fragments in the selections with and without puromycin leaves several regions as IRES candidates, including the two containing the CrPV IRES. With IRES1 and IGR, smaller fragments corresponding to truncated forms of the IRES are cross-correlating with background signals (purple) and/or lost (blue) (Figures 9-10 A-B). In other words, with IRES1 and IGR regions, we observe two sets of smaller fragments that are not enriched. We therefore expect to observe the same kind of pattern with other regions which would have an IRES feature. This reasoning enables to point three additional regions that may contain RNA fragments of interest (Table 1).

4. Discussion

This work indicates that our sucrose gradient-based screening procedure can lead to the identification of IRES elements in a viral genome using various cell-free translation extracts. We used our strategy to screen the CrPV genome, which contains two well-characterized IRES. This experiment was performed using either RRL, a broadly used *in vitro* translation system, or S2 cell-free extracts, a more biologically relevant *in vitro* translation system because Drosophila is infected by CrPV (Manousis and Moore 1987; Nayak *et al.* 2010). The selected CrPV fragments isolated from either initiating or elongating 80S ribosomes both led to the identification of the two IRES as expected but with huge differences in sensitivity and resolution. This is at least partially due to highly variable coverage of the genome in the starting library. This is especially obvious with the IGR, for which its low abundancy in the starting library has a huge influence on its detectability after selection. Therefore, an important perspective will be to alleviate the coverage biases of the starting library.

Apart from the two IRES, screening with initiating 80S also reveals other fragments in the CrPV genome. This is problematic for assigning precisely IRES in a viral genome to narrow down the selection to specific regions that are important for translation. Therefore, we decided to isolate RNA fragments from elongating ribosomes. We suspect that such fragments can trap ribosomes without initiating translation thanks to putative secondary structures formed after random fragmentation of the CrPV genome. To discriminate translation-related from translated-unrelated fragments, namely unspecific ribosome binders, we used puromycin in a control selection. RNA fragments that are selected in the presence of puromycin are necessarily unspecific binders and are considered as background. This strategy was efficient to select specifically the two IRES although three non-IRES regions are also selected (Table 1). In other words, these RNA elements are located upstream of some viral ORFs. Therefore, these fragments might be real translation-activators that could also trigger translation of several viral proteins without the need of the two previously described IRES. This possibility requires further investigations although previous ribosome profiling experiments realized on S2 cells infected by CrPV strongly suggest that there are only two IRES in the CrPV genome (Khong *et al.* 2016). However, these regions require further experimental investigation to classify them as functional translational elements.

Concerning the selection of the two IRES, they are not found by our selection procedure with the same efficiency. IRES1 is easily found by both screening procedure using initiating or elongating ribosomes. One likely explanation is that the abundance of fragments containing IRES1 in the starting library is high. Another selective advantage of this IRES is linked to its molecular mechanism used for ribosome recruitment. IRES1 recruits the ribosome on its Domain III but the ribosome can efficiently initiate translation in a window of approximately 30 nucleotides downstream the Domain III (Gross *et al.* 2017). Therefore, fragments containing the entire IRES1 will be selected plus additional fragments that contain the IRES1 plus ~30 nucleotides downstream. In addition, the fragments in the three frames will be equally selected since the recruited ribosome is able to perform a 'local' scanning (Gross *et al.*, 2017). On the contrary, IGR requires further data analysis to be efficiently identified and that motivated the "Fragment Repartition Matrix" approach, with the aim of extracting signals from noise. The first reason for this low score is that IGR is very low abundant in the starting library. Moreover, the recruited ribosome requires a pseudo-knot that mimics a codon-anticodon interaction, which positions the next codon in the A-site of the ribosome (Pestova and Hellen 2003). A single deletion and most of non-Watson-Crick mutations in the codon-anticodon mimicking domain

impairs translation (Pernod *et al.* 2020). The consequence of this mechanism is that the recruited ribosome by IGR is in fact an elongating ribosome on the IRES itself. Consequently, the frame is determined by the IGR pseudoknot after ribosome recruitment. This implies that only one third of the fragments linked to the IGR will be in frame with the luciferase frame. Consequently, two third of the RNA fragments that contain the IGR will not be selected because there are out-of-frame with the Renilla luciferase coding sequence. In conclusion, the 3' end of the IGR present in the selected RNA fragments is critical in contrast to IRES1. Altogether, these features drastically reduce the probability of selecting a functional IGR-containing fragments. IGR is also poorly identified with the initiating ribosome approach because it is located in the vicinity of RNA fragments that are putatively related to translation and which are very abundant in the starting library. IGR is however better highlighted with the elongating ribosomes strategy.

Concerning the identification of the 5' end of both IRES, we observed that they are less precise than the 3' ends. This is also due to molecular mechanism used by the IRES. Indeed, IRES are able to recruit a ribosome in the middle of an RNA molecule, therefore the sequences upstream of the IRES do not impair ribosome recruitment. RNA fragments containing the minimal IRES will be selected in addition to longer fragments that encompass the whole IRES with additional upstream sequences. In agreement with this statement, the CrPV IRES1 is located in the 5'UTR of the genome, however the minimal IRES1 is located between nucleotides 357 and 709 (CrPV genome coordinates), meaning that the 5' proximal 356 first nucleotides are not required for the IRES activity but does not interfere with the ribosome recruitment (Gross *et al.* 2017). Similarly, the IGR recruits the ribosome in the middle of the CrPV genome showing that the upstream sequences do not impact on the IGR activity. Altogether, this explains why the identification of the 5' end of the IRES is less precise than the 3' end.

Even though the use of puromycin allowed to better discriminate translation-related from translation-unrelated fragments and thus better highlighted the expected IRES fragments, we propose an alternative experimental strategy for the same purpose. We would cross our data with *in vitro* ribosome profiling data obtained on the CrPV starting library. Ribosomal footprints recovered after *in vitro* assembly of translation complexes in the presence of harringtonine on the RNA starting library followed by RNase digestion would give the translation initiation sites locations on the reference sequence. Combining both datasets to identify fragments featuring an initiation site would be an interesting improvement of our method to get rid of unspecific ribosomal binders.

5. References

Gross L, Vicens Q, Einhorn E, Noireterre A, Schaeffer L, Kuhn L, Imler J-L, Eriani G, Meignin C, Martin F. 2017. The IRES5′UTR of the dicistrovirus cricket paralysis virus is a type III IRES containing an essential pseudoknot structure. *Nucleic Acids Res* **45**: 8993–9004.

Jaafar ZA, Kieft JS. 2019. Viral RNA structure-based strategies to manipulate translation. *Nat Rev Microbiol* **17**: 110–123.

Khong A, Bonderoff J, Spriggs R, Tammpere E, Kerr C, Jackson T, Willis A, Jan E. 2016. Temporal Regulation of Distinct Internal Ribosome Entry Sites of the Dicistroviridae Cricket Paralysis Virus. *Viruses* **8**: 25.

Lee K-M, Chen C-J, Shih S-R. 2017. Regulation Mechanisms of Viral IRES-Driven Translation. *Trends Microbiol* **25**: 546–561.

Mailliot J, Martin F. 2018. Viral internal ribosomal entry sites: four classes for one goal: Viral internal ribosomal entry sites. *Wiley Interdiscip Rev RNA* **9**: e1458.

Manousis T, Moore NF. 1987. Cricket Paralysis Virus, a Potential Control Agent for the Olive Fruit Fly, *Dacus oleae* Gmel. *Appl Environ Microbiol* **53**: 142–148.

Nayak A, Berry B, Tassetto M, Kunitomi M, Acevedo A, Deng C, Krutchinsky A, Gross J, Antoniewski C, Andino R. 2010. Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila. *Nat Struct Mol Biol* **17**: 547–554.

Pelham HRB, Jackson RJ. 1976. An Efficient mRNA-Dependent Translation System from Reticulocyte Lysates. *Eur J Biochem* **67**: 247–256.

Pernod K, Schaeffer L, Chicher J, Hok E, Rick C, Geslain R, Eriani G, Westhof E, Ryckelynck M, Martin F. 2020. The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic Acids Res* **48**: 6170–6183.

Pestova TV, Hellen CUT. 2003. Translation elongation after assembly of ribosomes on the Cricket paralysis virus internal ribosomal entry site without initiation factors or initiator tRNA. *Genes Dev* **17**: 181–186.

Walsh D, Mohr I. 2011. Viral subversion of the host protein synthesis machinery. *Nat Rev Microbiol* **9**: 860–875.

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E. 2016. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**: aad4939.

Wilson JE, Pestova TV, Hellen CUT, Sarnow P. 2000. Initiation of Protein Synthesis from the A Site of the Ribosome. *Cell* **102**: 511–520.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

6. Acknowledgments

This work is funded by *Agence Nationale pour la Recherche* (ANR-17-CE12-0025-01, ANR-17-CE11-0024, ANR-20-COVI-0078), by *Fondation pour la Recherche Médicale* (project CoronaIRES), by *Fondation Bettencourt Schueller*, by University of Strasbourg and by the *Centre National de la Recherche Scientifique*. The pCrPV-3 plasmid was kindly provided by Eric Jan. We would like to thank the IBMP AEG sequencing platform.

7. Figure legends

Figure 1: overview of the experimental strategy

(**A**) A plasmid containing the whole CrPV genome is randomly fragmented and each resulting dsDNA fragment is inserted upstream the Renilla luciferase coding sequence and downstream a T7 promoter by TA-ligation into a home-made T-vector. (**B**) In order to avoid leaderless mRNA translation, we inserted a stable hairpin at the 5' end of the reporter mRNA library followed by flexible linker (in blue). (**C**) After PCR amplification, the amplified DNA reporter library is *in vitro* transcribed from a T7 promoter using a recombinant T7 RNA polymerase. (**D**) Next, translation complexes are assembled on the resulting mRNA reporter library using either RRL or S2 cell-free extracts and (**E**) separated on a sucrose density gradient. (**F**) RNA extraction is performed on 80S-containing fractions and the CrPV fragments thus selected are reverse-transcribed (**G**), PCR-amplified for adding Illumina indexes (**H**), sequenced by Illumina sequencing (**J**) and analysed (**K**). Data analysis is done using a self-made Python algorithm (**L**).

Figure 2: the CrPV genome contains two IRES elements that can initiate translation *in vitro* with cell-free translation extracts

A. Schematic representation of the Cricket Paralysis Virus genome.

B. The two CrPV IRES can efficiently initiate translation *in vitro* with S2 cell-free translation extracts with 2mM magnesium acetate.

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with S2 cell-free translation extracts using 1 mM, 1.5 mM or 2 mM magnesium acetate (MgAc). Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution.

Figure 3: the insertion of a 5' hairpin efficiently prevents cap-independent translation without impairing IRES translation in model RNA reporters

A. Schematic representation of the reporter RNA used in B.

The blue part corresponds to the 5' hairpin followed by the flexible linker and the T-vector sequence that remains after T-A ligation.

B. *In vitro* translation assays with RRL.

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase from RNA reporters using RRL which feature: the 5'UTR of β-globine, the m^7G-capped 5'UTR of β-globine, the IGR of CrPV, which are preceded or not by the 5' hairpin sequence followed by the flexible linker (blue).

Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution.

Figure 4: effect of increasing potassium concentration on sucrose gradient fractionation of ribosomal complexes

Plots representing the percentage of input RNA (total counts in cpm) of ^{32}P-radiolabeled m^{7}G 5'UTR β-globine RNA in each fraction of the gradient. Free refers to unbound RNAs.

Figure 5: starting library obtained with the truncated Renilla coding sequence

A. Schematic representation of the reporter library.

The truncated, 36 nucleotides long Renilla luciferase (Rluc) coding sequence (cds) is followed by a short 3'UTR (grey) to avoid ribosome stalling after termination.

B. Raw alignment profiles of the recovered fragments on the pCrPV-3 reference sequence.

The observed proportion of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

Vertical dashed black lines delimit the coordinates of the two minimal IRES regions.

C. Most of the recovered fragments have a length ranging from 100 to 800 nucleotides.

Top: scatterplot showing the observed proportions of each fragment as a function of their length.

Bottom: magnification of the part of the plot but showing only the fragments with a length smaller the 75th percentile of their length= distribution.

Figure 6: starting library obtained with the full-length Renilla coding sequence

A. Schematic representation of the reporter library.

The Renilla luciferase (Rluc) coding sequence (cds) is followed by a short 3'UTR (grey) to avoid ribosome stalling.

B. Raw alignment profiles of the recovered fragments on the pCrPV-3 reference sequence.

The observed proportion of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

Vertical dashed black lines delimit the coordinate of the two minimal IRES regions.

C. Most of the recovered fragments have a length ranging from 100 to 700 nucleotides.

Top: scatterplot showing the observed proportions of each fragment as a function of their length.

Bottom: magnification of the part of the plot but showing only the fragments with a length smaller the 75th percentile of their length distribution.

Figure 7: Signal Convergence Matrices obtained after RNA isolation from initiating 80S ribosomes

Because most of the fragments have a length between 100 and 800bp, the y-axis range is set to 100-800bp. Similarly, the x-axis range is set to [700-10000] which are the coordinates of the CrPV genome on the pCrPV-3 plasmid. The length of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence. Color intensities represent the sum of the observed proportions of fragments of size i that cover position j on the reference genome.

Only positive values are shown, which denote enriched regions after gradient selections.

Vertical dashed black lines delimit the coordinates of the two minimal IRES regions.

Vertical dashed green lines correspond to the 5' N-terminal coordinates of the viral proteins after proteolytic cleavage of the corresponding polyprotein.

A: after one round of selection with RRL.

B: after two rounds of selection with RRL.

C: after one round of selection with S2 cell-free translation extracts.

D: after two rounds of selection with S2 cell-free translation extracts.

Figure 8: the selected libraries are *in vitro* translated in RRL

A. Overview of the gradient-based selections performed with RRL and S2 cell-free translation extracts.

R0: starting library.

R1: library selected after one round of selection.

R2: library selected after two rounds of selection.

B. *In vitro* translation assays performed with RRL on the RNA libraries that have been selected with RRL (left) or S2 cell-free translation extracts (right) after one round of selection.

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with RRL. Relative Luminescence Units (R.L.U.) were normalized to those obtained with the R0 RNA reporters.

Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution.

Figure 9: graphical cross-correlation between Signal Convergence Matrices and background Fragment Repartition Matrices with RRL-selected fragments

Because most of the fragments have a length between 100 and 800bp, the y-axis range is set to 100-800bp. Similarly, the x-axis range is set to [700-10000] which are the coordinates of the CrPV genome on the pCrPV-3 plasmid. The length of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

Blue regions indicate regions only found in the presence of puromycin. Red regions indicate regions of increased fragments' concentration between the selected (without puromycin) and the initial libraries. Purple regions indicate cross-correlating enriched regions between the selected libraries in the absence and the presence of puromycin and can therefore be considered as false-positives. Regions of interest are thus the red ones.

Vertical dashed black lines delimit the coordinates of the two minimal IRES regions.

Vertical dashed green lines correspond to the 5' N-terminal coordinates of the viral proteins after proteolytic cleavage of the corresponding polyprotein.

A: after one round of selection with RRL.

B: after two rounds of selection with RRL.

Figure 10: graphical cross-correlation between Signal Convergence Matrices and background Fragment Repartition Matrices with S2-selected fragments

Because most of the fragments have a length between 100 and 800bp, the y-axis range is set to 100-800bp. Similarly, the x-axis range is set to [700-10000] which are the coordinates of the CrPV genome on the pCrPV-3 plasmid. The length of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

Blue regions indicate regions only found in the presence of puromycin. Red regions indicate regions of increased fragments' concentration between the selected (without puromycin) and the initial libraries. Purple regions indicate cross-correlating enriched regions between the selected libraries in the absence and the presence of puromycin and can therefore be considered as false-positives. Regions of interest are thus the red ones.

Vertical dashed black lines delimit the coordinate of the two minimal IRES regions

Vertical dashed green lines correspond to the 5' N-terminal coordinates of the viral proteins after proteolytic cleavage of the corresponding polyprotein.

A: after one round of selection with S2-cell free translation extracts.

B: after two rounds of selection with S2-cell free translation extracts.

8. Table legend

Table 1: table summarizing the CrPV regions identified as IRES candidates with both types of extracts

With IRES1 and IGR, smaller fragments corresponding to truncated forms of the IRES are cross-correlating with background signals (purple) and/or lost (blue). In other words, with IRES1 and IGR regions, we observe two sets of smaller fragments that are not enriched. We therefore expect to observe the same kind of pattern with other regions which would have an IRES feature. This strategy enables to point three additional regions that may contain RNA fragments of interest that are labelled “To investigate”.

Regions labelled with a ‘?’ are those that are enriched but do not feature the previously mentioned pattern and mix locally cross-correlating (purple) and/or lost (blue) fragments together with enriched fragments. There is therefore a greater uncertainty on those ones.

9. Supplementary Figures and table

Supplementary Figure S1: the two CrPV IRES are efficiently *in vitro* translated with Aag2 cell-free translation extracts

Real-time fluorescence analysis of *in vitro* translated GFP from IRES1-driven or IGR-driven RNA reporters using cell-free translation extracts prepared from Aag2 (blue) or S2 (orange) cells.

Supplementary Figure S2: selected libraries obtained with the full-length Renilla coding sequence without ribosome salt wash

Supplementary Figure S3: selected libraries obtained with the 36 nucleotides-long Renilla coding sequence using RRL (left) or S2 (right) cell-free translation extracts

Raw alignement profiles of the recovered pCrPV-3 fragments on the pCrPV-3 reference sequence. The observed proportion of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

R0: starting library.

R1: fragments recovered after one round of selection.

R2: fragments recovered after two rounds of selection.

Vertical dashed black lines delimit the coordinates of the two minimal IRES regions.

Supplementary Figure S4: polysomes are less abundant than monosomes in RRL, and undetectable in S2 cell-free translation extracts with absorbance measurements

Plots representing the absorbance profiles at 260 nm of the sucrose gradients collected after one round of selection with RRL or S2 cell-free extracts in ribosome salt wash conditions (700 mM potassium acetate). Free refers to unbound RNAs.

Supplementary Figure S5: the selected libraries are *in vitro* translated in RRL

A. Overview of the gradient-based selections performed with RRL and S2 cell-free translation extracts.

R0: starting library.

R1: library selected after one round of selection.

R2: library selected after two rounds of selection.

B. *In vitro* translation assays performed with RRL on the RNA libraries that have been selected with RRL (left) or S2 cell-free translation extracts (right) after one and two rounds of selection.

Histogram showing luminescence quantification of *in vitro* translated Renilla luciferase with RRL.

Error bars represent the 99% confidence interval calculated from three independent replicates using the t-distribution.

Supplementary Figure S6: selected libraries obtained with the full-length Renilla coding sequence using RRL (B) or S2 (C) cell-free translation extracts

Raw alignement profiles of the recovered fragments on the pCrPV-3 reference sequence. The observed proportion of each fragment in the starting library is shown as a function of the coordinates on the pCrPV-3 reference sequence.

Vertical dashed black lines delimit the coordinate of the two minimal IRES regions.

A. R0 starting library.

B-C-D-E. Raw alignement profiles obtained with RRL after one round (B and C) or two (D and E) rounds of selection. Left panels (B and D) show profiles obtained without puromycine and right panels (C and E) show profiles obtained with puromycine.

F-G-H-I. Raw alignement profiles obtained with S2 cell-free translation extracts after one (B and C) or two (D and E) rounds of selection. Left panels (F and H) show profiles obtained without puromycine and right panels (G and I) show profiles obtained with puromycine.

Supplementary Table S1: oligonucleotides' sequences

Figure 1: overview of the experimental strategy

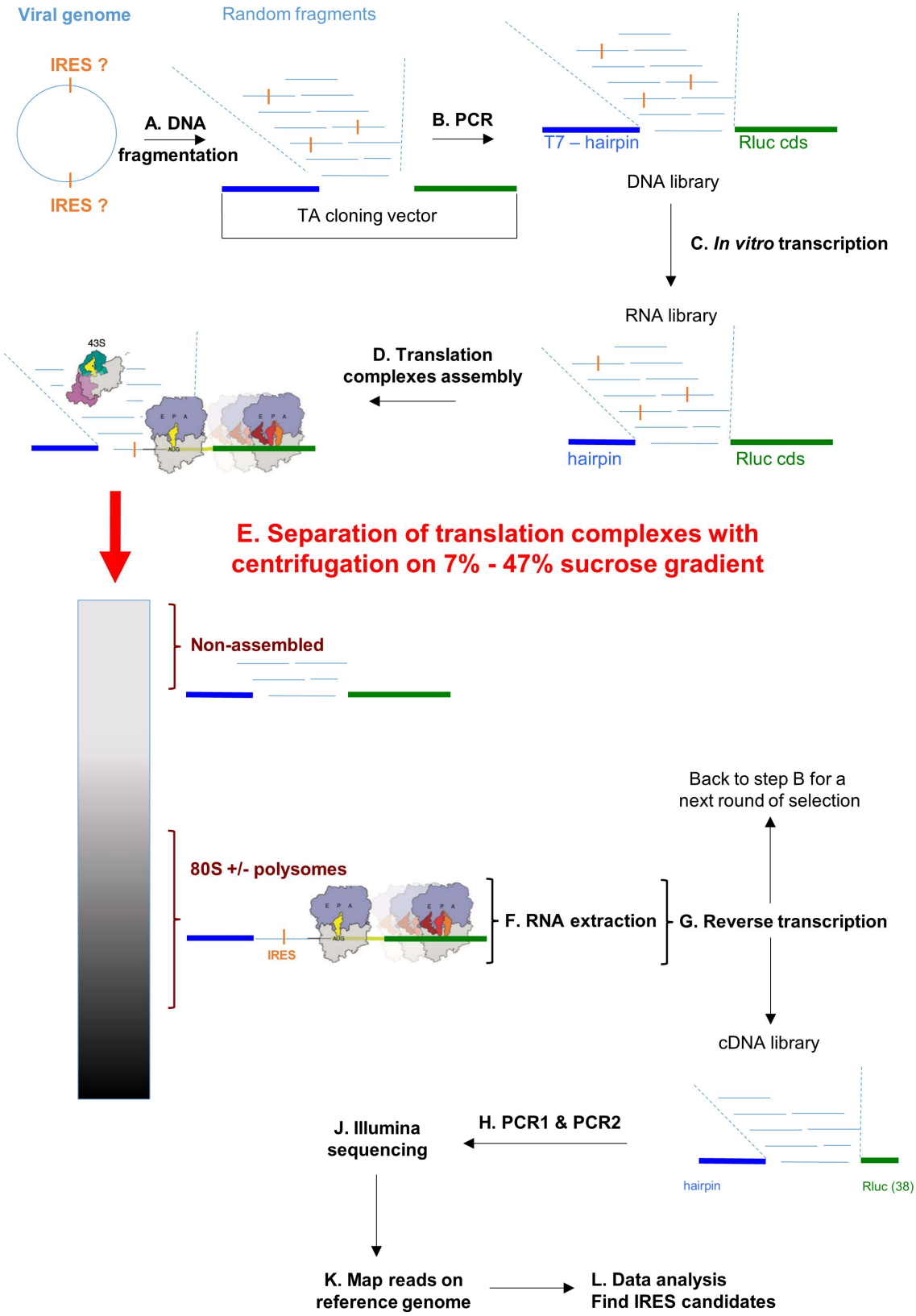


Figure 2: the CrPV genome contains two IRES elements that can initiate translation *in vitro* with cell-free translation extracts

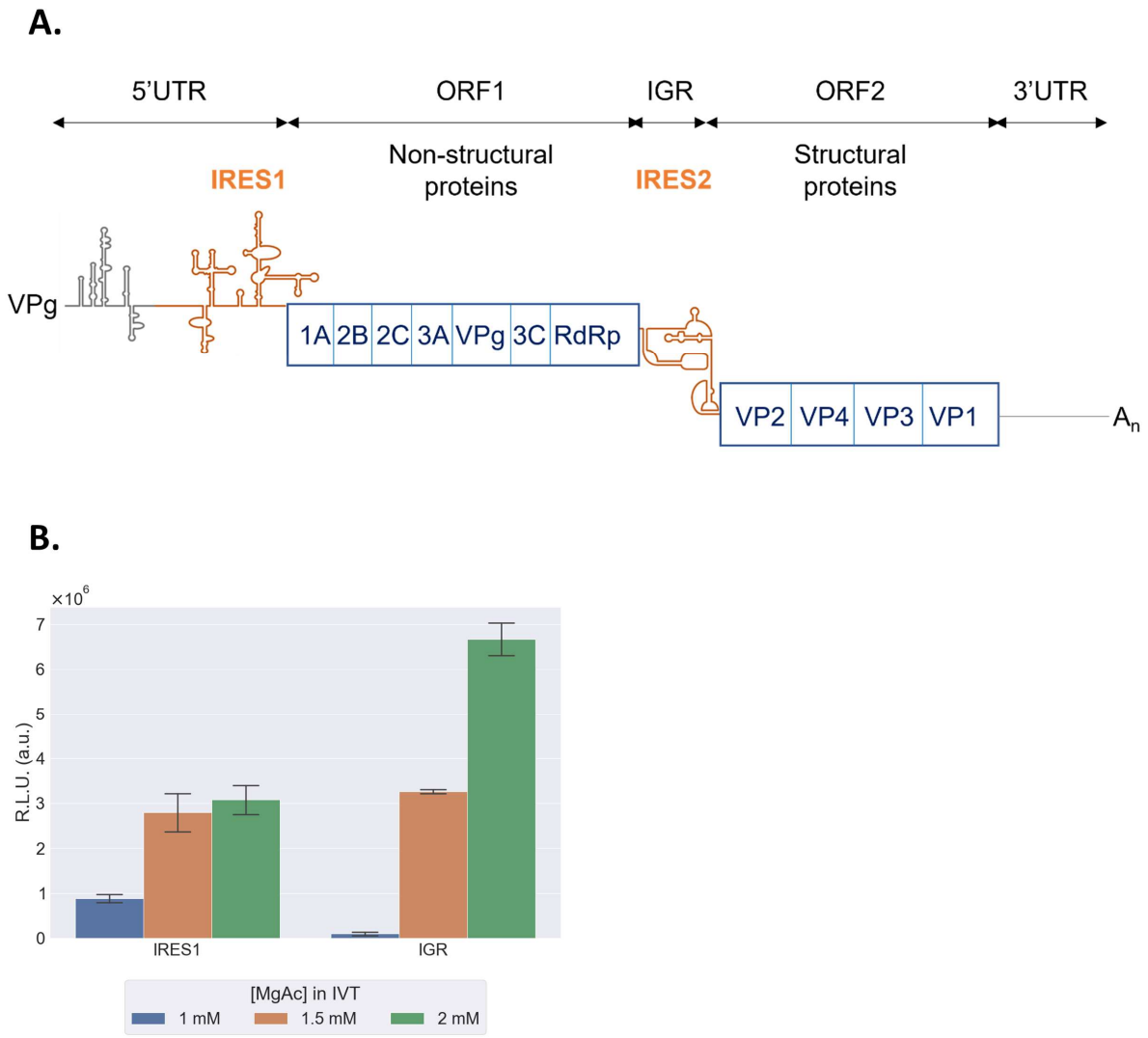
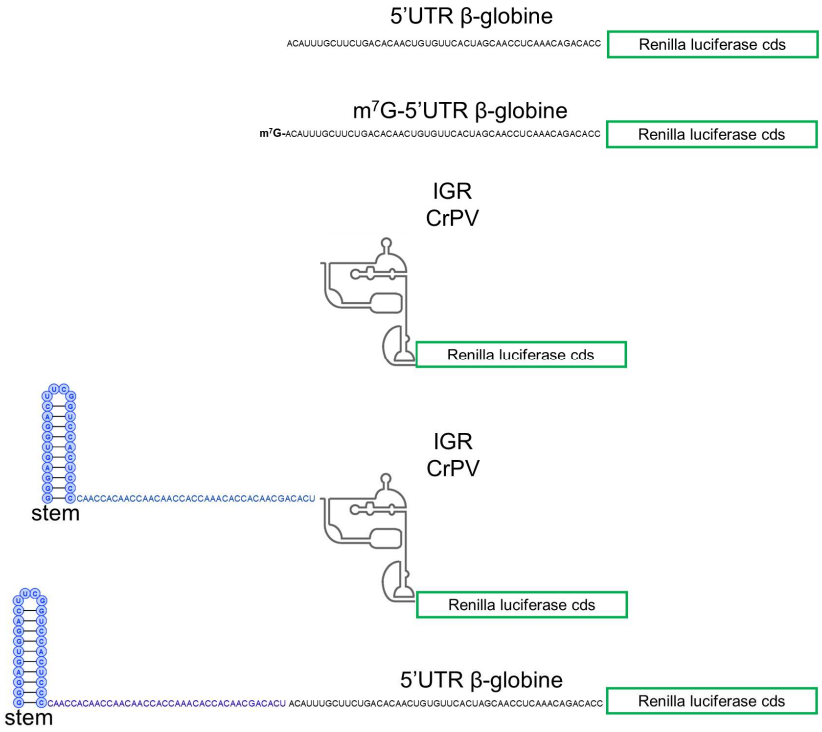


Figure 3: the insertion of a 5’ hairpin efficiently prevents cap-independent translation without impairing IRES translation in model RNA reporters

A.



B.

Figure 4: effect of increasing potassium concentration on sucrose gradient fractionation of ribosomal complexes

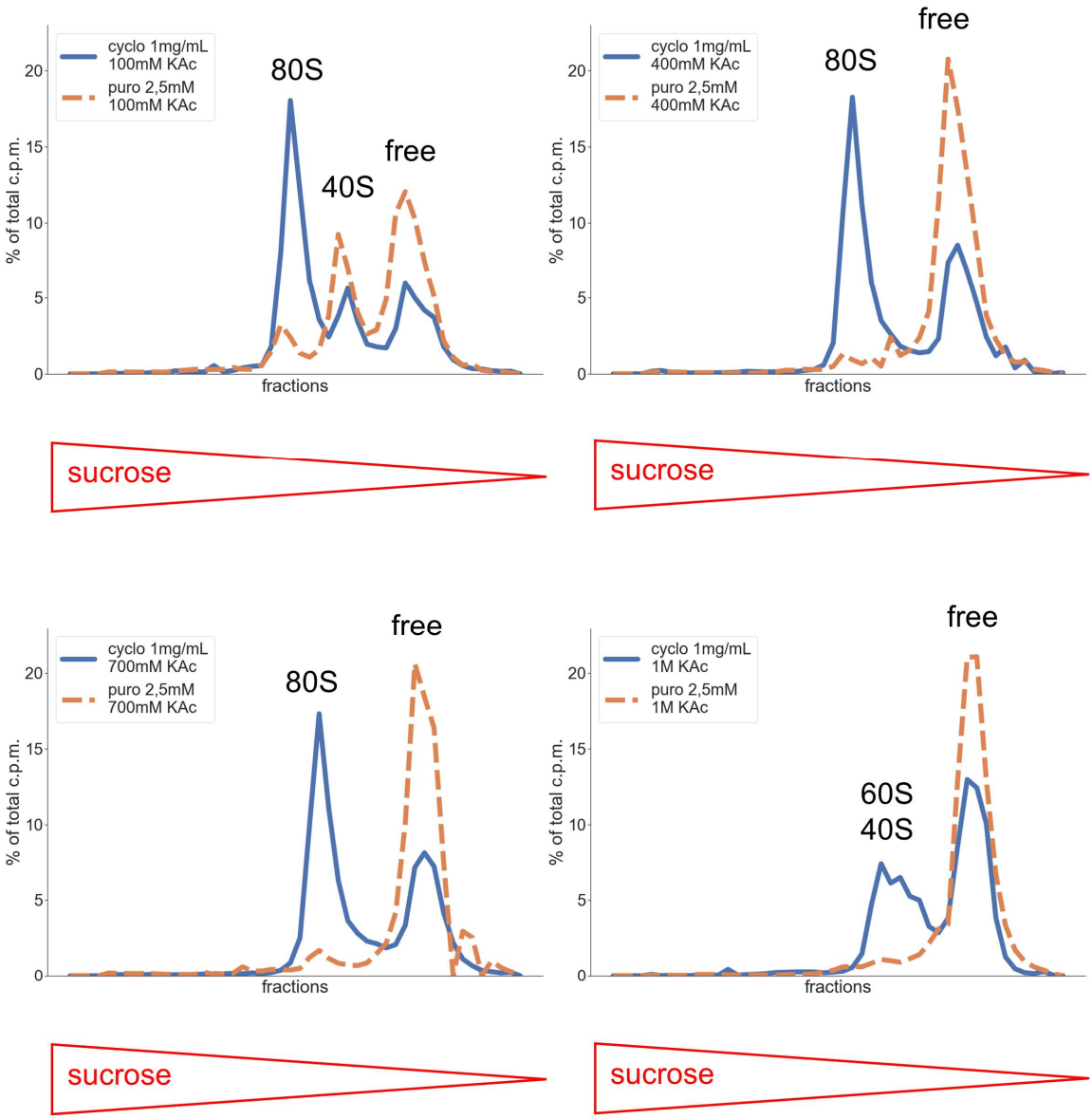


Figure 5: starting library obtained with the truncated Renilla coding sequence

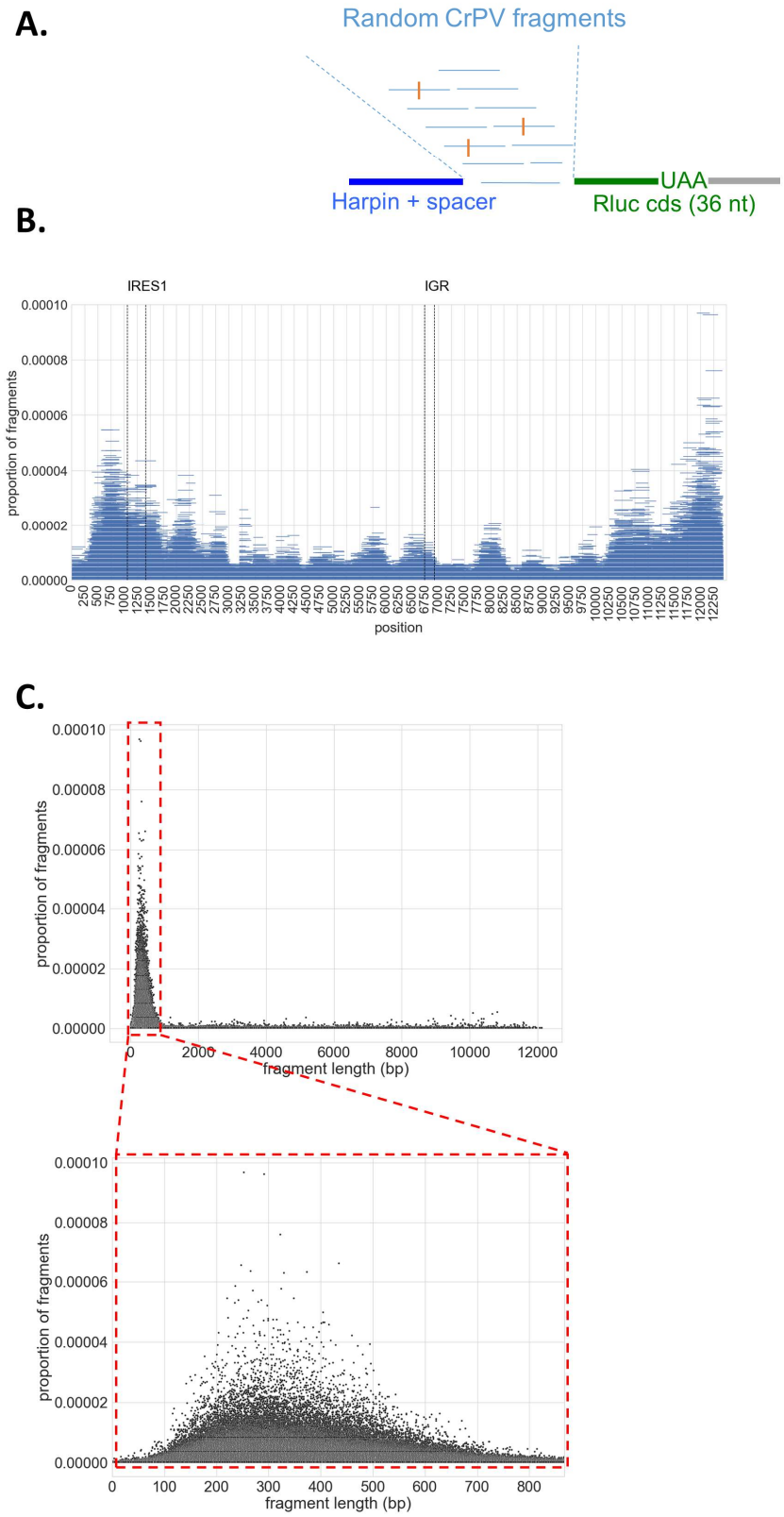


Figure 6: starting library obtained with the full-length Renilla coding sequence

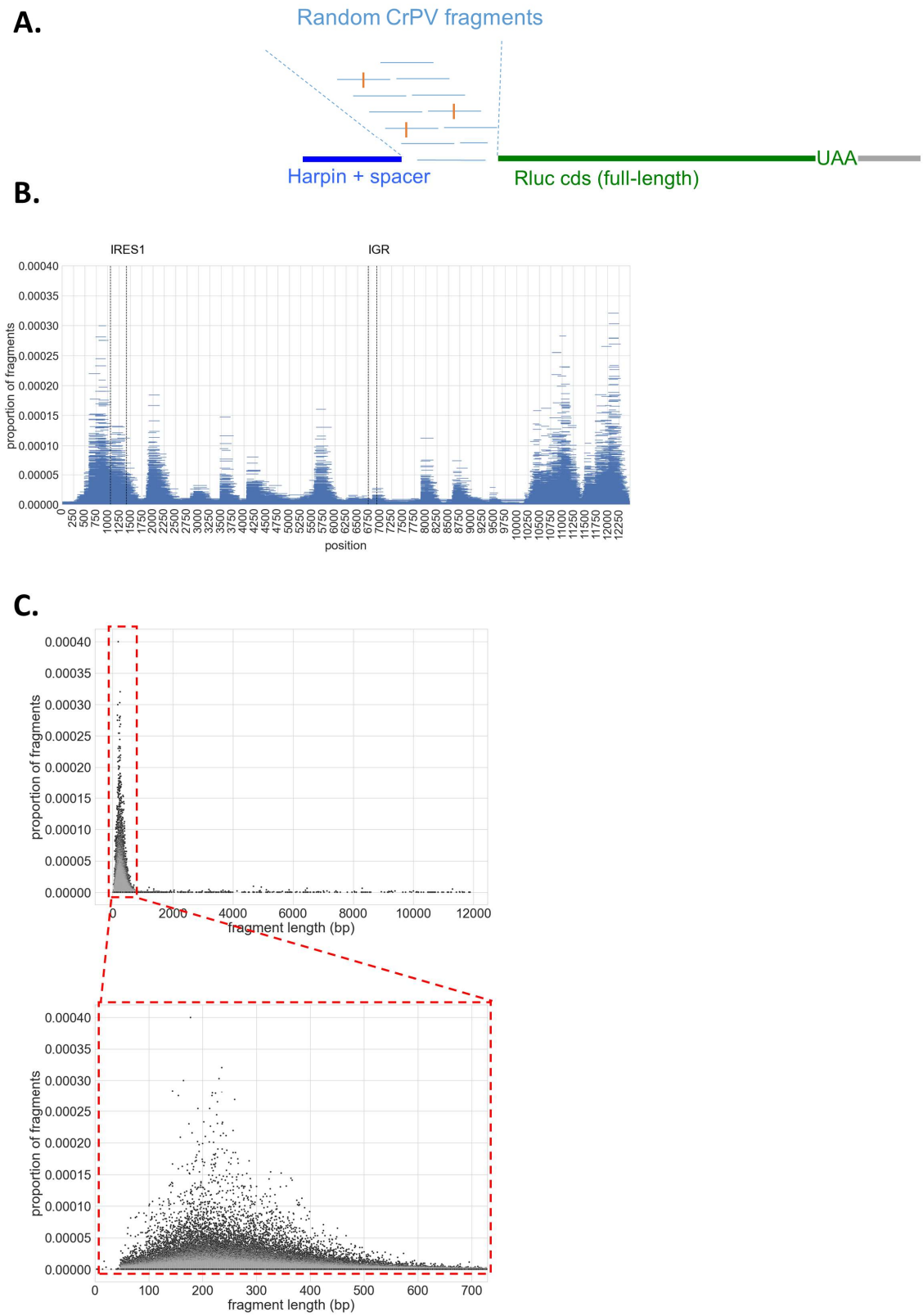


Figure 7: Signal Convergence Matrices obtained after RNA isolation from initiating 80S ribosomes

A.

B.

C.

D.

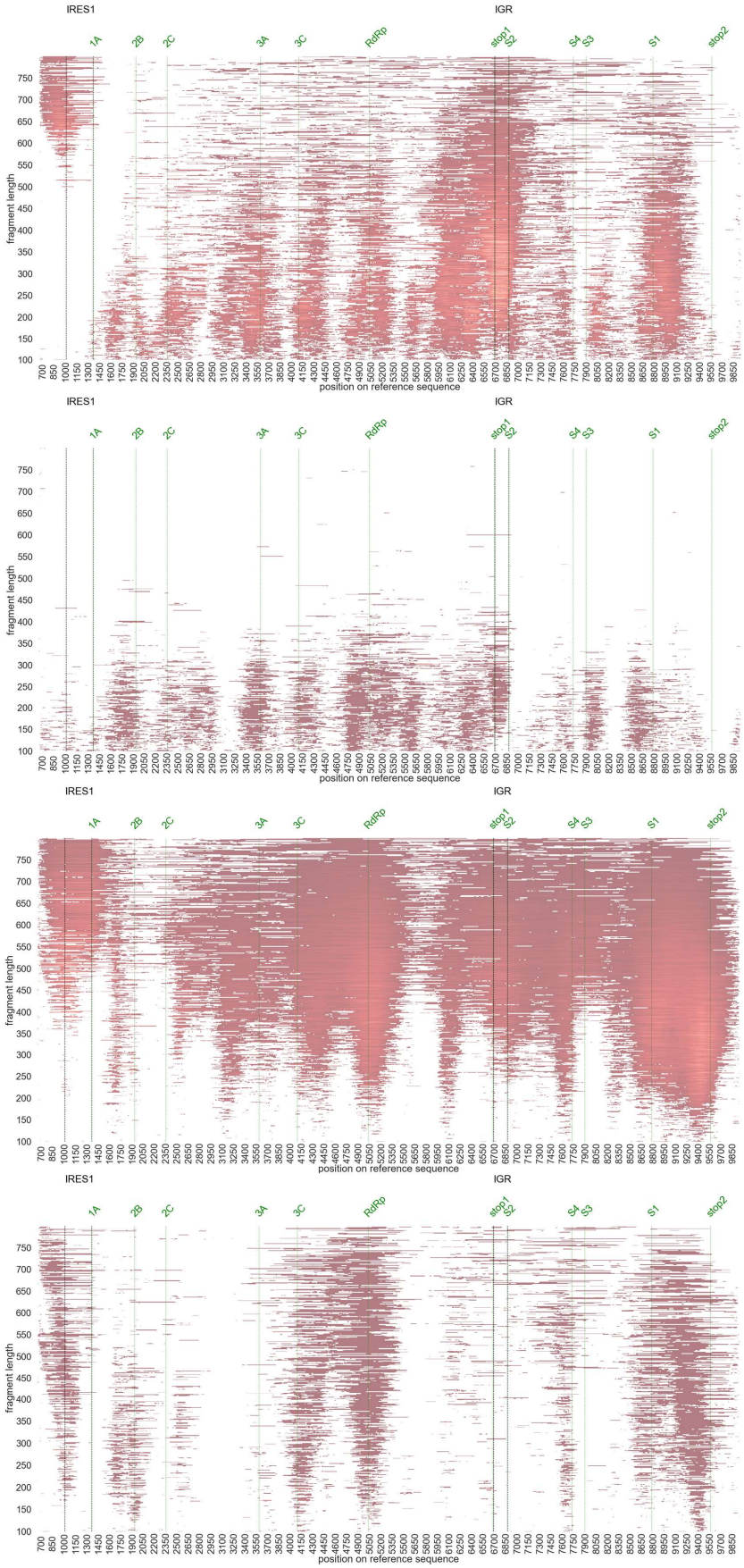


Figure 8: the selected libraries are *in vitro* translated in RRL

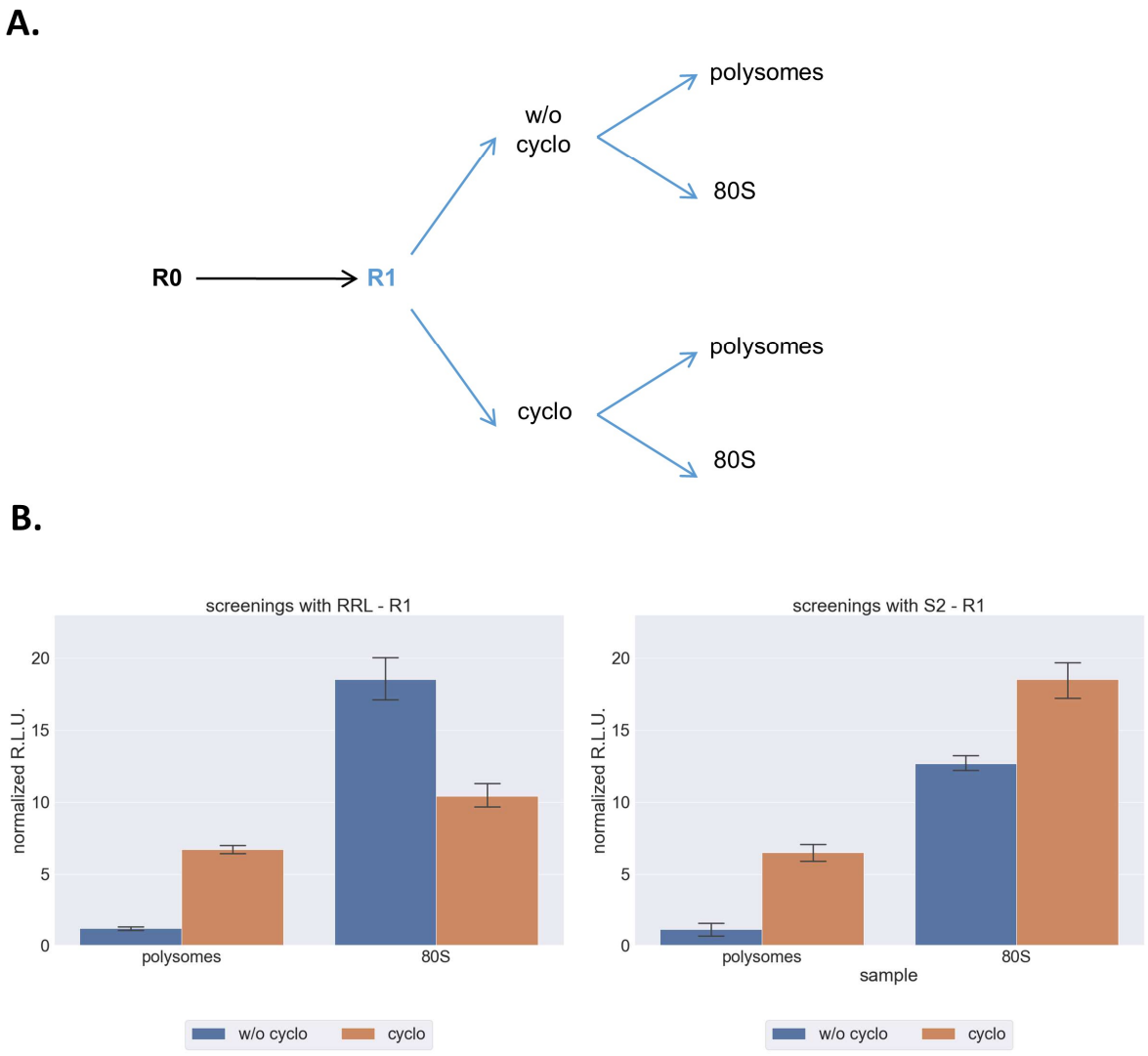
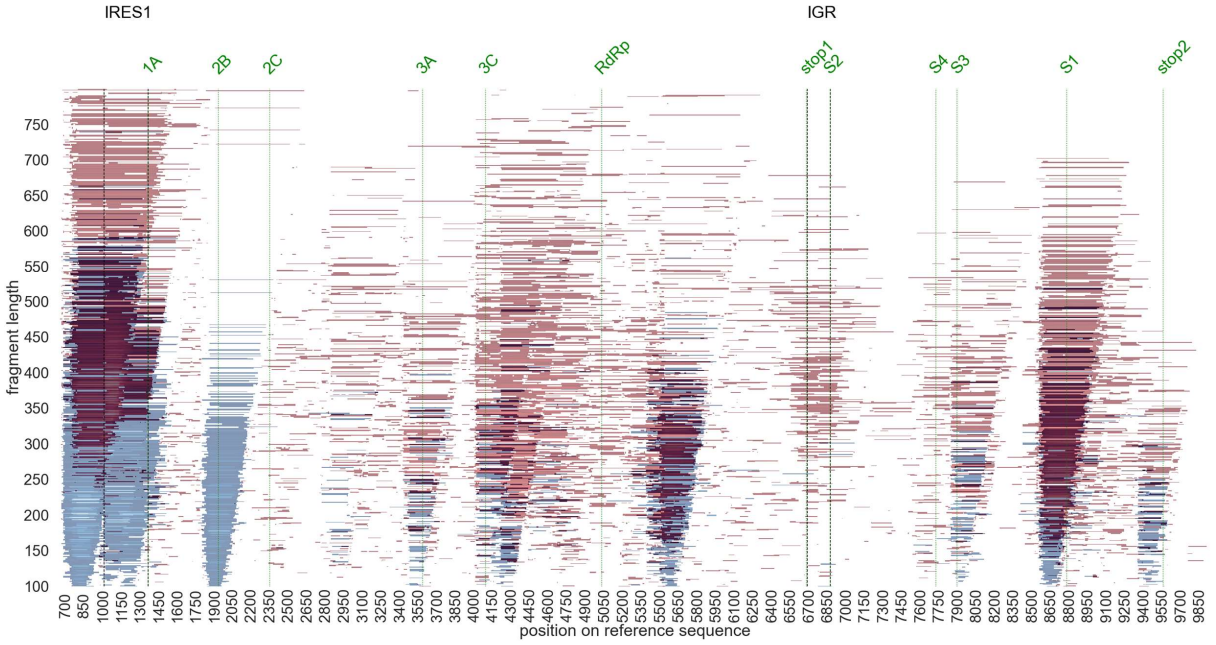


Figure 9: graphical cross-correlation between Signal Convergence Matrices and background Fragment Repartition Matrices with RRL-selected fragments

A.



B.

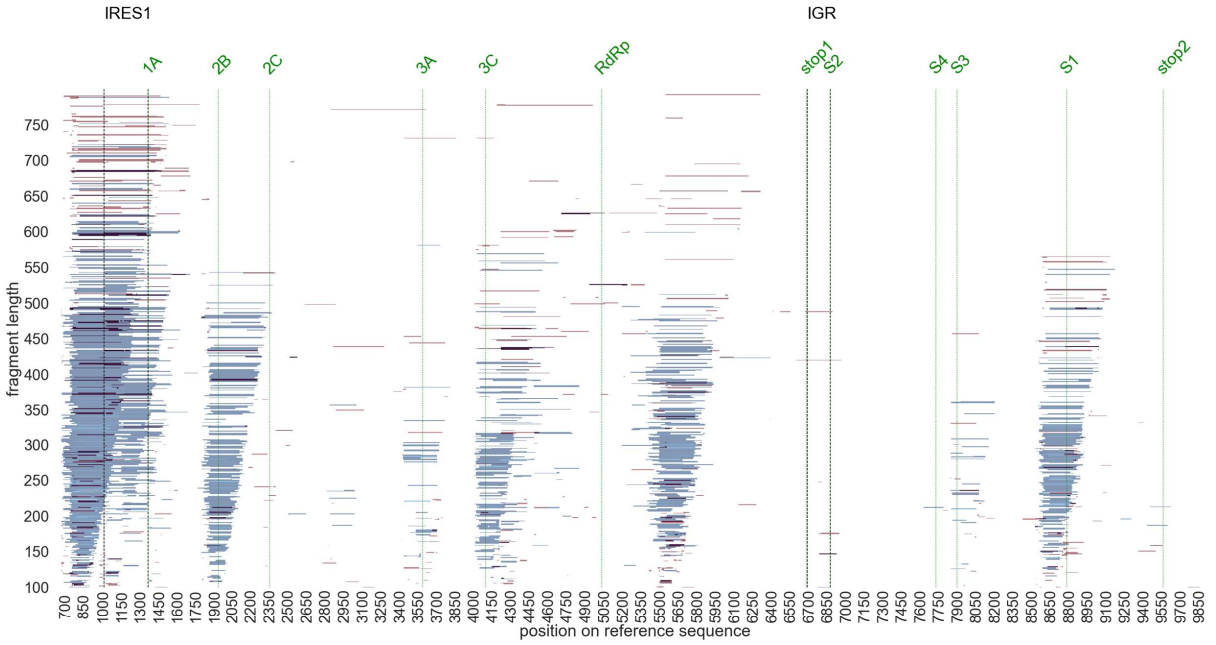
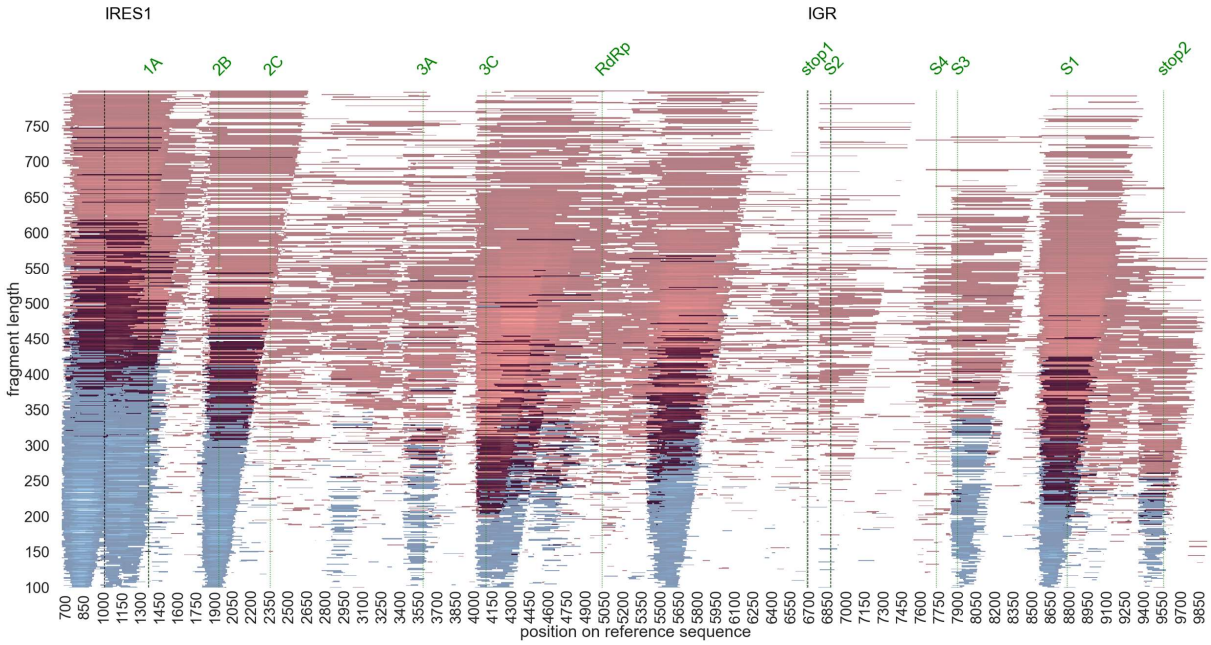


Figure 10: graphical cross-correlation between Signal Convergence Matrices and background Fragment Repartition Matrices with S2-selected fragments

A.



B.

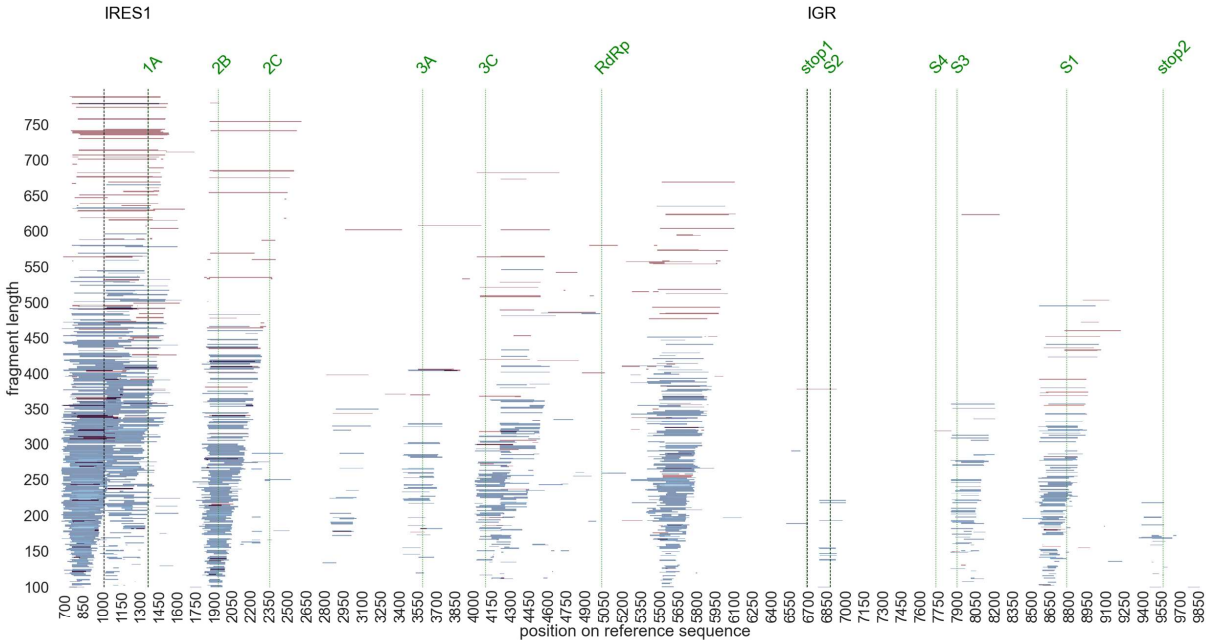
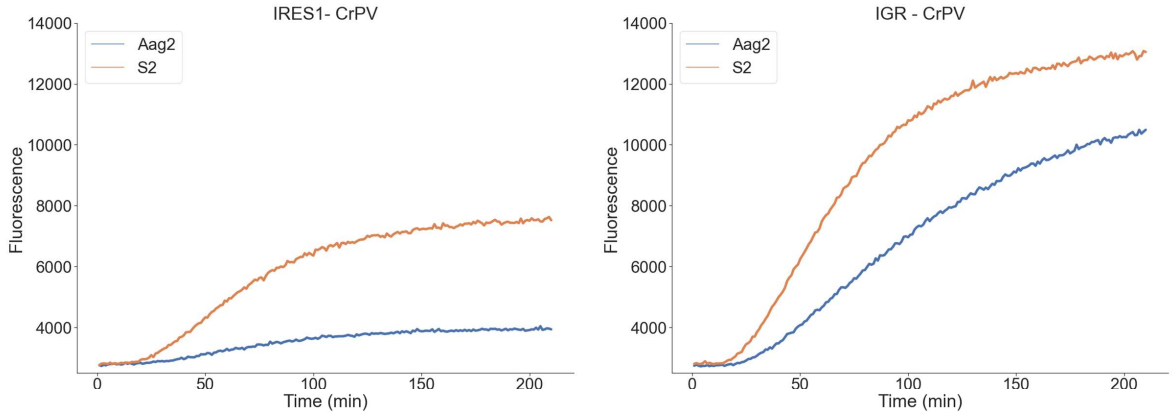


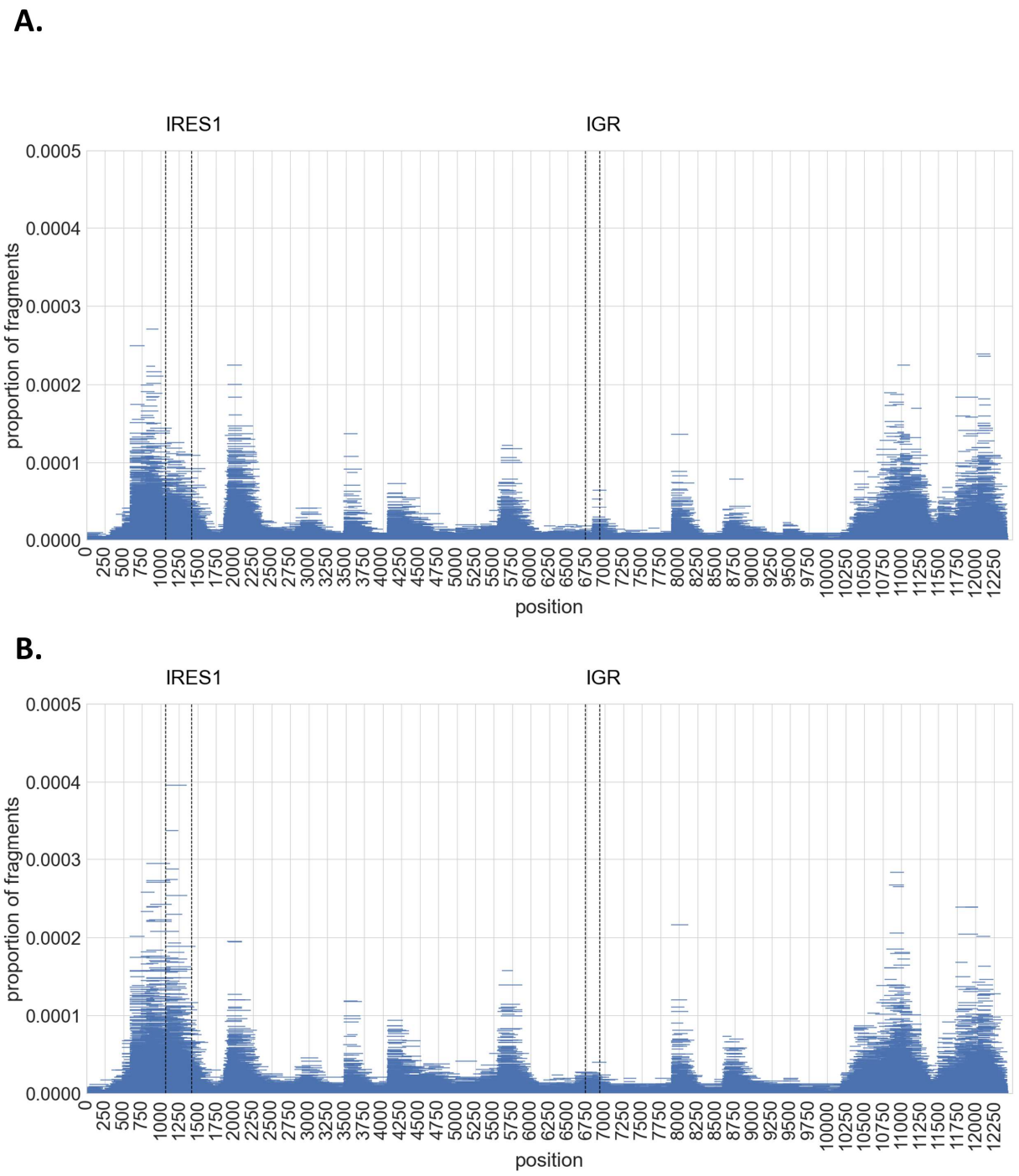
Table 1: table summarizing the CrPV regions identified as IRES candidates with both types of extracts

Region	Extract	Comment
700-1450	RRL & S2	IRES1
1800-2500	S2	To investigate
3000-3400	RRL & S2	To investigate
4000-4600	RRL & S2	To investigate
5400-5900	RRL & S2	?
6600-6900	RRL & S2	IRES2
8000-8200	RRL & S2	?
8500-9200	RRL & S2	?

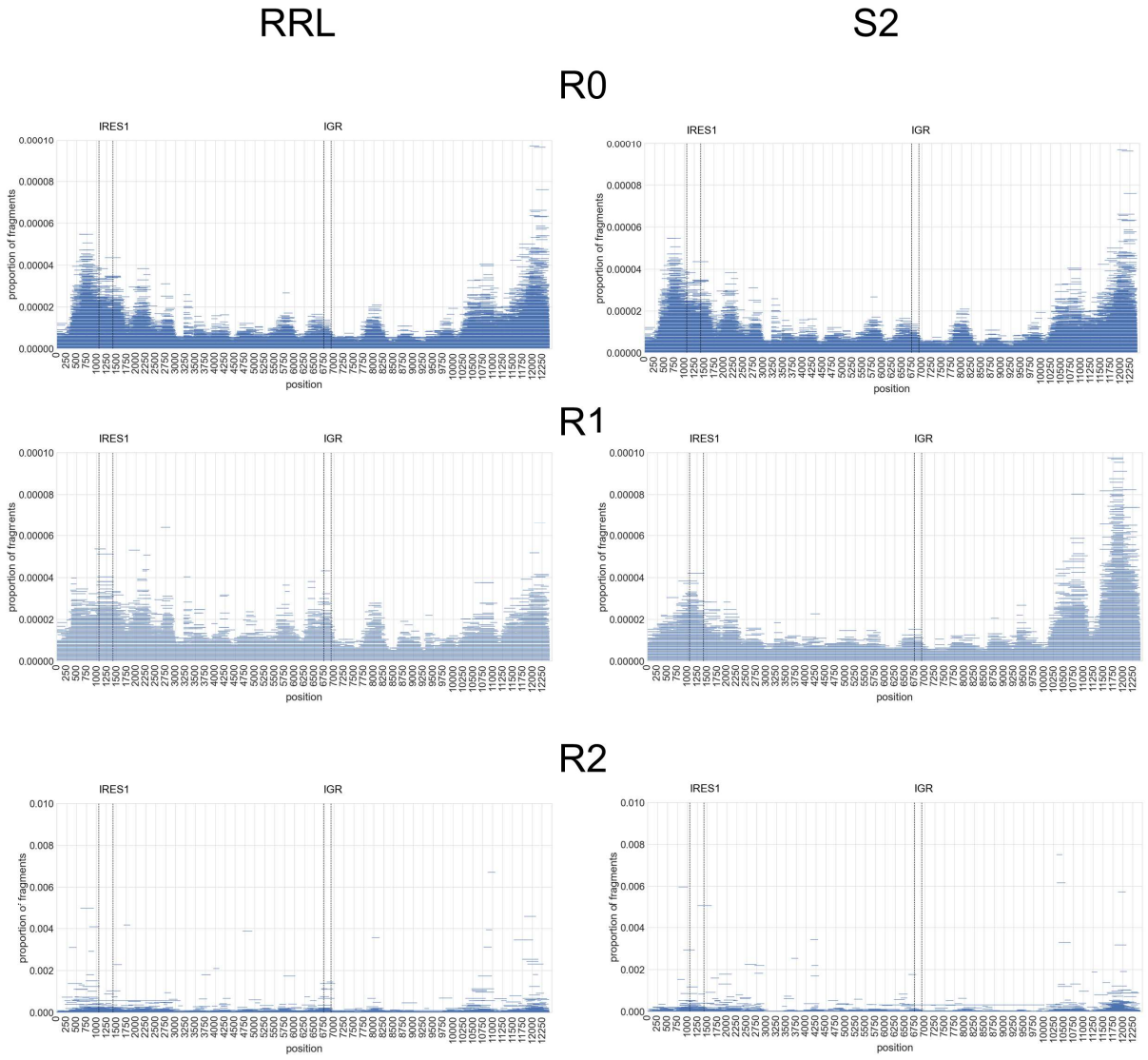
Supplementary Figure S1: the two CrPV IRES are efficiently *in vitro* translated with Aag2 cell-free translation extracts



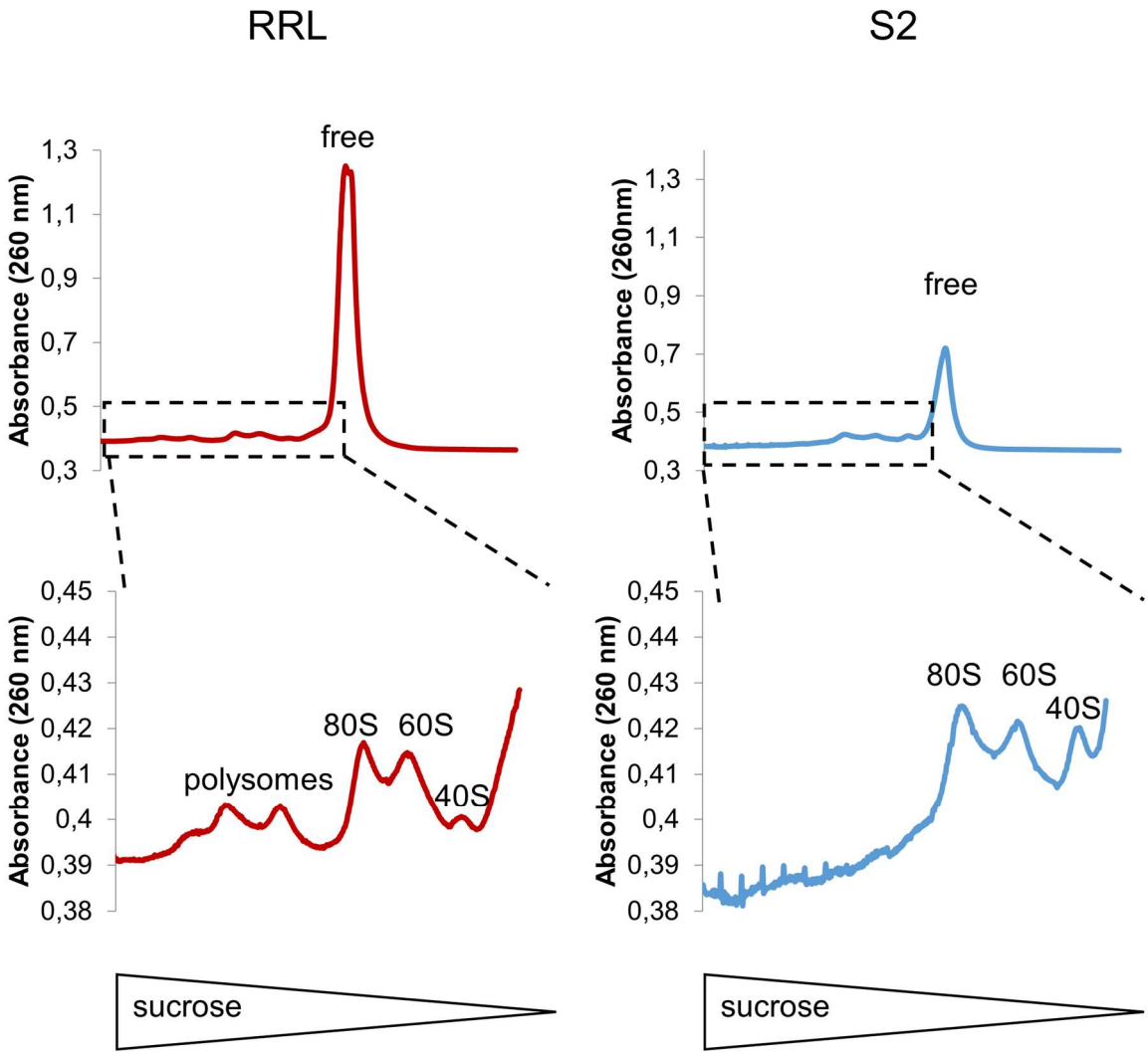
Supplementary Figure S2: selected libraries obtained with the full-length Renilla coding sequence without ribosome salt wash



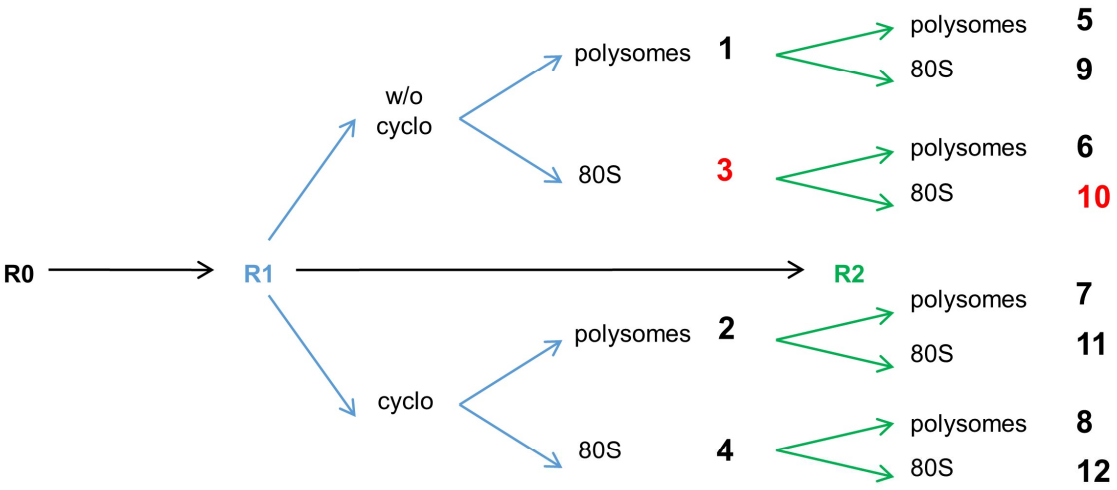
Supplementary Figure S3: selected libraries obtained with the 36 nucleotides-long Renilla coding sequence using RRL (left) or S2 (right) cell-free translation extracts

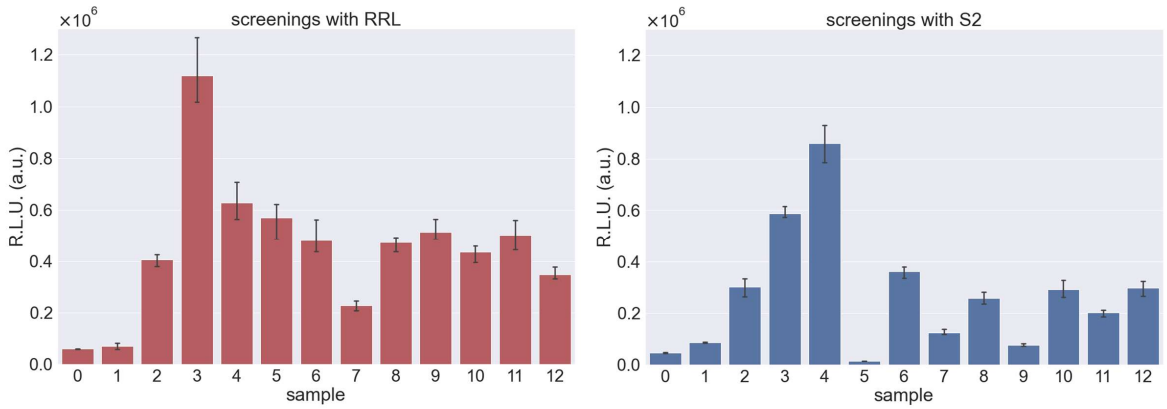


Supplementary Figure S4: polysomes are less abundant than monosomes in RRL, and undetectable in S2 cell-free translation extracts with absorbance measurements

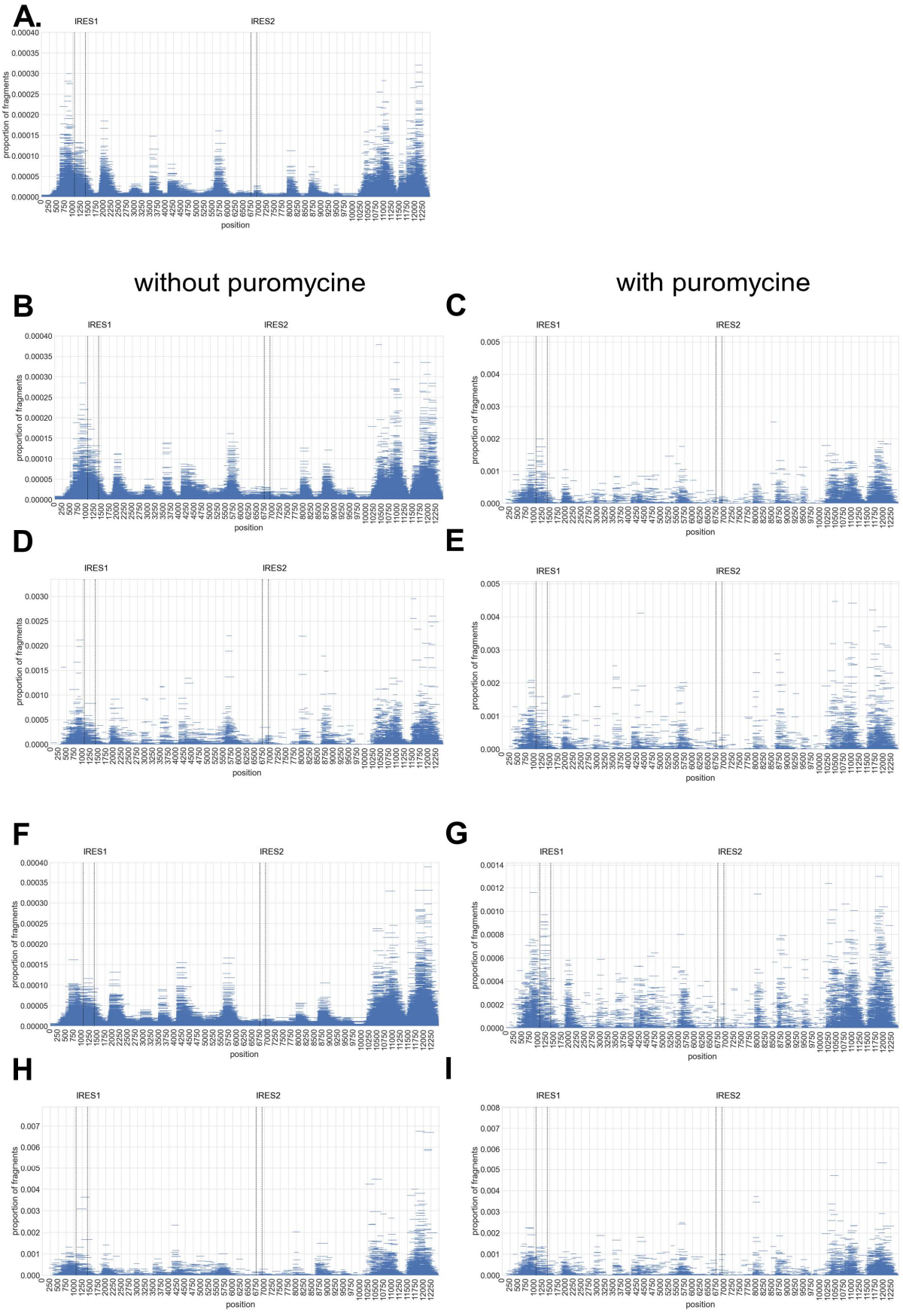


Supplementary Figure S5: the selected libraries are *in vitro* translated in RRL





Supplementary Figure S6: selected libraries obtained with the full-length Renilla coding sequence using RRL (B) or S2 (C) cell-free translation extracts



Supplementary Table S1: oligonucleotides' sequences

n° in Text	Name	Sequence 5'-3'
1	fw_T7_stem	CAAACAGGATCCTATATTAATACGACTCACTATAGGGAGTGGACTTCGGACTTC
2	rev_Renilla_NotI_pUC19(SpeI)	GGAAACAGCTATGACCATGATTACGCCAAGCGGCCGCTTGGTGTGGACGGTTTATACTTATTGTTCATTTTTGAGAACTCGC
3	rev_Renilla(36)_stop_CAA	GTGTTGTTGTTGTTGTTAGTTGTTATTTCCTTTGTTCTGGATCATAAACTTTCGAAGTCAT
4	rev_Renilla_38	CATCCGTTTCCTTTGTTCTGGATCATAAACTTTCGAAG
5	fw_Ilu2_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGCAAACACCACAACGACAC
6	fw_Ilu2_+1_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGTCAAACACCACAACGACAC
7	fw_Ilu2_+2_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGGTCAAACACCACAACGACAC
8	fw_Ilu2_+3_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGCGACAAACACCACAACGACAC
9	fw_Ilu2_+4_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGATGACAAACACCACAACGACAC
10	fw_Ilu2_+5_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGCGATTCAAACACCACAACGACAC
11	fw_Ilu2_+6_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGACTCGGCAAACACCACAACGACAC
12	fw_Ilu2_+7_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGTCGAGATCAAACACCACAACGACAC
13	fw_Ilu2_+8_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGCAGTCTCACAAACACCACAACGACAC
14	fw_Ilu2_+9_spacer_AhdI	TCGTCGGCACGCGTCAGATGTGTATAAGAGACAGCTAGCACCAAACACCACAACGACAC
15	rev_AhdI_Renilla(13)_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTTCGAAGTCATGACTG
16	rev_AhdI_Renilla(13)_+1_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTTTCGAAGTCATGACTG
17	rev_AhdI_Renilla(13)_+2_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCTTTCGAAGTCATGACTG
18	rev_AhdI_Renilla(13)_+3_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGACTTTCGAAGTCATGACTG
19	rev_AhdI_Renilla(13)_+4_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGACTTTCGAAGTCATGACTG
20	rev_AhdI_Renilla(13)_+5_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGATTCTTTCGAAGTCATGACTG
21	rev_AhdI_Renilla(13)_+6_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTCGGCTTTCGAAGTCATGACTG
22	rev_AhdI_Renilla(13)_+7_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGAGATCTTTCGAAGTCATGACTG
23	rev_AhdI_Renilla(13)_+8_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGTCTCACTTTCGAAGTCATGACTG
24	rev_AhdI_Renilla(13)_+9_Ilu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCTAGCACCTTTCGAAGTCATGACTG
25	T7_CrPV1(357)	CAACAAATATTAATACGACTCACTATAGGTTCGATACCAAGAGCTGGTGTCTGAGGG
26	T7_IGR	CAACAAATATTAATACGACTCACTATAGGCAAAAATGTGATCTTGCTT
27	T7_stem_IGR	CAACAAATATTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCCAACCACAACCACCAAACACCACAACGACACTCTAAAAAGCAAAAATGTGATCTTGCTTG
28	T7_5'UTR_beta-globine	CAACAAATATTAATACGACTCACTATAGGACATTTGCTTCTGACACAACTGTGTTCACTAGC
29	T7_stem_5'UTR-beta-globine	CAACAAATATTAATACGACTCACTATAGGGAGTGGACTTCGGTCCACTCCCCAACCACAACCACCAAACACCACAACGACACTACATTTGCTTCTGACACAAC

2.2.3. Perspectives

La suite de ce travail consistera dans un premier temps à déterminer si les régions identifiées par l'analyse des données qui ne correspondent pas aux deux IRES déjà connus du CrPV sont des artefacts liés à la méthodologie de criblage, ou si ces éléments exercent une réelle influence sur la traduction et par conséquent sont de véritables éléments *cis*-régulateurs. Ceci sera réalisé à l'aide de systèmes rapporteurs et d'extraits de traduction acellulaires.

En parallèle, nous chercherons à améliorer la sélectivité de la méthode de criblage afin de réduire la sélection de faux positifs à l'issue de l'analyse des données. Quelques pistes ont déjà été explorées, comme réduire de dix fois la quantité d'ARN provenant de la banque utilisée pour l'assemblage des complexes, ou réaliser l'extraction des ARN seulement à partir des fractions du gradient contenant les polysomes. L'utilisation de la puromycine après l'assemblage des complexes de traduction a permis de mieux identifier les faux positifs sélectionnés en raison de leur fixation aspécifique sur les ribosomes. Une autre approche complémentaire serait de croiser les résultats de nos sélections avec des données de ribosome profiling *in vitro* qu'il s'agit d'obtenir à partir des complexes assemblés sur la banque d'ARN en présence d'homo-harringtonine. Ceci permettra d'identifier les sites d'initiation de la traduction sur un génome viral. La combinaison des deux jeux de données permettra de caractériser sans ambiguïté de nouveaux IRES en identifiant les fragments qui contiennent un site d'initiation de la traduction.

Une fois pleinement validée, cette méthode sera appliquée à d'autres génomes viraux pour lesquels l'implication d'IRES dans la traduction de l'ARN viral est inconnue. Dans l'immédiat, les génomes des virus de la Dengue et du Zika sont prêts à être criblés au laboratoire. Dans cette optique, nous avons déjà mis au point des extraits acellulaires de traduction préparés à partir de cellules biologiquement pertinentes, c'est-à-dire provenant d'organismes infectés par ces virus, à savoir le moustique tigre *Aedes aegypti* qui véhicule ces deux virus, et l'homme, pour qui ces deux virus sont pathogènes. En effet, des extraits de traduction *in vitro* actifs ont été préparés à partir de cellules embryonnaires de moustique tigre Aag2 (présentés dans les figures supplémentaires de cette partie) et de cellules embryonnaires humaines de rein HEK293FT (présentés dans les études de ce travail de thèse). Il sera intéressant de déterminer si les mêmes résultats sont obtenus avec des extraits issus de deux hôtes différents, pour lesquels l'infection par le virus n'a pas les mêmes conséquences. L'identification d'IRES active(s) dans un type d'extrait précis pourrait permettre de comprendre pourquoi le virus n'est pas pathogène pour le moustique alors qu'il peut être fatal pour l'homme. Par ailleurs, les extraits de cellules HEK293FT préparés en conditions de stress peuvent également s'avérer pertinents dans le cadre de cette partie puisque l'infection virale est considérée comme un stress cellulaire.

A plus long terme, cette approche pourra être utilisée pour l'identification d'éléments *cis*-régulateurs de la traduction dans les régions non traduites humaines. En ce sens, une méthode de préparation d'une banque de régions 5'UTR humaines à déjà été imaginée mais n'a pas encore été testée.

2.3. Etude de l'impact de la protéine NSP1 sur la traduction des ARNm cellulaires et sur la traduction de l'ARN génomique du SARS-CoV-2

2.3.1. Introduction du projet

La traduction sélective de l'ARN génomique du SARS-CoV-2 nécessite la coopération d'éléments *cis*- et *trans*- régulateurs au cours de l'infection.

Les coronavirus du syndrome respiratoire aigu sévère (SARS-CoV) sont des virus à ARN de polarité positive. L'ARN génomique du SARS-CoV-2 est un long génome d'environ 30 kb : les deux tiers de la phase codante permettent la synthèse de deux polyprotéines (ORF1a et ORF1ab) qui seront clivées en protéines non structurales (NSP1 à NSP16). Le tiers restant du génome code pour les protéines structurales. La traduction de l'ORF1ab nécessite un décalage -1 du cadre de lecture du ribosome lors de la traduction de l'ORF1a. La traduction de l'ORF2 ne se fait pas à partir de l'ARN génomique mais requiert la synthèse d'ARN sous-génomiques synthétisés au cours de l'infection. La protéine NSP1 est une des premières protéines virales produites lors de l'infection. Après le clivage protéolytique de la polyprotéine produite à partir de l'ORF1a, NSP1 se replie en un domaine N-terminal globulaire séparé du domaine C-terminal par un domaine « linker » très flexible. Des études structurales ont montré que le domaine C-terminal de NSP1 se replie en deux hélices α qui viennent se loger dans le tunnel par lequel passe l'ARN messager dans le ribosome. Ce phénomène est stabilisé par des interactions entre ces deux hélices et les protéines ribosomiques uS3 et uS8 ainsi qu'avec l'hélice h18 de l'ARNr 18S. La gêne stérique induite par la présence de NSP1 dans le canal d'entrée de l'ARNm n'est pas compatible avec la prise en charge d'un ARNm dans un complexe de pré-initiation 43S et de ce fait inhibe la traduction cellulaire.

Pourtant, la traduction de l'ARN génomique et des ARN sous-génomiques viraux se poursuit malgré la présence de NSP1 qui rend les ribosomes non fonctionnels. Dès lors, l'objectif de ce projet est de comprendre les mécanismes moléculaires qui permettent de maintenir la traduction des ARN viraux en présence de NSP1. Nous avons d'abord étudié la structure secondaire de la région 5'UTR du SARS-CoV-2 par des méthodes de sondage chimique et enzymatique, ce qui a permis de proposer un modèle solide de la structure secondaire de sa région 5'UTR. Ensuite, des expériences de traduction *in vitro* couplées à des analyses de spectrométrie de masse ont permis de démontrer que la tige-boucle SL1 de la 5'UTR de l'ARN génomique du SARS-CoV-2 est essentielle à sa traduction lors de l'infection virale, en présence de NSP1. Nous proposons un modèle dans lequel SL1, située à l'extrémité 5' de l'ARN viral, permet de déloger NSP1 de l'entrée du tunnel de l'ARNm, permettant la traduction effective du génome viral. De manière remarquable, la tige-boucle SL1 est aussi présente dans les régions 5'UTR de tous les ARN sous-génomiques du virus. Ainsi, notre modèle permet d'expliquer la traduction virale pour la synthèse des protéines non-structurales pendant les phases précoces de l'infection et la synthèse des protéines structurales en fin du programme infectieux.

Nous avons également étudié le couple NSP1/SL1 dans d'autres coronavirus. A l'aide d'études basées sur l'échange de résidus clés dans les molécules NSP1 et SL1 des virus SARS-CoV-1 et SARS-CoV-2, nous avons mis en évidence un phénomène de coévolution entre ces deux facteurs *trans*- et *cis*- régulateurs de la traduction. De manière générale, ce mécanisme semble également conservé dans d'autres bêta-coronavirus.

2.3.2. Détermination de la structure secondaire de la 5'UTR du SARS-CoV-2

2.3.3. La tige boucle SL1 de la 5'UTR de l'ARN génomique du SARS-CoV-2 est essentielle pour sa traduction lors de l'infection virale, en présence de NSP1

2.3.4. Mise en évidence de coévolutions des séquences des tiges-boucles SL1 et des protéines NSP1 dans les coronavirus

Les résultats de ces trois parties seront présentés sous la forme de trois articles publiés dans les journaux RNA Biology et RNA.

ARTICLE 1

Détermination de la structure secondaire de la région 5'UTR de l'ARN génomique du SARS-CoV-2

RNA BIOLOGY
https://doi.org/10.1080/15476286.2020.1814556

Taylor & Francis
Taylor & Francis Group

BRIEF COMMUNICATION

OPEN ACCESS Check for updates

Secondary structure of the SARS-CoV-2 5′-UTR

Zhichao Miao [a,b,c], Antonin Tidu[d], Gilbert Eriani [d], and Franck Martin [d*]

[a]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, UK; [b]Translational Research Institute of Brain and Brain-Like Intelligence and Department of Anesthesiology, Shanghai Fourth People's Hospital Affiliated to Tongji University School of Medicine, Shanghai, China; [c]Newcastle Fibrosis Research Group, Institute of Cellular Medicine, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK; [d]Architecture Et Réactivité De l'ARN, Université De Strasbourg, Institut De Biologie Moléculaire Et Cellulaire Du CNRS, Strasbourg, France

ABSTRACT

The SARS-CoV-2, a positive-sense single-stranded RNA Coronavirus, is a global threat to human health. Thus, understanding its life cycle mechanistically would be important to facilitate the design of antiviral drugs. A key aspect of viral progression is the synthesis of viral proteins by the ribosome of the human host. In Coronaviruses, this process is regulated by the viral 5′ and 3′ untranslated regions (UTRs), but the precise regulatory mechanism has not yet been well understood. In particular, the 5′-UTR of the viral genome is most likely involved in translation initiation of viral proteins. Here, we performed inline probing and RNase V1 probing to establish a model of the secondary structure of SARS-CoV-2 5′-UTR. We found that the 5′-UTR contains stable structures including a very stable four-way junction close to the AUG start codon. Sequence alignment analysis of SARS-CoV-2 variants 5′-UTRs revealed a highly conserved structure with few co-variations that confirmed our secondary structure model based on probing experiments.

ARTICLE HISTORY
Received 18 June 2020
Revised 22 July 2020
Accepted 19 August 2020

KEYWORDS
SARS-CoV-2; 5′-UTR; secondary structure; probing

Introduction

Coronaviruses are found to infect a large variety of animals and humans. Besides enteric diseases, they mainly cause severe respiratory defects sometimes leading to death [1]. The recently emerged SARS-CoV-2 belongs to the betacoronavirus genome, subgenus Sarbecovirus [2]. Its genome is a positive single-stranded RNA molecule (+)ssRNA (Coronaviridae Study Group of the International Committee on Taxonomy of Viruses, 2020). The genomic sequence of SARS-CoV-2 was determined at the end of 2019 [3]. The RNA genome is capped at the 5′ end and polyadenylated at the 3′ end. The genome encodes two long open reading frames (ORF1a and ORF1b) at the 5′ end and several ORFs that are expressed in the late phase of infection from subgenomic RNAs (sgRNAs) [4]. After cell entry, the translation of ORF1a and ORF1b from the whole ss(+) RNA are the first events of the infectious process. The translation of ORF1b requires a − 1 frameshifting event [5,6]. The polyprotein synthesized from ORF1a is processed into eleven non-structural proteins (NSP1-NSP11). The first one, NSP1 binds to the host small ribosomal subunit 40S and recruits a yet unidentified cellular nuclease that triggers the degradation of the host mRNAs, while viral RNA is being translated [7,8]. Thus, the virus specifically degrades the cellular mRNAs that are translated by the canonical cap-dependent translation mechanism. Other studies have shown that NSP1 is also able to prevent 48S ribosomal complex formation by another so far uncharacterized mechanism [8,9]. Interestingly, it has been shown that IRES (Internal Ribosome Entry Site)-mediated translation initiation of class III- and IV–IRES (*e.g.*, in Hepatitis C Virus (HCV) and Cricket paralysis virus (CrPV)) respectively, is immune to the NSP1 inhibition. Instead, the translation initiation driven by the encephalomyocarditis virus (EMCV) class II–IRES is efficiently inhibited by NSP1 [8]. The NSP1 binding site to the ribosomal 40S subunit has not yet been determined.

Since the SARS-CoV-2 genomic RNA is capped at the 5′ end, it is generally believed that its translation initiation is canonical and cap-dependent. However, two major observations provide hints that the translation mechanism of SARS-CoV-2 RNA is in fact mediated by an unconventional translation initiation mechanism rather than a canonical one. First, the secondary structure of the SARS-CoV-2 5′-UTR is likely to be complex in the proximity of the 5′ cap, based on the experimental SHAPE structure of the 5′-UTR of Mouse Hepatitis Virus (MHV), a Coronavirus belonging to the Embecovirus subgroup, and on related structural predictions of the 5′-UTR of the SARS-CoV Sarbecovirus [10]. RNA structures proximal to the cap are known to inhibit the recruitment of cap-binding translation factors (eIF4E and consequently eIF4F), thus indicating canonical cap-dependent translation improbable [11,12]. Second, after the translation of ORF1a, the rapidly produced NSP1 protein would shut down the canonical cap-dependent translation of cellular mRNAs. Yet, the

*CONTACT Franck Martin f.martin@ibmc-cnrs.unistra.fr Architecture Et Réactivité De l'ARN, Université De Strasbourg, Institut De Biologie Moléculaire Et Cellulaire Du CNRS, Strasbourg, France

Supplemental data for this article can be accessed here.

translation of SARS-CoV-2 proteins is not inhibited by NSP1. Indeed, class III- and class IV- IRES are immune to this NSP1-mediated inhibitory mechanism [8,13]. This suggests that translation initiation of SARS-CoV-2 may be less dependent on eIF4F. Previous structural studies from other coronavirus 5′-UTRs have shown that stable hairpin structures are found in the proximity of the cap structure. When these structures are present in the vicinity of the cap,

Figure 1. SARS-CoV-2 5′-UTR.
An RNA transcript containing the 301 5′ proximal nucleotides from SARS-CoV-2 is represented. The positions of the uORF and the N-terminal coding sequence of NSP1 are shown in green.

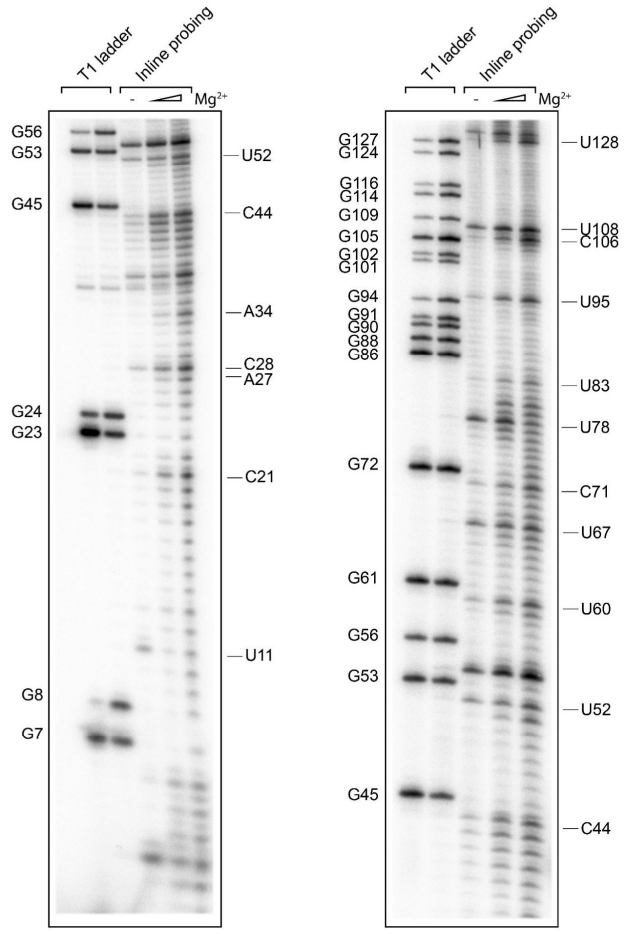


Figure 2. Inline probing of the SARS-CoV-2 5′-UTR from nucleotides 1 to 128.
Inline probing of the SARS-CoV-2 without Mg^{2+}, with 1 and 10 mM Mg^{2+}. The cuts are mapped using an RNase T1 denaturing ladder that cuts after G residues. The left panel shows the reactivity of nucleotides 1–56 and the right panel shows the reactivity of nucleotides 30–128.

translation initiation efficiency is significantly modulated by these secondary structures [11]. In both cap-dependent and cap-independent mechanisms, the secondary structure of the 5′-UTR is critical for translation initiation efficiency.

In order to better understand the translation initiation mechanism of viral translation during SARS-CoV-2 infection, the first step is to determine the secondary structure of the 5′-UTR. Here, we report the first experimental determination of SARS-CoV-2 5′-UTR structure using inline probing and RNAse V1 enzymatic probing.

Material and methods

Coronavirus complete genome sequence

The complete genome sequence of SARS-CoV-2 was downloaded from NCBI nucleotide database Genbank [14] MN908947.3.

Secondary structure probing

The RNA transcripts have been synthesized by *in vitro* transcription. The RNA transcripts were then separated on PAGE containing 8 M urea and purified by electro-elution using Bio Trap apparatus and Schleicher & Schuell membranes. The purified transcript was then ^{32}P-labelled by 5′ capping using the ScriptCap m^{7}G Capping System kit from CELLSCRIPT™. The RNA transcripts were probed directly after purification without any denaturation-renaturation step. Briefly, for inline probing, 50 000 cpm of radiolabelled RNA was incubated in 50 mM Tris-HCl pH 8.8, 100 mM KCl without $MgCl_2$ or with 1 or 10 mM $MgCl_2$ for 72 h at room temperature. The cuts in the RNA backbone were analysed on denaturating PAGE containing 8 M urea. For V1 probing, the RNA was incubated with serial dilutions of RNase V1 in order to have statistically one digestion cut per molecule for 10 min at room temperature as previously described [15]. The cuts were

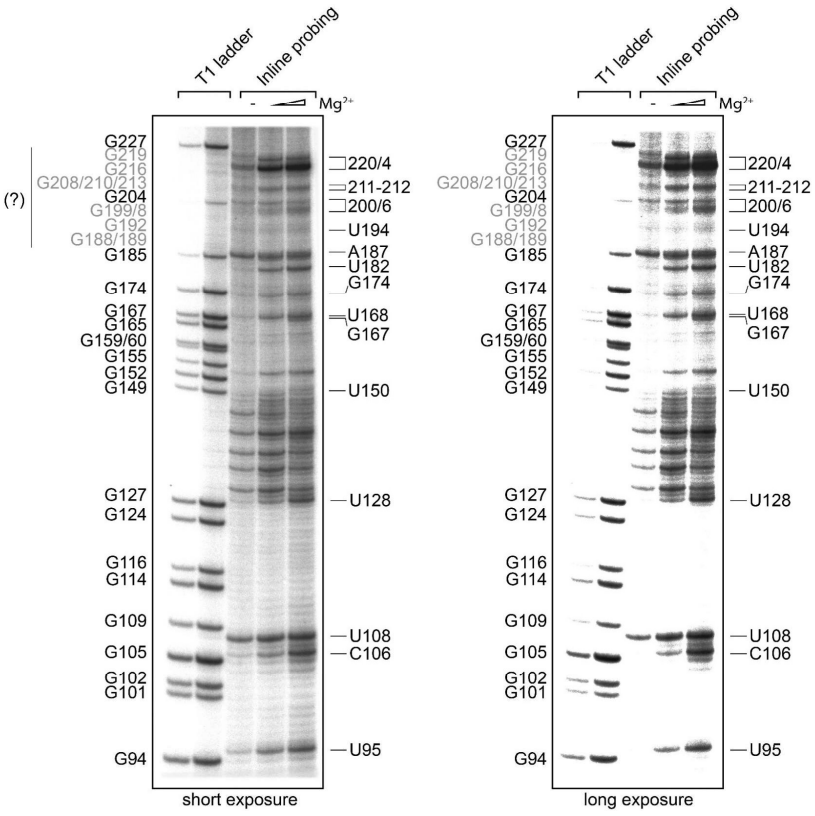


Figure 3. Inline probing of the SARS-CoV-2 5′-UTR from nucleotides 94 to 295.
(A) Inline probing of the SARS-CoV-2 without Mg^{2+}, with 1 and 10 mM Mg^{2+}. The cuts are mapped using an RNase T1 denaturing ladder that cuts after G residues. The left panel shows the reactivity of nucleotides 94–227 and the right panel shows the same gel with a longer exposure. **(B)** Inline probing of the SARS-CoV-2 without Mg^{2+}, with 1 and 10 mM Mg^{2+}. The upper panel shows the reactivity of nucleotides 165–262, and the lower panels show the reactivity of nucleotides 220–295 (short exposure on the left and long exposure on the right).

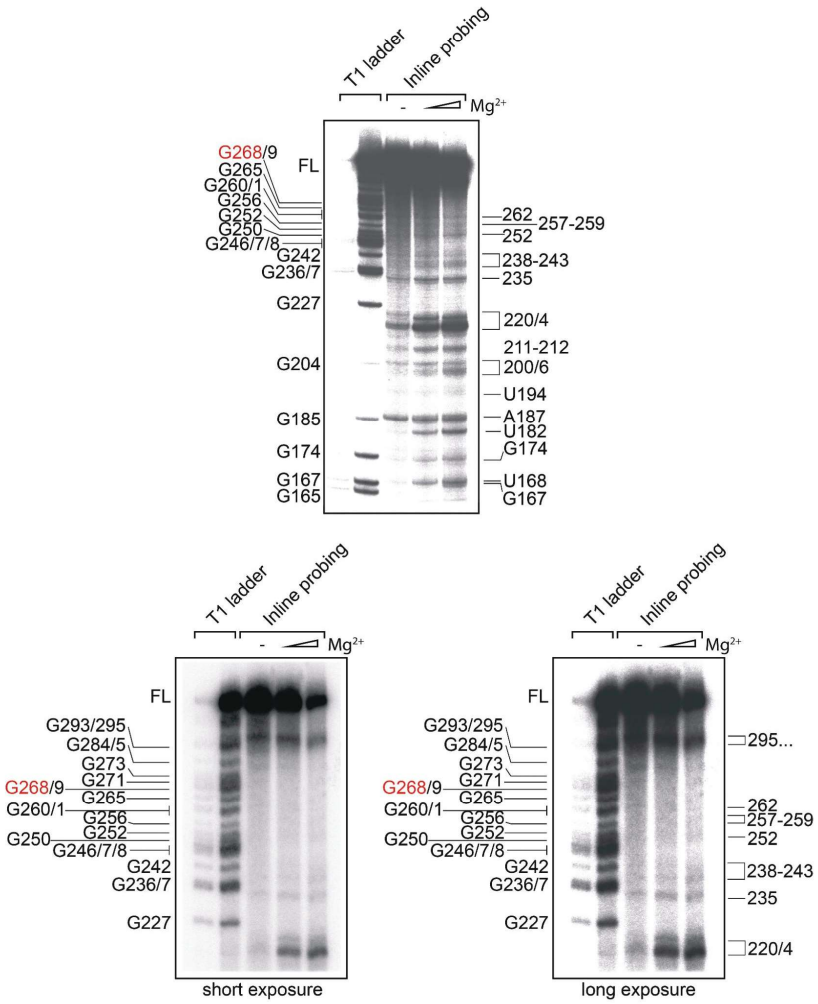


Figure 3. (continued).

mapped by an RNase T1 ladder performed in a denaturing buffer according to previously established protocol [16]. Each segment of the 5′-UTR has been probed at least twice and representative gels for all the parts of the 5′-UTR are shown in the figures and supplemental figures. The inline reactivities have been classified as 'accessible' or 'not accessible' for inline probing. For V1 probing, the reactivities are shown as 'weak' or 'strong' according to the band intensities.

Sequence alignments

Homologous sequences, most of which were from SARS-CoV-2, were retrieved by BLAST [17] search. Sequences were aligned with ClustalW [18] before the alignment-based prediction of RNAalifold [19]. Forna [20] and R2R [21] were used to visualize the secondary structures.

Results and discussion

SARS-CoV-2 5′-UTR inline probing

Using *in vitro* transcription, we synthesized a transcript encompassing nucleotides 1 to 301 from the SARS-CoV-2 variant (GenBank: MN908947.3). It contains the whole 5′-UTR and the sequence coding for the 12 N-terminal codons of NSP1 Fig. 1. Since the viral genomic RNA is capped at the 5′ end, we labelled the transcript at the 5′ end

with a radioactive m^7G cap. In order to determine the secondary structure of this transcript, we first performed inline probing in the absence and presence of 1 or 10 mM Mg^{2+}. We then analysed the cuts in the RNA backbone by migration on denaturing polyacrylamide gel electrophoresis. The cuts were mapped using an RNase T1 ladder performed in denaturing conditions. We analysed regions from nt 1 to 128 and 94 to 295 Figs. 2 and 3, respectively.

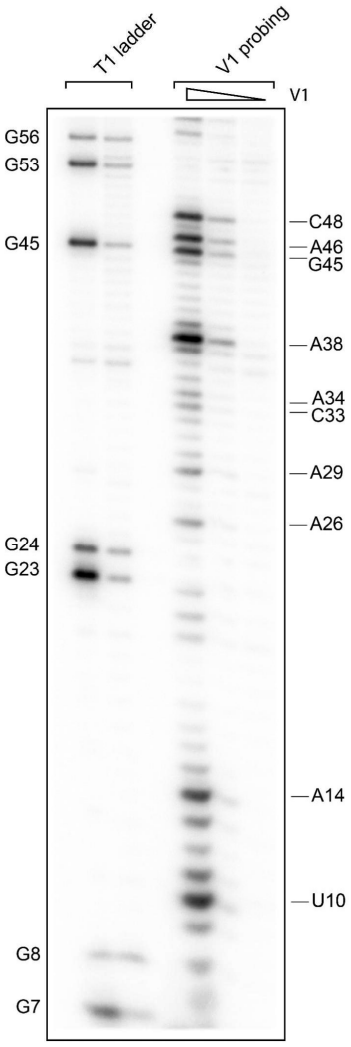


Figure 4. Enzymatic probing by RNase V1 of the SARS-CoV-2 5′-UTR from nucleotides 1 to 56.

Enzymatic probing of the SARS-CoV-2 which are performed by decreasing amount of RNase V1. The cuts are mapped using an RNase T1 denaturing ladder that cuts after G residues.

Interestingly, RNase T1 digestion of G residues from 188 to 219 was inefficient even though digestion was performed in denaturing conditions (Fig. 3, right panel). This indicates that these residues are embedded in a highly stable structural region that is still efficiently folded in denaturing conditions thereby preventing the access of RNase T1. Inline probing with and without Mg^{2+} allowed us to map the structurally accessible regions of the RNA that generally correspond to single-stranded regions. Using this method, we could also detect inaccessible areas that are potentially forming base pairs.

In order to confirm these putative stems, we performed RNase V1 probing. The V1 enzyme specifically cuts in stem regions Figs. 4 and 5. We confirmed that most of the inaccessible regions by inline probing do actually correspond to areas containing base pairs. Altogether, these data allowed us to establish a solid model of the 2D

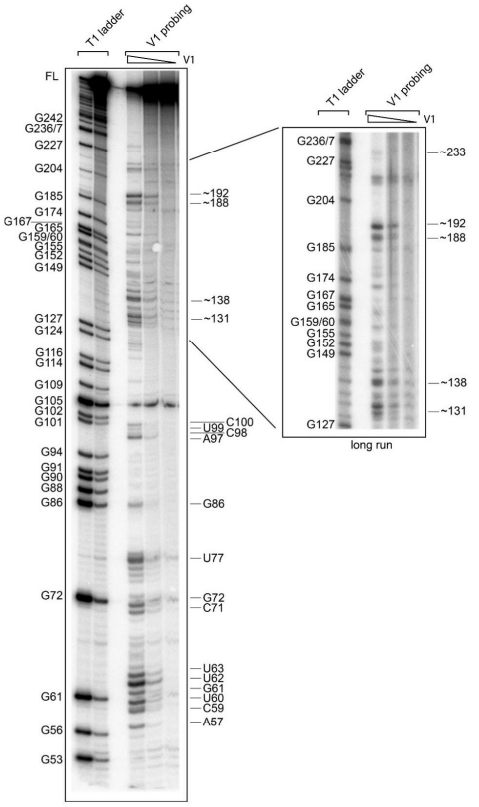


Figure 5. Enzymatic probing by RNase V1 of the SARS-CoV-2 5′-UTR from nucleotides 105 to 295.

Enzymatic probing of the SARS-CoV-2 which are performed by decreasing amount of RNase V1. The cuts are mapped using an RNase T1 denaturing ladder that cuts after G residues. The panel shows the reactivity of nucleotides 53–295. The right panel shows a long run of the same samples to improve the resolution of the T1 ladder 127 to 236.

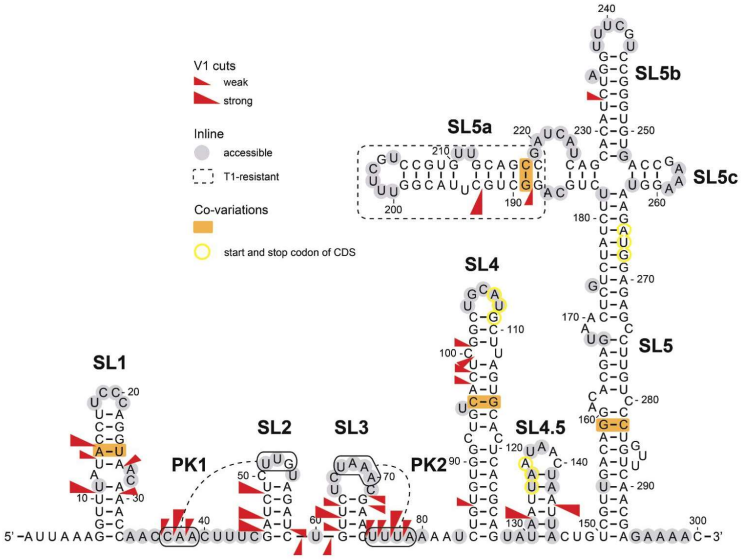


Figure 6. Secondary structural model of SARS-CoV-2 5′-UTR.

The model of the 5′-UTR is represented based on Inline probing and enzymatic probing by RNase V1. Reactivity determined by inline probing experiments is shown in grey and V1 cuts are shown by red triangles. A dashed line boxes the area that is resistant to RNase T1 digestion in denaturing conditions. The positions of the uAUG and the AUG codons are highlighted in yellow.

structure of the whole SARS-CoV-2 5′-UTR Fig. 6. The 5′-UTR is highly structured with a few accessible bulges and loops. It contains five simple hairpin structures that were named SL1, SL2, SL3, SL4 and SL5 in good agreement with bioinformatic secondary structure predictions for the SARS-CoV-2 [22] and also from other coronaviruses [23]. Our model is also highly similar to the models obtained by probing of the whole SARS-CoV-2 genome *in vitro* [24] and *in vivo* [25,26]. However, we found an additional hairpin located between SL4 and SL5 that we named SL4.5. As predicted, SL1 is located close to the 5′ extremity. It has been proposed that the low overall stability of SL1, due to a high proportion of A-U and U-A base pairs, is important for replication in Mouse Hepatitis Virus MHV [27]. The loop of SL1 is not conserved in SARS-CoV-2 variants and the two bulged nucleotides in the middle of SL1 can be involved in base pairs as observed in some variants suggesting that the loop and the bulge of SL1 are not required for efficient viral propagation Fig. 7. The SL2 hairpin domain is comparable to the SL3 domain in Bovine coronavirus (BCoV), which is known to form a hairpin structure according to NMR spectroscopy [28] as well as enzymatic probing [29,30]. This domain is expected to be involved in the replication complex formation Fig. 8. The loop (CUUGY) of SL2 is important for Mouse Hepatitis Virus replication [28]. It contains a U-turn like motif. In contrast, the SARS-CoV SL2 rather adopts a typical tetraloop structure [31]. This is in good agreement with our inline reactivity profile of the SL2 loop since C50 and G53 are not accessible in the loop but the U54 is highly accessible, a well-characterized feature of a tetraloop structure [31]. Hairpin SL3 is known to encompass the leader TRS (TRS-L) sequence as previously observed for group IIb coronavirus [23] Fig. 8. The SL4 structure is a relatively stable and long hairpin and three base pairs in the stem region have been described to contain covariations (R_{90}-Y_{121}, R_{97}-U_{115} and G_{101}-Y_{111}), indicating structural conservation. SL4 contains the start codon of a uORF that is present in all coronaviruses although there is no evidence of translation of this uORF so far [23]. The integrity of the upper part of SL4 is important for replication of BCoV [29] Fig. 9. It has also been proposed that SL4 is involved in the synthesis of subgenomic RNA fragments [23]. In the 3′ part, the 5′-UTR contains a more complex structure named SL5 that comprises a four-way junction formed by SL5a, SL5b and SL5c. This four-way junction is found in all the coronavirus 5′-UTRs that have been probed so far [23,32–35] Fig. 9. Importantly, in the SARS-CoV-2, the NSP1 AUG start codon is part of the SL5. We observed the same reactivity pattern in the identical loops from SL5a and SL5b, the inline reactivities are in good agreement with classical U-turns [36]. In contrast, SL5c contains a typical GNRA loop as previously observed in group IIb coronaviruses [23]. The loops of SL5 are most likely involved in RNA packaging as previously suggested [37]. These stems are separated by short single stranded regions that are generally

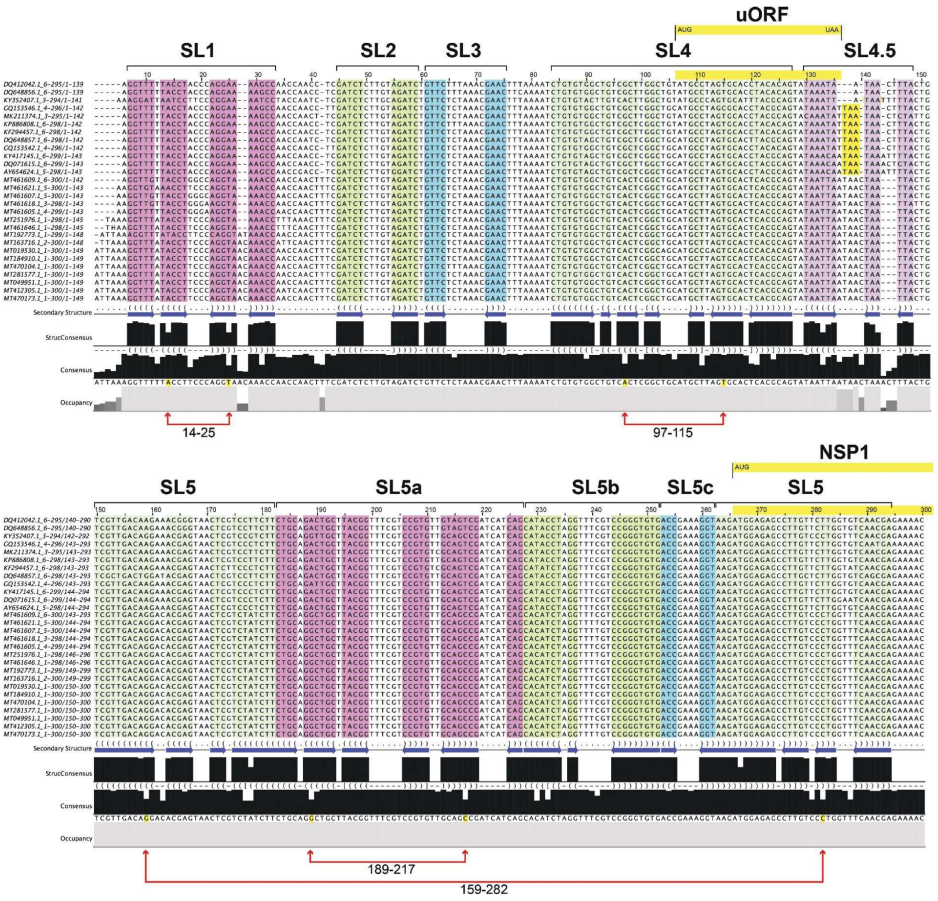


Figure 7. Sequence alignments of 5′-UTRs from SARS-CoV-2 variants.

The sequence alignments of 5′-UTRs from 28 SARS-CoV-2 variants are presented in two panels from nt #1 to 151 (top) and from nt 1#50 to 300 (bottom). The stems of SL1 to SL5 are underlined by distinct colours. The positions of uORF and the beginning of NSP1 ORF are indicated in light green on top of the sequences. In several variants, the stop codon of the uORF is shifted from one codon (also shown in green). Covariations in stems are indicated by red arrows and the corresponding nucleotides are shown in yellow on the consensus sequence.

accessible except for two short regions (A_{34}-C_{40}) and (C_{75}-A_{80}) that are accessible in our inline experiments but also contain V1 cuts indicating the potential forming of base pairs. These apparently contradictory results can be explained as follows: these regions are dynamic and can be considered as single-stranded regions in a so-called 'open' state or double-stranded in a 'closed' state. Interestingly, the sequences can fold into putative pseudo-knots structures with loop of SL2 and SL3 that we named, respectively, PK1 and PK2. The position of SL1 close to the m^7G cap might interfere with the binding of eIF4F complex thereby preventing canonical cap-dependent translation [12]. This observation is important for some SARS-CoV-2 variants that are shorter in which the 5′ extremity is located just one nucleotide upstream of SL1 Fig. 7. We also probed a truncated version lacking the 5′ proximal half of the 5-UTR (Δ105). Interestingly, the folding of SL4 and SL5 in this shorter version is identical to the folding found in the full-length 5′-UTR (Supplemental Figure 1). This also allowed us to confirm the highly stable structure of SL5abc three-way junction. Moreover, these experiments demonstrate that SL5 can fold without nt 1–105 which suggests the existence of independent motifs in the 5′-UTR that do not interact with each other. Importantly, sequence alignments of recently sequenced SARS-CoV-2 variants also enabled the discovery of co-variations in SL1, SL4 SL5 and SL5a thereby validating our secondary structure model for these structural regions Fig. 7. SL5a is

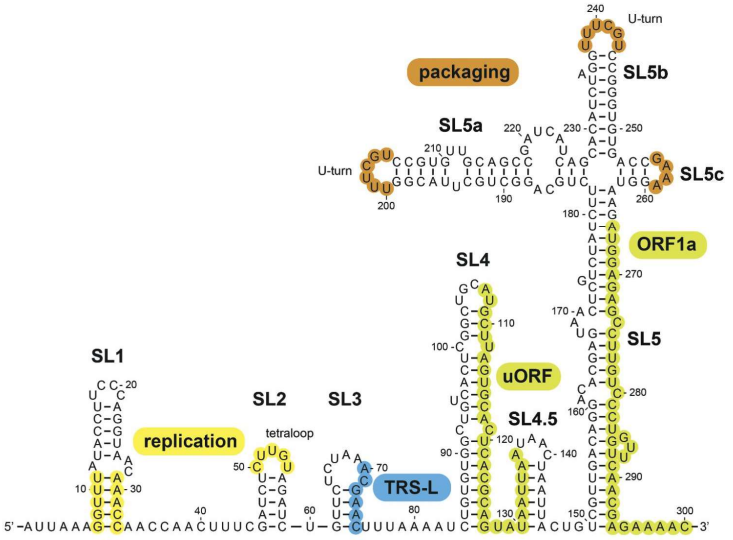


Figure 8. Functional motifs in the 5′-UTR of SARS-CoV-2.
The previously described functional motifs for replication, transcription, viral packaging and translation are shown on our secondary structural model. These motifs have been characterized in the coronavirus family.

a structurally stable region that is not cut by RNase T1 in denaturing conditions. This hairpin contains 10 G-C base pairs and its minimal free energy is predicted to be −17.30 kcal.mol^{-1}, probably explaining its resistance to RNase T1 digestion even in denaturing conditions. The 5′-UTR also contains an upstream opened reading frame (uORF) that is overlapping with SL4 and SL4.5. The location and the sequence of uAUG are absolutely conserved. On the contrary, the UAA stop codon is mutated in a few variants but another in-frame UAA stop is present immediately downstream, implying that the uORF is present and conserved in all variants Fig. 7. The NSP1 AUG start codon is embedded in the four-way junction structure. For an efficient translation initiation, the sequences surrounding the AUG start codon have to be unfolded. The mechanism used by the virus is still unknown and an important issue to investigate is the role and the putative function of the stable SL5 structure (SL5a, SL5b, SL5c) in the translation of the SARS-CoV-2 polyprotein. Although the viral genome is capped at its 5′ end, the translation initiation mechanism used to locate the AUG start codon in SL5 remains elusive. The presence of the 5′ m^{7}G cap and hairpins SL1, SL2, SL3, SL4 and SL5 suggest that a canonical cap-dependent scanning mechanism would require the eIF4A helicase [38,39]. On the other hand, the fact that the AUG start codon is located in the vicinity and downstream of a four-way junction structure is reminiscent of similar structures found in the HCV IRES [40] (and references therein). Indeed, IRES elements are typically highly structured RNA motifs [40]. In addition, the 5′-UTR contains a uORF that is conserved in SARS-CoV-2 variants, which indicates that it can be translated by the host ribosome. The use of uORF is another way of translation regulation [41,42]. It is possible that the viral translation may use all these mechanisms at distinct stages of the infectious process. We hope that our investigations of the 5′-UTR structure in SARS-CoV-2 translation will pave the way to further studies in understanding its functions in the viral infection life cycle.

Acknowledgments

ZM is supported by Single Cell Gene Expression Atlas grant from the Wellcome Trust (108437/Z/15/Z). FM, GE and AT are funded by 'Agence Nationale pour la Recherche' (ANR-17-CE12-0025-01, ANR-17-CE11-0024, ANR-20-COVI-0078), by 'Fondation pour la Recherche Médicale' (project CoronaIRES), by University of Strasbourg and by the 'Centre National de la Recherche Scientifique'. We gratefully thank Eric Westhof for helpful discussion.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Agence Nationale de la Recherche (FR) [ANR-17-CE12-0025-01]; Agence Nationale de la Recherche (FR) [ANR-17-CE11-0024]; Welcome Trust [108437/Z/15/Z]; Fondation pour la

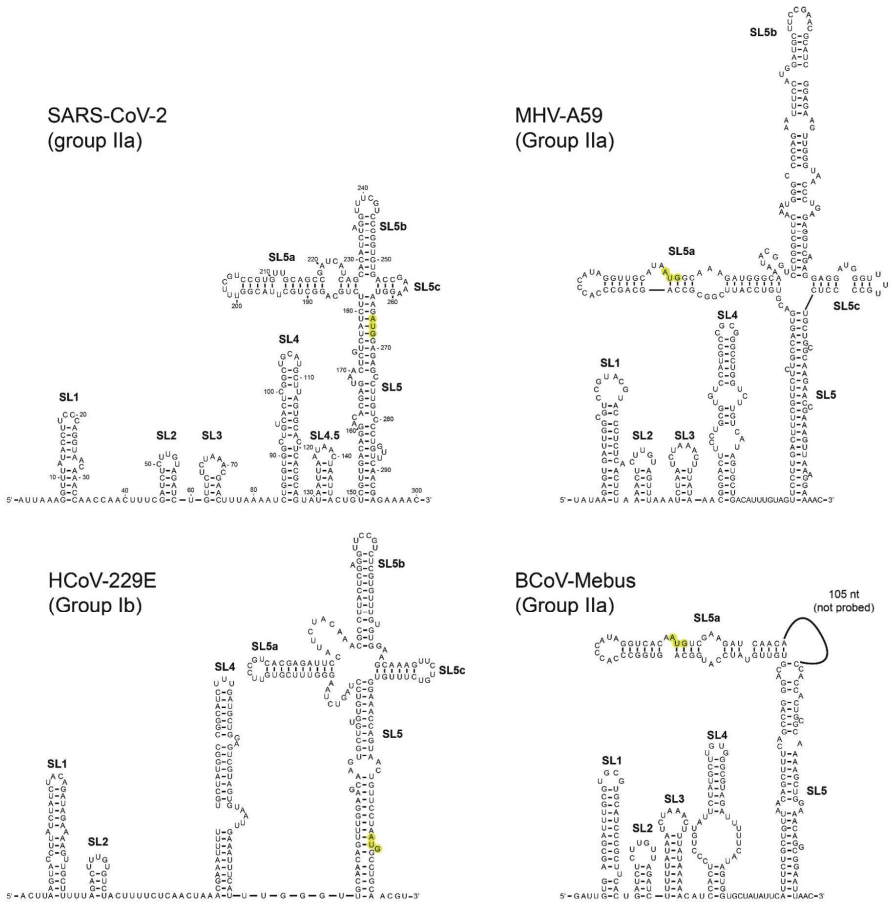


Figure 9. Comparison of the 2D structure of SARS-CoV-2 5′UTR with other probed coronavirus 5′-UTRs.

Our model for the SARS-CoV-2 5′-UTR is compared with other coronavirus 5′-UTR that have been determined by experimental probing. The 5′-UTR secondary structures from Mouse Hepatitis Virus (MHV) [34], Human coronavirus (HCoV) [35] and Bovine coronavirus (B-CoV) [23,32,33] are shown. The position of the NSP1 AUG start codon is highlighted in green.

Recherche Médicale [CoronaIRES]; Agence Nationale de la Recherche (FR) [ANR-20-COVI-0078].

ORCID

Zhichao Miao http://orcid.org/0000-0002-5777-9815
Gilbert Eriani http://orcid.org/0000-0002-8518-4675
Franck Martin http://orcid.org/0000-0001-9724-4025

References

[1] Weiss SR, Navas-Martin S. Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. Microbiol Mol Biol Rev. 2005;4(4):635–664.
[2] Andersen KG, Rambaut A, Lipkin WI, et al. The proximal origin of SARS-CoV-2. Nat Med. 2020;26(4):450–452.
[3] Wu A, Peng Y, Huang B. et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe. 2020;27(3):325–328. .
[4] Brian DA, Baric RS. Coronavirus genome structure and replication. Curr Top Microbiol Immunol. 2005;287:1–30.
[5] Plant EP, Rakauskaite R, Taylor DR, et al. Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. J Virol. 2010;84(9):4330–4340.
[6] Plant EP, Sims AC, Baric RS, et al. Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. Viruses. 2013;5(1):279–294.
[7] Kamitani W, Narayanan K, Huang C, et al. Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. Proc Natl Acad Sci U S A. 2006;103(34):12885–12890.
[8] Lokugamage KG, Narayanan K, Huang C, et al. Severe acute respiratory syndrome coronavirus protein nsp1 is a novel eukaryotic translation inhibitor that represses multiple steps of translation initiation. J Virol. 2012;86(24):13598–13608.

[9] Narayanan K, Ramirez SI, Lokugamage KG, et al. Coronavirus nonstructural protein 1: common and distinct functions in the regulation of host and viral gene expression. Virus Res. 2015;202:89–100.
[10] Yang D, Leibowitz JL. The structure and functions of coronavirus genomic 3′ and 5′ ends. Virus Res. 2015;206:120–133.
[11] Babendure JR, Babendure JL, Ding JH, et al. Control of mammalian translation by mRNA structure near caps. RNA. 2006;5 (5):851–861.
[12] Pestova TV, Kolupaeva VG. The roles of individual eukaryotic translation initiation factors in ribosomal scanning and initiation codon selection. Genes Dev. 2002;16(22):2906–2922.
[13] Kamitani W, Huang C, Narayanan K, et al. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. Nat Struct Mol Biol. 2009;16 (11):1134–1140.
[14] Clark K, Karsch-Mizrachi I, Lipman DJ, et al. Genbank. Nucleic Acids Res. 2016;44(D1):D67–72.
[15] Jaeger S, Martin F, Rudinger-Thirion J, et al. Binding of human SLBP on the 3′-UTR of histone precursor H4-12 mRNA induces structural rearrangements that enable U7 snRNA anchoring. Nucleic Acids Res. 2006;34(17):4987–4995.
[16] Peattie DA, Gilbert W. Chemical probes for higher-order structure in RNA. Proc Natl Acad Sci U S A. 1980;77(8):4679–4682.
[17] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410.
[18] Thompson JD, Higgins DG, Gibson TJ, et al. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673–4680.
[19] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.
[20] Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. Bioinformatics. 2015;31(20):3377–3379.
[21] Weinberg Z, Breaker RR. R2R - software to speed the depiction of aesthetic consensus RNA secondary structures. BMC Bioinformatics. 2011;12(1):3.
[22] Rangan R, Zheludev IN, Das R. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. RNA. 2020;26(8):937–959.
[23] Chen SC, Olsthoorn RCL. Group-specific structural features of the 5′-proximal sequences of coronavirus genomic RNAs. Virology. 2010;401(1):29–41.
[24] Manfredonia I, Nithin C, Ponce-Salvatierra A. et al. Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. bioRxiv 2020. 2020;06(15):151647.
[25] Lan TCT, Allan MF, Malsick L, et al. Structure of the full SARS-CoV-2 RNA genome in infected cells. bioRxiv 2020. 2020;2020(6):29.178343.
[26] Huston NC, Wan H, Tavares R de CA, et al. Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. bioRxiv 2020. 2020;7 (10):197079.
[27] Li L, Kang H, Liu P, et al. Structural lability in stem-loop 1 drives a 5′ UTR-3′ UTR interaction in coronavirus replication. J Mol Biol. 2008;377(3):790–803.
[28] Liu P, Li L, Millership JJ, et al. A U-turn motif-containing stem-loop in the coronavirus 5′ untranslated region plays a functional role in replication. RNA. 2007;13(5):763–780.
[29] Raman S, Bouma P, Williams GD, et al. Stem-loop III in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. J Virol. 2003;77 (12):6720–6730.
[30] Raman S, Brian DA. Stem-loop IV in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. J Virol. 2005;79(19):12434–12446.
[31] Liu P, Li L, Keane SC, Yang D, Leibowitz JL, Giedroc DP. Mouse Hepatitis Virus Stem-Loop 2 Adopts a uYNMG(U)a-Like Tetraloop Structure That Is Highly Functionally Tolerant of Base Substitutions. J Virol. 2009;83(23):12084–12093.
[32] Guan B-J, Wu H-Y, Brian DA. An Optimal cis-Replication Stem-Loop IV in the 5′ Untranslated Region of the Mouse Coronavirus Genome Extends 16 Nucleotides into Open Reading Frame 1. J Virol. 2011;85(11):5593–5605.
[33] Guan B-J, Su Y-P, Wu H-Y, Brian DA. Genetic Evidence of a Long-Range RNA-RNA Interaction between the Genomic 5′ Untranslated Region and the Nonstructural Protein 1 Coding Region in Murine and Bovine Coronaviruses. J Virol. 2012;86 (8):4631–4643.
[34] Yang D, Liu P, Wudeck E V, Giedroc DP, Leibowitz JL. Shape analysis of the rna secondary structure of the mouse hepatitis virus 5′ untranslated region and n-terminal nsp1 coding sequences. Virology. 2015;475:15–27.
[35] Madhugiri R, Karl N, Petersen D, Lamkiewicz K, Fricke M, Wend U, Scheuer R, Marz M, Ziebuhr J. Structural and functional conservation of cis-acting RNA elements in coronavirus 5′-terminal genome regions. Virology. 2018;517:44–55.
[36] Gutell RR, Cannone JJ, Konings D, et al. Predicting U-turns ribosomal RNA with comparative sequence analysis. J Mol Biol. 2000;300 (4):791–803.
[37] Escors D, Izeta A, Capiscol C, et al. Transmissible gastroenteritis coronavirus packaging signal is located at the 5′ end of the virus genome. J Virol. 2003;77(14):7890–7902.
[38] Boussemart L, Malka-Mahieu H, Girault I. et al. eIF4F is a nexus of resistance to anti-BRAF and anti-MEK cancer therapies. Nature. 2014;513(7516):105–109. .
[39] Hershey JWB, Sonenberg N, Mathews MB. Principles of translational control. Cold Spring Harb Perspect Biol. 2019;11(9): a032607.
[40] Mailliot J, Martin F. Viral internal ribosomal entry sites: four classes for one goal. Wiley Interdiscip. Rev. 9. RNA2018. e1458. doi: 10.1002/wrna.1458
[41] Leppek K, Das R, Barna M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018;19:158–174.
[42] Orr MW, Mao Y, Storz G, et al. Fs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. 2020;48(3):1029–1042.

ARTICLE 2

La tige boucle SL1 de la région 5'UTR de l'ARN génomique du SARS-CoV-2 est essentielle pour sa traduction lors de l'infection virale, en présence de NSP1

The viral protein NSP1 acts as a ribosome gatekeeper for shutting down host translation and fostering SARS-CoV-2 translation

ANTONIN TIDU,[1] AURÉLIE JANVIER,[1] LAURE SCHAEFFER,[1] PIOTR SOSNOWSKI,[1] LAURIANE KUHN,[2] PHILIPPE HAMMANN,[2] ERIC WESTHOF,[1] GILBERT ERIANI,[1] and FRANCK MARTIN[1]

[1]Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, "Architecture et Réactivité de l'ARN" CNRS UPR9002, F-67084 Strasbourg, France

[2]Institut de Biologie Moléculaire et Cellulaire, Plateforme Protéomique Strasbourg—Esplanade, CNRS FRC1589, Université de Strasbourg, F-67084 Strasbourg, France

ABSTRACT

SARS-CoV-2 coronavirus is responsible for the Covid-19 pandemic. In the early phase of infection, the single-strand positive RNA genome is translated into nonstructural proteins (NSP). One of the first proteins produced during viral infection, NSP1, binds to the host ribosome and blocks the mRNA entry channel. This triggers translation inhibition of cellular translation. Despite the presence of NSP1 on the ribosome, viral translation proceeds, however. The molecular mechanism of the so-called viral evasion to NSP1 inhibition remains elusive. Here, we confirm that viral translation is maintained in the presence of NSP1 and we show that the evasion to NSP1-inhibition is mediated by the *cis*-acting RNA hairpin SL1 in the 5′UTR of SARS-CoV-2. Only the apical part of SL1 is required for viral translation. We further show that NSP1 remains bound on the ribosome during viral translation. We suggest that the interaction between NSP1 and SL1 frees the mRNA accommodation channel while maintaining NSP1 bound to the ribosome. Thus, NSP1 acts as a ribosome gatekeeper, shutting down host translation and fostering SARS-CoV-2 translation in the presence of the SL1 5′UTR hairpin. SL1 is also present and necessary for translation of subgenomic RNAs in the late phase of the infectious program. Consequently, therapeutic strategies targeting SL1 should affect viral translation at early and late stages of infection. Therefore, SL1 might be seen as a genuine "Achilles heel" of the virus.

Keywords: SARS-CoV-2; NSP1; SL1; 5′UTR; translation; ribosome

INTRODUCTION

The SARS-CoV-2, of the beta-coronavirus family, recently emerged as responsible for the Covid-19 world pandemic (Andersen et al. 2020; Zhou et al. 2020). Its genome is a positive single strand RNA molecule containing 29,903 nt entirely sequenced at the end of 2019 (Chan et al. 2020; Lu et al. 2020). The viral genomic RNAs of coronaviruses are capped at their 5′ end and polyadenylated at their 3′ end (Nakagawa et al. 2016; Hartenian et al. 2020). After entry into the infected cell, the viral genome from SARS-CoV-2 hijacks the host translation machinery in order to produce the viral proteins required for the viral infectious program and the production of novel viral particles (Hartenian et al. 2020). Like many viruses, SARS-CoV-2 orchestrates viral translation concomitantly with the specific shutdown of cellular mRNA translation. The goal of this silencing is dual. First, general cellular translation inhibition generates the large pool of ribosomes necessary to ensure efficient and massive synthesis of viral proteins. Interestingly, the cellular mRNAs coding for protein components of the translational machinery, such as ribosomal proteins and translation factors, are preserved from the overall translation inhibition presumably in order to maintain a functional translational machinery during viral translation (Rao et al. 2020). Secondly, the cellular translation silencing inhibits more specifically mRNA subsets that are involved in cellular immune responses to viral infection.

In coronaviruses, viral translation begins with the expression from ORF1a that is translated into a polyprotein further processed by proteolytic cleavages to produce nonstructural proteins (NSP) involved in the multiple steps

Corresponding author: f.martin@ibmc-cnrs.unistra.fr

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.078121.120. Freely available online through the *RNA* Open Access option.

of the general viral infectious program (Masters 2006). The amino-terminal proximal protein NSP1 is one of the first viral proteins that is produced at the onset of the infectious program. NSP1 is required for efficient host cellular translation inhibition. For example, in SARS-CoV, NSP1 first recruits a yet unidentified cellular endonuclease that promotes specific mRNA degradation on translated mRNAs (Kamitani et al. 2006, 2009). These cleavages occur on the ribosome during translation elongation. Importantly, viral mRNA transcripts are resistant to NSP1-mediated cleavages (Huang et al. 2011a). Secondly, the SARS-CoV NSP1 prevents translation initiation by interfering with the preinitiation complex formation at multiple steps (Lokugamage et al. 2012). The SARS-CoV NSP1 is thus directly responsible for general translation inhibition (Narayanan et al. 2008; Tohya et al. 2009; Huang et al. 2011b). Similarly, the NSP1 protein from Mouse Hepatitis Virus (MHV), another member of the beta-coronavirus family, is also important for cellular translation inhibition (Lei et al. 2013). However, the impact of NSP1 on translation is stronger on specific mRNAs. Among the targets of NSP1, mRNA subsets involved in specific cellular immune responses are primarily shut off. Thus, NSP1 suppresses type I interferon responses during infection by SARS-CoV (Narayanan et al. 2008) and SARS-CoV-2 (Lei et al. 2020; Xia et al. 2020). The SARS-CoV-2 viral protein NSP1 binds to the host 40S ribosomal subunit with high affinity and a K_d in the nanomolar range (Lapointe et al. 2020). Actually, NSP1 proteins from SARS-CoV and SARS-CoV-2 are highly homologous. The amino-terminal domain of SARS-CoV was determined by NMR (Almeida et al. 2007). The carboxy-terminal domain of SARS-CoV-2 NSP1 contains an intrinsically disordered domain from residues 130 to 180 (Kumar et al. 2020). However, when bound to the ribosome, the SARS-CoV-2 NSP1 carboxy-terminal domain is folded and binds tightly to the mRNA entry channel (Schubert et al. 2020; Thoms et al. 2020). NSP1 carboxy-terminal residues from 148 to 180 interact with ribosomal proteins uS3 and uS5 and with helix h18 from the 18S rRNA (Schubert et al. 2020; Thoms et al. 2020). Interestingly, the SARS-CoV-2 NSP1 binding site overlaps with the binding sites of the initiation factors eIF1 and eIF3j (Lapointe et al. 2020) and, thus, the binding of NSP1 to the 40S prevents the formation of the 48S preinitiation complex necessary for efficient translation (Brito Querido et al. 2020). Single molecule approaches have shown that SARS-CoV-2 NSP1 is competing with the mRNA in the mRNA channel (Lapointe et al. 2020).

Recently, it has been shown that the amino-terminal domain of SARS-CoV-2 NSP1 is required for viral translation by NSP1-bound ribosomes (Shi et al. 2020). The linker length between the amino- and carboxy-terminal domains is critical for viral translation (Shi et al. 2020). Beside the positive single strand genomic RNA, nine subgenomic RNAs (S, 3a, E, M, 6, 7a, 7b, 8, and N) are produced during the late phase of the infection by SARS-CoV-2 (Kim et al. 2020). Similar to other coronaviruses, all the viral transcripts are capped at their 5′ end with the first two nucleotides being ribose methylated and polyadenylated at their 3′ end (Yogo et al. 1977; Lai and Stohlman 1981). In the case of SARS-CoV-2, the median length of the polyA tail on viral transcripts is 47 A residues (Kim et al. 2020). In coronaviruses, all viral transcripts contain the common so-called 5′ leader sequence (nucleotides 1 to 75) that forms the hairpins SL1, SL2, and SL3 (Sola et al. 2015). This is also the case for SARS-CoV-2 (Kim et al. 2020; Miao et al. 2020).

Here we show that SARS-CoV-2 viral translation is evading NSP1-mediated inhibition because the viral transcripts, genomic and subgenomic RNAs, all contain a specific region of the leader sequence. More specifically, the sole hairpin SL1 is promoting NSP1 evasion by acting on the NSP1 carboxy-terminal domain in order to enable viral RNAs accommodation in the ribosome for their translation. The interaction between the SL1 RNA hairpin and the NSP1 carboxy-terminal domain occurs while NSP1 remains bound on the ribosome. Therefore, NSP1 acts as a ribosome gatekeeper to impair cellular translation and specifically promote viral translation.

RESULTS

The evasion from NSP1 inhibition is due to *cis*-acting elements located in the SARS-CoV-2 5′UTR

To measure the impact of the SARS-CoV-2 5′UTR on viral translation, we inserted in a reporter construct the 5′UTR upstream of the *Renilla* luciferase coding sequence. As a control, a similar reporter containing the EMCV IRES was used. Using rabbit reticulocyte lysates (RRL), we measured translation efficiency of *Renilla* luciferase in the absence and presence of increasing concentrations of recombinant SARS-CoV-2 NSP1 (Fig. 1). Translation efficiency of both constructs is reduced by NSP1. However, translation is significantly less affected with the SARS-CoV-2 5′UTR construct, indicating that the SARS-CoV-2 5′UTR allows evasion from NSP1-mediated inhibition. This is in good agreement with previous studies that showed that NSP1 is indeed inhibiting EMCV-driven translation but not SARS-CoV translation (Lokugamage et al. 2012). Moreover, since we are using 5′ labeled capped RNAs, the fact that we monitor the formation of radioactive translation complexes in the presence of NSP1 demonstrates that the SARS-CoV-2 mRNA is not degraded. To investigate further the evasion of SARS-CoV-2 viral translation from NSP1 inhibition, we analyzed the formation of the ribosomal preinitiation complex by fractionation on sucrose gradients. The RNA transcripts were radiolabeled at their 5′ ends and incubated in RRL. We used a minimal RNA containing the whole SARS-CoV-2 5′UTR and the first 12 codons of the NSP1 coding sequence (nt 1–300) fused to a minimal portion of

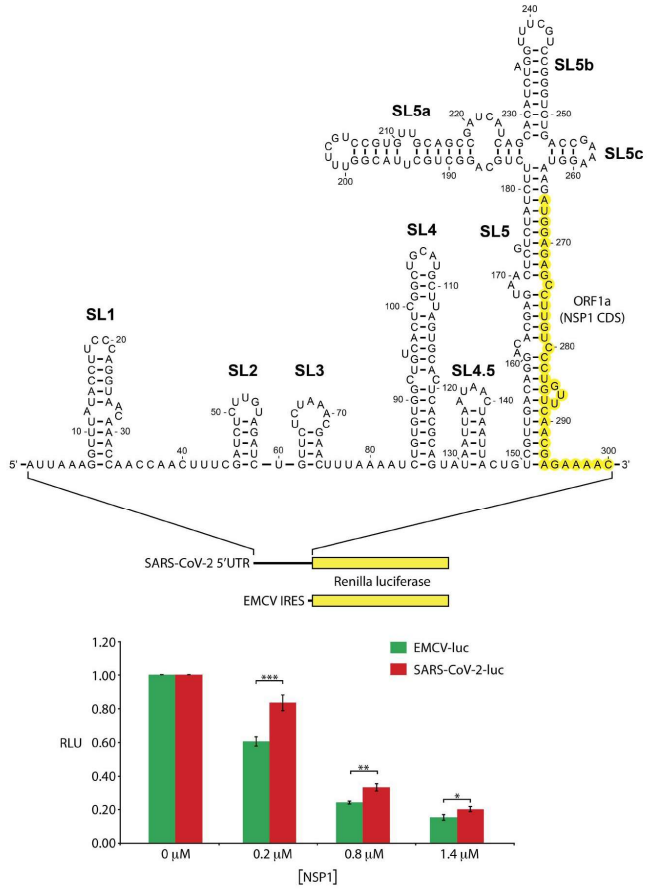


FIGURE 1. Translation inhibition by viral NSP1. The first reporter construct contains the SARS-CoV2 5′UTR plus the 12 N-terminal codons of NSP1 (nucleotides 1–300) fused to *Renilla* luciferase coding sequence. The second reporter contains the EMCV IRES upstream of the *Renilla* luciferase coding sequence. Translation efficiency was measured in the absence or presence of 0.2, 0.8, and 1.4 μM of recombinant NSP1. Standard deviations or translational activity for each transcript are shown and calculated from eight independent experiments. (*) 0.05 > *P*-value >0.01; (**) 0.01 > *P*-value >0.001; (***) *P*-value <0.001; based on Student's *t*-test.

the luciferase coding sequence (Fig. 2). The formed preinitiation complexes are then fractionated on sucrose gradient and collected in separate fractions. The presence of the radioactive RNAs in the collected fractions was then monitored by Cerenkov counting. With this experimental set-up, we detected the formation of 48S, 80S, and disomes for EMCV and SARS-CoV-2 constructs (Fig. 2A). With SARS-CoV-2, the 48S complex is almost absent indicating that the initiation process in this 5′UTR might be different from the one driven by EMCV. However, in the presence of GMP–PNP, a nonhydrolyzable GTP-analog that blocks the association of the large 60S ribosomal subunit (Gray and Hentze 1994), the accumulation of a 48S complex is observed, as expected (Supplemental Fig. 1). In the presence of 0.8 μM NSP1, the formation of these complexes is drastically reduced with the EMCV transcript. In contrast, the SARS-CoV-2 5′UTR transcript allows for the formation of 80S complexes to the same extent and with only a slight reduction of disomes. This experiment confirmed that the evasion from NSP1 inhibition is due to *cis*-acting elements located in the SARS-CoV-2 5′UTR.

The apical part of SL1 is absolutely required for NSP1 evasion

In order to identify precisely the *cis*-acting elements, we repeated the experiments with truncated SARS-CoV-2 5′UTRs. With the 5′ proximal 40 nt deleted (Δ40), the protection toward NSP1-mediated inhibition is totally abrogated, which indicates that this part of the 5′UTR contains essential *cis*-acting elements (Fig. 2B; Supplemental Fig. 2). This region of the 5′UTR contains the predicted hairpin SL1 (Rangan et al. 2020) that was confirmed by probing experiments (Miao et al. 2020). Next, we introduced four mutations in the upper part of SL1. These mutations prevent the formation of the apical stem–loop of SL1 (called SL1 mut) (Fig. 2C). We have verified that the introduced mutations do induce the opening of SL1 (Supplemental Fig. 3). Again, the formation of preinitiation complexes is totally prevented with SL1 mut, indicating that evasion to NSP1-mediated inhibition is abrogated. Thus, the apical part of SL1 is absolutely required for NSP1 evasion.

A fully functional NSP1 is necessary for translation inhibition

Mutations of residues K164 and H165 of NSP1 into alanines totally abolish the binding to the 40S subunit

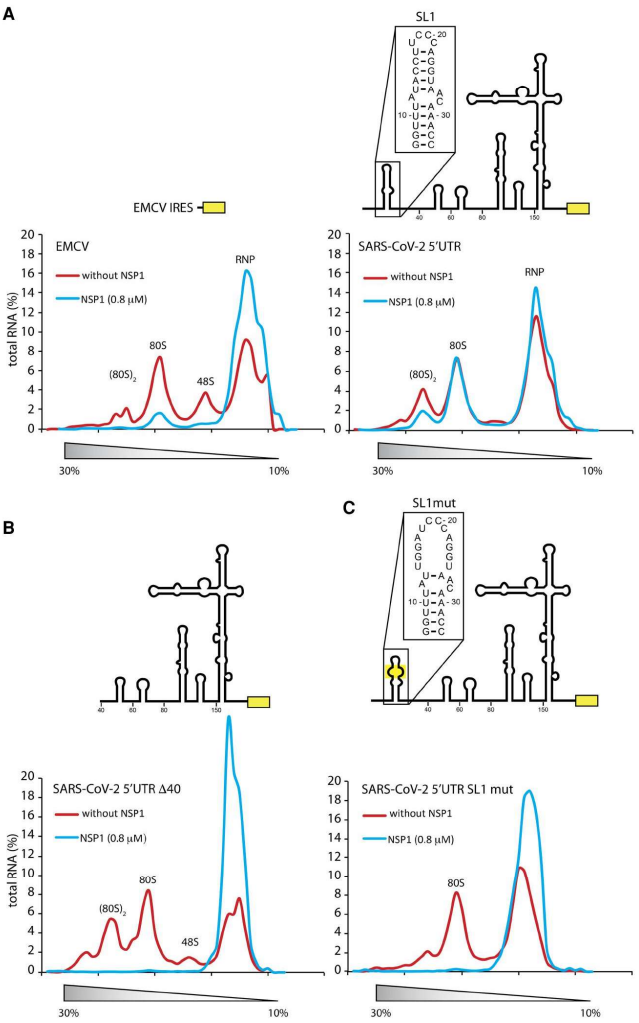


FIGURE 2. SL1 is required in the SARS-CoV-2 5′UTR for NSP1 evasion. Preinitiation complex formation analysis on 10%–30% sucrose gradients using radiolabeled RNA at their 5′ end. The SARS-CoV-2 RNA transcript was radiolabeled with a radioactive m7G cap at its 5′ end. The preinitiation complexes were fractionated on 10%–30% sucrose gradients. The presence of radioactive RNA was monitored by Cerenkov counting of all the fractions. The plots represent the percentage of radioactive RNA that was used for complex formation. The preinitiation complexes were formed in the absence (red line) or presence of 0.8 μM of recombinant viral NSP1 (blue line). The positions of ribonucleoproteins (RNP), the 48S, 80S, and disomes are indicated *above* the curves. (*A*) Translation initiation complexes analysis on sucrose gradients with EMCV reporter mRNA (*left* panel) and with SARS-CoV-2 reporter mRNA (*right* panel). Translation initiation complexes analysis on sucrose gradients with a truncated SARS-CoV-2 (*B*) or a SARS-CoV-2 5′UTR containing a mutation that disrupts SL1 (*C*).

(Kamitani et al. 2009). Likewise, the mutations R124A and K125A inhibit NSP1 ability to promote translation inhibition (Lokugamage et al. 2012). We tested the effect of these mutations in SARS-CoV-2 NSP1 on the preinitiation complex formation with the SARS-CoV-2 containing a mutated SL1. As expected, the wild-type NSP1 severely reduces the formation of the preinitiation complexes when SL1 is mutated. In contrast, none of the two NSP1 mutants affects the translation, indicating that NSP1 has to be bound to the ribosome to inhibit translation efficiently (Supplemental Fig. 4). Altogether, these experiments demonstrate that the presence of the *cis*-acting element SL1 and a NSP1 protein able to bind ribosomes is necessary to promote a viral translation resistant to NSP1-mediated inhibition.

The apical part of the SL1 RNA hairpin is solely responsible for NSP1 resistance

Next, we tested if SL1 is solely responsible for NSP1-resistance. For that purpose, we measured the impact of NSP1 on translation of another reporter construct containing the β-globin 5′UTR upstream of the *Renilla* coding sequence. Translation is significantly inhibited in the presence of increasing concentrations of NSP1, confirming the general inhibitory effect of NSP1 bound on the ribosome (Fig. 3A). We then transplanted the 5′ proximal 40 nt of the SARS-CoV-2 5′UTR, which contain the SL1 hairpin, upstream of the β-globin 5′UTR. The sole presence of SL1 allowed a significant protection against NSP1 inhibition. In contrast, when SL1 mut is transplanted, no protection is observed. Since the subgenomic RNAs contain the so-called leader sequence that encompasses SL1, SL2, and SL2 in their 5′UTR, we checked whether SL2 and SL3 are also required for NSP1 evasion. Indeed, addition of SL2 and SL3 only slightly improves the translation efficiency in the presence of NSP1 (Fig. 3B). This experiment indicates that

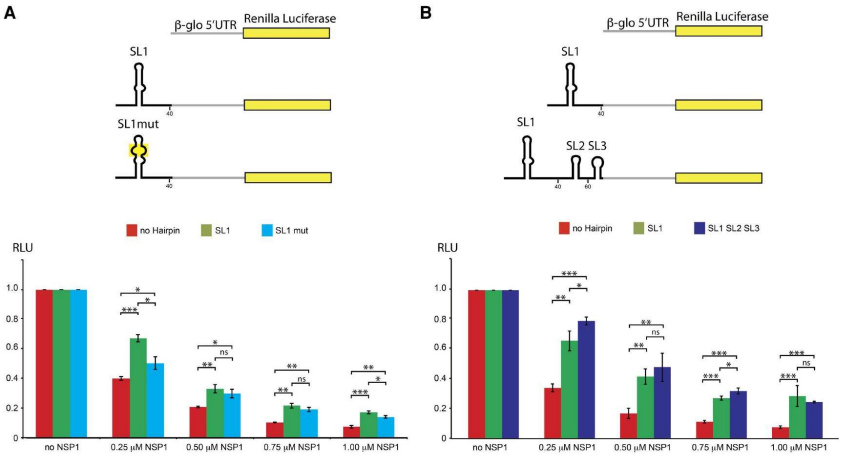


FIGURE 3. SL1 is sufficient to confer resistance to NSP1 inhibition. (*A*) Luciferase reporter mRNAs containing β-globin 5′UTR, SL1-β-globin 5′UTR, and SL1mut-β-globin 5′UTR were used to measure their translation efficiency in RRL in the absence or presence of 0.25, 0.50, 0.75, and 1 μM of recombinant NSP1. The average relative activities from three independent experiments are represented in the histogram. The activity of the reporter β-globin 5′UTR luciferase in the absence of NSP1 is used as a control for normalization. (*B*) Luciferase reporters containing β-globin 5′UTR or with SL1, or with SL1–SL2–SL3 in their 5′UTR. Standard deviations or translational activity for each transcript are shown and calculated from three independent experiments. (ns) nonsignificant; (*) 0.05 > *P*-value >0.01; (**) 0.01 > *P*-value >0.001; (***) *P*-value <0.001; based on Student's *t*-test.

subgenomic RNAs can also evade to NSP1-mediated inhibition because they harbor SL1 in their 5′UTR. These data confirm that the apical part of SL1 is essential for NSP1 evasion. Altogether, these results indicate that the evasion of the SARS-CoV-2 RNAs to NSP1 inhibition is due solely to the presence of SL1 in the 5′UTR leader sequence.

Free NSP1 has no affinity for RNA

A putative mechanism would be that SL1 directly interacts with NSP1 and thereby removes NSP1 from the ribosome to allow access to the mRNA channel of the ribosome during the initiation process of mRNA accommodation. To evaluate such a mechanism, we tested the RNA binding ability of NSP1 (Fig. 4A). We used radiolabeled RNAs from SARS-CoV-5′UTR with EMVC and HCV IRES as negative controls. None of the RNAs tested are bound by NSP1 even when 20 μM of NSP1 was used for Electrophoretic Mobility Shift Assay (EMSA). We concluded from these experiments that free NSP1 has no affinity for RNA even when it contains SL1 like the SARS-CoV-2 5′UTR.

The SARS-CoV-2 5′UTR does not bind to any component of the ribosome on its own

However, NSP1 has a strong affinity for the 40S ribosomal subunit, and its carboxy-terminal domain binds into the mRNA channel with a K_d in the nanomolar range (Lapointe et al. 2020). We therefore tested whether a SARS-CoV-2 transcript is able to bind to pure ribosomal subunits by EMSA. To validate our assay, we used the HCV IRES as a positive control, since it was shown earlier that it interacts specifically with the 40S ribosomal subunit (Filbin et al. 2013; Fuchs et al. 2015; Quade et al. 2015; Yamamoto et al. 2015) and even with full 80S (Yokoyama et al. 2019). As expected, the HCV IRES interacts with purified human 40S ribosomal subunit and with the 40S subunit from the complete 80S ribosome. In contrast to the HCV IRES, the SARS-CoV-2 5′UTR is not able to bind the 40S, and neither to the 60S nor the 80S particles (Fig. 4B). In summary, NSP1 has a strong affinity for the ribosomal 40S subunit but the SARS-CoV-2 5′UTR cannot bind to any component of the ribosome on its own.

The SARS-CoV-2 5′UTR promotes the assembly of preinitiation complexes in the presence of NSP1: the interaction between NSP1 and SL1 frees the mRNA accommodation channel while maintaining NSP1 bound to the ribosome

However, our results also indicate that the SARS-CoV-2 5′UTR promotes the assembly of preinitiation complexes in the presence of NSP1. First, we carefully checked that all the ribosomal complexes that are assembled on the SARS-CoV-2 5′UTR are fully functional and do not

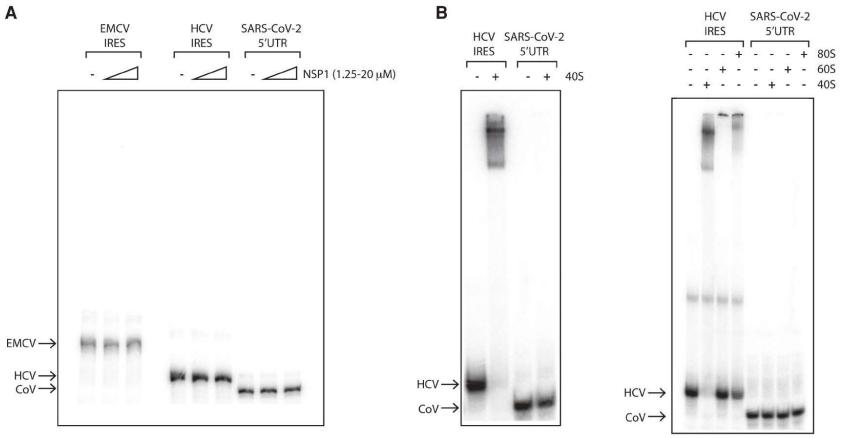


FIGURE 4. NSP1 has no RNA binding ability on its own. (*A*) Electrophoretic mobility shift assay using radiolabeled RNA containing EMCV IRES, HCV IRES, and the SARS-CoV-2 5′UTR. The RNAs were incubated in the presence of 1.25 to 20 μM recombinant NSP1 and loaded on native polyacrylamide gel. (*B*) The SARS-CoV-2 5′UTR was also used to test its binding to ribosomal 40S (*left* panel) and to 40S, 60S, and 80S (*right* panel). The HCV IRES was used as a positive control. The positions of the free RNA are indicated by arrows.

correspond to nonproductive ribosome complexes. Indeed, in the presence of GMP–PNP, we see efficient accumulation of 48S particles. In contrast, in the presence of edeine, which interferes specifically in the P-site with the codon–anticodon interaction between the initiator tRNA and the start codon, no 80S ribosomal complexes are assembled and only 43S complexes, presumably scanning complexes, are observed (Supplemental Fig. 1). These experiments demonstrate that the ribosome complexes that are assembled on SARS-CoV-2 5′UTR are fully functional and that nonproductive complexes are absent. In order to determine whether NSP1 is removed from the assembled preinitiation complexes, we used a previously established protocol that yields purified preinitiation complexes programmed with SARS-CoV-2 5′UTR (Prongidi-Fix et al. 2013; Chicher et al. 2015; Martin et al. 2016). The principle is to use a chimeric molecule composed on one hand by the RNA region encompassing the SARS-CoV-2 5′UTR followed by a small coding sequence and a DNA oligonucleotide coupled to Biotin at its 3′ end (Fig. 5A). The hybrid molecules are then immobilized on magnetic streptavidin beads and incubated in RRL in the presence of cycloheximide. Cycloheximide blocks the first translocation step and, therefore, incubation of SARS-CoV-2 5′UTR in RRL, previously treated with cycloheximide, leads to the accumulation of 80S ribosomes that are stalled on the start codon. The complexes are then eluted by DNase digestion that removes the Biotin and the DNA linker. The composition of the eluted ribosomal complexes is determined by mass spectrometry analysis. We performed in parallel two experiments with the SARS-CoV-2 5′UTR in the presence or in the absence of NSP1. Each experiment was repeated three times (Fig. 5B). Since cycloheximide induces the stalling of 80S complexes on the start codon, the 5′ cap of the mRNA is still accessible. Therefore, another scanning complex can also be present on the mRNA, meaning that we can in fact purify 43S, 48S, and 80S complexes at the same time. Concerning elongation, it is well described that cycloheximide does not allow a 100%-blockage and that a small proportion of disomes is always observed, which explains the presence of elongation factors in our complexes. The important point here is that NSP1 is still present together with disomes, 80S and scanning complexes. Since all these complexes are assembled with the mRNA in the mRNA channel, we have to conclude on the presence of NSP1 on these ribosomal complexes but with its carboxy-terminal domain displaced from the mRNA channel to enable the presence of mRNA.

Two actin-binding proteins, MYH9 and MYH10, are depleted from the complexes when NSP1 is present. Interestingly, it has been reported that depletion or inactivation of the myosin nonmuscle proteins MYH9 and MYH10 leads to renal failure (Otterpohl et al. 2020) or pulmonary disease (Kim et al. 2018), characteristic symptoms of SARS-CoV-2 infection.

In these two experiments, preinitiation complexes were efficiently purified as attested by the presence of 40S and 60S ribosomal proteins and eukaryotic initiation factors (eIFs) (Fig. 5C). In both complexes, we found the full set of ribosomal proteins from the 60S and the 40S ribosomal

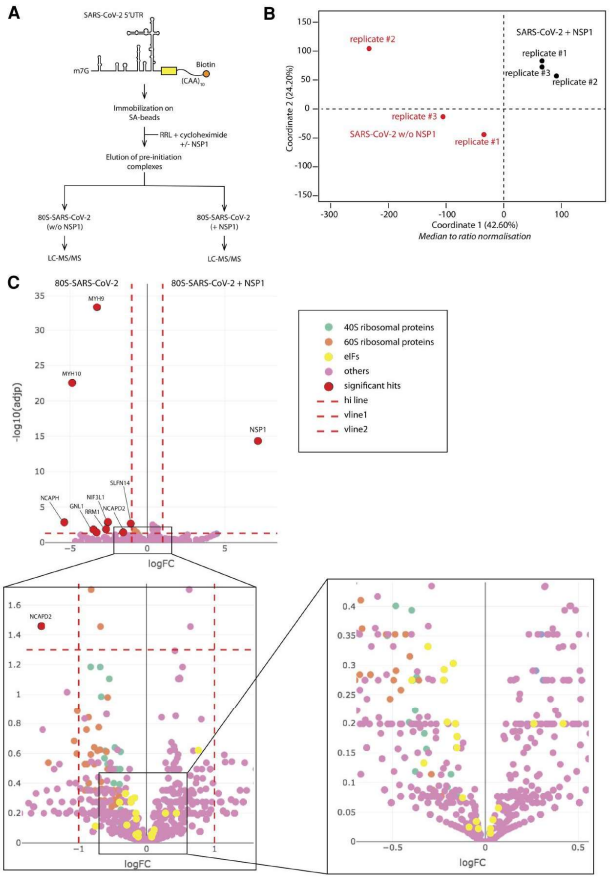


FIGURE 5. NSP1 remains bound to the preinitiation complex programmed with SARS-CoV-2 5′UTR. (*A*) Experimental strategy adapted from Chicher et al. (2015) to purify translation initiation complexes programmed with SARS-CoV-2 5′UTR in the absence or presence of recombinant NSP1. (*B*) Multidimensional scaling plot illustrating global variance and similarities between the SARS-CoV-2 (w/oNSP1) and SARS-CoV-2 (+NSP1) populations detected in the replicates, after a median-to-ratio normalization. (C) Volcano plot showing the proteins copurified with NSP1 as compared to the control condition performed without NSP1. *Y*- and *x*-axes display adjusted *P*-values and fold changes, respectively. The proteins indicated by a red circle are enriched in either the SARS-CoV-2 (+NSP1) condition ($Log_2FC > 1$) or in the SARS-CoV-2 (w/oNSP1) condition ($Log2FC < -1$). The dashed line indicates the threshold above which proteins are significantly enriched ($adjP < 0.05$). Magnifications of the nonsignificant hits are shown in order to visualize better the proteins from the translation machinery. Green and orange circles label the 40S and 60S ribosomal proteins, respectively, yellow circles label the initiation factors, and purple circles correspond to other proteins. The source data are available in Supplemental Table 1.

subunits (50 proteins and 36 proteins, respectively) (Supplemental Table 1). The fact that proteins from the translation machinery were found in both complexes confirms that NSP1 does not inhibit translation of SARS-CoV-2 mRNAs. Importantly, we found that NSP1 is still present in the preinitiation complexes formed in the presence of NSP1. This indicates that the purified preinitiation complexes contain NSP1 still bound on the ribosome. Thus, these experiments imply that the mRNA channel is accessible to the mRNA and that the NSP1 carboxy-terminal domain must have been remodeled in order to allow mRNA accommodation. Since NSP1 is present in these preinitiation complexes, this suggests that NSP1 remains attached to the ribosomal subunit with its amino-terminal domain (according to data from Shi et al. 2020). The addition of purified transcripts containing SL1 or SL1–SL2–SL3 in *trans* (with a 10-fold excess) does not rescue the translation in the presence of NSP1 (data not shown). Therefore NSP1-evasion does require the presence of SL1 in *cis* on the mRNA.

DISCUSSION

Altogether, our data enable us to propose the following model (Fig. 6). During the early stages of the SARS-CoV-2 infectious program, ORF1a is translated by canonical cap-dependent translation. The corresponding protein is then processed into NSP proteins. Among these, NSP1 binds to the ribosome with a high affinity. Its carboxy-terminal domain interacts with the mRNA channel entry site and thereby blocks the access to mRNAs. Cellular translation is consequently drastically shut down. However, viral translation escapes this blockage and still goes on. We have confirmed that viral translation is proceeding further in the presence of NSP1 in a so-called viral evasion. We demonstrated here that SL1 is required for this viral evasion. Deletion of SL1 abrogates NSP1 evasion which is in good agreement with a previously published model in which viral translation is also inhibited by NSP1 (Schubert et al. 2020). Indeed, according to the primers described in this publication, the reporter

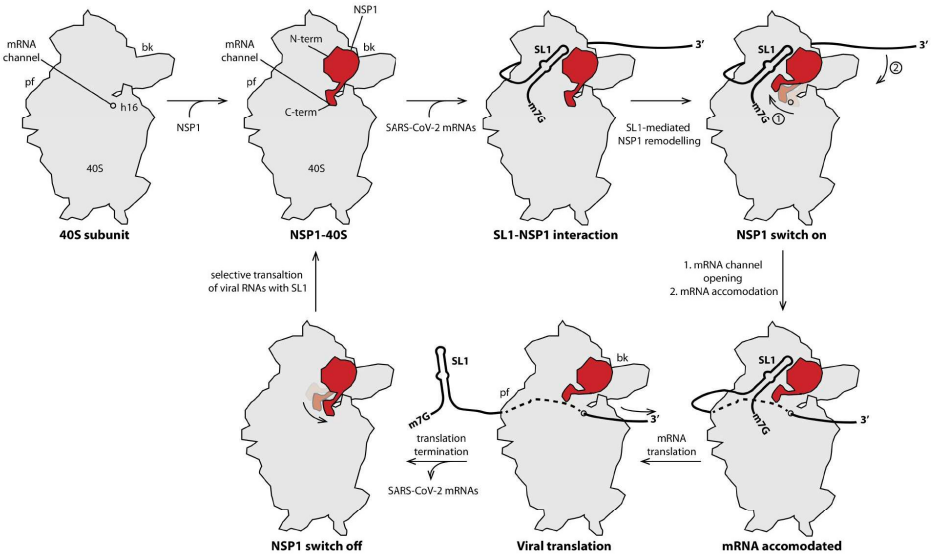


FIGURE 6. Model for NSP1 acting as a gatekeeper to ensure NSP1 evasion by SARS-CoV-2 5′UTR. In the early phase of infection, NSP1 protein is produced and binds with high affinity to the 40S ribosomal subunit. The images of the 40S subunits are shown from the solvent side. NSP1 binds on the beak of the 40S by interacting with the amino-terminal part of NSP1; this interaction places the carboxy-terminal part of NSP1 in the mRNA channel entry site and thereby prevents any mRNA accommodation. The viral mRNA transcripts contain in their 5′UTR SL1, that by interacting with the 40S-NSP1 complex enables mRNA accommodation and the formation of translation initiation complexes (the mRNA is accommodated in the decoding site on the 40S intersubunit side and is shown with dashed line). This required the removal of the carboxy-terminal domain of NSP1 to open access to the mRNA channel. Then, translation initiation, elongation, and termination can proceed further. After termination, the mRNA is released, the NSP1 carboxy-terminal can refold in front of the mRNA channel and prevent any de novo cellular mRNA translation. Only viral mRNA transcripts can access the mRNA channel, thanks to the SL1 that is present in the 5′UTR of genomic and subgenomic RNAs.

used in the later study did not contain SL1, which in fact confirms our results obtained with Δ40, Δ60, Δ80, and Δ150 (Supplemental Fig. 2) that showed that viral evasion is abrogated when SL1 is deleted.

In the model deduced from the available and the present data (Fig. 6), the SARS-CoV-2 5′UTR contains a *cis*-acting element, the SL1 hairpin, that induces during translation initiation a structural rearrangement of NSP1, especially of its carboxy-terminal domain. This frees the access to the mRNA channel and allows viral translation to proceed further. This model is in good agreement with two models recently proposed by others (Banerjee et al. 2020; Shi et al. 2020). They found that the amino-terminal part of NSP1 is required for evasion from NSP1 inhibition, that the length of the linker between the amino- and carboxy-terminal domains is critical and that the 5′UTR is required for viral evasion. In addition, they proposed that NSP1 is removed from the ribosome and retained by an interaction between the amino-terminal domain of NSP1 and the 5′UTR (Banerjee et al. 2020; Shi et al. 2020). Although we cannot rule out this possibility, the fact that we have not been able to detect any RNA binding ability with free NSP1 is a strong argument against this suggestion. Moreover, if NSP1 is released from the ribosome, this will lead to free ribosomes that can start again to translate cellular mRNAs. To ensure efficient shut down of host translation, it is more efficient for the virus to maintain NSP1 on the ribosome. Therefore, we rather suggest that NSP1 stays bound on the ribosome. Within this model, at the end of translation, the carboxy-terminal domain of NSP1 folds back into the mRNA channel and prevents any de novo cellular translation. In the late phase of infection, nine subgenomic RNAs are produced. Interestingly, all these viral transcripts contain the so-called leader body junction that contains SL1, SL2, and SL3 (Kim et al. 2020). The presence of SL1 in all the SARS-CoV-2 transcripts probably ensures efficient NSP1 evasion while still allowing efficient translation. This is especially important for the translation of subgenomic RNAs that is required in the late phases of the infectious process when the concentration of NSP1 is high and when most if not all the ribosomes are blocked by NSP1. In the model presented

in Figure 6, NSP1 acts as a gatekeeper to control selectively the access to the mRNA channel, preventing cellular translation and restricting translation to the sole viral transcripts. The release of the NSP1 gatekeeper is controlled by SL1 from viral transcripts. In conclusion, the sole presence of SL1 in all the SARS-CoV-2 transcripts is a prerequisite to complete the viral infectious program. Therefore, targeting SL1 for therapeutic purposes would be an elegant approach to impair viral translation, in the early phase but also in the late phases of SARS-CoV-2 infection.

Targeting NSP1 proteins from coronaviruses for the development of novel therapeutic strategies has been proposed earlier (Kamitani et al. 2006, 2009; Wathelet et al. 2007; Züst et al. 2007; Narayanan et al. 2008; Tohya et al. 2009; Huang et al. 2011a; Lokugamage et al. 2012; Tanaka et al. 2012; Jauregui et al. 2013; Jimenez-Guardeño et al. 2015; Wu et al. 2020). Another attractive alternative would be to target SL1. Indeed, SL1 being present in all the viral transcripts, drug-design against SL1 would allow to target specifically viral translation in the early phase of infection by impairing genomic RNA translation and in the late phase of infection by blocking translation of subgenomic RNAs. Altogether, SL1 might be seen as a genuine "Achilles heel" of SARS-CoV-2.

MATERIALS AND METHODS

In vitro transcription

The different variants of reporter constructs were transcribed by run-off in vitro transcription with T7 RNA polymerase. Uncapped RNAs were separated on denaturing PAGE (4%) and RNA were recovered from the gel slices by electroelution. The resulting pure RNA transcripts were capped at their 5′ end with the ScriptCap m7G Capping System (Epicenter Biotechnologies).

In vitro translation

In vitro translation with cell-free translation extracts were performed using self-made rabbit reticulocyte lysates (RRL) as previously described (Martin et al. 2011). Briefly, reactions were incubated at 30°C for 60 min and included 200 nM of each transcript and 10.8 µCi [^{35}S]Met. Aliquots of translation reactions were analyzed for *Renilla* luciferase activity on a luminometer.

Sucrose gradient analysis

For sucrose-gradient analysis, 5′-^{32}P-labeled mRNA were incubated in self-made RRL, in the presence of recombinant NSP1. NSP1 was incubated with RRL 5 min at 30°C prior to addition of radiolabeled mRNAs. The translation initiation complexes were separated on a 10%–30% linear sucrose gradient in buffer (25 mM Tris–HCl [pH 7.4], 50 mM KCl, 5 mM $MgCl_2$, 1 mM DTT). The reactions were loaded on the gradients and spun (23,411*g* for 2.5 h at 4°C) in a SW41 rotor. mRNA sedimentation on sucrose gradients was monitored by Cerenkov counting after fractionation.

DNA templates and primers used

A DNA fragment containing the 900 first nucleotides of SARS-COV-2 (accession number: MN908947.3) was ordered from Integrated DNA Technologies (IDT). The fragment was used as template for all the subsequent PCR amplifications. The DNA fragment corresponding to NSP1 (179 amino acid first NSP1 residues after TEV cleavage) was cloned in pET-His-GST-TEV-LIC-(2GT) (Addgene) with Gibson Assembly Cloning Technology (Gibson et al. 2009). The plasmids containing the EMCV IRES and the β-globin sequence have been kindly provided by G. Goodall and by B. Sargueil, respectively. The primers used in this work are listed in Supplemental Table 2.

Mass spectrometry analysis and data processing

Proteins were digested with sequencing-grade trypsin (Promega) and analyzed by nano LCMS/MS as previously described (Chicher et al. 2015). Digested proteins were then analyzed on a QExactive + mass spectrometer coupled to an EASY-nanoLC-1000 (Thermo Fisher Scientific). MS data were searched against the Rabbit UniProtKB subdatabase (release 2020_05, taxon 9986, 43454 sequences) with a decoy strategy. Peptides were identified with Mascot algorithm (version 2.5, Matrix Science) and data were imported into Proline 1.4 software (Bouyssié et al. 2020). Proteins were validated on Mascot pretty rank equal to 1, Mascot score above 25 and 1% FDR on both peptide spectrum matches (PSM score) and protein sets (Protein Set score). The total number of MS/MS fragmentation spectra was used to quantify each protein in each condition performed in three replicates. The statistical analysis based on spectral counts was performed using a home-made R package that calculates fold change and *P*-values using the quasi-likelihood negative binomial generalized log-linear model implemented in the edgeR package (https://github.com/hzuber67/IPinquiry4). The size factors used to scale samples were calculated according to the DESeq2 normalization method (i.e., median of ratios method). Volcano plots display the adjusted *P*-values and fold changes in the *y*- and *x*-axis, respectively, and show the enrichment of proteins in both conditions. *P*-values were adjusted using Benjamini–Hochberg method from stats R package.

NSP1 overexpression and purification

NSP1 and derivatives (R124A+K125A—Inhibits translation, no mRNA degradation—(Lokugamage et al. 2012); K164A + H165A—biologically inactive—(Narayanan et al. 2008) were cloned in plasmid pET-His-GST-TEV-LIC-(2GT). The plasmid pET-His-GST-TEV-LIC-(2GT) was purchased from Addgene. This vector overexpresses fusion proteins carrying a 6-His-tag on the amino-terminal GST domain followed by TEV protease cleavable site and by the NSP1 native protein or mutants. The fusion proteins were expressed in *E. coli* BL21 Rosetta (DE3) pLysS cells. Cells were grown at 37°C to a cell density of $OD_{600} = 0.6$. Temperature was decreased to 20°C and cells were induced by addition of 0.1 mM IPTG. Twelve hours after induction, pelleted cells were resuspended in EQ/W buffer (40 mM Na phosphate pH 7.2, 500 mM NaCl, 30 mM imidazole) supplemented with 0.1% Triton X-100, cOmplete Protease Inhibitor Cocktail

(Merck) and incubated on ice for 30 min with 1 mg/mL lysozyme. After lysis by sonification, the cell lysate was centrifuged at 105,000*g* for 1 h 30 min and the supernatant was applied to Ni-NTA Superflow resin (QIAGEN) equilibrated in buffer EQ/W. After column washing, NSP1 proteins were eluted from the resin by buffer EQ/W containing 250 mM imidazole. The NSP1 fraction was dialyzed against buffer EQ/W without imidazole overnight. The NSP1 fraction was loaded on Glutathione HiCap resin (Qiagen) equilibrated with the dialysis buffer and proteins were eluted by the same buffer supplemented with 50 mM glutathione. The purified 6-His-GST-TEV-NSP1 fusion proteins were subjected to TEV protease cleavage overnight at 4°C (50/1 fusion/TEV molar ratios). NSP1 proteins were separated from the 6-His-GST domain using a last purification step on the Ni-NTA resin that retained His-tagged GST and TEV proteins. The pure NSP1 proteins were concentrated, and stored in buffer that contains 50% glycerol at −20°C.

Ribosome purification

Ribosomes from HeLa cells were purified according to the previous protocol established for rabbit reticulocyte lysate (Pestova et al. 1996). Briefly, 826×10^6 HeLa cells (IGBMC) were resuspended in 40 mL of lysis buffer (15 mM Tris–HCl pH 7.5, 300 mM NaCl, 6 mM $MgCl_2$, 1% Triton X100) and incubated at 0°C for 20 min. All the following steps were performed at 4°C. Cell debris were removed by centrifugation at 6000*g* for 1 h. The supernatant was centrifuged at 45,000 rpm in a Ti50.2 rotor (Beckman) for 5.5 h. The ribosome pellet was resuspended in 6 mL buffer A (20 mM Tris–HCl pH 7.5, 50 mM KCl, 4 mM $MgCl_2$, 2 mM DTT, 250 mM sucrose) and puromycin was added to the final concentration of 1 mM, and incubated for 10 min at 0°C then for 10 min at 37°C before addition of KCl to a final concentration of 500 mM. The solution was then centrifuged in a Ti50.2 rotor at 45,000 rpm for 4.5 h, yielding a pellet of ribosomes that was cleared from initiation and other cellular factors. Ribosomes were resuspended in buffer A supplemented with 500 mM KCl and resolved by centrifugation through a 10%–30% sucrose gradient in buffer A (+500 mM KCl) in SW32Ti rotor (22,000 rpm for 17 h, 100 A_{260}/tube). Gradients were fractionated in 1-mL fractions that were analyzed by electrophoretic migration on agarose gel. Fractions containing 40S and 60S were assembled separately and centrifuged at 130,000 rpm in S140AT rotor (Hitachi) for 2 h. The resulting pellets of 40S and 60S subunits were resuspended in buffer B (20 mM Tris–HCl pH 7.5, 2 mM DTT, 2 mM $MgCl_2$, 100 mM KCl, 250 mM sucrose) for storage. In order to get pure 80S subunits, a similar purification was performed with the exception of the dissociation step at 500 mM KCl that was omitted as well as the addition of 500 mM KCl in the sucrose gradient.

Electrophoretic mobility shift assays (EMSA)

To detect RNA–protein interactions by EMSA, recombinant NSP1 or pure human 40S, 60S, and 80S ribosomal fractions were incubated with 50 fmol of 5′ ^{32}P-labeled RNA transcripts. Briefly, proteins and RNA were mixed with 20 µg of yeast total tRNA (Merck Sigma-Aldrich) and incubated for 20 min in 10 mM Tris–HCl (pH 7.5), 50 mM KCl, 1 mM DTT, 10% glycerol in 20 µL at 0°C. The RNA–protein complexes were analyzed by electrophoresis on native 5% polyacrylamide gels using Tris-50 mM glycine as buffer system and visualized by phosphor imaging.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work is funded by Agence Nationale pour la Recherche (ANR-17-CE12-0025-01, ANR-17-CE11-0024, ANR-20-COVI-0078), Fondation pour la Recherche Médicale (project CoronaIRES), Fondation Bettencourt Schueller, University of Strasbourg, and the Centre National de la Recherche Scientifique.

Received October 14, 2020; accepted November 29, 2020.

REFERENCES

Almeida MS, Johnson MA, Herrmann T, Geralt M, Wüthrich K. 2007. Novel β-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J Virol* **81:** 3151–3161. doi:10.1128/JVI.01939-06

Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. 2020. The proximal origin of SARS-CoV-2. *Nat Med* **26:** 450–452. doi:10.1038/s41591-020-0820-9

Banerjee AK, Blanco MR, Bruce EA, Honson DD, Chen LM, Chow A, Prashant B, Noah O, Quinodoz SA, Loney C, et al. 2020. SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. *Cell* **183:** 1325–1339.e21. doi:10.1016/j.cell.2020.10.004

Bouyssié D, Hesse AM, Mouton-Barbosa E, Rompais M, MacRon C, Carapito C, Gonzalez De Peredo A, Couté Y, Dupierris V, Burel A, et al. 2020. Proline: an efficient and user-friendly software suite for large-scale proteomics. *Bioinformatics* **36:** 3148–3155. doi:10.1093/bioinformatics/btaa118

Brito Querido J, Sokabe M, Kraatz S, Gordiyenko Y, Skehel JM, Fraser CS, Ramakrishnan V. 2020. Structure of a human 48S translational initiation complex. *Science* **369:** 1220–1227. doi:10.1126/science.aba4904

Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, Yuen KY. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* **9:** 221–236. doi:10.1080/22221751.2020.1719902

Chicher J, Simonetti A, Kuhn L, Schaeffer L, Hammann P, Eriani G, Martin F. 2015. Purification of mRNA-programmed translation initiation complexes suitable for mass spectrometry analysis. *Proteomics* **15:** 2417–2425. doi:10.1002/pmic.201400628

Filbin ME, Vollmar BS, Shi D, Gonen T, Kieft JS. 2013. HCV IRES manipulates the ribosome to promote the switch from translation initiation to elongation. *Nat Struct Mol Biol* **20:** 150–158. doi:10.1038/nsmb.2465

Fuchs G, Petrov AN, Marceau CD, Popov LM, Chen J, O'Leary SE, Wang R, Carette JE, Sarnow P, Puglisi JD. 2015. Kinetic pathway of 40S ribosomal subunit recruitment to hepatitis C virus internal ribosome entry site. *Proc Natl Acad Sci* **112:** 319–325. doi:10.1073/pnas.1421328111

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to

several hundred kilobases. *Nat Methods* **6:** 343–345. doi:10.1038/nmeth.1318

Gray NK, Hentze MW. 1994. Iron regulatory protein prevents binding of the 43S translation pre-initiation complex to ferritin and eALAS mRNAs. *EMBO J* **13:** 3882–3891. doi:10.1002/j.1460-2075.1994.tb06699.x

Hartenian E, Nandakumar D, Lari A, Ly M, Tucker JM, Glaunsinger BA. 2020. The molecular virology of coronaviruses running title: the molecular virology of coronaviruses. *J Biol Chem* **295:** 12910–12934. doi:10.1074/jbc.REV120.013930

Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S. 2011a. SARS coronavirus nsp1 protein induces template-dependent endonucleolytic cleavage of mRNAs: viral mRNAs are resistant to nsp1-induced RNA cleavage. *PLoS Pathog* **7:** e1002433. doi:10.1371/journal.ppat.1002433

Huang C, Lokugamage KG, Rozovics JM, Narayanan K, Semler BL, Makino S. 2011b. Alphacoronavirus transmissible gastroenteritis virus nsp1 protein suppresses protein translation in mammalian cells and in cell-free HeLa cell extracts but not in rabbit reticulocyte lysate. *J Virol* **85:** 638–643. doi:10.1128/JVI.01806-10

Jauregui AR, Savalia D, Lowry VK, Farrell CM, Wathelet MG. 2013. Identification of residues of SARS-CoV nsp1 that differentially affect inhibition of gene expression and antiviral signaling. *PLoS One* **8:** e62416. doi:10.1371/journal.pone.0062416

Jimenez-Guardeño JM, Regla-Nava JA, Nieto-Torres JL, DeDiego ML, Castaño-Rodriguez C, Fernandez-Delgado R, Perlman S, Enjuanes L. 2015. Identification of the mechanisms causing reversion to virulence in an attenuated SARS-CoV for the design of a genetically stable vaccine. *PLoS Pathog* **11:** e1005215. doi:10.1371/journal.ppat.1005215

Kamitani W, Narayanan K, Huang C, Lokugamage K, Ikegami T, Ito N, Kubo H, Makino S. 2006. Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc Natl Acad Sci* **103:** 12885–12890. doi:10.1073/pnas.0603144103

Kamitani W, Huang C, Narayanan K, Lokugamage KG, Makino S. 2009. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nat Struct Mol Biol* **16:** 1134–1140. doi:10.1038/nsmb.1680

Kim HT, Yin W, Jin YJ, Panza P, Gunawan F, Grohmann B, Buettner C, Sokol AM, Preussner J, Guenther S, et al. 2018. Myh10 deficiency leads to defective extracellular matrix remodeling and pulmonary disease. *Nat Commun* **9:** 4600. doi:10.1038/s41467-018-06833-7

Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* **181:** 914–921.e10. doi:10.1016/j.cell.2020.04.011

Kumar A, Kumar A, Kumar P, Garg N, Giri R. 2020. SARS-CoV-2 NSP1 C-terminal region (residues 130-180) is an intrinsically disordered region. bioRxiv doi:10.1101/2020.09.10.290932

Lai MM, Stohlman SA. 1981. Comparative analysis of RNA genomes of mouse hepatitis viruses. *J Virol* **38:** 661–670. doi:10.1128/JVI.38.2.661-670.1981

Lapointe CP, Grosely R, Johnson AG, Wang J, Fernández IS, Puglisi JD. 2020. Dynamic competition between SARS-CoV-2 NSP1 and mRNA on the human ribosome inhibits translation initiation. bioRxiv doi:10.1101/2020.08.20.259770

Lei L, Ying S, Baojun L, Yi Y, Xiang H, Wenli S, Zounan S, Deyin G, Qingyu Z, Jingmei L, et al. 2013. Attenuation of mouse hepatitis virus by deletion of the LLRKxGxKG region of Nsp1. *PLoS One* **8:** e61166. doi:10.1371/journal.pone.0061166

Lei X, Dong X, Ma R, Wang W, Xiao X, Tian Z, Wang C, Wang Y, Li L, Ren L, et al. 2020. Activation and evasion of type I interferon responses by SARS-CoV-2. *Nat Commun* **11:** 3810. doi:10.1038/s41467-020-17665-9

Lokugamage KG, Narayanan K, Huang C, Makino S. 2012. Severe acute respiratory syndrome coronavirus protein nsp1 is a novel eukaryotic translation inhibitor that represses multiple steps of translation initiation. *J Virol* **86:** 13598–13608. doi:10.1128/JVI.01958-12

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395:** 565–574. doi:10.1016/S0140-6736(20)30251-8

Martin F, Barends S, Jaeger S, Schaeffer L, Prongidi-Fix L, Eriani G. 2011. Cap-assisted internal initiation of translation of histone H4. *Mol Cell* **41:** 197–209. doi:10.1016/j.molcel.2010.12.019

Martin F, Ménétret JF, Simonetti A, Myasnikov AG, Vicens Q, Prongidi-Fix L, Natchiar SK, Klaholz BP, Eriani G. 2016. Ribosomal 18S rRNA base pairs with mRNA during eukaryotic translation initiation. *Nat Commun* **7:** 12622. doi:10.1038/ncomms12622

Masters PS. 2006. The molecular biology of coronaviruses. *Adv Virus Res* **66:** 193–292. doi:10.1016/S0065-3527(06)66005-3

Miao Z, Tidu A, Eriani G, Martin F. 2020. Secondary structure of the SARS-CoV-2 5′-UTR. *RNA Biol* **23:** 1–10. doi:10.1080/15476286.2020.1814556

Nakagawa K, Lokugamage KG, Makino S. 2016. Viral and cellular mRNA translation in coronavirus-infected cells. *AdvVirus Res* **96:** 165–192. doi:10.1016/bs.aivir.2016.08.001

Narayanan K, Huang C, Lokugamage K, Kamitani W, Ikegami T, Tseng C-TK, Makino S. 2008. Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells. *J Virol* **82:** 4471–4479. doi:10.1128/JVI.02472-07

Otterpohl KL, Busselman BW, Ratnayake I, Hart RG, Hart K, Evans C, Phillips CL, Beach JR, Ahrenkiel P, Molitoris B, et al. 2020. Conditional *Myh9* and *Myh10* inactivation in adult mouse renal epithelium results in progressive kidney disease. *JCI Insight* **5:** 138530. doi:10.1172/jci.insight.138530

Pestova TV, Hellen CU, Shatsky IN. 1996. Canonical eukaryotic initiation factors determine initiation of translation by internal ribosomal entry. *Mol Cell Biol* **16:** 6859–6869. doi:10.1128/MCB.16.12.6859

Prongidi-Fix L, Schaeffer L, Simonetti A, Barends S, Ménétret J-F, Klaholz BP, Eriani G, Martin F. 2013. Rapid purification of ribosomal particles assembled on histone H4 mRNA: a new method based on mRNA-DNA chimaeras. *Biochem J* **449:** 719–728. doi:10.1042/BJ20121211

Quade N, Boehringer D, Leibundgut M, van den Heuvel J, Ban N. 2015. Cryo-EM structure of hepatitis C virus IRES bound to the human ribosome at 3.9-Å resolution. *Nat Commun* **6:** 7646. doi:10.1038/ncomms8646

Rangan R, Zheludev IN, Das R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA* **26:** 937–959. doi:10.1261/rna.076141.120

Rao S, Hoskins I, Daniela Garcia P, Tonn T, Ozadam H, Cenik S, Cenik C. 2020. Genes with 5′ terminal oligopyrimidine tracts preferentially escape global suppression of translation by the SARS-CoV-2 NSP1 protein. bioRxiv doi:10.1101/2020.09.13.295493

Schubert K, Karousis ED, Jomaa A, Scaiola A, Echeverria B, Gurzeler LA, Leibundgut M, Thiel V, Mühlemann O, Ban N. 2020. SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* **27:** 959–966. doi:10.1038/s41594-020-0511-8

Shi M, Wang L, Fontana P, Vora S, Zhang Y, Fu T-M, Lieberman J, Wu H. 2020. SARS-CoV-2 Nsp1 suppresses host but not viral translation through a bipartite mechanism. bioRxiv doi:10.1101/2020.09.18.302901

Sola I, Almazán F, Zúñiga S, Enjuanes L. 2015. Continuous and discontinuous RNA synthesis in coronaviruses. *Annu Rev Virol* **2:** 265–288. doi:10.1146/annurev-virology-100114-055218

Tanaka T, Kamitani W, DeDiego ML, Enjuanes L, Matsuura Y. 2012. Severe acute respiratory syndrome coronavirus nsp1 facilitates efficient propagation in cells through a specific translational shutoff of host mRNA. *J Virol* **86:** 11128–11137. doi:10.1128/JVI.01700-12

Thoms M, Buschauer R, Ameismeier M, Koepke L, Denk T, Hirschenberger M, Kratzat H, Hayn M, Mackens-Kiani T, Cheng J, et al. 2020. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* **369:** 1249–1255. doi:10.1126/science.abc8665

Tohya Y, Narayanan K, Kamitani W, Huang C, Lokugamage K, Makino S. 2009. Suppression of host gene expression by nsp1 proteins of group 2 bat coronaviruses. *J Virol* **83:** 5282–5288. doi:10.1128/JVI.02485-08

Wathelet MG, Orr M, Frieman MB, Baric RS. 2007. Severe acute respiratory syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an attenuated strain. *J Virol* **81:** 11620–11633. doi:10.1128/JVI.00702-07

Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, Wang Q, Xu Y, Li M, Li X, et al. 2020. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharm Sin B* **10:** 766–788. doi:10.1016/j.apsb.2020.02.008

Xia H, Cao Z, Xie X, Zhang X, John Yun-Chung C, Wang H, Menachery VD, Rajsbaum R, Shi P-Y. 2020. Evasion of type-I interferon by SARS-CoV-2. *Cell Rep* **33:** 108234. doi:10.1016/j.celrep.2020.108234

Yamamoto H, Collier M, Loerke J, Ismer J, Schmidt A, Hilal T, Sprink T, Yamamoto K, Mielke T, Bürger J, et al. 2015. Molecular architecture of the ribosome-bound hepatitis C Virus internal ribosomal entry site RNA. *EMBO J* **34:** 3042–3058. doi:10.15252/embj.201592469

Yogo Y, Hirano N, Hino S, Shibuta H, Matumoto M. 1977. Polyadenylate in the virion RNA of mouse hepatitis virus. *J Biochem* **82:** 1103–1108. doi:10.1093/oxfordjournals.jbchem.a131782

Yokoyama T, Machida K, Iwasaki W, Shigeta T, Nishimoto M, Takahashi M, Sakamoto A, Yonemochi M, Harada Y, Shigematsu H, et al. 2019. HCV IRES captures an actively translating 80S ribosome. *Mol Cell* **74:** 1205–1214. doi:10.1016/j.molcel.2019.04.022

Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579:** 270–273. doi:10.1038/s41586-020-2012-7

Züst R, Cervantes-Barragán L, Kuri T, Blakqori G, Weber F, Ludewig B, Thiel V. 2007. Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS Pathog* **3:** e109. doi:10.1371/journal.ppat.0030109

ARTICLE 3

Mise en évidence de coévolutions des séquences des tiges-boucles SL1 et des protéines NSP1 dans les coronavirus

Correlated sequence signatures are present within the genomic 5′UTR RNA and NSP1 protein in coronaviruses

PIOTR SOSNOWSKI, ANTONIN TIDU, GILBERT ERIANI, ERIC WESTHOF, and FRANCK MARTIN

Université de Strasbourg, Institut de Biologie Moléculaire et Cellulaire, Architecture et Réactivité de l'ARN, CNRS UPR9002, F-67084 Strasbourg, France

ABSTRACT

The 5′UTR part of coronavirus genomes plays key roles in the viral replication cycle and translation of viral mRNAs. The first 75–80 nt, also called the leader sequence, are identical for genomic mRNA and subgenomic mRNAs. Recently, it was shown that cooperative actions of a 5′UTR segment and the nonstructural protein NSP1 are essential for both the inhibition of host mRNAs and for specific translation of viral mRNAs. Here, sequence analyses of both the 5′UTR RNA segment and the NSP1 protein have been done for several coronaviruses, with special attention to the betacoronaviruses. The conclusions are: (i) precise specific molecular signatures can be found in both the RNA and the NSP1 protein; (ii) both types of signatures correlate between each other. Indeed, definite sequence motifs in the RNA correlate with sequence motifs in the protein, indicating a coevolution between the 5′UTR and NSP1 in betacoronaviruses. Experimental mutational data on 5′UTR and NSP1 from SARS-CoV-2 using cell-free translation extracts support these conclusions and show that some conserved key residues in the amino-terminal half of the NSP1 protein are essential for evasion to the inhibitory effect of NSP1 on translation.

Keywords: SARS-CoV-2; NSP1; SL1; 5′UTR; translation; ribosome

INTRODUCTION

The coronaviruses belong to the *Coronaviridae* family (order *Nidovirales,* kingdom *Orthornavirae*), which is subdivided into four Genera: *alpha, beta, gamma*, and *deltacoronavirus*. Here we analyze in some detail the betacoronaviruses. The betacoronaviruses are subdivided into four Subgenera: the *Embecovirus*, the *Hibecovirus*, the *Merbecovirus*, and the *Sarbecovirus* (Lefkowitz et al. 2018; Gorbalenya et al. 2020; Gulyaeva and Gorbalenya 2021). Among those, we will focus on the *Sarbecovirus,* the subgenus to which the SARS-CoV-1 and the SARS-CoV-2 belong. In *Sarbecovirus*, the genomic material is a positive single-stranded RNA molecule, 26.2–31.7 kb. The RNA strand is 5′-capped and polyadenylated at the 3′ end (Yogo et al. 1977; Lai and Stohlman 1981). The genome encodes two long open reading frames (ORF1a and ORF1b) at the 5′ side and several ORFs that are expressed in the late phase of infection from subgenomic RNAs (sgRNAs) (Brian and Baric 2005). Translation of ORF1a and ORF1b from the whole ss(+)RNA are the first events of the infectious process. The polyproteins synthesized from ORF1a and ORF1ab are processed into 16 nonstructural proteins (NSP1–NSP16). Besides the positive single strand genomic RNA, nine subgenomic RNAs (S, 3a, E, M, 6, 7a, 7b, 8, and N) are produced during the late phase of the infection by SARS-CoV-2 (Kim et al. 2020) and they all contain the common so-called 5′ leader sequence (nucleotides 1 to 75) (Kim et al. 2020).

The secondary structure of the 5′UTR of SARS-CoV-2 has been studied theoretically (Rangan et al. 2020) and in solution (Miao et al. 2021). The first 80 nt common to all viral transcripts fold in a series of three hairpins: SL1, SL2, and SL3. Recent structural studies have shown that NSP1 binds tightly to the small ribosome subunit and blocks the entry of the mRNA channel, thereby inhibiting cellular translation (Schubert et al. 2020; Thoms et al. 2020; Yuan et al. 2020; Lapointe et al. 2021). Experiments with reporter systems have further shown that with the sole presence of SL1 in the 5′UTR, the translation inhibition induced by NSP1 is relieved, thereby allowing translation (Tidu et al. 2021). SL1 promotes NSP1 evasion by acting on the NSP1 carboxy-terminal domain enabling viral RNA accommodation in the ribosome for translation. The interaction between the

Corresponding authors: f.martin@ibmc-cnrs.unistra.fr, e.westhof@ibmc-cnrs.unistra.fr

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.078972.121. Freely available online through the *RNA* Open Access option.

SL1 RNA hairpin and the NSP1 carboxy-terminal domain occurs while NSP1 remains bound on the ribosome. It was therefore concluded that NSP1 acts as a ribosome gatekeeper to impair cellular translation and specifically promote viral translation (Tidu et al. 2021). For the 5′ leader to overcome the translational inhibition imposed by NSP1, specific interactions between SL1 and NSP1 are thus expected. Here, we show that the analysis of sequence comparisons points to a coevolution of the sequences of SL1 and NSP1 in coronaviruses. In addition, in vitro translation experiments are supportive of the idea that the amino-terminal region of NSP1 contains conserved residues required to evade the translation inhibition imposed by NSP1.

The available sequence comparisons focus on the viral proteins and rarely consider the noncoding segments of the RNA genome. It was thus also interesting to compare the deduced relationships and phylogenies between strains. Comparisons between the 5′UTR of betacoronaviruses allowed us to establish a classification of SL1 into four classes in sarbecoviruses, SARS-CoV-1 harbors a type I SL1 whereas SARS-CoV-2 has a type III SL1. Concomitantly, we analysed the NSP1 proteins of betacoronavirus. NSP1 from SARS-CoV-1 and -2 present few but significant differences in the three domains of the proteins. By swapping key residues from SARS-CoV-1 into both SL1 and NSP1 from SARS-CoV-2, we bring experimental evidence that NSP1 and SL1 from the 5′UTR coevolved as suggested by sequence alignments in the Sarbecovirus family.

RESULTS

Sequence alignments for the 5′UTR RNA

To analyze the genome variability of the 5′UTRs of coronaviruses, we have aligned the 5′ proximal 75–80 nt from the corresponding genomic RNAs. Figure 1 displays sequence alignments for chosen subsets of betacoronaviruses infecting various species (a rough tree of the genus and subgenus considered here with some typical ID for the sequenced strains is shown in Fig. 1A), and Figure 2 shows the sequence alignments of the sarbecovirus subfamily that contains SARS-CoV-1 and SARS-CoV-2. The general organization of the 5′UTR into three hairpins SL1–SL2–SL3 is highly conserved in coronaviruses. However, the highest variability is observed in SL1. This allows the establishment of a classification of the coronaviruses according to their type of SL1. For simplicity, for the sarbecovirus subfamily, we refer to type I to type IV, with type I gathering the SARS-CoV-1 sequences, type III, the SARS-Cov-2, and type II the sequences displaying intermediate situations (see below), and type IV other distant sequences. The resulting consensus secondary structures of examples of the main SL1 structures are shown in Figure 3. Unfortunately, many sequences present in databases do not contain these most 5′ ends of the genomes. This is especially disturbing when attempting to delineate a temporal evolution of the viruses. However, as will be shown later, the very high conservation of the NSP1 sequences allows determining with a high probability the corresponding sequence of the 5′UTR.

The alphacoronaviruses are the oldest known coronaviruses infecting humans (Smith et al. 2014). Here, we will discuss neither the alphacoronaviruses, nor the gammacoronaviruses (Fig. 1A). The *Embecovirus* OC43 strain emerged between the end of the nineteenth and the beginning of the twentieth centuries (Vijgen et al. 2005). The HKU1, the origin of which is unknown, belongs also to this subgenus (Woo et al. 2005a,b). SARS-CoV (also called SARS-CoV-1) (Peiris et al. 2003), a *Sarbecovirus*, was identified in 2003 and MERS-CoV (Zaki et al. 2012), a *Merbecovirus*, in 2012. It is difficult to discuss the *Hibecovirus* subgenus because the sequence is unique (Wu et al. 2016). However, one can note that the hairpin loop of SL1 has similarities (UGC) with that of the *Nobecoviruses*, the hairpin loop of SL2 is identical to that of *Nobecoviruses*, and the TRS segment AACGAAC is present in both with an absence of the strong base pairs that form SL3. These two subgenera have been observed only in bats.

The *Embecovirus* subgenus displays a larger variation in sequence types. This may reflect the fact that the first sequence dates to 1997. One can distinguish two subgroups depending on the apical loop of SL1, either UGC (like in the *Nobe-* and *Hibecovirus*) or CGU (Fig. 1B). In some sequences, these residues could be part of the stable tetraloop of the type UNCG, but not in all; thus, some alternative or dynamic conformations may exist. The SL3 stem forms three to four base pairs. Correlated variations appear between the SL2 stem and the single strand between SL2 and SL3. The sequences of the *Merbecovirus* subgenus display a much larger conservation with a 3-nt loop, often -CUA (or five, since we cannot be sure that the framing nucleotides form a usual Watson–Crick pair). Again, SL3 does not appear to form a stable helix and the TRS is most probably single-stranded. The SL1 stem is stronger than in the *Embecovirus*. The TRS sequences are identical with those of the *Nobe-* and *Merbecovirus* but different from those of the *Hibe-* and *Embecovirus*.

Figure 2 shows sequences related to the *Sarbecovirus* subgenus. Two subgroups can be clearly distinguished, the SARS-CoV-1 and SARS-CoV-2. The sequence conservation is high and for SARS-CoV-1 it extends over 15 years. The similarities between the two subgroups are extensive: SL2 and SL3 are identical and the stem of SL1 presents only a change from a UoG to a U–A pair. The main differences are in SL1B, the first nucleotide of the SL1 hairpin loop, two positions 3′ of SL2; following the 3′ end of SL1, after the conserved stretch AACCAAC, there is a conserved UU in SARS-CoV-2 that is replaced by a single C in SARS-CoV-1. Figure 3 presents the resulting consensus secondary structures for the SL1 hairpin in the SARS-CoV-2,

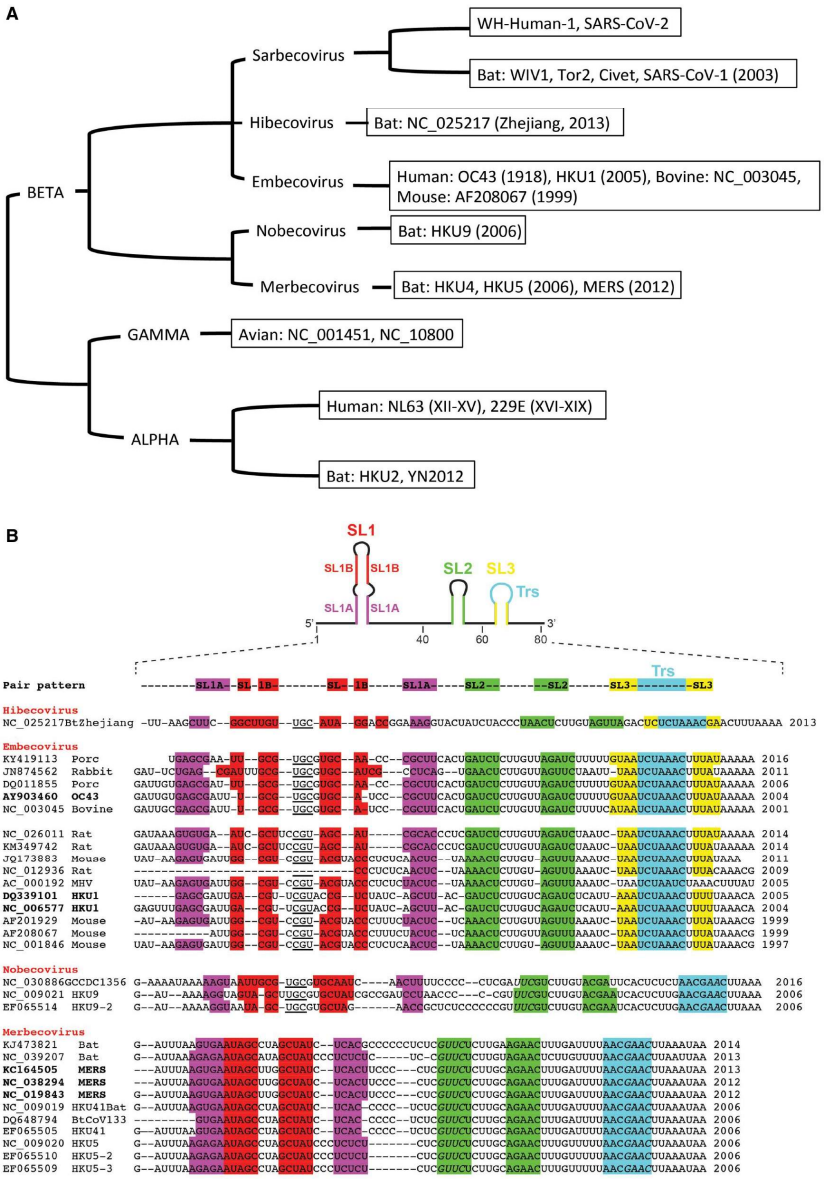


FIGURE 1. (*A*) A rough phylogenetic tree of the coronavirus family. Some strain names are indicated along the branches. Sequences can be found in the alignments shown in Figures 1B and 2. Based in part on Smith et al. (2014). (*B*) Selected sets of aligned sequences of the first 80 nt of betacoronavirus genomes, except the *Sarbecovirus*, which are shown in Figure 2 (for a rough tree of coronaviruses, see Supplemental Figure S1). The color code is such that base paired segments are colored identically (purple, SL1A, red, SL1B, green, SL2, and yellow, SL3, with the TRS sequence in cyan). Sequences in bold were isolated from infected humans. The dates are those of database deposition.

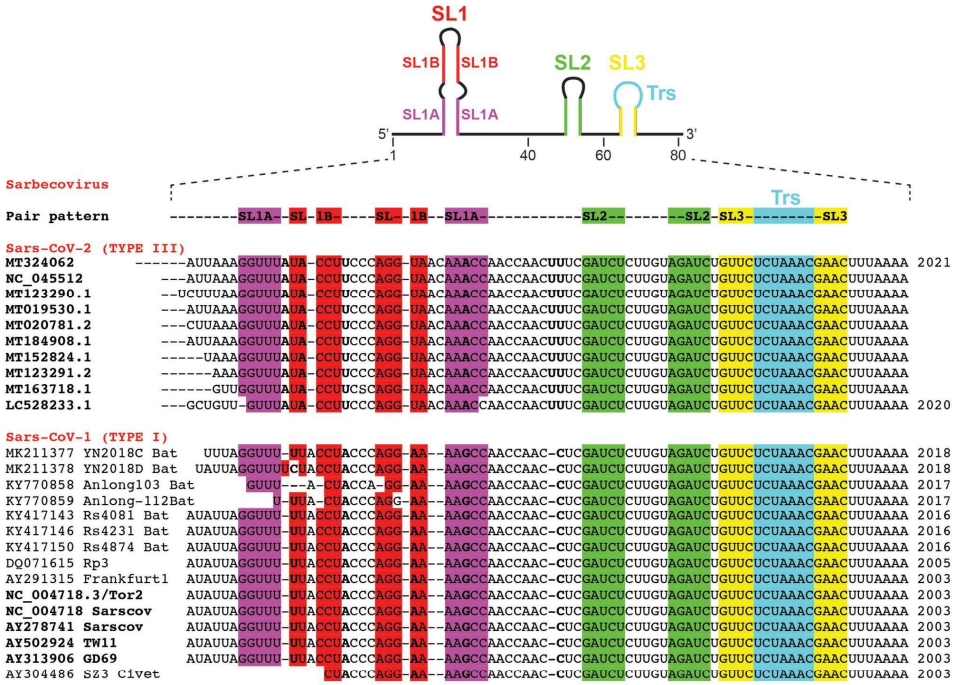


FIGURE 2. Selected sets of aligned sequences of the first 80 nt of SARS-CoV-1 and SARS-CoV-2 genomes from the betacoronavirus *Sarbecovirus*. Same annotations as in Figure 1. The nucleotides in bold vary between types I and III.

SARS-CoV-1, *Merbecovirus*, and *Embecovirus*. Some variations departing from the main types of sequences can be found. The number of sequences, from SARS-CoV-2 isolated from infected humans, with variations is small compared to the huge number of available sequences (search done using BLAST against the Coronavirus data base of Genbank). The variations occur mainly in SL1 (additional base pair in SL1B) (Supplemental Fig. S1). The variations of SARS-CoV-1 sequences occur either in the segment preceding SL2 (addition of C or changes between C and U) or in the loop of SL3 (C to U change). All those sequences were from viruses isolated from bats (Supplemental Fig. S1).

However, some sequences appear to be in-between SARS-CoV-1 and SARS-CoV-2 (Fig. 4). We named them type II sarbecoviruses. The names of the ID strains are given in Figure 4 since these are often discussed regarding the origins of the SARS-CoV-2 virus (see discussion below). The sequences deviate from the SARS-CoV-2 type between one and nine mutations. Five recent sequences come from infected humans, and all the other sequences are from bats with one from a pangolin (Guangdong). The sequences isolated from infected humans deviate by one mutation (a U to C mutation in the upstream sequence of SL2) like the Pangolin from Guandong (Liu et al. 2020) and, interestingly, two recent sequences from viruses isolated from *R. shameli* bats (Hul et al. 2021). The famous RaTG13 (Zhou et al. 2020b) sequence deviates by three mutations (like ZC45 and ZXC21 [Hu et al. 2018]), while RpYN06 and RmYN02 (Zhou et al. 2020a) deviate by two mutations only. The recently published sequence TG15 (Guo et al. 2021) deviates at least by three and maybe five mutations like RsYN04, RmYN05, or RmYN08 (Zhou et al. 2021). The sequences from the Guangxi pangolins (Wacharapluesadee et al. 2021) depart still further (these are called type IV for convenience, see Supplemental Fig. S2). For ease of comparison, the consensus secondary structures for types I to IV SL1 hairpins are shown in Supplemental Figure S3.

Sequence alignments for the NSP1 protein

The NSP1 sequences for the *Sarbecovirus* subgenus align very well. On the other hand, the sequences of the *Sarbecovirus* do not align well with those of either the

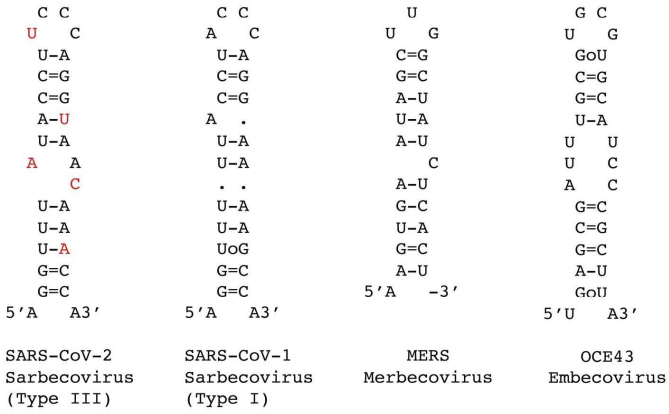


FIGURE 3. Consensus secondary structures derived from the sequence alignments for the SL1 hairpin present in SARS-CoV-2 (also called type III), SARS-CoV-1 (also called type I), *Merbecovirus*, and *Embecovirus*. The nucleotides in red show the five point-mutations between the type I and type III SL1 hairpins.

Merbecovirus or the *Embecovirus* (a Clustal-W multiple alignment is shown in Supplemental Fig. S4). The HKU1 sequence (*Embecovirus*) has a long additional carboxy-terminal end that is not shown here. The alignment yields 16 aligned identical residues (highlighted in cyan in Supplemental Fig. S4). The 16 aligned identical residues were not observed to vary in the alignments of *Sarbecovirus* types I, II, and III. Among those residues known experimentally to be functionally important, residues Y/F157 and R/K175 (indicated by #) align meaningfully.

Residues R124, K125, Y154, F157, K164, H165, R171, R175 are conserved throughout SARS-CoV-1 and SARS-CoV-2. Mutations K164A and H165A of SARS-CoV-1 NSP1 abolish binding to the 40S subunit (Kamitani et al. 2009). This is also true for SARS-CoV-2 NSP1; mutations of the key residues within the region 153–178 result in the loss of NSP1 binding to the ribosome (Supplemental Fig. S5; Schubert et al. 2020; Thoms et al. 2020). Likewise, the mutations R124A and K125A inhibit SARS-CoV-1 NSP1 ability to promote translation inhibition (Lokugamage et al. 2012). The NSP1 and SL1 sequences are therefore coupled so that a NSP1 sequence is correlated with a specific type of SL1 hairpin in *Sarbecovirus*. In the *Sarbecovirus*, most variations in the 5′UTR occur between type I (like SARS-CoV-1) and types II or III (like SARS-CoV-2) (Supplemental Fig. S3). Clearly, as shown by the NSP1 alignments, type II (Supplemental Figs. S6, S7) is closer to type III than to type I NSP1 (9 NSP1 mutations between II and III with up to 22 NSP1 mutations between I and II). This is also the case for type II and type III SL1 (Supplemental Fig. S3). However, there are conservations between type I and type II NSP1 sequences that do not occur in type III NSP1 sequences. Figure 5A illustrates the conservations of the NSP1 proteins on the alignments and Figure 5B on the available three-dimensional structures from SARS-CoV-1 and SARS-CoV-2. Further variants are discussed in the Supplemental Material section. Further descriptions and discussions about sequence variants in the 5′UTR and NSP1 protein can be found in the Supplemental Section.

Amino-acid swapping between SARS-CoV-1 and SARS-CoV-2 NSP1

Previous mutations have shown that conserved amino acids in the linker (R124A/K125A) and carboxy-terminal (K164A/H165A) regions prevent NSP1 to inhibit mRNA translation in both SARS-CoV-1 (Kamitani et al. 2009; Lokugamage et al. 2012) and SARS-CoV-2 (see Supplemental Fig. S4 of Tidu et al. 2021) and, thus, following the structural data (Schubert et al. 2020; Thoms et al. 2020; Yuan et al. 2020) should inhibit the binding of NSP1 to the mRNA channel of ribosomes.

NSP1 inhibits translation by blocking the mRNA channel on the ribosome. However, translation evasion is mediated by SL1 in the 5′UTR of viral mRNA transcripts (Tidu et al. 2021). We first purified recombinant NSP1 from SARS-CoV-1 and SARS-CoV-2 (Fig. 5B,C) and, using a canonical β-globin reporter mRNA, we found that the level of translation inhibition of SARS-CoV-1 NSP1 is significantly less than that with SARS-CoV-2 NSP1 (Supplemental Fig. S8A). Similarly, the translation evasion of cognate viral mRNA constructs was also less efficient with SARS-CoV-1 NSP1 than with SARS-CoV-2 NSP1 (Supplemental Fig. S8B,C). Considering these differences and the overall folding differences present in the available structural data, we undertook residues swapping experiments between NSP1

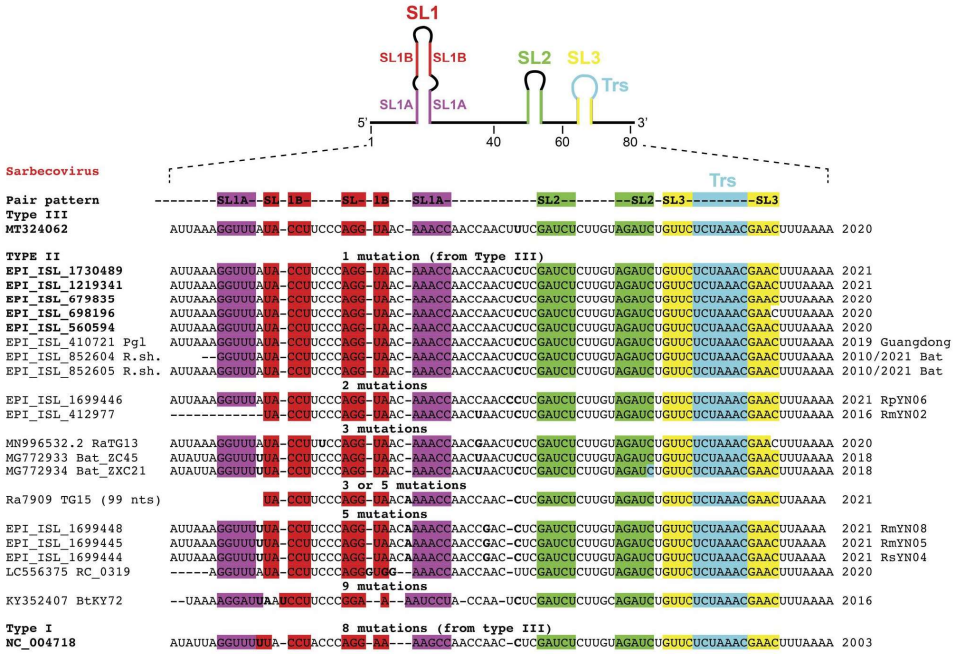


FIGURE 4. Aligned sequences of the first 80 nt of SARS-related genomes, called type II, isolated from various sources. They display variations with respect to both SARS-CoV-1 (type I) and SARS-CoV-2 (type III) sequences. The variations occur either in SL1 or in the upstream segment before SL2. Same annotations as in Figure 1.

proteins and SL1 from both viruses to assess the functional links between them. We decided to keep the SARS-CoV-2 NSP1 background, because it is the most efficient both in translation inhibition and in viral translation evasion, and residues from SARS-CoV-1 NSP1 were inserted into the SARS-CoV-2 NSP1 background. In total, eight SARS-CoV-2 recombinant NSP1 proteins with residues from SARS-CoV-1 were expressed and purified.

The key positions were chosen in the following way. We looked at two sides of NSP1 by structural comparisons between the available PDB models of SARS-CoV-1 and SARS-CoV-2 NSP1 proteins (Fig. 5B): the front side (left panels of Fig. 5B) and the back side (right panels of Fig. 5B). The previously studied R99A mutation (Mendez et al. 2021) demonstrated that this mutation abrogates viral translation evasion while reducing the translation inhibition activity of NSP1 against cellular mRNA. Since R99 is located on the front side of the NSP1, this suggests that SL1 might interact by contacts with residues located on the same front side. Our hypothesis is that the SARS-CoV-2 NSP1 surface coevolved with its 5′UTR mRNA by adjusting this front side for better SL1 recognition. In agreement with this hypothesis, there are more positive charges (susceptible to interact with negatively charged RNA) present on the front side of the protein than on the back side (see electrostatic surfaces of the right panels of Fig. 5B) for both SARS-CoV-1 and SARS-CoV-2 NSP1. Therefore, all residues swapped from SARS-CoV-1 into a SARS-CoV-2 background are located on the front side of NSP1. The remaining amino acid differences between SARS-CoV-1 and SARS-CoV-2 NSP1 proteins are located on back side of the proteins (right panels of Fig. 5B) and for that reason were not considered in this study.

Previous experiments had shown that the first twelve amino-terminal amino acids are critical for the cellular functions of NSP1 during viral replication (Yuan et al. 2020). Therefore, we also generated a truncated version of SARS-CoV-2 NSP1 called Δ12. Similarly, we introduced SARS-CoV-1 residues into the SARS-CoV-2 background at positions 6 and 8, both localized within the deleted part of Δ12. Figure 5D shows the purified eight recombinant NSP1 mutants.

Next, we tested their ability to promote translation in cell-free translation assays of a reporter containing the

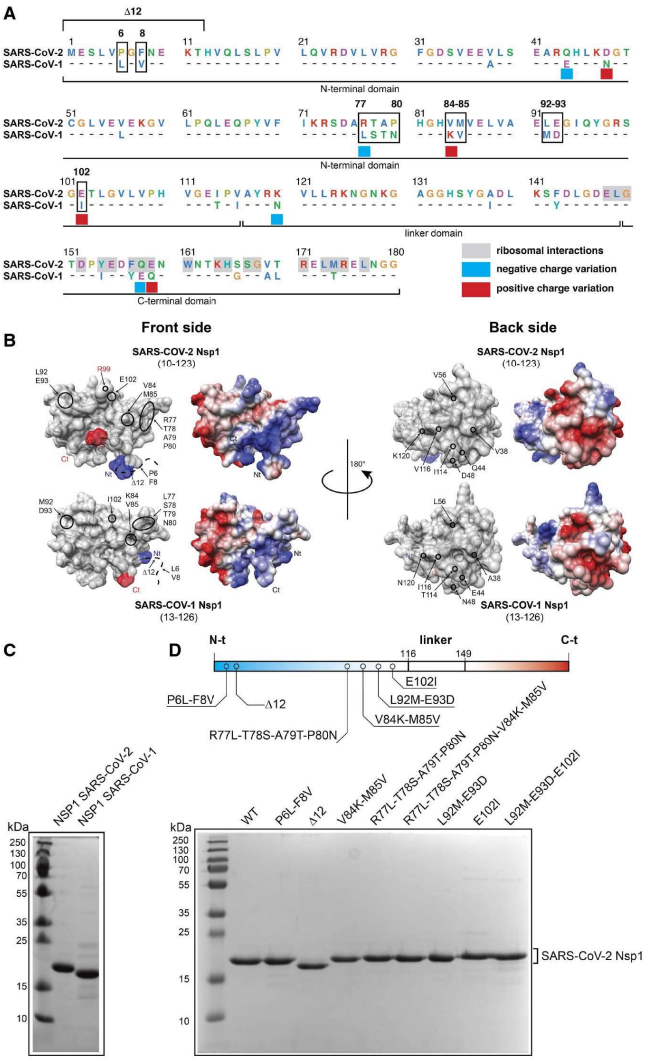


FIGURE 5. NSP1 proteins from SARS-CoV-1 and SARS-CoV-2. (*A*) Protein sequence alignment of SARS-CoV-2 and SARS-CoV-1 NSP1 proteins. For SARS-CoV-1, only the divergent amino acids are shown. The amino acids are shown according to the following color code: negatively charged amino acids in pink, hydrophobic amino acids in blue, positively charged amino acids in green, aromatic amino acids in cyan, glycines and prolines in orange. Residues involved in interactions with ribosomal components are boxed in gray in SARS-CoV-2 (Schubert et al. 2020; Thoms et al. 2020; Yuan et al. 2020). Negative charge variations from SARS-CoV-2 to SARS-CoV-1 are indicated by blue squares, and positive charge variations are indicated by red squares. Residues that have been mutated in this study are boxed in black. The NSP1 proteins are subdivided in three domains, the amino-terminal domain, the central linker domain, and a carboxy-terminal domain. (*B*) Surface representation of crystal structure of SARS-CoV-2 NSP1 from residues E10 to L123 (PDB: 7K7P) (Clark et al. 2021) and NMR structure of SARS-CoV-1 NSP1 from residues H13 to G126 (PDB: 2HSX) (Almeida et al. 2007) from the front side (*left* panels) and from the back side (*right* panels). Both structures have been aligned and represented with the amino-terminal end in blue and the carboxy-terminal end in red. Divergent residues from SARS-CoV-1 that have been inserted in SARS-CoV-2 are circled in black. Next to each structure, the electrostatic surfaces of the protein with negative and positive charges are colored in red and blue, respectively. The position of residue R99 in SARS-CoV-2 NSP1 is indicated in red. (*C*) SDS PAGE analysis of purified recombinant wild-type SARS-CoV-1 and SARS-CoV-2 NSP1 proteins. (*D*) Linear representation of the three domains of NSP1, the Nt-domain in blue, the linker domain and the Ct-domain in red. The SARS-CoV-2 residues that have been mutated are marked. SDS PAGE analysis of purified recombinant wild-type and mutant SARS-CoV-2 NSP1 proteins.

Renilla sequence with the WT 5′UTR SL1 of SARS-CoV-1, SARS-CoV-2, and a one point-mutation in the SARS-CoV-2 sequence of the apical loop to mimic the SARS-CoV-1 apical loop (Mut3, see Fig. 6A). First, we noticed a slight stimulation with 0.1 µM WT SARS-CoV-2 NSP1 with mRNA containing the WT SL1. Although we cannot offer a direct explanation of this reproducible phenomenon, we speculate that the slight boost in translation results from the increased pool of ribosomes released from endogeneous cellular mRNAs that were specifically inhibited by NSP1. With WT SARS-CoV-2 NSP1, the SL1 mutant Mut3 is significantly less efficient for translation evasion to NSP1 and behaves like the SARS-CoV-1 SL1, as could be expected from its similarity with the SARS-CoV-1 hairpin loop (Fig. 6B; Supplemental Fig. S9). Previous experiments had shown that the first twelve amino-terminal amino acids are critical for the cellular functions of NSP1 during viral replication (Yuan et al. 2020). In agreement with these studies, the removal of the first 12 amino-terminal amino acids of the SARS-CoV-2 NSP1 protein (Fig. 6C; Supplemental Fig. S9) leads to a marked decrease in the efficiency of the translational evasion to inhibition with a less pronounced differential effect on the three tested SL1 hairpins. Thus, the deletion of the amino-terminal region does not prevent ribosomal binding, but it does prevent the evasion from the inhibition mediated by SL1 hairpin. In the presence of the NSP1 mutant P6L/F8V, where the amino acid reversals follow the conservation of the NSP1 sequence observed in SARS-CoV-1, the evasion to inhibition is improved especially for Mut3 and the SARS-CoV-1 hairpins (Fig. 6D; Supplemental Fig. S9). These latter experiments show that the amino-terminal region is necessary for

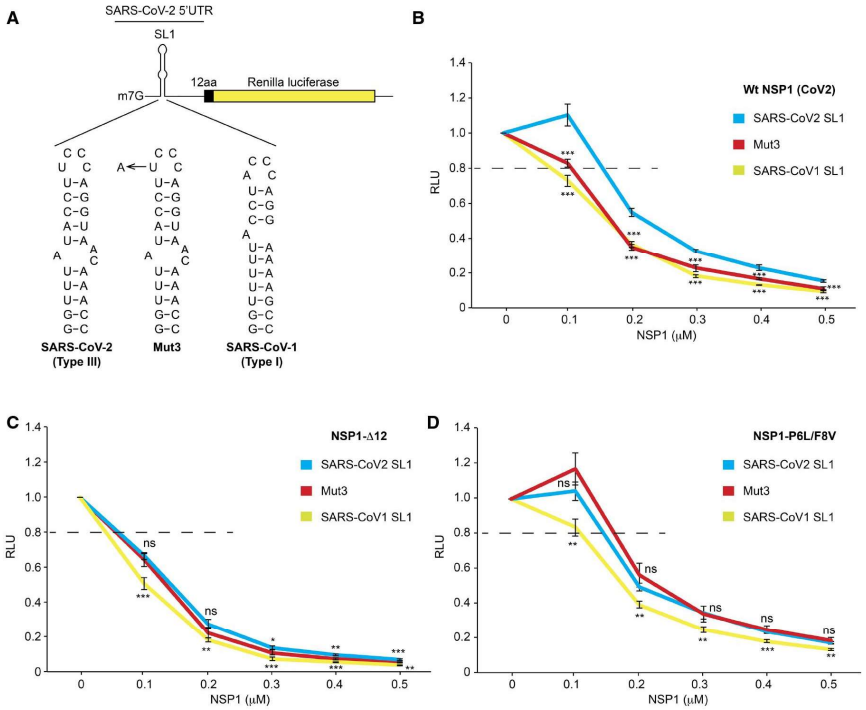


FIGURE 6. Evasion to NSP1-mediated translation inhibition with NSP1 variants containing mutations in the amino-terminal domain of NSP1. (*A*) Cartoon depicting the reporter mRNAs used for translation in RRL, they contain the SARS-CoV-2 5′UTR and the first 12 amino-terminal codons of NSP1 fused to the *Renilla* luciferase coding sequence. The reporter mRNA contains the wild-type SL1 hairpin or mut3 which has a U to A substitution in the loop or the SARS-CoV-1 SL1. (*B*) Curves representing the average relative activities measured after translation in RRL in the absence or presence of 0.1, 0.2, 0.3, 0.4, or 0.5 µM of wild-type NSP1, (*C*) Δ12, and (*D*) P6L/F8V. The averages are obtained from four independent experiments. Standard deviations or translational activity for each transcript are shown and calculated from four independent experiments. ns: non-significant; (*) 0.05 > *P*-value >0.01; (**) 0.01 > *P*-value >0.001; (***) *P*-value <0.001; based on Student's *t*-test. To facilitate comparisons between NSP1 proteins, the results have also been presented according to the hairpin SL1 tested (Supplemental Fig. S9).

evasion of the inhibitory effect and that the two residues at 6 and 8 either specifically directly interact with or contribute to interactions between NSP1 and the SL1 hairpin. We also tested the mutant P6L/F8V with the whole 5′UTR of SARS-CoV-1 and observed that the viral translation evasion is as low as with both SARS-CoV-1 and SARS-CoV-2 NSP1. This confirms that the SARS-CoV-1 SL1 in its 5′UTR context does not promote efficient viral evasion (Supplemental Fig. S10).

The NSP1 mutant V84K/M85V displays stronger effects than the NSP1 mutant R77L/T78S/A79T/P80N since their combination gives similar values than the double mutant V84K/M85V (compare Fig. 7A–C; Supplemental Fig. S11). Interestingly, the NSP1 mutant R77L/T78S/A79T/

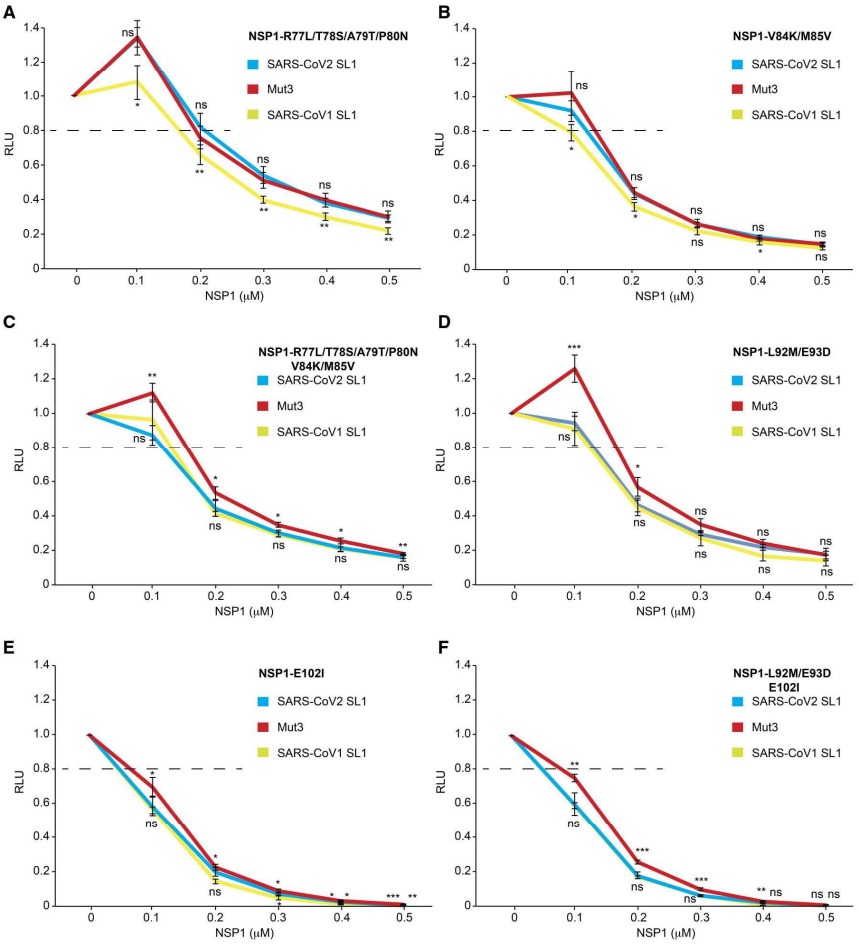


FIGURE 7. Evasion to NSP1-mediated translation inhibition with NSP1 variants containing mutations in the middle of NSP1. (*A*) Curves showing relative light unit (RLU) measured after translation in RRL in the absence or presence of 0.1, 0.2, 0.3, 0.4, or 0.5 µM of R77L/T78S/A79T/P80N, (*B*) V84K/M85V, (*C*) R77L/T78S/A79T/P80N/V84K/M85V, (*D*) L92M/E93D, (*E*) E102I, and (*F*) L92M/E93D/E102I. The averages are obtained from four independent experiments. Standard deviations or translational activity for each transcript are shown and calculated from four independent experiments. ns: nonsignificant; (*) 0.05 > *P*-value >0.01; (**) 0.01 > *P*-value >0.001; (***) *P*-value <0.001; based on Student's *t*-test. To facilitate comparisons between NSP1 proteins, the results have also been presented according to the hairpin SL1 tested (Supplemental Fig. S11).

P80N is the only mutant with an apparent reduction of the blockage of the mRNA channel (Fig. 7A; Supplemental Fig. S11).

We next introduced other amino acids typical of SARS-CoV-1 into the SARS-CoV-2 NSP1 protein in the region necessary for the conformational change that facilitates evasion of the blocking of the mRNA channel. Thus, the mutation E102I (Fig. 7E; Supplemental Fig. S11), at the end of the amino-terminal segment before the linker region (Fig. 5B,D), strongly prevents evasion from the inhibitory effect of NSP1, meaning again that the NSP1–E102I mutant binds tightly to the initiation complex and is not displaced by the SL1 5′UTR region. The latter mutant has a stronger dominant effect than the double NSP1 mutant L92M/E93D, since the combination of the three point-mutations, L92M/E93D and E102I, has effects like those observed with the single mutant E102I (compare Fig. 7D–F). Interestingly, E102 is very close to R99 (Fig. 5B) and our result is in good agreement with the previously described R99A mutation, which has a lower propensity to promote viral translation evasion (Mendez et al. 2021). Altogether, Mut3 is promoting better evasion to NSP1 inhibition when specific residues from SARS-CoV-1 are introduced into SARS-CoV-2 NSP1 thereby confirming the functional link between these specific residues and the first nucleotide of the apical loop of SL1.

DISCUSSION

Here, sequence analysis has been performed on a short non-coding RNA segment at the 5′UTR segment of the beta-coronaviruses, and especially the *Sarbecovirus* family responsible for a devastating pandemic that started at the end of 2019 in China. Although such sequence elements are not known to contribute to either infectivity or cell penetration, they are essential for the translation of the viral genome and for its replicative proliferation. The RNA analysis is complemented by the sequence analysis of the first non-structural protein translated by the virus, NSP1, that has been shown to be involved, together with the 5′UTR RNA, in the control of ribosomal translation initiation (Schubert et al. 2020; Thoms et al. 2020; Mendez et al. 2021; Tidu et al. 2021). We contribute further data showing that both the 5′UTR and the NSP1 protein are linked functionally and therefore are forced to coevolve to interact cooperatively with highly conserved elements of the ribosomal machinery and initiation cofactor proteins. The selection pressure on both the RNA and the protein is patent in the analyzed sequences, as revealed by the patterns of conservation and variation in the NSP1 sequences for each type of 5′UTR and by the experimental mutational data on both the NSP1 protein and the SL1 hairpin of the 5′UTR that are presented. The experimental mutational data confirm that the carboxy-terminal region is central to NSP1 binding to the mRNA channel, while the amino-terminal domain and the linker region contributes both to the conformational rearrangement and to binding to the SL1 hairpin of the 5′UTR that promotes evasion from the NSP1 inhibition. In short, the data concur with a model in which the Nt-domain interacts directly or indirectly with SL1 and thereby promotes remodeling of the Ct-domain out of the mRNA channel of the ribosome. The data support also the view that the linker domain mediates the communication between these two domains during viral evasion.

The question of the origins of the coronavirus that initiated the present pandemic is much debated. Although the data presented here are limited, they do set strong molecular constraints on the viral sequences for successful translation and replication in specific living cells. The various types of the 5′UTR first 80 nt reproduce the main branches of the various phylogenetic trees deduced from whole genome comparisons or functional genomic segments. Figure 8 shows an overall summary of key 5′UTR consensus sequences following the accepted phylogenetic tree.

Finally, one may remark the following. In the 5′UTR, there are seven mutations between type I and II: six mutations in SL1 and a 1-nt U insertion in the 3′ segment preceding the SL2 hairpin. Between type II and III sequences derived from infected humans, there is a single mutation of a C into a U in the same 3′-end part of SL2 (but there are more mutations from bat derived sequences, see Fig. 4). The differences may appear minimal, but their maintenance within each subgroup indicate that they are meaningful. These structures must fold, interact with proteins, and unfold with an appropriate dynamic that depends on the free energies of the hairpins, which themselves depend on the cellular environment (ions, temperature, …) that vary between species. Besides, and importantly, the formation of alternative structures in dynamic equilibrium, some of which are also key for functional initiation, cannot be excluded and should be investigated.

The experimental data show that the coevolved complex between NSP1 and the 5′UTR has virus specificity and that mutations in one component may affect their mutual interactions. In the absence of precise structural information, many unanswered questions remain on the complex interactions involving the 5′UTR leader sequence, the NSP1 protein and the ribosomal machinery, like (i) why did the 5′UTR and NSP1 sequences mutate from the SARS-CoV-1 in the SARS-CoV-2 genomes; (ii) how to explain the weak variations in the SARS-CoV-2 5′UTR genomes? For SARS-CoV-1, similar sequences of the 5′UTR and NSP1 proteins were observed from viruses isolated from bats or infected humans and, thus, both are adapted to ribosomal translation cofactors in both species. However, the SARS-CoV-2 5′UTR and NSP1 protein are only observed in infected humans (with later some other mammals like minks, probably infected by humans) and, up to now, they were not observed in bats (see Supplemental Material). In this work, we did not study all

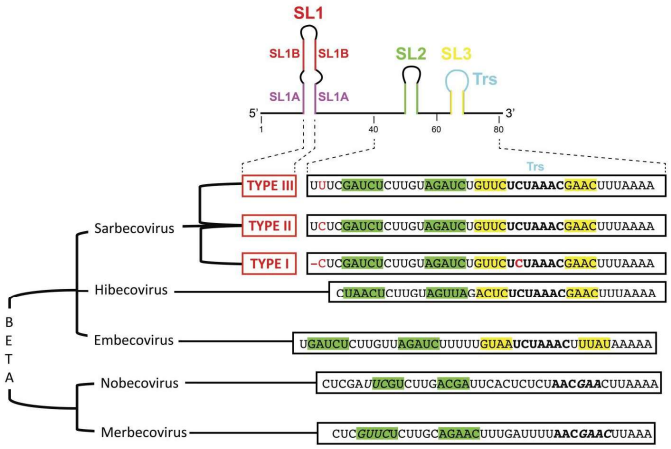


FIGURE 8. A rough phylogenetic tree of the Genus beta of the *Coronaviridae* with the main subgenus species (for more details, see Supplemental Fig. S1A). The consensus sequences derived here for the first 80 nt of the 5′ leader are shown on each branch, with the SL1 nomenclature indicated on the *right*. The color code is the same as in the alignments (green, SL2; yellow and cyan, SL3). TRS sequences are in bold. In italics are shown potential alternative pairings. In red are shown nucleotides that vary between types I, II, and III in the 5′ segment upstream of SL2.

the steps in which the first 80 nt of the 5′UTR are involved, and some of the observed conserved elements may play a role in processes not experimentally probed yet.

MATERIALS AND METHODS

Sequences

Viral sequences were retrieved form NCBI Genbank (http://www.ncbi.nlm.nih.gov), the Gisaid repository (http://www.GISAID.org) (Elbe and Buckland-Merrett 2017), or the National Genomics Data Center, China National Center for Bioinformation (Beijing Institute of Genomics, Chinese Academy of Sciences) (https://bigd.big.ac.cn/) and aligned manually. For the RNA, the alignments were done on the first 80 nt of the genome (when available) and the complete sequence was considered for the protein. The NSP1 protein sequences were derived using the EMBOSS program (https://www.ebi.ac.uk/Tools/st/emboss_transeq/) based on the published genomes. The correspondence between the accession numbers and the usual or common ID strain is given in the figures and tables.

In vitro transcription

The three reporter constructs were transcribed by run-off in vitro transcription with T7 RNA polymerase from DNA templates amplified by PCR. A DNA fragment containing the 900 first nucleotides of SARS-COV-2 (accession number: MN908947.3) was used as a template for PCR amplifications. The transcription was performed in the presence of m7GpppG cap analog (Jena Bioscience) to generate capped mRNAs as previously described (Tidu et al. 2021).

In vitro translation

In vitro translation with cell-free translation extracts were performed using self-made rabbit reticulocyte lysates (RRL) as previously described (Tidu et al. 2021). Briefly, reactions were incubated at 30°C for 60 min and included 200 nM of each transcript. Aliquots of translation reactions were analyzed for *Renilla* luciferase activity on a luminometer.

NSP1 overexpression and purification

NSP1 and derivatives (P6L/F8V-Δ12-R77L/T78S/A79T-P80N-V84K/M85V-L92M/E93D) were produced as previously described (Tidu et al. 2021) with slight modifications. Briefly, the plasmid pET-His-GST-TEV-LIC-(2GT) was purchased from Addgene. The fusion proteins were expressed in *E. coli* BL21 Rosetta (DE3) pLysS cells. Cells were grown at 37°C to a cell density of OD600 = 0.6 and then induced with 0.2 mM IPTG during 3 h at 37°C. After induction, cells were harvested and the cell pellets were frozen at −80°C. Frozen pellets were resuspended in EQ/W buffer (40 mM Na phosphate pH 7.2, 500 mM NaCl, 30 mM imidazole) supplemented with 0.1% Triton X-100, cOmplete Protease Inhibitor Cocktail (Merck) and incubated on ice for 30 min with 1 mg/mL lysozyme. After lysis by sonication, the cell lysate was centrifuged at 3200*g* at 4°C for 10 min to remove the large debris, and the supernatant was further centrifuged at 150,000*g* at 4°C for 30 min. The supernatant was applied to Ni-NTA Superflow resin (Qiagen) equilibrated in buffer EQ/W.

After column washing, 6xHis-GST-TEV-NSP1 proteins were eluted from the resin by buffer EQ/W with gradient increase of imidazole from 30 to 500 mM. Fractions containing 6xHis-GST-TEV-NSP1 were pooled and dialyzed against buffer EQ/W without imidazole overnight. Next, the 6xHis-GST-TEV-NSP1 fraction was loaded on Glutathione HiCap resin (Qiagen) equilibrated with the dialysis buffer, and proteins were eluted by the same buffer with a gradient of glutathione from 0 to 50 mM. Fractions containing 6xHis-GST-TEV-NSP1 fusion proteins were pooled and subjected to TEV protease cleavage overnight at 4°C (50/1 fusion/TEV molar ratio). NSP1 proteins were separated from the 6xHis-GST domains using a last purification step on Ni-NTA resin that retained His-tagged GST and TEV proteins. The pure NSP1 proteins were concentrated and stored in buffer that contains 50% glycerol at −20°C.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work is funded by Agence Nationale pour la Recherche (ANR-17-CE12-0025-01, ANR-17-CE11-0024, ANR-20-COVI-0078), Fondation pour la Recherche Médicale (project CoronaIRES), Fondation Bettencourt Schueller, University of Strasbourg and the Centre National de la Recherche Scientifique.

Received September 2, 2021; accepted February 8, 2022.

REFERENCES

Almeida MS, Johnson MA, Herrmann T, Geralt M, Wüthrich K. 2007. Novel β-barrel fold in the nuclear magnetic resonance structure of the replicase nonstructural protein 1 from the severe acute respiratory syndrome coronavirus. *J Virol* **81:** 3151–3161. doi:10.1128/JVI.01939-06

Brian DA, Baric RS. 2005. Coronavirus genome structure and replication. *Curr Top Microbiol Immunol* **287:** 1–30. doi:10.1007/3-540-26765-4_1

Clark LK, Green TJ, Petit CM. 2021. Structure of nonstructural protein 1 from SARS-CoV-2. *J Virol* **95:** e02019-20. doi:10.1128/JVI.02019-20

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Challenges* **1:** 33–46. doi:10.1002/gch2.1018

Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, Haagmans BL, Lauber C, Leontovich AM, Neuman BW, et al. 2020. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5:** 536–544. doi:10.1038/s41564-020-0695-z

Gulyaeva AA, Gorbalenya AE. 2021. A nidovirus perspective on SARS-CoV-2. *Biochem Biophys Res Commun* **538:** 24–34. doi:10.1016/j.bbrc.2020.11.015

Guo H, Hu B, Si H-R, Zhu Y, Zhang W, Li B, Li A, Geng R, Lin H-F, Yang X-L, et al. 2021. Identification of a novel lineage bat SARS-related coronaviruses that use bat ACE2 receptor. *Emerg Microbes Infect* **10:** 1507–1514. doi:10.1080/22221751.2021.1956373

Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, Yang L, Ding C, Zhu X, Lv R, et al. 2018. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. *Emerg Microbes Infect* **7:** 154. doi:10.1038/s41426-018-0155-5

Hul V, Delaune D, Karlsson EA, Hassanin A, Tey PO, Baidaliuk A, Gámbaro F, Tu VT, Keatts L, Mazet J, et al. 2021. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *bioRxiv* doi:10.1101/2021.01.26.428212

Kamitani W, Huang C, Narayanan K, Lokugamage KG, Makino S. 2009. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nat Struct Mol Biol* **16:** 1134–1140. doi:10.1038/nsmb.1680

Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* **181:** 914–921.e10. doi:10.1016/j.cell.2020.04.011

Lai MM, Stohlman SA. 1981. Comparative analysis of RNA genomes of mouse hepatitis viruses. *J Virol* **38:** 661–670. doi:10.1128/jvi.38.2.661-670.1981

Lapointe CP, Grosely R, Johnson AG, Wang J, Fernández IS, Puglisi JD. 2021. Dynamic competition between SARS-CoV-2 NSP1 and mRNA on the human ribosome inhibits translation initiation. *Proc Natl Acad Sci* **118:** e2017715118. doi:10.1073/pnas.2017715118

Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* **46:** D708–D717. doi:10.1093/nar/gkx932

Liu P, Jiang JZ, Wan XF, Hua Y, Li L, Zhou J, Wang X, Hou F, Chen J, Zou J, et al. 2020. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLoS Pathog* **16:** e1008421. doi:10.1371/journal.ppat.1008421

Lokugamage KG, Narayanan K, Huang C, Makino S. 2012. Severe acute respiratory syndrome coronavirus protein nsp1 is a novel eukaryotic translation inhibitor that represses multiple steps of translation initiation. *J Virol* **86:** 13598–13608. doi:10.1128/JVI.01958-12

Mendez AS, Ly M, González-Sánchez AM, Hartenian E, Ingolia NT, Cate JH, Glaunsinger BA. 2021. The *N*-terminal domain of SARS-CoV-2 nsp1 plays key roles in suppression of cellular gene expression and preservation of viral gene expression. *Cell Rep* **37:** 109841. doi:10.1016/j.celrep.2021.109841

Miao Z, Tidu A, Eriani G, Martin F. 2021. Secondary structure of the SARS-CoV-2 5′-UTR. *RNA Biol* **18:** 447–456. doi:10.1080/15476286.2020.1814556

Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, Nicholls J, Yee WKS, Yan WW, Cheung MT, et al. 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet* **361:** 1319–1325. doi:10.1016/S0140-6736(03)13077-2

Rangan R, Zheludev IN, Hagey RJ, Pham EA, Wayment-Steele HK, Glenn JS, Das R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *RNA* **26:** 937–959. doi:10.1261/rna.076141.120

Schubert K, Karousis ED, Jomaa A, Scaiola A, Echeverria B, Gurzeler LA, Leibundgut M, Thiel V, Mühlemann O, Ban N. 2020. SARS-CoV-2 Nsp1 binds the ribosomal mRNA channel to inhibit translation. *Nat Struct Mol Biol* **27:** 959–966. doi:10.1038/s41594-020-0511-8

Smith EC, Sexton NR, Denison MR. 2014. Thinking outside the triangle: replication fidelity of the largest RNA viruses. *Annu Rev Virol* **1:** 111–132. doi:10.1146/annurev-virology-031413-085507

Thoms M, Buschauer R, Ameismeier M, Koepke L, Denk T, Hirschenberger M, Kratzat H, Hayn M, MacKens-Kiani T, Cheng J, et al. 2020. Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2. *Science* **369:** 1249–1256. doi:10.1126/science.abc8665

Tidu A, Janvier A, Schaeffer L, Sosnowski P, Kuhn L, Hammann P, Westhof E, Eriani G, Martin F. 2021. The viral protein NSP1 acts

as a ribosome gatekeeper for shutting down host translation and fostering SARS-CoV-2 translation. *RNA* **27:** 253–264. doi:10.1261/rna.078121.120

Vijgen L, Keyaerts E, Moës E, Thoelen I, Wollants E, Lemey P, Vandamme A-M, Van Ranst M. 2005. Complete genomic sequence of human coronavirus OC43: molecular clock analysis suggests a relatively recent zoonotic coronavirus transmission event. *J Virol* **79:** 1595–1604. doi:10.1128/JVI.79.3.1595-1604.2005

Wacharapluesadee S, Tan CW, Maneeorn P, Duengkae P, Zhu F, Joyjinda Y, Kaewpom T, Chia WN, Ampoot W, Lim BL, et al. 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat Commun* **12:** 972. doi:10.1038/s41467-021-21240-1

Woo PCY, Huang Y, Lau SKP, Tsoi HW, Yuen KY. 2005a. *In silico* analysis of ORF1ab in coronavirus HKU1 genome reveals a unique putative cleavage site of coronavirus HKU1 3C-like protease. *Microbiol Immunol* **49:** 899–908. doi:10.1111/j.1348-0421.2005.tb03681.x

Woo PCY, Lau SKP, Huang Y, Tsoi HW, Chan KH, Yuen KY. 2005b. Phylogenetic and recombination analysis of coronavirus HKU1, a novel coronavirus from patients with pneumonia. *Arch Virol* **150:** 2299–2311. doi:10.1007/s00705-005-0573-2

Wu Z, Yang L, Ren X, He G, Zhang J, Yang J, Qian Z, Dong J, Sun L, Zhu Y, et al. 2016. Deciphering the bat virome catalog to better understand the ecological diversity of bat viruses and the bat origin of emerging infectious diseases. *ISME J* **10:** 609–620. doi:10.1038/ismej.2015.138

Yogo Y, Hirano N, Hino S, Shibuta H, Matumoto M. 1977. Polyadenylate in the virion RNA of mouse hepatitis virus. *J Biochem* **82:** 1103–1108. doi:10.1093/oxfordjournals.jbchem.a131782

Yuan S, Peng L, Park JJ, Hu Y, Devarkar SC, Dong MB, Shen Q, Wu S, Chen S, Lomakin IB, et al. 2020. Nonstructural protein 1 of SARS-CoV-2 is a potent pathogenicity factor redirecting host protein synthesis machinery toward viral RNA. *Mol Cell* **80:** 1055–1066.e6. doi:10.1016/j.molcel.2020.10.034

Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med* **367:** 1814–1820. doi:10.1056/NEJMoa1211721

Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, Wang P, Liu D, Yang J, Holmes EC, et al. 2020a. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr Biol* **30:** 2196–2203.e3. doi:10.1016/j.cub.2020.05.023

Zhou P, Lou YX, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, et al. 2020b. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579:** 270–273. doi:10.1038/s41586-020-2012-7

Zhou H, Ji J, Chen X, Bi Y, Li J, Wang Q, Hu T, Song H, Zhao R, Chen Y, et al. 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* **184:** 4380–4391.e14. doi:10.1016/j.cell.2021.06.008

2.3.5. Perspectives

Le mécanisme précis de la traduction de l'ARN viral n'est pas encore élucidé. Pour l'heure, notre laboratoire et d'autres ont clairement démontré la nécessité de la présence de SL1 à proximité de l'extrémité 5' des ARN viraux en présence de NSP1 pour leur traduction efficace. Notre modèle suggère une interaction entre SL1 et NSP1 au moment du recrutement de la particule 43S à l'extrémité 5' de l'ARN qui délogerait NSP1 du canal d'entrée de l'ARNm dans le ribosome. Seule une étude structurale permettrait de comprendre la mécanique moléculaire de ce modèle. Une étude par cryo-microscopie électronique est cours au laboratoire pour établir un modèle structural à partir de complexes de pré-initiation 48S assemblés sur le codon d'initiation de l'ARN viral en présence de NSP1. Etudier le positionnement relatif de l'ARN viral et des hélices C-terminales de NSP1 à proximité du canal d'entrée de l'ARNm au niveau du site A du ribosome permettra d'établir si NSP1 est délogée directement ou indirectement pour autoriser la traduction de l'ARN viral. Par ailleurs, cette étude fournira peut-être des informations structurales sur le domaine N-terminal de NSP1 dont l'interaction avec des facteurs *cis* ou *trans* pourrait être impliquée dans ce mécanisme. Jusqu'à présent, la flexibilité de ce domaine n'a pas permis de pouvoir l'observer à résolution atomique.

Par ailleurs, les premières études de la protéine NSP1 ont démontré qu'elle induit la dégradation des ARNm cellulaires. La question est de savoir si cette activité nucléase est portée par NSP1 elle-même ou par une autre protéine cellulaire qu'elle recrute au niveau du ribosome, possiblement par son domaine N-terminal. Pour le moment, les analyses de spectrométries de masse réalisées n'ont pas permis d'identifier formellement une nucléase recrutée au niveau des complexes purifiés. Bien que la nucléase XRN1 soit apparue de manière significative dans certaines de nos expériences, cette piste n'a pas encore été explorée. Les ARNm cellulaires concernés par la dégradation médiée par NSP1 comprennent notamment les ARNm qui codent pour des protéines impliquées dans la réponse interféron, et font l'objet d'un projet de thèse démarré cette année au laboratoire.

Enfin, les aspects évolutifs des protéines NSP1 et de la tige-boucle SL1 seront davantage approfondis. Nous avons démontré que le couple NSP1-SL1 est conservé au sein des bêta-coronavirus. En revanche, dans d'autres coronavirus, la protéine NSP1 est significativement différente, et certains coronavirus possèdent une protéine NSP1 dépourvue du domaine de liaison au ribosome. Pourtant, ces virus éteignent également la traduction cellulaire lors de l'infection. Il sera alors intéressant de déterminer les mécanismes moléculaires mis en jeu pour inhiber la traduction cellulaire.

2.4. Contributions à d'autres projets de recherche

2.4.1. Mesure de la vitesse de scanning de la particule 43S

Ce projet a été initié en collaboration avec l'équipe de Karen Perronet du laboratoire Lumière, Matière et Interface (LuMIn, UMR 9024). L'objectif de ce travail est de mesurer la vitesse de scanning de la particule 43S avec une méthode d'étude en molécules uniques observées par microscopie. Une approche similaire a déjà été mise en œuvre par notre collaboratrice dans le cadre d'une étude pour mesurer la vitesse d'élongation (Bugaud *et al.* 2017).

Pour déterminer la vitesse de scanning, nous avons mis au point un système rapporteur qui utilise une version modifiée de l'IRES de EMCV pour permettre le scanning. Ce système a déjà été utilisé la partie 2.1.1.1 de ce travail. La région 5'UTR à scanner est linéaire et contient deux séquences d'hybridation pour deux oligonucléotides marqués chacun par un fluorophore de couleur différente. Ces deux séquences sont placées de part et d'autre d'un site qui permet l'insertion d'une région structurée d'intérêt, ce qui autorise l'étude de l'influence des structures secondaires de la région 5'UTR sur la vitesse de scanning. Le signal de fluorescence est émis lorsque les nucléotides se décrochent de l'ARNm. Un troisième oligonucléotide, dit d'accroche, est biotinylé et permet l'ancrage et l'immobilisation de l'ARN rapporteur sur une lame de microscopie recouverte de streptavidine. La lame est ensuite mise en présence de lysats de réticulocytes de lapin qui permettent d'assembler des complexes d'initiation de la traduction. Cette étude est réalisée par microscopie TIRF (Total Internal Reflection Fluorescence), chaque molécule d'ARN se trouvant dans un même plan focal. L'ARNm rapporteur utilisé pour l'étude est schématisé dans la Figure 31.

Après son recrutement au niveau de l'IRES, la particule 43S en cours de scanning déplace séquentiellement les oligonucléotides 1 et 2, ce qui résulte en deux émissions de fluorescence successives rouge puis verte. Les signaux de fluorescence émis par chaque fluorophore sont enregistrés pour chaque point correspondant à une molécule unique d'ARN sur la lame. Le temps séparant ces deux émissions est mesuré grâce à la prise d'images à intervalles de temps réguliers, et, connaissant la distance séparant les deux oligonucléotides, permet ensuite de calculer la vitesse de scanning de la région séparant les deux oligonucléotides.

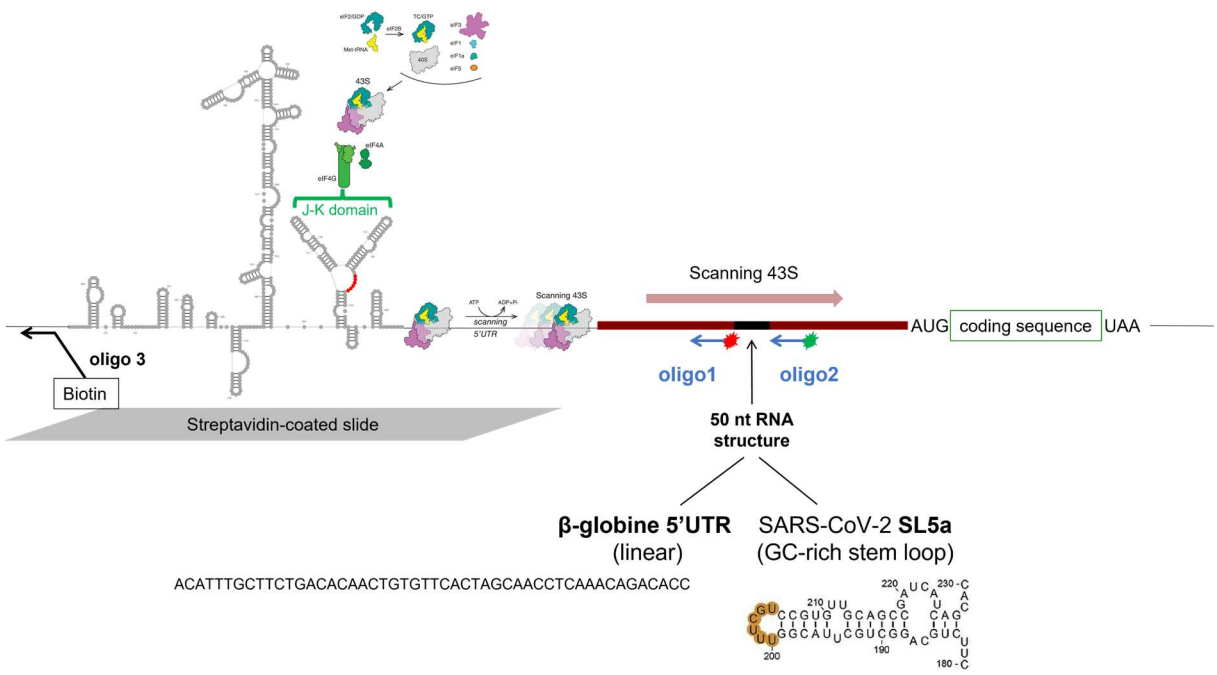


Figure 31 : schéma de l'ARN rapporteur utilisé pour mesurer la vitesse de scanning. L'initiation a lieu par l'intermédiaire de l'IRES EMCV dépourvu de son codon d'initiation AUG natif, obligeant ainsi le scanning de la région 5'UTR par la particule 43S recrutée. La région 5'UTR à scanner (rouge foncé) est principalement constituée de poly-CAA et est donc en principe non structurée. La zone d'insertion d'une région adoptant un repliement différent du reste de cette région 5'UTR est indiquée en noir. Y ont été insérées la région 5'UTR de la β-globine, décrite comme n'étant pas structurée, et la tige-boucle SL5a de la région 5'UTR du SARS-CoV-2, qui a été identifiée dans le projet d'étude de ce virus. Ici, l'ARN rapporteur est accroché côté 5' afin d'empêcher la traduction « leaderless ». Oligo : oligonucléotide.

Dans un premier temps, nous avons choisi de mesurer la vitesse de scanning de deux régions au repliement distinct mais de longueur identique : d'une part, la région 5'UTR de la β-globine, décrite comme n'étant pas structurée, et d'autre part la tige-boucle SL5a de la région 5'UTR du SARS-CoV-2, qui a été caractérisée dans le projet d'étude de la région 5'UTR de ce virus (partie 2.3.2). On s'attend à ce que la vitesse mesurée soit plus lente en présence de la structure secondaire.

Bien que les expériences soient encore en cours, une vitesse de scanning a déjà pu être déterminée. Lorsque la région 5'UTR n'est pas structurée, la vitesse de scanning est de 23 nucléotides par seconde, ce qui est à peu près comparable à la vitesse de la traduction (5,6 codons par seconde, soit presque 17 nucléotides par seconde) qui a été déterminée *in vivo* par des expériences de 'ribosome profiling' (Ingolia *et al.* 2011). Cette vitesse de scanning est la première mesure pour un système mammifère. Pendant la rédaction de ce mémoire, d'autres mesures ont été effectuées mais avec des extraits de traduction de levure (Wang *et al.* 2022).

Par la suite, une fois les expériences contrôles validées, nous étudierons l'influence d'autres structures secondaires sur la vitesse de scanning, ainsi que d'éventuels facteurs *trans*-régulateurs comme des hélicases.

2.4.2. Etude de l'impact sur la traduction de la restrictocine, une mycotoxine sécrétée par *Aspergillus fumigatus*, lors de l'infection de drosophiles

Dans cette étude, nous avons été sollicités par d'autres collaborateurs pour déterminer si la résistance des drosophiles à une toxine nommée restrictocine, un inhibiteur de la traduction synthétisé par le champignon *Aspergillus fumigatus* qui les infecte, est liée au niveau traductionnel par l'induction de la voie Toll et à la synthèse par la drosophile de peptides antimicrobiens nommés bomamines. Pour ce faire, nous avons préparé des extraits acellulaires de cellules embryonnaires de drosophile (S2) avec ou sans induction préalable de la voie Toll. Ensuite, des tests de traduction *in vitro* en présence de restrictocine et/ou de bomamines purifiées ont été réalisés à l'aide de ces extraits pour déterminer si l'induction de la voie Toll et/ou la présence des bomamines permettent de rétablir la traduction en présence de restrictocine. Les résultats montrent que ni l'un, ni l'autre, ni les deux combinés n'ont d'effet, ce qui suggère que la neutralisation de la restrictocine lors de l'infection a lieu de manière extracellulaire, par l'intermédiaire de molécules sécrétées par la drosophile. Ces résultats ont été publiés dans un article de la revue EMBO Reports dont nous sommes co-auteurs (Xu *et al.* 2022).

ARTICLE 4

L'induction de la voie Toll et la synthèse de bomamines qui s'ensuit permettent la résistance des drosophiles aux mycotoxines sécrétées par *Aspergillus fumigatus*

Article

TRANSPARENT PROCESS OPEN ACCESS

EMBO reports

The Toll pathway mediates *Drosophila* resilience to *Aspergillus* mycotoxins through specific Bomanins

Rui Xu[1,2,3,†], Yanyan Lou[1,2,3,†], Antonin Tidu[2,4], Philippe Bulet[5,6], Thorsten Heinekamp[7], Franck Martin[2,4], Axel Brakhage[7,8], Zi Li[1], Samuel Liégeois[1,2,3,*] & Dominique Ferrandon[1,2,3,**]

Abstract

Host defense against infections encompasses both resistance, which targets microorganisms for neutralization or elimination, and resilience/disease tolerance, which allows the host to withstand/tolerate pathogens and repair damages. In *Drosophila*, the Toll signaling pathway is thought to mediate resistance against fungal infections by regulating the secretion of antimicrobial peptides, potentially including Bomanins. We find that *Aspergillus fumigatus* kills *Drosophila* Toll pathway mutants without invasion because its dissemination is blocked by melanization, suggesting a role for Toll in host defense distinct from resistance. We report that mutants affecting the Toll pathway or the 55C *Bomanin* locus are susceptible to the injection of two *Aspergillus* mycotoxins, restrictocin and verruculogen. The vulnerability of 55C deletion mutants to these mycotoxins is rescued by the overexpression of Bomanins specific to each challenge. Mechanistically, flies in which *BomS6* is expressed in the nervous system exhibit an enhanced recovery from the tremors induced by injected verruculogen and display improved survival. Thus, innate immunity also protects the host against the action of microbial toxins through secreted peptides and thereby increases its resilience to infection.

Keywords *Drosophila melanogaster*; fumitremorgin/verruculogen; fungal infections; resilience/disease tolerance; restrictocin

Subject Categories Immunology; Microbiology, Virology & Host Pathogen Interaction; Signal Transduction

DOI 10.15252/embr.202256036 | Received 26 August 2022 | Revised 6 October 2022 | Accepted 14 October 2022

EMBO Reports (2022) e56036

Introduction

The outcome of an infection depends on the interactions between a host and a pathogen with its armory of multifarious virulence factors. In the case of fungal pathogens, several hundred potential virulence factors are known to be secreted (Gao *et al*, 2011; Lebrigand *et al*, 2016). The host confronts the invading microorganism through the multiple arms of its immune system. There are also varied strategies that counteract the effect of toxins and more generally withstand and repair damages inflicted directly by the pathogen or indirectly by the host through its own immune response (Medzhitov *et al*, 2012; Ferrandon, 2013; Soares *et al*, 2017). Fungal infections represent a widespread major health threat worldwide affecting more than 150 million patients and cause directly or indirectly at least one and a half million deaths each year (Bongomin *et al*, 2017; Rodrigues & Nosanchuk, 2020). Our current understanding of fungal infections relies on the study of the host's innate and adaptive immune responses and in parallel on investigations of fungal virulence factors (Scharf *et al*, 2014; van de Veerdonk *et al*, 2017). *Aspergillus fumigatus* can synthesize and secrete a vast array of toxins and secondary metabolites, the *in vivo* functions of which are just starting to be deciphered (Frisvad *et al*, 2009; Macheleidt *et al*, 2016; Raffa & Keller, 2019). Whereas some fungal virulence factors allow *A. fumigatus* to elude detection by the immune system, a few mycotoxins such as gliotoxin or fumagillin are known to interfere with immune signaling and help neutralize immune cell functions (Cramer *et al*, 2006; Kupfahl *et al*, 2006; Konig *et al*, 2019). However, it is currently poorly known whether the innate immune system is able to detect and counteract the actions of mycotoxins through specific effectors.

Drosophila melanogaster represents a genetically amenable model system that is well-suited to study infections and innate immunity as there is no vertebrate-like adaptive immunity. Its innate immune system comprises several arms: a cellular response mediated by plasmatocytes in the adult, melanization, which depends on the cleavage of prophenol oxidases by Hayan, and the humoral systemic immune

1 Sino-French Hoffmann Institute, Guangzhou Medical University, Guangzhou, China
2 Université de Strasbourg, Strasbourg, France
3 Modèles Insectes de l'Immunité Innée, UPR 9022 du CNRS, Strasbourg, France
4 Architecture et Réactivité de l'ARN, UPR 9002 du CNRS, Strasbourg, France
5 CR Université Grenoble Alpes, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Grenoble, France
6 Platform BioPark Archamps, Archamps, France
7 Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (Leibniz-HKI), Jena, Germany
8 Institute of Microbiology, Friedrich Schiller University Jena, Jena, Germany
*Corresponding author. Tel: +33 388417107; E-mail: s.liegeois@unistra.fr
**Corresponding author. Tel: +33 388417017; E-mail: d.ferrandon@ibmc-cnrs.unistra.fr
†These authors contributed equally to this work

response (Lemaitre & Hoffmann, 2007; Liegeois & Ferrandon, 2022). A landmark study published 25 years ago established the central role of the Toll pathway in mediating the humoral immune response against fungal infections, as represented by *A. fumigatus* (Lemaitre *et al*, 1996). This observation has since been extended to a variety of other filamentous fungi or pathogenic yeast infections and also to several Gram-positive bacterial infections. The current paradigm is that upon sensing infections, a MyD88 adaptor-dependent intracellular signaling pathway gets activated by the binding of the processed Spätzle cytokine to the Toll receptor and stimulating the transcription of effector genes (such as antimicrobial peptides (AMPs)) that mediate its role in host defense. Genes encoding antifungal peptides such as Drosomycin, Metchnikowin, and Daisho are Toll-regulated AMPs active on filamentous fungi (Fehlbaum *et al*, 1995; Levashina *et al*, 1995; Cohen *et al*, 2020). However, in contrast to the other NF-κB signaling pathway that mediates host defense against Gram-negative bacteria, Immune deficiency (IMD), it is less clear whether Toll-dependent AMPs provide the bulk of the protection against Gram-positive or fungal infections. Indeed, the deletion of a locus encoding ten Toll-dependent secreted peptides at the 55C locus known as Bomanins largely phenocopies the *Toll* mutant phenotype (Uttenweiler-Joseph *et al*, 1998; Clemmons *et al*, 2015). This suggests that these peptides are somehow involved in mediating the defenses resulting from Toll pathway activation, which regulates the expression of more than 150 immune-responsive genes (De Gregorio *et al*, 2002). A majority of Bomanins at the 55C locus are short (BomS), the secreted form of which essentially contains a single domain characteristic of this family of peptides. Other members include a tail after the Bomanin domain (BomT) whereas bicipital members are characterized by the inclusion of two domains separated by a linker domain (BomBc). Although a recent study suggests that some BomS are likely active against *Candida glabrata* and can function somewhat redundantly (Lindsay *et al*, 2018), the exact function of most Bomanins in host defense remains uncertain.

How exactly *Drosophila* host defense confronts *A. fumigatus* and fungal virulence factors in general remains unknown despite our knowledge of the generic role of the Toll pathway in antifungal defense. Here, we revisit *A. fumigatus* infections obtained by injecting a limited number of conidia into the thorax and find that the fungus is unable to invade flies, including Toll pathway *MyD88* mutants, due to melanization, a distinct host defense, which is mediated by the Hayan protease and the PPO2 phenol oxidase. Our data suggest that Toll pathway immuno-deficient flies succumb to *A. fumigatus* secreted toxins, some of which target the nervous system. We report here that Toll mediates resilience to particular mycotoxins through specific Bomanins that do not function as classical AMPs in this setting but neutralize the effects of these fungal virulence factors. Our data illustrate that evolution has selected a specialized defense partially mediated by secreted peptides that allow the host to elude or counteract the action or effects of the attack by mycotoxins.

Results

Defense against *A. fumigatus* depends on the Toll pathway independently of its role in controlling AMP expression

Homozygous or hemizygous *MyD88* but not wild-type flies were highly sensitive to an *A. fumigatus* challenge with various strains and succumbed to as few as five injected conidia (Figs 1A and EV1A–C). Mutations in the *Drosophila* Toll pathway genes *spätzle* (*spz*) and *Toll* (*Tl*) led to an *A. fumigatus* susceptibility phenotype similar to that of *MyD88* (Fig EV1D and E).

Unexpectedly and in contrast to other relevant microbial infections in Toll pathway mutants (Alarco *et al*, 2004; Apidianakis *et al*, 2004; Quintin *et al*, 2013; Duneau *et al*, 2017), the fungal burden did not reach values higher than 200–300 colony-forming units (cfus) in *MyD88* flies challenged with 50 conidia (Fig 1B) or even upon the injection of 5,000 conidia (Fig EV1F). The lack of growth of *A. fumigatus* in *MyD88* flies was confirmed by measuring the fungal load upon death (FLUD; Fig EV1G). Monitoring a GFP-expressing *A. fumigatus* strain revealed the formation of mycelia only next to the injection site of 50 conidia in both wild-type and *MyD88* flies (Figs 1C and D, and EV1H). Puzzlingly, the injection of a higher number of conidia led to the formation of fewer hyphae (Fig EV1I). To exclude the possibility that death might be caused by another microorganism, possibly deriving from the microbiota, we confirmed the sensitivity of *MyD88* mutant flies to *A. fumigatus* challenge on antibiotics-treated flies and on axenic flies (Fig EV1J–L). We conclude that *MyD88* flies succumb to a low *A. fumigatus* burden (lower than 200 *A. fumigatus* cfus at death; Fig EV1G).

A septic injury with the Gram-positive bacterium *Micrococcus luteus* induces the expression of *Drosomycin* and all 55C *Bomanin* genes, *BomS4*-excepted (Fig EV2A and B). In contrast, the injection of even high doses of live or killed conidia did not induce the expression of *Drosomycin* steady-state transcripts measured by conventional RT–qPCR (Fig 1E). Only a mild induction of *Drosomycin* and the small secreted peptide-encoding gene *BomS1* were detected using digital PCR (RT–dPCR) in wild-type flies challenged with 5,000 conidia (Fig EV2C and D), which was confirmed by mass spectrometry detection of the induction of some BomS peptides but not Drosomycin in collected hemolymph (Fig EV2E and F, Appendix Fig S1). *Aspergillus fumigatus* infection thus induces weakly at the transcriptional level the expression of classical Toll pathway activation read-outs such as *BomS1* or *Drosomycin*. Surprisingly, only the short Bomanins and one Daisho peptide (DIM4) were reliably detected in the hemolymph via MALDI-TOF mass spectrometry. Their levels in the hemolymph were rather independent of the size of the inoculum (Fig EV2G), in keeping with the relatively stable fungal load measured (Fig EV1F). The expression of these peptides in the hemolymph tended to actually decrease upon injection of an inoculum $> 1{,}000$ conidia, possibly correlating with the decreased formation of hyphae and likely higher levels of gliotoxin.

Thus, the SPZ/Toll/MyD88 cassette is required for host defense against *A. fumigatus* infections, even though this pathogen only mildly stimulates the Toll pathway. Strikingly, the major read-out of the Toll pathway *Drosomycin* steady-state transcript levels and its encoded peptide, are only weakly induced.

Drosophila melanization curbs *A. fumigatus* invasion

As melanization is a host defense of insects effective against fungal infections, we tested *Hayan* mutant flies defective for this arm of innate immunity (Nam *et al*, 2012). These mutant flies were sensitive to *A. fumigatus* infection but less susceptible than *MyD88* mutant flies (Fig 1F). In contrast to *MyD88*, the fungal burden was

increased in these mutants (Fig 1G and H). Further, the melanization response was dependent on Prophenoloxidase 2 (PPO2) but not PPO1 nor on the Sp7 protease (Fig EV3A–D). Thus, in contrast to a previous study that demonstrated a role for PPO1 and Sp7 in the host defense against low inocula of *Enterococcus faecalis* (Dudzic *et al*, 2019), it appears that a relevant melanization response downstream of Hayan can be mediated through PPO2, which like PPO1 is cleaved by Hayan (Nam *et al*, 2012). Interestingly, *A. fumigatus* disseminated throughout the body in *Hayan* mutants but was restricted to the thorax in *MyD88* flies (Fig EV3E). In keeping with these results, the fungus erupted in cadavers from all three tagmata, including the legs, of *Hayan* mutants (Fig 1I). In contrast, the fungus

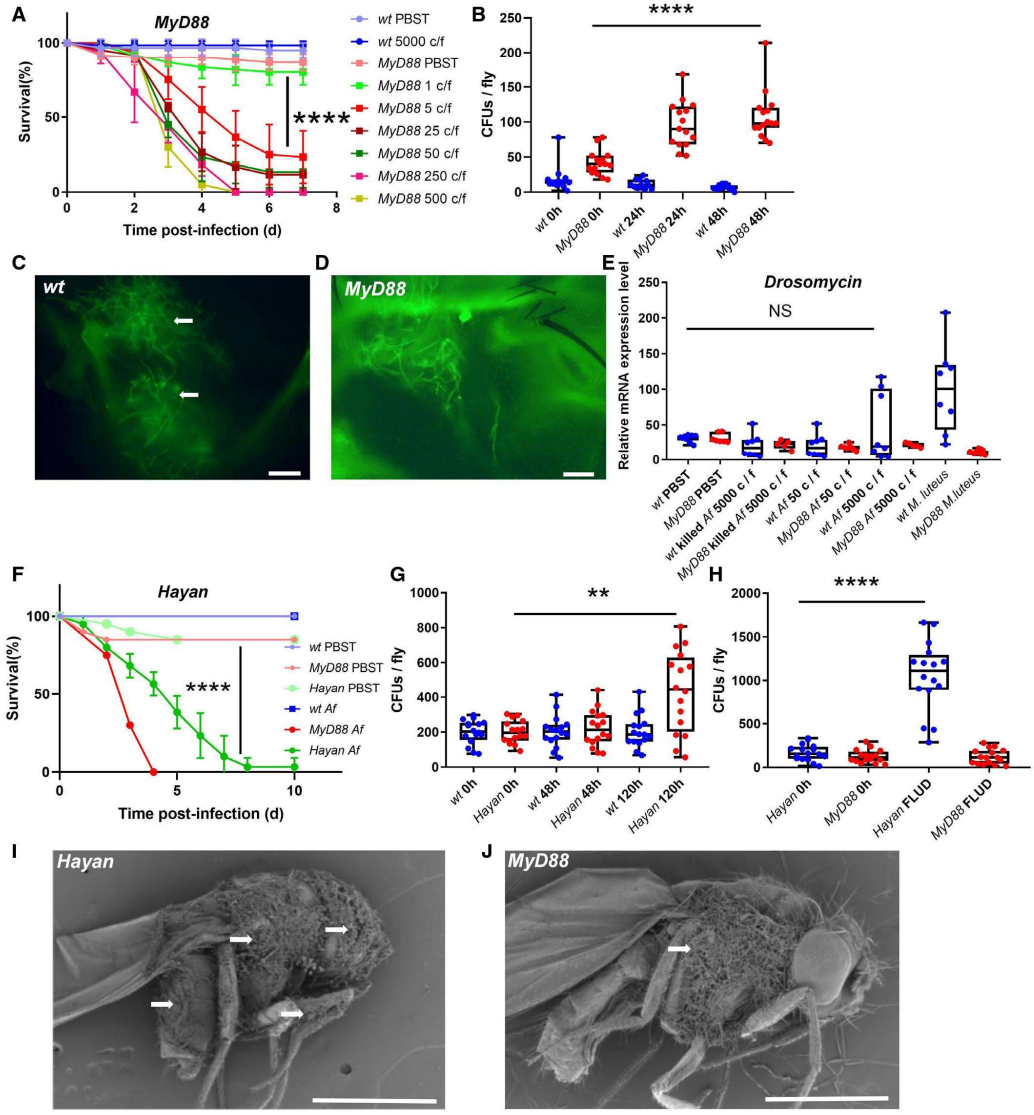


Figure 1.

Figure 1. Toll pathway mutants succumb to *Aspergillus fumigatus* infection even though it is not required to limit the proliferation and dissemination of the pathogen, an immune function mediated by melanization.

A Survival curves of *MyD88* flies injected with different doses of *A. fumigatus* conidia (c/f: conidia injected per fly; error bars represent mean ± SD of the survival of biological triplicates of 20 flies each).

B Fungal loads of single *MyD88* mutant and wild-type flies (50 conidia injected per fly).

C, D GFP-labeled *A. fumigatus* (50 conidia per fly) injected in wild-type (C) or *MyD88* mutant (D) flies form hyphae in the thorax of the flies (arrows). Scale bars 50 μm.

E Expression level of *Drosomycin* induced by different doses of injected *A. fumigatus* conidia measured by RT–qPCR; *M. luteus* (OD = 10) represents the positive control (pooled data of $n = 3$ experiments, biological replicates).

F–H *Hayan* flies are susceptible to *A. fumigatus*. Survival (F), time course of fungal loads of single *Hayan* mutant flies (500 conidia per fly) (G), and fungal load upon death (500 conidia per fly) (H) of *Hayan* mutant flies; *Hayan* 0 versus 120 h, (error bars represent mean ± SD of the survival of biological triplicates of 20 flies each).

I, J *A. fumigatus* hyphae extrusion (arrows) from *Hayan* (I) and *MyD88* (J) mutant cadavers; scanning electron micrographs. Scale bars 750 μm.

Data information: In (B, E, G, H), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima; data were analyzed using the Mann–Whitney statistical test. Survival curves were analyzed using the log-rank test. $^{**}P = 0.004$; $^{****}P < 0.0001$, and NS: not significant.

only broke through the cuticle in the thorax where it had been injected in *MyD88*, *Toll* or *spz* mutants (Figs 1J and EV3F). Of note, the fungus did not erupt from infected wild-type flies killed mechanically. We conclude that melanization limits the proliferation and the dissemination of *A. fumigatus* injected into wild-type flies yet does not eradicate it at the injection site, where a melanization plug forms.

We also tested the contribution of the cellular immune response either by challenging *eater* mutant flies lacking a major phagocytosis receptor, presaturating the phagocytic apparatus by injection of latex beads, or by genetically ablating hemocytes. In each case, no enhanced sensitivity to *A. fumigatus* infection was observed (Appendix Fig S2A–C).

A. fumigatus secondary metabolism is required for its virulence in *Drosophila*

The finding that *A. fumigatus* killed *MyD88* immuno-deficient flies with a low fungal burden and limited dissemination in conjunction with the observation of the modest induction of the Toll pathway suggested that Toll pathway mutants could be sensitive to some of the many diffusible mycotoxins known to be secreted by this fungus (Frisvad *et al*, 2009). We first tested this hypothesis using an *A. fumigatus* mutant strain lacking the phosphopantetheinyl transferase (*pptA*) gene required for the biosynthesis of all secondary metabolites, including most mycotoxins (Johns *et al*, 2017). This *A. fumigatus* mutant strain was not virulent when its conidia were injected into *MyD88* flies (Fig 2A); its fungal burden was somewhat reduced after 48 h (Fig 2B). Importantly, the Δ*pptA* mutant managed to form a limited mycelium at the injection site (Fig 2C and D). These findings indicated that one or several mycotoxins are responsible for the observed phenotypes. However, gliotoxin was not required to kill *MyD88* mutant flies as a gliotoxin deletion mutant, Δ*gliP*, was still as virulent as wild-type *A. fumigatus* (Fig 2E). As expected from the analysis of the gliotoxin mutant strain, the injection of commercially-available gliotoxin killed wild-type and *MyD88* flies at a similar rate, but only when injected at sufficiently high concentrations (Fig 2F). By contrast, other antimicrobial compounds secreted by *A. fumigatus* (Raffa & Keller, 2019), namely fumagillin and helvolic acid, did not kill wild-type or *MyD88* flies at the tested concentrations (Fig 2G and H).

The Toll pathway is required in the host defense against some *A. fumigatus* tremorgenic mycotoxins

The *ftm* gene cluster of *A. fumigatus* is involved in the biosynthesis of secondary metabolites belonging to the tremorgenic toxins such as the fumitremorgins and verruculogen. The *ftmA* gene encodes the first enzyme of this biosynthetic pathway (Kato *et al*, 2013). As shown in Fig 3A, a Δ*ftmA* mutant was slightly but reproducibly less virulent than the Δ*akuB* genetic background control strain, which is deficient for the nonhomologous end-joining DNA repair pathway. Whereas *MyD88* and wild-type flies behaved similarly after the injection of either low or high doses of verruculogen, *MyD88* flies were more sensitive than wild-type to this toxin injected at a 1 or 5 mg/ml concentration (or introduced as a powder thereby bypassing the need for dissolution in a DMSO-containing solvent), in conventional or microbe-free conditions (Figs 3B–D and EV4A). *Toll* and *spz* mutant flies also succumbed to injected verruculogen (Fig EV4B and C). *MyD88* flies were also sensitive to fumitremorgin C injected at concentrations greater than or equal to 1 mg/ml (Fig 3E).

Most wild-type and *MyD88* flies injected with verruculogen exhibited seizures as early as half an hour after injection and by 3 h all flies suffered from tremors (Fig 3F and Movies EV1–EV4). Interestingly, wild-type flies started recovering from seizures after verruculogen injection from 15 h onward; all surviving flies had recovered after about a day whereas *MyD88* flies never recovered (Fig EV4D). Of note, when challenging directly with verruculogen powder, *MyD88* flies did recover, but slower, likely because in this mode a lower effective dose of the mycotoxin is delivered (Fig EV4E). Upon closer inspection, we found that *MyD88*, but not wild-type flies, exhibited tremors after 2 days of *A. fumigatus* infection (Movie EV5). The Toll pathway is constitutively activated in $Toll^{10B}$ flies. As expected, $Toll^{10B}$ flies survived verruculogen injection like wild-type flies (Fig 3G). Remarkably, about 50% of these flies did not exhibit tremors at 3 h postinjection of verruculogen (Fig 3H). These data indicate that wild-type flies undergo the tremorgenic action of verruculogen and, in contrast to *MyD88* flies, are able to overcome the effects of the toxin in a resilience process that involves *spz*, *Toll*, and *MyD88*. Melanization and hemocytes did not appear to be involved in resilience to verruculogen action (Fig EV4F–H).

The Toll pathway is required in the host defense against a ribotoxin

We next tested the contribution of another mycotoxin, restrictocin, a ribotoxin protein secreted by *A. fumigatus* and other pathogenic fungi. Restrictocin cleaves 28S ribosomal RNA and thereby inhibits host cell translation (Fando *et al*, 1985; Lamy *et al*, 1991; Nayak *et al*, 2001). Injection of *A. fumigatus* Δ*aspf1* conidia, which lack the restrictocin biosynthesis locus, resulted in a modest but reproducible reduction in virulence as compared to the *ΔakuB* genetic background control strain when injected into *MyD88* mutants (Fig 4A). Strikingly, the injection of restrictocin killed *MyD88* but not wild-type flies in untreated, antibiotics-treated or axenic flies (Fig 4B–D). Whereas the injection of restrictocin led to the demise

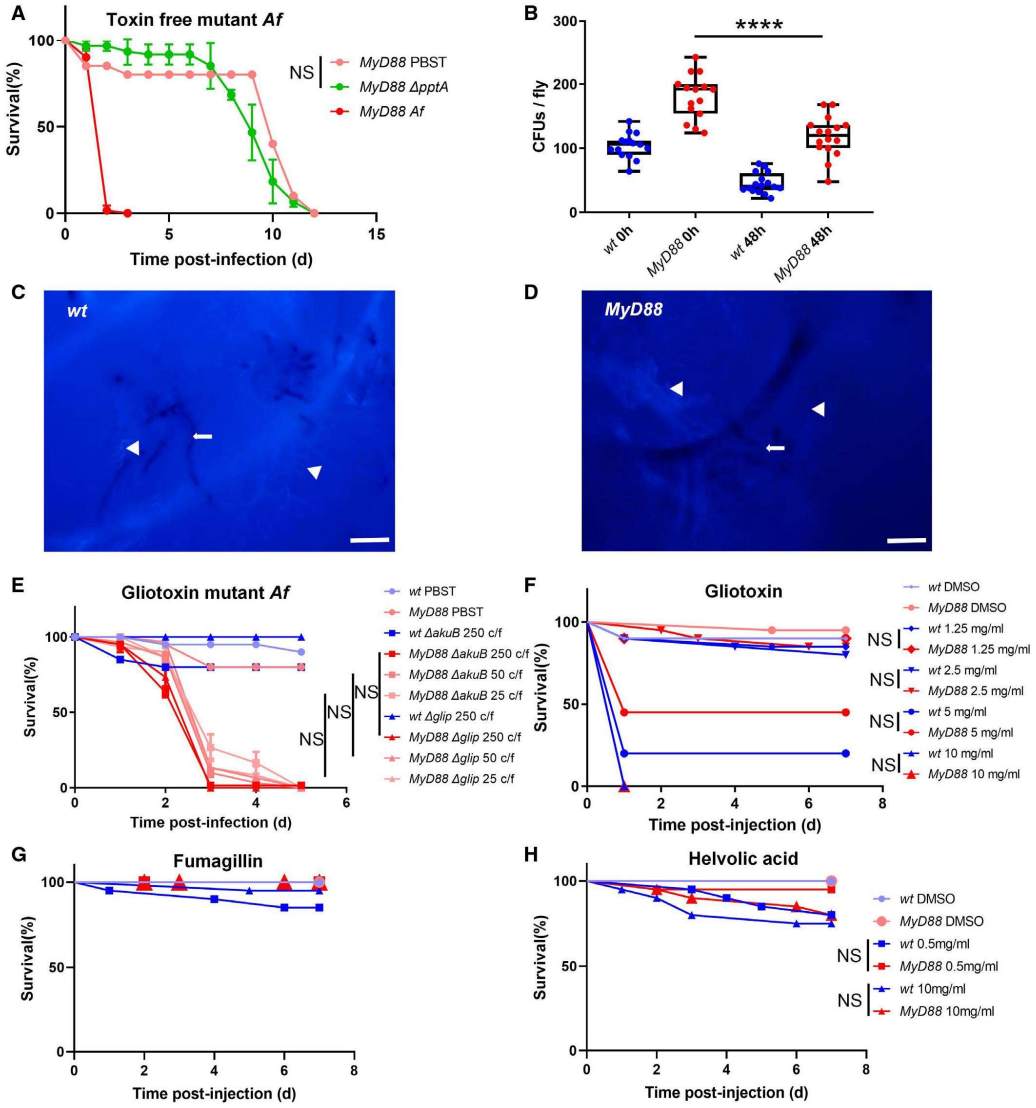


Figure 2.

Figure 2. Secondary metabolism is critical for the virulence of *Aspergillus fumigatus* in *Drosophila MyD88* immuno-deficient flies.

A Survival of *MyD88* flies injected with *ΔpptA* and wild-type (*Af*) *A. fumigatus* conidia (error bars represent mean ± SD of the survival of biological triplicates of 20 flies each).

B Fungal loads of single flies after the injection of 500 *ΔpptA* conidia.

C, D Hyphae of *ΔpptA A. fumigatus* observed in the thorax of wild-type and *MyD88* flies (arrow) after Uvitex-B negative staining, air sacs and tracheae are stained by Uvitex-B (arrowheads). Scale bars 50 μm.

E Dose response of *MyD88* flies after *ΔgliP* (gliotoxin) mutant or wild-type [*ΔakuB*] *A. fumigatus* infection; error bars represent mean ± SD of the survival of biological triplicates of 20 flies each; wild-type flies are used as a control for the dose of 250 conidia.

F–H Dose response of *MyD88* and wild-type flies after gliotoxin (F), fumagillin (G), and helvolic acid (H) injection at the indicated concentrations (20 flies per condition).

Data Information: In (B), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima; data were analyzed using the Mann–Whitney test. Survival curves were analyzed using the log-rank test. ****$P < 0.0001$, and NS: not significant.

of *spz* and *Toll* mutant flies, it did not impact flies deficient for either melanization or the cellular immune response (Fig EV4I–M).

The cleavage by restrictocin of the 28S RNA between G_{4325} and A_{4326} yields a fragment of about 500 nucleotides known as the α-sarcin fragment (Gluck *et al*, 1994). When we analyzed total RNA extracted from *MyD88* restrictocin-injected flies, we observed a fragment of the expected size, which was not detected in PBS-injected flies. The α-sarcin peak was also detected upon the injection of restrictocin in wild-type flies. The cleavage of the 28S RNA was, however, much less pronounced in wild-type flies as compared to *MyD88* (Fig 4E and F). These observations suggest that the MyD88-mediated response is able to counteract restrictocin *in vivo* prior to its action on rRNA. In agreement with these results, the GFP fluorescence emitted from a transgene-induced ubiquitously at the time of the challenge was reduced 42 h after restrictocin injection. In addition, GFP fluorescence was lower upon *A. fumigatus* challenge than upon a mock infection (Fig 4G). Taken together, these data suggest that restrictocin is able to inhibit translation to a detectable degree *in vivo*, likely through the cleavage of the ribosomal 28S α-sarcin/ricin loop as described *in vitro*.

We therefore checked in a rabbit reticulocyte translation assay that restrictocin is blocking translation *in vitro* (Fig 4H), as previously reported (Nayak *et al*, 2001). This observation was extended to *Drosophila* S2 cell extracts. Since the Toll pathway cannot be induced in regular S2 cells, we used a stable line that expresses a chimeric Toll receptor (ERTL) that can be activated by adding Epidermal Growth Factor (EGF) to the growth medium (Sun *et al*, 2004). In extracts from noninduced cells, eGFP *in vitro* translation was inhibited by the addition of restrictocin in a dose-dependent manner (Appendix Fig S3A and B). Even though the Toll pathway was indeed activated by the addition of EGF, translation with an extract made from induced ERTL-S2 cells was nevertheless inhibited by the addition of restrictocin almost as efficiently as with an extract made from noninduced ERTL-S2 cells (Fig 4I and J). Thus, the Toll pathway may not act at the intracellular level but possibly through secreted effectors as detailed further below.

The Spätzle/Toll/MyD88 cassette is thus required for host defense against both verruculogen, a secondary metabolite, and restrictocin, a protein ribotoxin.

Bomanins mediate resilience to mycotoxins

The Toll pathway regulates the expression of at least 150 genes, including some Bomanins initially identified as *Drosophila* immune-induced molecules (Uttenweiler-Joseph *et al*, 1998; De Gregorio *et al*, 2002). Strikingly, the deletion of the 55C locus (Fig EV2A) that spans 10 *Bomanin* genes yields a phenotype as strong as Toll pathway mutants in several infection models (Clemmons *et al*, 2015). In the case of *A. fumigatus*, we found the $Bom^{\Delta 55C}$ deletion mutant to be only somewhat less susceptible to this infection than *MyD88* flies (Fig 5A). The fungal burden remained low during and after the infection (Fig 5B, Appendix Fig S4A). Interestingly, the $Bom^{\Delta 55C}$ mutant was also sensitive to the injection of verruculogen and restrictocin (Fig 5C and D; green curve). Only 25% of $Bom^{\Delta 55C}$ flies versus more than 50% for isogenized wild-type survived verruculogen injection after day 1. In the case of restrictocin, $Bom^{\Delta 55C}$ flies succumbed to this challenge, which was not the case for control flies. To exclude the possibility of a nonspecific sensitivity of *MyD88* or of $Bom^{\Delta 55C}$ flies to stress, we submitted these mutant flies and their isogenized controls to a variety of stresses such as heat shock at 37°C or 29°C or the injection of salt solution or H_2O_2 (Appendix Fig S4B–E). The injection of 4.6 nl of 8% NaCl solution or of 2% H_2O_2 did not reveal a differential susceptibility of the immune-deficient flies to these challenges. In contrast, we did observe a mild susceptibility of *MyD88* but not of $Bom^{\Delta 55C}$ flies to a continuous exposure to 37°C. Similar results were obtained for an exposure to 29°C with *MyD88* flies displaying an enhanced sensitivity but only after 12–15 days, that is, much later than the usual time frame of our experiments (Figs 1A, 3B and 4B).

We attempted to identify the relevant 55C cluster genes involved in host defense against injected mycotoxins using a genetic rescue strategy in which we overexpressed single 55C locus *Bomanin* genes in the background of the $Bom^{\Delta 55C}$ deficiency. Overexpression of either *BomBc1*, *BomS3*, or *BomS6* provided a significant degree of protection (comparable to wild-type flies) against restrictocin whereas *BomS6* and, to a variable extent, *BomS1* protected $Bom^{\Delta 55C}$ mutant flies from verruculogen (Fig 5C and D). Of note, we still observed the induction of tremors in verruculogen-injected rescue flies.

To determine whether BomS3 interacts with restrictocin *in vitro*, we tested whether the preincubation of restrictocin with a BomS3 synthetic peptide would decrease the inhibition of translation in ERTL-S2 cells. As shown in Appendix Fig S3C and D, it was as inefficient as control synthetic BomS1 peptide in blocking the inhibition of translation mediated by restrictocin. Thus, BomS3 is unlikely to act directly and independently against restrictocin and might act extracellularly.

As monitored by RT–dPCR, the expression of *BomBc1*, *BomS1*, and especially *BomS3* was induced by an *A. fumigatus* challenge

(Figs 5E and F, and EV2D). Some other *Bom* genes located in 55C also exhibited a weak nonsignificant induction (Appendix Fig S5). As the injection of vehicle alone induced a significant response of some 55C locus genes, it was not possible to determine whether verruculogen is also able to induce their expression in this experimental series. We did find that the expressions of *BomBc1*, *BomS3*, and *BomS4* were induced to a low level by the injection of restrictocin (Fig 5E and F, Appendix Fig S5C), which was not the case for other *Bomanin* genes (Appendix Fig S5A, B and D–H).

BomS6 can protect flies from the toxic effects of verruculogen when expressed in the brain

That the ubiquitous overexpression of Tl^{10B} protects 50% of wild-type flies from the tremorgenic effects of verruculogen provided a convenient method to investigate which tissues mediate this effect. When we expressed the Tl^{10B} transgene in neurons, there was also a dominant protection of 50% of the flies from the tremors measured 3 h after the injection of verruculogen, an effect similar to its

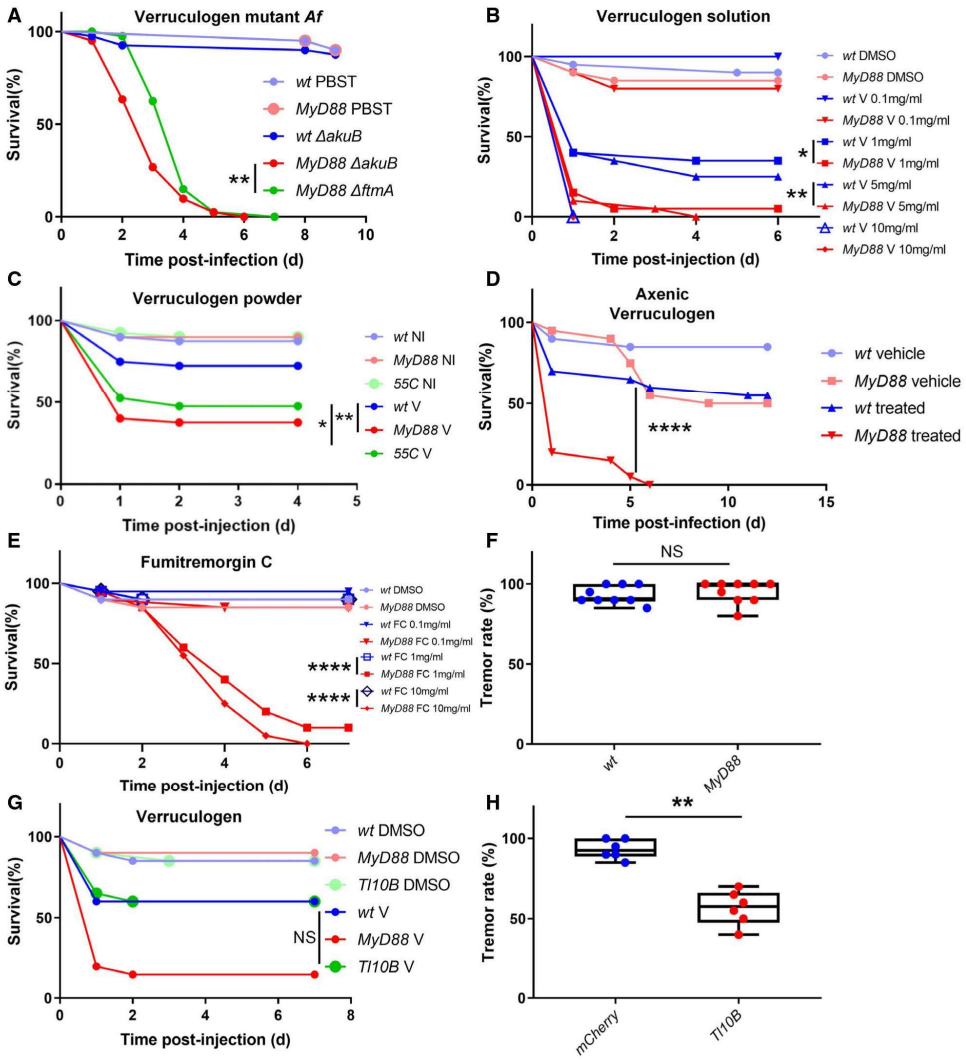


Figure 3.

Figure 3. The Toll pathway mediates *Drosophila* resilience to *Aspergillus fumigatus* tremorgenic secondary metabolites of the fumitremorgin/verruculogen biosynthesis pathway.

A, B Survival of *MyD88* or wild-type flies after injection of 250 conidia of *ΔftmA* (verruculogen and fumitremorgins biosynthesis pathway mutant) or wild-type [*ΔakuB*] *A. fumigatus* (A) or verruculogen (V) (B) (20 flies per condition); *ΔakuB* versus *ΔftmA*, $^{**}P = 0.002$ (A); wt vs. *MyD88* (1 mg/ml verruculogen, $^{*}P = 0.015$; 5 mg/ml, $^{**}P = 0.008$ (B)).

C, D Survival of *MyD88* mutant flies after verruculogen powder challenge (C), and axenic *MyD88* mutant flies after verruculogen solution injection (D); wt V versus 55C V, $^{*}P = 0.02$, versus *MyD88* V, $^{**}P = 0.002$ for verruculogen powder challenge (20 flies per condition).

E Survival of *MyD88* mutant flies after fumitremorgin C injection at different concentrations (20 flies per condition).

F Each dot corresponds to the tremor rate measured in a batch of 20 wild-type or *MyD88* flies 3 h after verruculogen injection (biological replicates).

G *Ubi-Gal4 > UAS-Toll^{10B}* flies survive like wild-type flies to verruculogen injection (20 flies per condition).

H Rate of *Ubi-Gal4 > UAS-Toll^{10B}* flies exhibiting tremors 3 h after injection of verruculogen in batches of 20 flies, each dot representing one batch; $^{**}P = 0.002$ (pooled data of $n = 3$ experiments, biological replicates).

Data Information: In (F, H), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima, and data were analyzed using the Mann–Whitney test. Survival curves were analyzed using the log-rank test. $^{****}P < 0.0001$, and NS: not significant. Except indicated otherwise (B), the concentration of injected verruculogen was 1 mg/ml.

ubiquitous expression (Figs 3H and 6A). We have shown above that 40% of flies in which Tl^{10B} was expressed ubiquitously succumbed to verruculogen injection, like wild-type flies (Fig 3G). In contrast, full protection was conferred to flies in which Tl^{10B} was expressed only in neurons (Fig 6B), which survived much better than wild-type flies. Similar observations were made when Tl^{10B} was expressed in glial cells, except that the degree of protection was weaker. This may reflect a side-effect of the overexpression strategy with a gene, the product of which is secreted: a cell type located in the vicinity of the physiologically relevant target cell type may partially achieve a biological activity (Fig 6C and D). We next tested the ectopic expression in a wild-type background of *BomS6*, which is the only peptide gene we found to reliably rescue the sensitivity phenotype of $Bom^{\Delta 55C}$ flies in survival experiments after a verruculogen challenge. When *BomS6* was ectopically expressed in the nervous system, all of the flies displayed tremors 3 h after injection and there was no enhanced protection of this phenotype (Fig EV5A and B). However, when we measured the time it took for those flies to recover from tremors, we did find that they recovered faster, at a pace similar to that obtained by its ubiquitous ectopic expression (Fig 6E and F). Interestingly, flies in which *BomS6* was expressed in neurons or ubiquitously were fully protected against the noxious effects of verruculogen in survival experiments (Fig 6G). When *BomS6* was expressed in glial cells, the improved recovery from verruculogen challenge was nearly significant (Fig 6H, $P = 0.058$) and flies did survive significantly better than wild-type flies but nevertheless were less protected than when *BomS6* was expressed in neurons (Fig 6I). When we tried to repeat the experiment by ectopically expressing *BomS4*, no protection was conferred to those flies suggesting a degree of specificity of the *BomS* genes (Fig EV5C and D).

To determine whether *Bomanins* can be induced by mycotoxins, we monitored their expression by dRT–PCR on head samples after verruculogen powder challenge or restrictocin injection. Strikingly, we found that only the expression of *BomS6* and *BomS4* was increased by the verruculogen powder challenge (Fig 6J and K, Appendix Fig S6A–H). Unexpectedly, these two genes were also the only ones to be induced in the head by the injection of restrictocin (Fig 6L and M). In contrast, all 55C *Bomanin* genes were induced in the head after the injection of 500 *A. fumigatus* conidia, except for *BomS3* and *BomS4* (Fig 6L and M, Appendix Fig S7A–H). Of note, *Drosomycin* expression was induced in the head after an *A. fumigatus* challenge but not by restrictocin or verruculogen (Appendix Figs S6I and S7I). Thus, only two *Bomanin* genes are induced in the head in response to restrictocin or verruculogen injection.

Discussion

Here, we observed that *A. fumigatus* remains confined to its injection site in both wild-type and Toll pathway mutant flies due to the restriction of fungal dissemination by melanization, not annihilation. Thus, this rare occurrence of a localized infection together with the analysis of mycotoxin mutants of *A. fumigatus* confirms the fundamental role of mycotoxins in the virulence of *A. fumigatus* and reveals an unanticipated role for the Toll pathway in the protection against various secreted poisonous molecules. In the course of evolution, host defense effectors able to effectively neutralize the action or effects of mycotoxins have been selected independently of classical xenobiotic detoxification pathways that protect the host through modification and elimination of the compounds.

Toll pathway mutants succumb directly to *A. fumigatus* or mycotoxin challenge

The microbiota plays an important role in various aspects of the biology of *Drosophila* (Lesperance & Broderick, 2020). Besides, bacteria, viruses may also participate in killing Toll pathway mutants as has been shown to be the case for ingested *Drosophila* C virus (Ferreira *et al*, 2014). As we did observe a significant lethality in some of our control experiments with the injection of PBS, especially on *MyD88* mutants, it was important to exclude the possibility of the microbiota playing a role in the observed susceptibility phenotypes by using either antibiotics treatment or axenic flies, which still succumbed to our experimental challenges; these experiments showed that the death of Toll pathway mutant flies is caused by the treatment and not an auxiliary infection. Of note, the antibiotics treatment was as effective as using axenic flies in suppressing the lethality observed sometimes in PBS-injected flies, which suggests that some bacteria escape from the digestive tract upon injury of the exoskeleton.

Induction of the expression of specific *Bomanin* genes upon mycotoxin challenge

The induction of *Drosomycin* transcripts by injected *A. fumigatus* conidia appears at best to be very mild as compared to that

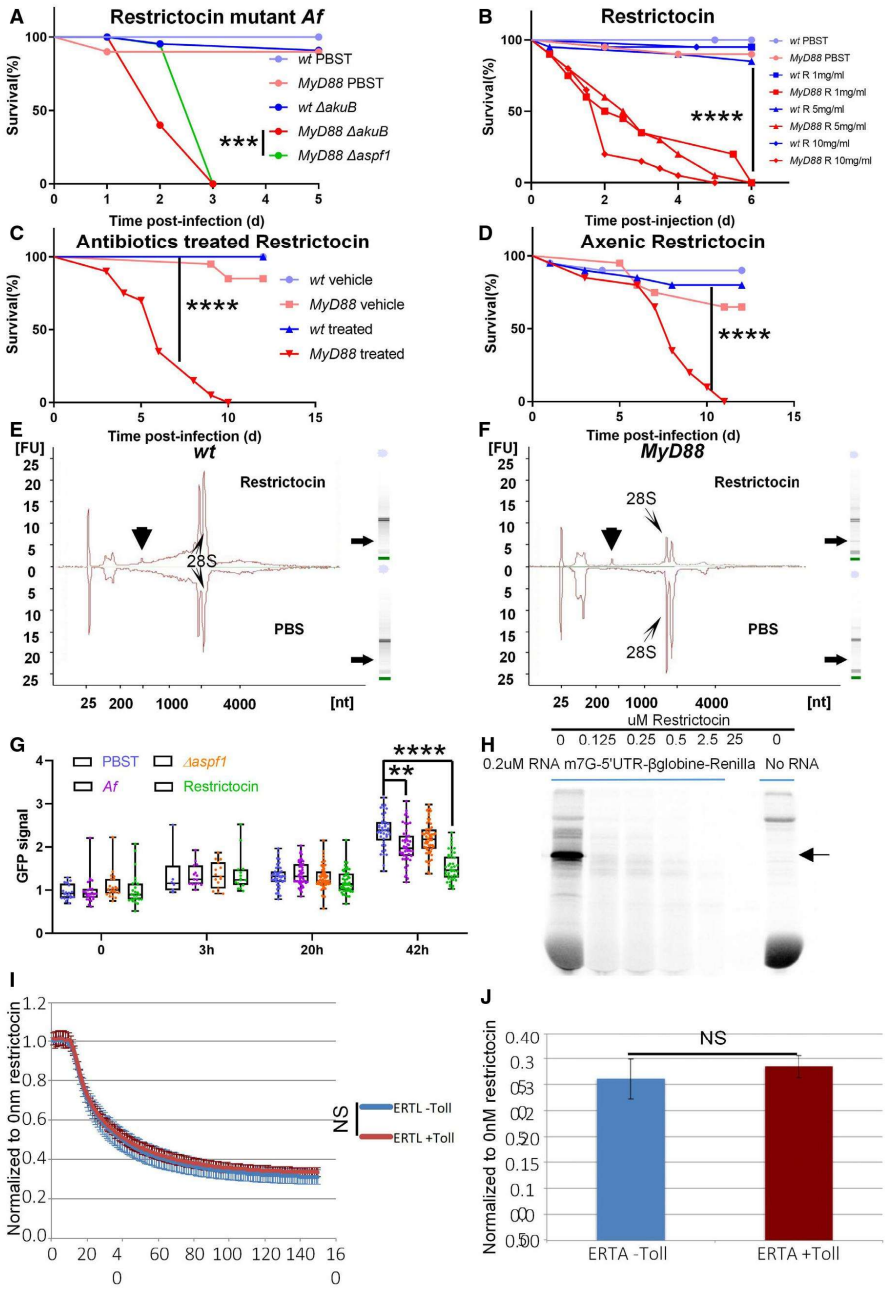


Figure 4.

Figure 4. Restrictocin functions as a ribotoxin *in vitro* and *in vivo* and affects *MyD88* mutant and not wild-type flies.

A Survival of *MyD88* or wild type flies to 250 injected *Δaspf1* (restrictocin mutant) or wild type [*ΔakuB*] *A. fumigatus* conidia (20 flies per condition); *MyD88: Δaspf1* versus *ΔakuB* (***$P = 0.0007$).

B Survival of *MyD88* flies after the injection of different concentrations of restrictocin (R) (20 flies per condition).

C, D Survival of antibiotics-treated (C) and axenic (D) *MyD88* mutant flies after restrictocin injection (20 flies per condition).

E, F Ribosomal RNA cleavage measurement after restrictocin or PBS injection in wild-type (E) and *MyD88* (F) flies; the arrowheads show the position of the 28S RNA-derived α-sarcin fragments whereas arrows on the right show its electrophoretic band position.

G Fluorescence (arbitrary units) emitted by transgenic p*Ubi-Gal4-Gal80ts* > *UAS-GFP* whole flies induced at the same time as the challenge and measured at the indicated time points; PBST versus *Af*: **$P = 0.002$ (pooled data of $n = 3$ experiments, biological replicates).

H SDS–PAGE analysis of ^{35}S-labeled translated proteins produced in a rabbit reticulocytes lysate from a m^{7}G-capped reporter RNA containing the 5′UTR of β-globin followed by the Renilla luciferase coding sequence (arrow), in the presence of increasing concentrations (0.125–25 nM) of restrictocin.

I, J Fluorescence analysis from *in vitro* translated eGFP from an IGR (CrPV)-driven reporter in noninduced (blue) and Toll-induced (red) ERTL lysates: translation kinetics of *in vitro* synthesized eGFP in the presence of 1 nM restrictocin showing fluorescence values normalized to untreated translation reactions (I) and a histogram representing the end-point fluorescence quantification of I (J).

Data Information: In (G), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima; data were analyzed using the Kruskal–Wallis test and Dunn's *post hoc* test. In (I), error bars represent mean ± SEM ($n = 3$, technical replicates); in (J), error bars represent mean ± SEM ($n = 3$, technical replicates). Survival curves were analyzed using the log-rank test. ****$P < 0.0001$, and NS: not significant.

induced by the monomorphic yeasts *C. glabrata* or *Saccharomyces cerevisiae* that were easily detected by regular RT–qPCR (Quintin *et al*, 2013). This may be linked to the masking of ß-(1-3) glucans by hydrophobin proteins or melanin on the conidial cell wall (van de Veerdonk *et al*, 2017; Blango *et al*, 2019). The secretion of inhibitors of NF-κB signaling such as gliotoxin may also be at work (Pahl *et al*, 1996). It is, however, perplexing that secreted short Bomanins but not Drosomycin were detected by mass spectrometry in the hemolymph even though this peptide is massively produced during the systemic immune response to injected bacteria, at an estimated concentration of 0.3 µM (Uttenweiler-Joseph *et al*, 1998).

We find that the injection of restrictocin in flies leads to a modest yet significant induction of only a subset of *Bomanins*, *BomBc1*, *BomS3*, and *BomS4* (Fig 5E and F, Appendix Fig S5C). In contrast, all of them but *BomS4*—which has the lowest basal expression—are induced by a systemic immune challenge by *M. luteus* (Fig EV2B). Only *BomS4* and *BomS6* were induced in the head after the injection of verruculogen (Fig 6J and K, Appendix Fig S6A–H). This differential expression of *Bomanin* genes upon mycotoxin injection, especially that of *BomS4* that is not induced in the systemic immune response, suggests that the observed induction is not due to peptidoglycan contaminating the mycotoxin preparations as it would have induced almost all *Bomanins*. Of note, we have likely employed higher concentrations of mycotoxins than actually released during infection. Indeed, whereas *BomS4* is induced in heads by verruculogen powder challenge, it is not induced there after *A. fumigatus* infection. Taken together, these data then suggest that a process akin to the immune surveillance of core cellular processes first described in *Caenorhabditis elegans* may also exist in *Drosophila*. For instance, toxins that affect the translation machinery lead to the induction of varied host defenses, a situation similar to that encountered with restrictocin that indirectly inhibits translation by targeting the ribosomal 28S RNA (Dunbar *et al*, 2012; McEwan *et al*, 2012; Melo & Ruvkun, 2012). We conclude that restrictocin and verruculogen induce a response limited to two *BomS* genes in the head, which is distinct from that induced in the framework of the Toll-dependent systemic immune response.

Functions and specificity of Bomanins encoded at the 55C locus

The current paradigm for insect immune-induced secreted peptides is that they primarily represent AMPs (Hanson & Lemaitre, 2020; Lazzaro *et al*, 2020; Lin *et al*, 2020). This has been checked experimentally by the deletion of multiple AMP genes loci; the deletion of AMP genes regulated mostly by the IMD pathway phenocopied the susceptibility to Gram-negative bacteria of IMD pathway mutants (Hanson *et al*, 2019). This was, however, less clear as regards the deletion of AMP genes regulated by the Toll pathway. It appears that 55C locus *Bomanin* genes play a predominant role in the host defense against Gram-positive bacteria, yeasts, and fungi (Clemmons *et al*, 2015; Hanson *et al*, 2019). It is clear that some of these genes are required in the resistance against *E. faecalis*, suggesting that some Bomanins may function as AMPs (Clemmons *et al*, 2015). $Bom^{\Delta 55C}$ deficiency flies are susceptible to *C. glabrata* infection and this susceptibility was rescued by overexpressing *BomS* genes such as *BomS3* (Lindsay *et al*, 2018). However, no *C. glabrata* killing activity of synthetic Bom peptides could be found in *in vitro* assays (Lindsay *et al*, 2018), even though hemolymph collected from wild-type but not mutant flies was fungicidal. The lack of fungicidal activity in the hemolymph of $Bom^{\Delta 55C}$ flies was partially restored in the hemolymph of $Bom^{\Delta 55C}$ mutant flies overexpressing *BomS5*, suggesting that at least this peptide may have some candicidal activity when combined with other Toll-dependent gene product(s) (Lindsay *et al*, 2018). While that study suggested that BomS peptides are interchangeable against *C. glabrata* provided they are expressed at sufficiently high levels, we report here that the BomS peptides appear to be much more specific with respect to the activity against mycotoxin action. Indeed, in the setting of the $Bom^{\Delta 55C}$ deficiency, only overexpressed BomS6 or BomS1 show some activity against verruculogen, whereas the forced expression of BomS6, BomS3, or BomBc1 appears to be able to counteract restrictocin. An antimicrobial role has been proposed for BomS3 against *C. glabrata* by Lindsay *et al* (2018). It cannot be, however, formally excluded that BomS3 might also act against an unidentified *C. glabrata* secreted toxin. In this respect, it has recently been reported that *C. glabrata* is able to invade the brain (Benmimoun *et al*, 2020) where we suspect that the activation of the Toll pathway signaling is also taking place.

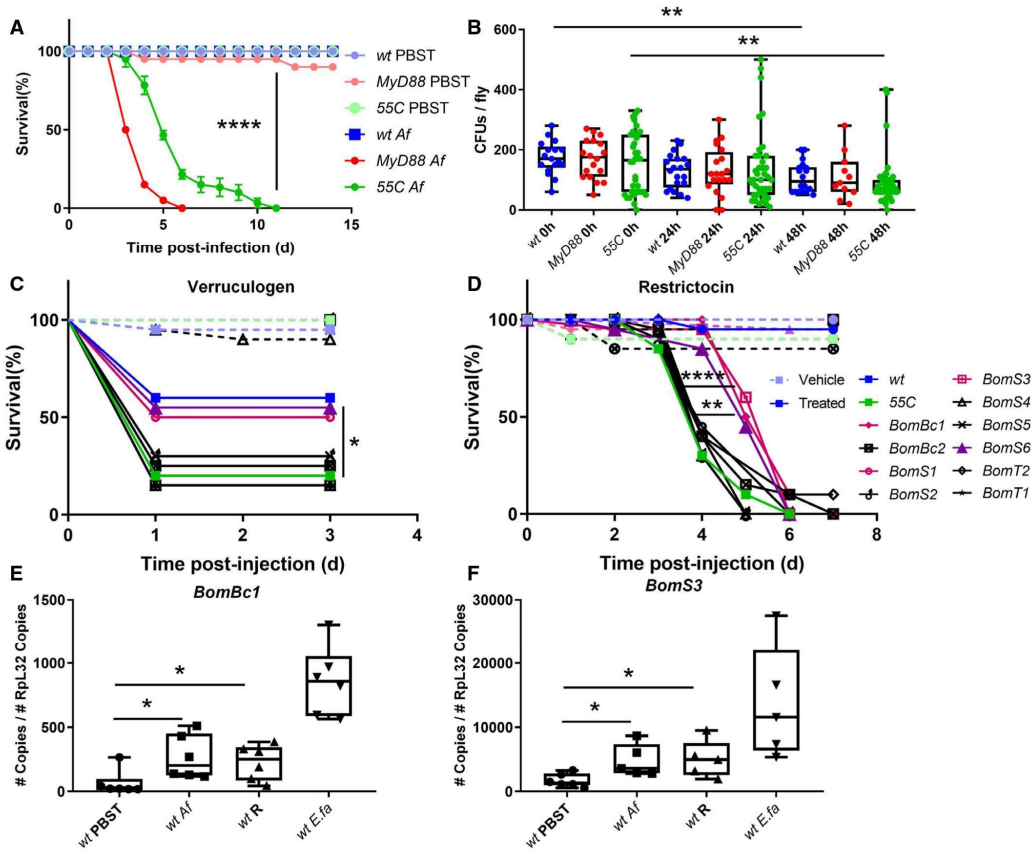


Figure 5. Distinct *Bomanins* mediate resilience to specific *Aspergillus fumigatus* mycotoxins.

A, B Survival (A) and fungal load (B) of $Bom^{\Delta 55C}$ (55C) deficient flies compared with wild-type and *MyD88 flies* after injection of 250 conidia (error bars represent mean ± SD of the survival of biological triplicates of 20 flies each); ****$P < 0.0001$. (B) The fungal burden does not increase in $Bom^{\Delta 55C}$-deficient flies; wt 0 versus 48 h, **$P = 0.001$; 55C 0 versus 48 h, **$P = 0.007$; pooled data of $n = 3$ experiments, biological replicates.

C, D Rescue of the sensitivity of $Bom^{\Delta 55C}$ flies to verruculogen (C) or to restrictocin (D) by the transgenic expression of individual 55C locus genes (caption in D also applies to (C)). 55C flies versus *BomS1*, *$P = 0.0495$, versus *BomS6* *$P = 0.011$ for verruculogen assay (C); 55C flies versus *BomBc1* or *BomS3*, ****$P < 0.0001$, 55C flies versus *BomS6*, **$P = 0.0028$ for restrictocin assay (D) (20 flies per condition).

E, F Expression levels of *BomBc1*, and *BomS3* measured by RT–digital PCR 48 h after challenge; *BomBc1* PBST versus *Af*, *$P = 0.015$, PBST versus restrictocin (R), *$P = 0.015$; *BomS3*: PBST versus *Af*, *$P = 0.02$, PBST versus R, *$P = 0.03$ (pooled data of $n = 3$ experiments, biological replicates).

Data Information: In (B, E, F), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima; data were analyzed using the Mann–Whitney statistical test. Survival curves were analyzed using the log-rank test.

With respect to host defense against restrictocin, our partial rescue data of the $Bom^{\Delta 55C}$ susceptibility phenotype by three 55C *Bomanins*, including one encoding a bicipital Bomanin might be accounted for by some form of redundancy. We cannot, however, exclude that these Bomanins provide a degree of protection through separate mechanisms, especially since the two Bomanin domains of Bc1 are rather divergent when compared to the high degree of conservation exhibited by BomS domains (Clemmons *et al*, 2015). Yet, *BomS6* is the sole 55C *Bomanin* providing protection against both restrictocin and verruculogen. BomS6 contains a lysine residue at position 10 of its Bomanin domain, like BomS2 but unlike BomS4 that contains a valine whereas other BomS peptides have an arginine at this position. The other difference is an isoleucine instead of valine at position 14 of the Bomanin domain. Thus, these

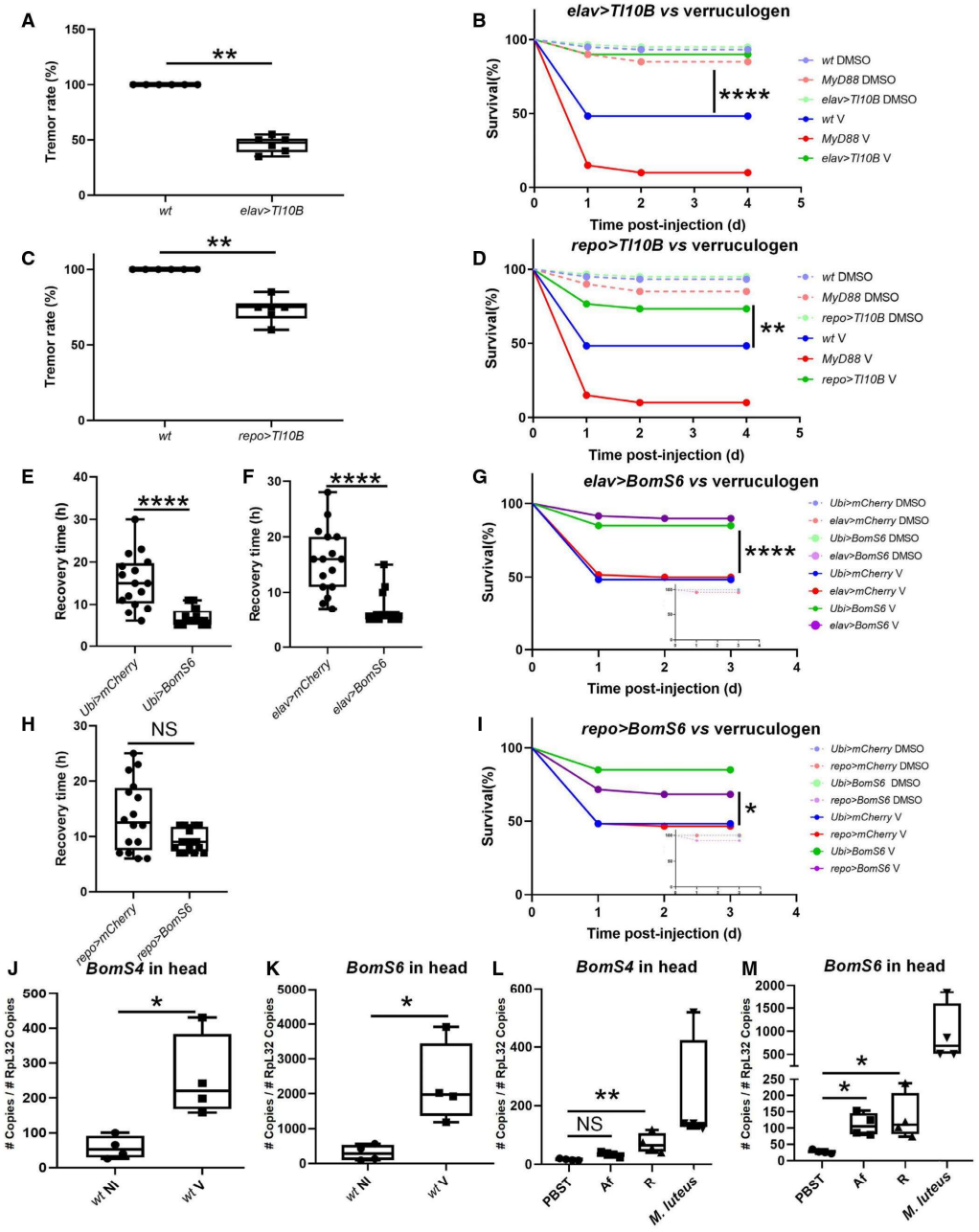


Figure 6

Figure 6. *Bomanin S6* mediates resilience to verruculogen in the nervous system of *Drosophila*.

A, B Tremor rate (A) and survival (B) of flies (20 flies per condition) overexpressing *Tl[10B]* in neurons compared with wild-type after injection of verruculogen. (A) Each dot corresponds to the tremor rate measured in a batch of 20 flies; tremor rate *wt* versus *elav* > *UAS-Toll^{10B}*, $^{**}P = 0.002$.

C, D Tremor rate (C) and survival (D) of flies (20 flies per condition) overexpressing *Tl[10B]* in glia compared with wild-type after injection of verruculogen. (C) Each dot corresponds to the tremor rate measured in a batch of 20 flies; tremor rate *wt* versus *repo* > *UAS-Toll^{10B}*, $^{**}P = 0.002$. (D) Survival *wt* V versus *repo* > *UAS-Toll^{10B}* V, $^{**}P = 0.005$.

E–G Recovery time from tremor (E, F) and survival (G) of single flies overexpressing *BomS6* ubiquitously (E, G) or in neurons (F, G) (biological replicates) compared with wild-type after injection of verruculogen; in (G) the inset represents the survival of vehicle control groups.

H, I Recovery time from tremor (H) and survival (I) of single flies overexpressing *BomS6* in glia (pooled data from $n = 3$ experiments, biological replicates) compared with wild-type after injection of verruculogen; in (I) the inset represents the survival of vehicle control groups. Recovery time (H) *repo* > *mCherry* versus *repo* > *BomS6*, $P = 0.058$, survival (I) *repo* > *mCherry* V versus *repo* > *BomS6* V, $^{*}P = 0.016$.

J–M Expression of *BomS4* (J) and *BomS6* (K) in the head after verruculogen powder challenge, and *BomS4* (L) and *BomS6* (M) in the head after *A. fumigatus* (*Af*), restrictocin or *M. luteus* injection (pooled data from $n = 3$ experiments, biological replicates).

Data Information: In (A, C, E, F, H, J–M), the middle bar of box plots represents the median and the upper and lower limits of boxes indicate, respectively, the first and third quartiles; the whiskers define the minima and maxima; data were analyzed using the Mann–Whitney statistical test. Survival curves were analyzed using the log-rank test. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$, and NS: not significant. In this figure, the concentration of injected verruculogen or restrictocin was 1 mg/ml.

biochemical differences along the capacity to be induced in the heads by *A. fumigatus* and verruculogen account for the unique function of BomS6 among Bomanins.

Whereas we propose here a specific function for some Bomanins in counteracting the effects of restrictocin or verruculogen, we cannot formally exclude an AMP function in other contexts. Indeed, it has previously been reported that mammalian alpha-defensin AMPs are also able to directly neutralize secreted bacterial virulence factors such as pore-forming toxins or enzymes that need to penetrate inside eukaryotic cells to act on their intracellular target (Kudryashova *et al*, 2014 and references therein). These proteins are inherently thermodynamically unstable as they need to change their conformations to insert or go through the mammalian cell plasma membrane. The amphipathic properties of alpha-defensins allow them to destructure these secreted virulence factors through hydrophobic interactions and thereby inactivate them (Kudryashova *et al*, 2014). Such a mechanism might be at play as regards a potential interaction of Bomanins with restrictocin, which does cross the plasma membrane. Indeed, the N-terminal part of mature BomS6 appears to be rather hydrophobic (38% of residues are hydrophobic) and uncharged. As regards verruculogen, it acts through hydrophobic interactions with one of its molecular targets (see further below), and possibly, BomS6 might also directly interact with verruculogen through hydrophobic interactions, although other BomS peptides (e.g., BomS3) exhibit similar or higher hydrophobicity.

The mechanism of action of restrictocin in inhibiting translation is well established and it crosses the plasma membrane of insects easily. Thus, it may act ubiquitously on all cell types and organs of the host. Wild-type flies tolerate exposure to a relatively large range of restrictocin concentrations whereas *MyD88* flies succumb faster to a high dose of 10 mg/ml (Fig 4B); the Toll-dependent response to *A. fumigatus* is not dose-dependent (Fig EV2G). Toll in wild-type flies blocks to a large extent the action of restrictocin since it prevents the cleavage of 28S rRNA (Fig 4E and F). BomS3, however, does not appear to directly bind to restrictocin (Appendix Fig S3C and D). Taken together, these observations suggest an indirect mode of action of Bomanins, at least for BomS3 and BomS6.

Our understanding of the action(s) of verruculogen on the nervous system is less clear as multiple effects are reported in the literature. These include increased spontaneous release of glutamate and aspartate from cerebrocortical synaptosomes (Hotujac *et al*, 1976; Norris *et al*, 1980), inhibition of the $GABA_A$ receptor (Gant *et al*, 1987) or inhibition of calcium-activated K^+ channels (Knaus *et al*, 1994) such as *Drosophila* Slowpoke, for which a detailed structural understanding of its interaction with verruculogen is available (Raisch *et al*, 2021). In contrast to the effect of restrictocin, verruculogen induces tremors also in wild-type flies, but unlike *MyD88* flies, these are able to reverse this effect, a situation also observed in cattle (Norris *et al*, 1980; Gant *et al*, 1987). The constitutive activation of the Toll pathway neutralizes to a significant extent the tremorgenic effects of injected verruculogen early on. Interestingly, BomS6 appears to function somewhat differently when overexpressed in the nervous system of wild-type flies. It does not prevent the initial tremors induced by verruculogen but allows the host to recover more rapidly, which suggests that two distinct processes are at play. Thus, Tl^{10B} may function through an effector that is distinct from BomS6 in the early protection against tremors; alternatively, this other effector may act in concert with BomS6. Future studies should determine whether the actions of Bomanins are direct or indirect, the latter being more likely given the different targets of restrictocin, verruculogen, and fumitremorgins. One may wonder whether Bomanins might alter the permeability of the plasma membrane for instance. The recent finding of a role for another Toll pathway effector, BaramicinA, in glial cells against a neurotoxin opened the possibility of an indirect role of BaraA in regulating the permeability of the Blood–Brain-Barrier (preprint: Huang *et al*, 2022).

Perspectives

Our work presented here and in a concurring study suggests that host defense has evolved to select mechanisms not only to directly fight off invading microorganisms such as AMPs but also to protect the host against the toxins they secrete (preprint: Huang *et al*, 2022). The identification of specific effectors of *Drosophila* innate immune signaling pathways evolved to counteract toxins is an important addition to our emerging understanding of host strategies implemented to cope with such microbial weapons, e.g., pore-forming toxins, fungal toxins in the gut, alpha-defensins (Kudryashova *et al*, 2014; Greaney *et al*, 2015; Lee *et al*, 2016; Chikina *et al*, 2020). It will be interesting to determine whether innate immunity also protects at least to some extent against mycotoxins that contaminate the food that present a major health threat for animals and humans (Brown *et al*, 2021).

Aspergillosis causes acute or chronic infections in an estimated 14 million patients (Kosmidis & Denning, 2015; Gago *et al*, 2019). Chronic infections represent major threats to the survival of patients with comorbidities. It will therefore be important to establish whether mammalian antifungal innate immune response pathways also contribute to resilience against mycotoxins as is the case for the Toll pathway in flies. Finally, our findings open the possibility of the existence of host defenses that protect immunocompetent animals or humans against some mycotoxins but leave individuals deficient for these defenses susceptible to disease.

Materials and Methods

Microbial strains

Aspergillus fumigatus was cultured on potato dextrose agar (PDA) medium supplemented with 0.1 g/l chloramphenicol in an incubator at 29°C. Conidia were harvested after 4–7 days of culture. The conidial suspension was purified by filtration on cheese cloth to eliminate hyphae and other impurities. The standard wild-type *A. fumigatus CEA17ΔakuB*^{Ku80} (CEA17) is a kind gift from Drs. Anne Beauvais and Jean-Paul Latge (Institut Pasteur, Paris) and is also the genetic background control for *ΔgliP*. Other wild-type strains include D141 (background for D141-GFP), Af293, ATCC46645, A1160 (background for *ΔpptA*), GFP-labeled strain (D141-GFP). *ΔpptA* (secondary metabolites free mutant) and *ΔgliP* (gliotoxin-free mutant) have been previously described (Hillmann *et al*, 2015; Johns *et al*, 2017).

For targeted deletion of *ftmA* (AFUB_086360) and *aspf1* (AFUB_050860) gene replacement cassettes were generated by three-fragment-based PCR as described previously (Szewczyk *et al*, 2006). In brief, deletion constructs were generated by amplifying around 1 kb up- and downstream sequences of the respective gene and insertion of the pyrithiamine resistance cassette (Kubodera *et al*, 2000) by fusion PCR. Protoplasts of *CEA17ΔakuB*^{Ku80} *A. fumigatus* strain (da Silva Ferreira *et al*, 2006) were transformed with purified PCR products. Transformants were selected for resistance to pyrithiamine. Homologous recombination and integration of the deletion cassette were validated by PCR. Phusion Flash High-Fidelity Master Mix (Thermo Scientific, Germany) was used for all reactions. *A. fumigatus* was cultivated in *Aspergillus* minimal medium (Jahn *et al*, 1997). Media were supplemented with 0.1 mg/l pyrithiamine (Merck, Germany) when required.

The sequence of primers is found in Appendix Table S1.

Micrococcus luteus CGMCC#1.2299 was cultured in Tryptic soy broth (TSB), and *E. faecalis* CGMCC#1.2135 was cultured in Luria-Bertani (LB) at 37°C for 24 h. The bacteria were then washed in PBS thrice and resuspended.

Fly strains

Fly lines were raised on food at 25°C with 65% humidity. For 25 l of fly food medium, 1.2 kg cornmeal (Priméal), 1.2 kg glucose (Tereos Syral), 1.5 kg yeast (Bio Springer), 90 g nipagin (VWR Chemicals) were diluted into 350 ml ethanol (Sigma-Aldrich), 120 g agar-agar (Sobigel) and water qsp were used.

w^{A5001} flies were used as wild-type control unless otherwise indicated. Canton-S (BDSC64349), w^{1118} (VDRC60000), and y^1w^1 were used as further wild-type controls as needed. The following mutant lines were used: *MyD88*^{c03881}, *Df (2R)3591*, *Hayan*^{34}, *w*, P{*ry*, *Dipt-LacZ*}, P{w^+, *drs-GFP*}; *spz*^{rm7}, *spz*^{ti5}, *Toll*^{632}. The following strains *Df(3R)Tl-I*, *e*^{1}*/TM3*, *Ser*^{1} (BDSC1911), *eater*^{1}, *SP7*, *PPO1*$^{\Delta}$, and *PPO2*$^{\Delta}$ were kind gifts from Dr. Bruno Lemaitre; *eater*$^{\Delta}$ mutants were obtained by crossing the two deficiency lines *Df(3R)Tl-I*, *e*^{1}*/TM3*, *Ser*^{1} (BDSC1911) and *Df(3R)D605/TM3*, *Sb*^{1} *Ser*^{1} (BDSC823). *spz* and *Tl* mutants used in this study were either transheterozygous or hemizygous mutants crossed at 25°C. *phmlΔ*-Gal4 > UAS-eGFP is a reporter line for hemocytes, *w*, P{UAS-*rpr*.C}, P{UAS-*hid*} flies (a kind gift of Akira Goto) were crossed to the *phmlΔ-Gal4* > *UAS-eGFP* line at 29°C to ablate hemocytes during development. *UAS-Toll*^{10B} flies were crossed to a *w*; *pUbi-Gal4*, *pTub-Gal80ts* (BDSC30140) or to *elav-Gal4* or *repo-Gal4* driver lines at 25°C; hatched adults were placed at 29°C for 5 days to activate the Toll pathway.

The *Bom*$^{\Delta 55C}$ deficiency was a kind gift of Steven Wasserman that was further isogenized in the w^{A5001} background. The transgenic lines expressing single *Bom* genes of the 55C locus under the *pUAS-hsp70* promoter control were generated as described (list of primers in Appendix Table S1) and checked by sequencing. The transgenic flies were crossed to a *w*; *pUbi-Gal4*, *pTub-Gal80*^{ts} driver line, in a homozygous *Bom*$^{\Delta 55C}$ mutant or w^{A5001} background. The expression of the transgenes was checked by RT–qPCR and mass spectrometry analysis on collected hemolymph of single flies as required.

Preparation of toxin or chemical stress solutions

Restrictocin (Sigma) was resuspended in phosphate buffer saline (PBS) pH = 7.2, gliotoxin (Abcam), helvolic acid (Abcam), fumagillin (Abcam), verruculogen (Abcam), fumitremorgin C (Sigma), were dissolved at 10 mg/ml in Dimethyl sulfoxide (DMSO; Molecular biology grade, Sigma) as stock solutions and stored at −20°C. A working concentration of 1 mg/ml in DMSO was used for injections of 4.6 nl of all toxin solutions unless otherwise indicated. Toxin solutions were thawed on ice for 1 h prior to use. As multiple freeze/thaw cycles reduce the potency of the toxins, care was taken not to use an aliquot more than five times and aliquots were not stored for more than 1 month. NaCl (Sigma) and 3% H_2O_2 solutions in PBS pH 7.2 were prepared freshly for each injection.

Axenic flies

To obtain axenic flies, eggs were collected, washed with water, and then 70% ethanol prior to dechorionation of eggs in a solution of 50% bleach until the chorion disappeared. Eggs were transferred into sterile vials containing media and a mix of antibiotics: ampicillin, chloramphenicol, erythromycin, and tetracycline. Once emerged, adult flies were crushed and tested on LB-, Brain-Heart infusion Broth-, MRS, and yeast peptone dextrose agar plates to observe any contamination by bacteria or fungi. Of note, no anaerobic microorganisms have been detected in the *Drosophila* microbiota.

Flies treated with antibiotics were fed on food containing ampicillin, tetracycline, chloramphenicol, erythromycin, and kanamycin at 50 μg/ml final concentration each. Females were collected after two generations on fly food with antibiotics and checked for sterility by plating.

A. fumigatus infections and injection of toxins or chemical stress solutions

For *A. fumigatus* infections, spores were prepared freshly for each infection. Unless otherwise stated, spores were injected into the thorax (mesopleuron) of adult flies, usually at a concentration of 250–500 spores in 4.6 nl PBS containing 0.01% Tween20 (PBST) unless indicated otherwise, using a microcapillary connected to a Nanoject II Auto-Nanoliter Injector (Drummond). The same volume of PBS-0.01% Tween20 (PBST) was injected for control experiments. All experiments were performed at 29°C unless otherwise indicated. Prior to all infection experiments, the flies were incubated in tubes containing only 100 mM sucrose solution for 2 days to eliminate traces of antifungal preservatives added to the regular food. Toxins were injected as for *A. fumigatus* injection, except that a toxin or a chemical stress solution was used instead of a spore suspension. As noted, verruculogen powder was directly introduced into flies as follows: the ethanol-cleaned needles were not filled but just dipped into the powder and then used to prick the flies. The procedure was reiterated for each fly. Whereas the injected quantity is not determined with accuracy, it nevertheless yielded reproducible results, which were, however, weaker than when injecting verruculogen initially dissolved into DMSO. Flies were kept on regular food without preservatives after injection.

Saturation of phagocytosis

Latex bead injection was performed as previously described (Nehme *et al*, 2011; Quintin *et al*, 2013). The injected flies were placed on 100 mM sucrose solution for 48 h prior to injections.

Survival tests

Survival tests were usually performed using 5–7-day old flies. Twenty flies per vial in biological triplicates were maintained at 25°C. The transgenic overexpression flies were transferred from 18 to 29°C for 7 days before the challenge to allow the expression of Gal4, which is repressed by the $Gal80^{ts}$ repressor at 18°C. Surviving flies were counted every day. Each experiment shown is representative of at least three independent experiments unless indicated otherwise. For the heat stress experiments, flies were placed either at 29 or 37°C.

Quantification of the fungal load

The fungal burden was determined using single adult flies per condition. Single flies were transferred into arrays of 8 tubes (Starstedt) containing two 1.4-mm ceramic beads (Dominique Dutcher) in 100 µl PBS-0.01% Tween20. Single flies were homogenized by shaking using a mixer mill 300 or 400 (F. Kurt Retsch GmbH & Co. KG) at a frequency of 30/min twice for 30 s and plated on potato dextrose agar (PDA) plates supplemented with antibiotics. After that, the plates were enclosed with Parafilm™ and cultured at 29°C with 65% humidity. Colony-forming units were counted after 48 h. FLUD was performed as described (Duneau *et al*, 2017).

Monitoring of *A. fumigatus* infection *in vivo*

Flies were sacrificed and dissected in 8-well diagnostic microscope slides (Thermo Scientific; Carl Zeiss). was used for negative staining of *ΔpptA's* hyphae, by adding to each well 5 µl Uvitex-2B for 30 s at room temperature. Flies injected by D141-GFP or *ΔpptA* were dissected and observed under an epifluorescent Zeiss axioscope microscope (Carl Zeiss) each hour after the injection.

Preparation UV-killed *A. fumigatus*

The conidial suspension at about 10^{10} conidia/ml was plated on dried potato dextrose agar (PDA) supplemented with 0.1 g/l chloramphenicol plates and exposed twice for 3 h to the UV-light of a microbiology safety hood. Plates were cultured at 29°C with 65% humidity, after 48 h to check for the absence of colonies. The dead conidia were resuspended and counted prior to injection.

Scanning electron microscope

Whole flies were incubated in 1 ml of a solution of 0.1 M phosphate buffer pH 7.2, glutaraldehyde 2.5%, and paraformaldehyde 2.4% final at room temperature for at least 1 h. The flies were embedded in resin prior to observation with a scanning electron microscope (Hitachi S 800).

Drosomycin and *Bomanins* expression measurement

Expression of *Drosomycin* and *Bom* genes was measured by RT–qPCR and RT–digital PCR as described previously (Gottar *et al*, 2006; Madic *et al*, 2016). With respect to digital PCR, the results (# copies/µl) are normalized by the counts of the *Rpl32* reference gene from the same reverse-transcribed sample, as also done for regular RT–qPCR. The sequences of primers are shown in Appendix Table S1.

Restrictocin-mediated inhibition of translation

28S RNA α-sarcin fragment

Total RNA of restrictocin or PBS-injected flies was extracted using Trizol reagent on samples of 2–3 flies. Samples were loaded in the RNA 6000 Nano chip (Agilent RNA 6000 Nano, 2100 electrophoresis Bioanalyzer) to detect the peak corresponding to the α-sarcin fragment.

Level of protein synthesis inhibition in vivo

The level of inhibition of protein by injected restrictocin or the infection by *A. fumigatus in vivo* was assessed by measuring GFP fluorescence in single fly extracts. *w*; *pUbi-Gal4*, $ptub\text{-}Gal80^{ts}$ were crossed to *w*; *UAS-GFP* flies at 18°C. The progeny from the cross was kept at 18°C until *A. fumigatus* or restrictocin injection; the flies were then placed at 29°C thereafter and analyzed at the indicated times. Each fly was homogenized in 200 µl PBS solution prior to measuring the GFP fluorescence using a Varioskan 2000 fluorometer (Thermo Fisher Scientific).

Preparation of in vitro translation extracts from noninduced and Toll-induced ERTL cells

ERTL cells were grown for 5 days at 25°C in 25 ml of culture medium. For the Toll-induced ERTL cells, the culture medium was supplemented with 2.5 µg/ml recombinant mouse EGF (Sigma-Aldrich) 16 h before harvesting.

After harvesting, cells were washed two times in cold 40 mM HEPES–KOH pH 8, 100 mM potassium acetate, 1 mM magnesium acetate, and 1 mM DTT solution, and resuspended at a concentration of 10^9 cells/ml in the same buffer supplemented with 1X Halt™ Protease Inhibitor Cocktail EDTA-free (Thermo Scientific™). Cell lysis was performed by nitrogen cavitation with a Cell Disruption Bomb (Parr Instrument Company). The lysate was cleared by centrifugations at 4°C with 10,000 *g*, aliquoted, frozen in liquid nitrogen, and stored at −80°C. The induction of the Toll pathway was checked by monitoring the transcript levels of *Drosomycin* by RT–qPCR.

In vitro translation assays in rabbit reticulocyte lysate and ERTL cell lysates

In vitro translation experiments in rabbit reticulocyte lysate were performed as previously described (Martin *et al*, 2011). *In vitro* translation experiments in ERTL cell lysates were performed as previously described for S2-cell lysates (Gross *et al*, 2017). eGFP *in vitro* translation was assessed by measuring fluorescence (λex = 485 nm; λem = 520 nm) every minute for 150 min.

Quantification of tremors and the recovery

Tremor quantification was performed after the injection of 4.6 nl of a 1 mg/ml verruculogen solution to batches of 20 w^{A5001} flies placed afterward in an empty vial. Biological triplicates were analyzed for each of the three independent experiments. The rate of tremor cases was measured in each tube 3 h after the injection. Tremor phenotypes are shown in videos available in the supplementary material of this article.

The tremor recovery was measured every 30 min: the flies that had recovered are the ones exhibiting no tremors and able to walk upwards on the sides of an empty vial. Observations were performed every 30 min and flies that had recovered (exhibiting no tremors and able to walk upwards on the sides of the vial) were removed. Three independent experiments were performed.

MALDI mass spectra analysis

MALDI spectra obtained from *Drosophila* hemolymph were acquired and analyzed using FlexControl and Flex Analysis (Bruker Daltonics) software or converted for analysis with the open-source mass spectrometry tool mMass (http://www.mmass.org). A sandwich sample preparation was used, which consists of deposition of (1) 0.5 µl of a saturated solution of 4HCCA in acetone, (2) 0.6 µl of acidified hemolymph in 0.1 µl of trifluoroacetic acid (TFA), and (3) 0.4 µl of a saturated solution of 4HCC in a solution of acetonitrile/0.1TFA (2:1). After a soft drying, spots were acquired in a linear positive mode at an attenuation maintained adjusted between 50 and 60 using a Bruker Daltonics UltraflexIII-Smartbeam instrument. Calibration of the measurements was made using the "DroCal" mixture containing the synthetic peptides BomS1, BomS2, BomS3, and BomS5 as well as a deuterated form of BomS1 (BomS1-ValD at 1,676.50 m/z with $z = 1$).

Quantification and statistical analysis

All statistical analyses were performed using Prism 7 or Prism 8 (GraphPad Software, San Diego, CA). The Mann–Whitney and/or Kruskal–Wallis tests were used unless otherwise indicated. The log-rank test was used to analyze survival experiments. When using parametric tests (analysis of variance (ANOVA) and *t*-test), a Gaussian distribution of data was checked using either D'Agostino-Pearson omnibus or Shapiro–Wilk normality tests. All experiments were performed at least three times, unless otherwise indicated. Significance values: $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

Data availability

In this study, no primary datasets have been generated or deposited in external repositories.

Expanded View for this article is available online.

Acknowledgements

We thank Anne Beauvais and Jean-Paul Latge for the *A. fumigatus* strain used in this study, Won-Jae Lee, Bruno Lemaitre, Jiyong Liu, Steven Wasserman, Akira Goto, Angela Giangrande, and the Guangzhou Drosophila Resource Center for fly stocks. Stocks obtained from the Bloomington Drosophila Stock Center (NIH P40OD018537) were also used in this study. We gratefully acknowledge the contributions of Valérie Demais from Plateforme d'Imagerie *in vitro* (UPS 3156-Université de Strasbourg) for scanning electron microscopy, Sébastien Voisin from the BioPark for MALDI-TOF analysis, and Miriam Yamba for expert technical help. We thank Adrian Acker for the gift of the ERTL S2 cells, controls, and advice on the *in vitro* experimental conditions. Finally, we are indebted to Matthew Blango and Robert Unckless for critical reading of the manuscript. RX and YL were, respectively, partially funded through the Sino-Foreign cooperative graduate education project of Guangzhou Medical University and the International Training Plan for young outstanding scientific research talents of Guangdong Province. This work was supported by the Deutsche Forschungsgemeinschaft collaborative research center/transregion 124 FungiNet (project A1) and the excellence cluster Balance of the Microverse to TH and AB, the Association Platform BioPark of Archamps on its Research & Development budget (PB), by grants from "Agence Nationale pour la Recherche" (ANR-17-CE12-0025) to AT and FM, from the 111 Project (#D18010; China), the Incubation Project for Innovative Teams of the Guangzhou Medical University, the Open Project from State Key Laboratory of Respiratory Diseases, China, and the China High-end Foreign Talent Program to DF.

Author contributions

Rui Xu: Conceptualization; resources; formal analysis; validation; investigation; methodology; writing – original draft; writing – review and editing. **Yanyan Lou:** Conceptualization; resources; formal analysis; investigation; methodology; writing – original draft; writing – review and editing; performed and analyzed the mass spectrometry analysis. **Antonin Tidu:** Formal analysis; validation; investigation; methodology; writing – original draft; designed, performed and analyzed the *in vitro* translation experiments. **Philippe Bulet:** Resources; formal analysis; funding acquisition; validation; investigation; methodology; performed and analyzed the mass spectrometry analysis; generated the *A. fumigatus* mutants reported in this study. **Thorsten Heinekamp:** Resources; writing – original draft; generated the *A. fumigatus* mutants reported in this study. **Franck Martin:** Conceptualization; resources; supervision; funding acquisition; writing – original draft; designed, performed and analyzed the *in vitro* translation experiments. **Axel Brakhage:** Conceptualization; resources; funding

acquisition; writing – original draft; writing – review and editing; generated the *A. fumigatus* mutants reported in this study. **Zi Li:** Conceptualization; resources; funding acquisition; project administration. **Samuel Liégeois:** Conceptualization; resources; formal analysis; supervision; validation; investigation; methodology; writing – original draft. **Dominique Ferrandon:** Conceptualization; resources; supervision; funding acquisition; validation; methodology; writing – original draft; project administration; writing – review and editing.

Disclosure and competing interests statement

The authors declare that they have no conflict of interest.

References

Alarco AM, Marcil A, Chen J, Suter B, Thomas D, Whiteway M (2004) Immune-deficient *Drosophila melanogaster*: a model for the innate immune response to human fungal pathogens. *J Immunol* 172: 5622 – 5628

Apidianakis Y, Rahme LG, Heitman J, Ausubel FM, Calderwood SB, Mylonakis E (2004) Challenge of *Drosophila melanogaster* with *Cryptococcus neoformans* and role of the innate immune response. *Eukaryot Cell* 3: 413 – 419

Benmimoun B, Papastefanaki F, Perichon B, Segklia K, Roby N, Miriagou V, Schmitt C, Dramsi S, Matsas R, Speder P (2020) An original infection model identifies host lipoprotein import as a route for blood-brain barrier crossing. *Nat Commun* 11: 6106

Blango MG, Kniemeyer O, Brakhage AA (2019) Conidial surface proteins at the interface of fungal infections. *PLoS Pathog* 15: e1007939

Bongomin F, Gago S, Oladele RO, Denning DW (2017) Global and multi-national prevalence of fungal diseases-estimate precision. *J Fungi* 3: 57

Brown R, Priest E, Naglik JR, Richardson JP (2021) Fungal toxins and host immune responses. *Front Microbiol* 12: 643639

Chikina AS, Nadalin F, Maurin M, San-Roman M, Thomas-Bonafos T, Li XV, Lameiras S, Baulande S, Henri S, Malissen B *et al* (2020) Macrophages maintain epithelium integrity by limiting fungal product absorption. *Cell* 183: e416

Clemmons AW, Lindsay SA, Wasserman SA (2015) An effector peptide family required for *Drosophila* toll-mediated immunity. *PLoS Pathog* 11: e1004876

Cohen LB, Lindsay SA, Xu Y, Lin SJH, Wasserman SA (2020) The Daisho peptides mediate *Drosophila* defense against a subset of filamentous fungi. *Front Immunol* 11: 9

Cramer RA Jr, Gamcsik MP, Brooking RM, Najvar LK, Kirkpatrick WR, Patterson TF, Balibar CJ, Graybill JR, Perfect JR, Abraham SN *et al* (2006) Disruption of a nonribosomal peptide synthetase in *Aspergillus fumigatus* eliminates gliotoxin production. *Eukaryot Cell* 5: 972 – 980

da Silva Ferreira ME, Kress MR, Savoldi M, Goldman MH, Hartl A, Heinekamp T, Brakhage AA, Goldman GH (2006) The akuB(KU80) mutant deficient for nonhomologous end joining is a powerful tool for analyzing pathogenicity in *Aspergillus fumigatus*. *Eukaryot Cell* 5: 207 – 211

De Gregorio E, Spellman PT, Tzou P, Rubin GM, Lemaitre B (2002) The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. *EMBO J* 21: 2568 – 2579

Dudzic JP, Hanson MA, Iatsenko I, Kondo S, Lemaitre B (2019) More than black or white: melanization and toll share regulatory serine proteases in *Drosophila*. *Cell Rep* 27: e1053

Dunbar TL, Yan Z, Balla KM, Smelkinson MG, Troemel ER (2012) *C. elegans* detects pathogen-induced translational inhibition to activate immune signaling. *Cell Host Microbe* 11: 375 – 386

Duneau D, Ferdy JB, Revah J, Kondolf H, Ortiz GA, Lazzaro BP, Buchon N (2017) Stochastic variation in the initial phase of bacterial infection predicts the probability of survival in *D. melanogaster*. *Elife* 6: e28298

Fando JL, Alaba I, Escarmis C, Fernandez-Luna JL, Mendez E, Salinas M (1985) The mode of action of restrictocin and mitogillin on eukaryotic ribosomes. Inhibition of brain protein synthesis, cleavage and sequence of the ribosomal RNA fragment. *Eur J Biochem* 149: 29 – 34

Fehlbaum P, Bulet P, Michaut L, Lagueux M, Brockaert WF, Hétru C, Hoffmann JA (1995) Septic injury of *Drosophila* induces the synthesis of a potent antifungal peptide with sequence homology to plant antifungal peptides. *J Biol Chem* 269: 33159 – 33163

Ferrandon D (2013) The complementary facets of epithelial host defenses in the genetic model organism *Drosophila melanogaster*: from resistance to resilience. *Curr Opin Immunol* 25: 59 – 70

Ferreira AG, Naylor H, Esteves SS, Pais IS, Martins NE, Teixeira L (2014) The Toll-dorsal pathway is required for resistance to viral oral infection in *Drosophila*. *PLoS Pathog* 10: e1004507

Frisvad JC, Rank C, Nielsen KF, Larsen TO (2009) Metabolomics of *Aspergillus fumigatus*. *Med Mycol* 47: S53 – S71

Gago S, Denning DW, Bowyer P (2019) Pathophysiological aspects of *Aspergillus* colonization in disease. *Med Mycol* 57: S219 – S227

Gant DB, Cole RJ, Valdes JJ, Eldefrawi ME, Eldefrawi AT (1987) Action of tremorgenic mycotoxins on GABAA receptor. *Life Sci* 41: 2207 – 2214

Gao Q, Jin K, Ying SH, Zhang Y, Xiao G, Shang Y, Duan Z, Hu X, Xie XQ, Zhou G *et al* (2011) Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. *PLoS Genet* 7: e1001264

Gluck A, Endo Y, Wool IG (1994) The ribosomal RNA identity elements for ricin and for alpha-sarcin: mutations in the putative CG pair that closes a GAGA tetraloop. *Nucleic Acids Res* 22: 321 – 324

Gottar M, Gobert V, Matskevich AA, Reichhart JM, Wang C, Butt TM, Belvin M, Hoffmann JA, Ferrandon D (2006) Dual detection of fungal infections in *Drosophila* via recognition of glucans and sensing of virulence factors. *Cell* 127: 1425 – 1437

Greaney AJ, Leppla SH, Moayeri M (2015) Bacterial exotoxins and the inflammasome. *Front Immunol* 6: 570

Gross L, Vicens Q, Einhorn E, Noireterre A, Schaeffer L, Kuhn L, Imler JL, Eriani G, Meignin C, Martin F (2017) The IRES 5′UTR of the dicistrovirus cricket paralysis virus is a type III IRES containing an essential pseudoknot structure. *Nucleic Acids Res* 45: 8993 – 9004

Hanson MA, Lemaitre B (2020) New insights on *Drosophila* antimicrobial peptide function in host defense and beyond. *Curr Opin Immunol* 62: 22 – 30

Hanson MA, Dostalova A, Ceroni C, Poidevin M, Kondo S, Lemaitre B (2019) Synergy and remarkable specificity of antimicrobial peptides *in vivo* using a systematic knockout approach. *Elife* 8: e44341

Hillmann F, Novohradska S, Mattern DJ, Forberger T, Heinekamp T, Westermann M, Winckler T, Brakhage AA (2015) Virulence determinants of the human pathogenic fungus *Aspergillus fumigatus* protect against soil amoeba predation. *Environ Microbiol* 17: 2858 – 2869

Hotujac L, Muftic RH, Filipovic N (1976) Verruculogen: a new substance for decreasing of GABA levels in CNS. *Pharmacology* 14: 297 – 300

Huang J, Lou Y, Liu J, Bulet P, Jiao R, Hoffmann JA, Liegeois S, Li Z, Ferrandon D (2022) A Toll pathway effector protects *Drosophila* specifically from distinct toxins secreted by a fungus or a bacterium. *bioRxiv* https://doi.org/10.1101/2020.11.23.394809 [PREPRINT]

Jahn B, Koch A, Schmidt A, Wanner G, Gehringer H, Bhakdi S, Brakhage AA (1997) Isolation and characterization of a pigmentless-conidium mutant of

Aspergillus fumigatus with altered conidial surface and reduced virulence. *Infect Immun* 65: 5110 – 5117

Johns A, Scharf DH, Gsaller F, Schmidt H, Heinekamp T, Strassburger M, Oliver JD, Birch M, Beckmann N, Dobb KS *et al* (2017) A nonredundant phosphopantetheinyl transferase, PptA, is a novel antifungal target that directs secondary metabolite, siderophore, and lysine biosynthesis in *Aspergillus fumigatus* and is critical for pathogenicity. *mBio* 8: e01504-16

Kato N, Suzuki H, Okumura H, Takahashi S, Osada H (2013) A point mutation in ftmD blocks the fumitremorgin biosynthetic pathway in *Aspergillus fumigatus* strain Af293. *Biosci Biotechnol Biochem* 77: 1061 – 1067

Knaus HG, McManus OB, Lee SH, Schmalhofer WA, Garcia-Calvo M, Helms LM, Sanchez M, Giangiacomo K, Reuben JP, Smith AB 3rd *et al* (1994) Tremorgenic indole alkaloids potently inhibit smooth muscle high-conductance calcium-activated potassium channels. *Biochemistry* 33: 5819 – 5828

Konig S, Pace S, Pein H, Heinekamp T, Kramer J, Romp E, Strassburger M, Troisi F, Proschak A, Dworschak J *et al* (2019) Gliotoxin from *Aspergillus fumigatus* abrogates leukotriene B4 formation through inhibition of leukotriene A4 hydrolase. *Cell Chem Biol* 26: e525

Kosmidis C, Denning DW (2015) The clinical spectrum of pulmonary aspergillosis. *Thorax* 70: 270 – 277

Kubodera T, Yamashita N, Nishimura A (2000) Pyrithiamine resistance gene (ptrA) of *Aspergillus oryzae*: cloning, characterization and application as a dominant selectable marker for transformation. *Biosci Biotechnol Biochem* 64: 1416 – 1421

Kudryashova E, Quintyn R, Seveau S, Lu W, Wysocki VH, Kudryashov DS (2014) Human defensins facilitate local unfolding of thermodynamically unstable regions of bacterial protein toxins. *Immunity* 41: 709 – 721

Kupfahl C, Heinekamp T, Geginat G, Ruppert T, Hartl A, Hof H, Brakhage AA (2006) Deletion of the gliP gene of *Aspergillus fumigatus* results in loss of gliotoxin production but has no effect on virulence of the fungus in a low-dose mouse infection model. *Mol Microbiol* 62: 292 – 302

Lamy B, Moutaouakil M, Latge JP, Davies J (1991) Secretion of a potential virulence factor, a fungal ribonucleotoxin, during human aspergillosis infections. *Mol Microbiol* 5: 1811 – 1815

Lazzaro BP, Zasloff M, Rolff J (2020) Antimicrobial peptides: application informed by evolution. *Science* 368: eaau5480

Lebrigand K, He LD, Thakur N, Arguel MJ, Polanowska J, Henrissat B, Record E, Magdelenat G, Barbe V, Raffaele S *et al* (2016) Comparative genomic analysis of *Drechmeria coniospora* reveals Core and specific genetic requirements for fungal endoparasitism of nematodes. *PLoS Genet* 12: e1006017

Lee KZ, Lestradet M, Socha C, Schirmeier S, Schmitz A, Spenle C, Lefebvre O, Keime C, Yamba WM, Bou Aoun R *et al* (2016) Enterocyte purge and rapid recovery is a resilience reaction of the gut epithelium to pore-forming toxin attack. *Cell Host Microbe* 20: 716 – 730

Lemaitre B, Hoffmann J (2007) The host defense of *Drosophila melanogaster*. *Annu Rev Immunol* 25: 697 – 743

Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA (1996) The dorsoventral regulatory gene cassette *spätzle/Toll/cactus* controls the potent antifungal response in *Drosophila* adults. *Cell* 86: 973 – 983

Lesperance DN, Broderick NA (2020) Microbiomes as modulators of *Drosophila melanogaster* homeostasis and disease. *Curr Opin Insect Sci* 39: 84 – 90

Levashina EA, Ohresser S, Bulet P, Reichhart J-M, Hétru C, Hoffmann JA (1995) Metchnikowin, a novel immune-inducible proline-rich peptide from *Drosophila* with antibacterial and antifungal properties. *Eur J Biochem* 233: 694 – 700

Liegeois S, Ferrandon D (2022) Sensing microbial infections in the *Drosophila melanogaster* genetic model organism. *Immunogenetics* 74: 35 – 62

Lin SJH, Cohen LB, Wasserman SA (2020) Effector specificity and function in *Drosophila* innate immunity: getting AMPed and dropping Boms. *PLoS Pathog* 16: e1008480

Lindsay SA, Lin SJH, Wasserman SA (2018) Short-form Bomanins mediate humoral immunity in *Drosophila*. *J Innate Immun* 10: 306 – 314

Macheleidt J, Mattern DJ, Fischer J, Netzker T, Weber J, Schroeckh V, Valiante V, Brakhage AA (2016) Regulation and role of fungal secondary metabolites. *Annu Rev Genet* 50: 371 – 392

Madic J, Zocevic A, Senlis V, Fradet E, Andre B, Muller S, Dangla R, Droniou ME (2016) Three-color crystal digital PCR. *Biomol Detect Quantif* 10: 34 – 46

Martin F, Barends S, Jaeger S, Schaeffer L, Prongidi-Fix L, Eriani G (2011) Cap-assisted internal initiation of translation of histone H4. *Mol Cell* 41: 197 – 209

McEwan DL, Kirienko NV, Ausubel FM (2012) Host translational inhibition by *Pseudomonas aeruginosa* exotoxin A triggers an immune response in *Caenorhabditis elegans*. *Cell Host Microbe* 11: 364 – 374

Medzhitov R, Schneider DS, Soares MP (2012) Disease tolerance as a defense strategy. *Science* 335: 936 – 941

Melo JA, Ruvkun G (2012) Inactivation of conserved *C. elegans* genes engages pathogen- and xenobiotic-associated defenses. *Cell* 149: 452 – 466

Nam HJ, Jang IH, You H, Lee KA, Lee WJ (2012) Genetic evidence of a redox-dependent systemic wound response via Hayan protease-phenoloxidase system in *Drosophila*. *EMBO J* 31: 1253 – 1265

Nayak SK, Bagga S, Gaur D, Nair DT, Salunke DM, Batra JK (2001) Mechanism of specific target recognition and RNA hydrolysis by ribonucleolytic toxin restrictocin. *Biochemistry* 40: 9115 – 9124

Nehme NT, Quintin J, Cho JH, Lee J, Lafarge MC, Kocks C, Ferrandon D (2011) Relative roles of the cellular and humoral responses in the *Drosophila* host defense against three gram-positive bacterial infections. *PLoS One* 6: e14743

Norris PJ, Smith CC, De Belleroche J, Bradford HF, Mantle PG, Thomas AJ, Penny RH (1980) Actions of tremorgenic fungal toxins on neurotransmitter release. *J Neurochem* 34: 33 – 42

Pahl HL, Krauss B, Schulze-Osthoff K, Decker T, Traenckner EB, Vogt M, Myers C, Parks T, Warring P, Muhlbacher A *et al* (1996) The immunosuppressive fungal metabolite gliotoxin specifically inhibits transcription factor NF-kappaB. *J Exp Med* 183: 1829 – 1840

Quintin J, Asmar J, Matskevich AA, Lafarge MC, Ferrandon D (2013) The *Drosophila* Toll pathway controls but does not clear *Candida glabrata* infections. *J Immunol* 190: 2818 – 2827

Raffa N, Keller NP (2019) A call to arms: mustering secondary metabolites for success and survival of an opportunistic pathogen. *PLoS Pathog* 15: e1007606

Raisch T, Brockmann A, Ebbinghaus-Kintscher U, Freigang J, Gutbrod O, Kubicek J, Maertens B, Hofnagel O, Raunser S (2021) Small molecule modulation of the *Drosophila* Slo channel elucidated by cryo-EM. *Nat Commun* 12: 7164

Rodrigues ML, Nosanchuk JD (2020) Fungal diseases as neglected pathogens: a wake-up call to public health officials. *PLoS Negl Trop Dis* 14: e0007964

Scharf DH, Heinekamp T, Brakhage AA (2014) Human and plant fungal pathogens: the role of secondary metabolites. *PLoS Pathog* 10: e1003859

Soares MP, Teixeira L, Moita LF (2017) Disease tolerance and immunity in host protection against infection. *Nat Rev Immunol* 17: 83 – 96

Sun H, Towb P, Chiem DN, Foster BA, Wasserman SA (2004) Regulated assembly of the Toll signaling complex drives *Drosophila* dorsoventral patterning. *EMBO J* 23: 100 – 110

Szewczyk E, Nayak T, Oakley CE, Edgerton H, Xiong Y, Taheri-Talesh N, Osmani SA, Oakley BR (2006) Fusion PCR and gene targeting in *Aspergillus nidulans*. *Nat Protoc* 1: 3111 – 3120

Uttenweiler-Joseph S, Moniatte M, Lagueux M, Van Dorsselaer A, Hoffmann JA, Bulet P (1998) Differential display of peptides induced during the immune response of *Drosophila*: a matrix-assisted laser desorption ionization time-of-flight mass spectrometry study. *Proc Natl Acad Sci USA* 95: 11342 – 11347

van de Veerdonk FL, Gresnigt MS, Romani L, Netea MG, Latge JP (2017) *Aspergillus fumigatus* morphology and dynamic host interactions. *Nat Rev Microbiol* 15: 661 – 674

CONCLUSION GENERALE

3. Conclusion générale

Les approches *in vitro* et de criblage à haut-débit mises au point dans le cadre de ces travaux de thèse ont permis la caractérisation systématique d'éléments *cis*-régulateurs de la traduction chez les eucaryotes supérieurs.

Le criblage par un système micro-fluidique des contextes nucléotidiques autour du codon initiateur AUG a suggéré qu'ils sont tous susceptibles d'exercer une influence significative qui leur est propre sur les interactions établies par la particule 43S avec l'ARNm, en fonction de la nature des nucléotides présents sur chaque position. Cette hétérogénéité d'influence sur l'initiation est davantage appuyée par la diversité des contextes nucléotidiques des codons initiateurs AUG trouvés dans les phases codantes humaines. Malgré la diversité des contextes nucléotidiques acceptés par le ribosome, ceux-ci ne représentent que 10% de l'ensemble des contextes de neuf nucléotides possibles. Dans l'ensemble, ces observations montrent que le contexte nucléotidique du codon d'initiation est un élément *cis*-régulateur majeur de l'initiation traduction chez les eucaryotes. Cependant, l'influence qu'il exerce sur cette dernière ne peut être expliquée par un nombre restreint de combinaisons de nucléotides minimales, comme cela est suggéré dans la littérature. Par ailleurs, l'initiation de la traduction sur un codon AUG-like est facilitée par la présence d'une structure d'ARN suffisamment stable située à une distance optimale, en aval du codon d'initiation. Cette structure permettrait de stopper la particule 43S pendant l'étape de scanning de manière à stabiliser l'interaction codon-anticodon dans le site P du ribosome. La présence de telles structures dans les régions 5'UTR des ARNm humains est actuellement à l'étude dans le but d'identifier de nouvelles uORFs dont la traduction ne démarre pas par un codon AUG.

La méthode de criblage basée sur le fractionnement de complexes de traduction sur gradients de saccharose a permis de retrouver les deux IRES présents dans le génome du virus de la paralysie du cricket. De plus, d'autres régions du génome viral ont été sélectionnées. Des expériences complémentaires sont nécessaires pour déterminer si ces régions sont bien des éléments *cis*-régulateurs de la traduction, ou plutôt des artefacts liés à la méthodologie de criblage. Dès lors, cette approche nécessite davantage de mises au point pour améliorer sa spécificité de sélection. Plusieurs pistes d'amélioration sont actuellement à l'étude et seront prochainement mises en œuvre. Une fois pleinement validée, cette méthode pourra être utilisée pour identifier dans un premier temps de potentiels IRES contenus les génomes des virus de la Dengue et du Zika dans des contextes biologiquement pertinents. Des extraits de traduction *in vitro* réalisés à partir de cellules embryonnaires de moustique tigre (Aag2) et humaines (HEK) ont été développés dans cette optique.

La traduction sélective de l'ARN génomique et des ARN sous-génomiques du SARS-CoV-2 fait intervenir la coopération d'éléments *cis*- et *trans*- régulateurs. En présence de la protéine virale NSP1 qui bloque les ribosomes cellulaires, la traduction des ARN du SARS-CoV-2 se poursuit grâce à la présence d'une structure en tige-boucle à son extrémité 5' qui délogerait NSP1 du ribosome, permettant ainsi la traduction sélective de l'ARN viral au détriment de la traduction des ARNm cellulaires. La détermination de la structure d'un complexe de pré-initiation en présence de l'ARN viral et de la protéine NSP1 par cryo-Microscopie Electronique devrait permettre d'élucider les mécanismes moléculaires à l'origine de cette évasion.

L'ensemble de ces études corrobore et précise le caractère multifactoriel de l'initiation de la traduction chez les eucaryotes. La stabilisation de l'interaction codon-anticodon dans le site P est dépendante de la contribution énergétique amenée par la coopération de facteurs *cis*- et *trans*- régulateurs. L'énergie d'activation requise pour l'initiation de la traduction peut être apportée par une grande variété de mécanismes et d'interactions moléculaires qui ne sont que partiellement compris à l'heure actuelle.

Les outils moléculaires qui ont été mis au point dans le cadre de ce travail de thèse ont été et seront mis à profit pour d'autres études de l'initiation de la traduction au sein du laboratoire.

Par exemple, le système rapporteur basé sur l'IRES modifié du virus EMCV mis au point pour le criblage des contextes du codon AUG a été adapté dans le cadre du projet de mesure de la vitesse de scanning de la particule 43S par microscopie TIRF. Ce rapporteur permet notamment d'étudier l'influence de structures d'ARN situées dans la région 5'UTR sur la vitesse de scanning. Cette étude des paramètres cinétiques de l'initiation la traduction est réalisée en collaboration avec l'équipe de Karen Perronet à Paris.

La méthode de préparation d'extraits de traduction *in vitro* à partir de cellules embryonnaires de drosophiles (S2) a été adaptée pour permettre d'y induire la voie Toll et ainsi étudier son rôle dans la traduction en présence de mycotoxines et de peptides antimicrobiens. Ces travaux ont été réalisés en collaboration avec le laboratoire de Dominique Ferrandon de l'UPR-9022 à Strasbourg et se sont soldés par une publication dans le journal *EMBO Reports*.

Des extraits de traduction acellulaires préparés à partir de cellules de neuroblastomes (SH-SY5Y) ont également été développés dans le cadre de ce travail de thèse. Ces extraits sont considérés comme un système pertinent pour l'étude de la traduction neuronale, et sont actuellement utilisés dans le cadre d'un projet de recherche qui s'articule autour de la traduction de l'ARNm codant pour la protéine Tau chez les patients atteints de la maladie d'Alzheimer. Ce projet est réalisé en collaboration avec le laboratoire de Luc Buée à Lille.

Par ailleurs, le développement et la caractérisation d'extraits de traduction réalisés à partir de cellules HEK cultivées en conditions physiologiques et en conditions de stress se sont avérés utiles dans d'autres projets du laboratoire, et ce protocole pourra être adapté en fonction du type de stress à étudier mais également en fonction du type cellulaire. Par exemple, l'utilisation de ces extraits a permis de mieux décortiquer les mécanismes moléculaires mis en jeu dans la traduction de l'ARNm C9ORF72 qui mène à la synthèse aberrante de poly-dipeptides toxiques qui sont retrouvés dans les motoneurones des patients atteints de la maladie de Charcot, également appelée sclérose latérale amyotrophique. Ce travail est réalisé en étroite collaboration avec le laboratoire de Clotilde Lagier-Tourenne à Harvard (USA).

Dans l'ensemble, les différents projets abordés au cours de ce travail de thèse ont permis le développement d'outils moléculaires et informatiques versatiles qui pourront être mis à profit, moyennant quelques adaptations, dans le cadre d'autres projets d'étude de l'initiation de la traduction chez les eucaryotes au sein du laboratoire.

BIBLIOGRAPHIE

4. Bibliographie

Abeyrathne PD, Koh CS, Grant T, Grigorieff N, Korostelev AA. 2016. Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome. *eLife* **5**: e14874.

Aitken CE, Lorsch JR. 2012. A mechanistic overview of translation initiation in eukaryotes. *Nat Struct Mol Biol* **19**: 568–576.

Alghoul F, Laure S, Eriani G, Martin F. 2021. Translation inhibitory elements from Hoxa3 and Hoxa11 mRNAs use uORFs for translation inhibition. *eLife* **10**: e66369.

Algire MA, Maag D, Lorsch JR. 2005. Pi Release from eIF2, Not GTP Hydrolysis, Is the Step Controlled by Start-Site Selection during Eukaryotic Translation Initiation. *Mol Cell* **20**: 251–262.

Ambrosini C, Destefanis E, Kheir E, Broso F, Alessandrini F, Longhi S, Battisti N, Pesce I, Dassi E, Petris G, et al. 2022. Translational enhancement by base editing of the Kozak sequence rescues haploinsufficiency. *Nucleic Acids Res* **50**: 10756–10771.

Andino R, Rieckhof GE, Achacoso PL, Baltimore D. 1993. Poliovirus RNA synthesis utilizes an RNP complex formed around the 5′-end of viral RNA. *EMBO J* **12**: 3587–3598.

Andino R, Rieckhof GE, Baltimore D. 1990. A functional ribonucleoprotein complex forms around the 5′ end of poliovirus RNA. *Cell* **63**: 369–380.

Andreev DE, Dmitriev SE, Zinovkin R, Terenin IM, Shatsky IN. 2012. The 5′ untranslated region of Apaf-1 mRNA directs translation under apoptosis conditions via a 5′ end-dependent scanning mechanism. *FEBS Lett* **586**: 4139–4143.

Asano K, Clayton J, Shalev A, Hinnebusch AG. 2000. A multifactor complex of eukaryotic initiation factors, eIF1, eIF2, eIF3, eIF5, and initiator tRNA ^{Met} is an important translation initiation intermediate in vivo. *Genes Dev* **14**: 2534–2546.

Au HH, Cornilescu G, Mouzakis KD, Ren Q, Burke JE, Lee S, Butcher SE, Jan E. 2015. Global shape mimicry of tRNA within a viral internal ribosome entry site mediates translational reading frame selection. *Proc Natl Acad Sci* **112**. https://pnas.org/doi/full/10.1073/pnas.1512088112 (Accessed November 25, 2022).

Babaylova ES, Gopanenko AV, Bulygin KN, Tupikin AE, Kabilov MR, Malygin AA, Karpova GG. 2020. mRNA regions where 80S ribosomes pause during translation elongation in vivo interact with protein uS19, a component of the decoding site. *Nucleic Acids Res* **48**: 912–923.

Ban N, Beckmann R, Cate JH, Dinman JD, Dragon F, Ellis SR, Lafontaine DL, Lindahl L, Liljas A, Lipton JM, et al. 2014. A new system for naming ribosomal proteins. *Curr Opin Struct Biol* **24**: 165–169.

Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, et al. 2014. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J* **33**: 981–993.

Behrmann E, Loerke J, Budkevich TV, Yamamoto K, Schmidt A, Penczek PA, Vos MR, Bürger J, Mielke T, Scheerer P, et al. 2015. Structural Snapshots of Actively Translating Human Ribosomes. *Cell* **161**: 845–857.

Belcourt MF, Farabaugh PJ. 1990. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell* **62**: 339–352.

Belsham GJ, McInerney GM, Ross-Smith N. 2000. Foot-and-Mouth Disease Virus 3C Protease Induces Cleavage of Translation Initiation Factors eIF4A and eIF4G within Infected Cells. *J Virol* **74**: 272–280.

Bertram G, Bell HA, Ritchie DW, Fullerton G, Stansfield I. 2000. Terminating eukaryote translation: Domain 1 of release factor eRF1 functions in stop codon recognition. *RNA* **6**: 1236–1247.

Bessa C, Matos P, Jordan P, Gonçalves V. 2020. Alternative Splicing: Expanding the Landscape of Cancer Biomarkers and Therapeutics. *Int J Mol Sci* **21**: 9032.

Betat H, Long Y, Jackman J, Mörl M. 2014. From End to End: tRNA Editing at 5'- and 3'-Terminal Positions. *Int J Mol Sci* **15**: 23975–23998.

Bhaskar V, Graff-Meyer A, Schenk AD, Cavadini S, von Loeffelholz O, Natchiar SK, Artus-Revel CG, Hotz H-R, Bretones G, Klaholz BP, et al. 2020. Dynamics of uS19 C-Terminal Tail during the Translation Elongation Cycle in Human Ribosomes. *Cell Rep* **31**: 107473.

Bhatt PR, Scaiola A, Loughran G, Leibundgut M, Kratzel A, Meurs R, Dreos R, O'Connor KM, McMillan A, Bode JW, et al. 2021. Structural basis of ribosomal frameshifting during translation of the SARS-CoV-2 RNA genome. *Science* **372**: 1306–1313.

Bohlen J, Harbrecht L, Blanco S, Clemm von Hohenberg K, Fenzl K, Kramer G, Bukau B, Teleman AA. 2020. DENR promotes translation reinitiation via ribosome recycling to drive expression of oncogenes including ATF4. *Nat Commun* **11**: 4676.

Bou-Nader C, Gordon JM, Henderson FE, Zhang J. 2019. The search for a PKR code—differential regulation of protein kinase R activity by diverse RNA and protein regulators. *RNA* **25**: 539–556.

Boyce M, Bryant KF, Jousse C, Long K, Harding HP, Scheuner D, Kaufman RJ, Ma D, Coen DM, Ron D, et al. 2005. A Selective Inhibitor of eIF2α Dephosphorylation Protects Cells from ER Stress. *Science* **307**: 935–939.

Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. 2015. Structural basis for stop codon recognition in eukaryotes. *Nature* **524**: 493–496.

Brunn GJ, Hudson CC, Sekulić A, Williams JM, Hosoi H, Houghton PJ, Lawrence JC, Abraham RT. 1997. Phosphorylation of the Translational Repressor PHAS-I by the Mammalian Target of Rapamycin. *Science* **277**: 99–101.

Budkevich T, Giesebrecht J, Altman RB, Munro JB, Mielke T, Nierhaus KH, Blanchard SC, Spahn CMT. 2011. Structure and Dynamics of the Mammalian Ribosomal Pretranslocation Complex. *Mol Cell* **44**: 214–224.

Bugaud O, Barbier N, Chommy H, Fiszman N, Le Gall A, Dulin D, Saguy M, Westbrook N, Perronet K, Namy O. 2017. Kinetics of CrPV and HCV IRES-mediated eukaryotic translation using single-molecule fluorescence microscopy. *RNA* **23**: 1626–1635.

Burke PC, Park H, Subramaniam AR. 2022. A nascent peptide code for translational control of mRNA stability in human cells. *Nat Commun* **13**: 6829.

Burrill CP, Westesson O, Schulte MB, Strings VR, Segal M, Andino R. 2013. Global RNA Structure Analysis of Poliovirus Identifies a Conserved RNA Structure Involved in Viral Replication and Infectivity. *J Virol* **87**: 11670–11683.

Buttgereit F, Brand MD. 1995. A hierarchy of ATP-consuming processes in mammalian cells. *Biochem J* **312**: 163–167.

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* **106**: 7507–7512.

Chen J, Brunner A-D, Cogan JZ, Nuñez JK, Fields AP, Adamson B, Itzhak DN, Li JY, Mann M, Leonetti MD, et al. 2020. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**: 1140–1146.

Chen YG, Chen R, Ahmad S, Verma R, Kasturi SP, Amaya L, Broughton JP, Kim J, Cadena C, Pulendran B, et al. 2019. N6-Methyladenosine Modification Controls Circular RNA Immunity. *Mol Cell* **76**: 96-109.e9.

Choi J, Ieong K-W, Demirci H, Chen J, Petrov A, Prabhakar A, O'Leary SE, Dominissini D, Rechavi G, Soltis SM, et al. 2016. N6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics. *Nat Struct Mol Biol* **23**: 110–115.

Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, Zhihao NT, Harmston N, D'Agostino G, Whiffin N, et al. 2022. A high-resolution map of human RNA translation. *Mol Cell* **82**: 2885-2899.e8.

Cimarelli A, Luban J. 1999. Translation Elongation Factor 1-Alpha Interacts Specifically with the Human Immunodeficiency Virus Type 1 Gag Polyprotein. *J Virol* **73**: 5388–5401.

Coldwell MJ, Mitchell SA, Stoneley M, MacFarlane M, Willis AE. 2000. Initiation of Apaf-1 translation by internal ribosome entry. *Oncogene* **19**: 899–905.

Cowling VH. 2010. Regulation of mRNA cap methylation. *Biochem J* **425**: 295–302.

Das S, Vera M, Gandin V, Singer RH, Tutucci E. 2021. Intracellular mRNA transport and localized translation. *Nat Rev Mol Cell Biol* **22**: 483–504.

de Breyne S, Bonderoff JM, Chumakov KM, Lloyd RE, Hellen CUT. 2008. Cleavage of eukaryotic initiation factor eIF5B by enterovirus 3C proteases. *Virology* **378**: 118–122.

Delatte B, Wang F, Ngoc LV, Collignon E, Bonvin E, Deplus R, Calonne E, Hassabi B, Putmans P, Awe S, et al. 2016. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351**: 282–285.

des Georges A, Dhote V, Kuhn L, Hellen CUT, Pestova TV, Frank J, Hashem Y. 2015. Structure of mammalian eIF3 in the context of the 43S preinitiation complex. *Nature* **525**: 491–495.

Despons L, Martin F. 2020. How Many Messenger RNAs Can Be Translated by the START Mechanism? *Int J Mol Sci* **21**: 8373.

Dever TE, Dinman JD, Green R. 2018. Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harb Perspect Biol* **10**: a032649.

Dever TE, Ivanov IP, Sachs MS. 2020. Conserved Upstream Open Reading Frame Nascent Peptides That Control Translation. *Annu Rev Genet* **54**: 237–264.

Diaz de Arce AJ, Noderer WL, Wang CL. 2018. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res* **46**: 985–994.

Dmitriev SE, Terenin IM, Andreev DE, Ivanov PA, Dunaevsky JE, Merrick WC, Shatsky IN. 2010. GTP-independent tRNA Delivery to the Ribosomal P-site by a Novel Eukaryotic Translation Factor. *J Biol Chem* **285**: 26779–26787.

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. 2012. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* **485**: 201–206.

Dominissini D, Nachtergaele S, Moshitch-Moshkovitz S, Peer E, Kol N, Ben-Haim MS, Dai Q, Di Segni A, Salmon-Divon M, Clark WC, et al. 2016. The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **530**: 441–446.

Donnelly MLL, Luke G, Mehrotra A, Li X, Hughes LE, Gani D, Ryan MD. 2001. Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip.' *J Gen Virol* **82**: 1013–1025.

Dorrello NV, Peschiaroli A, Guardavaccaro D, Colburn NH, Sherman NE, Pagano M. 2006. S6K1- and ßTRCP-Mediated Degradation of PDCD4 Promotes Protein Translation and Cell Growth. *Science* **314**: 467–471.

Dumas L, Herviou P, Dassi E, Cammas A, Millevoi S. 2021. G-Quadruplexes in RNA Biology: Recent Advances and Future Directions. *Trends Biochem Sci* **46**: 270–283.

Dunn EF, Connor JH. 2011. Dominant Inhibition of Akt/Protein Kinase B Signaling by the Matrix Protein of a Negative-Strand RNA Virus. *J Virol* **85**: 422–431.

Elfakess R, Dikstein R. 2008. A Translation Initiation Element Specific to mRNAs with Very Short 5′UTR that Also Regulates Transcription ed. S. Maas. *PLoS ONE* **3**: e3094.

Eriani G, Delarue M, Poch O, Gangloff J, Moras D. 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* **347**: 203–206.

Eriani G, Martin F. 2018. START: STructure-Assisted RNA Translation. *RNA Biol* **15**: 1250–1253.

Erzberger JP, Stengel F, Pellarin R, Zhang S, Schaefer T, Aylett CHS, Cimermančič P, Boehringer D, Sali A, Aebersold R, et al. 2014. Molecular Architecture of the 40S·eIF1·eIF3 Translation Initiation Complex. *Cell* **158**: 1123–1135.

Fraser CS, Hershey JWB, Doudna JA. 2009. The pathway of hepatitis C virus mRNA recruitment to the human ribosome. *Nat Struct Mol Biol* **16**: 397–404.

Fuchs G, Petrov AN, Marceau CD, Popov LM, Chen J, O'Leary SE, Wang R, Carette JE, Sarnow P, Puglisi JD. 2015. Kinetic pathway of 40S ribosomal subunit recruitment to hepatitis C virus internal ribosome entry site. *Proc Natl Acad Sci* **112**: 319–325.

Garreau de Loubresse N, Prokhorova I, Holtkamp W, Rodnina MV, Yusupova G, Yusupov M. 2014. Structural basis for the inhibition of the eukaryotic ribosome. *Nature* **513**: 517–522.

Gast LV, Völker S, Utzschneider M, Linz P, Wilferth T, Müller M, Kopp C, Hensel B, Uder M, Nagel AM. 2021. Combined imaging of potassium and sodium in human skeletal muscle tissue at 7 T. *Magn Reson Med* **85**: 239–253.

George A, Panda S, Kudmulwar D, Chhatbar SP, Nayak SC, Krishnan HH. 2012. Hepatitis C Virus NS5A Binds to the mRNA Cap-binding Eukaryotic Translation Initiation 4F (eIF4F) Complex and Up-regulates Host Translation Initiation Machinery through eIF4E-binding Protein 1 Inactivation. *J Biol Chem* **287**: 5042–5058.

Gingras AC, Svitkin Y, Belsham GJ, Pause A, Sonenberg N. 1996. Activation of the translational suppressor 4E-BP1 following infection with encephalomyocarditis virus and poliovirus. *Proc Natl Acad Sci* **93**: 5578–5583.

Godet A-C, David F, Hantelys F, Tatin F, Lacazette E, Garmy-Susini B, Prats A-C. 2019. IRES Trans-Acting Factors, Key Actors of the Stress Response. *Int J Mol Sci* **20**: 924.

Gradi A, Svitkin YV, Imataka H, Sonenberg N. 1998. Proteolysis of human eukaryotic translation initiation factor eIF4GII, but not eIF4GI, coincides with the shutoff of host protein synthesis after poliovirus infection. *Proc Natl Acad Sci* **95**: 11089–11094.

Gromadski KB, Schümmer T, Strømgaard A, Knudsen CR, Kinzy TG, Rodnina MV. 2007. Kinetics of the Interactions between Yeast Elongation Factors 1A and 1Bα, Guanine Nucleotides, and Aminoacyl-tRNA. *J Biol Chem* **282**: 35629–35637.

Grosjean H, Westhof E. 2016. An integrated, structure- and energy-based view of the genetic code. *Nucleic Acids Res* **44**: 8020–8040.

Gross JD, Moerke NJ, von der Haar T, Lugovskoy AA, Sachs AB, McCarthy JEG, Wagner G. 2003. Ribosome Loading onto the mRNA Cap Is Driven by Conformational Coupling between eIF4G and eIF4E. *Cell* **115**: 739–750.

Gross L, Vicens Q, Einhorn E, Noireterre A, Schaeffer L, Kuhn L, Imler J-L, Eriani G, Meignin C, Martin F. 2017. The IRES5′UTR of the dicistrovirus cricket paralysis virus is a type III IRES containing an essential pseudoknot structure. *Nucleic Acids Res* **45**: 8993–9004.

Guenther U-P, Weinberg DE, Zubradt MM, Tedeschi FA, Stawicki BN, Zagore LL, Brar GA, Licatalosi DD, Bartel DP, Weissman JS, et al. 2018. The helicase Ded1p controls use of near-cognate translation initiation codons in 5′ UTRs. *Nature* **559**: 130–134.

Gutierrez E, Shin B-S, Woolstenhulme CJ, Kim J-R, Saini P, Buskirk AR, Dever TE. 2013. eIF5A Promotes Translation of Polyproline Motifs. *Mol Cell* **51**: 35–45.

Haimov O, Sinvani H, Dikstein R. 2015. Cap-dependent, scanning-free translation initiation mechanisms. *Biochim Biophys Acta BBA - Gene Regul Mech* **1849**: 1313–1318.

Harding HP, Zhang Y, Bertolotti A, Zeng H, Ron D. 2000. Perk Is Essential for Translational Regulation and Cell Survival during the Unfolded Protein Response. *Mol Cell* **5**: 897–904.

Hashem Y, des Georges A, Dhote V, Langlois R, Liao HY, Grassucci RA, Pestova TV, Hellen CUT, Frank J. 2013. Hepatitis-C-virus-like internal ribosome entry sites displace eIF3 to gain access to the 40S subunit. *Nature* **503**: 539–543.

Hayek H, Eriani G, Allmang C. 2022. eIF3 Interacts with Selenoprotein mRNAs. *Biomolecules* **12**: 1268.

Hayek H, Gross L, Janvier A, Schaeffer L, Martin F, Eriani G, Allmang C. 2021. eIF3 interacts with histone H4 messenger RNA to regulate its translation. *J Biol Chem* **296**: 100578.

Hellen CUT. 2018. Translation Termination and Ribosome Recycling in Eukaryotes. *Cold Spring Harb Perspect Biol* **10**: a032656.

Henras AK, Plisson-Chastang C, O'Donohue M-F, Chakraborty A, Gleizes P-E. 2015. An overview of pre-ribosomal RNA processing in eukaryotes: Pre-ribosomal RNA processing in eukaryotes. *Wiley Interdiscip Rev RNA* **6**: 225–242.

Hentze MW, Caughman SW, Rouault TA, Barriocanal JG, Dancis A, Harford JB, Klausner RD. 1987. Identification of the Iron-Responsive Element for the Translational Regulation of Human Ferritin mRNA. *Science* **238**: 1570–1573.

Hernández G, Osnaya VG, Pérez-Martínez X. 2019. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem Sci* **44**: 1009–1021.

Hetz C, Zhang K, Kaufman RJ. 2020. Mechanisms, regulation and functions of the unfolded protein response. *Nat Rev Mol Cell Biol* **21**: 421–438.

Hinnebusch AG, Lorsch JR. 2012. The Mechanism of Eukaryotic Translation Initiation: New Insights and Challenges. *Cold Spring Harb Perspect Biol* **4**: a011544–a011544.

Hoffman EP. 2020. The discovery of dystrophin, the protein product of the Duchenne muscular dystrophy gene. *FEBS J* **287**: 3879–3887.

Hoffman MA, Palmenberg AC. 1995. Mutational analysis of the J-K stem-loop region of the encephalomyocarditis virus IRES. *J Virol* **69**: 4399–4406.

Holz MK, Ballif BA, Gygi SP, Blenis J. 2005. mTOR and S6K1 Mediate Assembly of the Translation Preinitiation Complex through Dynamic Protein Interchange and Ordered Phosphorylation Events. *Cell* **123**: 569–580.

Hug N, Longman D, Cáceres JF. 2016. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res* **44**: 1483–1495.

Hussain T, Llácer JL, Fernández IS, Munoz A, Martin-Marcos P, Savva CG, Lorsch JR, Hinnebusch AG, Ramakrishnan V. 2014. Structural Changes Enable Start Codon Recognition by the Eukaryotic Translation Initiation Complex. *Cell* **159**: 597–607.

Hussey KA, Hadyniak SE, Johnston RJ. 2022. Patterning and Development of Photoreceptors in the Human Retina. *Front Cell Dev Biol* **10**: 878350.

Igarashi J, Sasaki T, Kobayashi N, Yoshioka S, Matsushita M, Shimizu T. 2011. Autophosphorylation of heme-regulated eukaryotic initiation factor 2α kinase and the role of the modification in catalysis: Autophosphorylation of an HRI. *FEBS J* **278**: 918–928.

Imai S, Kumar P, Hellen CUT, D'Souza VM, Wagner G. 2016. An accurately preorganized IRES RNA structure enables eIF4G capture for initiation of viral translation. *Nat Struct Mol Biol* **23**: 859–864.

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.

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.

Iwakawa H, Tomari Y. 2022. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. *Mol Cell* **82**: 30–43.

Jaafar ZA, Kieft JS. 2019. Viral RNA structure-based strategies to manipulate translation. *Nat Rev Microbiol* **17**: 110–123.

Jaafar ZA, Oguro A, Nakamura Y, Kieft JS. 2016. Translation initiation by the hepatitis C virus IRES requires eIF1A and ribosomal complex remodeling. *eLife* **5**: e21198.

Jacks T, Power MD, Masiarz FR, Luciw PA, Barr PJ, Varmus HE. 1988. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* **331**: 280–283.

Jackson RJ, Hellen CUT, Pestova TV. 2010. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**: 113–127.

Jan E, Sarnow P. 2002. Factorless Ribosome Assembly on the Internal Ribosome Entry Site of Cricket Paralysis Virus. *J Mol Biol* **324**: 889–902.

Jang SK, Wimmer E. 1990. Cap-independent translation of encephalomyocarditis virus RNA: structural elements of the internal ribosomal entry site and involvement of a cellular 57-kD RNA-binding protein. *Genes Dev* **4**: 1560–1572.

Janzen DM, Frolova L, Geballe AP. 2002. Inhibition of Translation Termination Mediated by an Interaction of Eukaryotic Release Factor 1 with a Nascent Peptidyl-tRNA. *Mol Cell Biol* **22**: 8562–8570.

Juszkiewicz S, Chandrasekaran V, Lin Z, Kraatz S, Ramakrishnan V, Hegde RS. 2018. ZNF598 Is a Quality Control Sensor of Collided Ribosomes. *Mol Cell* **72**: 469-481.e7.

Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou M-N, Sonenberg N. 2005. Mammalian poly(A)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes Dev* **19**: 104–113.

Kaminski A, Pöyry TAA, Skene PJ, Jackson RJ. 2010. Mechanism of Initiation Site Selection Promoted by the Human Rhinovirus 2 Internal Ribosome Entry Site. *J Virol* **84**: 6578–6589.

Kapp LD, Lorsch JR. 2004. GTP-dependent Recognition of the Methionine Moiety on Initiator tRNA by Translation Factor eIF2. *J Mol Biol* **335**: 923–936.

Kashiwagi K, Yokoyama T, Nishimoto M, Takahashi M, Sakamoto A, Yonemochi M, Shirouzu M, Ito T. 2019. Structural basis for eIF2B inhibition in integrated stress response. 5.

Kessler SH, Sachs AB. 1998. RNA Recognition Motif 2 of Yeast Pab1p Is Required for Its Functional Interaction with Eukaryotic Translation Initiation Factor 4G. *Mol Cell Biol* **18**: 51–57.

Khong A, Bonderoff J, Spriggs R, Tammpere E, Kerr C, Jackson T, Willis A, Jan E. 2016. Temporal Regulation of Distinct Internal Ribosome Entry Sites of the Dicistroviridae Cricket Paralysis Virus. *Viruses* **8**: 25.

Kim JH, Park SM, Park JH, Keum SJ, Jang SK. 2011. eIF2A mediates translation of hepatitis C viral mRNA under stress conditions: eIF2A mediates eIF2-independent translation. *EMBO J* **30**: 2454–2464.

Klinge S, Woolford JL. 2019. Ribosome assembly coming into focus. *Nat Rev Mol Cell Biol* **20**: 116–131.

Kolupaeva VG, Pestova TV, Hellen CUT, Shatsky IN. 1998. Translation Eukaryotic Initiation Factor 4G Recognizes a Specific Structural Element within the Internal Ribosome Entry Site of Encephalomyocarditis Virus RNA. *J Biol Chem* **273**: 18599–18604.

Komar AA, Merrick WC. 2020. A Retrospective on eIF2A—and Not the Alpha Subunit of eIF2. *Int J Mol Sci* **21**: 2054.

Kozak M. 1987. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J Mol Biol* **196**: 947–950.

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

Kozak M. 1990. Downstream secondary structure facilitates recognition of initiator codons by eukaryotic ribosomes. *Proc Natl Acad Sci* **87**: 8301–8305.

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

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

Kozutsumi Y, Segal M, Normington K, Gething M-J, Sambrook J. 1988. The presence of malfolded proteins in the endoplasmic reticulum signals the induction of glucose-regulated proteins. *Nature* **332**: 462–464.

Krishnamoorthy T, Pavitt GD, Zhang F, Dever TE, Hinnebusch AG. 2001. Tight Binding of the Phosphorylated α Subunit of Initiation Factor 2 (eIF2α) to the Regulatory Subunits of Guanine Nucleotide Exchange Factor eIF2B Is Required for Inhibition of Translation Initiation. *Mol Cell Biol* **21**: 5018–5030.

Kwan T, Thompson SR. 2019. Noncanonical Translation Initiation in Eukaryotes. *Cold Spring Harb Perspect Biol* **11**: a032672.

Lamper AM, Fleming RH, Ladd KM, Lee ASY. 2020. A phosphorylation-regulated eIF3d translation switch mediates cellular adaptation to metabolic stress. *Science* **370**: 853–856.

Lancaster AM, Jan E, Sarnow P. 2006. Initiation factor-independent translation mediated by the hepatitis C virus internal ribosome entry site. *RNA* **12**: 894–902.

Lapointe CP, Grosely R, Sokabe M, Alvarado C, Wang J, Montabana E, Villa N, Shin B-S, Dever TE, Fraser CS, et al. 2022. eIF5B and eIF1A reorient initiator tRNA to allow ribosomal subunit joining. *Nature* **607**: 185–190.

Lee ASY, Kranzusch PJ, Doudna JA, Cate JHD. 2016. eIF3d is an mRNA cap-binding protein that is required for specialized translation initiation. *Nature* **536**: 96–99.

Lee K-M, Chen C-J, Shih S-R. 2017. Regulation Mechanisms of Viral IRES-Driven Translation. *Trends Microbiol* **25**: 546–561.

Lee S-J, Depoortere I, Hatt H. 2019. Therapeutic potential of ectopic olfactory and taste receptors. *Nat Rev Drug Discov* **18**: 116–138.

Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, Fatica A, Santini T, Andronache A, Wade M, et al. 2017. Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Mol Cell* **66**: 22-37.e9.

Leppek K, Das R, Barna M. 2018. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol* **19**: 158–174.

Li G-W, Burkhardt D, Gross C, Weissman JS. 2014. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell* **157**: 624–635.

Lind C, Åqvist J. 2016. Principles of start codon recognition in eukaryotic translation initiation. *Nucleic Acids Res* **44**: 8425–8432.

Liu N, Dai Q, Zheng G, He C, Parisien M, Pan T. 2015. N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* **518**: 560–564.

Liu S, Suragani RNVS, Wang F, Han A, Zhao W, Andrews NC, Chen J-J. 2007. The function of heme-regulated eIF2α kinase in murine iron homeostasis and macrophage maturation. *J Clin Invest* **117**: 3296–3305.

Locker N, Easton LE, Lukavsky PJ. 2007. HCV and CSFV IRES domain II mediate eIF2 release during 80S ribosome assembly. *EMBO J* **26**: 795–805.

Lomakin IB, Steitz TA. 2013. The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* **500**: 307–311.

López de Quinto S, Martínez-Salas E. 1997. Conserved structural motifs located in distal loops of aphthovirus internal ribosome entry site domain 3 are required for internal initiation of translation. *J Virol* **71**: 4171–4175.

Loveland AB, Demo G, Grigorieff N, Korostelev AA. 2017. Ensemble cryo-EM elucidates the mechanism of translation fidelity. *Nature* **546**: 113–117.

Lozano G, Martínez-Salas E. 2015. Structural insights into viral IRES-dependent translation mechanisms. *Curr Opin Virol* **12**: 113–120.

Luke GA, de Felipe P, Lukashev A, Kallioinen SE, Bruno EA, Ryan MD. 2008. Occurrence, function and evolutionary origins of '2A-like' sequences in virus genomes. *J Gen Virol* **89**: 1036–1042.

Lykke-Andersen S, Jensen TH. 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nat Rev Mol Cell Biol* **16**: 665–677.

Lyons SM, Cunningham CH, Welch JD, Groh B, Guo AY, Wei B, Whitfield ML, Xiong Y, Marzluff WF. 2016. A subset of replication-dependent histone mRNAs are expressed as polyadenylated RNAs in terminally differentiated tissues. *Nucleic Acids Res* gkw620.

Ma XM, Blenis J. 2009. Molecular mechanisms of mTOR-mediated translational control. *Nat Rev Mol Cell Biol* **10**: 307–318.

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

Mailliot J, Martin F. 2018. Viral internal ribosomal entry sites: four classes for one goal: Viral internal ribosomal entry sites. *Wiley Interdiscip Rev RNA* **9**: e1458.

Manousis T, Moore NF. 1987. Cricket Paralysis Virus, a Potential Control Agent for the Olive Fruit Fly, *Dacus oleae* Gmel. *Appl Environ Microbiol* **53**: 142–148.

Manzella JM, Blackshear PJ. 1990. Regulation of rat ornithine decarboxylase mRNA translation by its 5'-untranslated region. *J Biol Chem* **265**: 11817–11822.

Martin F, Barends S, Jaeger S, Schaeffer L, Prongidi-Fix L, Eriani G. 2011. Cap-Assisted Internal Initiation of Translation of Histone H4. *Mol Cell* **41**: 197–209.

Martin F, Ménétret J-F, Simonetti A, Myasnikov AG, Vicens Q, Prongidi-Fix L, Natchiar SK, Klaholz BP, Eriani G. 2016. Ribosomal 18S rRNA base pairs with mRNA during eukaryotic translation initiation. *Nat Commun* **7**: 12622.

Martin KC, Ephrussi A. 2009. mRNA Localization: Gene Expression in the Spatial Dimension. *Cell* **136**: 719–730.

Martin SA, Moss B. 1975. Modification of RNA by mRNA guanylyltransferase and mRNA (guanine-7-)methyltransferase from vaccinia virions. *J Biol Chem* **250**: 9330–9335.

Mateju D, Eichenberger B, Voigt F, Eglinger J, Roth G, Chao JA. 2020. Single-Molecule Imaging Reveals Translation of mRNAs Localized to Stress Granules. *Cell* **183**: 1801-1812.e13.

Mathew SF, Crowe-McAuliffe C, Graves R, Cardno TS, McKinney C, Poole ES, Tate WP. 2015. The Highly Conserved Codon following the Slippery Sequence Supports −1 Frameshift Efficiency at the HIV-1 Frameshift Site ed. H.-J. Wieden. *PLOS ONE* **10**: e0122176.

Mayo CB, Cole JL. 2017. Interaction of PKR with single-stranded RNA. *Sci Rep* **7**: 3335.

McCaughan KK, Brown CM, Dalphin ME, Berry MJ, Tate WP. 1995. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc Natl Acad Sci* **92**: 5431–5435.

McKenna MJ, Bangsbo J, Renaud J-M. 2008. Muscle K^+, Na^+, and Cl^- disturbances and Na^+-K^+ pump inactivation: implications for fatigue. *J Appl Physiol* **104**: 288–295.

Melnikov S, Mailliot J, Shin B-S, Rigger L, Yusupova G, Micura R, Dever TE, Yusupov M. 2016. Crystal Structure of Hypusine-Containing Translation Factor eIF5A Bound to a Rotated Eukaryotic Ribosome. *J Mol Biol* **428**: 3570–3576.

Merrick WC, Pavitt GD. 2018. Protein Synthesis Initiation in Eukaryotic Cells. *Cold Spring Harb Perspect Biol* **10**: a033092.

Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, Pestova TV, Qian S-B, Jaffrey SR. 2015. 5′ UTR m6A Promotes Cap-Independent Translation. *Cell* **163**: 999–1010.

Mitchell SA, Spriggs KA, Coldwell MJ, Jackson RJ, Willis AE. 2003. The Apaf-1 Internal Ribosome Entry Segment Attains the Correct Structural Conformation for Function via Interactions with PTB and unr. *Mol Cell* **11**: 757–771.

Mouzakis KD, Lang AL, Vander Meulen KA, Easterday PD, Butcher SE. 2013. HIV-1 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA entrance channel of the ribosome. *Nucleic Acids Res* **41**: 1901–1913.

Muckenthaler M, Gray NK, Hentze MW. 1998. IRP-1 Binding to Ferritin mRNA Prevents the Recruitment of the Small Ribosomal Subunit by the Cap-Binding Complex eIF4F. *Mol Cell* **2**: 383–388.

Murray J, Savva CG, Shin B-S, Dever TE, Ramakrishnan V, Fernández IS. 2016. Structural characterization of ribosome recruitment and translocation by type IV IRES. *eLife* **5**: e13567.

Nag N, Lin KY, Edmonds KA, Yu J, Nadkarni D, Marintcheva B, Marintchev A. 2016. eIF1A/eIF5B interaction network and its functions in translation initiation complex assembly and remodeling. *Nucleic Acids Res* gkw552.

Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K -i. 2007. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* **36**: 861–871.

Narasimhan J, Staschke KA, Wek RC. 2004. Dimerization Is Required for Activation of eIF2 Kinase Gcn2 in Response to Diverse Environmental Stress Conditions. *J Biol Chem* **279**: 22820–22832.

Natchiar SK, Myasnikov AG, Kratzat H, Hazemann I, Klaholz BP. 2017. Visualization of chemical modifications in the human 80S ribosome structure. *Nature* **551**: 472–477.

Nayak A, Berry B, Tassetto M, Kunitomi M, Acevedo A, Deng C, Krutchinsky A, Gross J, Antoniewski C, Andino R. 2010. Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila. *Nat Struct Mol Biol* **17**: 547–554.

Nedialkova DD, Leidel SA. 2015. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell* **161**: 1606–1618.

Neidermyer WJ, Whelan SPJ. 2019. Global analysis of polysome-associated mRNA in vesicular stomatitis virus infected cells ed. P. Sarnow. *PLOS Pathog* **15**: e1007875.

Nguyen PT, Deisl C, Fine M, Tippetts TS, Uchikawa E, Bai X, Levine B. 2022. Structural basis for gating mechanism of the human sodium-potassium pump. *Nat Commun* **13**: 5293.

Noderer WL, Flockhart RJ, Bhaduri A, Diaz de Arce AJ, Zhang J, Khavari PA, Wang CL. 2014. Quantitative analysis of mammalian translation initiation sites by FACS -seq. *Mol Syst Biol* **10**: 748.

Ogle JM, Brodersen DE, Clemons WM, Tarry MJ, Carter AP, Ramakrishnan V. 2001. Recognition of Cognate Transfer RNA by the 30 *S* Ribosomal Subunit. *Science* **292**: 897–902.

Oh WJ, Wu C -chih, Kim SJ, Facchinetti V, Julien L-A, Finlan M, Roux PP, Su B, Jacinto E. 2010. mTORC2 can associate with ribosomes to promote cotranslational phosphorylation and stability of nascent Akt polypeptide. *EMBO J* **29**: 3939–3951.

Orlova M, Yueh A, Leung J, Goff SP. 2003. Reverse Transcriptase of Moloney Murine Leukemia Virus Binds to Eukaryotic Release Factor 1 to Modulate Suppression of Translational Termination. *Cell* **115**: 319–331.

Otto GA, Puglisi JD. 2004. The Pathway of HCV IRES-Mediated Translation Initiation. *Cell* **119**: 369–380.

Park E-H, Walker SE, Lee JM, Rothenburg S, Lorsch JR, Hinnebusch AG. 2011. Multiple elements in the eIF4G1 N-terminus promote assembly of eIF4G1•PABP mRNPs *in vivo*: Functionally redundant elements in the eIF4G1 N-terminus. *EMBO J* **30**: 302–316.

Patil CK, Li H, Walter P. 2004. Gcn4p and Novel Upstream Activating Sequences Regulate Targets of the Unfolded Protein Response ed. Steven McKnight. *PLoS Biol* **2**: e246.

Paulin FEM, Campbell LE, O'Brien K, Loughlin J, Proud CG. 2001. Eukaryotic translation initiation factor 5 (eIF5) acts as a classical GTPase-activator protein. *Curr Biol* **11**: 55–59.

Pechmann S, Chartron JW, Frydman J. 2014. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat Struct Mol Biol* **21**: 1100–1105.

Pelham HRB, Jackson RJ. 1976. An Efficient mRNA-Dependent Translation System from Reticulocyte Lysates. *Eur J Biochem* **67**: 247–256.

Pernod K, Schaeffer L, Chicher J, Hok E, Rick C, Geslain R, Eriani G, Westhof E, Ryckelynck M, Martin F. 2020. The nature of the purine at position 34 in tRNAs of 4-codon boxes is correlated with nucleotides at positions 32 and 38 to maintain decoding fidelity. *Nucleic Acids Res* **48**: 6170–6183.

Pestova TV, Borukhov SI, Hellen CUT. 1998. Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. *Nature* **394**: 854–859.

Pestova TV, de Breyne S, Pisarev AV, Abaeva IS, Hellen CUT. 2008. eIF2-dependent and eIF2-independent modes of initiation on the CSFV IRES: a common role of domain II. *EMBO J* **27**: 1060–1072.

Pestova TV, Hellen CUT. 2003. Translation elongation after assembly of ribosomes on the Cricket paralysis virus internal ribosomal entry site without initiation factors or initiator tRNA. *Genes Dev* **17**: 181–186.

Pestova TV, Lomakin IB, Lee JH, Choi SK, Dever TE, Hellen CUT. 2000. The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature* **403**: 332–335.

Petrov A, Grosely R, Chen J, O'Leary SE, Puglisi JD. 2016. Multiple Parallel Pathways of Translation Initiation on the CrPV IRES. *Mol Cell* **62**: 92–103.

Pilipenko EV, Blinov VM, Romanova LI, Sinyakov AN, Maslova SV, Agol VI. 1989. Conserved structural domains in the 5′-untranslated region of picornaviral genomes: An analysis of the segment controlling translation and neurovirulence. *Virology* **168**: 201–209.

Pisarev AV, Kolupaeva VG, Pisareva VP, Merrick WC, Hellen CUT, Pestova TV. 2006. Specific functional interactions of nucleotides at key $^{-}$3 and $^{+}$4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Genes Dev* **20**: 624–636.

Pisarev AV, Skabkin MA, Pisareva VP, Skabkina OV, Rakotondrafara AM, Hentze MW, Hellen CUT, Pestova TV. 2010. The Role of ABCE1 in Eukaryotic Posttermination Ribosomal Recycling. *Mol Cell* **37**: 196–210.

Plant EP, Dinman JD. 2006. Comparative study of the effects of heptameric slippery site composition on −1 frameshifting among different eukaryotic systems. *RNA* **12**: 666–673.

Pochopien AA, Beckert B, Kasvandik S, Berninghausen O, Beckmann R, Tenson T, Wilson DN. 2021. Structure of Gcn1 bound to stalled and colliding 80S ribosomes. *Proc Natl Acad Sci* **118**: e2022756118.

Pöyry TAA, 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.

Prats A-C, David F, Diallo LH, Roussel E, Tatin F, Garmy-Susini B, Lacazette E. 2020. Circular RNA, the Key for Translation. *Int J Mol Sci* **21**: 8591.

Querido JB, Sokabe M, Kraatz S, Gordiyenko Y, Skehel JM, Fraser CS, Ramakrishnan V. 2020. Structure of a human 48S translational initiation complex. 9.

Ramanathan A, Robb GB, Chan S-H. 2016. mRNA capping: biological functions and applications. *Nucleic Acids Res* **44**: 7511–7526.

Ratje AH, Loerke J, Mikolajka A, Brünner M, Hildebrand PW, Starosta AL, Dönhöfer A, Connell SR, Fucini P, Mielke T, et al. 2010. Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. *Nature* **468**: 713–716.

Reid DW, Nicchitta CV. 2015. Diversity and selectivity in mRNA translation on the endoplasmic reticulum. *Nat Rev Mol Cell Biol* **16**: 221–231.

Ren Q, Wang QS, Firth AE, Chan MMY, Gouw JW, Guarna MM, Foster LJ, Atkins JF, Jan E. 2012. Alternative reading frame selection mediated by a tRNA-like domain of an internal ribosome entry site. *Proc Natl Acad Sci* **109**. https://pnas.org/doi/full/10.1073/pnas.1111303109 (Accessed November 25, 2022).

Rivera VM, Welsh JD, Maizel JV. 1988. Comparative sequence analysis of the 5′ noncoding region of the enteroviruses and rhinoviruses. *Virology* **165**: 42–50.

Robertson MEM, Seamons RA, Belsham GJ. 1999. A selection system for functional internal ribosome entry site (IRES) elements: Analysis of the requirement for a conserved GNRA tetraloop in the encephalomyocarditis virus IRES. *RNA* **5**: 1167–1179.

Rogers GW, Richter NJ, Lima WF, Merrick WC. 2001. Modulation of the Helicase Activity of eIF4A by eIF4B, eIF4H, and eIF4F. *J Biol Chem* **276**: 30914–30922.

Roper SD, Chaudhari N. 2017. Taste buds: cells, signals and synapses. *Nat Rev Neurosci* **18**: 485–497.

Ross JA, Dungen KV, Bressler KR, Fredriksen M, Khandige Sharma D, Balasingam N, Thakor N. 2019. Eukaryotic initiation factor 5B (eIF5B) provides a critical cell survival switch to glioblastoma cells via regulation of apoptosis. *Cell Death Dis* **10**: 57.

Rubio Gomez MA, Ibba M. 2020. Aminoacyl-tRNA synthetases. *RNA* **26**: 910–936.

Russell JB, Cook GM. 1995. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol Rev* **59**: 48–62.

Saxton RA, Sabatini DM. 2017. mTOR Signaling in Growth, Metabolism, and Disease. *Cell* **168**: 960–976.

Schaffrath R, Abdel-Fattah W, Klassen R, Stark MJR. 2014. The diphthamide modification pathway from *S accharomyces cerevisiae* – revisited. *Mol Microbiol* **94**: 1213–1226.

Schaffrath R, Leidel SA. 2017. Wobble uridine modifications–a reason to live, a reason to die?! *RNA Biol* **14**: 1209–1222.

Scheuner D, Song B, McEwen E, Liu C, Laybutt R, Gillespie P, Saunders T, Bonner-Weir S, Kaufman RJ. 2001. Translational Control Is Required for the Unfolded Protein Response and In Vivo Glucose Homeostasis. *Mol Cell* **7**: 1165–1176.

Schimmel P. 2018. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nat Rev Mol Cell Biol* **19**: 45–58.

Schlesinger PH, Blair HC, Beer Stolz D, Riazanski V, Ray EC, Tourkova IL, Nelson DJ. 2020. Cellular and extracellular matrix of bone, with principles of synthesis and dependency of mineral deposition on cell membrane transport. *Am J Physiol-Cell Physiol* **318**: C111–C124.

Schmeing TM, Voorhees RM, Kelley AC, Gao Y-G, Murphy FV, Weir JR, Ramakrishnan V. 2009. The Crystal Structure of the Ribosome Bound to EF-Tu and Aminoacyl-tRNA. *Science* **326**: 688–694.

Schmidt C, Becker T, Heuer A, Braunger K, Shanmuganathan V, Pech M, Berninghausen O, Wilson DN, Beckmann R. 2016. Structure of the hypusinylated eukaryotic translation factor eIF-5A bound to the ribosome. *Nucleic Acids Res* **44**: 1944–1951.

Schüler M, Connell SR, Lescoute A, Giesebrecht J, Dabrowski M, Schroeer B, Mielke T, Penczek PA, Westhof E, Spahn CMT. 2006. Structure of the ribosome-bound cricket paralysis virus IRES RNA. *Nat Struct Mol Biol* **13**: 1092–1096.

Schuller AP, Wu CC-C, Dever TE, Buskirk AR, Green R. 2017. eIF5A Functions Globally in Translation Elongation and Termination. *Mol Cell* **66**: 194-205.e5.

Schütz P, Bumann M, Oberholzer AE, Bieniossek C, Trachsel H, Altmann M, Baumann U. 2008. Crystal structure of the yeast eIF4A-eIF4G complex: An RNA-helicase controlled by protein–protein interactions. *Proc Natl Acad Sci* **105**: 9564–9569.

Sejersted OM, Sjøgaard G. 2000. Dynamics and Consequences of Potassium Shifts in Skeletal Muscle and Heart During Exercise. *Physiol Rev* **80**: 1411–1481.

Sen ND, Gupta N, K. Archer S, Preiss T, Lorsch JR, Hinnebusch AG. 2019. Functional interplay between DEAD-box RNA helicases Ded1 and Dbp1 in preinitiation complex attachment and scanning on structured mRNAs in vivo. *Nucleic Acids Res* gkz595.

Sen ND, Zhou F, Ingolia NT, Hinnebusch AG. 2015. Genome-wide analysis of translational efficiency reveals distinct but overlapping functions of yeast DEAD-box RNA helicases Ded1 and eIF4A. *Genome Res* **25**: 1196–1205.

Shao S, Murray J, Brown A, Taunton J, Ramakrishnan V, Hegde RS. 2016. Decoding Mammalian Ribosome-mRNA States by Translational GTPase Complexes. *Cell* **167**: 1229-1240.e15.

Sharma P, Yan F, Doronina VA, Escuin-Ordinas H, Ryan MD, Brown JD. 2012. 2A peptides provide distinct solutions to driving stop-carry on translational recoding. *Nucleic Acids Res* **40**: 3143–3151.

Shi Y. 2017. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nat Rev Mol Cell Biol* **18**: 655–670.

Shi Z, Fujii K, Kovary KM, Genuth NR, Röst HL, Teruel MN, Barna M. 2017. Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. *Mol Cell* **67**: 71-83.e7.

Shuman S, Hurwitz J. 1981. Mechanism of mRNA capping by vaccinia virus guanylyltransferase: characterization of an enzyme--guanylate intermediate. *Proc Natl Acad Sci* **78**: 187–191.

Shuvalov A, Shuvalova E, Biziaev N, Sokolova E, Evmenov K, Pustogarov N, Arnautova A, Matrosova V, Egorova T, Alkalaeva E. 2021. Nsp1 of SARS-CoV-2 stimulates host translation termination. *RNA Biol* **18**: 804–817.

Sidrauski C, Cox JS, Walter P. 1996. tRNA Ligase Is Required for Regulated mRNA Splicing in the Unfolded Protein Response. *Cell* **87**: 405–413.

Sidrauski C, Tsai JC, Kampmann M, Hearn BR, Vedantham P, Jaishankar P, Sokabe M, Mendez AS, Newton BW, Tang EL, et al. 2015. Pharmacological dimerization and activation of the exchange factor eIF2B antagonizes the integrated stress response. *eLife* **4**: e07314.

Sidrauski C, Walter P. 1997. The Transmembrane Kinase Ire1p Is a Site-Specific Endonuclease That Initiates mRNA Splicing in the Unfolded Protein Response. *Cell* **90**: 1031–1039.

Simonetti A, Guca E, Bochler A, Kuhn L, Hashem Y. 2020. Structural Insights into the Mammalian Late-Stage Initiation Complexes. *Cell Rep* **31**: 107497.

Singh CR, Watanabe R, Zhou D, Jennings MD, Fukao A, Lee B, Ikeda Y, Chiorini JA, Campbell SG, Ashe MP, et al. 2011. Mechanisms of translational regulation by a human eIF5-mimic protein. *Nucleic Acids Res* **39**: 8314–8328.

Skinner MA, Racaniello VR, Dunn G, Cooper J, Minor PD, Almond JW. 1989. New model for the secondary structure of the 5′ non-coding RNA of poliovirus is supported by biochemical and genetic data that also show that RNA secondary structure is important in neurovirulence. *J Mol Biol* **207**: 379–392.

Skou JChr. 1957. The influence of some cations on an adenosine triphosphatase from peripheral nerves. *Biochim Biophys Acta* **23**: 394–401.

Smith RCL, Kanellos G, Vlahov N, Alexandrou C, Willis AE, Knight JRP, Sansom OJ. 2021. Translation initiation in cancer at a glance. *J Cell Sci* **134**: jcs248476.

Sokabe M, Fraser CS, Hershey JWB. 2012. The human translation initiation multi-factor complex promotes methionyl-tRNA i binding to the 40S ribosomal subunit. *Nucleic Acids Res* **40**: 905–913.

Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite MF, Hemmings BA, Barford D. 2000. The Crystal Structure of Human Eukaryotic Release Factor eRF1—Mechanism of Stop Codon Recognition and Peptidyl-tRNA Hydrolysis. *Cell* **100**: 311–321.

Soto-Rifo R, Limousin T, Rubilar PS, Ricci EP, Decimo D, Moncorge O, Trabaud M-A, Andre P, Cimarelli A, Ohlmann T. 2012. Different effects of the TAR structure on HIV-1 and HIV-2 genomic RNA translation. *Nucleic Acids Res* **40**: 2653–2667.

Spahn CM, Gomez-Lorenzo MG, Grassucci RA, Jørgensen R, Andersen GR, Beckmann R, Penczek PA, Ballesta JP, Frank J. 2004. Domain movements of elongation factor eEF2 and the eukaryotic 80S ribosome facilitate tRNA translocation. *EMBO J* **23**: 1008–1019.

Spangle JM, Münger K. 2010. The Human Papillomavirus Type 16 E6 Oncoprotein Activates mTORC1 Signaling and Increases Protein Synthesis. *J Virol* **84**: 9398–9407.

Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung J-W, Kuchelmeister HY, Batista PJ, Torre EA, Kool ET, et al. 2015. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**: 486–490.

Starosta AL, Lassak J, Peil L, Atkinson GC, Virumäe K, Tenson T, Remme J, Jung K, Wilson DN. 2014. Translational stalling at polyproline stretches is modulated by the sequence context upstream of the stall site. *Nucleic Acids Res* **42**: 10711–10719.

Stewart M. 2019. Polyadenylation and nuclear export of mRNAs. *J Biol Chem* **294**: 2977–2987.

Sudhakar A, Ramachandran A, Ghosh S, Hasnain SE, Kaufman RJ, Ramaiah KVA. 2000. Phosphorylation of Serine 51 in Initiation Factor 2α (eIF2α) Promotes Complex Formation between eIF2α(P) and eIF2B and Causes Inhibition in the Guanine Nucleotide Exchange Activity of eIF2B †. *Biochemistry* **39**: 12929–12938.

Sun C, Querol-Audí J, Mortimer SA, Arias-Palomo E, Doudna JA, Nogales E, Cate JHD. 2013. Two RNA-binding motifs in eIF3 direct HCV IRES-dependent translation. *Nucleic Acids Res* **41**: 7512–7521.

Sun Y, Atas E, Lindqvist L, Sonenberg N, Pelletier J, Meller A. 2012. The eukaryotic initiation factor eIF4H facilitates loop-binding, repetitive RNA unwinding by the eIF4A DEAD-box helicase. *Nucleic Acids Res* **40**: 6199–6207.

Tabet R, Schaeffer L, Freyermuth F, Jambeau M, Workman M, Lee C-Z, Lin C-C, Jiang J, Jansen-West K, Abou-Hamdan H, et al. 2018. CUG initiation and frameshifting enable production of dipeptide repeat proteins from ALS/FTD C9ORF72 transcripts. *Nat Commun* **9**: 152.

Tang L, Morris J, Wan J, Moore C, Fujita Y, Gillaspie S, Aube E, Nanda J, Marques M, Jangal M, et al. 2017. Competition between translation initiation factor eIF5 and its mimic

protein 5MP determines non-AUG initiation rate genome-wide. *Nucleic Acids Res* **45**: 11941–11953.

Taylor DJ, Nilsson J, Merrill AR, Andersen GR, Nissen P, Frank J. 2007. Structures of modified eEF2·80S ribosome complexes reveal the role of GTP hydrolysis in translocation. *EMBO J* **26**: 2421–2431.

Terenin IM, Dmitriev SE, Andreev DE, Shatsky IN. 2008. Eukaryotic translation initiation machinery can operate in a bacterial-like mode without eIF2. *Nat Struct Mol Biol* **15**: 836–841.

Thakor N, Holcik M. 2012. IRES-mediated translation of cellular messenger RNA operates in eIF2α- independent manner during stress. *Nucleic Acids Res* **40**: 541–552.

Thakur A, Gaikwad S, Vijjamarri AK, Hinnebusch AG. 2020. eIF2α interactions with mRNA control accurate start codon selection by the translation preinitiation complex. *Nucleic Acids Res* **48**: 10280–10296.

Thakur A, Hinnebusch AG. 2018. eIF1 Loop 2 interactions with Met-tRNA $_i$ control the accuracy of start codon selection by the scanning preinitiation complex. *Proc Natl Acad Sci* **115**: E4159–E4168.

Tian B. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33**: 201–212.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* **141**: 344–354.

Uppala JK, Sathe L, Chakraborty A, Bhattacharjee S, Pulvino AT, Dey M. 2022. The cap-proximal RNA secondary structure inhibits preinitiation complex formation on HAC1 mRNA. *J Biol Chem* **298**: 101648.

Valášek LS, Zeman J, Wagner S, Beznosková P, Pavlíková Z, Mohammad MP, Hronová V, Herrmannová A, Hashem Y, Gunišová S. 2017. Embraced by eIF3: structural and functional insights into the roles of eIF3 across the translation cycle. *Nucleic Acids Res* **45**: 10948–10968.

Vasudevan D, Clark NK, Sam J, Cotham VC, Ueberheide B, Marr MT, Ryoo HD. 2017. The GCN2-ATF4 Signaling Pathway Induces 4E-BP to Bias Translation and Boost Antimicrobial Peptide Synthesis in Response to Bacterial Infection. *Cell Rep* **21**: 2039–2047.

Vasudevan D, Neuman SD, Yang A, Lough L, Brown B, Bashirullah A, Cardozo T, Ryoo HD. 2020. Translational induction of ATF4 during integrated stress response requires noncanonical initiation factors eIF2D and DENR. *Nat Commun* **11**: 4677.

Vattem KM, Wek RC. 2004. Reinitiation involving upstream ORFs regulates *ATF4* mRNA translation in mammalian cells. *Proc Natl Acad Sci* **101**: 11269–11274.

Ventoso I, Blanco R, Perales C, Carrasco L. 2001. HIV-1 protease cleaves eukaryotic initiation factor 4G and inhibits cap-dependent translation. *Proc Natl Acad Sci* **98**: 12966–12971.

Vicens Q, Kieft JS, Rissland OS. 2018. Revisiting the Closed-Loop Model and the Nature of mRNA 5′–3′ Communication. *Mol Cell* **72**: 805–812.

Villa N, Do A, Hershey JWB, Fraser CS. 2013. Human Eukaryotic Initiation Factor 4G (eIF4G) Protein Binds to eIF3c, -d, and -e to Promote mRNA Recruitment to the Ribosome. *J Biol Chem* **288**: 32932–32940.

Walsh D, Mohr I. 2011. Viral subversion of the host protein synthesis machinery. *Nat Rev Microbiol* **9**: 860–875.

Walter P, Blobel G. 1981. Translocation of proteins across the endoplasmic reticulum. II. Signal recognition protein (SRP) mediates the selective binding to microsomal membranes of in-vitro-assembled polysomes synthesizing secretory protein. *J Cell Biol* **91**: 551–556.

Walter P, Ibrahimi I, Blobel G. 1981. Translocation of proteins across the endoplasmic reticulum. I. Signal recognition protein (SRP) binds to in-vitro-assembled polysomes synthesizing secretory protein. *J Cell Biol* **91**: 545–550.

Wang J, Shin B-S, Alvarado C, Kim J-R, Bohlen J, Dever TE, Puglisi JD. 2022. Rapid 40S scanning and its regulation by mRNA structure during eukaryotic translation initiation. *Cell* **185**: 4474-4487.e17.

Wang X. 2001. Regulation of elongation factor 2 kinase by p90RSK1 and p70 S6 kinase. *EMBO J* **20**: 4370–4379.

Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. 2014. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**: 117–120.

Weingarten-Gabbay S, Elias-Kirma S, Nir R, Gritsenko AA, Stern-Ginossar N, Yakhini Z, Weinberger A, Segal E. 2016. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science* **351**: aad4939.

Wilson JE, Pestova TV, Hellen CUT, Sarnow P. 2000. Initiation of Protein Synthesis from the A Site of the Ribosome. *Cell* **102**: 511–520.

Wu CC-C, Peterson A, Zinshteyn B, Regot S, Green R. 2020a. Ribosome Collisions Trigger General Stress Responses to Regulate Cell Fate. *Cell* **182**: 404-416.e14.

Wu Q, Wright M, Gogol MM, Bradford WD, Zhang N, Bazzini AA. 2020b. Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J* **39**. https://onlinelibrary.wiley.com/doi/10.15252/embj.2020104763 (Accessed November 28, 2022).

Wurth L, Gribling-Burrer A-S, Verheggen C, Leichter M, Takeuchi A, Baudrey S, Martin F, Krol A, Bertrand E, Allmang C. 2014. Hypermethylated-capped selenoprotein mRNAs in mammals. *Nucleic Acids Res* **42**: 8663–8677.

Xi Q, Cuesta R, Schneider RJ. 2004. Tethering of eIF4G to adenoviral mRNAs by viral 100k protein drives ribosome shunting. *Genes Dev* **18**: 1997–2009.

Xu R, Lou Y, Tidu A, Bulet P, Heinekamp T, Martin F, Brakhage A, Li Z, Liégeois S, Ferrandon D. 2022. The Toll pathway mediates *Drosophila* resilience to *Aspergillus* mycotoxins through specific Bomanins. *EMBO Rep*. https://onlinelibrary.wiley.com/doi/10.15252/embr.202256036 (Accessed November 23, 2022).

Yamamoto H, Nakashima N, Ikeda Y, Uchiumi T. 2007. Binding Mode of the First Aminoacyl-tRNA in Translation Initiation Mediated by Plautia stali Intestine Virus Internal Ribosome Entry Site. *J Biol Chem* **282**: 7770–7776.

Yan LL, Simms CL, McLoughlin F, Vierstra RD, Zaher HS. 2019. Oxidation and alkylation stresses activate ribosome-quality control. *Nat Commun* **10**: 5611.

Yang Y, Fan X, Mao M, Song X, Wu P, Zhang Y, Jin Y, Yang Y, Chen L-L, Wang Y, et al. 2017. Extensive translation of circular RNAs driven by N6-methyladenosine. *Cell Res* **27**: 626–641.

Ye J, Palm W, Peng M, King B, Lindsten T, Li MO, Koumenis C, Thompson CB. 2015. GCN2 sustains mTORC1 suppression upon amino acid deprivation by inducing Sestrin2. *Genes Dev* **29**: 2331–2336.

Yu Y, Alwine JC. 2006. 19S Late mRNAs of Simian Virus 40 Have an Internal Ribosome Entry Site Upstream of the Virion Structural Protein 3 Coding Sequence. *J Virol* **80**: 6553–6558.

Yu Y, Kudchodkar SB, Alwine JC. 2005. Effects of Simian Virus 40 Large and Small Tumor Antigens on Mammalian Target of Rapamycin Signaling: Small Tumor Antigen Mediates Hypophosphorylation of eIF4E-Binding Protein 1 Late in Infection. *J Virol* **79**: 6882–6889.

Yusupova G, Yusupov M. 2015. Ribosome biochemistry in crystal structure determination. *RNA* **21**: 771–773.

Zhang G, Wang X, Rothermel BA, Lavandero S, Wang ZV. 2022. The integrated stress response in ischemic diseases. *Cell Death Differ* **29**: 750–757.

Zhou F, Zhang H, Kulkarni SD, Lorsch JR, Hinnebusch AG. 2020. eIF1 discriminates against suboptimal initiation sites to prevent excessive uORF translation genome-wide. *RNA* **26**: 419–438.

Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian S-B. 2015. Dynamic m6A mRNA methylation directs translational control of heat shock response. *Nature* **526**: 591–594.

Zhou M, Sandercock AM, Fraser CS, Ridlova G, Stephens E, Schenauer MR, Yokoi-Fong T, Barsky D, Leary JA, Hershey JW, et al. 2008. Mass spectrometry reveals modularity and a complete subunit interaction map of the eukaryotic translation factor eIF3. *Proc Natl Acad Sci* **105**: 18139–18144.

Zinzalla V, Stracka D, Oppliger W, Hall MN. 2011. Activation of mTORC2 by Association with the Ribosome. *Cell* **144**: 757–768.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

Zyryanova AF, Weis F, Faille A, Alard AA, Crespillo-Casado A, Sekine Y, Harding HP, Allen F, Parts L, Fromont C, et al. 2018. Binding of ISRIB reveals a regulatory site in the nucleotide exchange factor eIF2B. *Science* **359**: 1533–1536.

Antonin TIDU

Etude d'éléments *cis*-régulateurs de la traduction eucaryote par des approches *in vitro* et par des méthodes de criblage à haut-débit

Le premier volet de ce travail est dédié à la recherche de séquences qui modulent la reconnaissance du codon initiateur par le ribosome. Pour ce faire, nous avons réalisé des criblages de banques d'ARN rapporteurs portant des contextes nucléotidiques randomisés autour du codon AUG initiateur, et les contextes des codons AUG initiateurs des phases codantes humaines ont été analysés en parallèle. Ces deux études suggèrent l'absence de motif minimal autour du codon AUG permettant d'expliquer sa reconnaissance par le ribosome. Par ailleurs, nous avons démontré que l'initiation de la traduction sur un codon non-AUG est dépendante de la présence d'une structure d'ARN suffisamment stable localisée à une distance optimale en aval du codon initiateur qui permet d'y stabiliser le ribosome. Le second volet est dédié à la mise au point d'une méthode de criblage permettant d'identifier des IRES (Internal Ribosome Entry Site) dans un génome viral. Cette méthode est basée sur le fractionnement de complexes de traduction sur gradient de saccharose. Une expérience de preuve de concept sur le génome du virus de la paralysie du cricket a montré que la méthode permet d'identifier les deux IRES qui s'y trouvent mais nécessite encore des améliorations avant d'être appliquée à la recherche d'IRES dans d'autres génomes viraux. Le troisième volet est dédié à l'étude de la traduction de l'ARN génomique du SARS-CoV-2 lors de l'infection virale. Nous avons démontré qu'en présence de la protéine virale NSP1 qui invalide les ribosomes cellulaires, la traduction des ARN du SARS-CoV-2 est strictement dépendante de la présence d'une structure en tige-boucle à son extrémité 5' qui délogerait NSP1 du ribosome, permettant ainsi la traduction de l'ARN viral. Les outils moléculaires et informatiques mis au point dans le cadre de ce travail peuvent être mis à profit dans d'autres projets en lien avec l'initiation de la traduction chez les eucaryotes.

Mots-clés : initiation de la traduction, ribosome, *cis*-régulation, séquence Kozak, non-AUG, virus, IRES

The first part of this work is dedicated to the search for sequences that modulate the recognition of the initiator codon by the ribosome. For that purpose, we have performed high-throughput screenings of reporter RNA libraries carrying randomized nucleotide contexts around the initiator AUG codon, and the contexts of the initiator AUG codons of human coding sequences have been analysed in parallel. Both studies suggest that there is no minimal motif around the AUG codon that can explain its recognition by the ribosome. Furthermore, we demonstrated that the initiation of translation on a non-AUG codon is dependent on the presence of a stable-enough downstream RNA structure located at an optimal distance from the initiator codon that stabilizes the scanning ribosome. The second part is dedicated to the development of a screening method for identifying IRES (Internal Ribosome Entry Site) in a viral genome. This method is based on the fractionation of translation complexes on a sucrose gradient. A proof-of-concept experiment on the genome of the cricket paralysis virus showed that the method can identify the two IRESs found there, but further improvements are required before it can be applied to the search for IRESs in other viral genomes. The third part is dedicated to the study of the translation of SARS-CoV-2 genomic RNA during viral infection. We demonstrated that in the presence of the viral protein NSP1, which invalidates cellular ribosomes, the translation of SARS-CoV-2 RNAs is strictly dependent on the presence of a stem-loop structure at its 5' end that would displace NSP1 from the ribosome, thus allowing the translation of the viral RNA. The molecular and computational tools developed in this work can be used in other projects related to translation initiation in eucaryotes.

Key words: translation initiation, ribosome, *cis*-regulation, Kozak sequence, non-AUG, virus, IRES

Antonin TIDU

Etude d'éléments *cis*-régulateurs de la traduction eucaryote par des approches *in vitro* et par des méthodes de criblage à haut-débit

Le premier volet de ce travail est dédié à la recherche de séquences qui modulent la reconnaissance du codon initiateur par le ribosome. Pour ce faire, nous avons réalisé des criblages de banques d'ARN rapporteurs portant des contextes nucléotidiques randomisés autour du codon AUG initiateur, et les contextes des codons AUG initiateurs des phases codantes humaines ont été analysés en parallèle. Ces deux études suggèrent l'absence de motif minimal autour du codon AUG permettant d'expliquer sa reconnaissance par le ribosome. Par ailleurs, nous avons démontré que l'initiation de la traduction sur un codon non-AUG est dépendante de la présence d'une structure d'ARN suffisamment stable localisée à une distance optimale en aval du codon initiateur qui permet d'y stabiliser le ribosome. Le second volet est dédié à la mise au point d'une méthode de criblage permettant d'identifier des IRES (Internal Ribosome Entry Site) dans un génome viral. Cette méthode est basée sur le fractionnement de complexes de traduction sur gradient de saccharose. Une expérience de preuve de concept sur le génome du virus de la paralysie du cricket a montré que la méthode permet d'identifier les deux IRES qui s'y trouvent mais nécessite encore des améliorations avant d'être appliquée à la recherche d'IRES dans d'autres génomes viraux. Le troisième volet est dédié à l'étude de la traduction de l'ARN génomique du SARS-CoV-2 lors de l'infection virale. Nous avons démontré qu'en présence de la protéine virale NSP1 qui invalide les ribosomes cellulaires, la traduction des ARN du SARS-CoV-2 est strictement dépendante de la présence d'une structure en tige-boucle à son extrémité 5' qui délogerait NSP1 du ribosome, permettant ainsi la traduction de l'ARN viral. Les outils moléculaires et informatiques mis au point dans le cadre de ce travail peuvent être mis à profit dans d'autres projets en lien avec l'initiation de la traduction chez les eucaryotes.

Mots-clés : initiation de la traduction, ribosome, *cis*-régulation, séquence Kozak, non-AUG, virus, IRES

The first part of this work is dedicated to the search for sequences that modulate the recognition of the initiator codon by the ribosome. For that purpose, we have performed high-throughput screenings of reporter RNA libraries carrying randomized nucleotide contexts around the initiator AUG codon, and the contexts of the initiator AUG codons of human coding sequences have been analysed in parallel. Both studies suggest that there is no minimal motif around the AUG codon that can explain its recognition by the ribosome. Furthermore, we demonstrated that the initiation of translation on a non-AUG codon is dependent on the presence of a stable-enough downstream RNA structure located at an optimal distance from the initiator codon that stabilizes the scanning ribosome. The second part is dedicated to the development of a screening method for identifying IRES (Internal Ribosome Entry Site) in a viral genome. This method is based on the fractionation of translation complexes on a sucrose gradient. A proof-of-concept experiment on the genome of the cricket paralysis virus showed that the method can identify the two IRESs found there, but further improvements are required before it can be applied to the search for IRESs in other viral genomes. The third part is dedicated to the study of the translation of SARS-CoV-2 genomic RNA during viral infection. We demonstrated that in the presence of the viral protein NSP1, which invalidates cellular ribosomes, the translation of SARS-CoV-2 RNAs is strictly dependent on the presence of a stem-loop structure at its 5' end that would displace NSP1 from the ribosome, thus allowing the translation of the viral RNA. The molecular and computational tools developed in this work can be used in other projects related to translation initiation in eucaryotes.

Key words: translation initiation, ribosome, *cis*-regulation, Kozak sequence, non-AUG, virus, IRES