Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

ÉCOLE DOCTORALE (414) DES SCIENCES DE LA VIE ET DE LA SANTÉ

UPR9002 (CNRS) – Architecture et Réactivité de l'arN (ARN)

THÈSE présentée par :

Natacha DENTZ

Soutenue le : **10 Septembre 2021**

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

Discipline/ Spécialité : Aspects moléculaires et cellulaires de la biologie

Caractérisation de la machinerie traductionnelle à l'aide de criblages ultra haut-débit en gouttelettes microfluidiques

THÈSE dirigée par :

Prof. RYCKELYNCK Michaël	Professeur des universités, IBMC, Université de Strasbourg (France)

RAPPORTEURS :

Dr. GUILLIER Maude	Maitre de Conférence, IBPC, Université de Paris Diderot (France)
Prof. GAUTIER Arnaud	Professeur des universités, UMR 7203, Université Paris 1 Sorbonne (France)

AUTRES MEMBRES DU JURY :

Prof. BECKER Hubert	Professeur des universités, IPCB, Université de Strasbourg (France)
Prof. FAULON Jean-Loup	Professeur des universités, UMR 1319 (France)

Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

ÉCOLE DOCTORALE (414) DES SCIENCES DE LA VIE ET DE LA SANTÉ

UPR9002 (CNRS) – Architecture et Réactivité de l'arN (ARN)

THÈSE présentée par :

Natacha DENTZ

Soutenue le : **10 Septembre 2021**

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

Discipline/ Spécialité : Sciences de la vie, Biologie, Biochimie

Caractérisation de la machinerie traductionnelle à l'aide de criblages ultra haut-débit en gouttelettes microfluidiques

THÈSE dirigée par :

Prof. RYCKELYNCK Michaël	Professeur des universités, IBMC, Université de Strasbourg (France)

RAPPORTEURS :

Dr. GUILLIER Maude	Maitre de Conférence, IBPC, Université de Paris Diderot (France)
Prof. GAUTIER Arnaud	Professeur des universités, UMR 7203, Université Paris 1 Sorbonne (France)

AUTRES MEMBRES DU JURY :

Prof. BECKER Hubert	Professeur des universités, IPCB, Université de Strasbourg (France)
Prof. FAULON Jean-Loup	Professeur des universités, UMR 1319 (France)

Remerciements

De manière assez surprenante et émouvante, il ne s'agit pas de la partie la plus simple à écrire, en espérant oublier personne et en terminant ce long moment de rédaction je vais remercier l'ensemble des personnes qui m'ont suivie et soutenue pendant ces trois années voire plus !

Je vais commencer par remercier les membres de mon jury : Pr. Arnaud Gautier, Dr. Maud Guillier, Pr. Jean-Loup Faulon et Pr. Hubert Becker qui ont accepté d'évaluer mon travail de thèse.
Bien évidemment, je remercie l'ensemble du personnel et des personnes de l'institut ainsi que mes collaborateurs Franck Martin et Antonin Tidu.

Maintenant, je poursuis avec les membres de mon équipe qui m'auront bravement soutenue au fil de ces trois ans et demi. Un grand merci à Michaël, qui m'a aidé, soutenue et poussée à ne jamais baisser les bras, Cédric qui, à peine arrivé, a participé à la correction, Stéphanie génialissime Maman de substitution au laboratoire, Roger (ma vieille loutre catalane) qui aura eu les bons mots pour me faire positiver et me re-booster lorsque je suis venu imposer mes trêves quotidiennes dans la caverne, Farah (Belzeb) qui m'a permis de relativiser sur ma thèse et m'a transformée en démon certifié, Clair(e)asite pour les cafés du matin, les séances psy simples et efficaces sans oublier ton apprivoisement (LOL), Tous mes petits stagiaires, avec Martin (PionPion / alien) qui était vraiment exceptionnel, Blandine et Alexandre. Et…Emilie (optimale et superbe cagole) qui est autant une collègue qu'une amie proche avec qui j'ai pu rigoler, me plaindre, ou encore me ressourcer et me faire nourrir lorsque j'étais seule ou en télétravail...SANS oublier Arya sa fidèle acolyte qui a toujours su me jeter des feuilles ou sa souris trempée pour me remonter le moral.

Je vais remercier maintenant ma famille (parents, frère, parrain, marraine, oncles et tantes !) mais plus particulièrement mes parents qui m'ont encouragée tout le long de mes études sans qui rien n'aurait été possible. Merci à ma maman pour le supply dans son magasin outlet personnel, les repas et les séances de piscine/psy, et à mon papa pour tous les défis sportifs qui m'ont permis de garder la forme comme le moral et l'aide en toutes circonstances (vélo et réponses aux appels, LUI).

Je remercie Baptiste, mon (futur) mari qui, à vrai dire, mérite bien son propre paragraphe ici. Je vais être simple et direct l'ensemble de ces dures années et surtout de cette rédaction m'aurait paru bien plus compliquée sans ton aide quotidienne et ton soutien permanent. Tu m'as aidée, motivée et recentrée sur ce qui était essentiel quand je m'éparpillais. Un grand merci à toi, tu mérites tout autant un titre honorifique pour tout ce que tu as fait. Je vais tout de même ajouter à ce petit paragraphe le deuxième assistant le plus important, je parle bien évidemment de Fluffy ! Cette adorable « petite » boule de poils de 30kg qui, par sa joie de vivre et son amour débordant, m'a obligée à relativiser lors des moments difficiles.

La fin de ces remerciements est dédiée à l'ensemble de mes ami(e)s qui sont à mes yeux une deuxième famille sans qui tout serait bien plus fade. Incluant l'ensemble des copains du Master (David, Robin, Claire, Kevin) toujours disponibles pour une pinte et une soirée. Il y a obligatoirement les amis à distance (Eliot & Anko) toujours présents aux temps forts. Ceux des snaps du matin (Yann, Arnaud et Clara) qui boostent quotidiennement, ceux qui sont des membres de la famille (Orianne (dont j'attends toujours les tractions), Maureen, Nenelle et aussi Bribri), et ceux qui prennent des nouvelles et sont toujours là (Pierro et Lola). Je remercie (xoxo) par de gros bisous plus particulièrement Marion, qui est un chat de compétition et même incroyable qui m'a apportée un soutien de dingue toutes ces soirées et nuits. Puis celui qui m'a et me fais toujours mourir de rire, Orian The virus, toujours sur la même longueur d'onde avec qui j'ai passé les meilleures pauses café à râler, rire qui vont beaucoup trop me manquer !

Je termine par un grand merci aux aventures de Harry Potter qui m'ont permis de m'échapper, rêvasser et même de glisser des petits mots aux connaisseurs…
Merci à la bande son d'Europa Park et à J-J G, qui a chanté mon quotidien et m'a boostée : «Bien qu'un grand nombre de matins me parût pour rien, j'ai 'saigné' sur mes Gilsons pour aller au bout de mes rêves.»

NOX.

SOMMAIRE

Introduction :

1. Expression des gènes

2. Régulation de l’expression des gènes

3. Expression des gènes en systèmes acellulaires

4. La microfluidique en gouttelettes et l’expression des gènes

Introduction

1. Expression des gènes

L'expression des gènes est décrite comme le cœur de la biologie moléculaire dès 1958 par Francis Crick, la proposant comme le dogme central de la discipline qu'il va lui-même renverser dans le but de l'ajuster et compléter dans les années 70 (Crick, 1970)

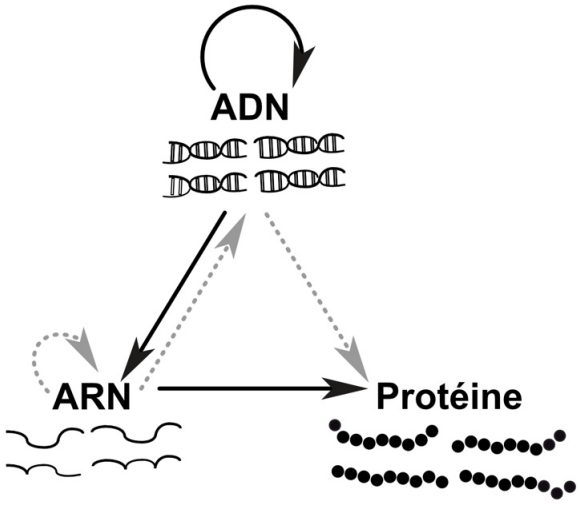


Figure 1 : Dogme de la biologie moléculaire selon Francis Crick dans les années 70. Les flèches en noires correspondent aux liens existant en 1958 (réplication, transcription et traduction), tandis que les flèches en pointillé correspondent aux transferts possibles en 1970. Représentation adaptée de (Crick, 1970).

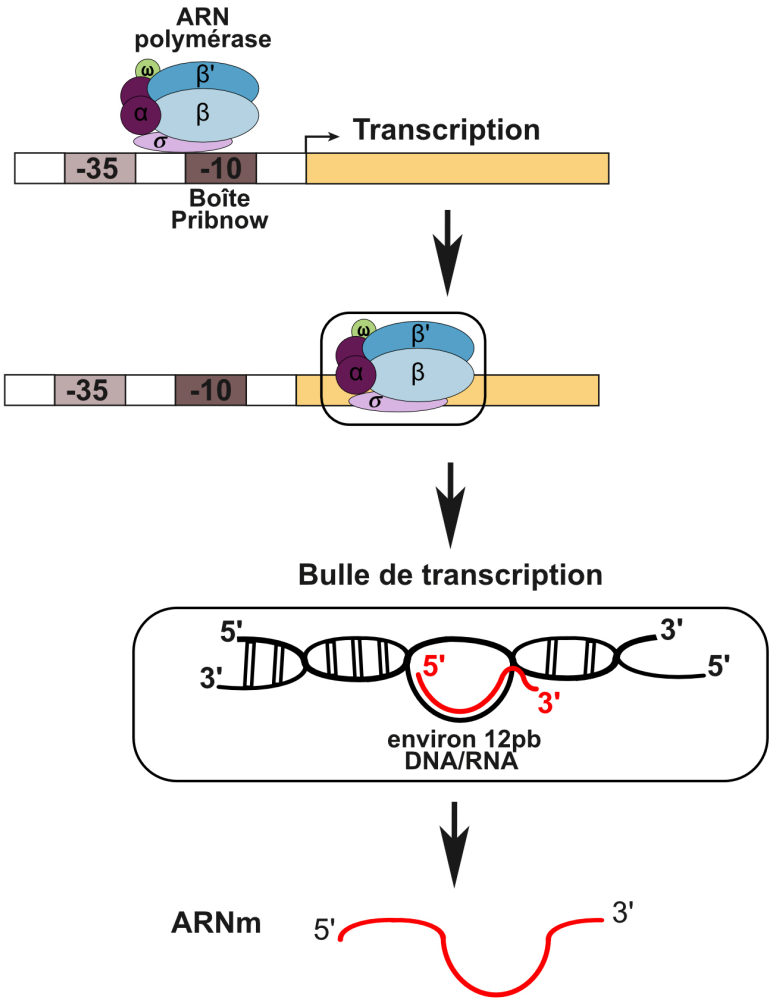
En bref, l'ADN est le support de l'information génétique contenu dans le noyau chez les eucaryotes et dans le nucléoïde chez les procaryotes. Cette molécule est exprimée au travers de deux grands mécanismes : la transcription et la traduction selon que les gènes soient codants ou non. En effet, une partie non négligeable des gènes transcrits est non codante, occupant fréquemment des fonctions indispensables au bon fonctionnement de la cellule comme par exemple, les ARN de transfert (ARNt), ribosomique (ARNr) ou régulateurs de manière général, pour les procaryotes comme pour les eucaryotes. Les gènes codants quant à eux vont être transcrits puis traduits en protéines par la machinerie traductionnelle. La transcription comme la traduction sont toutes deux des mécanismes complexes et finement régulés en fonction des besoins de la cellule. La mécanistique de chacun de ces processus varie d'ailleurs selon les deux grands types cellulaires que sont les procaryotes et les eucaryotes (Krebs *et al.*, 2018).

1.1 La transcription

Lors de la transcription de l'ARN est synthétisé par l'activité de l'ARN polymérase ADN dépendante à partir d'un ADN matrice. Ce processus a lieu en trois grandes étapes : l'initiation, l'élongation et la terminaison. Une fois l'ARN transcrit, il peut être modifié, tel que les ARN dit messager (ARNm) qui chez les eucaryotes sont ensuite coiffés et polyadénylés (Harvey Lodish *et al.*, 2000; Krebs *et al.*, 2018) ou simplement les ARNt (ex : pseudouridinilation, méthylation…) chez les eucaryotes et les procaryotes (Krebs *et al.*, 2018). Puis, les ARN dit non-codants vont assurer leurs fonctions et les ARNm seront dirigés vers les ribosomes pour être traduit en protéines.

1.1.1 La transcription procaryote

En moyenne, une cellule d'*E. coli* contient 13 000 ARN polymérases permettant de transcrire l'ensemble des 4 641 gènes codants ou non-codants du génome de cette bactérie (Blattner *et al.*, 1997). Si la taille des génomes, le nombre de gènes, ou la quantité exacte d'ARN polymérase varient selon les organismes procaryotes, l'ARN polymérase est, quant à elle, toujours composée des mêmes éléments. Le cœur enzymatique est constitué de 5 sous-unités : α_2 servant d'échafaudage à la polymérase et facilitant son assemblage, ω jouant de même un rôle dans l'assemblage et enfin, β et β' qui forment le cœur de l'activité enzymatique (Murakami, 2015) (Figure 2).



<u>Figure 2 :</u> Représentation généraliste de l'initiation de la transcription chez les procaryotes. Les différentes sous-unités formant l'ARN polymérase ($\alpha_2\beta\beta'\omega\sigma$), sont respectivement représentées en mauve, bleu clair, bleu foncé, vert, et violet clair. Le gène transcrit est quant à lui représenté en jaune tandis que dans la bulle de transcription le brin d'ADN est représenté en noir et celui de l'ARN néo transcrit en rouge.

Le terme d'holoenzyme désigne l'association du core enzymatique avec le facteur sigma, qui va lui conférer la capacité de reconnaitre un promoteur définit selon le facteur sigma associé (Fredrick and Helmann, 1997; Krebs *et al.*, 2018). Par exemple, le facteur sigma 70 va permettre la reconnaissance des gènes de ménage par la polymérase. A l'inverse le facteur sigma 32 va être sollicité lors d'un choc thermique pour la reconnaissance de promoteurs précis placés en amont de gènes permettant l'adaptation de la cellule à ce stress environnemental (David B. Strauss, William A. Walter and Carol A. Gross, 1987; Roncarati and Scarlato, 2017). La fréquence à laquelle la polymérase initie la transcription dépend du promoteur en amont du gène, allant d'une initiation par seconde pour des gènes très sollicités, tels que ceux codant pour les ARN ribosomiques, à une initiation toutes les 30 minutes pour des gènes moins sollicités (Krebs *et al.*, 2018). Lorsque l'initiation démarre et que la

polymérase entame le processus d'élongation, un complexe tertiaire est formé incluant, l'ADN, l'ARN et l'enzyme (Figure 2). Durant l'élongation, le cœur enzymatique se sépare du facteur sigma, afin de pouvoir se déplacer vers l'extrémité 5' du brin transcrit de l'ADN. Les ribonucléotides sont ajoutés à la suite et associés entre eux par une liaison phosphodiester formant ainsi un ARN néosynthétisé. La transcription s'achève par la terminaison qui est la dernière des trois étapes. Cette dernière peut avoir lieu selon deux mécanismes différents (Krebs *et al.*, 2018) :

- La terminaison Rho-indépendante
- La terminaison Rho-dépendante

Dans le premier cas, la polymérase va reconnaitre un terminateur intrinsèque, correspondant à une séquence formant une tige boucle riche en paires de bases G-C, suivie d'une longue séquence d'uridines favorisant la dissociation du complexe ARN polymérase/ARN fraichement synthétisé sans l'implication d'autres facteurs (Farnham and Platt, 1981; Wilsont and von Hippel, 1995). Dans le deuxième cas, la polymérase a besoin d'une protéine endogène appelée Rho pour terminer la transcription, qui reconnaît une séquence spécifique appelée rut (de l'anglais Rho UTilization) dans l'ARN néosynthétisé en amont de la tige boucle. Rho va alors utiliser son activité hélicase ATP dépendante pour rejoindre l'ARN polymérase, induisant ainsi la dissociation du complexe et l'arrêt de la transcription après contact (Richardson, 2013). Enfin, une caractéristique fondamentale de l'expression des gènes chez les procaryotes est que la transcription est couplée à la traduction. Ainsi, le temps nécessaire au repliement des structures des ARN est corrélé à la vitesse de transcription et de traduction, pouvant moduler l'efficacité de la transcription elle-même ou la traduction lorsqu'il s'agit d'un ARNm.

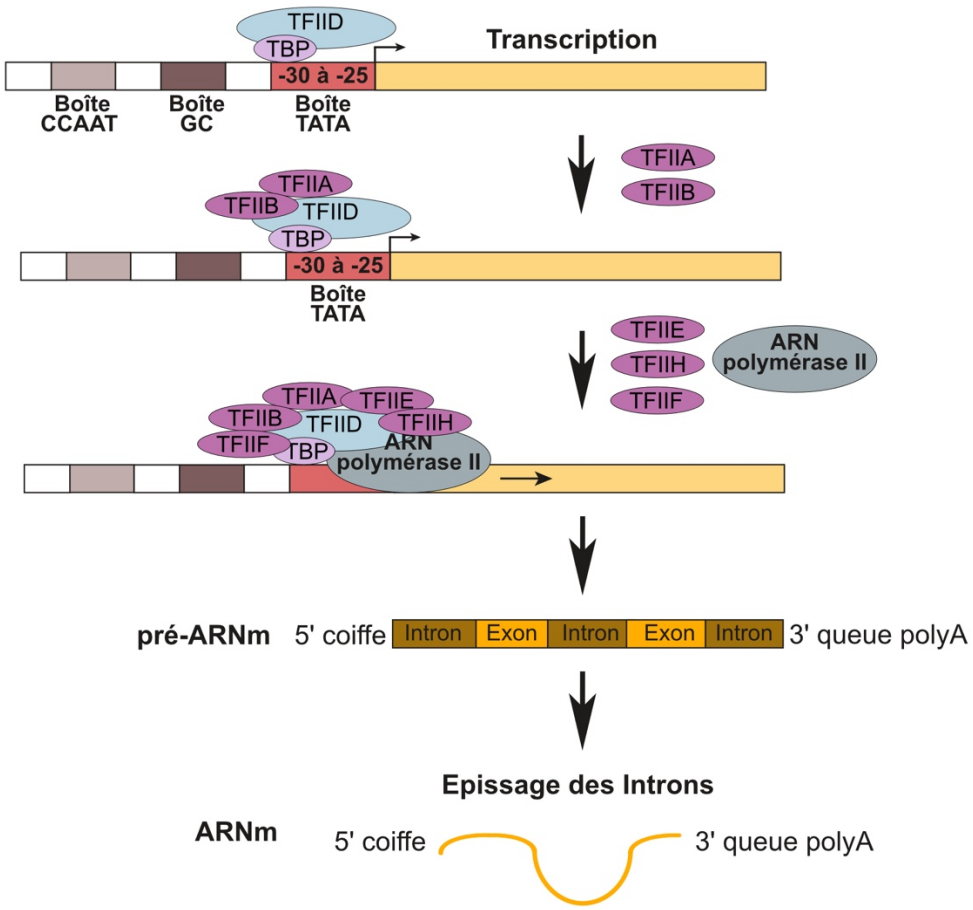
1.1.2 La transcription eucaryote

La transcription chez les eucaryotes peut être réalisée par quatre ARN polymérases différentes selon le type de gènes concernés. Trois d'entre elles, les polymérases I, II et III, sont présentes dans l'ensemble des organismes eucaryotes tandis que la polymérase IV n'est retrouvée que chez les plantes (Barba-Aliaga, Alepuz and Pérez-Ortín, 2021). Chacune a ses propres caractéristiques et localisation cellulaire dans le but de transcrire l'ensemble des gènes :

- L'ARN polymérase I transcrit les ARNr tels que l'ARN 28S et 18S indispensables à la traduction, et ce au niveau du nucléole.
- L'ARN polymérase II transcrit les ARNm au niveau du nucléoplasme.
- L'ARN polymérase III transcrit les petits ARN dont l'ARNr 5S ou encore les ARNt dans le nucléoplasme.

- L'ARN polymérase IV transcrit les petits ARN interférents, un mécanisme propre aux plantes.

D'une manière générale, les ARN polymérases eucaryotes dans leur ensemble sont composées de douze sous-unités dont certaines communes à différentes polymérases, formant un grand complexe d'environ 500 kDa. Une différence notable entre ces complexes est leur localisation, conduisant à la spécificité des gènes transcrits, ainsi que les promoteurs dirigeant l'initiation de la transcription (Krebs *et al.*, 2018; Barba-Aliaga, Alepuz and Pérez-Ortín, 2021). Brièvement, la polymérase I reconnait un promoteur ainsi qu'une séquence en amont de celui-ci appelé "UPE" (de l'anglais : « Upstream Promoter Element »). La polymérase III peut utiliser deux types de promoteurs, soit internes soit en amont de la séquence à transcrire. Enfin, la polymérase II nécessite des facteurs généraux de la transcription tels que $TF_{II}X$ (de l'anglais : « Transcription Factor II X ») pour initier la transcription des ARNm et reconnait un promoteur pouvant comporter différentes portions de séquences conservées (Krebs *et al.*, 2018; Schier and Taatjes, 2020).



<u>Figure 3 :</u> Représentation généraliste de l'initiation de la transcription chez les eucaryotes avec la polymérase II. Le gène transcrit est en jaune. Le facteur général de la transcription $TF_{II}D$ en bleu, (de l'anglais : « Transcription Factor II D ») possède un domaine capable de se lier à la boîte TATA,

appelé TBP (de l'anglais : « TATA Binding Protein ») représenté en rose. Les autres facteurs de la transcription sont représentés en violet, l'ARN polymérase II est en gris. Dans le pré-ARNm les introns sont représentés en marron les exons en orange, et l'ARNm maturé issu de l'épissage des introns du pré-ARN est lui aussi représenté en orange.

Premièrement le promoteur contient une séquence initiatrice conservée dénotée Inr (de l'anglais : « Initiator ») (Figure 3). Deuxièmement, ce dernier peut aussi contenir en supplément une portion de séquence riche en adénines et thymines appelée boîte TATA, ou bien une séquence spécifique présente après le promoteur, appelée "DPE" (de l'anglais : « Downstream Promoter »). Brièvement, l'initiation de la transcription par la polymérase II démarre avec la liaison du facteur $TF_{II}D$ à la séquence initiatrice du promoteur ou à la boite TATA s'il y en a une (Figure 3). Puis de façon chronologique, les autres facteurs de transcription vont se lier à leur tour selon un ordre établi au complexe d'initiation, jusque l'interaction de la polymérase avec ce complexe d'initiation. Cette première étape de la transcription se poursuit en élongation avec le recrutement de différents facteurs ($TF_{II}B$, $TF_{II}E$ et $TF_{II}H$) supplémentaires conduisant au déplacement de la polymérase, notamment par la phosphorylation de son domaine carboxy-terminal (CTD) (Krebs *et al.*, 2018; Schier and Taatjes, 2020) (Figure 3). Lors de l'élongation, les ribonucléotides sont incorporés un à un et associés par une liaison phosphodiester. La dernière étape, la terminaison, se déroule de deux manières différentes selon les gènes, comme chez les procaryotes : avec ou sans l'implication de facteurs supplémentaires. La terminaison intrinsèque se résume à la reconnaissance d'une longue série de thymines dans la séquence transcrite induisant une pause et le détachement de la polymérase. La terminaison facteur dépendante, se réalise à l'aide du domaine CTD et des facteurs "CPSF" (de l'anglais : « Cleavage and Polyadenylation Specificity Factor ») ainsi que les facteurs "CSTF" (de l'anglais : « Cleavage Stimulation Factor ») qui reconnaissent le signal de polyadénylation au sein de la séquence d'ARN néotranscrit, et recrutent les protéines nécessaires au clivage de l'ARN ainsi que sa polyadénylation (maturation de l'ARN). Une fois la transcription du pré-ARNm terminée, ce dernier est épissé par un complexe de ribonucléoprotéines appelé spliceosome, (de l'anglais : « splicing ») dans le but de retirer les portions de séquences non codantes appelées introns, de l'ARNm (Berget, Moore and Sharp, 1977; Chow *et al.*, 1977). Par la suite, l'ARNm épissé est maturé avec d'une part l'ajout d'environ 200 adénines en 3', et d'autre part l'ajout d'une coiffe en 5' pour une grande partie des ARNm (Furuichi, Lafiandra and Shatkin, 1977; Shimotohnot *et al.*, 1977; Krebs *et al.*, 2018). Une fois maturé, l'ARNm est exporté vers le réticulum endoplasmique rugueux (REG) dans le but d'être traduit par les ribosomes.

1.2 La traduction

Que ce soit chez les procaryotes ou les eucaryotes, la traduction se déroule en trois grandes étapes : l'initiation, l'élongation et la terminaison. De manière très générale, l'élongation et la terminaison sont similaires chez les procaryotes et eucaryotes (Krebs *et al.*, 2018). À l'inverse, l'initiation diffère fondamentalement selon l'organisme, mais aussi selon les mécanismes impliqués par chacun (initiation dite canonique ou non canonique) (Kozak, 1999). De manière brève, l'élongation correspond à l'étape durant laquelle les acides aminés sont incorporés de façon séquentielle suivant la séquence d'ARNm. Une fois le premier codon, dit initiateur (codant majoritairement pour la méthionine) incorporé, de manière cyclique les ARNt aminoacylés par leur acide aminé respectif vont venir au niveau du ribosome pour permettre leur ajout selon le codon lu dans le site de décodage. Le déchiffrage du codon stop va induire l'arrêt de la traduction et permettre la libération de la protéine néo synthétisée ainsi que le recyclage du ribosome (Krebs *et al.*, 2018) (Figure 4). L'initiation est l'étape limitante et donc la plus régulée. Notre intérêt se portera tout particulièrement sur cette étape, qui sera également le centre d'intérêt de mes travaux de thèse.

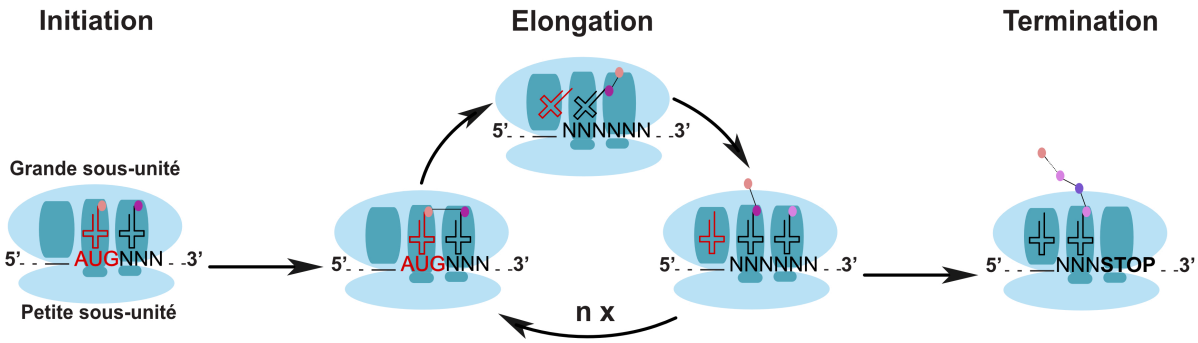


Figure 4 : Représentation globale des trois étapes de la traduction. Le ribosome est représenté en bleu, l'ARNt initiateur aminoacylé par la méthionine initiatrice est représenté en rouge tout comme le codon initiateur (AUG). Les codons suivants ainsi que l'ARNt aminoacylé sont représentés en noir. La liaison peptidique entre les acides aminés est représentée par un trait noir tandis que les acides aminés sont représentés par une gamme de rose.

1.2.1 L'Initiation de la traduction chez les procaryotes

L'initiation est l'étape limitante de la traduction et aussi une étape clef dans la régulation de l'expression des gènes. Chez les procaryotes, pour la majorité des gènes, l'initiation démarre par le recrutement de la petite sous-unités 30S (SU 30S) du ribosome associée au facteur d'initiation IF3 (de l'anglais : « Initiation Factor 3 ») qui empêche son association à la grande sous-unité 50S (SU 50S). La SU 30S reconnait alors une séquence complémentaire à l'ARN 16S dont un consensus (AGGAGGUAA) a été décrit par les chercheurs Shine et Dalgarno dans les années 70 chez *E. coli* (J. Shine and Dalgarno, 1974). En effet, ces derniers sont parvenus à caractériser une séquence consensus optimale pour l'entrée du ribosome RBS (de l'anglais : « Ribosome Binding Site ») dont les caractéristiques ont par la suite été complétées par de nombreuses études (Ringquist *et al.*, 1992; del Campo *et al.*, 2015; Hecht *et al.*, 2017; Cambray, Guimaraes and Arkin, 2018; Komarova *et al.*, 2020; Kuo *et al.*, 2020). Ces caractéristiques peuvent moduler (améliorer ou inhiber) l'efficacité d'initiation de la traduction et correspondent à la proximité de séquence avec le consensus, sa taille, la distance avec le codon initiateur, le contenu des séquences ainsi que les structures possibles environnantes (Hecht *et al.*, 2017; Cambray, Guimaraes and Arkin, 2018; Komarova *et al.*, 2020; Kuo *et al.*, 2020). Ainsi, une fois que le ribosome a reconnu le RBS il est positionné sur le codon initiateur à une distance optimale de 5 à 10 nucléotides de cette séquence consensus (Rodnina, 2018).

Le premier codon se localise alors au niveau du site P (peptidyl-transferase) du ribosome, et sera reconnu par le complexe entre l'ARNt initiateur aminoacylé par une formylméthionine et le facteur d'initiation IF2 (de l'anglais Initiation Factor 2) couplé au GTP. L'hydrolyse du GTP par IF2 entraîne alors la libération du site A (accepteur) du facteur IF3 et la petite sous-unité est à présent capable de s'associer à la SU 50S afin d'accueillir un nouvel ARNt aminoacylé correspondant au second codon (Krebs *et al.*, 2018; Rodnina, 2018). Une fois le ribosome assemblé (50 et 30S), le codon initiateur et le premier acide aminé (méthionine) sont dans le site P et un deuxième ARNt aminoacylé peut apporter le deuxième acide aminé au niveau du site A. La liaison peptidique est catalysée entre les deux acides aminés par l'activité peptidyl transférase de l'ARNr 23S. Puis, le ribosome réalise une translocation à l'aide du facteur EF-G (de l'anglais : « Elongation Factor G »), faisant passer le premier ARNt vers le site E (de l'anglais : « Exit ») permettant sa sortie, celui du site A au site P permettant alors l'entrée du prochain ARNt aminoacylé guidé par le facteur EF-tu (de l'anglais : « Elongation Factor thermo unstable ») qui par l'hydrolyse de son GTP aide au décodage du codon (Krebs *et al.*, 2018; Rodnina, 2018). Ces actions sont répétées de manière cyclique lors de l'ensemble de la phase d'élongation de la traduction (Figure 5).

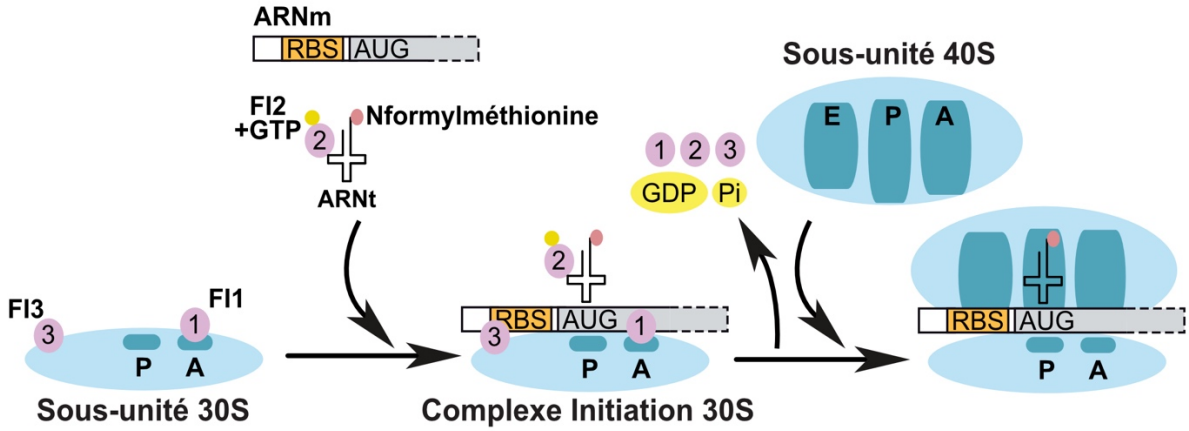


Figure 5 : Représentation schématique de l'initiation de la traduction procaryote. Les facteurs de l'initiation (FI) sont représentés en rose, le ribosome en bleu. Le site d'entrée du ribosome ou RBS en orange interagit avec sa séquence complémentaire au niveau de l'ARN 16S de la SU-30S. La guanine tri-phosphate puis di-phosphate avec le pyrophosphate sont représentés en jaune, pour finir la méthionine initiatrice est représentée en rose.

Cependant, un nombre non négligeable de gènes ne comportent pas de séquence correspondant à un RBS (absence de séquence favorisant l'arrimage du ribosome) et sont appelés "Leaderless" (Zheng *et al.*, 2011). L'initiation dans ce cas a lieu en absence de structures secondaires dans la zone juste avoisinante le codon initiateur (Scharff *et al.*, 2011; del Campo *et al.*, 2015) ou encore par l'intervention de protéines limitant ces structures telles que la protéine S1 chez *E. coli* (Boni *et al.*, 1990; Duval *et al.*, 2013; Qureshi *et al.*, 2018).

1.2.2 L'initiation de la traduction chez les eucaryotes

Chez les eucaryotes, de même que pour les procaryotes, l'initiation reste une étape limitante, finement régulée par divers éléments affectant l'efficacité de son initiation, détaillés en 2.1.2 . Il existe deux grands mécanismes d'initiation, coiffe-dépendant et coiffe-indépendant (Merrick, 2004; Krebs *et al.*, 2018; Shirokikh and Preiss, 2018; Kwan and Thompson, 2019).

L'initiation canonique est le mécanisme coiffe-dépendant. Pour ce type d'initiation, l'ARNm doit impérativement, lors de sa maturation, être coiffé en 5' et polyadénylé en 3'. En effet, parmi le grand nombre de facteurs appelés eIF (de l'anglais : « eukaryotic Initiation Factor ») impliqués tout au long de l'initiation, certains interagissent directement avec ces deux éléments (coiffe et queue polyA) pour circulariser l'ARNm (Shirokikh and Preiss, 2018). De façon plus détaillée, la petite sous-unité 40S (SU 40S) est associée à deux facteurs ; eIF3 empêchant son association avec la grande sous-unité 60S (SU 60S) et eIF1 bloquant le site A du ribosome afin de guider l'ARNt initiateur aminoacylé de la méthionine initiatrice

directement dans le site P. Une fois ce complexe de pré-initiation formé, d'autres facteurs (eIF4 A, B, E et G) se fixent au niveau de l'ARNm afin de former un complexe liant la coiffe et la queue polyA par l'intermédiaire des protéines liant la queue polyA ou PABP (de l'anglais : « PolyA Binding Protein »). Le complexe de pré-initiation interagit alors avec l'ARNm cyclisé et démarre un scanning jusqu'au codon initiateur (Krebs *et al.*, 2018; Shirokikh and Preiss, 2018). Chez les eucaryotes, une séquence consensus à proximité du codon initiateur a été mis en avant par M. Kozak dans les années 70, qui bien que peu précise et caractérisé semble favoriser l'initiation de la traduction (Kozak, 1981). Une fois la SU 40S positionnée au niveau du codon initiateur, l'hydrolyse du GTP de eIF2 induit la dissociation des différents facteurs d'initiation de cette sous-unité, tels que eIF1 et 3 afin de permettre l'association avec la SU 60S et donc la formation du ribosome. Un deuxième ARNt aminoacylé peut alors se placer au niveau du deuxième codon de l'ARNm (Krebs *et al.*, 2018; Shirokikh and Preiss, 2018) (Figure 6). Puis, de manière cyclique, une liaison peptidique est formée entre les acides aminés au sein du site P grâce à l'activité peptidyl-transférase de l'ARNr 28S. Le ribosome subi ensuite une translocation du premier ARNt aminoacylé dans le site E, du deuxième dans le site P suivi de la venue d'un nouvel ARNt aminoacylé correspondant au codon suivant dans le site A. L'élongation se poursuit alors jusqu'à la fin de la traduction (Krebs *et al.*, 2018; Shirokikh and Preiss, 2018).

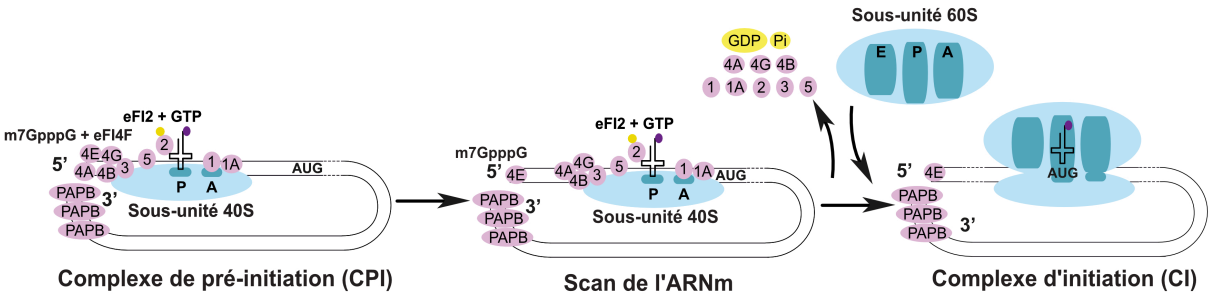


Figure 6 : Représentation schématique de l'initiation cap-dépendante de la traduction chez les eucaryotes. Les facteurs de l'initiation sont représentés en rose et numérotés selon leur appellation, de même que les protéines liant la queue poly adénilée. Le ribosome est représenté en bleu, la guanosine tri-phosphate (GTP) devenant du diphosphate (GDP) accompagnée de phosphate inorganique après hydrolyse sont représentés en jaune, la méthionine initiatrice en gris clair.

L'initiation dite coiffe-indépendante est guidée par des régions très structurées capables de capter le ribosome, et ont été initialement caractérisés dans les ARNm codés par certains virus de cellules eucaryotes. Ces éléments sont appelés IRES (de l'anglais Internal Ribosome Entry Site) qui, comme leur nom l'indique, permettent l'entrée du ribosome au niveau ou à proximité du codon initiateur (Kwan and Thompson, 2019). 4 classes d'IRES ont été décrites selon leurs tailles, séquences, structures secondaires et tertiaires (Mailliot and

Martin, 2018; Kwan and Thompson, 2019). Les IRES de type IV ne nécessitent aucun facteur de l'initiation ni même de scanning puisque l'IRES guide directement le ribosome au niveau du codon initiateur en mimant la boucle anti-codon de l'ARNt. Les IRES de type III nécessitent uniquement eIF2 et 3 pour interagir avec la SU 40S. Enfin, les IRES de type I et II ont besoin de plus de facteurs de l'initiation et vont même pour certaines promouvoir un scanning (Mailliot and Martin, 2018; Kwan and Thompson, 2019). Les IRES sont des éléments moins bien décrits à l'heure actuelle, mais capables d'initier la traduction de manière alternative à l'initiation canonique les rendant d'autant plus intéressantes (Merrick, 2004; Komar and Hatzoglou, 2011; Sriram, Bohlen and Teleman, 2018a) (Figure 7).

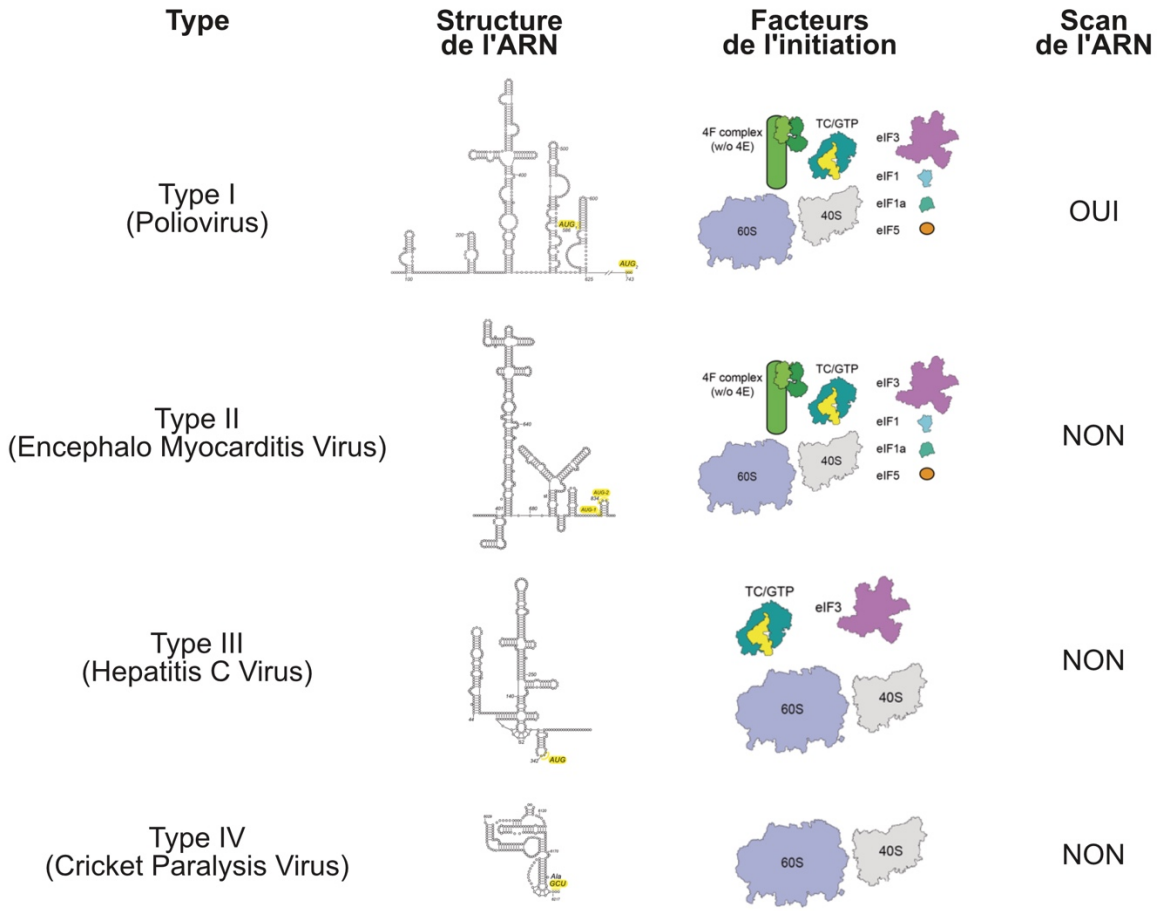


Figure 7 : Récapitulatif des différents types d'IRES. Est représenté, de gauche à droite, le type d'IRES (avec le virus associé), un exemple de structure secondaire (AUG en jaune), les facteurs de la traductions nécessaires à l'expression du gène associé, et la nécessité d'une étape de scanning ou non. Figure adaptée de (Mailliot and Martin, 2018).

L'expression des gènes est un mécanisme complexe et varié qu'il faut finement réguler de façon à ajuster de manière précise l'expression aux besoins de la cellule. De nombreux mécanismes de régulations ont ainsi pu être mis en évidence. (Mailliot and Martin, 2018)

2. Régulation de l'expression des gènes

L'expression coordonnée et finement adaptée des gènes est indispensable à la survie de n'importe quel organisme. Ce processus comprend différentes phases allant de la transcription à la dégradation des ARN, en passant par la traduction des ARNm à la dégradation des protéines, et chaque étape est finement régulée par divers moyens. Dans un souci de clarté, je décrirais ici exclusivement les processus de régulation de la transcription et de l'initiation de la traduction chez les procaryotes comme les eucaryotes, introduits précédemment qui constituent l'essentiel de mon travail de thèse.

2.1 Régulation chez les procaryotes

La modulation de l'expression des gènes chez les procaryotes peut aussi bien avoir lieu lors de la transcription que de la traduction voire plus simplement dans l'organisation des gènes eux-mêmes.

2.1.1 Régulation transcriptionnelle et organisation des gènes

Chez les procaryotes la taille moyenne des génomes est aux alentours du millions de paires de bases (Mpb) avec par exemple *E. coli* composé de 4,5 Mpb (Blattner *et al.*, 1997), ce qui malgré tout est nettement inférieure aux génomes eucaryotes pouvant atteindre des milliers de millions de paires de bases (Mpb) tel que le génome humain atteignant 3400 Mpb (International Human Genome Sequencing Consortium, 2004). Les génomes procaryotes sont donc organisés de manière stratégique. En général, plusieurs gènes peuvent être rassemblés dans des systèmes de régulation appelés opéron sous le contrôle d'un promoteur unique (Monod, 1949; Pardee, François Jacob and Jacques Monod, 1959; Lewis, 2013). Ainsi, l'initiation de la transcription, aura lieu conjointement pour l'ensemble des gènes constituant l'opéron, permettant une économie de temps et d'énergie. De manière intéressante, les opérons regroupent l'ensemble des gènes nécessaires à l'expression et / ou le transport d'une même voie métabolique tel que l'opéron lactose (Pardee, François Jacob and Jacques Monod, 1959; Lewis, 2013). Les opérons peuvent être inductibles ou répressibles selon leur fonctionnement, et ce, en présence ou absence de petites molécules. Par exemple, l'opéron lactose régule les gènes impliqués dans le transport et le métabolisme du lactose chez *E. coli.* En absence de cette petite molécule, une protéine appelée répresseur se lie dans la région promotrice de l'opéron empêchant sa transcription. À l'inverse, la présence d'allolactose et sa

liaison au répresseur induit un changement de conformation qui est alors incapable de lier au promoteur, entraînant l'expression de l'opéron (Pardee, François Jacob and Jacques Monod, 1959; Lewis, 2013). Cette organisation est judicieuse car elle permet d'orchestrer l'expression de l'ensemble de la voie.

De manière plus simpliste, un élément indispensable à la transcription sert de levier de contrôle de cette dernière, le promoteur (Harley and Reynolds, 1987). Ce dernier est reconnu par la polymérase lors de l'initiation de la transcription et peut contenir une séquence définit comme capable de favoriser l'initiation de façon plus ou moins efficace, mais aussi d'attirer certains facteurs sigma. Par exemple, deux séquences désignées comme des "boîtes" situées 35 nucléotides et 10 nucléotides en amont du site d'initiation de la transcription ont été décrites comme des éléments favorisant la transcription avec l'exemple le plus répandus de la boîte TATA, localisée au niveau des gènes de ménages (Harley and Reynolds, 1987; Krebs *et al.*, 2018).

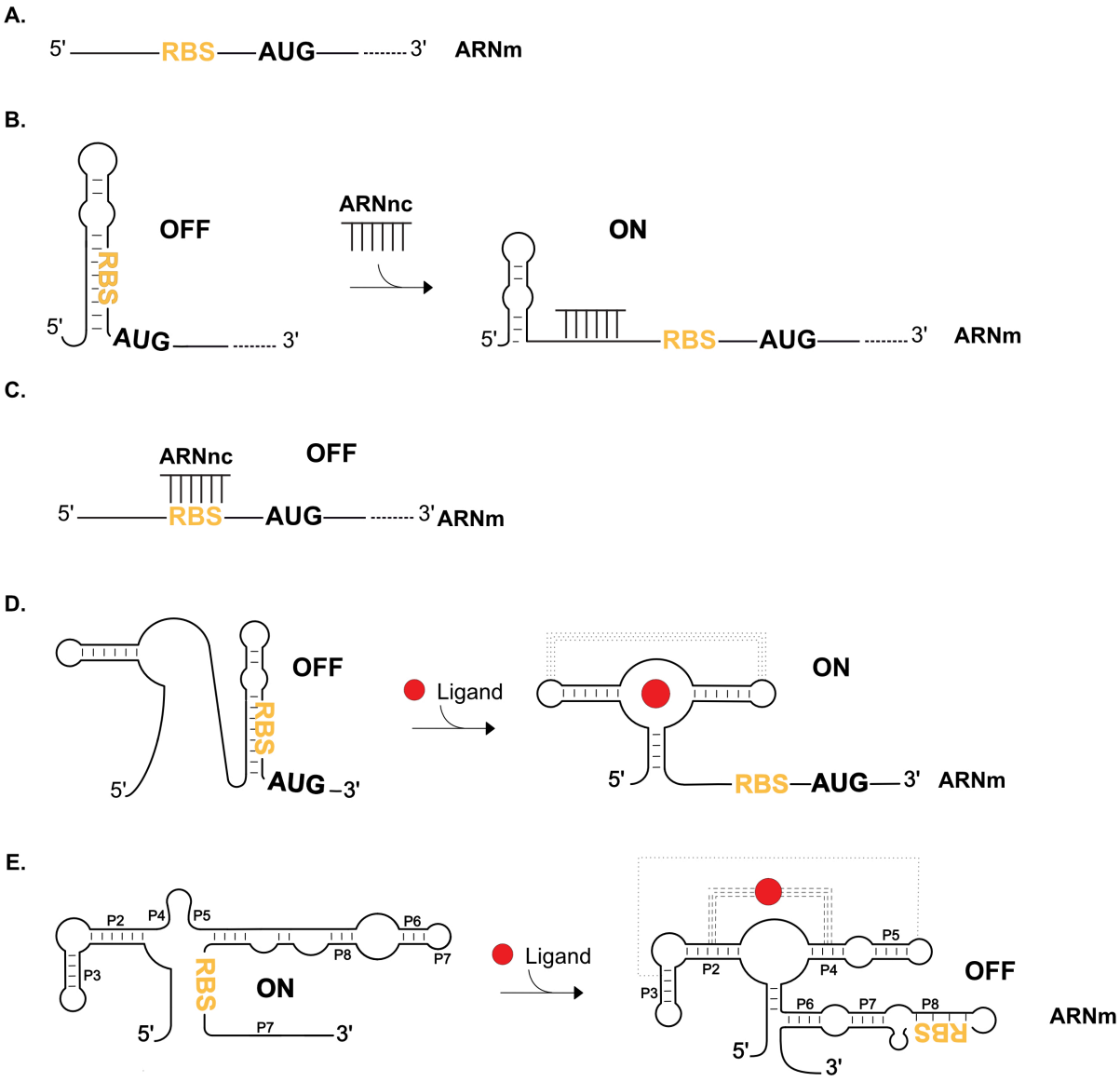
Enfin la terminaison peut, elle aussi, être la cible de régulations par l'intermédiaire d'éléments tels que les riboswitches. Les riboswitches ARN sont des séquences d'ARN, qui une fois transcrites, vont adopter une structure spécifique en fonction de la présence ou l'absence d'un ligand (Husser, Dentz and Ryckelynck, 2021). Cette structuration peut conduire à la stabilisation d'un terminateur de transcription ou une structure anti-terminatrice modulant ainsi l'expression des gènes en aval. Il est à noter que les riboswitches contrôlent souvent l'expression d'opérons régulant les gènes impliqués dans la même voie métabolique que le métabolite reconnu par la structure ARN. Leurs fonctionnalités et ingénieries sont présentées plus en détails dans le paragraphe 2.1.3.

2.1.2 Régulation traductionnelle

La majorité des mécanismes de régulation de la traduction impacte l'étape de l'initiation, bien que d'autres facteurs puissent tout de même affecter l'efficacité globale de la traduction d'un gène tel que la stabilité de l'ARNm par exemple (Arraiano, 1993). En effet, certains ARNm sont peu stables et peuvent être rapidement dégradés dans la cellule, ne laissant que peu de temps à la machinerie traductionnelle pour produire la protéine correspondante.

L'étape d'initiation de la traduction reste le point limitant, et la cellule dispose de multiples moyens pour moduler son efficacité. Pour commencer, comme décrit plus haut, la nature de la séquence complémentaire à l'ARN 16S qui chez *E.coli*, appelée séquence de Shine et Dalgarno (SD), impacte l'efficacité de la traduction, de par son degré de complémentarité avec l'ARN 16S, sa distance avec le codon initiateur, la présence de structure dans la séquence environnante etc... (Stormo, Schneider and Gold, 1982; Ringquist *et al.*,

1992; Hecht *et al.*, 2017; Cambray, Guimaraes and Arkin, 2018) (Figure 8 A). En outre, d'autres éléments peuvent venir moduler l'accessibilité du RBS (Figure 8 B, C, D, et E). Cette modulation peut être réalisée par des éléments agissant en *cis* ou en *trans* (Dutta and Srivastava, 2018; Desgranges *et al.*, 2019; Chiaruttini and Guillier, 2020). Les éléments agissant en *trans* sont souvent de petits ARN capable de venir s'apparier avec le RBS de l'ARNm qui n'est alors plus accessible au ribosome (Dutta and Srivastava, 2018; Desgranges *et al.*, 2019; Chiaruttini and Guillier, 2020) (Figure 8 B et C). Les éléments en *cis* correspondent aux riboswitches précédemment décrits pour la régulation de la transcription. Ici, le changement de conformation de l'ARN module l'accessibilité du RBS selon la présence ou l'absence du ligand (Chiaruttini and Guillier, 2020; Husser, Dentz and Ryckelynck, 2021) (Figure 8 D et E). Différents exemples et mécanismes de fonctionnement sont détaillés dans la revue que j'ai co-écrite et qui est présentée dans le paragraphe suivant en 2.1.3 (Husser, Dentz and Ryckelynck, 2021)



<u>Figure 8 :</u> Exemples de mécanisme de régulation de l'initiation de la traduction procaryote. A. Impact de la séquence de l'ARNm elle-même. Le RBS est représenté en orange et l'ARNm en noir. **B.** Activation de la traduction par un petit ARN régulateur. Le RBS est représenté en orange, l'ARNm en noir. **C.** Inhibition de la traduction par un petit ARN régulateur. Le RBS est représenté en orange, l'ARNm en noir. **D.** Activation de la traduction par un riboswitch. Le RBS est représenté en orange, l'ARNm en noir et le ligand du riboswitch en rouge. **E.** Inhibition de la traduction par un riboswitch. Le RBS est représenté en orange, l'ARNm en noir et le ligand du riboswitch en rouge. Les figures **D** et **E** sont adapté de (Husser, Dentz and Ryckelynck, 2021).

2.1.3 Les riboswitches des éléments au cœur de la régulation

Les riboswitches sont des éléments très utilisés par les bactéries pour réguler l'expression de gènes impliqués dans une même voie métabolique ou encore la réponse au stress. Ils ne sont en revanche que rarement retrouvés chez les eucaryotes (uniquement chez les plantes dans de rare cas). En plus de leur rôle biologique naturel clé, les riboswitches sont de plus en plus exploités comme plateforme synthétique de contrôle de l'expression des gènes tant chez les procaryotes que les eucaryotes. Ces outils peuvent avoir diverses applications allant de l'étude de l'expression des gènes à la détection de molécules. Que l'application soit de suivre l'expression de certains gènes ou de détecter de petites molécules, les riboswitches sont fréquemment utilisés en tandem de séquences codantes pour des protéines rapportrices. Pour la localisation de protéines ou le suivi de la traduction, la protéine fluorescente verte ou GFP (de l'anglais : « Green fluorescente Protein ») est la plus couramment utilisée ou la Rénillase luciférase permettant alors un suivi par luminométrie. La transcription peut être suivie *via* l'utilisation d'aptamères fluorogènes détaillés dans le paragraphe (3.2.1) ainsi que dans la revue suivante (Husser, Dentz and Ryckelynck, 2021).

Les riboswitches sont donc des éléments capables de flexibilité et souplesse à la suite d'une interaction avec une molécule et sont au cœur de nombreuses expérimentations. Le détail de leur fonctionnement et de leur capacité d'applications a été détaillé dans la revue que j'ai co-signé et qui est reproduite ci-après.

2.1.3.1 Revue : “Structure-Switching RNAs : From Gene Expression Regulation to Small Molecule Detection”

Structure-Switching RNAs: From Gene Expression Regulation to Small Molecule Detection

Claire Husser, Natacha Dentz and Michael Ryckelynck

2021

Small Structure Journal

small structures
www.small-structures.com

Structure-Switching RNAs: From Gene Expression Regulation to Small Molecule Detection

*Claire Husser, Natacha Dentz, and Michael Ryckelynck**

RNA is instrumental to cell life in many aspects, especially gene expression regulation. Among the various known regulatory RNAs, riboswitches are particularly interesting *cis*-acting molecules as they do not need cellular factor to achieve their function and are therefore highly portable from one organism to the other. These molecules usually found in the 5′ untranslated region of bacterial messenger RNAs are able to specifically sense a target ligand *via* an aptamer domain prior to transmitting this recognition event to an expression platform that turns on, or off, the expression of downstream genes. In addition to their obvious scientific interest, these modular molecules can also serve for the development of synthetic RNA devices with applications ranging from the control of transgene expression in gene therapy to the specific biosensing of small molecules. The engineering of such nanomachines is greatly facilitated by the proper understanding of their structure as well as the introduction of new technologies. Herein, a general overview of the current knowledge on natural riboswitches prior to explaining the main strategies used to develop new synthetic structure-switching molecules (riboswitches or biosensors) controlled by small molecules is given.

1. Introduction

As originally proposed by Francis Crick, the central dogma of molecular biology places RNA at the center of gene expression, first as being a labile intermediate (messenger RNA) of this expression.[1] Many structural functions were also early attributed to these nucleic acids upon finding their key involvement in RNA post-transcriptional modification (e.g., snoRNA),[2] genetic information translation (i.e., transfer RNAs),[3] or as scaffold allowing the assembly of complex molecular machines (e.g., ribosome and spliceosome).[4] Later on, the discovery of the catalytic potential of RNA, in the form of ribonucleic enzymes (ribozymes in short),[5] marked a turn in RNA (and even in molecular) biology by providing evidences that nanomachines such as RNase P,[6] the spliceosome, and even the ribosome are themselves ribozymes. Such complex functions are primarily made possible by the great structural plasticity of RNA allowing the molecule to explore a vast repertoire of 3D folding and the emergence of structures in which key residues involved in reaction catalysis are precisely placed in space (the reader is redirected to excellent reviews on the topic and the references within).[7] The discovery of ribozymes also brought the concept of "RNA World" in which Walter Gilbert[8] proposed that early life may have appeared and evolved on Earth from systems in which RNA (or a similar polymer) not only stored the genetic information but also carried the enzymatic activity allowing energy to be produced and genetic information to be replicated, two functions later, respectively, transferred to DNA and proteins in contemporary systems.

Finally, over the past decades, the set of functions attributed to RNA kept on growing up with the discovery of a plethora of gene expression regulatory RNAs in every kingdom of life. Most of these RNAs act *in trans* (e.g., microRNAs or long noncoding RNAs in eukaryotes and small noncoding RNAs in bacteria) by directly annealing to a target RNA.[9] Moreover, bacteria also make extensive use (regulation of at least 2% of the genes in *Bacillus subtilis*) of regulatory RNAs directly embedded within the 5′ untranslated region (5′ UTR) of the messenger RNA (mRNA) and acting *in cis*. These regulatory elements adjust gene expression by locally modulating mRNA structure upon change in pH or temperature,[10] the binding of a tRNA (in the so-called T-box system),[11] or the specific direct recognition of small molecules. RNAs entering in the latter category were discovered nearly 20 years ago and named riboswitches.[12] As most riboswitches are unique to bacteria, these RNAs are also very attractive targets for the development of new generations of antibiotics.[13] Moreover, as they work in a stand-alone manner, riboswitches can be seen as nanofunctional modules that can be inserted within a transcription unit and used as simple synthetic gene expression regulators for synthetic biology applications. Furthermore, they also represent an extremely valuable starting point in the design of specific small molecule biosensors that can be easily used in vitro or in living cells. The broad application range of these RNA nanoblocks stimulated a very active field of research and motivated the writing of this article that does not pretend to be a comprehensive review of the field but rather aims at giving a rapid overview of the current knowledge on natural riboswitches and how they can be harnessed for various applications, prior to introducing the main

C. Husser, N. Dentz, Prof. M. Ryckelynck
Architecture et Réactivité de l'ARN
Université de Strasbourg, CNRS
UPR 9002, Strasbourg F-67000, France
E-mail: m.ryckelynck@unistra.fr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/sstr.202000132.

DOI: 10.1002/sstr.202000132

strategies used to develop artificial riboswitches or convert them into innovative biosensing probes.

2. Natural Riboswitches

2.1. Discovery of Natural Riboswitches

Natural riboswitches were originally discovered within unusually long 5′ UTR of mRNA whose expression was known to be regulated by a mechanism involving a metabolite of the pathway the gene belongs to, but for which no protein effector can be identified by conventional genetics. In addition, these mRNAs were also known to possess a highly conserved motif (or box) embedded within strongly structured regions and shared between many, sometimes quite evolutionary distant, bacteria species. First experimental evidences that these motifs can autonomously regulate downstream gene expression came from the study of mRNAs coding for enzymes involved in the biosynthesis of cobalamin (vitamin B12),[12] thiamine pyrophosphate (TPP, a derivative of vitamin B1), and flavin mononucleotide (FMN, a derivative of vitamin B2).[14] Soon after, the concept was rapidly extended to several additional ligands including S-adenosylmethionine (SAM) and purines.[15] Excitingly, these first ligands either contained a nucleotide-derived moiety or are related to nucleotide metabolism which led to the proposition that riboswitches may be relics of the RNA World in which they could have acted as primitive regulatory systems.[16] If true, this observation would elegantly explain the wide distribution of some motifs observed within the bacterial reign even between distant species.

The discovery rate of riboswitches then strongly accelerated with the explosion of bioinformatics and comparative genomics allowing the exploration of the ever-growing microbe genome databases. Searching for noncoding sequences shared by several genes, within the same or between different organisms, and containing a conserved motif embedded within a conserved 2D structure led to the discovery of dozens of new motifs.[17] Studying the sequence context surrounding the motif can help identifying the type of regulation (e.g., transcriptional or translational control) at work. Moreover, in some cases, the identity of the gene downstream the motif can help inferring the identity of riboswitch ligand as this was the case, for instance, with the glycine riboswitch found upstream a gene encoding an enzyme involved in glycine metabolism.[18] Yet, in many cases, there is such a large diversity of genes (sometimes wrongly annotated) controlled by the same riboswitch motif that it becomes extremely challenging to properly infer ligand identity, turning such riboswitches orphans as recently nicely reviewed by Breaker's lab.[19] In such a case, identifying the biologically relevant ligand may require great ingenuity, especially when the molecule was not previously known to be present in the cell. This is typically illustrated by the case of guanidine-I riboswitch (formerly known as the orphan ykkC-group 1) for which it took a decade to identify guanidine as being the natural ligand.[20] Among other limitations, this identification was complexified by erroneous gene annotations and by the ignorance that guanindine (a well-known protein denaturant) can be produced by the cell as a final product of guanine degradation. Although many riboswitches are still orphans, nearly two decades after their first functional description, this class of RNA now encompasses nearly 50 natural motifs responding to ligands as various as ions, nucleotides and their derivatives, amino acids, vitamins, and various other metabolites (**Table 1**).[21] An open question concerns the total number of different natural riboswitches that one may still expect to be discovered in the coming years. Indeed, it is likely that the most abundant riboswitches have now been identified.[21a] Nevertheless, assuming that riboswitch abundance follows a power-law distribution,[22] it has been proposed that hundreds (if not thousands) of unabundant motifs may still be left to be discovered;[21a,23] these RNAs representing a vast reservoir of new motifs, and likely new specificities, with as many potential applications. Nevertheless, these putative riboswitches are also expected to be moderately to poorly abundant which may significantly delay their identification.

2.2. Gene Expression Control by Natural Riboswitches

As stressed earlier, most of the riboswitches do not require the assistance of protein factors and are stand-alone regulatory elements with both the capacity to sense a target molecule and to modulate the expression of downstream genes. To do so, a riboswitch is usually made of two parts: an aptamer domain and an expression platform that, respectively, specifically recognize the ligand and turn-on/off gene expression upon a structure switching event.[14b] Interestingly, riboswitches are modular genetic elements and a given type of aptamer can be found associated with different expression platforms in different organisms (or even within the same organism), whereas the same type of expression platform can be found associated with different aptamers. Moreover, and further adding to their value, several riboswitches can be simultaneously present upstream the same gene allowing Boolean logic regulation to take place.

2.2.1. Structures and Features of Aptamer Domains

The ligand-recognition module was named aptamer domain by analogy with synthetic nucleic acids isolated by in vitro selection methods (e.g., Systematic Evolution of Ligands by EXponential enrichment [SELEX], see later) for their capacity to specifically interact with a target molecule.[14b] As it carries the ligand recognition function, the aptamer is the most conserved part of the riboswitch with a strict sequence conservation of the residues directly contacting the ligand by H-bond, electrostatic or water-mediated interactions, or even π-staking. These residues usually reside between secondary structure elements (i.e., helices and pseudoknots) that are more tolerant to sequence variations provided the structure is preserved. Interestingly, identifying such sequence covariations (i.e., the simultaneous presence of a mutation and its compensatory) constitutes a solid argument to validate secondary structure models generated in silico. There is no precise rule regarding the size and the structural complexity of aptamer domains. Indeed, on the one hand, the length of the molecule can vary from a few tens of nucleotides in the case of small aptamers such as the motifs found in fluoride (made of pseudoknot and two paired regions P1 and 2),[24] guanidine-II (made of two paired regions, P1 and P2, whom the apical loops establish loop–loop interaction),[25] and pre-Q1-I (made

Table 1. Classes of experimentally validated riboswitches and Protein Data Bank (PDB) number of the crystal structures (when available).

	Ligand	Class riboswitch	Ref.	PDB
Coenzyme	Adenosyl-cobalamin (Ado-Cbl), vitamin B12	Cbl-I	[129]	4GMA, 4GXY
		Cbl-IIb	[130]	–
	Methylcobalamin (MeCbl), Aquocobalamine (AqCbl)	Cbl-IIa	[129]	4FRN, 4FRG
	Glucosamine-6-phosphate (GlcN6P)	Glms	[131]	2HO7, 2N74
	Flavinmononucleotide (FMN)	FMN subtype 1	[132]	3F4E
		FMN subtype 2	[133]	–
	Thiamin pyroPhosphate (TPP)	THI (Thi-box)	[14b]	2GDI, 4NYG
	Thiamine precursor (HMP-PP)	ThiS	[51]	–
	Molybdenum-cofactor (Moco), Tungsten-cofactor (Tuco), WCo	Moco	[134]	–
	S-adenosylhomocysteine (SAH)	SAH	[135]	3NPN, 3NPQ
	S-adenosylmethionine (SAM)	SAM-I	[15a]	2GIS, 3IQN
		SAM-II (SAM-α)	[136]	2QWY
		SAM-III (SMK box)	[137]	3E5C
		SAM-IV	[138]	6UES, 6UET
		SAM-V	[139]	6FZ0
		SAM-VI	[140]	6LAS
	SAM/SAH	SAM-SAH	[17c]	6HAG
	Tetrahydrofolate (THF)	THF-I	[141]	3SD1, 3SD3
		THF-II	[17d,142]	–
	Nicotinamide-adenine dinucleotide (NAD^+)	NAD^+-I	[143]	7D7V
		NAD^+-II	[144]	–
Amino Acid	Glycine	glycine	[18]	3DIY, 3P49 3OXE
	L-Lysine	L-box	[28,52]	3D0U, 3D0X
	L-Glutamine	Glutamine-I	[145]	5DDO, 5DDP
		Glutamine-II	[146]	6QN3
Metal	Magnesium (Mg^{2+})	Mg2+-I (M-box)	[27b]	2QBZ
		Mg2+-II	[27a]	–
	Manganese (Mn^{2+})	Mn	[32]	4Y1I, 4Y1M
	Nickel (Ni^{2+}), cobalt (Co^{2+}), iron (Fe^{2+})	NiCo	[120]	4RUM
	Fluor (F^-)	Fluoride	[24]	4ENC
Nucleotide and derivative	Guanine/hypoxanthine	Purine	[15b]	1Y27, 4FE5
	Adenine	Purine	[147]	1Y26
	2′-deoxyguanosine (2′-dG)	2′-dG-I	[148]	3SKI
		2′-dG-II	[17e]	6P2H
	PreQueuosine1 (PreQ1)	PreQ1-I	[26]	2L1V
		PreQ1-II	[149]	2MIY
		PreQ1-III	[150]	4RZD
	3′-5′-Cyclic-di-GMP (c-di-GMP)	c-di-GMP-I, GEMM-I	[39]	3IRW, 3MXH
		c-di-GMP-II, GEMM-II	[151]	3Q3Z
	3′-5′-Cyclic-di-AMP (c-di-AMP)	c-di-AMP	[152]	4W90 4W92
	cyclic AMP-GMP (cAMP-GMP; cGAMP)	GEMM-Ib	[42113a]	4YAZ, 4YB0
	8-Azaxanthine, xanthine, and uric acid	NMT1	[153]	–
	Guanidine	ykkC 1: Guanidine-I	[20b]	5T83, 5U3G
		mini-ykkC: Guanidine-II	[25]	5NDI, 5NOM 5NDH
		ykkC-III: Guanidine-III	[154]	5NWQ
	ppGpp	ykkC 2a	[43]	6DMC
	PRPP	ykkC 2b	[44]	6CK5
	ADP, dADP, CDP, and dCDP	ykkC 2c	[155]	–
	ZMP/ZTP (5-Aminoimidazole-4-carboxamide riboside 5′-triphosphate)	ZTP	[156]	6OD9
	Aza-aromatics	yjdF	[157]	–

of a simple pseudoknot and a single paired region, P1) riboswitches (**Figure 1**A–C),[26] to molecules approaching, or even exceeding, hundred nucleotides in aptamers responding to Mg^{2+} (made of six paired regions, P1 to P6 on Figure 1D),[27] to lysine (made of five paired regions P1 to P5 on Figure 1E),[28] or even targeting guanidine-I (made of three paired regions, P1 to P3 on Figure 1F),[20b] just as a few examples. On the other hand, the aptamer structure can be as simple as two kissing stem-loops (e.g., in guanidine-II),[29] or a pseudoknot (e.g., in fluoride and pre-Q1)[30] but, in some cases, it can also be rather complex and compact, involving the stacking of several helices connected through a three-way (e.g., purine and TPP riboswitches),[31] a four-way (e.g. Mn^{2+} riboswitch),[32] or even a five-way junction (e.g., lysine riboswitch),[33] an organization completed by numerous tertiary interactions (Figure 1A–I). As

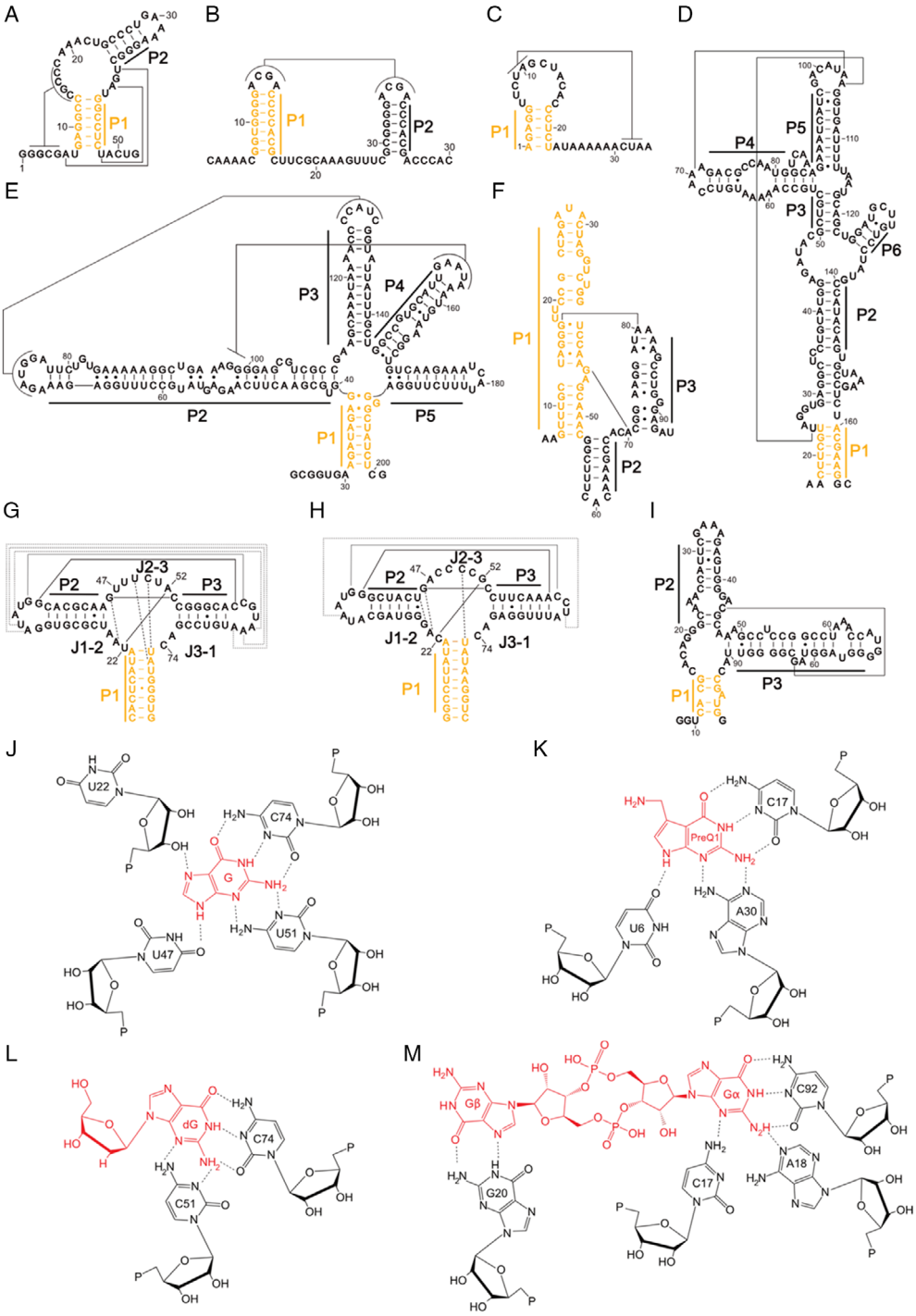


Figure 1. Size and structural diversity of natural riboswitches and their interaction mode with their cognate ligand. A–I) Sequences and 2D structure of natural riboswitches. P1 helices are shown in orange, whereas other paired (P) and unpaired joining (J) regions are shown in black. A) Fluoride riboswitch from *Pseudomonas syringae*.[30a] B) Guanidine-II riboswitch from *Gloeobacter violaceus*.[29] C) PreQ1-I riboswitch from *B. subtilis*.[30b] D) Magnesium-I riboswitch from *B. subtilis*.[27b] E) Lysine riboswitch from *B. subtilis*.[125] F) Guanidine-I riboswitch from *Sulfobacillus acidophilus*.[126] G) Guanine riboswitch from *B. subtilis*.[40] H) 2′-dG-I riboswitch from *Mesoplasma florum*.[127] I) c-di-GMP-I riboswitch from *Vibrio cholerae*.[39] J–M) Ligand recognition by the aptamer domain of the riboswitches. The H-bonds are represented by dotted lines, ligand is shown red, and the nucleotides from the riboswitch are shown in black. Nucleotides in direct interaction with the ligand are identified and numbered according to the 2D structure shown earlier. J) Guanine recognition by its specific riboswitch in *B. subtilis*. The specificity for guanine (G, in red) is brought by a Watson–Crick interaction with C74.[35] K) PreQ1-I riboswitch from *B. subtilis*. The specificity for PreQ1 (PreQ1, in red) is brought by a Watson–Crick interaction with C17.[30b] L) 2′-dG-I riboswitch from *M. florum*. The specificity for 2′-dG (dG, in red) is brought by a Watson–Crick interaction with C74 and a noncanonical interaction with C51.[127] M) c-di-GMP-I riboswitch from *V. cholerae*. The specificity for c-di-GMP is brought by a Watson–Crick pairing between the Gα C92 (Gα, in red) and an Hoogsteen interaction between Gβ (Gβ, in red) and G20.[39]

for nearly any RNA, the function of a riboswitch is intimately linked to its structure making the knowledge of the 3D structure of the RNA in complex with its ligand a prerequisite to the proper understanding at a molecular level of the amazing riboswitches selectivity.[34] Indeed, numerous works demonstrated that, in cell relevant conditions, aptamer domains are able to efficiently discriminate their cognate ligand from highly similar analogues.

The purine-sensing aptamer found upstream *xpt–pbux* operon of *B. subtilis* was the first riboswitch element whose crystal structure was determined in complex with a ligand.[35] Though rather small in size and simple in structure (Figure 1G), this first structure laid important bases in the study of riboswitches function. The aptamer is organized in three helices, two of which (P2 and P3) are closed by apical loops, joined together at a three-way junction (3WJ) made of unpaired stretches of nucleotides (joining nucleotides J1-2, J2-3, and J3-1 regions on Figure 1G,H). P2 and P3 are rather tolerant to sequence variations and are only constrained in their length.[31b,36a,b] On contrary, the nucleotides forming their apical loops are highly conserved and form a series of tight tertiary interactions important for ligand-binding pocket structuration by bringing and holding P2 and P3 together. Moreover, residues in unpaired regions of the 3WJ (J1-2, J2-3, and J3-1) are also highly conserved as they establish direct contacts with the ligand. In the guanine-responsive riboswitch, the first interaction is established between the Watson–Crick face of the guanine ligand and an invariant C74 residue (Figure 1J) a feature shared by other purine-responsive riboswitches such as pre-Q1-I,[37] 2′-dG-I,[38] or c-di-GMP-I,[39] just as a few examples (Figure 1J–M). This interaction was shown to be the main driver of aptamer specificity since mutating C74 to U74 was sufficient to render the molecule responsive to adenine.[40] Moreover, three other residues (U51, U47, and U22) establish specific contacts on the Hoogsteen and the sugar faces of the ligand. Finally, the guanine is stacked between base triples forming the floor and the ceiling of the ligand-binding pocket. This tight accommodation (≈98% of the ligand inaccessible to the solvent) explains not only the high specificity and affinity of the aptamer for guanine but also raises important questions. First, the complete enclosement of the ligand points that the ligand-binding pocket is, at least partly, unfolded and acquires its structure only upon guanine binding. Second, the very high affinity of some purine-binding aptamers for their ligand (e.g., the guanine aptamer has a K_D of 5 nM for its ligand, whereas c-di-GMP-I riboswitch aptamer has an amazing K_D of 10 pM for c-di-GMP)[15b,41] is somehow stunning as such molecules would always be saturated in physiological conditions. Yet, it is believed that this elevated affinity actually compensates for the short time some riboswitches have to make their decision by placing the molecule under kinetic control.[31b] Indeed, in the case of RNA devices controlling transcription through the formation of a Rho-independent terminator, ligand binding to the aptamer is in direct competition with the progression of the RNA polymerase (RNAP) and regulatory decision should be made before the RNAP leaves the regulatory region. Thought an aptamer displays a very high affinity for its ligand when the structure has reached its conformational equilibrium (the condition in which K_D values are usually determined), this value may not be relevant to the effective regulatory concentration required for a cotranscriptional regulation to take place and that may be significantly higher than K_D due to the aforementioned competition. Interestingly, it was also found that translation-regulating riboswitches tend to display K_D values closer to cellular ligand concentration, which is more in line with a thermodynamic control.

2.2.2. Aptamer Structural Diversity and Repurposing

In addition to the mechanistic considerations discussed earlier and applicable to any riboswitch, another important feature highlighted by purine-responsive riboswitches is the possibility of repurposing aptamer structure to sense new ligands.[31b,36a,b] Indeed, a simple point mutation ($C_{74}U$) allows to convert a guanine-responsive riboswitch into an adenine-responsive one.[40] Moreover, acquiring a handful of mutations was also found to convert guanine riboswitch into an RNA able to respond to 2′-deoxyguanosine (2′-dG) while preserving the same overall structural organization (Figure 1H,L).[38] Introducing more changes throughout the molecule (especially in the apical part of P2 and P3) while still maintaining an overall 3WJ-based structure allows to obtain riboswitches responding to the second messengers bis-(5′-3′)-cyclic dimeric guanosine monophosphate (c-di-GMP, Figure 1I,M),[41] and bis-(3′-3′)-cyclic dimeric guanosine adenosine monophosphate (cGAMP).[42] Interestingly, in both cases, a guanine moiety of the ligand is recognized through a direct Watson–Crick base-pairing with a highly conserved C residue (Figure 1M), a situation reminiscent to that of the guanine aptamer introduced earlier.

As discussed later, this possibility of reusing a scaffold and changing its specificity by introducing just a few mutations opens really exciting new possibilities to design and engineer new synthetic reprogramed riboswitches and biosensors. Structure repurposing is not restricted to purine aptamers and it was also found to occur within other families as typically exemplified by the ykkC riboswitches. After being an orphan for nearly 10 years, the ykkC motif was finally identified as a guanidine-responsive riboswitch.[20b] Yet, it was rapidly found that only a subset of this riboswitch family (ykkC-subtype 1, later renamed guanidine-I aptamer) actually responds to guanidine. ykkC-subtype 2 (nonresponsive to guanidine) riboswitches have the same overall structure as ykkC-subtype 1 but display subtle differences (e.g., the formation of an additional helix and some discrete point mutations). These small changes are sufficient to reprogram ykkC-subtype 2 specificity toward at least four new ligands, especially the ppGpp alarmone (ykkC-subtype 2a),[43] and PRPP (ykkC-subtype 2b),[44] an intermediate in purine synthesis. These two examples are really interesting since, at a first look, it may sound rather counterintuitive that such strongly electronegative phosphate-rich compounds can be specifically recognized by polyanionic RNA molecules. As before, an explanation to this dilemma came with the structural characterization of the aptamers in complex with their ligands. Indeed, two independent studies revealed that RNA performs this complicated task by making extensive use of Mg^{2+} divalent cations to mediate phosphate recognition.[45] Closely looking at ykkC-subtype 2a structure also revealed that ppGpp is specifically recognized through a Watson–Crick base-pairing between the guanine moiety and a conserved C residue and that simply mutating

this C in U was enough to change riboswitch specificity to ppApp,[45b] further demonstrating the ease with which riboswitches can be reprogrammed in comparison to their protein counterparts. The key role of Mg^{2+} in ligand recognition is not unique to ykkC-subtype 2 and is actually found in several other riboswitches, the most impressive example being the fluoride-responsive riboswitch crcB.[24] Indeed, highly electronegative fluoride anions have the unique capacity to establish strong interactions with three Mg^{2+} displayed at the heart of the riboswitch and stabilize its structure.[30a]

Finally, the great structural plasticity of RNA allows the molecule to adopt a wide range of folding and to find several solutions to the difficult task of specifically recognizing a small target molecule. As a consequence, ligands such as SAM,[46] guanidine,[47] or pre-Q1[30b] are recognized by two to six different classes of riboswitches (Table 1), each displaying a different overall structure and exploiting sometimes different recognition modes, suggesting that these motifs were reinvented several times over the course of evolution.

2.2.3. Gene Expression Switching by the Expression Platform

Whereas the sensing function is carried by the aptamer domain, the regulatory action is performed at the level of the expression platform upon transduction of the binding information. Since it is subjected to a lower evolutive selection pressure, the sequence of the expression platform is usually less conserved than the aptamer domain and may strongly vary from one riboswitch to the other, depending on the type of regulation involved.[48] In a large fraction of the known riboswitches, the information is transmitted *via* the so-called P1 helix that is found between the aptamer and the expression platform, and whose 3' arm is shared by another mutually exclusive secondary structure involved, for instance, in the control of transcription termination or translation initiation (**Figure 2**). Indeed, depending on the type of control (activation or repression), the P1 helix will be in equilibrium with the formation of a Rho-independent transcription antiterminator (ON switch, Figure 2A) or a terminator (OFF switch, Figure 2B), or a Shine–Dalgarno antisequestrator (ON switch, Figure 2C) or a sequestrator (OFF switch, Figure 2D). In this scheme, Gram-positive bacteria make more extensive use of transcriptional control, whereas gram-negative bacteria tend to rather use translational control.[48a] The formation of P1 is usually the rate-limiting step in aptamer folding and the point at which regulation decision is made.[49] The ligand binding is expected to take place within a more or less narrow time window depending on the type of regulation at work. Indeed, as highlighted earlier, transcriptional regulations are usually under kinetic control and the decision should be made within a short time window, whereas in translation regulation, the riboswitch is under thermodynamic control and has more time to make its decision. Yet, the situation might be slightly more complex, since the translation initiation control performed by some riboswitches was found to be further coupled with Rho-dependent transcription termination.[48a,50] Though a P1 stem is often found to act as a communication module (CM) in the region where the aptamer domain and expression platform overlap, not all riboswitches make use of it and some of them are even deprived of such CM as their expression platform is also part of the aptamer domain as recently shown for 4-amino-5-hydroxymethyl-2-methylpyrimidine diphosphat (HMP-PP)-responsive ThiS riboswitch.[51]

In addition to the widespread mechanisms introduced earlier, riboswitches were also found to control RNA stability either by exposing sequences targeted by a cellular RNase (e.g., RNaseE in *Escherichia coli*) (Figure 2E).[48c,52] Other regulation strategies involving riboswitch-controlled small RNAs sequestering, metabolite-assisted self-cleaving ribozymes (i.e., GlmS ribozyme),[53] or trans-acting riboswitches have also been described but will not be further discussed in this review.[48c,54] Finally, riboswitches are not restricted to prokaryotes since TPP-responsive riboswitches regulating intron splicing were also identified near splicing sites in plants and many fungal species.[55]

2.3. Complex Arrangement of Riboswitches and Logic Decision Making

Whereas the majority of 5'UTRs possess a single copy of the riboswitch, more complex regulatory arrangements were also discovered in which several riboswitches coexist within the same mRNA.[56] In the simplest case, two,[18] or three,[57] repeats of an aptamer domain targeting the same ligand are found in tandem. Each aptamer can behave independently and control its own expression platform (e.g., TPP tandem) allowing a more digital response to ligand recognition.[58] Alternatively, they can work in concert and regulate a single expression platform (e.g., glycine tandem)[18] allowing a cooperative binding of the ligand and a better dynamic range. By analogy with electronics, riboswitches may also be viewed as single input YES (ON riboswitches) or NOT (OFF riboswitches) logic gates.

In more complex regulation schemes, tandems are made of two different aptamer domains and tune gene expression in a Boolean logic manner. Such arrangements in two-input logic gates enable to adjust gene expression to two independent parameters and the response can be summarized in a truth table (**Figure 3**). Yet, whereas in electronics, logic gates adopt a digital (completely ON or OFF) behavior, riboswitch-based gates allow a smoother tuning of gene expression as they respond to ligands concentration (continuous variable). So far, three types of riboswitch-mediated Boolean regulation have been identified. When the activation of any riboswitch leads to gene expression repression, the tandem behaves as a NOR gate,[59] whereas in an AND gate, the activation of both riboswitches is required to stimulate gene expression.[43,60] Finally, in Firmicutes, a guanine and a PRPP riboswitches were found to work in tandem as an IMPLY gate in nutrient starvation conditions.[44] This arrangement allows the production of enzymes involved in IMP (inosine monophosphate) synthesis that, in turn, supports adenosine triphosphate (ATP) production during the stringent response. Interestingly, whereas at least eight different basic logic gates (i.e., AND, NAND, OR, NOR, XOR, XNOR, IMPLY, and NIMPLY) can theoretically be implemented, so far, only three of them were identified in bacteria. Altogether, this possibility offered by riboswitches to easily and rapidly conceive complex decision-making systems has opened wide avenues in synthetic biology, biocomputing, and bioengineering (see in the following section).

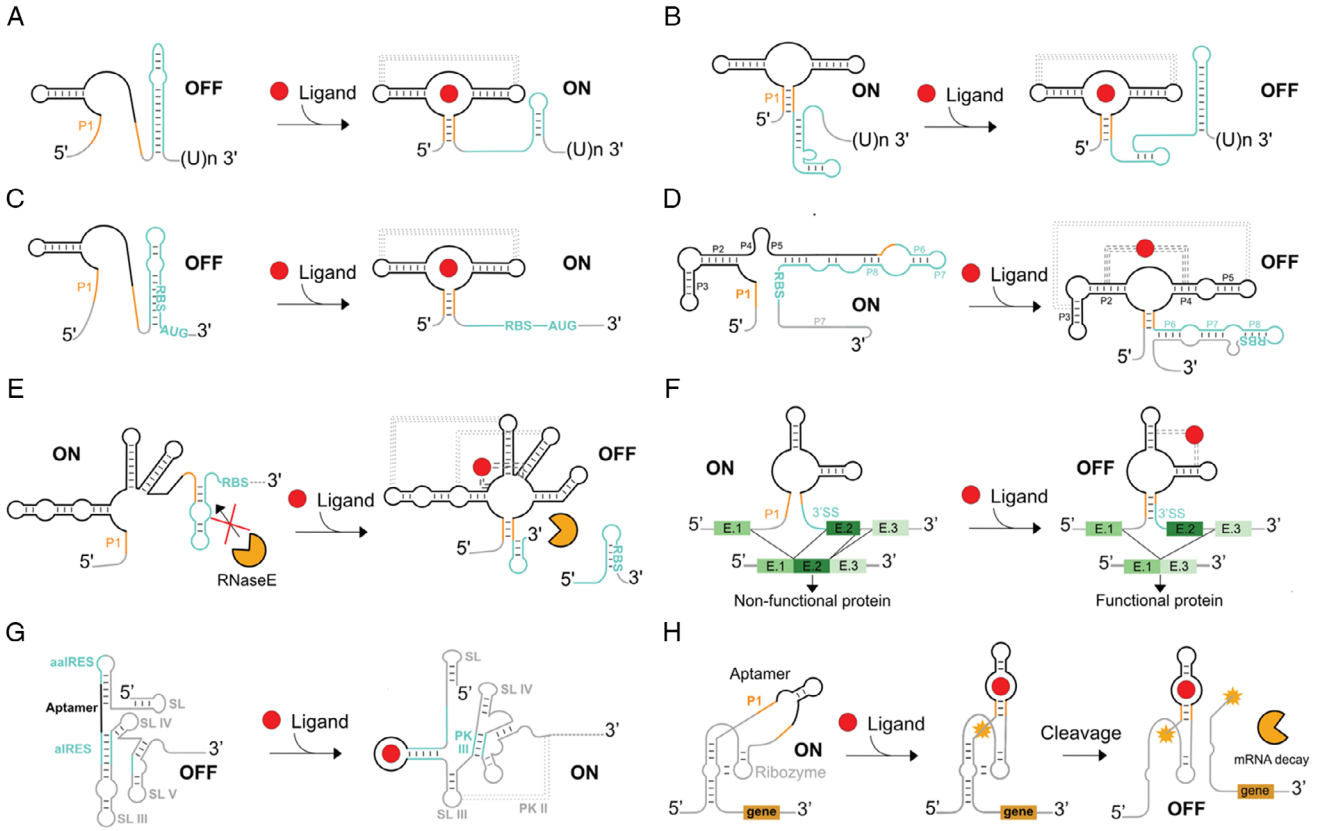


Figure 2. Examples of gene expression regulation mediated by riboswitches. In every case, the secondary structure of the riboswitch is schematized with the P1 helix shown in orange, the expression platform shown in blue, the sensing aptamer domains is shown in black, and long-range tertiary interactions represented by dotted lines. Ligands are represented by red dots and special interactions with the RNA by dashed lines. A,B) Transcription regulation using Rho-independent terminator. A) Transcription activation (ON switch) by the purine riboswitch. Upon ligand binding, a transcription antiterminator stem-loop is favored allowing transcription to further proceed.[40] B) Transcription repression (OFF switch) by purine riboswitch. Upon ligand binding, a Rho-independent terminator is favored, leading to premature transcription termination.[40] C,D) Translation regulation using RBS sequestration. C) Translation activation (ON switch) by the purine riboswitch.[40] Upon ligand binding, a antisequestrator is folding, leading to RBS exposure and translation initiation. D) Translation repression (OFF switch) by the TPP riboswitch. The binding of a ligand leads to the formation of a structure sequestering the RBS, therefore inhibiting ribosome recruitment.[128] E) Control of RNA stability by lysine riboswitch. Ligand-binging induces both RBS sequestration and the exposition of a sequence targeted by the cellular RNaseE leading to the RNA degradation.[52] F) Control of splicing by tetracycline-specific synthetic riboswitch. In absence of ligand, the 3′SS is accessible to the spliceosome leading to the inclusion of a synthetic exon 2 (E.2) into the mature mRNA and the production of a nonfunctional protein. On contrary, in presence of ligand, the 3′SS is kept inaccessible causing the skipping of the synthetic exon and the synthesis of a functional protein.[84] G) Control of translation initiation using artificial IRESs with an aptamer domain embedded. Whereas in the absence of ligand, the IRES is held partly unstructured, so inactive at recruiting the ribosome, the structure switching induced by the presence of the ligand restores IRES folding (i.e., formation of key stem-loops [SLs] and pseudoknot [PK] structures) and ribosome recruitment to initiate translation.[87] H) Control of mRNA stability by artificial aptazymes. An aptazyme is a self-cleaving ribozymes placed under the control of a structure-switching aptamer, so the allosteric of ligands. Although the ribozyme is maintained in an inactive conformation in the absence of ligand, the addition of ligand stabilizes aptamer folding which, in turn, stabilizes and reactivates ribozyme. Upon self-cleavage, the resulting cleaved mRNA is rapidly targeted for degradation by cellular RNase activities.[91]

3. Development and Use of Synthetic Riboswitches

The discovery of natural riboswitches was a breakthrough that opened exciting possibilities in bioengineering and synthetic biology by making it possible to precisely control gene expression in a protein-free manner, using very simple and portable RNA modules. These natural modules can be directly used not only in experiments aiming at improving the synthesis/secretion rate of natural compounds (e.g., amino acids) by microorganisms[61] but also to characterize biological pathways or even in the conception of biosensors.[62] Indeed, riboswitches have been widely used in metabolic engineering to regulate metabolic flow of microorganisms as recently and comprehensively reviewed.[63] Yet, they are less relevant for experiments aiming at reprogramming gene expression as their ligands are naturally present in cells and may compromise control accuracy, making it necessary to develop synthetic riboswitches responding to bio-orthogonal ligands (i.e., molecules absent from living systems). Ideally, such ligands should display functions allowing interactions with RNA,

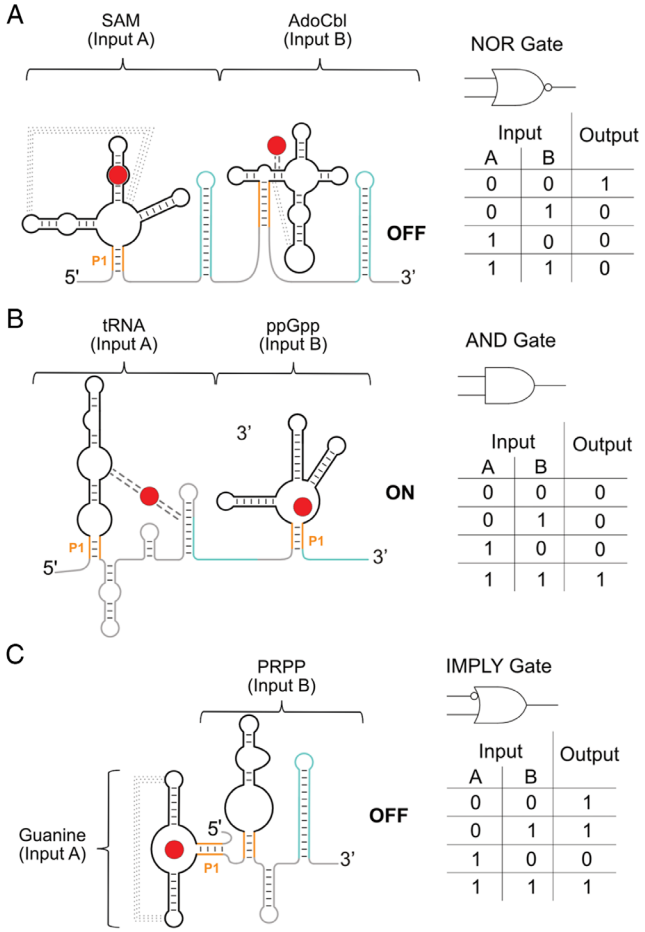


Figure 3. Tandem riboswitches allow to tune gene expression through Boolean logic operations. A–C) The 2D structure of each tandem riboswitch is shown on the left with the expression platforms (here, Rho-independent terminators) shown in blue, P1 helices shown in orange, the aptamer domains in black, and long-range tertiary interactions by dotted lines. Ligands are represented by red dots and special interactions with the RNA by dashed lines. The corresponding truth table is shown on the right, with "A" and "B" columns representing the first and the second input ligands respectively. A) SAM (left) and adenosyl-cobalamin (AdoCbl, right) riboswitches are each associated to their own independent expression platform allowing NOR logic operation to take place. B) Transfer RNA (left) and ppGpp (right) riboswitches are each associated to their own independent expression platform allowing AND logic operation to take place. C) Guanine (left) and PRPP (right) riboswitches share the same expression platform allowing IMPLY logic operation to take place.

be nontoxic to the cell and have a good cell membrane permeability. Interestingly, the potential of RNA to be applied for synthetic gene expression switching was already demonstrated several years before the first natural riboswitch was identified, by showing that inserting repeats of a synthetic aptamer in the 5′ UTR region of an mRNA can interfere with proper gene expression in the presence of the cognate ligand (i.e., an antibiotic or a dye).[64] Whereas the exact mechanism at work in this seminal study stayed poorly characterized at that time, it laid the foundation to a new field aiming at developing synthetic riboswitches.[65]

3.1. Development of Synthetic Structure-Switching RNA Aptamers

As for their natural counterparts, the aptamer domain is the core part of synthetic riboswitches as it carries the ligand recognition function. A first way of developing a synthetic aptamer, consists in reprogramming the specificity of a natural one toward a bio-orthogonal synthetic ligand. This was, for instance, nicely demonstrated by the successful conversion of an adenine-responsive RNA (three point mutations were introduced in the ligand-binding site) into riboswitches able to respond to synthetic heterocyclic molecules (e.g., azacytosine and pyrimido[4,5-*d*]pyrimidine-2,4-diamine, PPDA) and that lost affinity for their original ligand (i.e., adenine).[66] Yet, such a process may somehow be restricted to ligands with a structure sufficiently close to that of the natural one and its applicability to other natural riboswitches is left to be demonstrated.

An alternative approach consists in isolating new aptamer domains through in vitro selection/evolution prior to grafting them onto a natural or a synthetic expression platform. Synthetic aptamers targeting small molecules are typically isolated using the SELEX technology[67] in which large libraries of RNA molecules randomized over an extended region (25 to >50 nucleotide-long stretches) are challenged to interact with a ligand immobilized on a surface. Adjusting selection conditions (e.g., wash and elution) enables to identify dozens of synthetic aptamers displaying both great affinity and specificity for their target.[68] Yet, these criteria are not sufficient to make an aptamer a relevant building block for riboswitch design as elegantly illustrated by neomycin aptamers.[65b] Indeed, the most abundant aptamer (R23) identified at the end of a conventional SELEX,[69] also revealed to be completely inactive to control gene expression despite its great affinity and specificity for its target,[70] a phenomenon primarily attributed to the important preformation of the ligand-binding site and the absence of structure-switching element acting as a CM.[71] An additional limitation of using SELEX for isolating structure-switching aptamers comes with the necessity of immobilizing the ligand through a linker that may prevent the isolation of highly specific aptamers expected to be able to completely envelop the ligand similarly to what a natural riboswitch would do (see aforementioned). These strong limitations can be partly overcome by the use of Capture-SELEX, an in vitro selection strategy in which RNAs are selected for their capacity to switch their structure in the presence of a native unmodified ligand.[72]

A last issue faced by in vitro selection methods is linked to the limited chemical complexity of the selection medium. Indeed, a typical SELEX is performed in simple buffers whose pH, salt, and magnesium content are controlled, but in which no competitor is usually added. On the contrary, it is thought that the great specificity of natural riboswitches comes in part from their evolution in a rather complex environment in which a vast diversity of small molecules (especially close analogues of the target) coexist. This limitation can nevertheless be overcome by expressing and selecting aptamer in cellular environments. To this end, upon a few rounds of SELEX, the enriched library can be cloned into a vector under the control of a constitutive promoter and embedded within a simple expression platform placed upstream

a fluorescent reporting [e.g., green Fluorescent Protein-coding gene, (*gfp*)] or a selective marker (e.g., *tetA*) gene.[73] Growing cells alternatively in the presence and in the absence of the target ligand then allows to rapidly select and identify those constructs carrying the most relevant aptamer variants. Such a strategy allowed, for instance, the identification of aptamers able to efficiently switch their structure in the presence of neomycin,[70] and tetracycline that,[73a] together with theophylline-binding aptamer,[74] constitutes the few synthetic aptamers found to be well suited for the development of synthetic riboswitches, indicating that efforts are still necessary to devise robust selection strategies. Perhaps, a new solution may come with the finding that riboswitch function can be recapitulated in cell-free extracts,[75] offering a superior assay-to-assay reproducibility and a better control over reaction conditions. Though the cost of a large-scale screening (more than thousands variants) would rapidly become cost-prohibitive using conventional microtiter plates, the strong miniaturization made possible by the use of droplet-based microfluidics may represent a new efficient way of developing synthetic structure-switching aptamers, by making it possible to rapidly and functionally screen large sequence diversities as we recently demonstrated.[76]

3.2. Synthetic Riboswitches and Their Applications

Once an aptamer module has been developed, it can be associated with an expression platform in several ways. In the case of bacterial RNAs, the aptamer can be embedded within a rationally designed sequence that allows, for instance, the sequestration of the ribosome-binding site (RBS) upstream of a bacterial target gene in a ligand-dependent manner.[77] Alternatively, since the modularity of many natural riboswitches has now been experimentally demonstrated, the aptamer can also be combined with a natural expression platform.[78] Synthetic sequences exploiting natural concepts such as Rho-independent termination can also be appended to synthetic aptamers in a computer-assisted manner as recently done with different model molecules.[79] Finally, the expression platform can be developed using a selection strategy where,[80] for example, the region spacing the aptamer and the RBS (or the terminator stem-loop) is randomized and subjected to a dual genetic selection,[73b] or to a fluorescence-based screening.[73a] Once developed and implemented in living bacteria, such RNA devices can be used to reprogram cell behavior to make them able to sense and track small molecules,[81] change cell morphology,[66b] or even report on the presence of a target metabolite.[82]

In eukaryotes, even more appealing applications can be envisioned, but the implementation of such structure-switching RNA is also more challenging since, except for a few examples of TPP-responsive RNAs found in plants and fungi,[83] no natural riboswitch has been identified so far in mammals to guide the design of such devices. Inspired by the few natural cases, a synthetic RNA device was designed to modulate intron splicing in mammalian cells.[84] To do so, the Suess group interrupted the target gene with a synthetic exon (exon 2 on Figure 2F) bordered by a tetracycline-responsive aptamer on its 5′ side. In absence of tetracycline, the aptamer does not fold and leaves the 3′ splicing site (3′SS) accessible to the spliceosome. This leads to the inclusion of exon 2 into the mature mRNA and the production of a non-functional protein. On the contrary, addition of tetracycline stabilizes the structure of the aptamer that sequesters the 3′SS, causing the skipping of the synthetic exon and the synthesis of a functional protein. Inserting such a device into a suicide gene makes it possible, for instance, to control cell survival and might be of great interest in gene therapies.

Many other innovative strategies were proposed over the past decades.[65d,e,85] Initial attempts introduced simple structure-switching aptamers near the cap,[86] or the start codon of the messenger RNA.[73a] Upon interaction of the aptamer domain with its ligand, the complex formed a roadblock interfering with ribosome recruitment or mRNA scanning. In addition to conventional cap-dependent initiation, translation can also be initiated from highly structured RNA known as internal ribosome entry sites (IRESs) that attract the ribosome and prime it for translation. As the function of an IRES relies on its structure, placing the proper folding of a key element under the allosteric control of a structure-switching aptamer (Figure 2G) offers an alternative way of modulating ribosome recruitment.[87] Introducing such a construct upstream GFP-coding region allowed to control gene expression with various ligands (e.g., theophylline, FMN, or tetracycline) in wheat germ extracts in vitro expression systems, establishing the feasibility of the concept. In addition to direct control of translation initiation, regulation can also be mediated by miRNAs guiding the interference machinery to target mRNAs. Placing the release of a miRNA, or the accessibility of a miRNA-binding site of the mRNA, under the control of a structure-switch aptamer offers an efficient way of dosing gene expression by small molecules and represents another strategy to control transgene expression in gene therapy.[88]

Last but not least, aptazymes represent so far, the most explored and used strategy to control mammalian gene expression.[85,89] An aptazyme is basically a self-cleaving ribozyme (e.g., hammerhead, hepatitis delta virus, or twister ribozyme) whom a key helix is fused to and placed under the control of a structure-switching aptamer (Figure 2H), a concept already introduced by the Breaker lab 5 years before the discovery of natural riboswitches.[90] Introducing such a construct in 3′ untranslated region of a target mRNA allows to trigger its cleavage on-demand by adding the small molecule ligand which leads to the rapid degradation of the mRNA and gene extinction. First demonstrated in yeast, by the Smolke lab,[91] aptazymes allow the precise control of gene expression and are now used in synthetic biology programs aiming at generating yeast strains able to produce large amounts of natural plant products.[92] The concept was also recently applied in mice to control the expression of transgenes delivered by adeno-associated virus (AAV) and represents another strategy to assist gene therapy.[93] Finally, aptazymes were also recently exploited in plant science by showing that theophylline-controlled hammerhead ribozyme allows theophylline-modulated expression of *gfp* in *Arabidopsis thaliana*.[94] Though aptazymes can be optimized using genetic screens introduced before, the recent development of high-throughput methodologies exploiting next generation sequencing coupled,[95] or not,[96] with cell sorting cytometry now allows the rapid identification of optimized sequences from large variant libraries. Finally, whereas most of the examples reported so far used aptazymes controlled by a single ligand, the modification of two

helices of the ribozyme domain by two different structure-switching aptamers turns the whole construct into a two-input logic gate able to perform Boolean computations (AND, OR, NAND, or AND NOT),[97] further expanding the set of operations already identified in living organisms and discussed earlier.

Aptazymes can be seen as simple molecular machines activated by an external input signal. Yet, much more complex machines can be designed by embedding structure-switching aptamers within supramolecular assemblies as those typically developed in nucleic acids nanotechnology and nicely reviewed in previous studies.[98] Moreover, higher complexity may arise from the integration of structure-switching aptamers into molecular circuits such as those used in biocomputing and enzyme-free amplification biosensing circuits as discussed by others.[99]

4. Sensing of Small Molecules

Beyond their direct use for controlling gene expression, riboswitches can also be harnessed for small molecule detection. Indeed, except for ions, it is quite challenging to synthesize specific fluorogenic (or chromogenic) chemical reporters of small molecules whose detection usually requires the use of rather large pieces of equipment. Instead, the uses of genetically encoded systems in which a biological polymer (i.e., protein or RNA) both senses and reports on the presence of the small molecule are really attractive as they are not invasive, compatible with live-cell (and even single-cell) analyses and they can display very high specificity. Though cell-based biosensors have obvious interests,[82] these systems can be strongly simplified and better controlled using a cell-free approach. For example, this was recently achieved by expressing, in cell-free extracts, constructs in which a protein-coding reporter gene was placed under the control of the fluoride riboswitch crcB. This simple synthetic molecular circuit was then validated for on-site detection of trace amounts of fluoride contaminants in environmental samples.[100] An even simpler small molecule-sensing technology called SPRINT was also recently introduced in which a riboswitch is used to control the transcription of a downstream tag sequence later quantified by the Cas13a-mediated technology SHERLOCK.[101] Yet, even though these methods are sensitive, their readout is rather indirect, a limitation that can be overcome by using more direct RNA fluorogenic reporters. No naturally fluorescent RNA molecule has been discovered so far. Nevertheless, several synthetic light-up RNA aptamers have been developed by SELEX and microfluidic-assisted functional screening over the past decade, opening new perspectives in RNA biology.[102] These RNAs are able to specifically bind and strongly activate the fluorescence of small fluorogenic molecules (also called fluorogens) to form a fluorescent complex. Interestingly, one of these systems (i.e., Riboglow) was even developed using the natural cobalamin riboswitch aptamer as a starting point.[103]

4.1. Development of Direct RNA-Based Fluorogenic Biosensors

Although a light-up aptamer/fluorogen complex is always fluorescent, the system can be converted into a fluorogenic RNA-based biosensor (FRB) emitting a fluorescence only in the presence of a target ligand. To do so, a sensing aptamer is fused to a transiently destabilized light-up RNA (**Figure 4**). In the absence of target ligand, the light-up moiety is unable to bind its cognate fluorogen. However, the binding of the target analyte to the sensing domain triggers a structure switching that ultimately restores the fluorogen-binding capacity of the reporting domain and leads to the formation of a fluorescent complex. The Jaffrey lab described a first strategy mimicking a natural situation in which a riboswitch (e.g., TPP, guanine, or adenine) is entirely repurposed, with the aptamer domain sensing the target molecule and the expression platform controlling the proper folding of the light-up aptamer Spinach (Figure 4A).[104] Yet, most of the FRB developed so far, exploit a principle first described with the malachite green (MG)-binding aptamer.[105] This pioneer work demonstrated that a light-up aptamer can be converted into an FRB by directly connecting it to a structure-switching sensing aptamer *via* a transiently destabilized helix (Figure 4B) acting as a CM (also called transducer, transmitter, or modulation sequence).[106] Up to now, the majority of FRB are based on DFHBI-binding aptamers (i.e., Spinach, Spinach2, iSpinach, or Broccoli)[107] connected to natural (e.g., SAM, guanine, glycine, etc.) or synthetic sensing aptamers, allowing the specific detection of a wide range of ligands (**Table 2**). Moreover, some natural aptamers were also reprogrammed either by rational design (e.g., a single point mutation extended the specificity of the cyclic-di-GMP aptamer toward cyclic-Guanosine monophosphate-Adenosine MonoPhosphate [cGAMP]),[108] or by more complex in vitro selection approaches. For instance, randomizing the 3WJ region of natural aptamers (e.g., xpt–pbuX aptamer) while preserving their secondary and tertiary structure allowed preparing libraries from which aptamers with specificity toward new ligands (e.g., 5HTP or L-

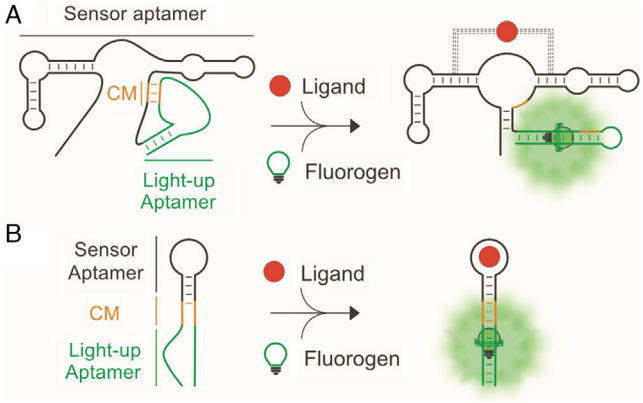


Figure 4. Engineering strategies to develop fluorogenic RNA-based biosensor (FRB). A) TPP riboswitch-based biosensors. A transiently destabilized version of the light-up RNA aptamer Spinach (green) was fused to a TPP-sensing aptamer (black) *via* a CM (orange) mimicking a natural riboswitch expression platform. Upon ligand binding, TPP aptamer switches its conformation and stabilizes the light-up aptamer structure fluorogen binding and fluorescence emission. B) The light-up RNA aptamer Spinach (green) can also be converted into an FRB by directly connecting a transiently destabilized form of Spinach to a structure-switching sensing aptamer (black) *via* a CM (orange). As before, ligand binding leads to a switch of the sensing aptamer conformation transmitted to the light-up aptamer *via* the CM and leading to a fluorescence emission.

Table 2. Fluorogenic RNA-based biosensors and their applications.

Ligand	Sensing aptamer	N/S[a)]	Light-up	Claimed application	Ref.
ATP, FMN, theophylline	ATP, FMN, or theophylline-binding aptamer	S	Malachite green	In vitro sensing	[105]
Adenosine 5-diphosphate (ADP)	ADP-binding aptamer	S	Spinach	Live-cell imaging	[111]
S-Adenosylmethionine (SAM)	SAM-binding aptamer	S	Spinach	Live-cell imaging	[111]
Adenosine, guanine, SAM, GTP	Adenosine, guanine, SAM, or GTP-binding aptamer	S	Spinach	In vitro sensing	[111]
Adenosine 5-diphosphate (ADP)	ADP-binding aptamer	S	Spinach	Live-cell imaging	[158]
Cyclic di-GMP (c-di-GMP)	GEMM-I aptamer	N	Spinach	In vitro sensing	[159]
Cyclic di-GMP (c-di-GMP)	GEMM-I aptamer	N	Spinach	Live-cell imaging	[108]
c-AMP-GMP, c-di-GMP	G20A-GEMM-I aptamer	N/S	Spinach	Live-cell imaging to monitor production of c-AMP-GMP by DncV enzyme	[108]
cyclic di-GMP (c-di-GMP)	GEMM-I aptamer	N	Spinach2	Live-cell imaging under both aerobic and anaerobic conditions	[112]
2′,3′-cGAMP	G103A GEMM-II aptamer	N/S	Spinach2	High-throughput analysis of the cGAS-cGAMP-STING pathway	[118a]
Cyclic AMP-GMP (3′-5′ cGAMP)	GEMM-Ib aptamer	N	Spinach	Riboswitch discovery and visualization of cAG signaling in living bacteria	[113a]
Cyclic di-AMP (cdiA)	yuaA aptamer	N	Spinach2	Flow cytometry analysis of diadenylate cyclase activity	[160]
3'-5'-cyclic AMP (cAMP)	class II cAMP-binding aptamer	S	Spinach	Quantitative imaging of cAMP in compartments	[161]
Thiamine-pyrophosphate (TPP)	Thi-box aptamer	N	Spinach	Live-cell imaging and identification of agonists and antagonists	[104]
SAM, guanine, adenine	SAM-I, purine aptamer	N	Spinach	In vitro sensing	[104]
Theophylline, guanine, c-di-GMP, SAM	Theophylline, guanine, c-di-GMP or SAM-binding aptamer	S/N	iSpinach	In vitro sensing	[76]
Nickel (Ni^{2+}), cobalt (Co^{2+}), iron (Fe^{2+})	czcD (NiCo) aptamer	N	Spinach2	Identification of candidate ligands for czcD riboswitch in vitro and in vivo	[121]
Glycine	Tandem glycine aptamer	N	Spinach	In vitro high-throughput screening	[162]
S-adenosyl-l-homocysteine (SAH)	SAH aptamer	N	Spinach2	In vitro monitoring of methyl transferase activity and live-cell detection of SAH	[118b]
SAM	SAM-III aptamer	N	Corn	Quantification of SAM in vitro and in living mammalian cell	[163]
5-hydroxytryptophan, 3,4-dihydroxyphenylalanine	3WJ aptamers	N/S	Broccoli	Live-cell imaging	[109]
SAM	SAM aptamer	N	Red Broccoli	SAM imaging in living mammalian cells	[164]
c-di-GMP, Tetracycline	GEMM-I and Tetracycline-binding aptamer	S/N	DNB/Broccoli	Live-cell imaging	[118]
Tyrosine, phenylalanine, tryptophane	Tyrosine, phenylalanine, tryptophane-binding aptamer	S	Spinach2	RAPID: high-throughput screening of living cells for secretory phenotypes	[122]

[a)]N or S, respectively, stands for natural or synthetic sensing aptamer.

DOPA neurotransmitters) emerged upon in vitro selection.[109] This demonstrates that, as anticipated earlier, structural scaffold widely exploited by natural riboswitches can serve as starting point in the development of new aptamers. Finally, FRET-based FRB can also be designed by combining a structure-switching aptamer (e.g., SAM aptamer) with two light-up RNAs forming a FRET pair (e.g., Spinach/DFHBI and Mango/YO3).[110] In this scheme, the structure switching triggered by ligand binding changes the relative position of both light-up aptamers and induces FRET fluorescence emission.

The CM is the instrumental part of FRBs as it dictates response amplitude (i.e., the ratio of the fluorescence measured in the presence over that measured in absence of ligand).[111] As the sequence of the optimal CM is directly influenced by that of the connected aptamers, a universal solution is not expected to exist, which implies that the module may have to be reoptimized

for each new sensor. A first solution consists in using the natural P1 stem of the riboswitch from which the aptamer has been collected.[108,112] Interestingly, preserving the natural environment of the aptamer minimizes the risk of structural perturbation and the loss of affinity for the ligand. Yet, these natural elements may be limited in their activation capacity. Therefore, artificial CM can also be rationally designed using thermodynamic parameters (e.g., free energy values) as a guide through a trial-and-error strategy in which a small number of sequences is analyzed.[111] Alternatively, optimal CM can be identified by testing every possible sequence permutation using an ultrahigh-throughput screening approach such as the microfluidic-assisted in vitro compartmentalization coupled with high-throughput sequencing (μIVC-seq) strategy we recently introduced and that enables to functionally analyze millions of variants in a single experiment.[76]

4.2. Applications of RNA-Based Fluorogenic Biosensors

A great advantage of FRB modules is the possibility to genetically encode them, making it possible to devise simple and noninvasive biosensors allowing the real-time monitoring of cell content in a target small molecule (e.g., SAM, TPP, etc.) by live-cell imaging.[104,111] Moreover, these sensors can be used in flow cytometry screening to characterize the variation of small molecule concentration and its impact on enzyme dynamics in different conditions and with single-cell resolution. For instance, this strategy allowed to monitor the impact of Zn^{2+} treatment on the activity of diguanylate cyclase DgcZ and on the cellular content in cyclic-di-GMP.[113] In a different way, the use of cGAMP-specific FRPs enabled the identification of the enzymes responsible for the synthesis of the ligand,[114] demonstrating the fantastic potential of these biosensors in shedding light on poorly characterized biological mechanisms. A potential limitation of RNA devices is linked to their degradation by endogenous RNase activities. Nevertheless, inserting the molecule into a stable RNA (e.g., F30) acting as a scaffold,[115] or circularizing it,[116] enables to significantly increase sensor half-life in bacteria and mammalian cells, respectively. In addition to assisting the proper folding of an FRB and protecting it from RNase-mediated degradation, 3WJ-containing scaffolds such as F30 may also serve as an assembly platform to design more complex RNA devices. For example, connecting two different FRBs to the same scaffold turns the construct into a two-input logic gate able to perform Boolean (e.g., "OR") operations.[117] In a different way, associating a light-up RNA aptamer to an FRB *via* the scaffold yields a ratiometric biosensor allowing to normalize sensor fluorescence to its concentration, leading to a gain of sensitivity.[118]

FRBs can also have plenty of in vitro applications. Just as a few examples, FRBs specific of S-adenosyl homocysteine (SAH) and 2′,3′-cGAMP were designed and used to monitor the activity of methyltransferase and cytosolic DNA sensor cGAMP synthase (cGAS) enzymes, respectively, in a high-throughput screening compatible manner.[119] Moreover, FRBs can be used to identify natural and alternative ligands of natural riboswitches. Indeed, fusing the aptamer domain of the czcD riboswitch (also known as NiCo)[120] to Spinach2 light-up aptamer yielded an FRB whom the specificity can be tested. Surprisingly, this work demonstrated that the best aptamer ligand was neither Ni^{2+} nor Co^{2+} as initially proposed but rather Fe^{2+} ions, a result in better agreement with the known biochemistry of transition metal ions in the cell.[121] FRBs can also be used to identify new ligands (agonist and antagonist) of already well-characterized riboswitches (e.g., TPP) in view of discovering new potential antimicrobial drugs.[104] Finally, the capacity of FRBs to report in real time on the presence of extracellular ligands can also be exploited to screen microorganism libraries to search for cells with improved capacity to synthesize and secrete compounds of industrial interest. Using droplet-based microfluidics, the Abate lab showed that variants of a mutant yeast library can be individualized within picoliter water-in-oil droplets together with an FRB and that those cells with an improved synthesis/secretion capacity of the target metabolite can be rapidly identified and isolated.[122]

5. Conclusion

Since its original description, RNA has gained an ever-increasing place in molecular and cell biology and is seen today as an instrumental molecule that orchestrates gene expression by acting at a structural, functional, and regulatory level. Among the variety of RNA types currently known, the riboswitches introduced and discussed in this review are fascinating in many aspects. First, they are a window on our remote past as they are thought to be remnants of primitive living systems while life was evolving in the RNA World. Second, riboswitches often control key bacterial metabolic or adaptation pathways and their aptamer domains are highly conserved both in structure and sequence throughout many genera. In this view, they constitute an ideal drug target since it is very difficult for them to acquire a mutation bringing resistance without compromising their function. Moreover, given that several copies of the same aptamer can be present in the genome, it is unlikely that every copy will acquire a mutation, a prerequisite to the appearance of a resistant cell. Therefore, extensive efforts are now turned to riboswitches for the development of new antibiotics as exemplified by Ribocil, a molecule targeting FMN riboswitch.[123] Moreover, riboswitches, aptazymes, and biosensors also hold great promise as new therapeutics or to assist their development. Yet, this vast field is out of the scope of this review and the reader is redirected to several excellent reviews on the topic.[124] Third, riboswitches taught us a lot about how RNA folding can bring complex functions through the precise and specific recognition of small molecules contained in a sea of close analogues. Structural characterizations are instrumental to the proper comprehension of the recognition mechanisms at work and also give important clues on how to engineer these molecules to endow them with new properties/specificities. Finally, as largely developed in this review, riboswitches inspire or even serve as starting points in the development of many artificial structure switching aptamers with applications as diverse as transgene expression control, the optimization of microorganisms, or the development of biosensors targeting a wide array of ligands. Given the estimation of the number of riboswitches that may be left to be discovered together with the possibility of developing new synthetic molecules, it is likely that the field is still in its infancy and that many more exciting applications will appear in the near future.

Acknowledgements

C.H. and N.D. contributed equally to this work. This work has been published under the framework of the LabEx NetRNA ANR-10-LABX-0036 and of ANR-17-EURE-0023, as funding from the state managed by Agence Nationale de la Recherche (ANR) as part of the investments for the future program. This work was also supported by the ANR program Ribofluidix (ANR-17-CE12-0025-02), the "Université de Strasbourg," and the "Centre National de la Recherche Scientifique" (CNRS). The authors wish to dedicate this article to the memory of Jean-Pascal RYCKELYNCK who suddenly left them while this article was under revision.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

biosensing, gene expression regulation, riboswitches, RNA aptamers, synthetic biology

Received: November 12, 2020
Revised: December 22, 2020
Published online: January 25, 2021

[1] F. Crick, *Nature* **1970**, *227*, 561.
[2] T. Bratkovic, J. Bozic, B. Rogelj, *Nucleic Acids Res.* **2020**, *48*, 1627.
[3] P. Fernandez-Millan, C. Schelcher, J. Chihade, B. Masquida, P. Giege, C. Sauter, *Arch. Biochem. Biophys.* **2016**, *602*, 95.
[4] a) H. F. Noller, *Science* **2005**, *309*, 1508; b) C. L. Will, R. Luhrmann, *Cold Spring Harb. Perspect. Biol.* **2011**, *3*, a003707.
[5] a) T. R. Cech, A. J. Zaug, P. J. Grabowski, *Cell* **1981**, *27*, 487; b) C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, S. Altman, *Cell* **1983**, *35*, 849.
[6] M. W. Gray, V. Gopalan, *J. Biol. Chem.* **2020**, *295*, 2313.
[7] a) J. A. Doudna, T. R. Cech, *Nature* **2002**, *418*, 222; b) J. F. Atkins, R. F. Gesteland, T. Cech, in *RNA Worlds: From Life's Origins to Diversity in Gene Regulation*, Cold Spring Harbor Laboratory Press, New York, NY **2011**; c) T. J. Wilson, Y. Liu, D. M. J. Lilley, *Front. Chem. Sci. Eng.* **2016**, *10*, 178; d) S. Muller, *Molecules* **2017**, *22*, 789; e) C. E. Weinberg, Z. Weinberg, C. Hammann, *Nucleic Acids Res.* **2019**, *47*, 9480.
[8] W. Gilbert, *Nature* **1986**, *319*, 618.
[9] a) T. R. Mercer, M. E. Dinger, J. S. Mattick, *Nat. Rev. Genet.* **2009**, *10*, 155; b) E. G. Wagner, P. Romby, *Adv. Genet.* **2015**, *90*, 133; c) D. P. Bartel, *Cell* **2018**, *173*, 20.
[10] J. Kortmann, F. Narberhaus, *Nat. Rev. Microbiol.* **2012**, *10*, 255.
[11] J. Zhang, *Wiley Interdiscip. Rev. RNA* **2020**, *11*, e1600.
[12] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, R. R. Breaker, *Chem. Biol.* **2002**, *9*, 1043.
[13] a) K. F. Blount, R. R. Breaker, *Nat. Biotechnol.* **2006**, *24*, 1558; b) R. R. Breaker, *Future Microbiol.* **2009**, *4*, 771; c) K. E. Deigan, A. R. Ferre-D'Amare, *Acc. Chem. Res.* **2011**, *44*, 1329.
[14] a) A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A. Perumov, E. Nudler, *Cell* **2002**, *111*, 747; b) W. Winkler, A. Nahvi, R. R. Breaker, *Nature* **2002**, *419*, 952; c) W. C. Winkler, S. Cohen-Chalamish, R. R. Breaker, *Proc. Natl. Acad. Sci.* **2002**, *99*, 15908.
[15] a) V. Epshtein, A. S. Mironov, E. Nudler, *Proc. Natl. Acad. Sci.* **2003**, *100*, 5052; b) M. Mandal, B. Boese, J. E. Barrick, W. C. Winkler, R. R. Breaker, *Cell* **2003**, *113*, 577.
[16] a) R. R. Breaker, *Cold Spring Harb. Perspect. Biol.* **2012**, *4*, a003566; b) J. W. Nelson, R. R. Breaker, *Sci. Signal.* **2017**, *10*, eaam8812.
[17] a) J. E. Barrick, K. A. Corbino, W. C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J. K. Wickiser, R. R. Breaker, *Proc. Natl. Acad. Sci.* **2004**, *101*, 6421; b) Z. Weinberg, J. E. Barrick, Z. Yao, A. Roth, J. N. Kim, J. Gore, J. X. Wang, E. R. Lee, K. F. Block, N. Sudarsan, S. Neph, M. Tompa, W. L. Ruzzo, R. R. Breaker, *Nucleic Acids Res.* **2007**, *35*, 4809; c) Z. Weinberg, J. X. Wang, J. Bogue, J. Yang, K. Corbino, R. H. Moy, R. R. Breaker, *Genome Biol.* **2010**, *11*, R31; d) Z. Weinberg, C. E. Lunse, K. A. Corbino, T. D. Ames, J. W. Nelson, A. Roth, K. R. Perkins, M. E. Sherlock, R. R. Breaker, *Nucleic Acids Res.* **2017**, *45*, 10811; e) Z. Weinberg, J. W. Nelson, C. E. Lunse, M. E. Sherlock, R. R. Breaker, *Proc. Natl. Acad. Sci.* **2017**, *114*, E2077.
[18] M. Mandal, M. Lee, J. E. Barrick, Z. Weinberg, G. M. Emilsson, W. L. Ruzzo, R. R. Breaker, *Science* **2004**, *306*, 275.
[19] M. E. Sherlock, R. R. Breaker, *RNA* **2020**, *26*, 675.
[20] a) R. R. Breaker, R. M. Atilho, S. N. Malkowski, J. W. Nelson, M. E. Sherlock, *Biochemistry* **2017**, *56*, 345; b) J. W. Nelson, R. M. Atilho, M. E. Sherlock, R. B. Stockbridge, R. R. Breaker, *Mol. Cell* **2017**, *65*, 220.
[21] a) P. J. McCown, K. A. Corbino, S. Stav, M. E. Sherlock, R. R. Breaker, *RNA* **2017**, *23*, 995; b) N. Pavlova, D. Kaloudas, R. Penchovsky, *Gene* **2019**, *708*, 38.
[22] M. E. J. Newman, *Contemporary Physics* **2005**, *46*, 323.
[23] T. D. Ames, R. R. Breaker, *The Chemical Biology of Nucleic Acids* (Ed: G. Mayer), Wiley Online Books, Chichester, UK **2010**, p. 433.
[24] J. L. Baker, N. Sudarsan, Z. Weinberg, A. Roth, R. B. Stockbridge, R. R. Breaker, *Science* **2012**, *335*, 233.
[25] M. E. Sherlock, S. N. Malkowski, R. R. Breaker, *Biochemistry* **2017**, *56*, 352.
[26] A. Roth, W. C. Winkler, E. E. Regulski, B. W. Lee, J. Lim, I. Jona, J. E. Barrick, A. Ritwik, J. N. Kim, R. Welz, D. Iwata-Reuyl, R. R. Breaker, *Nat. Struct. Mol. Biol.* **2007**, *14*, 308.
[27] a) M. J. Cromie, Y. Shi, T. Latifi, E. A. Groisman, *Cell* **2006**, *125*, 71; b) C. E. Dann 3rd, C. A. Wakeman, C. L. Sieling, S. C. Baker, I. Irnov, W. C. Winkler, *Cell* **2007**, *130*, 878.
[28] N. Sudarsan, J. K. Wickiser, S. Nakamura, M. S. Ebert, R. R. Breaker, *Genes Dev.* **2003**, *17*, 2688.
[29] a) L. Huang, J. Wang, D. M. J. Lilley, *Cell Chem. Biol.* **2017**, *24*, 695; b) C. W. Reiss, S. A. Strobel, *RNA* **2017**, *23*, 1338.
[30] a) A. Ren, K. R. Rajashankar, D. J. Patel, *Nature* **2012**, *486*, 85; b) C. D. Eichhorn, M. Kang, J. Feigon, *Biochim. Biophys. Acta* **2014**, *1839*, 939.
[31] a) A. Serganov, A. Polonskaia, A. T. Phan, R. R. Breaker, D. J. Patel, *Nature* **2006**, *441*, 1167; b) R. T. Batey, *Q. Rev. Biophys.* **2012**, *45*, 345.
[32] I. R. Price, A. Gaballa, F. Ding, J. D. Helmann, A. Ke, *Mol. Cell* **2015**, *57*, 1110.
[33] A. D. Garst, A. Heroux, R. P. Rambo, R. T. Batey, *J. Biol. Chem.* **2008**, *283*, 22347.
[34] a) A. D. Garst, A. L. Edwards, R. T. Batey, *Cold Spring Harb. Perspect. Biol.* **2011**, *3*, a003533; b) M. M. Meyer, *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1370.
[35] R. T. Batey, S. D. Gilbert, R. K. Montange, *Nature* **2004**, *432*, 411.
[36] a) J. N. Kim, R. R. Breaker, *Biol. Cell* **2008**, *100*, 1; b) E. B. Porter, J. G. Marcano-Velazquez, R. T. Batey, *Biochim. Biophys. Acta* **2014**, *1839*, 919.
[37] a) D. J. Klein, T. E. Edwards, A. R. Ferre-D'Amare, *Nat. Struct. Mol. Biol.* **2009**, *16*, 343; b) R. C. Spitale, A. T. Torelli, J. Krucinska, V. Bandarian, J. E. Wedekind, *J. Biol. Chem.* **2009**, *284*, 11012.
[38] A. L. Edwards, R. T. Batey, *J. Mol. Biol.* **2009**, *385*, 938.

[39] N. Sudarsan, E. R. Lee, Z. Weinberg, R. H. Moy, J. N. Kim, K. H. Link, R. R. Breaker, *Science* **2008**, *321*, 411.
[40] A. Serganov, Y. R. Yuan, O. Pikovskaya, A. Polonskaia, L. Malinina, A. T. Phan, C. Hobartner, R. Micura, R. R. Breaker, D. J. Patel, *Chem. Biol.* **2004**, *11*, 1729.
[41] K. D. Smith, S. V. Lipchock, T. D. Ames, J. Wang, R. R. Breaker, S. A. Strobel, *Nat. Struct. Mol. Biol.* **2009**, *16*, 1218.
[42] A. Ren, X. C. Wang, C. A. Kellenberger, K. R. Rajashankar, R. A. Jones, M. C. Hammond, D. J. Patel, *Cell Rep.* **2015**, *11*, 1.
[43] M. E. Sherlock, N. Sudarsan, R. R. Breaker, *Proc. Natl. Acad. Sci.* **2018**, *115*, 6052.
[44] M. E. Sherlock, N. Sudarsan, S. Stav, R. R. Breaker, *Elife* **2018**, *7*, e33908.
[45] a) A. J. Knappenberger, C. W. Reiss, S. A. Strobel, *Elife* **2018**, *7*, e36381; b) A. Peselis, A. Serganov, *Nat. Chem. Biol.* **2018**, *14*, 887.
[46] R. T. Batey, *Wiley Interdiscip. Rev. RNA* **2011**, *2*, 299.
[47] R. A. Battaglia, A. Ke, *Wiley Interdiscip. Rev. RNA* **2018**, *9*, e1482.
[48] a) W. C. Winkler, R. R. Breaker, *Annu. Rev. Microbiol.* **2005**, *59*, 487; b) R. R. Breaker, *Cold Spring Harb. Perspect. Biol.* **2018**, *10*, a032797; c) A. V. Bedard, E. D. M. Hien, D. A. Lafontaine, *Biochim. Biophys. Acta Gene Regul. Mech.* **2020**, *1863*, 194501.
[49] F. Aboul-Ela, W. Huang, M. Abd Elrahman, V. Boyapati, P. Li, *Wiley Interdiscip. Rev. RNA* **2015**, *6*, 631.
[50] K. Hollands, S. Proshkin, S. Sklyarova, V. Epshtein, A. Mironov, E. Nudler, E. A. Groisman, *Proc. Natl. Acad. Sci.* **2012**, *109*, 5376.
[51] R. M. Atilho, G. Mirihana Arachchilage, E. B. Greenlee, K. M. Knecht, R. R. Breaker, *Elife* **2019**, *8*, e45210.
[52] M. P. Caron, L. Bastet, A. Lussier, M. Simoneau-Roy, E. Masse, D. A. Lafontaine, *Proc. Natl. Acad. Sci.* **2012**, *109*, E3444.
[53] A. R. Ferre-D'Amare, *Q. Rev. Biophys.* **2010**, *43*, 423.
[54] J. R. Mellin, P. Cossart, *Trends Genet.* **2015**, *31*, 150.
[55] a) M. T. Cheah, A. Wachter, N. Sudarsan, R. R. Breaker, *Nature* **2007**, *447*, 497; b) S. Li, R. R. Breaker, *Nucleic Acids Res.* **2013**, *41*, 3022.
[56] R. R. Breaker, *Mol. Cell* **2011**, *43*, 867.
[57] H. Zhou, C. Zheng, J. Su, B. Chen, Y. Fu, Y. Xie, Q. Tang, S. H. Chou, J. He, *Sci. Rep.* **2016**, *6*, 20871.
[58] R. Welz, R. R. Breaker, *RNA* **2007**, *13*, 573.
[59] N. Sudarsan, M. C. Hammond, K. F. Block, R. Welz, J. E. Barrick, A. Roth, R. R. Breaker, *Science* **2006**, *314*, 300.
[60] A. G. Chen, N. Sudarsan, R. R. Breaker, *RNA* **2011**, *17*, 1967.
[61] J. Yang, S. W. Seo, S. Jang, S. I. Shin, C. H. Lim, T. Y. Roh, G. Y. Jung, *Nat. Commun.* **2013**, *4*, 1413.
[62] Z. Ghazi, S. Jahanshahi, Y. Li, *PLoS One* **2017**, *12*, e0188399.
[63] S. G. Kim, M. H. Noh, H. G. Lim, S. Jang, S. Jang, M. A. G. Koffas, G. Y. Jung, *FEMS Microbiol Lett* **2018**, *365*, fny187.
[64] G. Werstuck, M. R. Green, *Science* **1998**, *282*, 296.
[65] a) F. Groher, B. Suess, *Biochim. Biophys. Acta* **2014**, *1839*, 964; b) C. Berens, B. Suess, *Curr. Opin. Biotechnol.* **2015**, *31*, 10; c) M. Etzel, M. Morl, *Biochemistry* **2017**, *56*, 1181; d) C. M. Schmidt, C. D. Smolke, *Cold Spring Harb. Perspect. Biol.* **2019**, *11*, a032532; e) M. Sporing, M. Finke, J. S. Hartig, *Curr. Opin. Biotechnol.* **2020**, *63*, 34.
[66] a) N. Dixon, J. N. Duncan, T. Geerlings, M. S. Dunstan, J. E. McCarthy, D. Leys, J. Micklefield, *Proc. Natl. Acad. Sci.* **2010**, *107*, 2830; b) C. J. Robinson, H. A. Vincent, M. C. Wu, P. T. Lowe, M. S. Dunstan, D. Leys, J. Micklefield, *J. Am. Chem. Soc.* **2014**, *136*, 10615.
[67] a) A. D. Ellington, J. W. Szostak, *Nature* **1990**, *346*, 818; b) C. Tuerk, L. Gold, *Science* **1990**, *249*, 505.
[68] F. Pfeiffer, G. Mayer, *Front Chem.* **2016**, *4*, 25.
[69] M. G. Wallis, U. von Ahsen, R. Schroeder, M. Famulok, *Chem. Biol.* **1995**, *2*, 543.
[70] J. E. Weigand, M. Sanchez, E. B. Gunnesch, S. Zeiher, R. Schroeder, B. Suess, *RNA* **2008**, *14*, 89.
[71] J. E. Weigand, S. R. Schmidtke, T. J. Will, E. Duchardt-Ferner, C. Hammann, J. Wohnert, B. Suess, *Nucleic Acids Res.* **2011**, *39*, 3363.
[72] A. Boussebayle, F. Groher, B. Suess, *Methods* **2019**, *161*, 10.
[73] a) B. Suess, S. Hanson, C. Berens, B. Fink, R. Schroeder, W. Hillen, *Nucleic Acids Res.* **2003**, *31*, 1853; b) Y. Nomura, Y. Yokobayashi, *J. Am. Chem. Soc.* **2007**, *129*, 13814.
[74] R. D. Jenison, S. C. Gill, A. Pardi, B. Polisky, *Science* **1994**, *263*, 1425.
[75] a) D. M. Mishler, J. P. Gallivan, *Nucleic Acids Res.* **2014**, *42*, 6753; b) A. Ogawa, *Methods Enzymol.* **2015**, *550*, 109.
[76] A. Autour, F. Bouhedda, R. Cubi, M. Ryckelynck, *Methods* **2019**, *161*, 46.
[77] a) B. Suess, B. Fink, C. Berens, R. Stentz, W. Hillen, *Nucleic Acids Res.* **2004**, *32*, 1610; b) S. Topp, J. P. Gallivan, *J Am Chem Soc* **2007**, *129*, 6807.
[78] P. Ceres, A. D. Garst, J. G. Marcano-Velazquez, R. T. Batey, *ACS Synth. Biol.* **2013**, *2*, 463.
[79] S. Hammer, C. Gunzel, M. Morl, S. Findeiss, *Methods* **2019**, *161*, 54.
[80] S. V. Harbaugh, J. A. Martin, J. Weinstein, G. Ingram, N. Kelley-Loughnane, *Methods* **2018**, *143*, 77.
[81] D. M. Mishler, S. Topp, C. M. Reynoso, J. P. Gallivan, *Curr. Opin. Biotechnol.* **2010**, *21*, 653.
[82] G. S. Hossain, M. Saini, R. Miyake, H. Ling, M. W. Chang, *Trends Biotechnol.* **2020**, *38*, 797.
[83] N. Sudarsan, J. E. Barrick, R. R. Breaker, *RNA* **2003**, *9*, 644.
[84] M. Vogel, J. E. Weigand, B. Kluge, M. Grez, B. Suess, *Nucleic Acids Res.* **2018**, *46*, e48.
[85] Y. Yokobayashi, *Curr. Opin. Chem. Biol.* **2019**, *52*, 72.
[86] I. Harvey, P. Garneau, J. Pelletier, *RNA* **2002**, *8*, 452.
[87] A. Ogawa, *RNA* **2011**, *17*, 478.
[88] a) C. I. An, V. B. Trinh, Y. Yokobayashi, *RNA* **2006**, *12*, 710; b) H. Mou, G. Zhong, M. R. Gardner, H. Wang, Y. W. Wang, D. Cheng, M. Farzan, *Mol. Ther.* **2018**, *26*, 1277.
[89] a) S. Auslander, M. Fussenegger, *Curr. Opin. Biotechnol.* **2017**, *48*, 54; b) M. Felletti, J. S. Hartig, *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1395.
[90] J. Tang, R. R. Breaker, *Chem. Biol.* **1997**, *4*, 453.
[91] M. N. Win, C. D. Smolke, *Proc. Natl. Acad. Sci.* **2007**, *104*, 14283.
[92] A. Cravens, J. Payne, C. D. Smolke, *Nat. Commun.* **2019**, *10*, 2142.
[93] B. Strobel, M. J. Duchs, D. Blazevic, P. Rechtsteiner, C. Braun, K. S. Baum-Kroker, B. Schmid, T. Ciossek, D. Gottschling, J. S. Hartig, S. Kreuz, *ACS Synth. Biol.* **2020**, *9*, 1292.
[94] N. Shanidze, F. Lenkeit, J. S. Hartig, D. Funck, *Plant Physiol.* **2020**, *182*, 123.
[95] a) B. Townshend, A. B. Kennedy, J. S. Xiang, C. D. Smolke, *Nat. Methods* **2015**, *12*, 989; b) J. S. Xiang, M. Kaplan, P. Dykstra, M. Hinks, M. McKeague, C. D. Smolke, *Nat. Commun.* **2019**, *10*, 4327.
[96] B. Strobel, M. Sporing, H. Klein, D. Blazevic, W. Rust, S. Sayols, J. S. Hartig, S. Kreuz, *Nat. Commun.* **2020**, *11*, 714.
[97] M. Felletti, J. Stifel, L. A. Wurmthaler, S. Geiger, J. S. Hartig, *Nat. Commun.* **2016**, *7*, 12834.
[98] a) D. Jasinski, F. Haque, D. W. Binzel, P. Guo, *ACS Nano* **2017**, *11*, 1142; b) J. Li, A. A. Green, H. Yan, C. Fan, *Nat Chem* **2017**, *9*, 1056.
[99] a) J. Chen, L. Tang, X. Chu, J. Jiang, *Analyst* **2017**, *142*, 3048; b) J. Liu, Y. Zhang, H. Xie, L. Zhao, L. Zheng, H. Ye, *Small* **2019**, *15*, e1902989; c) X. Zhou, Q. Zhu, Y. Yang, *Biosens. Bioelectron.* **2020**, *165*, 112422.
[100] W. Thavarajah, A. D. Silverman, M. S. Verosloff, N. Kelley-Loughnane, M. C. Jewett, J. B. Lucks, *ACS Synth. Biol.* **2020**, *9*, 10.
[101] R. S. Iwasaki, R. T. Batey, *Nucleic Acids Res.* **2020**, *48*, e101.
[102] a) F. Bouhedda, A. Autour, M. Ryckelynck, *Int. J. Mol. Sci.* **2018**, *19*, 44; b) E. Braselmann, C. Rathbun, E. M. Richards, A. E. Palmer, *Cell Chem. Biol.* **2020**, *27*, 891.

[103] E. Braselmann, A. J. Wierzba, J. T. Polaski, M. Chrominski, Z. E. Holmes, S. T. Hung, D. Batan, J. R. Wheeler, R. Parker, R. Jimenez, D. Gryko, R. T. Batey, A. E. Palmer, *Nat. Chem. Biol.* **2018**, *14*, 964.
[104] M. You, J. L. Litke, S. R. Jaffrey, *Proc. Natl. Acad. Sci.* **2015**, *112*, E2756.
[105] M. N. Stojanovic, D. M. Kolpashchikov, *J. Am. Chem. Soc.* **2004**, *126*, 9266.
[106] Y. Su, M. C. Hammond, *Curr. Opin. Biotechnol.* **2020**, *63*, 157.
[107] a) J. S. Paige, K. Y. Wu, S. R. Jaffrey, *Science* **2011**, *333*, 642; b) R. L. Strack, M. D. Disney, S. R. Jaffrey, *Nat. Methods* **2013**, *10*, 1219; c) G. S. Filonov, J. D. Moon, N. Svensen, S. R. Jaffrey, *J. Am. Chem. Soc.* **2014**, *136*, 16299; d) A. Autour, E. Westhof, M. Ryckelynck, *Nucleic Acids Res.* **2016**, *44*, 2491.
[108] C. A. Kellenberger, S. C. Wilson, J. Sales-Lee, M. C. Hammond, *J. Am. Chem. Soc.* **2013**, *135*, 4906.
[109] E. B. Porter, J. T. Polaski, M. M. Morck, R. T. Batey, *Nat. Chem. Biol.* **2017**, *13*, 295.
[110] M. D. E. Jepsen, S. M. Sparvath, T. B. Nielsen, A. H. Langvad, G. Grossi, K. V. Gothelf, E. S. Andersen, *Nat. Commun.* **2018**, *9*, 18.
[111] J. S. Paige, T. Nguyen-Duc, W. Song, S. R. Jaffrey, *Science* **2012**, *335*, 1194.
[112] X. C. Wang, S. C. Wilson, M. C. Hammond, *Nucleic Acids Res.* **2016**, *44*, e139.
[113] J. Yeo, A. B. Dippel, X. C. Wang, M. C. Hammond, *Biochemistry* **2018**, *57*, 108.
[114] a) C. A. Kellenberger, S. C. Wilson, S. F. Hickey, T. L. Gonzalez, Y. Su, Z. F. Hallberg, T. F. Brewer, A. T. Iavarone, H. K. Carlson, Y. F. Hsieh, M. C. Hammond, *Proc. Natl. Acad. Sci.* **2015**, *112*, 5383; b) Z. F. Hallberg, X. C. Wang, T. A. Wright, B. Nan, O. Ad, J. Yeo, M. C. Hammond, *Proc. Natl. Acad. Sci.* **2016**, *113*, 1790.
[115] G. S. Filonov, C. W. Kam, W. Song, S. R. Jaffrey, *Chem. Biol.* **2015**, *22*, 649.
[116] J. L. Litke, S. R. Jaffrey, *Nat. Biotechnol.* **2019**, *37*, 667.
[117] S.-F. Yuan, H. S. Alper, *Biotechnology Notes* **2020**, *1*, 2.
[118] R. Wu, A. Karunanayake Mudiyanselage, F. Shafiei, B. Zhao, Y. Bagheri, Q. Yu, K. McAuliffe, K. Ren, M. You, *Angew. Chem., Int. Ed.* **2019**, *58*, 18271.
[119] a) D. Bose, Y. Su, A. Marcus, D. H. Raulet, M. C. Hammond, *Cell Chem. Biol.* **2016**, *23*, 1539; b) Y. Su, S. F. Hickey, S. G. Keyser, M. C. Hammond, *J. Am. Chem. Soc.* **2016**, *138*, 7040.
[120] K. Furukawa, A. Ramesh, Z. Zhou, Z. Weinberg, T. Vallery, W. C. Winkler, R. R. Breaker, *Mol. Cell* **2015**, *57*, 1088.
[121] J. Xu, J. E. Johnson Jr., *Biochemistry* **2020**, *59*, 1508.
[122] J. Abatemarco, M. F. Sarhan, J. M. Wagner, J. L. Lin, L. Liu, W. Hassouneh, S. F. Yuan, H. S. Alper, A. R. Abate, *Nat. Commun.* **2017**, *8*, 332.
[123] J. A. Howe, H. Wang, T. O. Fischmann, C. J. Balibar, L. Xiao, A. M. Galgoci, J. C. Malinverni, T. Mayhood, A. Villafania, A. Nahvi, N. Murgolo, C. M. Barbieri, P. A. Mann, D. Carr, E. Xia, P. Zuck, D. Riley, R. E. Painter, S. S. Walker, B. Sherborne, R. de Jesus, W. Pan, M. A. Plotkin, J. Wu, D. Rindgen, J. Cummings, C. G. Garlisi, R. Zhang, P. R. Sheth, C. J. Gill, et al., *Nature* **2015**, *526*, 672.
[124] a) J. Mulhbacher, P. St-Pierre, D. A. Lafontaine, *Curr Opin Pharmacol* **2010**, *10*, 551; b) C. H. Lee, S. R. Han, S. W. Lee, *Nucleic Acid Ther.* **2016**, *26*, 44; c) E. Mehdizadeh Aghdam, M. S. Hejazi, A. Barzegar, *Gene* **2016**, *592*, 244; d) T. H. T. Chau, D. H. A. Mai, D. N. Pham, H. T. Q. Le, E. Y. Lee, *Int. J. Mol. Sci.* **2020**, *21*.
[125] S. Blouin, R. Chinnappan, D. A. Lafontaine, *Nucleic Acids Res.* **2011**, *39*, 3373.
[126] a) R. A. Battaglia, I. R. Price, A. Ke, *RNA* **2017**, *23*, 578; b) C. W. Reiss, Y. Xiong, S. A. Strobel, *Structure* **2017**, *25*, 195.
[127] O. Pikovskaya, A. Polonskaia, D. J. Patel, A. Serganov, *Nat. Chem. Biol.* **2011**, *7*, 748.
[128] K. Lang, R. Rieder, R. Micura, *Nucleic Acids Res.* **2007**, *35*, 5370.
[129] J. E. Johnson Jr., F. E. Reyes, J. T. Polaski, R. T. Batey, *Nature* **2012**, *492*, 133.
[130] J. T. Polaski, S. M. Webster, J. E. Johnson Jr., R. T. Batey, *J. Biol. Chem.* **2017**, *292*, 11650.
[131] W. C. Winkler, A. Nahvi, A. Roth, J. A. Collins, R. R. Breaker, *Nature* **2004**, *428*, 281.
[132] A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, M. S. Gelfand, *Nucleic Acids Res.* **2002**, *30*, 3141.
[133] R. M. Atilho, K. R. Perkins, R. R. Breaker, *RNA* **2019**, *25*, 23.
[134] E. E. Regulski, R. H. Moy, Z. Weinberg, J. E. Barrick, Z. Yao, W. L. Ruzzo, R. R. Breaker, *Mol. Microbiol.* **2008**, *68*, 918.
[135] J. X. Wang, E. R. Lee, D. R. Morales, J. Lim, R. R. Breaker, *Mol. Cell* **2008**, *29*, 691.
[136] K. A. Corbino, J. E. Barrick, J. Lim, R. Welz, B. J. Tucker, I. Puskarz, M. Mandal, N. D. Rudnick, R. R. Breaker, *Genome Biol.* **2005**, *6*, R70.
[137] R. T. Fuchs, F. J. Grundy, T. M. Henkin, *Nat. Struct. Mol. Biol.* **2006**, *13*, 226.
[138] Z. Weinberg, E. E. Regulski, M. C. Hammond, J. E. Barrick, Z. Yao, W. L. Ruzzo, R. R. Breaker, *RNA* **2008**, *14*, 822.
[139] a) M. M. Meyer, T. D. Ames, D. P. Smith, Z. Weinberg, M. S. Schwalbach, S. J. Giovannoni, R. R. Breaker, *BMC Genomics* **2009**, *10*, 268; b) E. Poiata, M. M. Meyer, T. D. Ames, R. R. Breaker, *RNA* **2009**, *15*, 2046.
[140] G. Mirihana Arachchilage, M. E. Sherlock, Z. Weinberg, R. R. Breaker, *RNA Biol.* **2018**, *15*, 371.
[141] T. D. Ames, D. A. Rodionov, Z. Weinberg, R. R. Breaker, *Chem. Biol.* **2010**, *17*, 681.
[142] X. Chen, G. Mirihana Arachchilage, R. R. Breaker, *RNA* **2019**, *25*, 1091.
[143] S. N. Malkowski, T. C. J. Spencer, R. R. Breaker, *RNA* **2019**, *25*, 1616.
[144] S. S. S. Panchapakesan, L. Corey, S. Malkowski, G. Higgs, R. R. Breaker, *RNA* **2020**, *27*, 99.
[145] T. D. Ames, R. R. Breaker, *RNA Biol.* **2011**, *8*, 82.
[146] S. Klahn, P. Bolay, P. R. Wright, R. M. Atilho, K. I. Brewer, M. Hagemann, R. R. Breaker, W. R. Hess, *Nucleic Acids Res.* **2018**, *46*, 10082.
[147] M. Mandal, R. R. Breaker, *Nat. Struct. Mol. Biol.* **2004**, *11*, 29.
[148] J. N. Kim, A. Roth, R. R. Breaker, *Proc. Natl. Acad. Sci.* **2007**, *104*, 16092.
[149] M. M. Meyer, A. Roth, S. M. Chervin, G. A. Garcia, R. R. Breaker, *RNA* **2008**, *14*, 685.
[150] P. J. McCown, J. J. Liang, Z. Weinberg, R. R. Breaker, *Chem. Biol.* **2014**, *21*, 880.
[151] E. R. Lee, J. L. Baker, Z. Weinberg, N. Sudarsan, R. R. Breaker, *Science* **2010**, *329*, 845.
[152] J. W. Nelson, N. Sudarsan, K. Furukawa, Z. Weinberg, J. X. Wang, R. R. Breaker, *Nat. Chem. Biol.* **2013**, *9*, 834.
[153] D. Yu, R. R. Breaker, *RNA* **2020**, *26*, 960.
[154] M. E. Sherlock, R. R. Breaker, *Biochemistry* **2017**, *56*, 359.
[155] M. E. Sherlock, H. Sadeeshkumar, R. R. Breaker, *Biochemistry* **2019**, *58*, 401.
[156] a) P. B. Kim, J. W. Nelson, R. R. Breaker, *Mol. Cell* **2015**, *57*, 317; b) J. J. Trausch, J. G. Marcano-Velazquez, M. M. Matyjasik, R. T. Batey, *Chem. Biol.* **2015**, *22*, 829.
[157] S. Li, X. Y. Hwang, S. Stav, R. R. Breaker, *RNA* **2016**, *22*, 530.
[158] R. L. Strack, W. Song, S. R. Jaffrey, *Nat. Protoc.* **2014**, *9*, 146.
[159] S. Nakayama, Y. Luo, J. Zhou, T. K. Dayie, H. O. Sintim, *Chem. Commun.* **2012**, *48*, 9059.
[160] C. A. Kellenberger, C. Chen, A. T. Whiteley, D. A. Portnoy, M. C. Hammond, *J. Am. Chem. Soc.* **2015**, *137*, 6432.

[161] S. Sharma, A. Zaveri, S. S. Visweswariah, Y. Krishnan, *Small* **2014**, *10*, 4276.
[162] S. Ketterer, L. Gladis, A. Kozica, M. Meier, *Nucleic Acids Res.* **2016**, *44*, 5983.
[163] H. Kim, S. R. Jaffrey, *Cell Chem. Biol.* **2019**, *26*, 1725.
[164] X. Li, L. Mo, J. L. Litke, S. K. Dey, S. R. Suter, S. R. Jaffrey, *J. Am. Chem. Soc.* **2020**, *142*, 14117.

Claire Husser received her master's degree in "Molecular Biology and Genetics" in 2020 from the University of Strasbourg (France). In parallel, she joined the "Integrative Molecular and Cellular Biology" Graduate School. She is currently a Ph.D. student in the team "Digital Biology of RNA" at the Institute of Molecular and Cellular Biology (Strasbourg, France) under the supervision of Michael Ryckelynck. Her current research interest focuses on the development of synthetic RNA by ultrahigh-throughput microfluidic screening.

Natacha Dentz is a Ph.D. student in the Digital Biology team at the Institute of Molecular and Cell Biology of Strasbourg. Her research focuses on the development of an ultrahigh-throughput screening platform to study translation initiation mechanisms in prokaryotes and eukaryotes.

Michael Ryckelynck received his Ph.D. in Molecular and Cellular Biology in 2005 from Université Louis Pasteur (Strasbourg-France). He then integrated the group of Professor. Andrew Griffiths where he contributed to the development of droplet-based microfluidics applied to biology, first as a postdoctoral fellow in 2006 and then Assistant Professor in 2007. In 2016, founded the team "Digital Biology of RNA" at the Institute of Molecular and Cell Biology (Strasbourg-France), a group that develops and exploits microfluidics to study RNA and develop new RNA-based tools. He was appointed full Professor of Biochemistry at University of Strasbourg in 2020.

2.2 Régulation chez les eucaryotes

2.2.1 Régulation transcriptionnelle et organisation des gènes

Contrairement au cas des procaryotes, les gènes eucaryotes ne présentent pas d'organisation en opéron, et d'autres mécanismes de régulation ont été mis en place au cours de l'évolution. L'ADN chez les eucaryotes est compacté de manière très dense au sein du noyau (Kensal van Holde and Jordanka Zlatanova, 1995; Woodcock and Ghosh, 2010). Les brins de chromatine sont super enroulés et ne permettent pas l'accès de l'ARN polymérase pour la transcription sur l'ensemble du génome. L'hétérochromatine représente les portions de chromatines les plus condensée et à l'inverse l'euchromatine est plus ouverte permettant l'accès au brin d'ADN. Certaines modifications épigénétiques au niveau des histones sont impliquées dans la compaction ou au contraire le relâchement de la chromatine. La chromatine est ouverte (euchromatine) dans les portions où les histones sont acétylées ou dé-méthylées par l'action de déméthylases, et où la transcription va donc être réalisable (Samantara *et al.*, 2021). Inversement, la présence de groupement méthyles sur les histones signalent une portion de chromatine inactive, aussi appelée hétérochromatine (Samantara *et al.*, 2021). Cependant, même une fois que la polymérase localise une portion d'euchromatine et entame la transcription, d'autres éléments peuvent jouer un rôle de régulation.

Chez les eucaryotes, une portion de séquence conservée riche en thymines et adénines appelée boite de Pribnow se localise en amont du premier nucléotide transcrit (Ozawa, Mizuno and Mizushima, 1987). Cette séquence va favoriser l'initiation de la transcription en facilitant l'interaction du complexe de pré-initiation de la transcription au niveau du promoteur qui, lui-même, peut permettre une interaction plus ou moins forte avec la polymérase modulant ainsi la fréquence d'initiation de la transcription. En effet, une interaction trop faible entre le promoteur et la polymérase peut limiter l'initiation de la transcription, tandis qu'une interaction trop forte aboutit à des initiations abortives (Li and Zhang, 2014).

Il existe également des séquences plus éloignées ayant la capacité de co-activer la transcription, appelé activateur (de l'anglais : « enhancer »). Ces séquences interagissent par l'intermédiaire de complexes protéiques tels que ''Mediator'', découvert en 1990 par Roger D. Kornberg qui interagit avec les facteurs généraux de la transcription et l'ARN polymérase II (Kim *et al.*, 1994). De la même manière, un autre complexe protéique, appelé SAGA (de l'anglais : « Spt-Ada-Gcn5-Acetyltransferase ») a été décrit comme étant capable d'acétylé des histones et donc de favoriser la transcription chez les levures (Roberts and Winston, 1997).

Une dernière alternative de modulation de la transcription réside dans la présence d'introns dans l'ARN pré-messager (pré-ARNm) fraîchement transcrit qui nécessite d'être

épissé avant de pouvoir être maturé puis traduit. L'épissage est une étape clé pouvant avoir des conséquences importantes selon l'alternative d'épissage réalisée post-transcriptionellement. En 1977, le premier épissage alternatif fut observé avec la mise en évidence de la transcription de différents ARNm à partir d'un même gène, codant pour différentes protéines, obtenues en fonction de la voie d'épissage employée par l'adénovirus de type 2 dans sa phase tardive (Chow *et al.*, 1977; Chow and Broker, 1978). Ce mécanisme complexe propose un autre panel d'alternatives pour l'expression des gènes très représenté chez les virus et eucaryotes (Woan-Yuh Tarn and Joan A. STeitz, 1997) non seulement pour la transcription mais aussi la traduction indirectement.

2.2.2 Régulation traductionnelle

La traduction eucaryote est elle aussi régulée à différents niveaux. Premièrement, la stabilité de l'ARNm lui-même va impacter sa traduction. L'ARNm peut ainsi être plus ou moins stable selon la présence de la coiffe en 5' et de la queue polyA en 3' conférant une certaine stabilité à l'ARNm mais aussi lui permettant d'être traduit selon le mécanisme de traduction coiffe-dépendant. Cette stabilité dépend aussi de la présence de silencing RNA (siRNA) ou de micro RNA (miRNA) qui en s'hybridant à l'ARNm peuvent induire son clivage et en conséquent sa dégradation (Valencia-Sanchez *et al.*, 2006; Carthew and Sontheimer, 2009) (Figure 9 A).

Par ailleurs, la séquence de l'ARNm elle-même, incluant le codon initiateur et la séquence de Kozak, module l'efficacité de l'initiation de la traduction. La séquence décrite dans l'ARNm par M. Kozak a été retrouvé en présence d'un codon initiateur de type AUG (Kozak, 1981). D'ailleurs, selon le codon initiateur employé l'efficacité de l'initiation de la traduction peut être aussi modulé, d'autres codons ayant déjà démontrés leur capacité à initier la traduction même s'ils tendent à être moins efficace (Kearse and Wilusz, 2017; Hernández, Osnaya and Pérez-Martínez, 2019) (Figure 9 B).

Enfin, quoique plus rare que chez les procaryotes, des riboswitches ont aussi été décrit chez les eucaryotes mais uniquement chez les plantes à ce jour. Le riboswitch répondant à la Thyamine PyroPhosphate ou TPP, module le splicing de pré-ARNm en présence ou en absence de son ligand, ce qui aura pour conséquence d'affecter la traduction de l'ARNm (Serganov and Nudler, 2013; Husser, Dentz and Ryckelynck, 2021) (Figure 9 C). Les alternatives de ce splicing peuvent avoir différentes conséquences allant de l'excision d'une micro ORFs à l'insertion d'un codon stop prématuré ou encore l'absence de signaux achevant la synthèse de la queue polyA déstabilisant alors l'ARNm. Ce mécanisme régulateur rarement retrouvé chez les eucaryotes est cependant de plus en plus représenté dans des expérimentations utilisant des riboswitches synthétiques (Weigand and Suess, 2007) qui

fonctionnent au sein d'environnement acellulaire eucaryote comme mentionné au prochain paragraphe (Husser, Dentz and Ryckelynck, 2021).

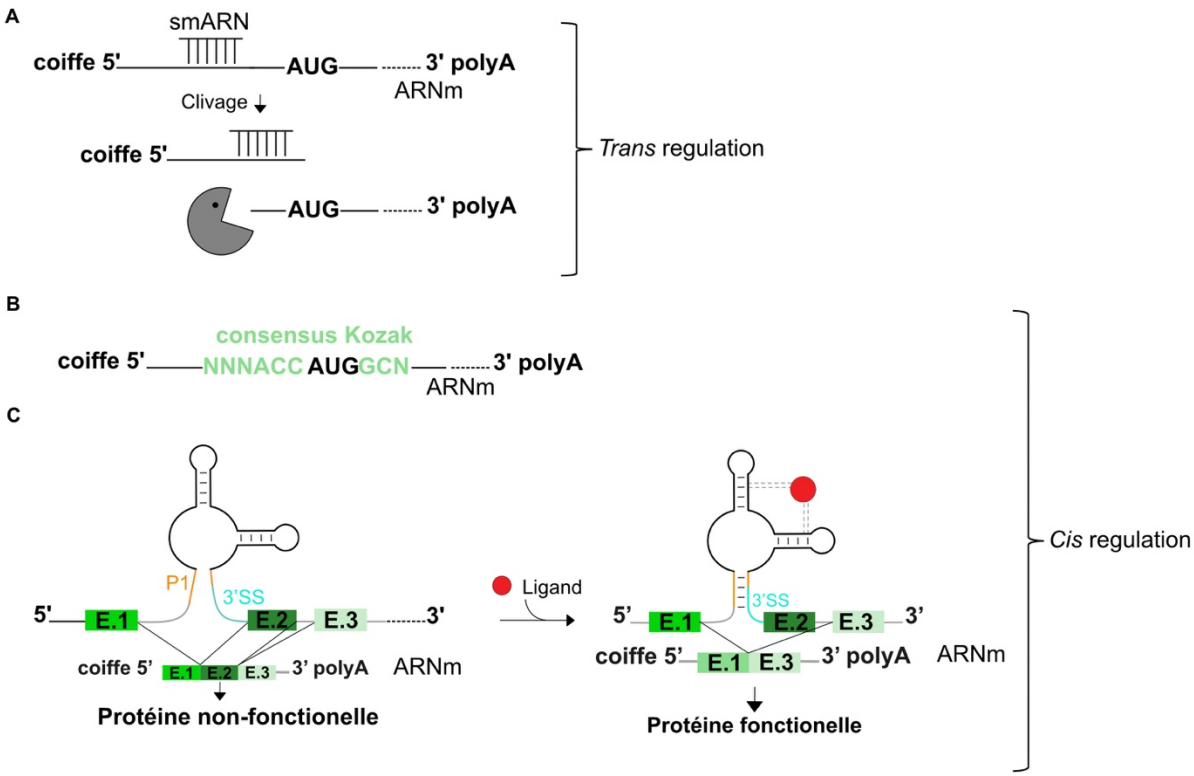


Figure 9 : Exemples de mécanisme de régulation de l'initiation de la traduction procaryote. A. Régulation par la présence de petit ARN régulateurs appelés smARN (de l'anglais : « smallRNA »). Le smRNA induit le clivage de l'ARNm qui est ensuite digéré par le dégradosome représenté par le ¾ de cercle gris **B.** Impact de la séquence de Kozak entourant le codon initiateur. La séquence de Kozak est décrite sur 9 positions représentée en vert clair avec 6 nucléotides en amont du codon initiateur et trois en aval **C.** Influence d'un riboswitch sur l'épissage d'un pré-ARNm. Selon la conformation du riboswitch l'épissage n'est pas le même et donc la protéine obtenue de même. Les exons sont dénommés par un E et le numéro de l'exon, l'hélice P1 est représenté en orange et le site de splicing en cyan. La figure C est adaptée de Husser, Dentz and Ryckelynck, 2021.

3. Expression de gènes en systèmes acellulaires

Dans le but de développer de nouveaux outils d'imagerie, de détection, de régulation ou plus simplement pour mieux comprendre certains systèmes biologiques, de nouvelles stratégies d'expression des gènes doivent être envisagées. Le besoin notamment de pallier des limites telles que les difficultés de culture cellulaire, les volumes réactionnelles ou encore le prix de ces expérimentations ont motivé le développement de nouvelles approches expérimentales qui sont alors mises en place à l'aide d'extraits cellulaires permettant de nouvelles méthodologies en condition acellulaire.

3.1 Les systèmes d'expression acellulaire

3.1.1 Historique

Au cours du siècle dernier, les systèmes d'expression *in vitro* aussi qualifiés d'acellulaires ou « cell-free » (de l'anglais : « sans cellule ») ont émergé et rapidement été améliorés pour s'étendre à de nombreuses applications énoncées plus bas (Silverman, Karim and Jewett, 2020). Ces systèmes d'expression permettent la réalisation d'expériences biologiques tels qu'exprimer des gènes codants (transcription/traduction) sans le besoin de la cellule et de s'affranchir de certaines limites. Avant même la première description de protocoles de préparation d'extraits cellulaires (Zubay, 1973), leur utilisation avait déjà permis l'élucidation de mécanismes biologiques clefs tels que : la fermentation alcoolique (Buchner Eduard, 1897) mais aussi le décryptage du code génétique (Marshall W. Nirenberg and J. Heinrich Matthaei, 1960). Dans la même optique, les facteurs généraux impliqués lors de l'initiation de la transcription ont eux aussi été étudiés au moyen d'extraits dénommés S30 dans un travail servant à l'heure actuelle encore de référence pour les préparations de kits (Lesley, Ann Brow and Burgess, 1991). L'utilisation de ces premiers extraits fût un réel avantage en permettant de s'affranchir de clonages ainsi que des cultures qui étaient, et sont encore à l'heure actuelle, des expérimentations plutôt complexes et/ou chronophages.

Par la suite, dans les années 2000, de nombreuses études se sont concentrées sur l'optimisation des protocoles de préparation des systèmes d'expression *in vitro*. L'objectif était alors d'améliorer les points critiques de ces kits tels que la reproductibilité, notamment grâce au développement d'extraits composés uniquement de facteurs de traduction purifiés comme avec la technologie d'expression PURE (de l'anglais : « Protein Synthesis Using Recombinant Elements ») Express (Shimizu et al., 2001). Dans ces kits d'expression, chaque facteur est

purifié à l'aide d'une étiquette 6-Histidines permettant par la suite de reconstituer à la demande et de manière complètement modulable et contrôlable la machinerie d'expression des gènes codants pour des protéines. À l'inverse de ce type d'extrait nécessitant un long processus de préparation incluant l'expression et la purification de chacun des éléments, d'autres méthodes ont suivi une stratégie alliant simplicité, rapidité et économie. En effet, une véritable vague de développement de procédés de préparation a déferlé dont l'un des objectifs majeurs était la mise en place de protocoles nécessitants peu ou pas de matériel spécifique, permettant une rapidité d'exécution et étant le plus économique possible, tout en maintenant des standards de qualité en termes de rendement et reproductibilité. L'idée était alors de rendre le plus accessible possible ce type de systèmes tout en répondant aux besoins des expérimentateurs (Kwon and Jewett, 2015). La gamme d'extraits s'est ensuite étendue à d'autres organismes, aussi bien procaryotes (bactéries à Gram positif ou négatif et les souches non modèles) (Kelwick *et al.*, 2016; Moore *et al.*, 2017; Wang, Li and Jewett, 2018) qu'eucaryotes dont des protozoaires, des levures, des plantes, des insectes et des mammifères (Zemella *et al.*, 2015). Au vu de la diversité des études menées, la variété de protocoles et d'organismes pouvant être employée devrait certainement s'étendre d'avantage au fil du temps.

Cependant, un autre aspect a fortement orienté le développement et l'amélioration de ces extraits : les rendements de production des protéines cibles ainsi que le degré de complexité de ces dernières. De manière générale, la capacité à produire des protéines en grande quantité est un objectif central que ce soit d'un point de vue industriel ou pour des applications de recherche nécessitant de grande quantité (2mg/mL) de protéine d'intérêt. Par exemple, l'étude d'un petit peptide cationique appelé hBD2 (de l'anglais : « humane Beta Defensine 2 ») a pu révéler des capacités antimicrobiennes intéressantes ne causant pas l'émergence de résistances (Chen *et al.*, 2005). Ainsi, des extraits comme myTXTL ont été développés par V. Noireaux et sont actuellement commercialisés avec une forte capacité de renouvellement d'énergie permettant une durée importante de réaction pouvant atteindre jusqu'à 16 heures sans ajout de substrat supplémentaire et avec des rendements pouvant dépasser les 2 mg/mL (Garamella *et al.*, 2016). De plus, une récente étude a établi qu'une marge d'optimisation était encore possible pour l'ensemble de ces protocoles et de ces kits suivant les tampons employés (Borkowski *et al.*, 2020).

3.1.2 Applications, avantages et limites

Il existe à ce jour un vaste panel d'extraits prêts à être utilisés pour un tout aussi large panel d'applications. Pour commencer, une grande partie des expérimentations en conditions acellulaires englobe des applications de production de protéines, que ce soit de façon massive avec des conditions d'expression de l'ordre d'une centaine de litres (Zawada *et al.*, 2011), ou pour la synthèse de protéines avec des structures particulières ou nécessitant des modifications post-traductionnelles. Malgré les avantages des conditions d'expression *in vivo,* qui sont indiscutablement représentatives du vivant, les kits d'expression acellulaire permettent quant à eux de s'affranchir de certaines contraintes, comme le temps de croissance des cellules, le stress induit par l'application de certaines molécules dans les milieux de cultures ou encore la surexpression de certaines molécules pouvant altérer et moduler le bon fonctionnement de l'ensemble de la machinerie cellulaire (Rosano and Ceccarelli, 2014). Les principaux avantages et les principales limites de chacune de ces stratégies d'études sont résumés ci-dessous (Figure 10).

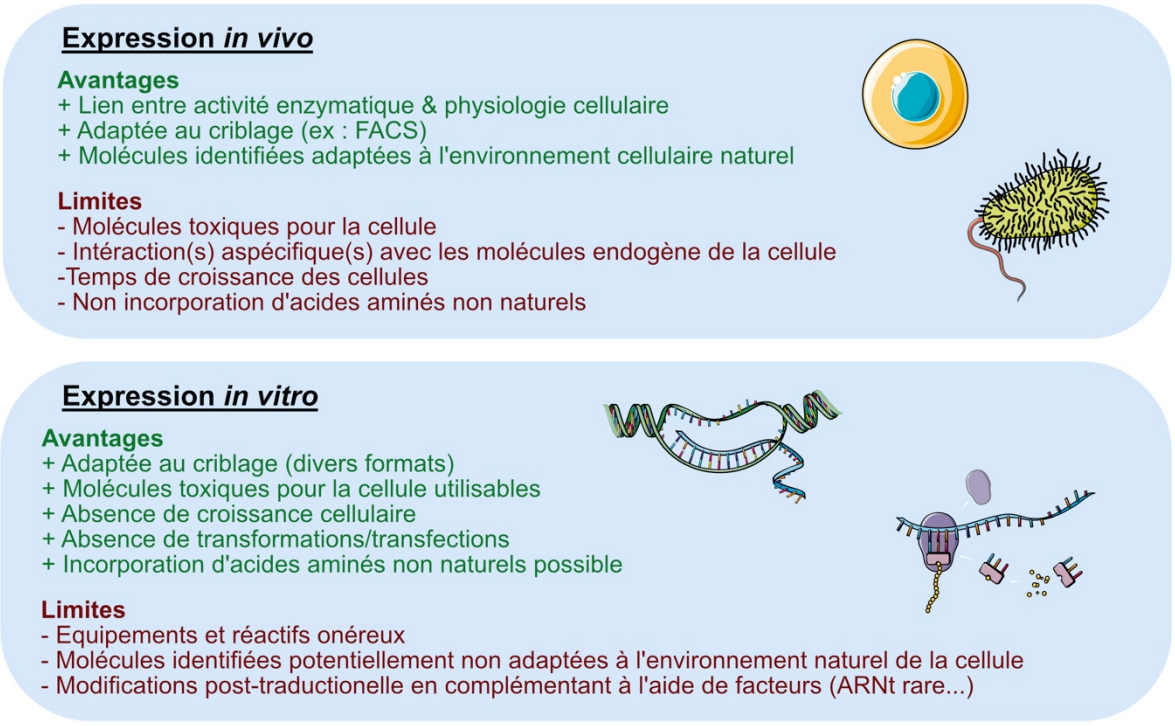


Figure 10 : Avantages et limites des deux principales conditions utilisées pour l'expression de gènes.

Les systèmes d'expression *in vitro*, sont variés et ajustables. Ils peuvent être utilisés sous forme d'extrait brut (de l'anglais : « crude extract ») ou bien sous forme pure. Les extraits bruts ne nécessitant pas de purification particulière sont obtenus plus rapidement et de manière moins coûteuse. De plus, leurs rendements sont nettement plus élevés que ceux obtenus avec des extraits ''PURE''. Néanmoins, les extraits purifiés offrent un parfait contrôle sur leur composition qui est adaptable à l'application souhaitées, et ils permettent une haute reproductibilité des expériences réalisées. Pour ces derniers, ce sont le faible rendement et le prix très élevé qui en limitent l'utilisation de manière plus étendue. D'un point de vue plus général, la variété d'organismes utilisables en conditions acellulaires sous forme PURE ou d'extrait brut, ouvre une large gamme de possibilités comportant des avantages tout comme des inconvénients propres à chacun, permettant d'adapter au mieux leur utilisation selon les travaux souhaités. Si par exemple, on en revient à la production de protéines, des extraits de germes de blé, ou WGE (de l'anglais : « Wheat Germ Extract ») seront pertinents pour l'introduction de modifications post-traductionnelles avec de bons rendements (Tableau 1) tandis que d'autres extraits seront plus appropriés à de la production massive de protéines tels que les extraits d'*E. coli*. De manière intéressante, afin de pouvoir maintenir un rendement élevé et d'y associer des caractéristiques particulières pour la production de protéines nécessitant une optimisation des codons ou des ponts di-sulfures, des optimisations ont été mises en place par l'ajout d'éléments et/ou additifs aux extraits d'*E. coli* (Smolskaya, Logashina and Andreev, 2020) (Tableau 1). Le tableau 1 ci-dessous résume de manière simple et non exhaustive les principaux avantages, inconvénients, rendements et applications des extraits (sans modifications/ajouts) les plus employés.

			Critères d'analyses					
Organismes			**Appelation**	**Avantages +**	**Limites -**	**Rendement (µg/mL de GFP)**	**Example d'applications**	**Références**
Procaryotes		*Archea*	*Archea* extract	Conditions extrêmes de production (haute température)	Faible rendement protéique	-	Production de protéine thermostable	Londei *et al.* Biochimie, (1991)
				Synthèse d'ARN très structuré				Ruggero *et al* .FEMS Microbiology
				Production de protéines structurées et thermostables				
	Gram -	*E. coli*	S30 extract, ECE	Rendement protéique élevé	Absence de modifications post traductionelle	2300	Production de protéines en grande quantité	Cashera *et al*. Biochimie, (2014)
				Modulable et flexible (incroporation d'acides aminés non conventionelles)	Mauvaise structuration de protéines complexe		Criblage à haut-débit	Levine *et al*. Journal of Visualized Experiments, (2019)
				Facilité d'obtention des extraits	Difficultés pour les protéines membranaires		Circuit génétique	
				Rentabilité prix de production/ prix de vente			Biologie synthétique	
				Commercialisé				
		Pseudomonas	*P. putida*-based extract	Utilisation de souche modèle autres que E. coli	Faible rendement protéique	200	Criblage haut débit	Wang *et al*. Synthetic Biology, (2018)
				Expression de gène riche en GC	Absence de modifications post-traductionelle			
					Difficultés pour les protéines membranaires			
					Mauvaise structuration de protéines complexe			
	Gram +	*Bacillus*	*B. subtilis* extract	Utilisation de souche modèle GRAM +	Rendement protéique à améliorer	420	Biologie synthétique	Kelwick *et al.* Metabolic Engineering, (2016)
					Absence de modifications post-traductionelles		Etude de la régulation de gènes	
					Mauvaise structuration de protéines complexes			
		Streptomyces	*S. venezuelae* TX-TL system	Utilisation de souche GRAM +	Absence de modifications post-traductionelle Difficultés pour les protéines membranaires et mauvaise structuration de protéines complexe	690	Biologie synthétique	Moore *et al.* Biotechnology Journal, (2017)
				Bon rendement protéique mais < rendement E. coli et Sreptomyces	Difficultés pour les protéines membranaires		Recherche microbiologie sur métabolisme et circuit génétique	
				Obtention de protéines propre au métabolisme de *streptomyces* avec la bonne structure	Mauvaise structuration de protéines complexes			

Eucaryotes	Levure	*Sacharomyces*	SCE	Simplicité de préparation et faible coût	Absence de modifications post traductionelles complexes	17	Production de protéine complexe	Hodgman *et al*. Biotechnology and Bioengineering, (2013)
				Co traduction et structuration de la protéine	Faible rendement protéique		Biologie synthétique	Rui Gan *et al* . Biotechnology Journal, (2014)
				Modifications post-traductionelle limitées dont glycolisation				
	Plantes	Germe de blé	WGE	Rendement protéique élevé en eucaryote	Extrait fastidieux à obtenir	1600	Criblage haut débit	Harber *et al*. FEBS Letters, *(2014)*
				Production de protéines eucaryote, membranaires, avec structuration et très soluble	Faible rendement protéique en comparaison aux procaryotes		Production de protéine (Vaccin)	
				Réaction jusque 60h sans modification			Etude du protéome ou de la structure des protéines	
	Mamifères	Lignées cellulaires	CHO extract/ HEK / HeLa	Modifications post-traductionelles dont glycolysation	Faible rendement protéique	Dépendant des lignées	Production de protéines complexes	Mikami *et al*. Protein Expression and Purification, (2006)
				Production de protéine membranaire	Besoin de grande quantité d'extraits (50% de la réaction)			Brödel *et al.* Biotechnol. Bioeng, *(2014)*
				Production des protéines humaines par les lignées de type humaines				
		Rétyculocytes de lapin	RRL	Systeme de mamifère plus simple à obtenir	Faible rendements protéique	-	Criblage à haut débit	Jackson *et al.* Methods in Enzymology, (1983)
					Absence de modifications post-traductionelle avec compléments		Biologie synthétique	Pelham *et al*. Eur. J. Biochem, (1976)
					Utilisation d'animaux vivants			
	Insecte		ICE	Facilité d'obtention de ces extraits	Faible rendement protéique	10 à 50 (YFP)	Production de protéines complexes	Brödel *et al.* Journal of Biotechnology, (2013)
				Modifications post-traductionelles dont glycolysation	Besoin de grande quantité d'extrait (50% de la réaction)		Biologie synthétique	
				Production de protéines membranaires				
				Commercialisé				
	rotozoaire			Production peu coûteuse	Absence de modifications post-traductionelles	-	Analyse à haut débit de PCR/CFPS	Mureev *et al.* Nature Biotechnology, (2009)
				Production de protéine très soluble	Peu utilisé et décris			
				Forte efficacité d'inititation selon la séquence d'ARN	Faible rendement protéique			
PURE Systeme		Personalisable	PURE express, PUREfrex	Absence de nuclease/protéase après purification	Prix élevé	380	Cellules minimales	Schimizu *et al.* Nature Biotechnology, (2001)
				Flexibilité et modulation de la composition	Pas d'activation du métabolisme endogène		Production de protéines complexes	
				Commercialisable	Purification par étiquette Histidine			

Tableau 1 : Récapitulatif des principaux extraits utilisés en expérimentation acellulaires. Chaque ligne décrie un extrait avec ses appellations les plus courantes, les avantages et inconvénients principaux, un ordre de grandeur de rendement de production protéique type ainsi que les applications les plus fréquentes (Pelham and Jackson, 1976; Richard Jackson and Tim Hunt, 1983; Londei et al., 1991; Ruggero, Creti and Londei, 1993; Shimizu et al., 2001b; Mikami et al., 2006; Mureev et al., 2009; Brödel et al., 2013; Hodgman and Jewett, 2013a; Brödel, Sonnabend and Kubick, 2014; Caschera and Noireaux, 2014; Harbers, 2014; Rui Gan and Michael C.Jewett, 2014; Kelwick et al., 2016; Moore et al., 2017; Wang, Li and Jewett, 2018; Levine et al., 2019).

Mise à part la synthèse de protéines, d'autres applications ont également été explorées, en amont de l'ensemble des optimisations de protocoles. Le fort potentiel de ces extraits, dans le domaine des applications thérapeutiques et de la santé, a été mis en avant avec des protéines telles que les anticorps (Ryabova *et al.*, 1997; Stech and Kubick, 2015) ou les peptides antibactériens (Martemyanov *et al.*, 2001). D'autres approches permettent de réduire le temps d'expérimentation, le prix de production tout en ouvrant la possibilité d'incorporer des acides aminés non naturels (dans le cadre des anticorps) ou de produire des protéines pouvant être cytotoxiques sous forme soluble, (dans le cadre des polypeptides antibactériens). Le caractère prometteur de ces futures applications dans le domaine médical a également motivé le développement de systèmes d'expression acellulaire dits "sur-demande". Ce concept est rendu accessible grâce au procédé de lyophilisation des extraits qui deviennent alors facilement transportables, conservés et aisément réhydratés (Pardee *et al.*, 2016), afin de produire, n'importe où et n'importe quand, une protéine d'intérêt, par exemple dans le cadre de kits diagnostics (Pardee *et al.*, 2016).

A côté de ces applications commerciales et industrielles, l'utilisation de ces systèmes est tout autant répandue pour des applications de recherche fondamentale. Les principales étapes de l'expression des gènes, par exemple, peuvent être étudiées à l'aide des systèmes acellulaires. Une étude a d'ailleurs joliment illustré la capacité d'étudier des éléments clefs de la régulation de l'expression des gènes que sont les promoteurs de transcription chez *Bacillus subtilis.* L'analyse d'une banque de promoteurs donna des résultats identiques dans différentes conditions d'étude, que ce soit *in vivo,* en microplaques ou en FACS, mais aussi *in vitro* à l'aide d'extraits acellulaires, démontrant la possibilité d'utiliser ces extraits pour l'obtention de résultats pertinents dans des conditions acellulaires mais de façon simplifiée, ouvrant la voie vers de l'ingénierie d'éléments régulateurs (Kelwick *et al.*, 2016). Des circuits génétiques peuvent aussi être mis en place et étudiés grâce à ces nouveaux systèmes d'expression, allant de ''simple'' cascade de promoteurs à des cellules artificielles. Ces circuits d'expression illustrent comment, au sein des cellules, les différentes molécules (ADN, ARN ou encore protéines) interagissent en réseaux. Ce type de dynamique se retrouve d'ailleurs fréquemment dans des mécanismes de régulations de l'expression des gènes. Des premiers circuits de complexité restreinte à deux constructions ont tout d'abord été testés *in vitro*. Ici,

l'holoenzyme de la RNA polymérase endogène d'*E. coli* et le facteur sigma 70 assurent la transcription d'une première construction codant pour un facteur sigma alternatif. L'association de l'holoenzyme à ce facteur permet alors d'initier la transcription d'une deuxième construction codante, elle, pour une protéine rapportrice (Figure 11 A). L'expression de ce circuit peut être suivie en temps réel à l'aide de la protéine rapportrice pouvant être fluorescente (ex : GFP) ou bien luminogène (ex : Renilla luciférase) (Shin and Noireaux, 2012). Ensuite, la complexité des circuits réalisés a augmenté en concevant des cascades d'expression impliquant bien plus de deux constructions et ressemblant davantage à ceux retrouvés dans la cellule. Ces circuits vont alors comporter des promoteurs pouvant s'activer de manière séquentielle (Figure 11 B). Il est également possible d'utiliser une construction comprenant un promoteur ainsi qu'un opérateur complexifiant le circuit par l'ajout d'une interaction répressive selon l'expression des autres constructions du circuit (Shin and Noireaux, 2012; Hori et al., 2017) (Figure 11 C).

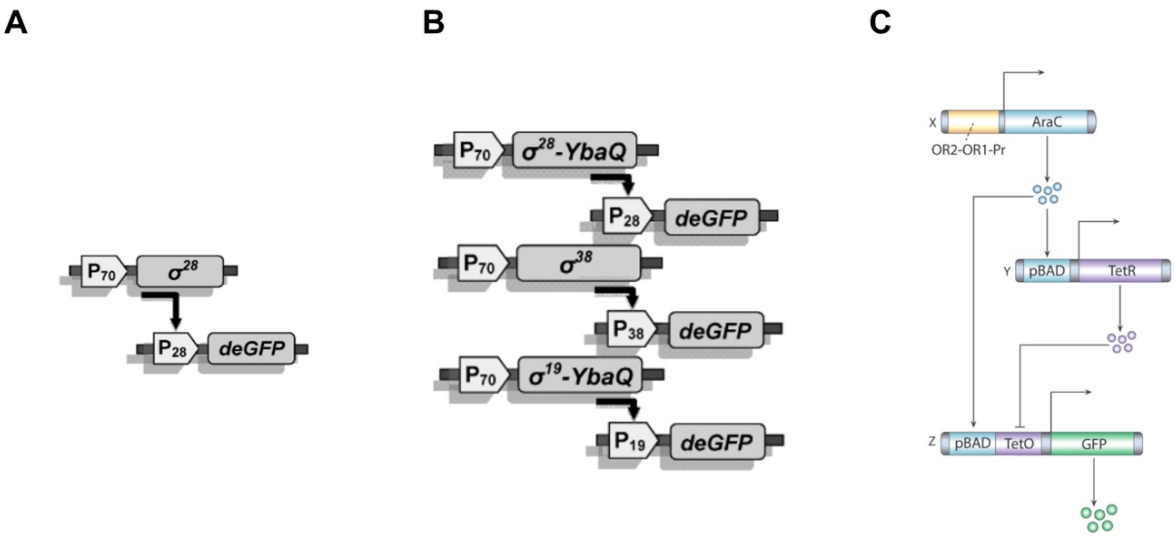


Figure 11 : Représentation schématique de circuits génétiques d'expression. A et B. Circuit d'expression une complexité moindre (A) et plus élevée (B). Les promoteurs sont représentés par les flèches et les gènes exprimés par les rectangles. Les représentations sont issues de la publication de Shin and Noireaux, 2012. **C.** Circuit d'expression incluant une séquence opératrice. Le gène (R), la molécule régulatrice (cercle) et la séquence opératrice (O) concernant la tétracycline sont représentés en violet. Le gène (C), la molécule régulatrice (cercle), et les promoteurs (p) concernant l'Arabinose sont représentés en bleu, le gène de la protéine rapportrice et cette dernière (cercle) sont représentés en vert. La représentation est issue de la publication de Hori et al., 2017.

Récemment, certaines équipes ont poussé les systèmes jusqu'à l'obtention de cellules artificielles par la compartimentation de systèmes acellulaires et ont démontré la possibilité de faire communiquer les différents compartiments formés (Dubuc *et al.*, 2019; Cho and Lu, 2020). L'utilisation d'extraits acellulaires est de plus en plus répandue et est appliquée à un vaste champ d'étude. D'une manière plus globale l'objectif des études fondamentales les utilisant très fréquemment est de mieux comprendre et d'ingénierer la régulation

de l'expression des gènes, que ce soit par l'intermédiaire de circuits génétiques ou d'expérimentation en conditions *in vitro*. Cependant, les avantages de ce type de manipulations sont contrebalancés par le bas débit des méthodes d'analyse compatibles avec ce format d'expression qui reste limité. Ainsi, les perspectives de ces extraits se trouveraient démultipliés s'ils pouvaient être utilisés pour la réalisation de criblage mais à haut-débit voir ultra haut-débit.

3.2 Criblage à haut débit en condition acellulaire et ses applications

Un criblage correspond à l'analyse individuelle et à la quantification (éventuellement suivi du tri) d'une ou plusieurs propriétés des éléments (gènes, cellules…) d'une population comme une activité enzymatique, de la fluorescence ou encore de la luminescence. La réalisation de tels criblages en conditions acellulaires permet de s'affranchir des limites inhérentes à l'utilisation de cellules : les conditions de culture *in vivo* (les cultures cellulaires), la variabilité des populations. L'utilisation de milieux de transcription *in vitro*, appelés IVT (de l'anglais : « *In Vitro* Transcription ») ou d'extraits acellulaires permettant la transcription et/ou la traduction *In Vitro* (TTIV) à la place de cultures cellulaires, permet l'étude de mécanismes moléculaires à l'aide de différents organismes pouvant être modèles, non modèles et même ceux complexes à cultiver. Il est possible d'utiliser des extraits d'*E. coli* dans l'idée de maintenir un fort rendement tout en complémentant ces extraits avec des facteurs spécifiques afin de les adapter et leur conférer des propriétés normalement absentes d'*E. coli* telles que, par exemple, certaines modifications post-traductionnelles comme les glycosylations (Tableau 1). Que le criblage implique une transcription, une traduction ou le couplage des deux (TTIV), le volume réactionnel ainsi que les débits de criblage peuvent varier selon la stratégie expérimentale employée pouvant aller du millilitre ou microlitre avec des compartiments rigides (microtubes et plaques) à des volumes réactionnels bien plus petits (de l'ordre du nano ou picolitre) à l'aide de microréacteurs en matière molle.

3.2.1 Stratégies conventionnelles de criblage acellulaires

Les approches de criblages sont grandement facilitées par l'utilisation de tests fluorogènes (c'est-à-dire conduisant à la génération d'un signal fluorescent). Ces derniers permettent la réalisation de mesures rapides, non invasives, facilement automatisables et multiplexables. La transcription *in vitro* peut facilement être suivie en temps réel grâce à l'utilisation d'aptamères fluorogènes. Les aptamères fluorogèniques sont de petites structures

ARN capable de se lier et d'activer spécifiquement la fluorescence de petites molécules appelées fluorogènes, peu ou pas fluorescentes à l'état libre. Mon laboratoire d'accueil a été particulièrement impliqué dans le développement de ces molécules, et plus particulièrement leurs composantes ARN (Autour, Westhof and Ryckelynck, 2016; Bouhedda, Autour and Ryckelynck, 2018). Ainsi, par exemple, une fluorescence verte (510 nm-534 nm) peut être obtenues avec des systèmes tels que Mango-III/TO1-Biotine (Trachman *et al.*, 2019) ou bien Broccoli/DFHBI-1T (Filonov *et al.*, 2014), tandis qu'une émission orange (560 nm-600 nm) pourra être obtenue avec des systèmes tels que o-Coral/Gemini 561 (Bouhedda, Autour and Ryckelynck, 2018). Le répertoire de couples aptamères/fluorogènes est bien plus vaste, et il couvre aujourd'hui presque l'entièreté du spectre visible (pour une liste exhaustive, voir la revue : Zhou and Zhang, 2021).

Pour le suivi de la traduction, plusieurs possibilités sont envisageables, avec tout d'abord la production de protéines naturellement fluorescentes, telles que la GFP (488 nm-535 nm) ou l'eGFP (488 nm - 510 nm) (Rodriguez *et al.*, 2017). Enfin, la protéine exprimée peut également activer spécifiquement un fluorogène comme le système FAST (de l'anglais : « Fluorescence - Activating absorption - Shifting Tag ») (Plamont *et al.*, 2016). La longueur d'onde émise par le complexe peut également être facilement ajustée en remplaçant le ligand utilisé (Plamont *et al.*, 2016; Tebo *et al.*, 2021). L'émission de fluorescence par une protéine d'intérêt peut aussi résulter d'une activité enzymatique ; soit l'activité enzymatique étudiée, soit celle d'un rapporteur générant un signal luminescent (la Renilla luciférase par exemple) ou de la fluorescence (dégradation de la Fluorescéine ßD di-galactopyranoside par la ß-galactosidase). Bien que permettant une quantification indirecte à contrario des protéines fluorescentes, l'approche enzymatique permet un gain substantiel de sensibilité du fait de l'amplification de signal liée à la réaction de conversion du substrat en produit.

En plus de disposer d'un test d'activité compatible avec le haut-débit, il est également nécessaire de définir le format dans lequel ce dernier est réalisé. La compartimentation de chaque variant (ARN ou protéine) peut se faire dans un compartiment rigide, comme dans un microtube ou un puit de microplaque. Les tubes Eppendorf offrent un volume réactionnel de 1,5 mL, cependant leur utilisation manuelle nécessite beaucoup de temps d'expérimentation (préparations des essais), de consommables et de réactifs, ce qui est donc moins compétitif d'un point de vue économique. Ce type de criblage peut être miniaturisé en microplaques (pouvant comporter jusqu'à 384 puits de 15 à 100 µL chacune), elles-mêmes manipulables en grande quantité par des processus automatisés (bras robotiques). Cette stratégie est compatible avec des analyses de criblages dits à haut débit, c'est-à-dire de l'ordre de 10 000 réactions par jour. Bien que l'association des systèmes de détection rapidement lisible en microplaque et les conditions acellulaires ouvrent la porte à divers criblages bien plus exhaustif que ceux réalisés initialement à la main, les essais restent limités en termes de nombre au vu

du prix des extraits commercialisés ou encore du temps nécessaire à leur obtention. Une alternative à cette limitation consiste à miniaturiser davantage les tests et ainsi à réaliser des économies substantielles de temps et d'argent, et d'atteindre le criblage à ultra haut-débit (au-delà de 100 000 tests par jour)

3.2.2 Compartimentation *in vitro*

Les principales limites de ces criblages *in vitro* concernent le coût élevé des extraits employés et le temps investit à la réalisation de ces expérimentations lorsque l'on souhaite tester un nombre important de variant. Une réduction des volumes des contenants rigides (tubes et microplaques) n'est pas envisageable du fait de contraintes liées aux procédés de fabrication et à la manipulation des liquides. La miniaturisation nécessite donc un changement de paradigme et l'adoption d'approches exploitant la matière molle. Ainsi, des compartiments renfermant des extraits d'expression *in vitro* peuvent être fabriqués sous forme d'hydrogel (Cho and Lu, 2020) ou encore de gouttelettes d'eau-dans-l'huile (Tawfik and Griffiths, 1998). La Compartimentation *In Vitro* (ou IVC), correspond à la génération de compartiments à très petite taille dans le but de contenir et exprimer un unique variant. L'une des premières utilisations de l'IVC fut la démonstration par S. Tawfik et A.D. Griffiths en 1998 de l'individualisation de gènes codant pour différentes enzymes au sein de gouttelettes d'eau-dans-l'huile de la taille de bactéries. Leur objectif était de permettre l'expression individuelle d'un grand nombre de gènes tout en maintenant le lien génotype/phénotype. Ils ont ainsi démontré la possibilité de sélectionner des gènes codant pour une protéine aux propriétés d'intérêt (dans leur cas la méthylation de l'ADN) parmi une large population de gènes (Tawfik and Griffiths, 1998). Bien que révolutionnaire en son temps, l'IVC souffrait de deux limitations majeures : la forte polydispersité (c'est à dire la variabilité de volume entre les gouttelettes) des gouttelettes générées et l'impossibilité d'en modifier le contenu une fois celles-ci formées. Ces limitations ont été contournées par la suite en adaptant l'IVC au format microfluidique, donnant ainsi naissance à la µIVC (Ryckelynck *et al.*, 2015). Il est ainsi possible d'individualiser et d'exprimer *in vitro* des gènes codant pour des protéines au sein de gouttelettes de l'ordre du picolitre. Cela a notamment permis l'expression de gènes codant pour la ß-galactosidase par suivi direct de la transformation par l'enzyme d'un substrat fluorogène en un produit fluorescent (Mazutis *et al.*, 2009), mais aussi d'une Pénicilline Acylase au moyen d'un test plus indirect (Woronoff *et al.*, 2015). Plus récemment, l'initiation de la traduction médiée par divers variant d'IRES (1.2.2) a pu être observée avec la détection de l'eGFP en µIVC, lors d'une étude visant à déterminer la nature des paires codons/anticodons acceptées par le ribosome dans des extraits de réticulocytes de lapin ou RRL (de l'anglais : « Reticulocyte Rabbit Lysate ») (Pernod *et al.*, 2020). Enfin, la µIVC permet également d'évoluer des protéines en

partant de vastes banques d'ADN. Il a ainsi été possible d'améliorer de 5 fois l'activité d'une protéase (Holstein, Gylstorff and Hollfelder, 2021). Au-delà de l'expression de gènes codants pour des protéines, la technologie a également été largement employée par mon équipe pour l'étude et l'amélioration d'ARN qu'ils soient catalytiques (Ryckelynck *et al.*, 2015) ou fluorogènes (Autour, Westhof and Ryckelynck, 2016; Bouhedda, Autour and Ryckelynck, 2018).

4. La microfluidique en gouttelettes et l'expression des gènes

La microfluidique est une discipline consistant en la manipulation de fluides dans des systèmes micrométriques. Il s'agit d'une technologie de plus en plus employée en biologie, principalement pour ses avantages évidents : la miniaturisation du système étudié, des cadences de débits pouvant atteindre l'ultrahaut-débit (millions d'analyses par jour). Par cette stratégie, il est possible d'analyser une grande population d'individus avec une résolution à l'individu unique ; donnant ainsi accès à une analyse digitale (Figure 12).

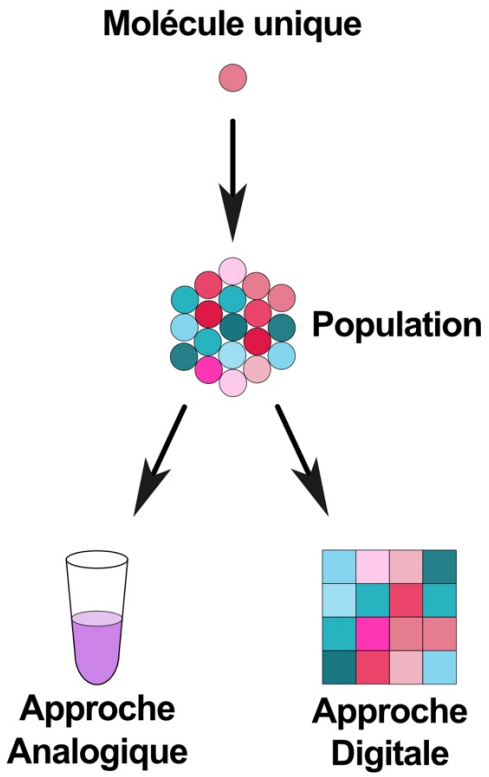


Figure 12 : Concept des analyses digitales d'une population. Des molécules (ADN, ARN, protéines, bactéries, cellules etc…) sont représentées par des petits cercles dont chaque couleur correspond à un variant. L'approche analogique représente une moyenne des variants et donc des couleurs. L'approche digitale représente chaque molécule par un carré et donc une analyse de chaque variant de manière

isolé. On obtient ainsi une analyse pertinente de la population avec notamment distinction des évènements rares au lieu d'une moyenne parfois moins pertinente.

4.1 La microfluidique en gouttelettes : concept et stratégie employée

4.1.1 Écoulement de liquides, production et stabilisation de gouttelettes

Les fluides circulant dans les canaux d'une puce microfluidique sont caractérisés par plusieurs paramètres distincts. Leur écoulement est associé au nombre de Reynolds, synonyme d'écoulement laminaire dont la conséquence directe est une faible efficacité de mélange. Cette propriété est une limitation lorsqu'il s'agit d'expériences requérant la réalisation de mélanges rapides. En revanche, elle donne également accès à un contrôle fin de l'écoulement liquides. Il devient ainsi possible de former des gouttelettes d'eau-dans-l'huile de façon extrêmement contrôlée avec une très faible variation de taille d'une gouttelette à l'autre. Pour ce faire, une phase aqueuse et une phase d'huile sont dirigées dans les canaux microfluidiques et forcées à se rencontrer au niveau d'une géométrie particulière. La phase d'eau se disperse alors dans la phase d'huile en gouttelettes régulières. Ainsi, par exemple, le dispositif par focalisation de flux permet la formation de gouttelettes en pinçant une phase aqueuse par deux flux d'huile et en forçant ce flux combiné à passer par un orifice (Figure 13).

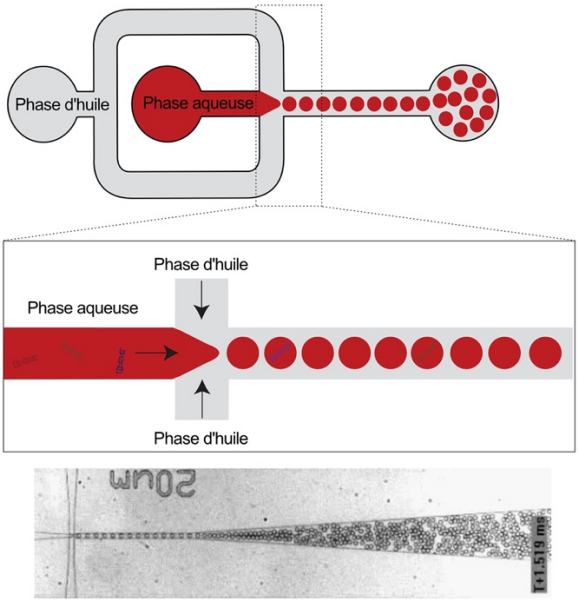


Figure 13 : Module microfluidique de production de gouttelettes. La phase d'huile représentée en grise vient pincer (flèches noires) la phase aqueuse en rouge contenant les molécules d'intérêts (exemple : ADN représenté dans les gouttelettes). Sous la représentation schématique se situe une capture d'écran de l'observation au microscope de ce module.

Le simple mélange d'eau et d'huile ne suffit pas à former une émulsion stable. En effet, la forte tension de surface existante entre ces deux composés tend naturellement à faire fusionner des gouttelettes non stabilisées, afin de minimiser la surface de contact entre les liquides. L'ajout d'un agent tensio-actif, ou surfactant, (une molécule amphiphile ayant une tête hydrophile à l'intérieur des gouttelettes et une queue hydrophobe enchâssée dans l'huile) permet néanmoins de réduire cette tension de surface et ainsi de prévenir la fusion non contrôlée des gouttelettes entre elles. Ces agents rendent également les gouttelettes résistantes à de hautes températures et donc compatibles avec des applications de thermocyclages.

La production et manipulation de gouttelettes se réalise au sein de puces microfluidiques composées de différents modules selon les actions souhaitées. Le module le plus basique consiste en la production des gouttelettes comme décrit plus haut. C'est également l'étape lors de laquelle les individus (molécules ou cellules) peuvent être individualisés en vue d'analyses ultérieures (amplification d'ADN, phénotypage…). La répartition d'objets dans des compartiments suit une distribution de Poisson (S. D. Poisson, 1837) selon la relation : $P_{(X=k)} = (e^{-\lambda}/k!) \times \lambda^k$, où P est la probabilité d'avoir k, le nombre exacte de molécules par compartiment et λ le nombre moyen de molécule par compartiment. La conséquence directe de cette propriété est que la minimisation des phénomènes d'encapsulation multiples nécessite de diluer fortement l'échantillon. Ainsi, si par exemple l'objectif tend à n'avoir qu'une molécule par gouttelettes (k = 1), l'occupation moyenne des gouttelettes sera d'environ 20 % (λ = 0,2, valeur à laquelle 81% des gouttelettes sont vides, 16% contiennent exactement 1 molécule, et le reste 2 molécules ou plus).

4.1.2 Manipulation des gouttelettes

En plus de présenter une forte homogénéité de taille, les gouttelettes microfluidiques peuvent être manipulées et leur contenu modifié à la demande après leur production, de différentes manières. Tout d'abord, deux séries de gouttelettes de tailles différentes peuvent être synchronisées une à une, permettant à la paire de circuler collées l'une à l'autre (Figure 14).

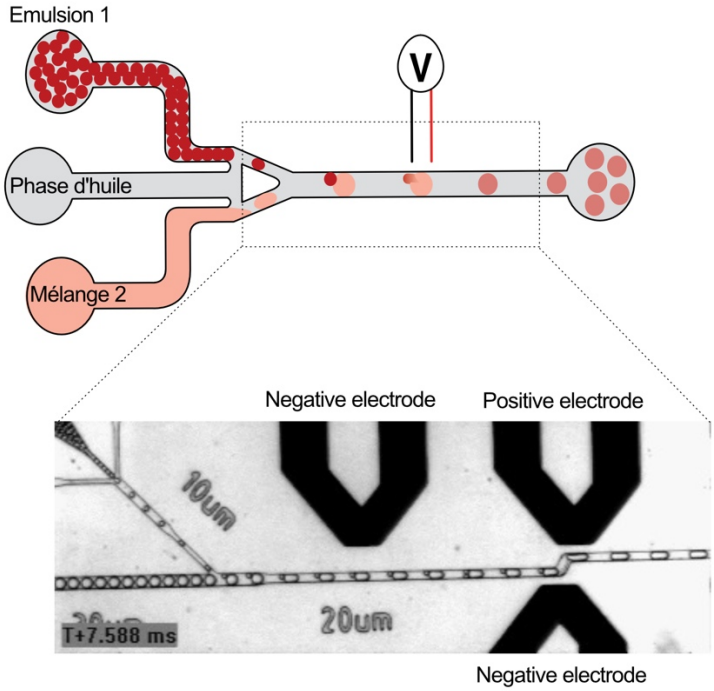


Figure 14 : Module microfluidique de fusion de gouttelettes. La phase d'huile représenté en grise aide à synchroniser une première émulsion en rouge (par exemple des gouttelettes de PCR) avec des gouttelettes dix fois plus volumineuse d'un deuxième mélange réactionnelle (milieu d'expression par exemple). L'application d'un champ électrique (V) permet la déstabilisation des gouttelettes favorisant la fusion de ces deux gouttelettes ensemble. Sous la représentation schématique se situe une capture d'écran de l'observation au microscope de ce module adaptée de la publication : J-C. Baret et al., 2009.

Le passage devant un champ électrique créé par une paire d'électrodes désorganise localement les couches de surfactant, les déstabilisant et provoquant la fusion contrôlée des gouttelettes. De ce fait, si une des gouttes contient l'ADN à exprimer et l'autre le milieu d'expression *in vitro*, cette action de fusion équivaut à transférer l'ADN dans le milieu d'expression et à déclencher la réaction de transcription des ADN en ARN. Ici, le volume de solution délivrée est strictement défini par la somme du volume de chaque gouttelette.

Alternativement, et suivant un principe analogue à la fusion de gouttelettes, une solution de composition définie peut être pico-injectée dans chaque gouttelette en introduisant cette dernière sous pression dans un canal arrivant de façon perpendiculaire à un second canal dans lequel circulent les gouttelettes à modifier. Lorsque les gouttelettes passent une à une en face du canal de solution, elles entrent transitoirement en contact avec cette dernière. La présence d'un jeu d'électrodes placé au niveau de la jonction entre les deux canaux permet de créer un champs électrique déstabilisant transitoirement l'interface gouttelette/solution et conduisant à l'injection d'un volume de solution proportionnel à la taille de la gouttelette ainsi qu'au temps qu'elle passe devant le canal de solution (le volume délivré peut donc être modulé par la vitesse de passage de la gouttelette et/ou de la pression exercée sur la solution dans le canal orthogonal) (Figure 15).

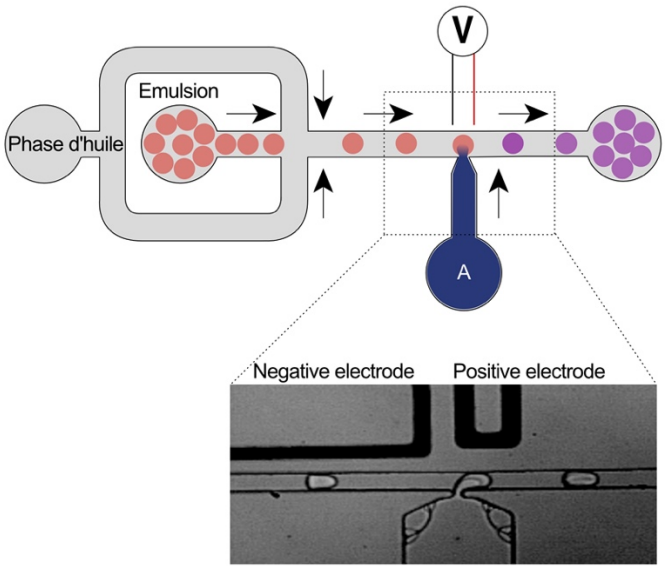
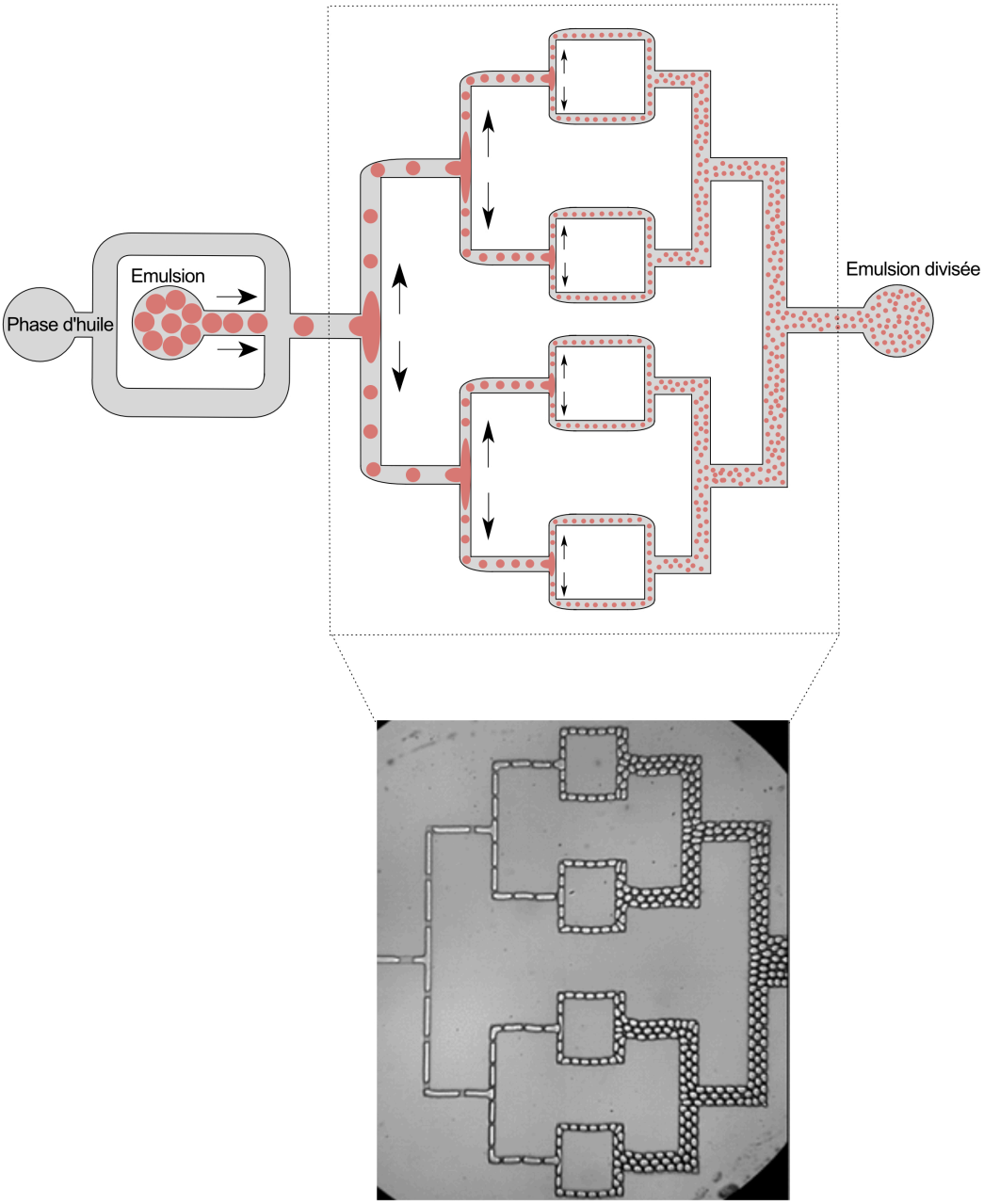


Figure 15 : Module microfluidique de pico injection de gouttelettes. La phase d'huile représenté en grise sépare les gouttelettes d'une première émulsion en rose qui sont ensuite pico injecté d'un mélange réactionnel A en bleu et ce, à l'aide d'un champ électrique favorisant la déstabilisation de la gouttelette pico-injectée. Sous la représentation schématique se situe une capture d'écran de l'observation au microscope de ce module.

Par ailleurs, les gouttelettes collectées peuvent être divisées en deux sous gouttelettes. Cette division a lieu lors du passage des gouttelettes au travers de modules microfluidiques particuliers forçant la division du volume en deux volumes égaux (Figure 16).



<u>Figure 16 :</u> Module microfluidique de division de gouttelettes (de l'anglais : « Splitter »). La phase d'huile représentée en grise sépare les gouttelettes de l'émulsion (en rose) qui sont ensuite divisées en plus petites gouttelettes de volumes équivalents de part et d'autre du canal (flèches noires). Sous la représentation schématique se situe une capture d'écran de l'observation au microscope de ce module adaptée de la publication : Link et al., 2004

Enfin, la fluorescence de chaque gouttelette peut être mesurée grâce à un dispositif optique dédié et utilisé pour trier les gouttelettes (Figure 17).

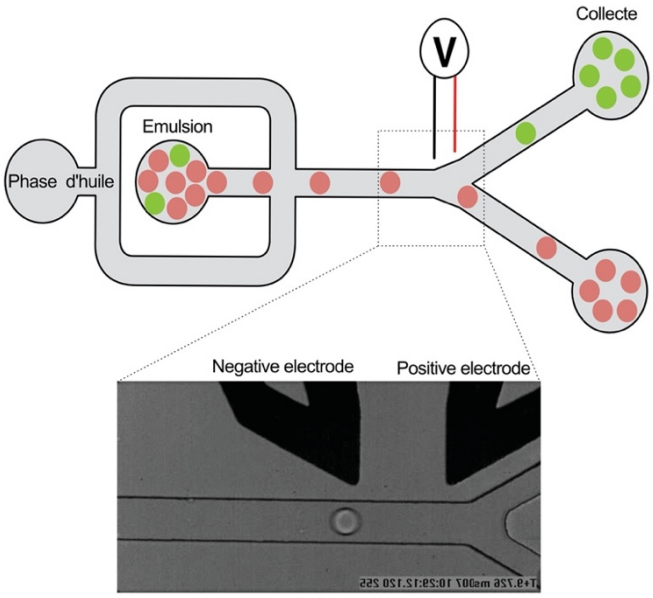


Figure 17 : Module microfluidique de tri de gouttelettes. La phase d'huile représenté en grise sépare les gouttelettes d'émulsion en rose ou vert selon leur profil fluorescent. Les gouttelettes sélectionnées avec le profil fluorescent souhaité en vert dans la représentation sont déflectées par le champs électrique appliqué (V), à l'inverse du reste des gouttelettes. Sous la représentation schématique se situe une capture d'écran de l'observation au microscope de ce module adaptée de la publication : Mazutis, Baret and Griffiths, 2009

Les gouttelettes peuvent être déflectées par application d'un gradient de champs électriques déplaçant alors ces dernières par diélectrophorèse. Le champ électrique déclenche une orientation des dipôles contenus dans la gouttelette qui vont ensuite se déplacer dans le gradient de champs électriques vers l'intensité la plus importante. Cela conduit au déplacement de la gouttelette et à son écoulement dans un autre canal. Ici, tout repose sur l'existence du champ électrique dont la formation peut être contrôlée par ordinateur et n'être activée que sous condition d'une émission de fluorescence détectée et utilisée par l'ordinateur pour gérer la décision d'appliquer le champ électrique ou non.

4.1.3 Le processus de criblage microfluidique employé

L'utilisation en série, voire combinée, des différents principes énoncés plus haut permet la conception de pipelines d'analyse microfluidique tels que celui qui sera employé pour les travaux réalisés dans le cadre de cette thèse. Ce pipeline se segmente en trois grandes étapes : l'individualisation des gènes codants les différents variant, l'expression de chaque variant amplifié, puis l'analyse de leur phénotype ainsi que la sélection des plus adaptés (Figure 18).

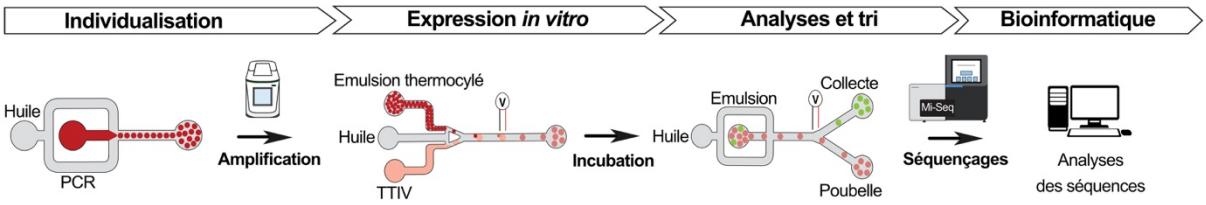


Figure 18 : Processus microfluidique utilisé pour les sélections.

La première étape réalisée consiste en l'individualisation des variants composant la banque étudiée. Cette banque est préalablement diluée dans un milieu d'amplification PCR de sorte à atteindre l'occupation de gouttelettes souhaitée. Le mélange est ensuite émulsionné avant que les gouttelettes ne soient thermocyclées. Chaque variant est alors amplifié de manière indépendante au sein de sa gouttelette. Après amplification, l'émulsion est réinjectée sur une nouvelle puce afin de fusionner une à une ces gouttelettes de PCR avec des gouttelettes dix fois plus volumineuse de milieu d'expression. Une fois la fusion réalisée entre une gouttelette de PCR et une de mixture permettant la transcription et/ou la traduction, l'ensemble est placé à température d'incubation afin de permettre une expression synchronisée dans chaque gouttelette. Par la suite, l'émulsion est réinjectée dans une dernière puce dans le but de mesurer le profil fluorescent de chacune des gouttelettes et de sélectionner celles comportant le profil recherché. Les gouttelettes d'intérêt sont enfin collectées, cassées et leur contenu d'ADN récupéré afin de l'analyser (par exemple par séquençage à haut débit) ou de l'utiliser pour initier un nouveau tour de criblage. En fin de processus, le contenu des banques est analysé par séquençage à haut débit et bio-informatiques (Bouhedda *et al.*, 2021).

5. Objectifs du travail de thèse

L'objectif principal de ma thèse a été d'établir une plateforme d'analyse à ultrahaut-débit dans le but de caractériser des séquences d'ARN impliquées dans la régulation de l'expression des gènes et plus particulièrement lors de la traduction, que ce soit chez les procaryotes ou les eucaryotes. Pour l'ensemble des objectifs, l'optimisation ainsi que la validation du processus de criblage ont été réalisées incluant l'ensemble des préparations depuis de la mise en place des banques de grandes tailles (>500pb) jusque l'envoie au séquençage et l'analyse des résultats.

J'ai tout d'abord mis en place et validé l'ensemble du processus à l'aide d'un système modèle déjà bien décrit dans la littérature, la séquence de recrutement du ribosome d'*E. coli*. Puis, après cette première validation, le processus a été étendu à des éléments plus

complexes comme les riboswitches tout en commençant par le système le plus simple encore une fois, en l'occurrence les riboswitches transcriptionnels.

Par ailleurs, un autre objectif de mon travail de doctorat a été de pouvoir étendre cette plateforme à l'étude des mécanismes impliqués dans le contrôle de la traduction chez les eucaryotes. Deux alternatives de l'initiation de la traduction chez les eucaryotes ont ainsi été explorées. Une première partie des travaux a porté sur la ré-exploration de la séquence dite de Kozak qui est impliquée dans l'initiation coiffe-dépendante. La nature des séquences impliquées dans cette initiation peut d'ailleurs varier selon le contexte environnant (codons initiateurs ou structures secondaires).

La deuxième étude a porté sur les initiations non canoniques qui impliquent des éléments structuraux de recrutement interne du ribosome, ou IRES. Ces éléments sont quant à eux impliqués lors d'initiation dite coiffe-indépendante bien souvent utilisée par les virus. L'idée a ici été de parvenir à identifier de nouvelles séquences initiatrices au sein de génomes viraux, séquences qui pourraient servir dans le futur de cibles thérapeutiques. Une preuve de concept a ainsi été réalisée en parvenant à sélectionner au sein du génome modèle du Cricket Paralysis Virus (CrPV), une des IRES contenue par ce dernier (Jang *et al.*, 1990). A terme, cette étude peut se projeter sur l'étude de génome viraux (tels que le Zika ou le SARS-Cov2) afin de mettre en évidence de nouvelles séquences IRES encore inconnues.

Résultats et Discussion

1. Etude de l'initiation de la traduction procaryote

2. Régulation de la transcription et de la traduction par les riboswitches

3. Etude de l'initiation de la traduction chez les eucaryotes

4. Initiation *via* les IRES

Résultats

L'objectif de ma thèse a été de mettre en place et de valider un processus de criblage le plus universel possible, dans le but d'étudier l'initiation de la traduction tant chez les procaryotes que chez les eucaryotes. Pour réaliser cet ambitieux projet et mettre au point le processus de criblage, l'ensemble des expérimentations a tout d'abord été réalisé à l'aide d'extraits d'*E. coli* (extrait S30 commercial). La traduction bactérienne étant moins complexe que la traduction eucaryote, elle servira de preuve de concept de notre stratégie expérimentale. Les extraits en provenance des cellules eucaryotes ne seront employés qu'une fois le procédé validé dans le but d'étendre le spectre d'applications de notre technologie.

Nous avons choisi de débuter la mise en place de la technologie à l'aide d'un système déjà bien caractérisé mais restant le plus simple possible. Notre choix s'est donc porté sur le site de liaison au ribosome (RBS) bactérien qui permettra d'établir une preuve de concept robuste avant de poursuivre avec des éléments régulateurs plus complexes comme par exemple les riboswitches.

Par la suite, ce processus sera adapté et appliqué à des extraits eucaryotes dans le but d'étudier les mécanismes d'initiation de la traduction qui sont plus complexe chez ces organismes. Un premier objectif a été l'étude de la séquence bordant le codon initiateur qui est impliquée dans l'initiation dite coiffe-dépendant. Un second objectif m'a amené à étudier un mécanisme alternatif de l'initiation eucaryote impliquant cette fois-ci des éléments structurés d'entrée interne des ribosomes (IRES) permettant d'initier la traduction de façon coiffe-indépendante.

1. Étude de l'initiation de la traduction procaryote

Comme mentionné plus haut, j'ai tout d'abord mis en place et validé la stratégie expérimentale avec l'étude de la séquence de recrutement du ribosome procaryote. Cependant, bien que la nature de cette séquence module l'efficacité d'initiation de la traduction d'un ARNm, ce n'est pas le seul élément impliqué et des séquences beaucoup plus complexes peuvent également entrer en jeu afin d'ajuster finement et de façon dynamique l'expression des gènes aux besoins de la cellule (2.1.2). Un cas particulièrement fascinant de séquences régulatrices en *cis* est représenté par les riboswitches qui seront étudiés un peu plus bas. Enfin, bien qu'ils ne seront pas étudiés plus en profondeur ici, une part importante des régulations bactériennes est assurée par de petits ARN régulateurs, souvent non-codants et

agissant en *trans* (Dutta and Srivastava, 2018; Desgranges *et al.*, 2019; Chiaruttini and Guillier, 2020).

1.1 Étude de l'initiation dite "canonique" *via* le RBS

Chez les procaryotes la séquence appelée RBS est le principal acteur contrôlant l'efficacité d'initiation de la traduction d'un ARNm (Rodnina, 2018). Comme mentionné dans l'introduction (1.2.1), ses caractéristiques (tailles, teneur en purines, distance du codon initiateur) permettent d'en moduler la capacité à recruter le ribosome et donc son efficacité à initier la traduction de la séquence en aval. Une séquence consensus a été décrite chez la souche d'E. *coli* par Shine et Dalgarno (SD) dans les années 70 suite à une longue caractérisation manuelle (J Shine and Dalgarno, 1974a); un travail chronophage et fastidieux, qui a mis en avant cet appariement de bases direct entre l'ARNm et l'extrémité 3' de l'ARNr 16S. Par la suite, d'autres travaux ont complété ce travail séminal, en particulier des études exploitant des stratégies d'analyse à haut-débit telles que la cytométrie en flux, notamment le FACS (de l'anglais "Fluorescence Activated Cell Sorter") (Kosuri *et al.*, 2013; del Campo *et al.*, 2015; Cambray, Guimaraes and Arkin, 2018; Komarova *et al.*, 2020; Kuo *et al.*, 2020). Ces études ont toutes été réalisées à l'aide de cellules vivantes et, bien que les résultats obtenus soient absolument pertinents biologiquement, ces travaux s'exposent aux difficultés liées au contexte cellulaire : complexité d'interprétation du phénotype observé due à l'interconnexion des voies d'expression génique, possible effets indirects du traitement appliqué, toxicité de séquences testées, nécessité de cloner et transformer chaque variant à tester. Ces différentes limitations peuvent cependant être résolues en utilisant des extraits acellulaires, pourvu qu'un format d'analyse adapté soit disponible (Contreras-Llano and Tan, 2018).

1.1.1 Choix de l'extrait

J'ai débuté mes travaux avec des extraits reconstitués (système PurExpress), car cette approche semblait très prometteuse en termes de contrôle des conditions et de reproductibilité des expérimentations. Pour rappel, chaque élément de ce système est purifié indépendamment permettant ainsi de contrôler la quantité et la qualité de chacun et de s'assurer un haut degré de reproductibilité. Cependant, la viscosité relativement importante de ce système de TTIV s'est trouvé être un facteur limitant lors de son utilisation en puces microfluidiques. En effet, je me suis retrouvée confrontée à de fortes difficultés à pincer et segmenter la phase de TTIV par la phase d'huile, conduisant à un phénomène de « Jetting ». De plus, d'importants agrégats ont été retrouvés dans les gouttelettes en fin de fusion et

encore lors des tris, conduisant à la formation de bouchons dans les canaux. Enfin, en dépit de la haute reproductibilité attendue, les signaux obtenus en systèmes microfluidiques lors des premiers essais d'analyse de banques modèles, ont montré une forte variabilité d'un tour à l'autre, et des doutes sur la stabilité des extraits ont donc été émis. Ces différents problèmes liés à l'utilisation des systèmes reconstitués m'ont donc conduit à préférer utiliser des extraits bruts de S30 plus conventionnels qui, pour leur part, se sont révélés facilement émulsifiables et ont permis l'obtention de données nettement plus reproductibles.

1.1.2 La preuve de concept

Lors de mes travaux de thèse le premier point a consisté à mettre au point mon approche microfluidique avec pour but d'être en mesure de distinguer des variants de RBS d'après leur capacité à initier la traduction de façon différentielle d'un ARNm codant pour une protéine rapportrice, la GFPmut2. L'ensemble des travaux réalisés et des résultats collectés sont recueillis dans le manuscrit reproduit ci-après.

1.1.2.1 Manuscrit en préparation : “Functional selection of sequences controlling translation initiation using droplet-based microfluidics”

Functional selection of sequences controlling translation initiation using droplet-based microfluidics

Natacha Dentz, Roger Cubì, Cédric Romilly et Michael Ryckelynck

Functional selection of sequences controlling translation initiation using droplet-based microfluidics

Natacha Dentz, Roger Cubì, Cédric Romilly et Michaël Ryckelynck

Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR 9002, Strasbourg, F-67000, France

Introduction

Over the past decades, cell-free expression systems (i.e., *in vitro* transcription and/or translation) gained an increasing place in biological sciences (Noireaux and Liu, 2020). Among other significant achievements, they allowed to describe biological mechanisms as important as alcoholic fermentation (Buchner Eduard, 1897) and were also instrumental to the elucidation of the genetic code (M. Nirenberg et al., 1965). Whereas original cell-free protein synthesis systems had limited performances (i.e., moderate batch-to-batch reproducibility, short reaction half-life and elevated cost) (Silverman et al., 2020), constant improvement in cell extract preparation and buffer formulation make current systems much more robust and widened their application scope (Perez et al., 2016). Cell extracts can basically be prepared from any cell type provided it can be cultured to produce the required starting material. Besides, reconstituted cell-free translation systems, in which each recombinant protein factor is purified prior to being formulated together with other components, are expected to offer even much higher reproducibility and control over reaction conditions (Shimizu et al., 2001). Though most cell-free applications are devoted to protein production, these systems can also be used, for instance, to study and decipher translation (and/or transcription) related mechanisms (Noireaux and Liu, 2020) as well as to implement synthetic genetic circuits (Kelwick et al., 2016; Shin and Noireaux, 2012).

Genetic circuits can be rationally designed and later optimized by modulating the sequence of key elements (e.g., transcription promoter, ribosome binding-site, regulatory regions). This would ideally require to functionally analyze a large number of sequence permutations (hundreds, thousands if not more; a number that exponentially increases with the length of the varied sequence) to maximize the chance of finding an element with the properties best matching those expected. Even though the use of microtiter plates and robotic liquid handlers may help assisting such experiments, they become rapidly very laborious, and cost prohibited. A first solution came with the introduction of *in vitro* Compartmentalization (IVC) by Tawfik and Griffiths, (Tawfik and Griffiths, 1998) a technology in which genes are individualized and expressed within small (i.e., femtoliter scale) water-in-oil droplets.

The strong miniaturization offered by this emulsion-based approach allows millions of genes to be analyzed in a single experiment in an inexpensive way. Whereas the original IVC was limited by emulsion polydispersity and the difficulty to modify droplet content after they have been formed, its adaptation to droplet-based microfluidics (also referred to as microfluidic-assisted IVC, or µIVC in short (Ryckelynck et al., 2015)) allowed the full potential of the method to be exploited. Indeed, and as deeper exemplified below in this article, the use of microfluidics allows to produce highly homogeneous picoliter (pL) droplets, but also to modify their content on demand (Sohrabi et al., 2020). Moreover, when used in tandem with cell-free expression systems (either reconstituted (Holstein et al., 2021), or based on bacterial (Hori et al., 2017; Mazutis et al., 2009a) or eukaryotic extracts (Pernod et al., 2020)) µIVC enables the ultrahigh-throughput production and functional analysis of protein synthesis (Hori et al., 2017) and/or enzyme activity (Mazutis et al., 2009a). Furthermore, the analysis can simply be applied to toxic or insoluble proteins, and unnatural or isotope-labeled amino acids can easily be added to the mixture (Contreras-Llano and Tan, 2018). Finally, complex multi-component molecular circuits can also be expressed, and their formulation optimized in such droplets (Hori et al., 2017).

In the present work we went one step beyond the state-of-the-art by evaluating the capacity of µIVC to isolate sequences for their capacity to modulate translation initiation efficiency. As a model system for this proof-of-concept, we reexplored the nature of the sequences able to attract the ribosome of *Escherichia coli*. Though sequence variations occur, the Ribosome-Binding Site (RBS) is best prototyped by the Shine and Dalgarno (SD) consensus sequence (5'- AGGAGGU-3') identified in early 70's (Shine and Dalgarno, 1974a) and that interact with a complementary region present at the 3' end of 16S ribosomal RNA (5'- ACCUCCU-3') to facilitate ribosome recruitment. Interestingly, non-SD sequences were also proposed to work as an RBS, though they were quite distant from the consensus (Omotajo et al., 2015; Saito et al., 2020). Beside the sequence itself, its direct environment (e.g., distance to the start codon, identity of the nucleotides surrounding the start codon, overall structuration of the region) can also directly influence the strength of an RBS (Barrick+ et al., 1994; del Campo et al., 2015; Komarova et al., 2020; Kuo et al., 2020a; Maarten H. De Smit and J. Van Duin, 1990; Ringquist et al., 1992). Compiling the knowledge collected on RBS and combining them with thermodynamic parameters (e.g., energy pairing with the 16S RNA) eventually led to the release of predictive algorithms evaluating the likelihood of a sequence to act as an RBS and scoring its strength (Salis et al., 2009). More recently, high-throughput methodologies in which cytometry was used in tandem with Next Generation Sequencing (NGS) shed further light on the overall translation initiation mechanism by characterizing the capacity of each of the 61 possible codons to behave as a start one (Hecht et al., 2017) or even the influence of RNA structuration and of the identity of the first codons (Cambray et al., 2018). Most of the data available so far were either derived from genome-wide bioinformatic analyses or collected from cell-expressed mutant libraries, and we thought that a cell-free ultrahigh-throughput approach like µIVC could nicely complement existing methodologies by enabling to comprehensively analyze large mutant libraries in a simple way and well controlled *in vitro* conditions benefiting from a

limited droplet-to-droplet variation. Therefore, we generated a library in which the 9 positions of the RBS were randomized, prior to functionally analyze the more than 260,000 different sequences permutations using µIVC-seq, an approach combining µIVC, NGS and bioinformatics (Bouhedda et al., 2021). Altogether, our data are in very good agreement with the knowledge from the literature, confirming that µIVC-seq is well-suited to characterize translation control mechanism, paving the way toward the characterization of more complex regulatory mechanisms.

Material and methods

Mutant library preparation

First, randomized RBS was added on the sequence of the *gfpmut2* (Cormack et al., 1996) by PCR, reaction mixture contains 15 pmol of each primer (forward RBS random: GAGACCACAACGGTTTCCCTCCTTTANNNNNNNNNTATACCATGACCAGCTACCCATACG ATGTTCC and reverse GFPmut2 (RevGFPmut2): TCTCGTGGGCTCGGAGATGTGTATAAGAGACAG), 1 ng of plasmid, 0.2 mM of eachs dNTPs (ThermoScientific 10 mM each), 4 U of Q5® High-Fidelity DNA Polymerase (New England Biology, NEB) and the corresponding buffer (NEB). The mixture was thermocycled starting with an initial step of denaturation of 2 min at 98 °C followed by 25 cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C to finish it. Then, PCR product was purified with DNA clean up kit (Wizard® DNA Clean-Up Promega) and 1 ng of it is amplified in a second PCR reaction with exactly the same condition that the first one, but with the use of another forward to add promoter T7 and UDI (Fwd UDIT7) (Unique Droplet Identifier) sequence (TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAANNNNNGATCNNNNNTGACNNNNN AACATCGTCCACATAATACGACTCACTATAGGGAGACCACAACGGTTTCCCTCC).

Microfluidics screening

Microfluidic chips were molded into polydimethylsiloxane (PDMS) and electrodes were fabricated as described in Ryckelynck et al., 2015.

For the Droplet digital PCR, the libraries were diluted in 200 µg/mL yeast total RNA solution (Ambion) to obtain the desired droplet occupancy. 1 µL of this dilution was introduced in 100 µL of PCR mixture containing 15 pmol of forward i2 (Fwd i2) (TCGTCGGCAGCGTCAGAT) and Rev GFPmut2, 0.2 mM of each dNTPs, 10 µM Cyanine 5 (Thermo Fisher), 0.1 % Pluronic F68 (Sigma), 2 U of Q5 DNA polymerase (NEB) and the corresponding buffer at the recommended dilution. The mixture was loaded into a PTFE tubing (Thermo) and infused into a droplet generator microfluidic chip where it was dispersed in 2.5 pL droplets carried by an HFE 7500 fluorinated oil (3 M) supplemented with 3 % of a surfactant. Droplet production frequency (~10,000 droplets per second) was monitored in

real time using an optical device and a software developed by the team (Ryckelynck et al., 2015). 2.5 pL droplets were generated by adjusting pump flow rates (MFCS, Fluigent). The emulsion was collected into 0.2 mL tube and subjected to an initial denaturation step of 2 min at 98 °C followed by 25 PCR cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C

For the addition of *in vitro* expression mixture (IVTT) by droplet fusion, PCR droplets were reinjected into a droplet fusion device at a rate of ~1500 droplets per second as described previously (Ryckelynck et al., 2015) PCR droplets were spaced by a stream of HFE 7500 fluorinated oil (3 M) to synchronized and paired one-to-one with a 17,5 pL *in vitro* transcription translation S30 extract (*E. coli* S30 Extract System for Linear Templates Promega) supplemented with 0.1 % of Pluronic F68, 1 µM of Cy5, 17.5 µg/mL T7 RNA polymerase (purified in the laboratory). *IVTT* mixture was loaded in a length of PTFE tubing and kept on ice during the experiment. IVTT droplets were produced using an HFE 7500 fluorinated oil (3 M) stream supplemented with 3 % (w/w) of surfactant. Flow rates (MFCS, Fluigent) were adjusted to generate 17,5 pL *IVTT* droplets and to maximize the synchronization of 1 PCR droplet with 1 *IVT* droplet. Pairs of droplets were then fused with an AC field (50 mV at 30 kHz) and the resulting emulsion was collected in a tube and then incubated for 2 h at 37 °C.

For the droplet fluorescence analysis and sorting, the emulsion was finally re-injected into an analysis and sorting microfluidic device (Ryckelynck et al., 2015). Droplets were injected at a frequency of ~150 droplets per second and spaced with a stream of HFE 7500 fluorinated oil (3 M). Green (Green fluorescent protein GFPmut2) and red (Cyanine 5) fluorescence of each droplet was analyzed. In rounds of selection, the greenest droplets displaying a red fluorescence corresponding to single fused droplets (red box in Figure 1) were targeted and deflected into the collection channel by applying an AC field (1200 mV 30 kHz) and collected into a 2 mL tube. Sorted droplets were recovered from the collection tubing by flushing 200µL of HFE 7500 fluorinated oil (3 M), then 150 µL of 1H, 1H, 2H, 2H-perfluoro-1-octanol (Sigma-Aldrich) and 200 µL of 200 µg/mL yeast total RNA solution. The droplets were broken by vortexing the mixture. DNA-containing aqueous phase was finally transferred to a new tube.

An aliquot of DNA-containing aqueous phase was treated as described previously but with the Fwd (bc) to both amplify the material and reset the UDI carried by each molecule before using these molecules to achieve a new round of screening.

Enrichment test

An aliquot (3 µL) of DNA-containing aqueous phase recovered at the end of each round of screening (or from the starting library R0) was amplified into 100 µL of PCR mixture containing 15 pmol of fwd i2 and Rev GFPmut2, 0.2 mM of each dNTP, 2 U of Q5 DNA polymerase and the corresponding buffer at recommended concentration. The mixture was thermocycled with an initial denaturation step of 1 min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 60 °C for 20 sec and 72 °C for 1 min 45, then 72 °C for 2 min. PCR products were finally purified using a “Sera-Mag" kit (Sera-Mag Select reagent Fisher) and quantified with a Nanodrop (Thermo Scientific).

2 µL of PCR product was introduced in an *in vitro* transcription translation (IVTT) mixture containing *in vitro* transcription traduction S30 extract (*E. coli* S30 Extract System for Linear Templates Promega) supplemented with 17.5 µg/mL T7 RNA polymerase (purified in the laboratory). The green fluorescence (ex: 488 nm/em: 535 nm) of GFP was then monitored every minute for 2 h at 37 °C on spectrophotometer (SpectraMax® iD3 Molecular Device).

Functional test

Some NGS analysis sequences of interest from all rounds of selection were TA-cloned, then PCR-amplified like in the preparation library part and GFPmut2 fluorescence is during two hours of transcription, translation monitoring at 37 °C like in enrichment test part. First, A base is added to the PCR product from reamplification of selection recovery sequence in a mixture which contains: 10 pmol of dntps, 5 U of DreamTaq™ (NEB) and the corresponding buffer, at 95°C for 5min and 72 °C for 20 min. Then, PCR-Adenylated was purified by a Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR. PCR purified products were inserted in pTZ57R/T vector following manufacturer's instructions (InsTAclone PCR cloning Kit, Thermo-Scientific). Ligation products were recovered by phenol/chloroform extraction and 400 ng of DNA used to transform Electro-10 blue bacteria (Agilent) placed in 2 mm electroporation (MicroPulser, Bio-Rad). After an hour at 37 °C under agitation in 2YT, bacteria were plated on 2YT-Amplicillin agar plate and incubated overnight at 37 °C. The colonies were picked, used to inoculate liquid 2YT and grown at 37 °C until saturation. Plasmids DNA were extracted using 'GenJet Plasmid Miniprep kit' (Thermo-Scientific), and sequences determined by Sanger approach (GATC Biotech).

NGS library preparation

An aliquot (to obtain the desired droplet occupancy) of DNA-containing aqueous phase recovered at the end of each round of screening (or from the starting library) was amplified into 100 µL of digital PCR mixture containing 15 pmol of Fwd i2 and Reverse addi1 (Rev addi1) GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTAATCTGGAACATCGTATGGG), 0.2 mM of each dNTP, 0.1 % of pluronic, 1 µM of Cyanine 5, 2 U of Q5 DNA polymerase and the corresponding buffer at recommended concentration (NEB). The emulsion of 2.5 pL droplets was thermocycled with an initial denaturation step of 1min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 60 °C for 20 sec and 72 °C for 1 min, then 72 °C for 2 min. PCR products was recovered after droplet broken with 1H, 1H, 2H, 2H-perfluoro-1-octanol (Sigma-Aldrich) and then, were finally purified using Sera-Mag™ kit (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR and quantified with a Nanodrop (Thermo Scientific). Then, Illumina-index (Nextera XT DNA Library Preparation Kit Illumina) was added in digital PCR mixture which contains 1 ng of digital PCR i1/i2added purified, 10 µL of both Index Illumina forward and reverse, 0.2 mM of eachs dNTPs (ThermoScientific 10 mM each), 4 U of Q5® High-Fidelity DNA Polymerase (NEB) and the

corresponding buffer (NEB). The mixture was thermocycled with an initial denaturation step of 1min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 62 °C for 20 sec and 72 °C for 1min15, then 72 °C for 2 min. PCR products was recovered after droplet broken with 1H, 1H, 2H, 2H-perfluoro-1-octanol (Sigma-Aldrich) and then, were finally purified using Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR and quantified with a Nanodrop (Thermo Scientific).

NGS analysis platform

Sequencing data were analyzed using a custom Python bioinformatic pipeline in 8 main steps (Supplemental Fig. 2). First, fastq files were parsed using the Biopython library and only reads with a Q-score > 30 were conserved for the rest of the analysis (step 1). Then, UDI and 9-mer randomized regions were extracted from each read (step 2). Sequences with an occurrence below that threshold were likely mutants (raised from PCR or sequencing errors) and were no longer considered for the rest of the analysis (step 3). Moreover, sequences displaying mutations outsides of randomized regions (i.e. the UDI and the randomized RBS) were also filtered out. Next, identical sequences with different UDI and the expected length were clustered together, and their occurrence measured (step 4) while the 9-mer randomized sequences of each selection round were isolated in parallel and used to analyze the enrichment of motifs using the MEME analysis suite (step 5). RNAup from the ViennaRNA Package was used to calculate the free energy and the number of nucleotide pairs formed between the randomized RBS sequences and the 3'-end of 16S *E. coli* rRNA (5'-GAUCACCUCCUUA-3') (step 6). Nucleotide sequences were codified in three-dimensional trajectories (TDT) vector as described in Lo et al., 2007 (step 7). Using the SOMPY python library, a Self-Organizing Map (SOM) of the sequence space was trained using the sequences TDT vectors generated in step 7, eventually appending a combinatorial mix of the parameters determined at the step 5 (presence of the motif identified at the step 5 codified in binary format (0 or 1)) and the step 6 (the free energy and the number of nucleotide pairs formed with the 16S rRNA) to the vector (step 8). A grid of 50x50 neurons with randomly weights generation was selected to represent the sequence space. To train the model we used a rough train with 40 iterations with a radius of 10 followed by 80 finetune train iterations with a radius of 4. Next, a fitness landscape was constructed from the nodes grid generated by the SOM algorithm (step 9). For each node a fitness value (Z axis of the fitness landscape) was calculated by adding the sum of the occurrence frequency of all the sequences present on that node. Finally, neighboring nodes sharing a high fitness were clustered together in view of further analyzing their sequence content and features.

Results and discussion

Mutant RBS gene library design and validation

To reach an ultrahigh-throughput regime, μIVC requires a rapid and easy to automate functional assay, ideally based on fluorescence. We therefore prepared a construct in which the DNA sequence coding for the Green Fluorescent Protein (GFP) was placed under the control of T7 RNA polymerase promoter (Figure 1A) to enable its efficient transcription. Next, we defined that the stretch of 9 base-pairs (bps) found 6 bps upstream the start codon will constitute the RBS, while the composition of the 6-bp linker was biased in favor of pyrimidines in the RNA. Finally, we introduced upstream the whole construct a Unique Droplet Identifier (UDI), a sequence (5'-ANNNNNGATCNNNNNTGACNNNNNAACAT-3') made of 3 stretches of 5 randomized bps interspersed by 4 constant bps and used for bioinformatic data treatment (Autour et al., 2019; Bouhedda et al., 2021).

We first validated the concept by using as RBS either the consensus sequence of Shine and Dalgarno (5'-AGGAGGAGT-3') or a pyrimidine-rich sequence (5'-TTCTTCTTC-3') as negative control (Figure 1A). Both constructs were expressed in a S30 extract-based *in vitro* coupled transcription/translation (IVTT) system, and the green fluorescence apparition monitored. As expected, a strong fluorescence was only observed in the presence of the consensus whereas the negative control did not yield a fluorescence emission significantly above the background (No DNA condition), validating our overall design strategy and that translation initiation mainly rely on the 9-mer region defined as RBS (Figure 1B). We next prepared a mutant gene library in which the 9 positions of the RBS were randomized, each position having an equal probability to contain any of the 4 nucleotides. Since most of the sequences should not be efficient at attracting the ribosome, one would expect this library to generate a fluorescence level close to the background, in great agreement with what we observed. The 262,144 (4^9) different sequence permutations contained in this library were then functionally screened for their capacity to drive fluorescence production, so to initiate *gfp* translation.

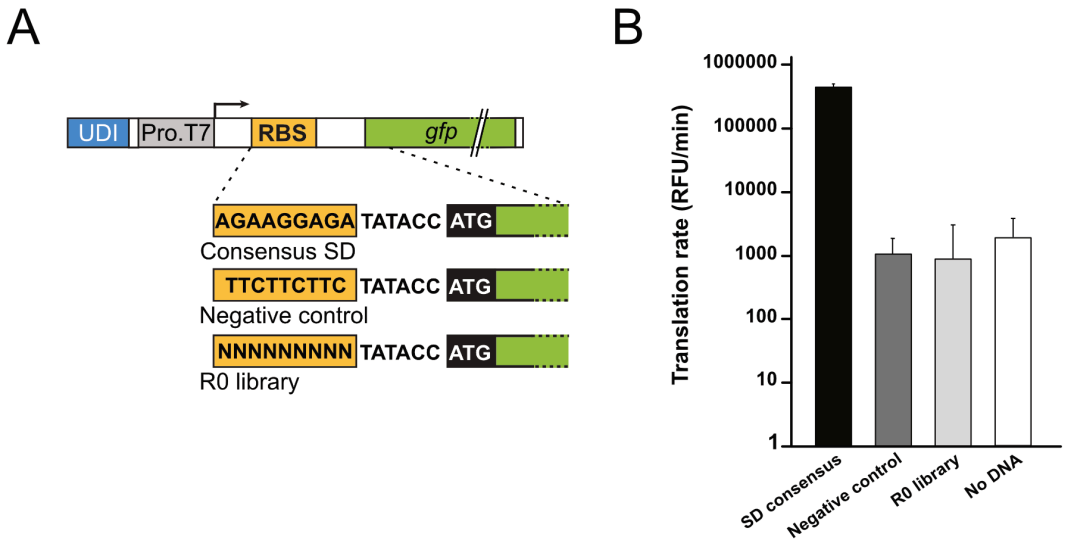


Figure 1: Validation of the library construction by expression in cell-free conditions and fluorescence monitoring. A. Schematic representation of DNA constructs and mutant library. The Unique Droplet Identifier (UDI) corresponding to three stretches of 5 consecutives random bases separated by 3 constant regions, the T7 RNA polymerase promoter drive the transcription of the constructs, while the Ribosome Binding Site (RBS) with the Shine and Dalgarno consensus sequence was used as the positive control or a pyrimidine-rich sequence was used as negative control. The starting "R0" library was made of a construct in which the 9 positions of the RBS were randomized. **B.** Fluorescence production monitoring. Each construct shown in A was *in vitro* expressed while monitoring GFP fluorescence apparition rate. A control without DNA added to the expression mixture was performed in parallel. The values are the mean of 3 independent experiments and the error-bars correspond to ± 1 standard deviation.

Ultrahigh-throughput in vitro functional screening of RBS randomized libraries

The mutant library was screened using a typical 3-step µIVC procedure (Figure 2A) (Mazutis et al., 2009a; Pernod et al., 2020). Briefly, the DNA molecules contained in the library were diluted in a PCR amplification mixture prior to being emulsified into ~ 2.5 pL droplets. Knowing that DNA molecules distributes in droplets following Poisson law (Mazutis et al., 2009a), the dilution of the DNA was adjusted to have an average of one DNA molecule per droplet for the first round of screening to maximize the fraction of occupied droplets (~ 63 %), so the total number of analyzed genes, while maintaining multiple encapsulation events at an acceptable rate (~ 60 % of the occupied droplets contained exactly one DNA molecule). The next rounds were performed using an average of 0.2 DNA molecule per droplet to gain in measurement accuracy (less than 2 % of the droplets contained more than one DNA molecule per droplet). Upon emulsification, droplets were thermocycled to PCR amplify their DNA content. Then, the droplets were reinjected into a droplet fusion device in which they were synchronized and fused one-to-one with larger (17.5 pL) droplets containing an IVTT mixture. Droplets were again collected and incubated to allow DNA to be transcribed and the resulting mRNAs to be translated into GFP all the better the sequence upstream the GFP-coding is an efficient RBS. Finally, droplets were reinjected into a fluorescence-activated droplet sorting device (Baret et al., 2009) in which the green (GFP) and the red (Cy5 added to droplets to discriminate droplets for the number of PCR droplets they were fused to (Ryckelynck et al., 2015)) fluorescence of each droplet was analyzed and

used to sort them accordingly. Depending on the round, the 0.5 to 1 % most green fluorescent droplets were sorted (red boxed populations on Figure 2B and Supplementary Figure 1), the DNA was recovered and used to prime the next round of screening. 4 rounds of screening were performed, and the experiment was made in duplicate.

The efficiency of each round of screening was validated prior to start the next one by PCR amplifying the DNA recovered from the sorted droplets and expressing it by IVTT while monitoring GFP fluorescence apparition rate (Figure 2C). In both replicates (A and B) the overall fluorescence gradually increased throughout the process. This observation was strongly supportive of an overall success of the process. Upon the fourth round of screening, the average fluorescence of the enriched library (Figure 2C) approached that measured with the SD consensus (Figure 1B). We decided to stop the process at this stage, we indexed the DNA contained in each library (the 8 enriched libraries and the starting one) prior to pooling then and to analyzing them on a MiSeq Next Generation Sequencing platform.

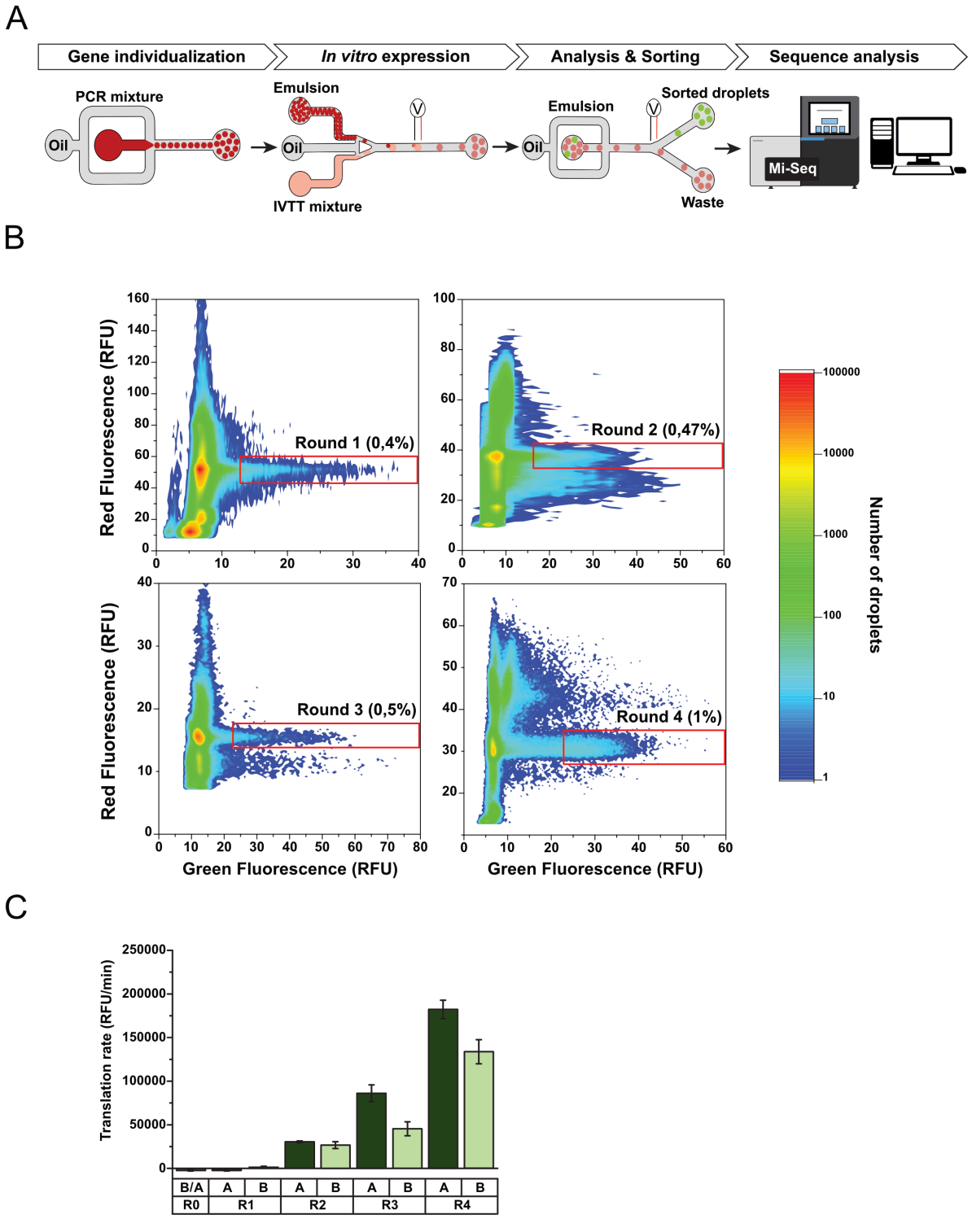


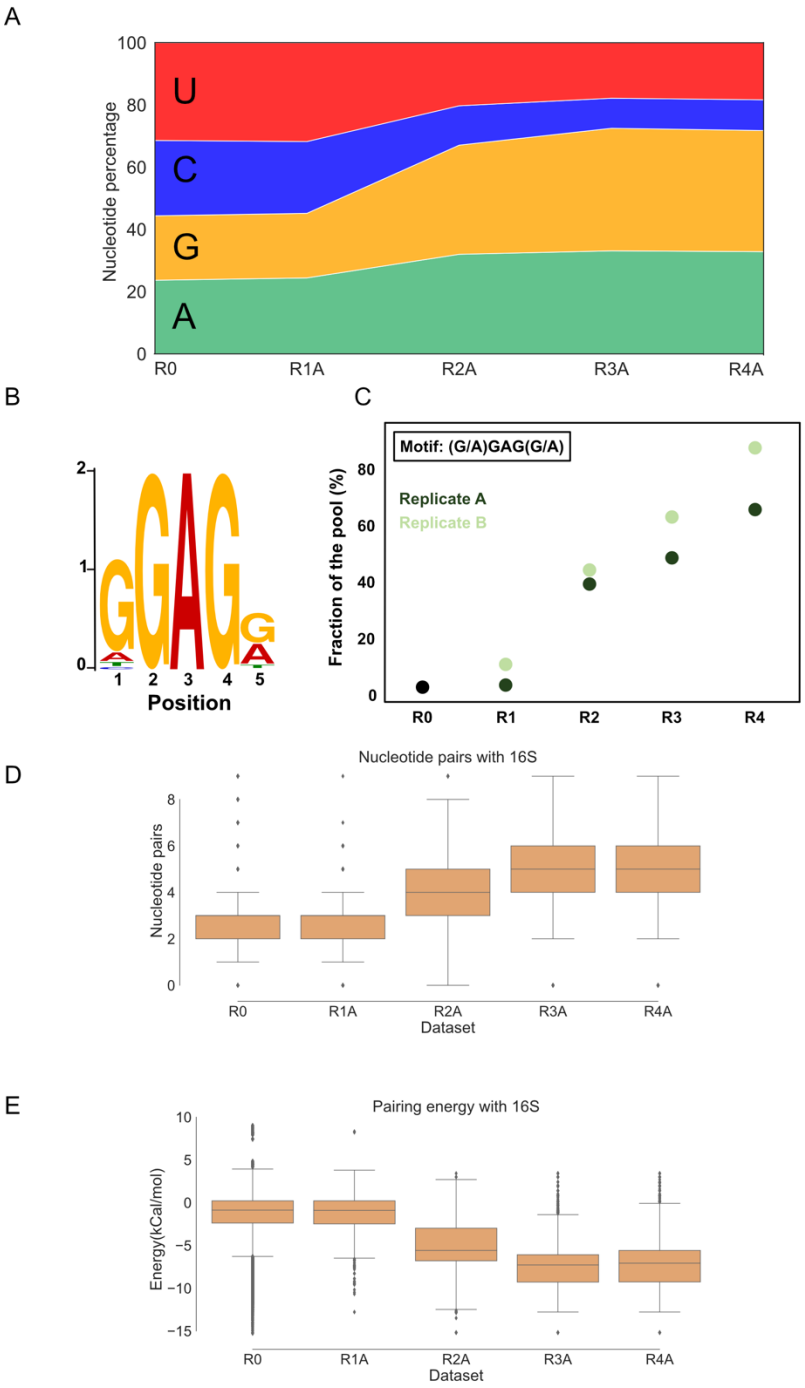
Figure 2: Functional screening of the gene library. A. Overview of the screening pipeline. The different steps are shown. DNA molecules contained in the library are individualized during PCR droplets production. Next, after thermocycling, each PCR droplet is merged with an IVTT one. Then, upon incubation, the red (Cy5) and green (GFP) fluorescence of each droplet is analyzed and used to sort them for the capacity of the tested sequence to initiated translation. Finally, at the end of the process, the DNA content of all the libraries is analyzed by NGS and bioinformatics. **B.** Fluorescence profiles recorded during the screening steps of replicate A. The sorted droplets are boxed in red and the percentage of the population they represent is given. **C.** Translation rate of the different libraries. Starting an enriched libraries of both replicates were in vitro expressed while monitoring GFP

fluorescence apparition. The translation rate was computed as RFU/min and the values are the mean of 3 independent experiments and the error-bars correspond to ± 1 standard deviation.

A motif fitting the Shine and Dalgarno consensus dominates the selected populations

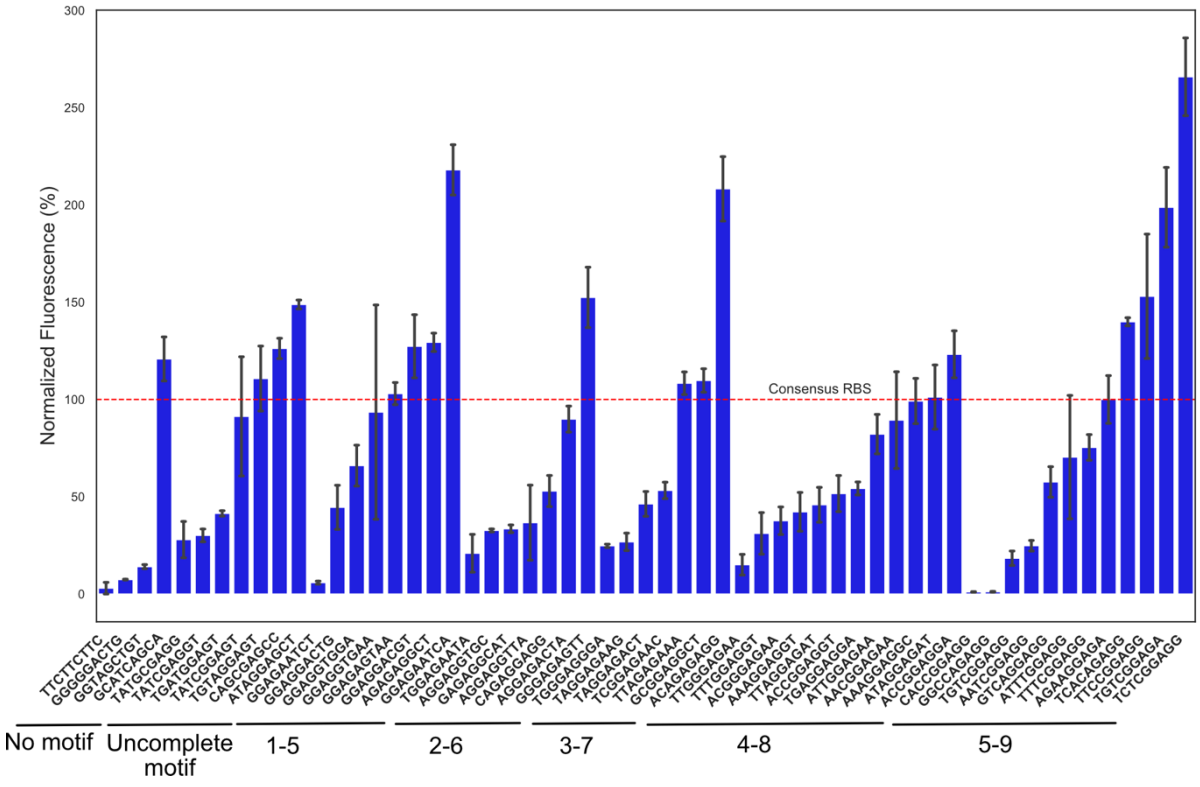
The sequence content of each library was analyzed using a bioinformatic pipeline (Supplementary Figure 2) similar to the one we previously described (Bouhedda et al., 2021). Briefly, the sequence of the UDI and that of the randomized 9-mer were extracted from each read free of unwanted mutations outside these regions. Using the UDI, we next removed the under-represented sequences expected to result from PCR or sequencing errors (Autour et al., 2019; Bouhedda et al., 2021). Importantly, since the UDI has a theoretical diversity (4^{15} or 10^9 possible sequence permutations) much higher than the number of analyzed droplets (less than 10^7) it was unlikely that the same UDI was present in two different droplets. Therefore, counting the number of different UDIs associated to each 9-mer enabled to determine the number of droplets that contained each 9-mer and, this way, precisely determine the occurrence frequency of each variant. This value allowed the enrichment of each sequence to be computed throughout the whole process.

A first rapid analysis revealed a progressive and significant enrichment of the 9-mer in purines, especially Gs (Figure 3A), as typically expected for sequences behaving as RBS. Further analyzing the 9-mer sequences conserved in each library with the MEME suite allowed to identified the motif 5'-(G/A)GAG(G/A)-3' as the dominant sequence in the different libraries of both replicates (Figure 3B). Indeed, starting from the second round of screening, this motif was found in nearly half of the sequences and reached 60 to 80 % by the fourth round (Figure 3C). Excitingly, this motif perfectly matches the consensus originally described by Shine and Dalgarno and later confirmed by others (Cambray et al., 2018; Komarova et al., 2020; Kuo et al., 2020; Ringquist et al., 1992; Shine and Dalgarno, 1974). Consistently, all along the process, the libraries of both replicates tended to get enriched in sequences forming the more and more stable interactions with the 16S rRNA as evidenced by the increasing number of pairing established between the 9-mer and the 16S rRNA together with the gradual decrease of free energy resulting from the interaction (Figure 3 D et E, Supplementary Figure 3). Interestingly, both the free energy of the duplex and the number of base pairs involved in its formation remains stable between the third and fourth rounds of selection. This is in correlation with previous studies showing that a too stable duplexes between the SD and the anti-SD prevents 30S from transitioning from the initiation and the elongation phase of translation (Takahashi et al., 2013). We further confirmed the importance of this motif by testing the functionality of variants possessing all or part of it (Figure 4). As expected, while sequences displaying the full-length motif allowed an efficient synthesis of GFP, single point mutations were enough to reduce it, confirming the instrumental role of the motif.



<u>Figure 3:</u> Analysis of the sequence contained in the different libraries. A. Evolution of the global nucleotide composition of the 9-mer over the selection process. **B.** Logo of the motif found enriched at the R4 selection round using the MEME analysis suite. **C.** Evolution of the fraction of the sequences containing the motif (5'-(G/A)GAG(G/A)-3'). **D.** Number of nucleotide pairs formed between the sequence 5'-GAUCACCUCCUUA-3' of the 3'-end of *E. coli* 16S rRNA and the 9-mer randomized sequences at the different selection rounds of the replicate A. **E.** Free energy of the RNA-RNA interaction between the sequence 5'-GAUCACCUCCUUA-3' of the 3'-end of *E. coli* 16S rRNA and the 9-mer sequences. Both, number of pairs and free energy were calculated using the Vienna RNA package RNAup.

Interestingly, a closer examination of the data sets revealed a different average position of the motif in the libraries in both replicates (Supplementary 4 and Figure 4). Indeed, whereas the motif tended to be located 8-7 nucleotides upstream the start codon in the first replicate, it was shifted 10 nucleotides back in the second replicate. The explanation of this difference would require additional investigations that were out of the scope of the present work. Yet, it is tempting to propose that a bias (e.g., a subsampling) occurred early in the process and that a handful of molecules then rapidly dominated their respective pool. It is noteworthy that despite a variation in its position, the same motif is found conserved in both replicates, confirming the prime importance of the sequence and the lower contribution of its distance to the start codon, an observation in agreement with the data from the literature (Ringquist et al., 1992). We further experimentally confirmed this assertion by testing variants displaying the motif at various positions and found that, as expected, the location of the motif had no significant effect on its functionality (Figure 4 and Supplementary 4).



<u>Figure 4:</u> Functional analysis of selected RBS variants with respect to the SD consensus sequence (5'-AGGAGGTAA-3'). Sequences are organized with respect to the motif (5'-(G/A)GAG(G/A)-3') position over the 9 randomized nucleotides. Values are the mean of 3 independent experiments and the error-bars correspond to ± 1 standard deviation.

We finally tried to identify additional factors that may drive the function of the motif by using a machine learning approach we recently successfully applied to guide the engineering of light-up RNA aptamers (Cubi et al., 2021). In this approach, called µIVC-Useq, the Self Organizing Map deep learning

algorithm uses a set of parameters supplied as a vector to organize the sequence space. As in our previous study, each nucleotide sequence was coded as a three-dimensional trajectory (TDT) and additional parameters (e.g., number of residues paired to the 16S rRNA, free energy pairing or presence of the motif) were then sequentially appended to the vector (Supplementary Figure 5). The TDT alone allowed by itself to distribute the 411 sequences found at the end of the fourth round into 42 clusters, highlighting similarities between the sequences of the pool. Adding the number of base pairs established with the 3' end 16S rRNA or the free energy of the 9-mer and the 3' end of 16S rRNA did not significantly change the degree of clustering whether these parameters were considered independently or together. However, adding a simple parameter informing on the presence, or on the absence, of the motif had a significant impact by reducing the number of clusters to 23, with 30 % of the sequences closely associated in the sequence space. Appending other parameters had only a limited effect on sequence clustering. Taken together, these data suggest that, in the context of the mRNA we designed here, the sequence of the 9-mer but also the presence of the 5'-(G/A)GAG(G/A)-3' motif were the main driving forces involved in the binding of the ribosome. Interestingly, this conclusion is in full agreement with a recent study that also evaluated the RBS fitness landscape but using living cells (Kuo et al., 2020), which further confirms the relevance of the data collected here in cell-free systems and brings a final validation of our overall ultrahigh-throughput analytical strategy.

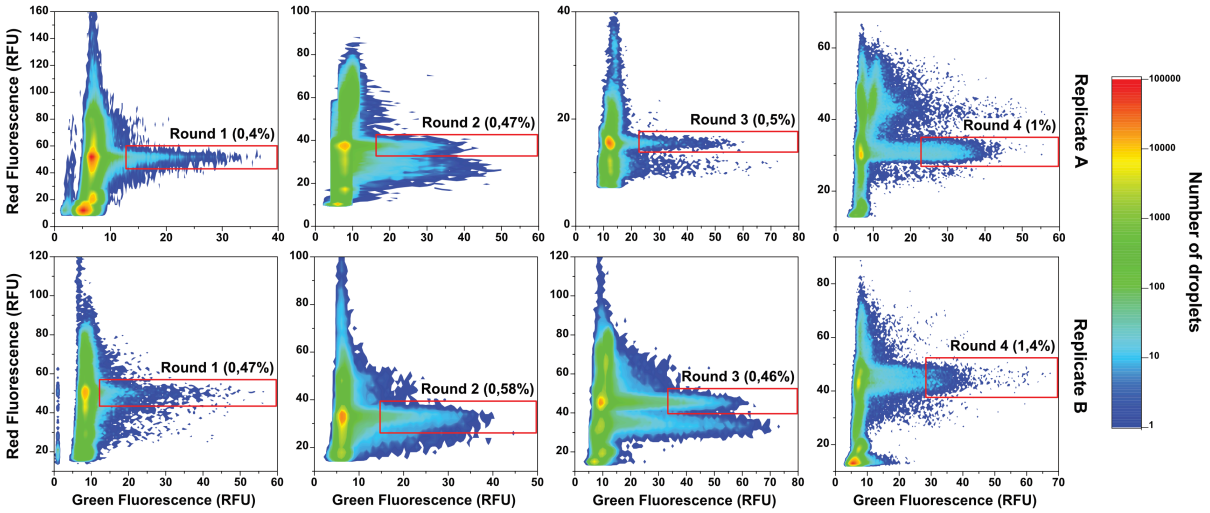
Conclusions

Over the past decade, microfluidics triggered several breakthroughs in biological sciences thanks to the strong volume reduction and parallelization it allows. Perhaps the biggest achievement so far has been the advent of next generation sequencing technologies that allow a whole genome to be rapidly analyzed in an inexpensive manner. A second acceleration has now started with droplet-based microfluidics and the extreme miniaturization it makes possible while preserving a prefect control over reaction conditions. Being able to sequentially and precisely act on droplets after their formation allows complex experimental procedures to be set-up. This is well exemplified by µIVC (Ryckelynck et al., 2015), a technology in which millions of genes of a library are individually amplified and *in vitro* expressed within water-in-oil droplets. Further devising a fluorogenic assay even allows to establish the phenotype of each gene and to sort the droplets accordingly. By dropping the cost of experiments while boosting their throughput, microfluidic-assisted technologies like µIVC allowed biology to enter a new dimension by making it possible to perform analyses otherwise impossible to perform.

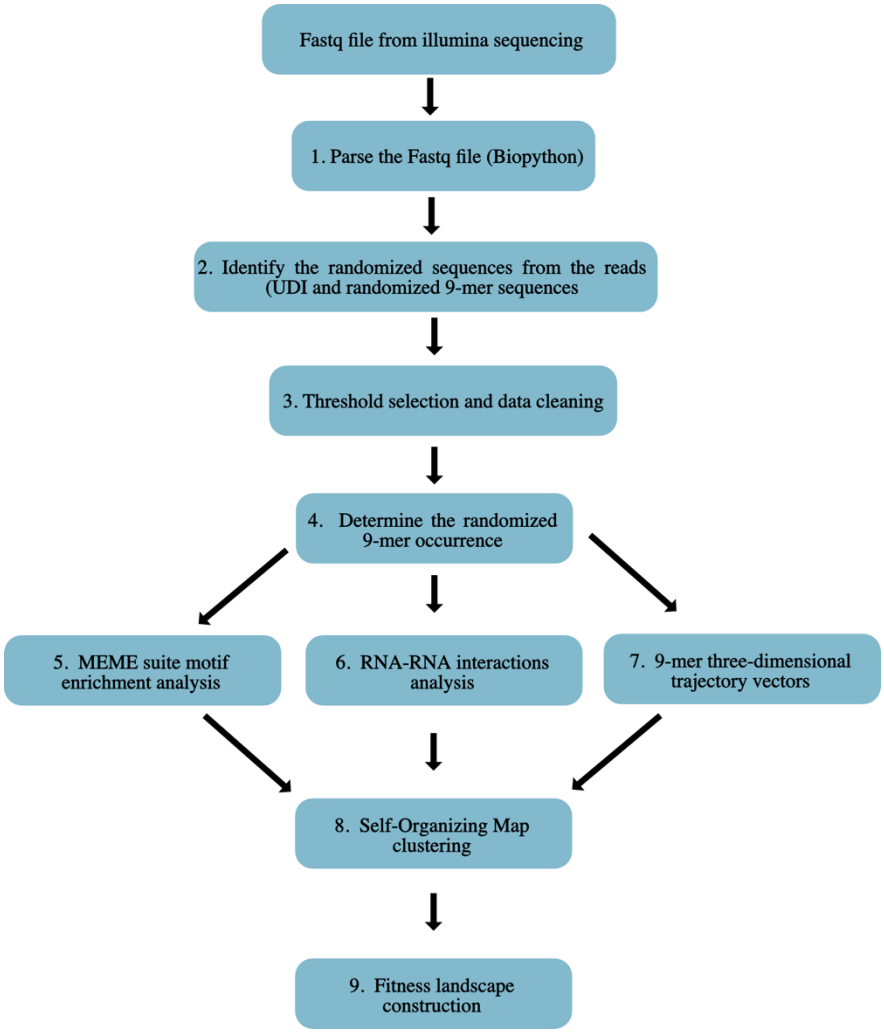
µIVC was initially developed to search for improved RNAs (Autour et al., 2016; Ryckelynck et al., 2015) and proteins (Holstein et al., 2021). Others used droplet microfluidics to assemble genetic circuits and optimize the formulation of their components (Dubuc et al., 2019; Hori et al., 2017). Yet, so far and to the best of our knowledge, the technology was never used to analyze and/or assist the

development of prokaryotic non-coding regulatory RNA sequences *in vitro*. As a first proof of principle, we re-explored the fitness landscape of *E. coli* ribosome-binding site using a construct in which the 9 nucleotides located 6 nucleotides upstream the start codon of *gfp* were randomized. Repeating the whole process twice led to the identification of the same motif that shares strong similarities with the consensus initially described by Shine and Dalgarno (Shine and Dalgarno, 1974) and tolerates variations both in its sequence and its distance from the start codon as already known in the literature (Cambray et al., 2018; del Campo et al., 2015; Hecht et al., 2017; Komarova et al., 2020; Ringquist et al., 1992). Moreover, though the central A of the motif appeared to be conserved, the sequence tended to display a strong bias toward G residues, a result consistent with a recent study that reexplored RBS landscape in living *E. coli*. Altogether, these data support the biological relevance of our cell-free approach and validate the whole process as a new methodology now available in the toolbox of synthetic biologists.

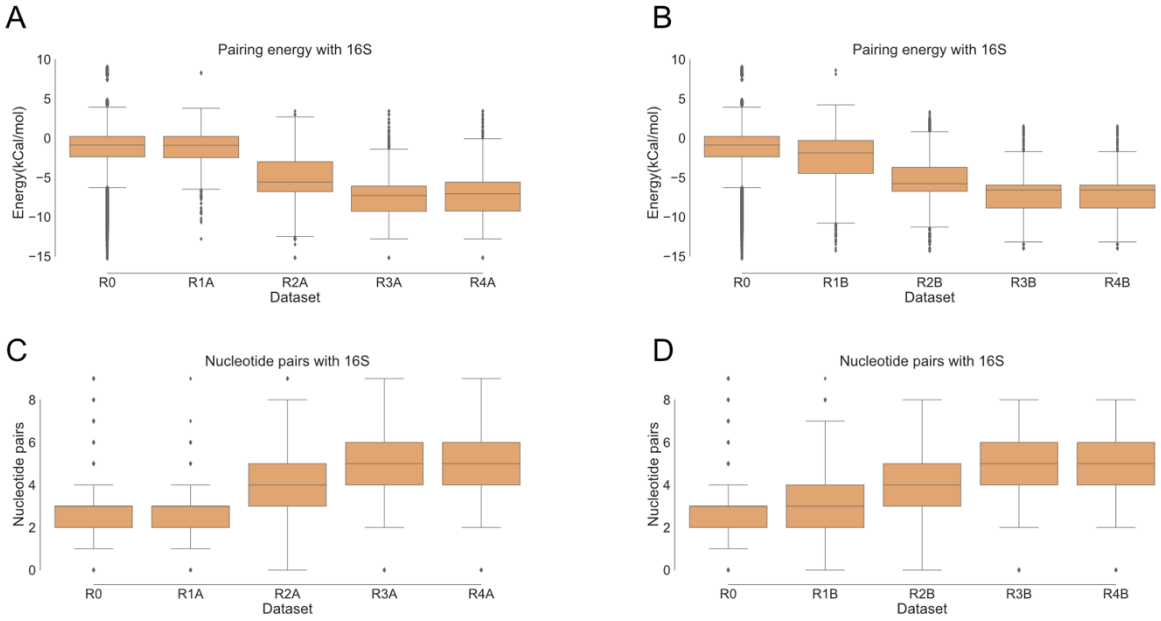
A first advantage of our approach over the use of living cells is the great theoretical droplet-to-droplet reproducibility or, said differently, the lack of cell-to-cell heterogeneity. Yet, the strongest benefits may come with the use of extracts produced from non-model cells, making dispensable the need for specific genetic tools to transform and manipulate the cells. Therefore, provided the organism can be cultured to produce an extract, the approach described in this article can easily by applied to any organism in a straightforward way. Though the technology was benchmarked using the well-characterized *E. coli* RBS to be able to take advance of the rich literature on the topic, it could easily be extended to the study of other regulatory RNAs acting either in *cis* (e.g., riboswitchs) or in *trans* (small non-coding RNAs) (Desgranges et al., 2019). The approach validated in this work offers an alternative way to rapidly collect data on existing systems or even efficiently assist the development of new synthetic regulatory molecules. Such data could also, for instance, be used to refine current predictive algorithms (Salis et al., 2009) and/or experimentally validate their predictions in non-model organisms, opening new axes of research in the rapidly expending field of synthetic biology.



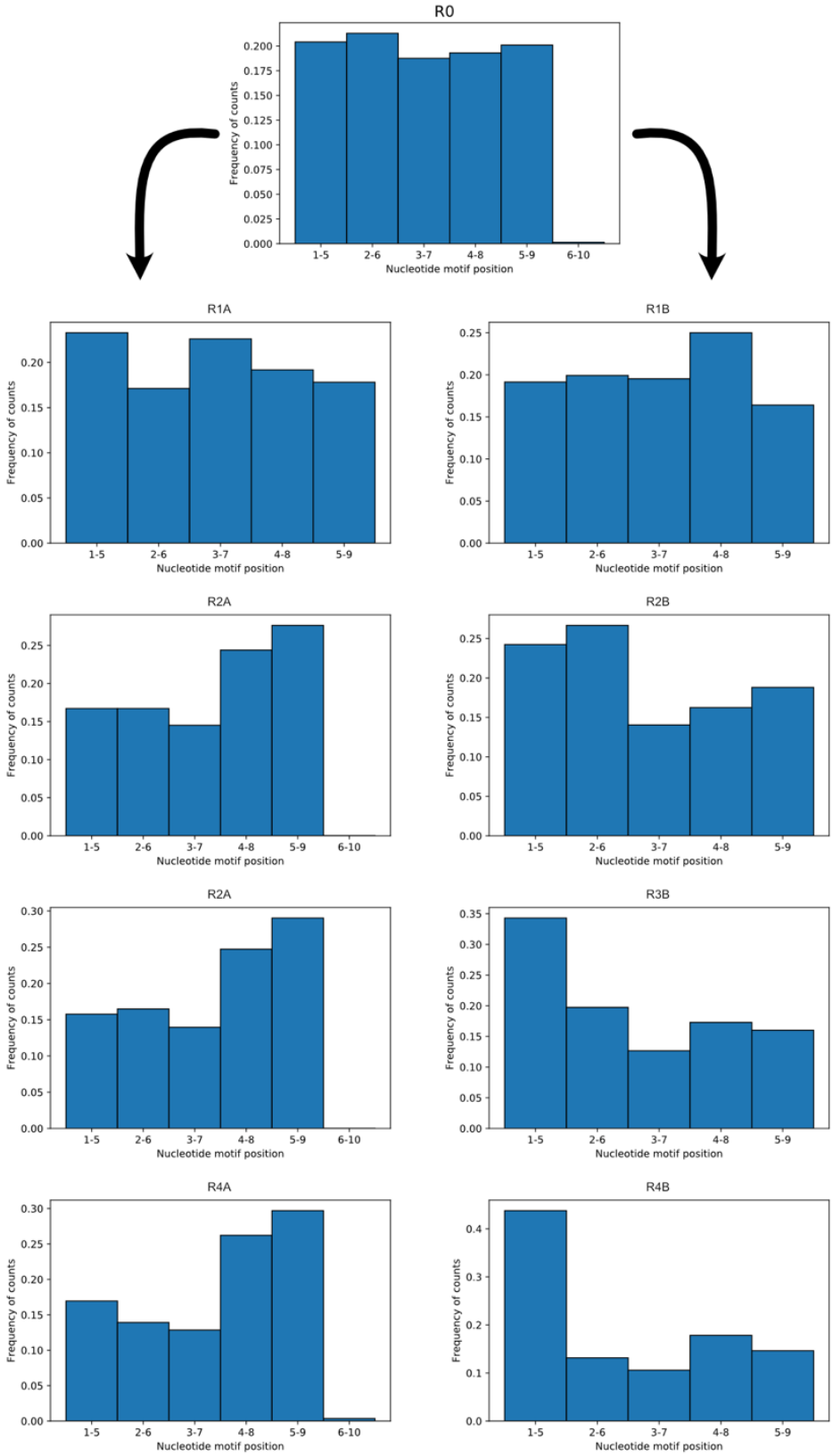
Supplementary Figure 1: Fluorescence profiles recorded during the screening steps of both replicate A and B. The sorted droplets are boxed in red and the percentage of the population they represent is given.



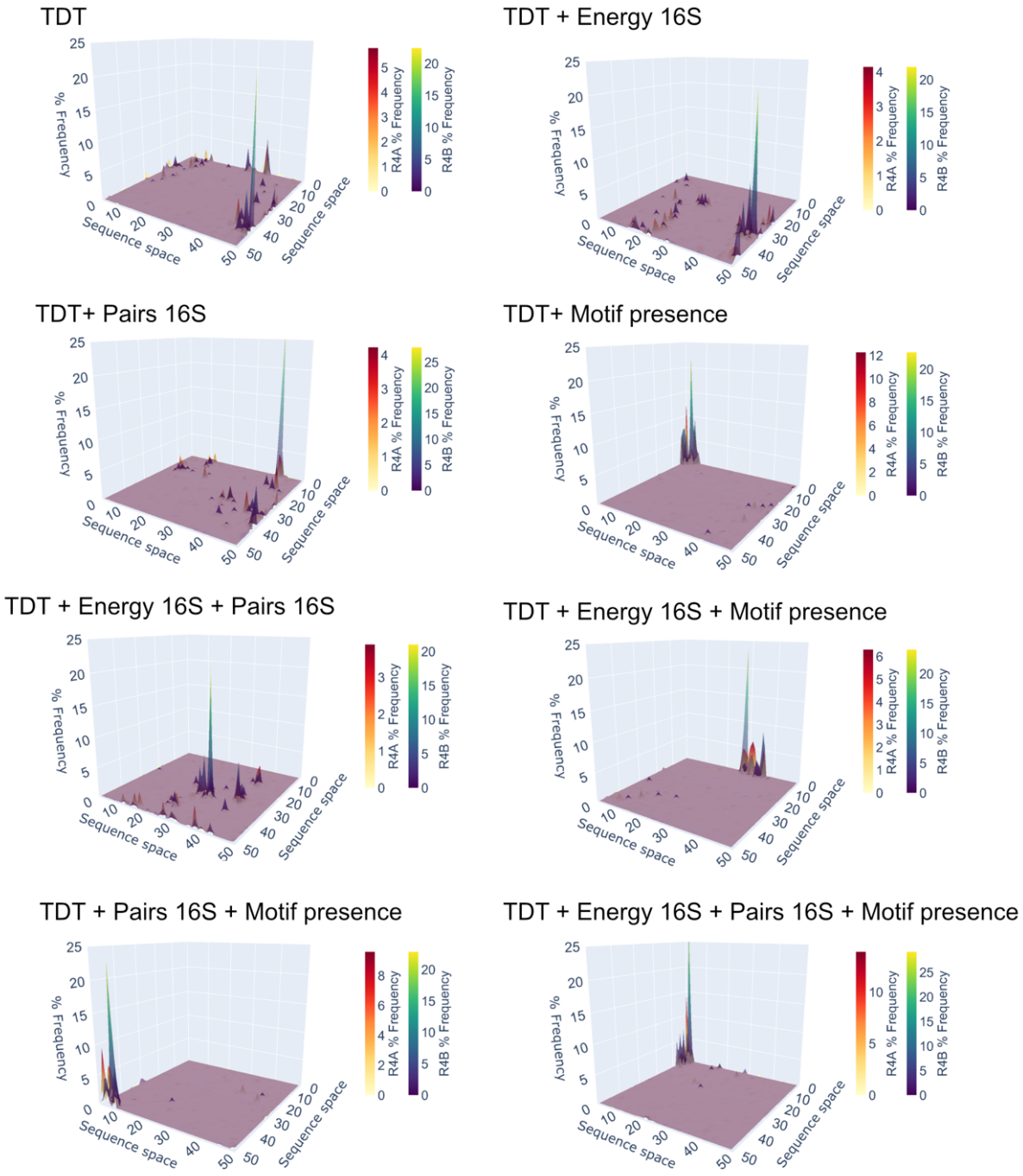
Supplementary Figure 2: Schematic of the bioinformatic pipeline used in the study.



<u>Supplementary Figure 3:</u> Free energy and number of pairs formed upon the RNA-RNA interaction of the sequence 5'-GAUCACCUCCUUA-3' of the 3'-end of *E. coli* 16S rRNA and the 9-mer randomized sequences. A and C correspond to the first replicate (A), whereas the B and C correspond to the second replicate (B). Both, number of pairs and free energy, were calculated using the ViennaRNA package RNAup.



<u>Supplementary Figure 4:</u> Position of the motif 5'-(G/A)GAG(G/A)-3' in with respect to the 9-mer randomized nucleotides thoughout the selection rounds in both duplicates.



<u>Supplementary Figure 5:</u> Fitness landscapes of the sequence contained in the round 4 of replicate A according to the parameters used to feed the SOM algorithm. The sequence was encoded as a Tridimensional trajectory (TDT) and coupled to other parameters in a single vector used to feed the SOM algorithm.

References

Autour, A., Westhof, E., and Ryckelynck, M. (2016). ISpinach: A fluorogenic RNA aptamer optimized for in vitro applications. Nucleic Acids Research *44*, 2491–2500.

Autour, A., Bouhedda, F., Cubi, R., and Ryckelynck, M. (2019). Optimization of fluorogenic RNA-based biosensors using droplet-based microfluidic ultrahigh-throughput screening. Methods *161*, 46–53.

Baret, J-C., Miller, Olivier J., Taly, V., Ryckelynck, M., El-Harrack, A., Frenz, L., Rick, C., Samules, M., Hutchison, J. B. (2009). Fluorescence-activated droplet sorting (FADS) efficient microfluidic cell sorting based on enzymatic activity. Lab on Chip *134,* 1092-1098

Barrick+, D., Villanueba, K., Childs, J., Kalilo, R., Schneiderl, T.D., Lawrence1, C.E., Gold, L., and Stormo, G.D. (1994). Quantitative analysis of ribosome binding sites in E.coli.

Bouhedda, F., Cubi, R., Baudrey, S., and Ryckelynck, M. (2021). µIVC-Seq: a method for ultrahigh-throughput development and functional characterization of small RNAs. Methods Mol. Biol. *2300*, 203–237.

Buchner Eduard (1897). Alkoholische Gärung ohne Hefezellen.

Cambray, G., Guimaraes, J.C., and Arkin, A.P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli. Nature Biotechnology *36*, 1005.

del Campo, C., Bartholomäus, A., Fedyunin, I., and Ignatova, Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. PLoS Genetics *11*.

Contreras-Llano, L.E., and Tan, C. (2018). High-throughput screening of biomolecules using cell-free gene expression systems. Synthetic Biology *3*.

Cormack, B.P., Valdivia, R.H., and Falkow, S. (1996). FACS-optimized mutants of the green fluorescent protein (GFP). Gene *173*, 33–38.

Cubi, R., Bouhedda, F., Collot, M., Klymchenko, A.S., and Ryckelynck, M. (2021). µIVC-Useq: a microfluidic-assisted high-throughput functionnal screening in tandem with next generation sequencing and artificial neural network to rapidly characterize RNA molecules. RNA.

Desgranges, E., Marzi, S., Moreau, K., Romby, P., and Caldelari, I. (2019). Noncoding RNA. Microbiology Spectrum *7*.

Dubuc, E., Pieters, P.A., van der Linden, A.J., van Hest, J.C., Huck, W.T., and de Greef, T.F. (2019). Cell-free microcompartmentalised transcription–translation for the prototyping of synthetic communication networks. Current Opinion in Biotechnology *58*, 72–80.

Hecht, A., Glasgow, J., Jaschke, P.R., Bawazer, L.A., Munson, M.S., Cochran, J.R., Endy, D., and Salit, M. (2017). Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Research *45*, 3615–3626.

Holstein, J.M., Gylstorff, C., and Hollfelder, F. (2021). Cell-free Directed Evolution of a Protease in Microdroplets at Ultrahigh Throughput. ACS Synthetic Biology acssynbio.0c00538.

Hori, Y., Kantak, C., Murray, R.M., and Abate, A.R. (2017). Cell-free extract based optimization of biomolecular circuits with droplet microfluidics. Lab on a Chip *17*, 3037–3042.

Kelwick, R., Webb, A.J., MacDonald, J.T., and Freemont, P.S. (2016). Development of a Bacillus subtilis cell-free transcription-translation system for prototyping regulatory elements. Metabolic Engineering *38*, 370–381.

Komarova, E.S., Chervontseva, Z.S., Osterman, I.A., Evfratov, S.A., Rubtsova, M.P., Zatsepin, T.S., Semashko, T.A., Kostryukova, E.S., Bogdanov, A.A., Gelfand, M.S., et al. (2020). Influence of

the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli. Microbial Biotechnology *13*, 1254–1261.

Kuo, S.T., Jahn, R.L., Cheng, Y.J., Chen, Y.L., Lee, Y.J., Hollfelder, F., Wen, J. der, and Chou, H.H.D. (2020a). Global fitness landscapes of the Shine-Dalgarno sequence. Genome Research *30*, 711–723.

Kuo, S.T., Jahn, R.L., Cheng, Y.J., Chen, Y.L., Lee, Y.J., Hollfelder, F., Wen, J. der, and Chou, H.H.D. (2020b). Global fitness landscapes of the Shine-Dalgarno sequence. Genome Research *30*, 711–723.

Lo, N.-W., Chang, H.T., Xiao, S.W., Li, C.H., and Kuo, C.J. (2007). Global Visualization and Comparison of DNA Sequences by Use of Three-Dimensional Trajectories *. JOURNAL OF INFORMATION SCIENCE AND ENGINEERING *23*, 1723–1736.

M. Nirenberg, P. Leder, M. Bernfield, R. Brimacombre, J. Trupin, F. Rottman, and C; O'Neal (1965). RNA codewords and protein synthesis, YII. On the general nature of the RNA code. J. Mol. Biol *53*, 1161–1168.

Maarten H. De Smit, and J. Van Duin (1990). Secondary structure of the ribosome binding site determines translational efficiency: A quantitative analysis (translational initlation/RNA-helfx stability). Proc. Nati. Acad. Sci. USA *87*, 7668–7672.

Mazutis, L., Araghi, A.F., Miller, O.J., Baret, J.C., Frenz, L., Janoshazi, A., Taly, V., Miller, B.J., Hutchison, J.B., Link, D., et al. (2009a). Droplet-based microfluidic systems for high-throughput single DNA molecule isothermal amplification and analysis. Analytical Chemistry *81*, 4813–4821.

Mazutis, L., Baret, J.C., and Griffiths, A.D. (2009b). A fast and efficient microfluidic system for highly selective one-to-one droplet fusion. Lab on a Chip *9*, 2665–2672.

Noireaux, V., and Liu, A.P. (2020). The New Age of Cell-Free Biology. Annual Review of Biomedical Engineering 52–68.

Omotajo, D., Tate, T., Cho, H., and Choudhary, M. (2015). Distribution and diversity of ribosome binding sites in prokaryotic genomes. BMC Genomics *16*.

Perez, J.G., Stark, J.C., and Jewett, M.C. (2016). Cell-free synthetic biology: Engineering beyond the cell. Cold Spring Harbor Perspectives in Biology *8*.

Pernod, K., Schaeffer, L., Chicher, J., Hok, E., Rick, C., Geslain, R., Eriani, G., Westhof, E., Ryckelynck, M., and 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 Research *48*, 6170–6183.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G.D., and Gold, L. (1992). Translation initiation in Escherichia coii: sequences within the ribosome-binding site. Molecular Microbiology *6*, 1219–1229.

Ryckelynck, M., Baudrey, S., Rick, C., Marin, A., Coldren, F., Westhof, E., and Griffiths, A.D. (2015). Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. RNA *21*, 458–469.

Saito, K., Green, R., and Buskirk, A.R. (2020). Translational initiation in E. Coli occurs at the correct sites genome-wide in the absence of mrna-rrna base-pairing. ELife *9*.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nature Biotechnology *27*, 946–950.

Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nature Biotechnology *19*, 751–755.

Shin, J., and Noireaux, V. (2012). An E. coli cell-free expression toolbox: Application to synthetic gene circuits and artificial cells. ACS Synthetic Biology *1*, 29–41.

Shine, J., and Dalgarno, L. (1974a). The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites (terminal labeling/stepwise degradation/protein synthesis/suppression).

Shine, J., and Dalgarno, L. (1974b). The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites (terminal labeling/stepwise degradation/protein synthesis/suppression).

Silverman, A.D., Karim, A.S., and Jewett, M.C. (2020). Cell-free gene expression: an expanded repertoire of applications. Nature Reviews Genetics *21*, 151–170.

Sohrabi, S., Kassir, N., and Keshavarz Moraveji, M. (2020). Droplet microfluidics: Fundamentals and its advanced applications. RSC Advances *10*, 27560–27574.

Takahashi, S., Furusawa, H., Ueda, T., and Okahata, Y. (2013). Translation enhancer improves the ribosome liberation from translation initiation. Journal of the American Chemical Society *135*, 13096–13106.

Tawfik, D.S., and Griffiths, A.D. (1998). Man-made cell-like compartments for molecular evolution. Nature Biotechnology *16*.

1.1.3 Perspectives directes découlant de ces travaux

Cette première étude s'est concentrée sur la séquence du RBS et plus précisément sur la séquence consensus de SD à proprement parler. Cependant, les autres caractéristiques clés de ce consensus (longueur, distance avec le codon initiateur) et l'impact de son environnement (proximité ou inclusion dans une structure secondaire, séquence codante en aval) pourraient également être ré-explorées au moyen de la même stratégie. En effet, ces différents éléments peuvent tout autant intervenir et potentiellement affecter l'initiation de la traduction laissant imaginer différents paramètres à tester, soit indépendamment, soit en combinant différentes variables comme la séquence de SD et sa distance avec le codon initiateur et ce, en mélangeant différentes constructions de départ incluant des séquences aléatoires placées plus ou moins loin du codon initiateur. Les sélections pourraient ensuite être répétées et servir de base de comparaison avec d'ancien travaux ayant déjà explorés certaines de ces caractéristiques comme, par exemple, l'étendue de la séquence consensus ou sa distance optimale du codon initiateur (Ringquist *et al.*, 1992; Komarova *et al.*, 2020). A ce jour, une étude a même combiné 4 paramètres différents dans la compilation des données (la composition en nucléotides, l'identité des codons de la séquence codante en aval, l'index d'hydropathie des acides aminés ainsi que la présence éventuelle de structures secondaires) permettant d'évaluer l'impact de chacun de ces paramètres indépendamment mais aussi en combinaison (Cambray, Guimaraes and Arkin, 2018). Ces travaux représenteraient une belle base de comparaison avec des données similaires que nous pourrions générer en utilisant la plateforme d'analyse que j'ai implémentée.

Sous sa forme actuelle, ma procédure permet uniquement de suivre la synthèse protéique à proprement parler. Or, il est tout à fait plausible que la nature d'une séquence donnée impacte non seulement la traduction de l'ARNm mais aussi la transcription du gène. Dans la configuration actuelle, les deux effets restent indiscernables. Une solution consisterait à intégrer dans la région 3' non traduite (3'UTR) de l'ARNm la séquence d'un aptamère fluorogène, en particulier un de ceux développés par l'équipe (Autour, Westhof and Ryckelynck, 2016; Trachman *et al.*, 2019; Bouhedda *et al.*, 2020). Ainsi, en choisissant un des aptamère et une protéine émettant dans une longueur d'onde orthogonale, il serait possible de suivre simultanément et de distinguer la transcription et la traduction. Cela permettrait notamment de normaliser l'efficacité de traduction de l'ARNm de chaque variant par celle de transcription de son gène et ainsi de discriminer d'éventuels cas où l'une des deux étapes est limitante, de cas où l'ensemble du processus est optimal. Deux essais préalables ont été réalisés dans ce sens à l'aide d'une construction où la phase codante pour la GFPmut2 était précédée de la séquence consensus AGGAGGU et possédait une séquence d'aptamère fluorogène dans la région 3' UTR (Figure 19).

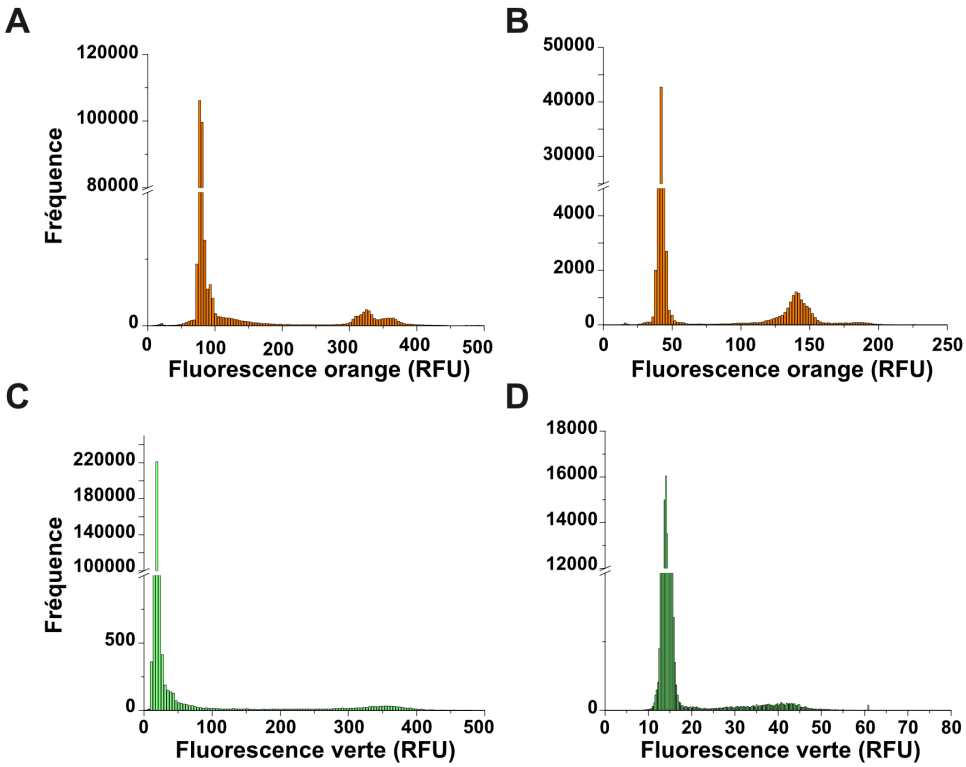


Figure 19 : Disparité de fluorescence après TTIV de la construction modèle contenant le témoin positif (AGGAGGU). A et B. Fréquence de la fluorescence orange obtenue pour le réplica A en A et B en B. **C et D.** Fréquence de la fluorescence verte obtenue pour le réplica A en C et B en D.

Étant donné que la GFPmut2 émet une fluorescence verte, j'ai ici choisi d'utiliser le système orange Gemini 561/o-Coral (λ_{ex} = 561 nm et λ_{em} = 630nm) récemment développé par l'équipe (Bouhedda *et al.*, 2020). Cependant, je n'ai pas réussi à obtenir de résultats présentant un degré de reproductibilité satisfaisant avec les extraits S30 d'*E. coli*, en particulier dans les systèmes microfluidiques, où comme on peut l'observer sur la Figure 19 les échelles et répartition de population sont variable entre les différents essais. Après analyse des données, une disparité de fluorescence est déterminée en divisant la fréquence moyenne de gouttelette ayant la fluorescence recherchée (exprimant la GFPmut2, mais aussi o-Coral et donc les valeurs de RFU les plus élevé) par l'écart-type de cette population d'intérêt. Cette disparité est alors d'environ 70 et 88% pour les fluorescences respectivement verte et orange du réplica A et de 164 et 160% pour le réplica B.

Plusieurs explications à ces faibles performances peuvent être proposées ici. Le couple o-Coral/Gemini 561 a déjà été utilisé pour la co-détection ARNm/protéine en intégrant l'aptamère o-Coral dans la région 3'UTR de l'ARN de l'eGFP, à 250 nucléotides en aval du codon stop (Bouhedda *et al.*, 2020), laissant amplement la place au ribosome pour arrêter la

traduction et quitter l'ARNm sans interférer avec la fonction de l'aptamère. Dans ma construction, le codon stop n'est distant de l'aptamère que d'une dizaine de nucléotides. Cette distance relativement réduite pourrait dans un premier temps conduire à une interférence entre l'aptamère et le ribosome lorsque ce dernier atteindrait la fin de la séquence à traduire. Si tel est le cas, une solution consisterait à utiliser un autre aptamère performant tel que Mango-III (Trachman *et al.*, 2019) ou iSpinach (Autour, Westhof and Ryckelynck, 2016) par exemple. Cependant, ces systèmes produisant des fluorescences vertes, il faudra également changer la protéine rapportrice pour une autre émettant dans le rouge telle que mScarlet ou mCherry. Une autre alternative intéressante serait l'utilisation d'un système protéique fluorogène tel que FAST (Plamont *et al.*, 2016). Entre autres intérêts, les longueurs d'ondes émises peuvent rapidement être modifiées en changeant le fluorogène présent dans le milieu. D'autre part, une fluorescence étant émissive dès la formation du complexe fluorogène/protéine, FAST permet de s'affranchir du temps de maturation nécessaire aux protéines fluorescentes conventionnelles qui peuvent atteindre plusieurs dizaines de minutes, voire des heures, en particulier avec les protéines émettant dans le rouge.

Une dernière perspective d'application immédiate de mon processus de criblage consisterait en son extension à d'autres organismes procaryotes, par exemple *Bacillus* ou *Staphylococcus aureus*. La possibilité d'approfondir les connaissances pour ces souches moins décrites et plus complexe à manipuler pourrait avoir un impact notoire dans les domaines de l'industrie (*Bacillus*) et médicaux (*Staphylococcus*).

Ainsi, ces premiers travaux au cours desquels j'ai ré-exploré la nature du site d'entrée de ribosome procaryote m'ont permis de mettre en place et valider ma procédure de criblage fonctionnel et d'en identifier les points qui nécessiteront de futures améliorations. Cette preuve de concept étant établie, il est à présent possible d'étendre cette stratégie de criblage à d'autres éléments de régulation plus complexes et moins bien caractérisés.

2. Régulation de la transcription et de la traduction par les riboswitches

La traduction procaryote est majoritairement régulée au niveau de son initiation, l'étape limitante du processus (1.2.1). Pour rappel, cette régulation peut avoir lieu grâce à divers mécanismes, notamment ceux impliquant des riboswitches. Ces structures sont présentes dans la région 5' UTR des ARNm et peuvent impacter l'étape de transcription ou de traduction. Les riboswitches transcriptionnels donnent lieu à un arrêt prématuré de la transcription par formation d'une tige boucle suivie d'une série d'uracile agissant comme un terminateur de

transcription Rho indépendante, selon la concentration de ligand et selon le type de riboswitch (un riboswitch appelé ON signifiant qu'il permet la transcription lorsqu'il interagit avec son ligand et à l'inverse, l'appellation OFF qu'il arrête prématurément la transcription lorsqu'il interagit avec son ligand). Les riboswitches traductionnels vont quant à eux moduler l'efficacité d'initiation de la traduction en séquestrant ou non le RBS selon le type de riboswitch (un riboswitch appelé ON signifiant qu'il permet l'accessibilité au RBS lorsqu'il interagit avec son ligand et à l'inverse, l'appellation OFF qu'il limite l'accessibilité au RBS lorsqu'il interagit avec son ligand) et la concentration de ligand (Husser, Dentz and Ryckelynck, 2021). Globalement, ces éléments permettent de répondre aux besoins de la cellule puisqu'ils sont souvent retrouvés en amont d'un opéron dont les gènes sont en relation avec la fonction ou la voie métabolique du ligand spécifiquement reconnu. Comme cela est présenté dans la revue que je cosigne (Husser, Dentz and Ryckelynck, 2021), les nombreux riboswitches découverts représentent non seulement des cibles extrêmement intéressantes pour la découverte de nouveaux antibiotiques, mais ils peuvent également constituer le point de départ du développement de versions synthétiques. Ces nouvelles versions peuvent opérer dans des cellules procaryotes mais aussi eucaryotes (par exemple pour le contrôle de l'expression de gènes de sécurité employés en thérapie génique) ou peuvent être reprogrammés pour répondre à de nouveaux ligands (avec un fort potentiel applicatif en biologie de synthèse par exemple). Notre technologie étant en mesure de sélectionner des molécules sur la base d'une modulation de l'expression génique *in vitro*, elle est donc toute indiquée pour la sélection de riboswitches synthétiques.

2.1 Une nouvelle stratégie expérimentale, de nouveaux acteurs : les riboswitches transcriptionels

L'étude de riboswitches régulant la traduction sous-entend que le contrôle de l'expression du gène en question doit aussi bien se faire au niveau de la transcription que de la traduction, rendant nécessaire de pouvoir suivre les deux phénomènes indépendamment. Cependant, les premiers essais peu satisfaisants et peu reproductibles de notre construction composée d'un rapporteur de la traduction (GFPmut2) et de la transcription (o-Coral) (1.1.3) m'ont conduit à me focaliser, dans un premier temps, uniquement sur l'utilisation d'éléments contrôlant la transcription. Ainsi, je me suis tout d'abord assurée de la possibilité de visualiser par fluorescence la modulation de la transcription médiée par un riboswitch en présence et absence de quantité suffisante de son ligand, avant d'essayer d'en améliorer éventuellement les performances par évolution *in vitro*. A terme, de telles expériences démontreraient la capacité de l'approche de sélection par microfluidique en gouttelettes à améliorer, voire à

développer de nouveaux riboswitches synthétiques en vue d'applications notamment en biologie de synthèse.

2.1.1 Riboswitch spécifique du FMN

Je me suis tout d'abord intéressée au riboswitch transcriptionnel répondant à la Flavin MonoNucléotide issus du génome de *B. subtilis* (connus sous le nom de RFN) et décrit comme exerçant un contrôle transcriptionnel de type OFF (Alexander S. Mironov *et al.*, 2002; Winkler, Cohen-Chalamish and Breaker, 2002) qui sera appelé dans ce manuscrit R-FMN. La présence de FMN conduit ainsi à l'extinction de l'expression de gènes codants pour des protéines impliquées dans la synthèse de la riboflavine (Vitreschak *et al.*, 2002; Atilho, Perkins and Breaker, 2019), suite à la formation d'une structure terminatrice de la transcription favorisée par la présence de FMN (Figure 20).

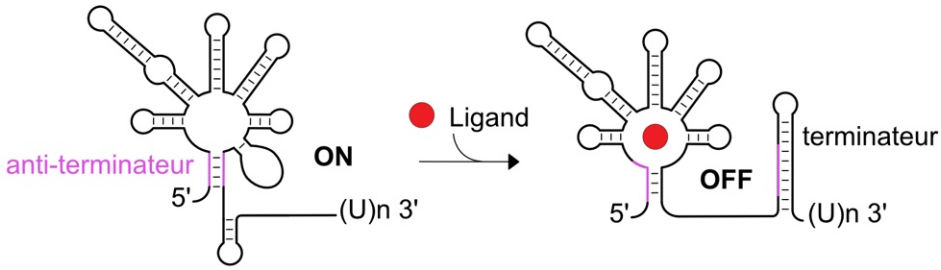


Figure 20 : Représentation schématique du fonctionnement du R-FMN. Le R-FMN est sous une conformation anti terminatrice en absence du ligand FMN (rond rouge) et à l'inverse en présence un terminateur transcriptionnel se forme en 3'. Les séquences impliquées pour former la structure anti-terminatrice sont en violet.

S'agissant d'une régulation transcriptionnelle, l'aspect cinétique est particulièrement important, l'ARN néo-transcrit doit disposer de suffisamment de temps pour adopter sa structure en interaction avec le ligand avant que la polymérase ne soit hors de portée de l'élément régulateur. De ce fait, le choix de la polymérase pour ces expérimentations est primordial. La polymérase la plus fréquemment utilisée en laboratoire est celle issue du phage T7, pouvant transcrire à une vitesse de 230 ribonucléotides (rNTP) par seconde contre seulement 49 rNTP par seconde pour l'ARN polymérase endogène d'*E. coli* (Proshkin *et al.*, 2010; Wang *et al.*, 2018). La polymérase phagique est donc vraisemblablement moins adaptée à ce type d'expérimentation, puisqu'il est fort probable que l'ARN n'aura pas le temps de se structurer convenablement et ainsi assurer sa fonction dans le court laps de temps octroyé par une transcription rapide. J'ai donc choisi de réaliser l'ensemble des transcriptions des

constructions contenant des riboswitches avec l'holoenzyme d'*E. coli (*soit le core enzymatique en association avec le facteur sigma 70*)* disponible commercialement.

2.1.2 Constructions et conditions d'expression

Le promoteur endogène fort T7A1, issu du génome d'*E. coli* (Sclavi *et al.*, 2005) a été ajouté en 5' de la séquence du R-FMN tandis que la construction a été flanquée en 3' de la séquence de l'aptamère fluorogène o-Coral 17 nucléotides en aval du R-FMN (Figure 21 A) afin de pouvoir suivre la transcription en temps réel grâce à l'émission d'une fluorescence orange après complexation de l'aptamère avec son fluorogène (Gemini-561) (Bouhedda *et al.*, 2020). Je me suis attelée premièrement à déterminer l'effet du ligand FMN sur l'ARN polymérase elle-même en réalisant le suivi de la transcription d'une construction comportant le promoteur T7A1 en amont de l'aptamère fluorogène o-Coral (Figure 21 A), et ce, en présence ou non de 100 µM de FMN. La fluorescence issue de o-Coral obtenue dans ces deux conditions (Figure 21 B) ne varie que très peu suggérant que le FMN ne pose pas de problème de compatibilité avec nos conditions de transcription et la polymérase (Figure 21 C et D). J'ai ensuite fait un premier essai de transcription de cette construction à l'aide d'une ARN polymérase holoenzyme d'*E. coli* commerciale (New England Biolabs) en présence ou en absence de 100 µM de FMN, et j'ai suivi l'apparition de fluorescence à l'aide d'un lecteur de microplaques (Figure 21 D). Ce suivi accompagné d'une vérification de la présence de transcrits à la bonne taille sur gel m'a permis de valider la capacité de cette polymérase à transcrire ma construction mais aussi a démontré la capacité du FMN à modifier la conformation du R-FMN limitant la transcription par un arrêt précoce de cette dernière avec l'observation sur gel de la présence de produit abortifs (Figure 21 C et D).

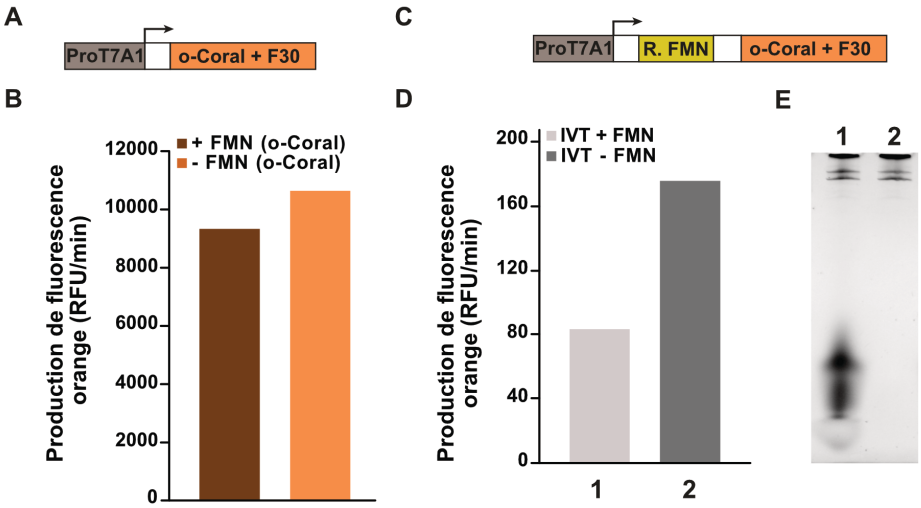


Figure 21 : Optimisation des conditions de transcription de la construction modèle par l'ARN polymérase d'*E. coli*. A. Représentation de la construction d'expression de o-Coral avec le promoteur T7A1 (boite marron) et en 3' l'aptamère fluorogène o-Coral (boite orange) et son scaffold F30 (de l'anglais : « échafaudage »). **B.** Suivi des effets du FMN sur l'efficacité de la transcription de l'aptamère o-Coral seul, sous le contrôle du promoteur T7A1. L'IVT, en microplaque, est réalisée avec (en orange) ou sans (marron) FMN. La transcription correspond à la fluorescence orange générée par l'aptamère fluorogène o-Coral, en fonction du temps (min). **C.** Représentation de la construction d'expression du riboswitch FMN. La construction contient le riboswitch FMN (boite jaune) représenté en jaune avec en 5' le promoteur T7A1 (boite marron) et en 3' l'aptamère fluorogène o-Coral (boite orange). **B.** Test de l'efficacité de transcription de la construction T7A1/R-FMN/o-Coral en présence ou absence du ligand FMN. La transcription est réalisée en microplaque et suivi en temps réel afin de suivre la transcription (ordonnée) en présence (1) ou en absence (2) de 100 µM de FMN. **C.** Gel de polyacrylamide dénaturant avec en 1 la production d'ARN en présence du ligand FMN et en 2 en absence de ce dernier.

J'ai ensuite cherché à déterminer l'effet de la concentration en FMN sur la transcription de la construction comportant le R-FMN (Figure 22). Ces essais préliminaires montrent une réactivité limitée de notre construction, puisque que des effets ne sont observés qu'à des concentrations importantes (10 et 100 µM) de FMN sachant que le Kd (mesure d'affinité) identifié pour l'aptamère des riboswitches FMN naturels de *E. coli* est de 5 nM (Wickiser *et al.*, 2005) De plus, même à la concentration la plus importante de FMN testée, une importante transcription de la matrice ADN (suivi par la fluorescence issue de o-Coral) est toujours observée, suggérant que les propriétés de la molécule peuvent être améliorées.

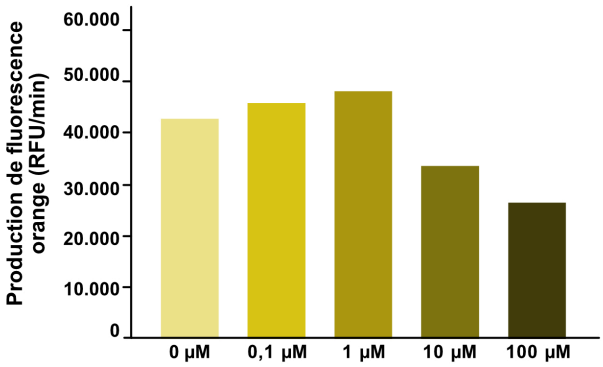


Figure 22 : Effet du FMN sur la transcription et sur le R-FMN selon la concentration. Différentes concentrations de FMN sont ajoutées dans le milieu d'IVT avec du vert plus clair au plus foncé, (0, 0.1, 1, 10, 100 µM). La transcription correspond à la fluorescence orange générée par l'aptamère fluorogène o-Coral situé en 3' du riboswitch, en fonction des minutes de transcription réalisées.

Je me suis ensuite concentré sur la possibilité d'optimiser les conditions expérimentales de la transcription en cherchant à améliorant la vitesse de transcription (réduction du temps d'expérimentation pour les manipulations futures en microfluidique) tout en maintenant l'effet régulateur de notre riboswitch. Comme on peut l'observer sur la Figure 23 A et B, le milieu minimum d'IVT conseillé par le fournisseur (MM NEB) semble bien permettre la production d'ARN à partir de notre construction d'ADN, confirmant encore une fois la fonctionnalité de notre construction avec l'holoenzyme d'*E. coli*. L'augmentation de la quantité d'enzyme ne semble pas affecter significativement l'efficacité de la transcription, indiquant que l'enzyme n'est pas le facteur limitant ici. En revanche, une augmentation de la concentration en rNTPs ou la suppression de la spermidine semble avoir un effet plus prononcé. Une quantité plus importante de ribonucléotides lors de la synthèse d'ARN pourrait être bénéfique en saturant davantage la polymérase et ainsi maximiser son activité tout en maintenant une vitesse de synthèse assez lente pour laisser au riboswitch le temps de se structurer. De manière étonnante l'absence de spermidine semble fortement favoriser la transcription (Figure 23 A), contrairement à ce qui est attendu sachant que la spermidine favorise la dissociation du complexe enzyme/ARN sans affecter le complexe enzyme/ADN, permettant une dynamique plus importante (Gumport, 1970). Cependant, en observant les produits de transcription sur gel aucune condition ne semble réellement se démarquer, suggérant le besoin de répéter les expérimentations (Figure 23 B). Enfin, dans des conditions combinant l'ensemble des paramètres soit, une quantité plus importante d'enzyme et de rNTPs et en absence de spermidine, la production d'ARN en temps réel semble moins efficace bien qu'encore une fois sur gel il est difficile d'observer des différences notoires (Figure 23 A et B). Il est possible que la quantité trop importante d'enzyme ainsi que les sels apportés par cette dernière limite fortement la transcription malgré les avantages apportés par les autres

paramètres. Dans l'ensemble de nos IVT impliquant l'holoenzyme d'*E. coli* la condition d'expression choisie ne comportera pas de spermidine mais une quantité de rNTP (6 mM) plus important que celle conseillée en attendant la répétition des données obtenues.

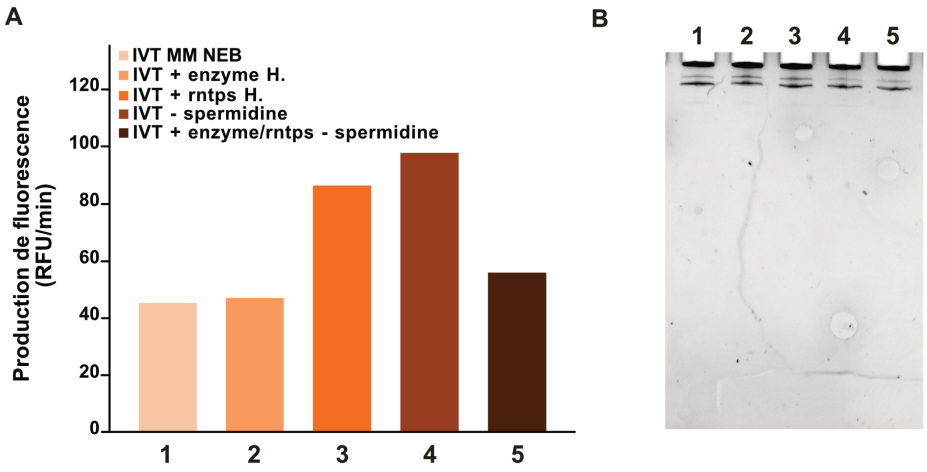


Figure 23 : Impact des conditions de transcription sur la production d'ARN et la réponse du R-FMN. **A.** Essais de transcription *in vitro* de R-FMN suivi d'o-Coral (Figure 21 A) en absence de FMN, suivis en temps réel en microplaque dans différentes conditions d'expression. Les couleurs désignent chaque condition utilisée avec du plus clair (1) au plus foncé (5) : (1) le mélange de transcription (MM NEB) recommandé par NEB (1 mM rNTP, 0.6 mM de Spermidine et 1.5 unités d'holoenzyme), (2) le MM NEB contenant une haute (H) concentration d'enzyme (3 unités) ; (3) le MM NEB avec une haute (H) concentration (6 mM) de rNTP ; (4) le MM NEB sans spermidine, et enfin (5) le MM NEB avec 3 unités d'enzyme, 6mM de rNTP et sans spermidine. En ordonné, la transcription est exprimée en fluorescence orange produite en RFU, par minute. Je précise que ces mesures de fluorescence ont été obtenues à l'aide de l'ancien spectrophotomètre (Agilent) moins sensible d'où la différence d'échelle des mesures. **C.** Gel de polyacrylamide dénaturant avec de 1 à 5 la production d'ARN des différentes conditions mentionnées lors du panel B.

2.1.3 Essais d'amélioration du riboswitch R- FMN

Comme démontré dans la section précédente, bien que fonctionnelle, la molécule R-FMN reste peu efficace dans sa capacité à arrêter la transcription en présence de ligand. Pour tenter d'améliorer ses propriétés, j'ai exploité une stratégie basée sur la microfluidique en gouttelette, mais en optant cette fois-ci pour une approche par évolution *in vitro* couplant mutagénèse et sélection *in vitro* (Figure 24).

Stratégie expérimentale

Ne connaissant pas les positions à muter pour améliorer la molécule, j'ai choisi une approche de mutagénèse aléatoire par PCR en présence d'analogues de nucléotides (dPTP

et 8-oxo-GTP) afin d'introduire en moyenne une à deux mutations par gène sur le riboswitch. La banque de mutants résultante a ensuite été analysée au moyen d'un pipeline de criblage similaire à celui utilisé précédemment pour la sélection des séquences de recrutement du ribosome. Cependant, s'agissant ici de sélectionner des molécules répondant à la présence d'un ligand, les sélections doivent être réalisées selon deux régimes : des sélections positives durant lesquelles sont sélectionnées les séquences permettant la transcription de l'aptamère fluorogène en absence de FMN (gouttelettes présentant une fluorescence orange forte), et des sélections négatives durant lesquelles sont sélectionnés les mutants inhibant efficacement la transcription de l'aptamère fluorogène en présence de FMN (gouttelettes présentant une fluorescence orange faible). Enfin, le résultat des sélections est ensuite analysé par séquençage haut débit et bio-informatique.

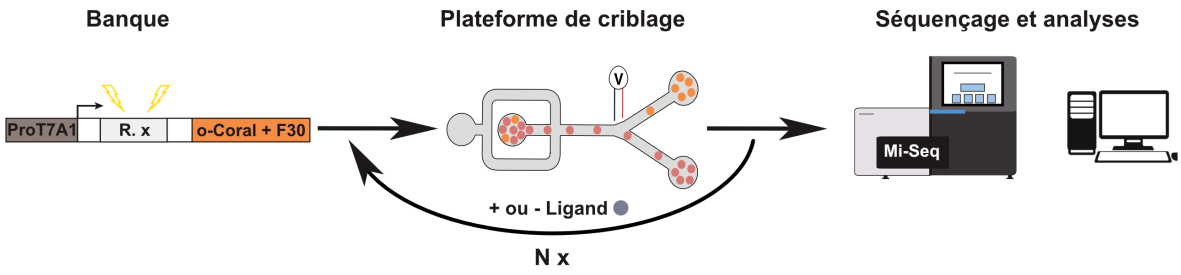


Figure 24 : Stratégie expérimentale employée pour la caractérisation de nouveaux variant de riboswitches. Le promoteur de l'ARN polymérase de *E. coli* est représenté en couleur marron, le riboswitch est représenté en gris et les mutations appliquées à ce dernier par les petits éclairs jaunes enfin, l'aptamère fluorogène o-Coral dans son scaffold F30 est représenté en orange. Le ligand du riboswitch employé est représenté par un cercle gris foncé.

Test de rétention du FMN

Dans un premier temps, la bonne rétention du FMN dans les gouttelettes a été vérifiée par un test simple tirant profit de la fluorescence verte naturelle du FMN. Une émulsion mixte composée de gouttelettes sans matrice ADN, de même volume contenant du PBS avec ou sans FMN ont été produites et collectées ensemble avant d'être réinjectées à différents intervalles de temps afin d'analyser leur fluorescence. L'absence de fuite du ligand entre les gouttelettes a pu être constatée après une longue incubation de 16 h où les deux populations de gouttelettes (contenant ou non du FMN) restent distinguables malgré des fuites observables (6,6% présentant une fluorescence verte intermédiaire) (Figure 25).

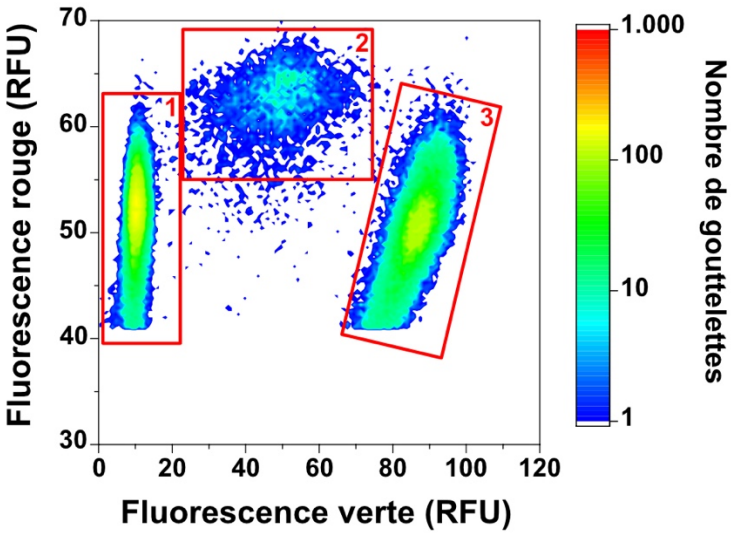


Figure 25 : Test de rétention du FMN dans les gouttelettes. En ordonné la fluorescence rouge indique la taille des gouttelettes grâce à la quantité de Cyanine 5 servant de traceur de gouttelettes, en abscisse la fluorescente verte émise naturellement par le FMN. Le profil représenté est obtenu après 16h d'incubation. La population (1) correspond aux gouttelettes sans FMN, la numéro (3) à celle contenant du FMN (100 µM), et la population (2) à une fraction de gouttelettes avec des niveaux intermédiaire de fluorescence correspondant probablement à du FMN s'étant échappé ou ajouté ou bien à un peu de coalescence entre les gouttelettes contenant du FMN.

Test d'expression du R- FMN en microfluidique

Un premier essai de transcription de la construction R-FMN non mutée a ensuite été réalisé en gouttelettes en absence de ligand (Figure 26). Une incubation de 12h à 37°C est utilisée pour obtenir une quantité suffisante d'ARN (donc de complexe Gemini-561/o-Coral fluorescent) d'après des tests préliminaires réalisés en microplaque (Figure 21 B). Après une incubation équivalente des gouttelettes, une population discrète avec une fluorescence orange plus importante résultant de la transcription d'o-Coral est nettement visible au sein des gouttelettes occupées par une molécule d'ADN, validant ainsi le temps d'incubation. La transcription semble également homogène.

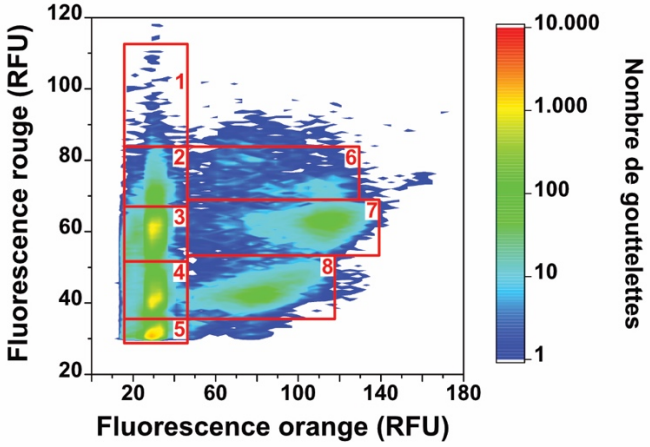


Figure 26 : Transcription *in vitro* par l'holoenzyme d'*E. coli* en gouttelettes. En ordonné la fluorescence rouge indique la taille des gouttelettes grâce à la quantité de Cyanine 5 servant de traceur de gouttelettes, en abscisse la fluorescence orange émise par les complexes o-Coral/Gemini 561. La population (1) correspond à des gouttelettes de PCR, la population (2) correspond aux gouttelettes d'*IVT* fusionnées à trois gouttelettes de PCR vides puis de manière décroissante la population (3) correspond aux gouttelettes d'*IVT* fusionnées à deux gouttelettes de PCR vides, la population (4) à une gouttelette de PCR vide et la population (5) non fusionnées à des gouttelettes de PCR. Enfin, la population (6) correspond aux gouttelettes d'*IVT* fusionnées à trois gouttelettes de PCR dont une ou deux peuvent être vides mais au moins une est occupée (la fluorescence orange varie selon les gouttelettes en question), la population (7) correspond aux gouttelettes d'*IVT* fusionnées à deux gouttelettes de PCR dont une peut être vide mais l'autre est occupée (ou toutes les deux) (la fluorescence orange varie selon les gouttelettes en question), la population (8) correspond aux gouttelettes d'*IVT* fusionnées à une gouttelette de PCR occupée.

L'expérience en gouttelettes a été ensuite répétée une fois en présence du ligand, le FMN. Cependant, le profil microfluidique fut ininterprétable dû à une superposition importante entre le signal orange émis par o-Coral et celui émis par la molécule de FMN collecté dans le canal de mesure orange. En effet, le couple o-Coral/Gemini 561 (λ_{ex} = 560 nm et λ_{em} = 600 nm) et le FMN (λ_{ex} = 450 nm et λ_{em} = 550 nm) sont excités par le laser bleu (λ_{ex} = 488 nm) et émettent tous deux dans les longueurs d'onde de la fluorescence orange (bien que, pour la FMN il ne s'agit que de 20 % de son signal) qui sont donc collecté dans le canal de mesure orange (580 - 620 nm) de la station microfluidique. Sachant que nous sommes en large excès de FMN avec 100 µM final contre seulement 100 nM de ligand fluorogène (Gemini 561) de l'aptamère o-Coral, le vrai signal issu de notre transcription est masqué et l'inhibition censée être observée suite à la modulation de la transcription par le FMN l'est de même par le ligand lui-même. Une proposition afin de poursuivre ce projet fut la modification de l'aptamère fluorogène utilisé. Le FMN émettant dans le vert et en partie dans le canal orange, l'emploi d'un couple aptamère/ligand fluorogène émettant de la fluorescence dans des longueurs d'onde plus éloigné du canal orange est à envisager comme mentionné lors du paragraphe (1.1.3). Une

dernière alternative pour poursuivre nos travaux fut le changement du riboswitch lui-même, permettant de s'affranchir de la limite technique de la fluorescence aspécifique du ligand (FMN) pour le moment.

2.1.4 Essais de riboswitches synthétiques transcriptionnels

Les difficultés techniques rencontrées avec le système FMN m'ont conduit à explorer des solutions alternatives sous la forme de riboswitches synthétiques co-développés par les équipes de Mario Mörl et Peter F. Stadler (Université de Leipzig - Allemagne). Ces molécules issues de prédictions *in silico* sont des riboswitches transcriptionnels de type ON (Figure 27) qui ont tous été testés et validés *in vivo* (Wachsmuth *et al.*, 2013; Etzel and Mörl, 2017; Findeiß *et al.*, 2017; Günzel *et al.*, 2021).

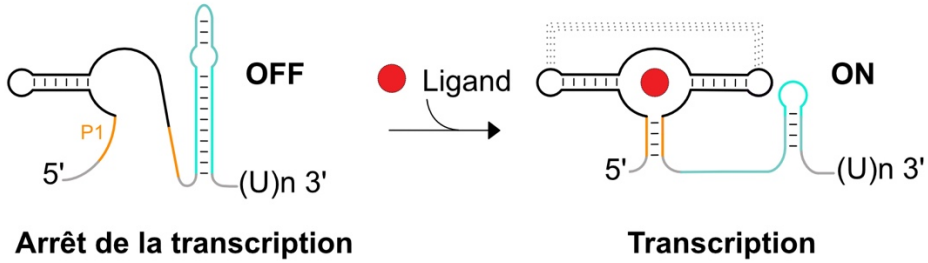
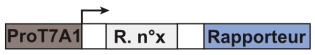


Figure 27 : Représentation schématique du mécanisme de régulation d'un riboswitch ''ON'' répondant à une purine. La première tige du riboswitch appelée P1 est représentée en orange et les séquences impliquées dans la formation de la tige-boucle terminatrice sont représentées en cyan, le ligand quant à lui est représenté en rouge. Figure adaptée de Husser, Dentz and Ryckelynck, 2021.

La stratégie d'analyse et de sélection sera la même que celle déjà mise en place pour le R-FMN, la différence majeure résidant ici dans le fait qu'il s'agit de riboswitches ON. Trois constructions répondant à différents ligands (théophylline (avec le R-Theo), tétracycline (avec le R-Tet) et la néomycine (avec le R-Neo)) nous ont été transmises par notre collaborateur Mario Möerl. Comme précédemment, le promoteur T7A1 d'*E. coli* a été ajouté en 5' et un aptamère fluorogène (Mango-III ou o-Coral) en 3' afin de pouvoir suivre la transcription au cours des différents essais (Figure 28 A). Deux aptamères fluorogènes sont cette fois-ci testés en parallèle pour minimiser les risques d'interférence fluorescentes comme observées précédemment : o-Coral et Mango-III, tous deux développés par l'équipe (Trachman *et al.*, 2019; Bouhedda *et al.*, 2020). Les ligands utilisés lors de ces travaux n'émettant pas de fluorescence verte ou orange il n'y a donc pas eu de problèmes de bruit de fond ou ''cross-talk''. Les mêmes conditions de transcription que pour le R-FMN sont utilisées (sans

spermidine et 6mM de rNTPS). Des travaux préliminaires ont permis de définir la quantité de ligand nécessaire pour une réponse minimum pour chaque construction utilisée : 10 µM de néomycine pour le R-Neo, 1 µM de tétracycline pour le R-Tet et 1mM de théophylline pour le R-Theo. Un récapitulatif des meilleures conditions pour chaque construction est donné ci-dessous (Figure 28 B et C).

A



B

R. n°	Ligand	Rapporteur
1	Néomycine	Mango-III
2		o-Coral + F30
3	tétracycline	Mango-III
4		o-Coral + F30
5	Théophylline	Mango-III
6		o-Coral + F30

C

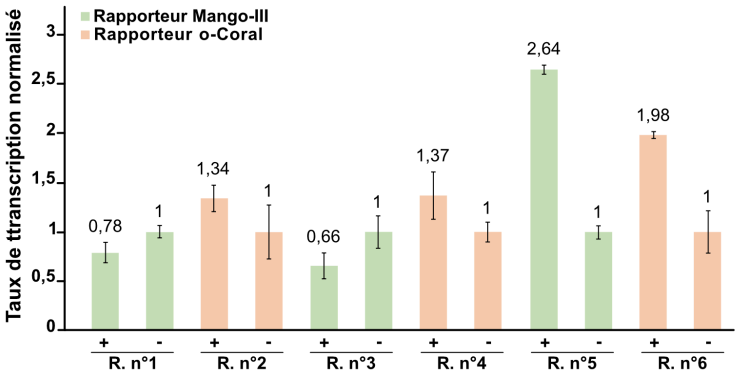


Figure 28 : Comparaison de l'efficacité de réponse selon les constructions employées. A. Représentation générique des constructions utilisées. Le promoteur T7A1 situé en 5' est représenté en marron, le riboswitch utilisé est représenté en gris, et le rapporteur de la transcription correspondant à un aptamère fluorogène est représenté en bleu. **B.** Tableau récapitulatif des riboswitches construits. Chaque riboswitch a été associé soit avec l'aptamère fluorogène Mango-III en vert, soit o-Coral en orange. **C.** Comparaison de l'efficacité de transcription par suivi de l'apparition de fluorescence (Mango-III en vert, o-Coral en orange) des différentes constructions utilisées (voir B), en présence (+) ou en absence (-) de ligand, Les valeurs sont la moyenne de 3 expériences indépendantes et les barres d'erreurs correspondent aux écart-types.

Aucun résultat concluant (aucune différence significative en présence de ligand) n'a pu être obtenu avec les constructions répondant à la tétracycline (R-Tet) et à la Néomycine (R-Neo), quel que soit l'aptamère fluorogène rapporteur utilisé (Figure 28B et C). L'absence de réponse pour le R-Tet pourrait être liée l'introduction d'éthanol dans le mélange transcriptionnel, nécessaire pour la solubilisation du ligand. L'absence de réponse de la construction R-Neo, quant à elle, pourrait être due au ligand lui-même. En effet, la néomycine entraine une inhibition de la transcription *in vitro* à partir d'une concentration de 10 µM (Dube and Palit, 1981). Le R-Neo utilisé dans ces travaux répond à partir de 275 µM de Néomycine en condition *in vivo* (à l'aide de souche d' *E.coli* résistante à la néomycine), n'annonçant pas de bonnes pré-dispositions de fonctionnement. Enfin, le riboswitch répondant à la théophylline (R-Theo) répond significativement à l'ajout de ligand quel que soit l'aptamère rapporteur utilisé (Figure 28C). Puisque l'aptamère Mango-III est significativement plus petit qu'o-Coral inclus dans un scaffold F30 (30 et 230 nucléotides, respectivement) et permet d'observer une meilleure

réponse à la théophylline, la construction R-Théo/Mango-III a été choisie pour la suite du projet.

Une fois l'aptamère rapporteur choisi, je me suis assurée que l'ARN polymérase d'*E. coli* était encore la meilleure solution pour transcrire la construction R-Theo-Mango/III. J'ai ainsi comparé la réponse de la construction à la théophylline suivant qu'elle soit transcrite à partir d'un promoteur de l'holoenzyme de *E.coli* ou de l'ARN polymérase du phage T7 (Figure 29).

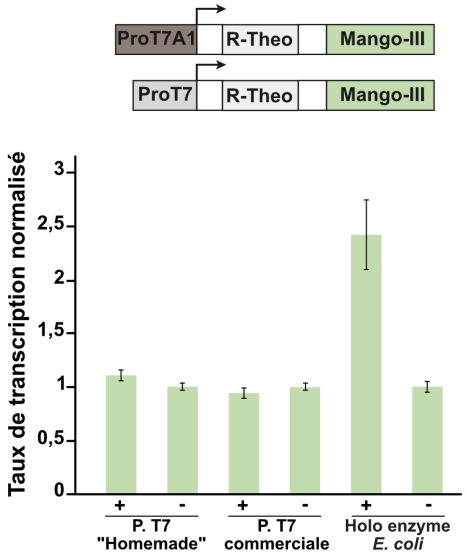


Figure 29 : Comparaison de l'efficacité de réponse du riboswitch répondant à la théophylline selon l'ARN polymérase employée. Les constructions utilisées sont représentées en haut avec en marron le promoteur T7A1 et en gris foncé le promoteur T7, le R-Theo en gris clair et Mango-III en vert. Les transcriptions *in vitro* ont été réalisées avec de l'ARN polymérase du phage T7 purifiée au laboratoire (P.T7 « homemade, haute concentration), commerciale (P. T7 commerciale, basse concentration) ou l'holoenzyme de *E. coli* commerciale, en présence (+) ou en absence (-) de théophylline.

Comme attendu, la seule polymérase permettant de maintenir l'activité du riboswitch est la polymérase de *E. coli* ayant une vitesse de transcription plus lente permettant à l'ARN de se structurer. La transcription par la polymérase T7, quelle que soit la concentration d'enzyme, n'a pas permis d'observer de réponse à la théophylline. Sa vitesse de synthèse très supérieure (100 nt / sec contre 40 nt / sec pour l'ARN polymérase de *E. coli*) ne permet pas au mécanisme régulateur de se mettre en place à temps pour entraîner l'arrêt de la transcription. Cependant, il peut être intéressant de noter que même si la polymérase d'*E. coli* (Gram négatif) est la seule à permettre l'observation d'une réponse pour les riboswitches testés, celle-ci n'est peut-être pas la plus adaptée car les riboswitches transcriptionnels sont retrouvés majoritairement chez

les bactéries à Gram positif. Il n'est donc pas à exclure que l'utilisation d'une polymérase d'un organisme Gram positif permettrait d'obtenir des résultats plus homogènes.

2.1.5 Essais d'optimisation de réponse du riboswitch R-Theo/Mango-III

Bien que répondant à son ligand, la construction R-Theo/Mango-III reste améliorable tant au niveau de sa capacité d'arrêt de la transcription en absence d'une concentration suffisante de ligand que dans sa capacité à permettre la transcription en présence d'une concentration suffisante de ligand. La construction a donc été testée en gouttelettes microfluidiques dans les mêmes conditions de transcription *in vitro* et en présence de théophylline (soit en condition permettant la transcription du rapporteur (ON-switch)) (Figure 30).

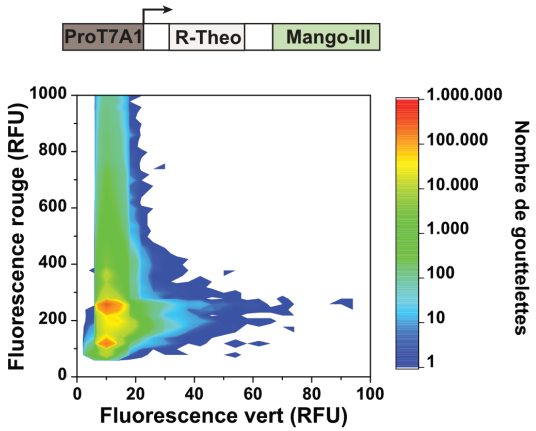


Figure 30 : Expression de la construction modèle R-Theo/Mango-III en gouttelettes microfluidique. A. La construction testée est le riboswitch R-Theo associé à Mango-III (en vert) sous le contrôle du promoteur T7A1 (en marron). **B.** Résultats de criblage après IVT en gouttelettes avec en ordonnée, la fluorescence rouge émise par la cyanine 5 (marqueur de la taille des gouttes) et en abscisse les valeurs de fluorescence verte émise par Mango-III (couplé au TO1-Biotine). Les populations encadrées en rouges correspondent à : (1) (2) (3) et (4) aux gouttelettes d'*IVT* fusionnées avec 0, 1, 2 ou plus de gouttelettes de PCR vides ou contenant une matrice incompatible avec la transcription, respectivement, et (5) (6), aux gouttelettes d'*IVT* fusionnées à deux ou une gouttelette(s) de PCR contenant au moins une matrice permettant la transcription.

Après 12h de transcription *in vitro*, seule une population de gouttelettes à la fluorescence très diffuse a été observée (encadrée 6 en rouge sur la Figure 30). Une telle dispersion n'est à priori pas attendus, puisque la matrice de départ isolée dans les gouttelettes devrait codée pour un phénotype donné. Pourtant, les expériences précédentes avec la construction R-FMN avaient montré que l'ARN polymérase d'*E. coli* était efficace en microgouttelettes (Figure 26). Ainsi, plusieurs hypothèses sont à envisager. La première conjecture écartée fut l'erreur

expérimentale après la réplication de la manipulation et obtention de résultats similaires. Puis, la présence de mutation au sein de la construction et donc la possibilité de différents mutants initialement fut de même écartées après séquençage de la construction de départ. Enfin, les hypothèses les plus probables restantes sont : la haute concentration en théophylline employé (1mM) qui comme déjà observé par l'équipe peut affecter négativement la transcription ou comme préalablement supputé, la polymérase ne serai pas la plus adapté pour ce riboswitch combiné à l'utilisation en microfluidique.

Pour le moment, la question du défaut de fonctionnement des différents riboswitches testés en gouttelettes microfluidiques n'est pas encore résolue. En ce qui concerne le système FMN, une modification du système optique avec l'ajout d'un laser jaune (λ_{ex} = 560 nm) ciblant l'excitation de o-Coral tout en limitant celle de la FMN pourrait permettre de s'affranchir des problèmes liés au « cross-talk » optique et ainsi poursuivre les essais d'amélioration de la molécule. Concernant les riboswitches artificiels, la question est plus délicate et nécessiterait de passer davantage de temps pour mieux comprendre l'origine de la dispersion de signal avec le système théophylline. En effet, le système a été montré comme étant fonctionnel en cellules vivantes, suggérant que la polymérase endogène d'*E. coli* devrait, en principe, être en mesure de récapituler le phénomène comme cela a été vérifié d'ailleurs en microplaques. Cependant, la réaction semble être handicapée lorsqu'elle est transférée au format microfluidique. Bien qu'aucune preuve n'ait pu alors être collectée, un suspect assez sérieux est le surfactant employé pour stabiliser les gouttelettes. En effet, le lot utilisé alors s'est trouvé être incriminé dans l'échec d'un certain nombre d'expériences réalisées par la suite par l'équipe. Il serait intéressant de répéter les essais d'expression/réponse de la matrice R-Theo/Mango-III avec de nouveaux lots de surfactants.

De façon globale, mes travaux sur les systèmes procaryotes ont démontré la possibilité d'utiliser la technologie de microfluidique en gouttelettes pour la sélection de séquences d'ARN modulant l'efficacité d'initiation de la transcription. A ce jour, il s'agit toujours du premier exemple où cette technologie a été employé pour ce type de sélection. Bien qu'il s'agisse d'une expérience modèle, celle-ci ouvre à présent la porte vers le développement et la sélection d'ARN régulateurs artificiels permettant de réguler l'expression de gène cibles. Une première étape dans ce sens a été l'évaluation de la capacité des riboswitches à être étudiés, voire améliorés, par une approche de microfluidique en gouttelettes. Pour les raisons techniques évoquées plus haut, je me suis focalisée sur les riboswitches transcriptionnels. Cela m'a d'ores et déjà permis de collecter des données (choix de la polymérase, identifications des points bloquants potentiels) qui seront extrêmement pertinents et

encourageants pour les futures expériences qui viseront à améliorer ces molécules et à appliquer la technologie à des riboswitches traductionnels.

3. Étude de l'initiation de la traduction chez les eucaryotes

L'initiation de la traduction chez les eucaryotes fait intervenir un plus grand nombre de facteurs que la traduction bactérienne, et reste considérée comme l'étape limitante du processus (Krebs *et al.*, 2018; Rodnina, 2018). Chez les eucaryotes, la traduction peut s'initier de différentes manières : i) la traduction dite coiffe-dépendante qui fait intervenir une pléthore de facteurs interagissant avec la coiffe en 5' de l'ARN messager (ARNm) et la queue poly-A à l'extrémité 3' pour former le CPI (Krebs *et al.*, 2018; Rodnina, 2018) et ii), l'initiation interne médiée par les IRES au cours de la laquelle le ribosome est directement recruté par une région d'ARN structurée et, en général, déposé à proximité du codon d'initiation (Mailliot and Martin, 2018; Kwan and Thompson, 2019).

Après avoir mis en place une plateforme de criblage pour mieux comprendre les mécanismes capables d'initier et de moduler l'initiation de la traduction chez les procaryotes, j'ai transposé cette stratégie à un système eucaryote. J'ai ainsi pu étudier chacun des modes d'initiation de la traduction eucaryote au travers de deux cas précis : i) la ré-exploration de la séquence bordant le codon initiateur, généralement appelée séquence de Kozak dans le cadre de l'initiation coiffe-dépendante, et ii) mettre au point une nouvelle méthode d'identification à haut débit d'éléments d'initiation coiffe-indépendante de la traduction, notamment des IRES, à l'échelle de génomes viraux.

3.1 Ré-exploration de la séquence de Kozak

L'initiation de la traduction coiffe-dépendante implique un grand nombre de facteurs qui interagissent avec la coiffe et la queue poly-A de l'ARNm permettant notamment sa circularisassions. Une fois le CPI formé, incluant la petite sous-unité (SU 40S) du ribosome eucaryote et les nombreux facteurs (eiF1,1A, 2, 3, 5, 4F : eiF4 E, G, A, B), et les PABP) (1.2.2), l'ARNm est alors scanné jusqu'à l'identification du codon initiateur et sa région environnante. Lors du scan de l'ARNm, l'identification du codon initiateur requiert une séquence préférentielle décrite par Marylin Kozak en 1981 (Kozak, 1981), une région comprenant les 6 nucléotides en amont du codon initiateur et les trois nucléotides en aval.

Malgré les progrès techniques et technologiques (profilage du ribosome, suivi de la traduction par rapporteur et NGS) et les différentes études réalisées (Jackson, Hellen and Pestova, 2010; Acevedo *et al.*, 2018; Benitez-Cantos *et al.*, 2020), cette séquence n'est pas définie par un consensus strict. Une autre difficulté réside dans l'évaluation précise de la contribution de chaque facteur d'initiation de la traduction dans la liaison et la reconnaissance de la séquence Kozak. Cependant, la mécanistique eucaryote étant complexe, différents éléments sont à prendre en compte (composition et structure de la séquence, stress cellulaire).

Ainsi, l'objectif de ces travaux a été de ré-explorer la nature de la séquence environnant le codon initiateur afin d'identifier d'éventuels biais de compositions et dans quelle mesure ceux-ci se conforment à la séquence originellement décrite par M. Kozak et si de nouvelles séquences plus efficaces peuvent être identifiées.

3.1.1 Stratégie expérimentale et impact de l'extrait

Stratégie expérimentale

Le but de ce projet était de mettre à profit la stratégie de criblage fonctionnel à ultrahaut-débit validée lors des travaux sur les systèmes procaryotes pour l'étude d'une banque de variant de la séquence environnant le codon initiateur. Pour cela, j'ai préparé une banque dans laquelle la phase codante pour l'eGFP (de l'anglais : « enhanced Green Fluorescent Protein »), a été placée sous le contrôle du promoteur de l'ARN polymérase du phage T7 (Figure 31). La protéine fluorescente eGFP a été choisie ici pour sa brillance plus importante que celle de la GFP (Zhang, Gurtu and Kain, 1996), mais aussi la fréquence d'utilisation des codons optimisée pour une utilisation en systèmes eucaryotes.

Afin de permettre une expression couplée (transcription et traduction), nos collaborateurs ont eu l'idée ingénieuse d'insérer l'IRES de l'EncephaloMyoCardite Virus (EMCV) (Jang *et al.*, 1990) dépourvu de son codon initiateur à l'extrémité 5' de l'ARNm. Cet ajout permet de s'affranchir de l'étape de coiffage de l'ARNm normalement nécessaire à la mise en place du CPI mais complexe à implémenter dans les microgouttelettes. En effet, cette IRES fonctionne comme une plateforme permettant l'ancrage des facteurs nécessaires à l'initiation de la traduction (eiF1, 1A, 2, 3, 4G, 4A, 4B, 5B), ainsi qu'au recrutement du ribosome à l'extrémité 5' de l'ARNm. Une fois le CPI recruté au niveau de l'IRES ‘'non initiatrice'', celui-ci peut alors scanner l'ARN messager jusqu'au codon initiateur. L'environnement de ce codon a quant à lui fait l'objet d'une dégénération complète des 6 nucléotides en amont et des 3 nucléotides en aval (Figure 31).

Enfin, pour être dans les conditions les plus pertinentes possibles, l'ensemble des expérimentations a été réalisé dans des extraits issus de cellules embryonnaires de rein

humain appelées HEK (de l'anglais : « Human Embryonic Kidney ») supplémentés avec de l'ARN polymérase du phage T7.

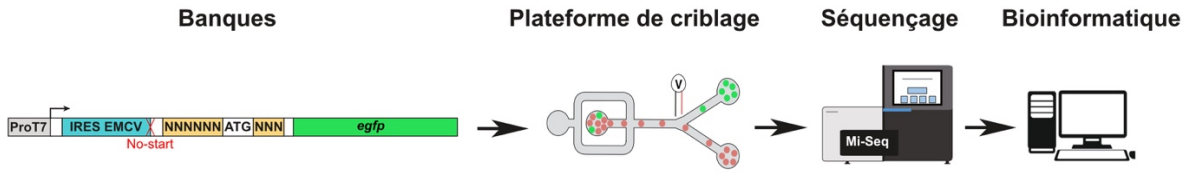


Figure 31 : Stratégie expérimentale employée pour la ré-exploration de la séquence de Kozak. Le promoteur de l'ARN polymérase du phage T7 est représenté en gris, la séquence de l'IRES de EMCV en bleu avec l'absence de codon initiateur en rouge, les séquences aléatoires entourant le codon initiateur en orange et la protéine rapportrice (l'eGFP) en vert.

Le processus de criblage microfluidique s'est déroulé en trois étapes successives (Figure 31) : (i) la banque de gènes a été diluée dans un milieu de PCR et les molécules d'ADN individualisées dans une première série de gouttelettes. (ii) Après thermocyclage, chaque gouttelette de cette première émulsion a été fusionnées avec une gouttelette contenant de l'extrait de cellule HEK contenant la polymérase T7 pour permettre la transcription et la traduction. Puis (iii), la fluorescence de chaque gouttelette a été mesurée et utilisée pour trier les gouttelettes selon leur contenu en eGFP, et donc la capacité du variant analysé à initier efficacement la traduction. Enfin, (iv) le contenu des banques enrichies a été séquencé par NGS avant d'être analysé par bio-informatique.

Impact de l'extrait

Lors de premiers essais de criblage de notre banque au sein de la plateforme microfluidique, une forte variation de la fluorescence verte a été observée pour l'ensemble des gouttelettes (Figure 32 A et B). A titre de comparaison, la Figure 32 A est un profil « modèle » obtenu à l'aide d'extraits de RRL qui présente des populations nettes et peu dispersées, que ce soit au niveau de la fluorescence rouge (marqueur de la taille des gouttelettes, avec les gouttelettes de PCR contenant dix fois plus de Cyanine 5 que celles de TTIV (RRL)) que pour la fluorescence verte. Ce profil correspond à l'expression de deux plasmides modèles, contenant le promoteur T7 suivi de la séquence d'une IRES (selon le plasmide correspond à la 5'UTR IRES ou l'IGR du CrPV) puis la séquence de la GFPmut2. Le point marquant à comparer est la population 4 de ce profil qui est très homogène à l'inverse de son équivalent encadré sur la Figure 32 B. Cette deuxième figure correspond au premier tour de sélection réalisé à partir de la banque conçue pour randomiser la séquence Kozak à l'aide d'extrait de HEK (Figure 31). Le profil obtenu est globalement plus diffus pour le signal fluorescent rouge, illustrant un manque de stabilité des gouttelettes qui ont probablement fusionnées entre-elles de manière passive. Par ailleurs, l'encadré rouge numéroté 1 du panel B, met en avant la

dispersion de la population par rapport à celle retrouvée dans l'encadré 4 du panel A (modèle). Lors du panel B, la matrice correspond à une banque qui donne lieu à différents degrés d'expression et donc de production d'eGFP formant une trainée de gouttelettes avec divers profils fluorescents verts. Cependant, si la population fusionnée aux gouttelettes de PCR vides ou contenant une matrice ne permettant pas la traduction (encadré rouge 1 panel B) est trop disperse, elle risque de masquer en partie la trainée attendue et en conséquence limiter la sélection de ces variants Figure 32 A et B.

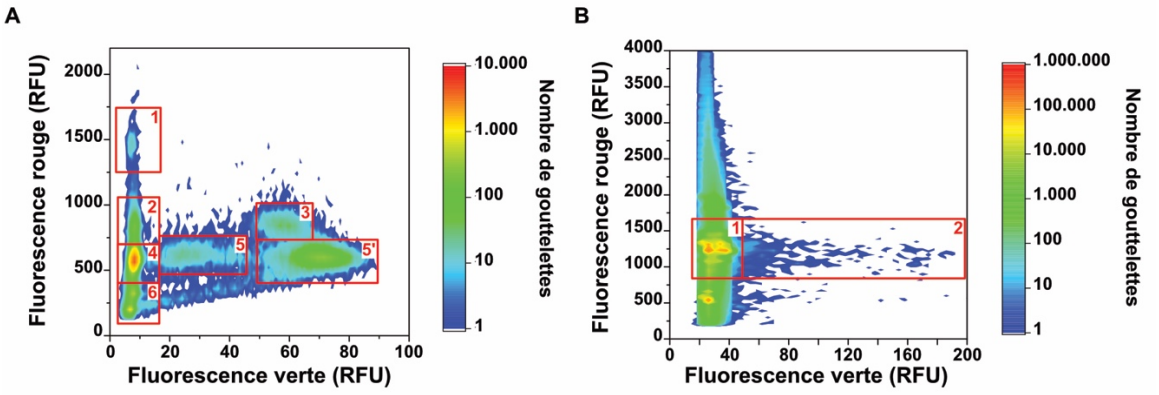


Figure 32 : Comparaison de lisibilité du profil microfluidique. A. Profil microfluidique « modèle ». Ce profil microfluidique est obtenu à partir de deux plasmides modèles comportant le promoteur T7 pour mener la transcription, et soit la 5'UTR IRES soit l'IGR du CrPV et la protéine fluorescente GFPmut2 comme rapportrice de la traduction. L'axe des ordonnées correspond à la fluorescence rouge issue de la Cyanine 5 (traceur de gouttelettes), tandis que l'axe des abscisses correspond à la fluorescence verte issue de la GFPmut2 (rapporteur de la traduction ou du bruit de fond des extraits). Chaque boite rouge est numérotée afin de désigner une population de gouttelettes avec en (1) les gouttelettes de PCR, en (2) les gouttelettes de TTIV fusionnées à deux gouttelettes de PCR vides, en (3) les gouttelettes de TTIV fusionnées à deux gouttelettes de PCR dont au moins une contient de l'ADN, en (4), les gouttelettes de TTIV fusionnées à une gouttelette de PCR vide, en (5) les gouttelettes de TTIV fusionnées à une gouttelette de PCR contenant la 5'UTR IRES, en (5') les gouttelettes de TTIV fusionnées à une gouttelette de PCR contenant l'IGR, et en (6) les gouttelettes de TTIV non fusionnées. B. Profil microfluidique obtenu lors du premier tour de sélection avec la banque aléatoire de Kozak (promoteur T7, IRES EMCV, séquence aléatoire/codon initiateur et eGFP) et de l'extrait de HEK. La population (1) correspond aux gouttelettes de TTIV fusionnées à une gouttelette de PCR vide ou contenant une matrice ne permettant pas l'expression de l'eGFP, la population (2) correspond aux gouttelettes de TTIV fusionnées à une gouttelette de PCR contenant une matrice permettant de manière plus ou moins efficace l'expression de l'eGFP (d'où la dispersité observée).

Une solution proposée pour limiter les variations d'expression des constructions au sein des gouttelettes a été de limiter le temps d'incubation dans la glace avant production des gouttelettes, ainsi que celui dans les tubings avant collection de l'émulsion afin de limiter les variations d'expression, sans succès. De plus, nous avons observés la formation d'agrégats avec les extraits HEK, qui en plus de déstabiliser les gouttelettes (Figure 33 A), ont bloqué les canaux microfluidiques, entraînant un arrêt de la manipulation (Figure 33 B et C).

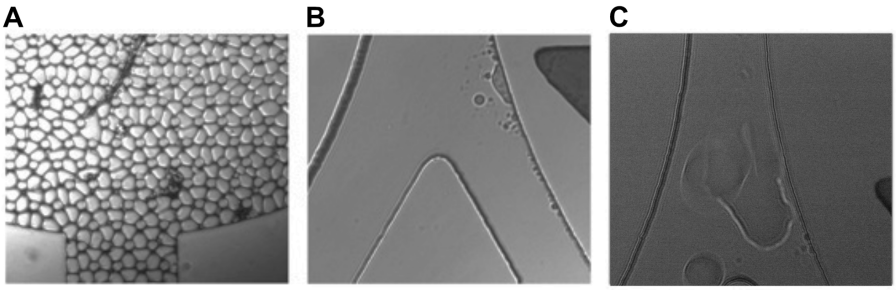


Figure 33 : Photographie de puces microfluidiques montrant les effets de la formation d'agrégats des extraits HEK. **A.** Observation d'agrégats au sein des gouttelettes après expression. **B et C.** Observation d'agrégats bloquant la fourche de sélection.

3.1.2 Criblage fonctionnel à ultra haut-débit en absence de couplage

Malgré l'ensemble des difficultés rencontrées dues à l'utilisation des extraits de HEK, trois tours de sélection consécutifs ont été réalisés en vue d'enrichir la banque en séquences permettant une synthèse efficace d'eGFP (Figure 34 A et B). Les profils de sélection obtenus restent très disperses pour la fluorescence rouge (marqueur de la taille des gouttelettes) laissant penser à un phénomène de coalescence de nos gouttelettes. Après le premier tour de sélection, les séquences obtenues ont été testées par TTIV en microplaque avec le suivi de l'émission la fluorescence verte correspondant à la traduction de l'eGFP sous le contrôle des séquences Kozak randomisées (Figure 34 C). Le premier tour de sélection a permis d'enrichir la banque de départ en séquences plus efficaces pour l'initiation de la traduction. Cependant, dès le deuxième tour de sélection le taux de traduction a commencé à chuter pour ne donner plus qu'une très faible population positive au troisième tour (Figure 34 B). Étant donné la faible fluorescence des gouttelettes dans ce dernier tour, deux populations distinctes ont été criblées : un premier tri où toutes les gouttelettes présentant un signal de fluorescence ont été sélectionnées (fenêtre de tri T) et un second tri plus strict où seules les gouttelettes les plus fluorescentes ont été collectées (fenêtre de tri M). Néanmoins, quel que soit le tri réalisé, un véritable effondrement du taux de traduction a été observé en microplaques (Figure 34 C).

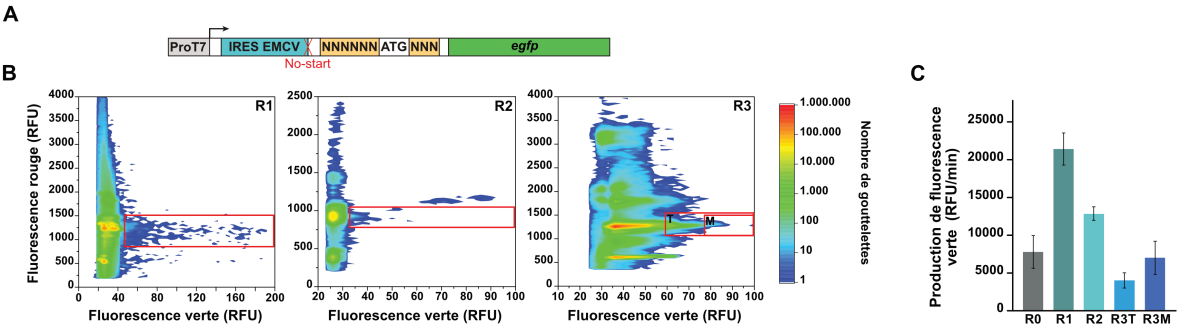


Figure 34 : Sélections et tests d'enrichissement de la banque. A. Représentation schématique de la banque utilisée, le promoteur T7 est représenté en gris, l'IRES de l'EMCV en bleu ainsi que l'absence de codon initiateur en rouge, les séquences aléatoires en orange et la séquence rapportrice en vert. **B.** Les trois tours de sélection consécutifs réalisés en microfluidique en gouttelette. L'axe des ordonnées correspond à la fluorescence rouge issue de la Cyanine 5 notre traceur de gouttelette, tandis que l'axe des abscisses correspond à la fluorescence verte issue de l'eGFP notre rapporteur de la traduction. Les rectangles rouges représentent les gouttelettes ayant un profil fluorescent d'intérêt qui sont sélectionnées. Lors du dernier tour de sélection (R3) deux fenêtres de sélections ont été réalisées annotées T (gouttelettes avec de la fluorescence verte) et M (fluorescence verte plus importante soit supérieur à 75 RFU) **C.** Test d'enrichissement par suivi fluorimétrique en temps réel. La traduction correspondant à la fluorescence verte produite au cours du temps est représenté en ordonné et en abscisse sont représentés les différents tours de sélection réalisés.

Le paradoxe de ces résultats suggère une contre-sélection. Cependant la vérification du logiciel contrôlant les stations microfluidiques ne corrobore pas cette hypothèse. D'autre part, ces criblages ont été handicapé par l'apparition des bandes d'ADN aspécifiques apparues lors des ré-amplifications des molécules d'ADN obtenues à l'issue des criblages (Figure 35).

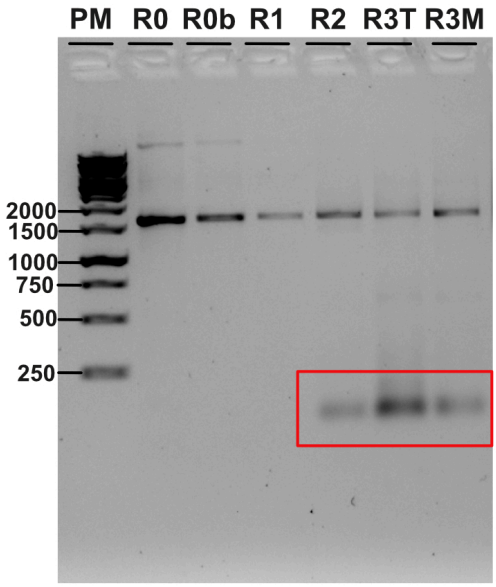


Figure 35 : Vérification sur gel d'agarose des ré-amplifications des séquences issues de chaque tour de sélection. 6 µL (R0) et 3 µL (R0b) de la banque initiale sont déposés sur gel d'agarose 1,5%. 3 µL de PCR de ré-amplification des séquences de chaque tour de sélection sont déposé de R1 (de l'anglais : « Round ») à R3. Pour le dernier tour de sélection (R3) le puit R3T (annoté T pour Total) contient la PCR des séquences du troisième tour de sélection avec la stringence la moins importante et le puit R3M (annoté M pour Maximum) correspond aux séquences ré-amplifiées par PCR issues du troisième tour de sélection avec la stringence la plus importante. Le fragment d'intérêt a une taille de 1540 paires de base. L'encadré rouge met en avant les molécules aspécifiques ré-amplifiées en parallèle de plus en plus présentes aux alentours de 150 paires de base.

Ces bandes aspécifiques s'accumulent au fil des amplifications et détournent la machinerie d'amplification à la façon de parasites, un phénomène bien connu survenu à plusieurs reprises lors de travaux précédents de l'équipe (sélection *in vitro* et d'évolution dirigée d'ARN par exemple).
En dépit des difficultés rencontrées et pour avoir une meilleure idée du contenu des banques enrichies, celles-ci ont été envoyées au séquençage à haut débit et je suis actuellement en attente des résultats de l'analyse bio-informatique en cours de réalisation par nos collaborateurs.

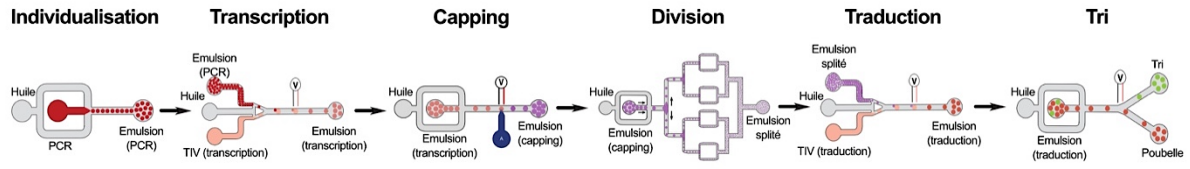
D'un point de vue plus critique, l'utilisation de l'IRES de l'EMCV pour s'affranchir du coiffage n'est pas optimale, puisqu'il est difficile de distinguer si l'initiation a lieu suite au scanning de l'ARNm ou si l'IRES elle-même a permis l'initiation de la traduction à un codon d'initiation non-optimale. Pour valider le mécanisme d'initiation, il est possible d'inhiber l'initiation coiffe dépendante en utilisant des facteurs tels que eIF4G1 et 4E-BP-1 interagissant avec eIF4E impliqué dans la reconnaissance de la coiffe (Vaklavas *et al.*, 2015), ou de bloquer l'initiation coiffe-indépendante grâce à de petites molécules interférant avec le fonctionnement

d'une IRES. Le blocage d'IRES par de telles molécules est une stratégie prometteuse pour le traitement anticancéreux tel que pour le cancer du sein où ce type d'IRES participe au développement du cancer (Vaklavas *et al.*, 2015).

3.1.3 Optimisation du processus microfluidique et perspectives

Bien que les résultats des analyses bio-informatiques soit en cours de traitement, un certain nombre de points peuvent d'ores et déjà être discutés en termes de futures applications de l'approche microfluidique dans l'étude des mécanismes de la traduction.

Tout d'abord, le processus microfluidique tel que proposé ici est adapté aux études de l'initiation de la traduction d'ARNm dépourvus de coiffe à leur extrémité 5' (Kwan and Thompson, 2019). La solution proposée par nos collaborateurs (F. Martin et A. Tidu) dans laquelle une construction comportant l'IRES EMCV en amont de la région à traduire permet de répondre partiellement à cette limitation et peut être utilisée directement avec des systèmes d'expression par transcription et traduction couplées. Cette approche reste cependant peu adaptée à l'étude de mécanismes dans lesquels la présence de la coiffe est directement impliquée. Un tel ajout serait envisageable en gouttelettes mais nécessiterait une modification majeure du système actuel. Ainsi, il faudrait tout d'abord découpler les étapes de transcription et traduction.



<u>Figure 36 :</u> Processus microfluidique en perspective pour la ré-exploration de la séquence de Kozak.

Dans ce schéma (Figure 36), après amplification PCR, les gouttelettes contenant l'ADN seraient d'abord fusionnées une à une avec d'autres contenant le milieu de transcription. Après synthèse des ARNs, une étape de pico-injection (4.1.2) serait réalisée pour délivrer à chaque gouttelette un mélange réactionnel commercial (protéine du Vaccinia virus appelé Vaccinia Capping Enzyme (VCE)) permettant l'ajout de la coiffe à l'extrémité 5' de l'ARN transcrit. Les gouttelettes contenant les ARN modifiés seraient ensuite divisées en gouttelettes plus petites par utilisation d'un module de vision des gouttelettes, ''Droplet splitter'' (Link *et al.*,

2004) pour obtenir des gouttelettes dont le volume est compatible avec les modules de fusion utilisés au laboratoire. Ces dernières seraient ensuite soumises à une seconde étape de fusion pour l'ajout de mélange réactionnel de traduction et permettre l'expression des protéines fluorescentes. Enfin, comme dans le processus original, les gouttelettes seraient triées selon leur fluorescence et leur contenu en séquences analysé par NGS et bio-informatique.

En ce qui concerne le possible manque d'homogénéité et de reproductibilité des extraits, plusieurs explications semblent plausibles. Ces extraits ne sont pas déplétés des ARNm endogènes. De ce fait, bien que le milieu d'expression soit stocké sur glace la majeure partie du temps avant la fusion des gouttelettes de TTIV avec celles de PCR, il ne peut être exclus que ces ARNm captent les ribosomes et entre en compétition avec nos ARN d'intérêts. Dans cette configuration, le délai entre la préparation des extraits, leurs émulsions et leurs fusions aux gouttelettes contenant les ADN de la banque à tester peut conduire à l'expression des ARNm endogènes. L'expression non contrôlée de ces ARNm peut potentiellement conduire à la dispersion de signaux observés. Un autre facteur pouvant influencer la traduction en gouttes est lié au « vieillissement » de l'extrait. En effet, bien que gardé sur glace, l'extrait peut commencer à s'inactiver au fil du temps pouvant aboutir à une traduction différenciée entre le début et la fin d'une fusion d'IVTT. Pour vérifier la stabilité des extraits dans la glace, des mesures d'efficacité de traduction pourront être effectuées en faisant varier les temps d'incubation sur la glace. Les résultats préliminaires collectés jusqu'à maintenant se sont révélés assez variables et ne permettent pas de conclure de façon claire.

Enfin, une fois l'ensemble de ces optimisations réalisées et la preuve de concept établie pour ce processus de criblage avec le codon initiateur le plus fréquemment utilisé (« AUG »), il sera possible de poursuivre les sélections en utilisant cette fois d'autres codons initiateurs comme le codon CUC et GUG, afin d'étudier l'impact de ce codon sur l'environnement de séquence optimal au démarrage de la traduction (Kearse and Wilusz, 2017; Hernández, Osnaya and Pérez-Martínez, 2019). L'effet de la présence de structures secondaires pourra également être évalué dans chaque cas (Leppek et al., 2018).

Dans une perspective plus lointaine, il sera également envisageable d'utiliser des extraits plus complexes obtenus, par exemple, à partir de cellules soumises à certains stress environnementaux. En effet la régulation de l'expression des gènes en réponse au stress peut être le résultat de facteurs alternatifs de la traduction induits dans ce type de conditions exclusivement (Madlung and Comai, 2004).

Un stress environnemental peut également être associé à une bascule dans l'équilibre entre les initiations coiffe-dépendante et coiffe-indépendante, permettant l'adaptation rapide de la cellule à son environnement (Komar and Hatzoglou, 2011; Sriram, Bohlen and Teleman,

2018b). C'est pourquoi nous avons aussi porté notre intérêt sur l'adaptation de notre plateforme de criblage pour un second projet dédié cette fois-ci à l'initiation coiffe-indépendante.

4. Initiation *via* les IRES

À l'inverse de la traduction canonique, qui requiert capping et poly-adénylation des ARN messagers, la traduction non-canonique est basée notamment sur les sites d'entrée interne du ribosome appelés "IRES". Ces structures ARN vont permettre le placement du ribosome au niveau du codon initiateur ou à proximité de ce dernier selon le type d'IRES en jeu (Mailliot and Martin, 2018; Kwan and Thompson, 2019). A l'heure actuelle quatre types d'IRES ont été décrits. Les IRES de type I et II nécessitent la majorité des facteurs de la traduction mise à part eIF4E interagissant avec la coiffe. La principale différence résidant dans le placement du ribosome qui se fait sur le codon initiateur pour les IRES II à l'inverse des IRES de type I impliquant le scanning de l'ARNm. Les IRES de type III ne nécessitent que les facteurs eIF2 et 3 et placent le ribosome immédiatement sur le codon initiateur tandis que les IRES de type IV n'ont besoin d'aucun facteur et placent le ribosome sur le codon initiateur. Malgré les différentes avancées technologiques (rapporteurs et suivi de l'expression, FACS, NGS…) et l'importance de ce type d'éléments, leur identification n'est pas toujours aisée et devient rapidement chronophage (Bonnal *et al.*, 2003; Baird *et al.*, 2006; Mokrejs *et al.*, 2006; Yang *et al.*, 2021). À l'heure actuelle, leur identification se base essentiellement sur des similarités de séquences ou sur les prédictions de structures théoriques de l'ARNm qui sont par la suite testées *in vitro* et/ou *in cellulo* (Baird *et al.*, 2006; Yang *et al.*, 2021). L'objectif de cette dernière partie de mon travail de thèse a donc été la mise au point d'une nouvelle technologie, permettant l'identification rapide d'éléments IRES au niveau de génomes par criblage fonctionnel à ultrahaut-débit couplé à la bio-informatique, en adaptant l'approche déjà validée dans les parties précédentes. Pour ce faire, il m'a tout d'abord fallu mettre au point une méthode de préparation de banques la moins biaisée possible et compatible avec les besoins de la technologie de microfluidique en gouttelettes.

4.1 Mise au point de la méthode de préparation de banques génomiques

L'objectif étant de trouver des séquences potentiellement capables d'initier la traduction au sein d'un génome, il a tout d'abord fallu déterminer un mode de fragmentation de celui-ci et une taille de fragments à générer pour maximiser les chances d'isoler une IRES fonctionnelle, tout en restant capable d'analyser la diversité de la banque ainsi produite.

Le plasmide utilisé que je détaillerai dans la partie concernant la preuve de concept plus bas (4.2) contient le génome du CrPV (12 496 pb ensemble) ainsi que deux IRES (Wilson *et al.*, 2000; Gross *et al.*, 2017a) : l'IRES située dans la 5'UTR (352 pb) (IRES 5'UTR) est de type III et nécessite les facteurs eIF2 et eIF3, tandis que l'IRES intergénique dite IGR (204 pb), de type IV, est fonctionnelle de manière autonome. J'ai donc décidé de réaliser une fragmentation aléatoire à l'aide d'un kit de fragmentase commercial qui utilise deux enzymes pour couper le double brin d'ADN de ce plasmide afin d'obtenir la plus large diversité possible de séquences. J'ai réalisé une fragmentation permettant l'obtention de fragments en moyenne autour des 700 ± 250 pb, contenant ainsi les IRES déjà caractérisées au sein de ce génome. Théoriquement, le protocole de fragmentation employé génèrera un fragment d'en moyenne 700 pb à chacune des 12 496 positions du génome. L'IRES IGR peut alors se situer à 500 positions différentes dans un fragments de 700 pb et l'IRES 5'UTR à 350 positions différentes. Toutefois en tenant compte du nombre total de fragments de la banque de départ, des différentes tailles de fragments obtenus, de la taille respective de chaque IRES (insertions et délétions tolérables) et la nécessité d'un cadre de lecture en phase (avec le gène dont elle contrôle la traduction), seul un nombre très faible de variants ont le potentiel d'initier la traduction.

Ensuite, pour permettre l'expression de ces fragments et évaluer leur capacité à initier la traduction il a fallu mettre au point une approche pour flanquer chaque fragment d'un promoteur de la transcription en 5' et de la séquence codant pour la GFPmut2 (rapporteur traductionnel) en 3'. La construction de cette banque a été le point le plus délicat de ce développement technologique. L'ensemble des mises au point a été réalisé à l'aide d'une séquence modèle comportant des adénines sortantes à l'extrémité 3' simulant un ADN obtenu après fragmentation et réparation selon les approches standards utilisées pour la préparation de banques pour le séquençage à haut débit. L'ensemble des stratégies évaluées sont détaillées plus bas (Figure 38 et Figure 39). L'efficacité et le rendement des différentes méthodes ont tout d'abord été évalués par qPCR (Figure 37). Pour cela, les signaux obtenus avec l'amplicon "interne" (jeu d'amorces 1) ont été comparés à ceux obtenus avec l'amplicon "externe" correspondant à la construction souhaitée. Le couple d'amorce 1 s'hybride aux extrémités de l'IRES modèle (simulant les fragments futurs utilisés) tandis que le couple 2 s'hybride au niveau du promoteur T7 et au début de la séquence codante pour la GFPmut2 (permettant d'amplifier la construction complète souhaitée). La comparaison des signaux donne un ordre de valeur du rendement entre les fragments flanqués du promoteur et de la GFPmut2 (construction souhaitée) et la matrice de départ non flanquée (l'IRES modèle). Bien que donnant des résultats difficilement comparables du fait de la différence significative de longueur des amplicons, ces mesures restent un accompagnement pour évaluer si une

modification de procédure tend à améliorer ou au contraire à décroitre l'efficacité de la méthode. Cette stratégie a cependant rapidement atteint ses limites, notamment liées à des problèmes d'amplification non spécifique. Ainsi, pour les étapes plus tardives du développement, je me suis plutôt appuyée sur l'utilisation d'oligonucléotides fluoromarqués qui ont permis le suivi de ma méthode de manière direct par révélation sur gel d'agarose sans étape intermédiaire de PCR.

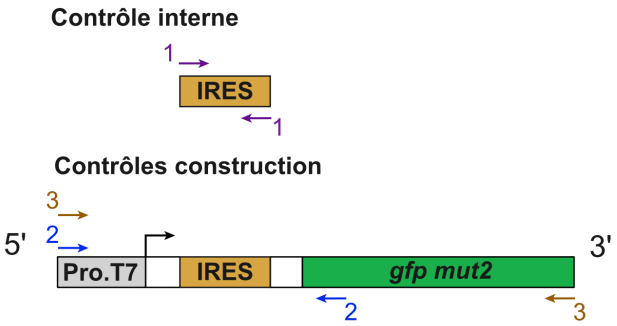


Figure 37 : Amplicons utilisés pour l'évaluation de l'efficacité de la méthode de préparation de banques. Le fragment contenant l'IRES modèle (IGR) est représenté en orange. La paire d'oligonucléotides numérotée 1 permet de détecter le fragment d'ADN étudié. La séquence du promoteur T7 est représentée en gris et la séquence codante pour la GFPmut2 en vert. La paire d'oligonucléotides numérotée 2 s'hybride au niveau du promoteur T7 et au début de la *gfpmut2*. La paire d'oligonucléotides numérotée 3 s'hybride quant à elle au niveau du promoteur T7 et à la fin de la *gfpmut2* permettant la réamplification de l'ensemble de la construction (T7/IRES/GFPmut2).

Pour réaliser cette banque modèle, la première idée testée fut l'ajout par ligation du promoteur T7 en 5' (qui permet la transcription de la construction en aval) ainsi que du gène de la *gfpmut2* aux fragments obtenus (Figure 38).

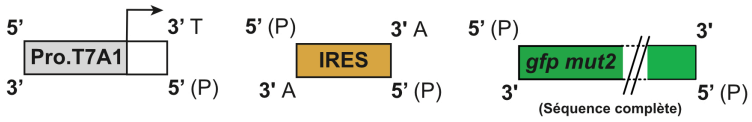


Figure 38 : Premier essais de préparation de la banque modèle par ligation.

Cependant, lors de l'évaluation du rendement de la ligation par qPCR comme mentionné ci-dessus, avec une paire d'oligo se fixant sur notre IRES modèle (couple 1) et une paire d'oligo se fixant sur le promoteur et le début de la GFPmut2 (couple 2) (Figure 37) le cycle seuil (cycle PCR à partir duquel la quantité d'amplicons produite est détectable par fluorescence) de la construction complète (du promoteur à la GFPmut2) présente un décalage de 10 cycles de PCR supplémentaires comparé au fragment témoin d'IRES. Ce qui par déduction illustre la

faible quantité de produit ligué dans le milieu réactionnel. Néanmoins en testant par qPCR avec l'oligo sens du couple 1 (s'hybridant au niveau du promoteur) et l'anti-sens du couple 2 (s'hybridant au niveau de l'IRES) (excluant le contrôle de la ligation de la GFPmut2), on obtient un cycle seuil similaire à celui de l'IRES et son couple d'oligos (couple 1). Ce qui permet de conclure que la ligation entre le fragment du promoteur T7 et du fragment contenant l'IRES a bien eu lieu et de manière efficace. J'ai donc supposé que liguer des fragments de grandes tailles n'était pas optimal en termes de rendement, d'où l'essai de liguer des fragments de plus petites tailles (Figure 39).

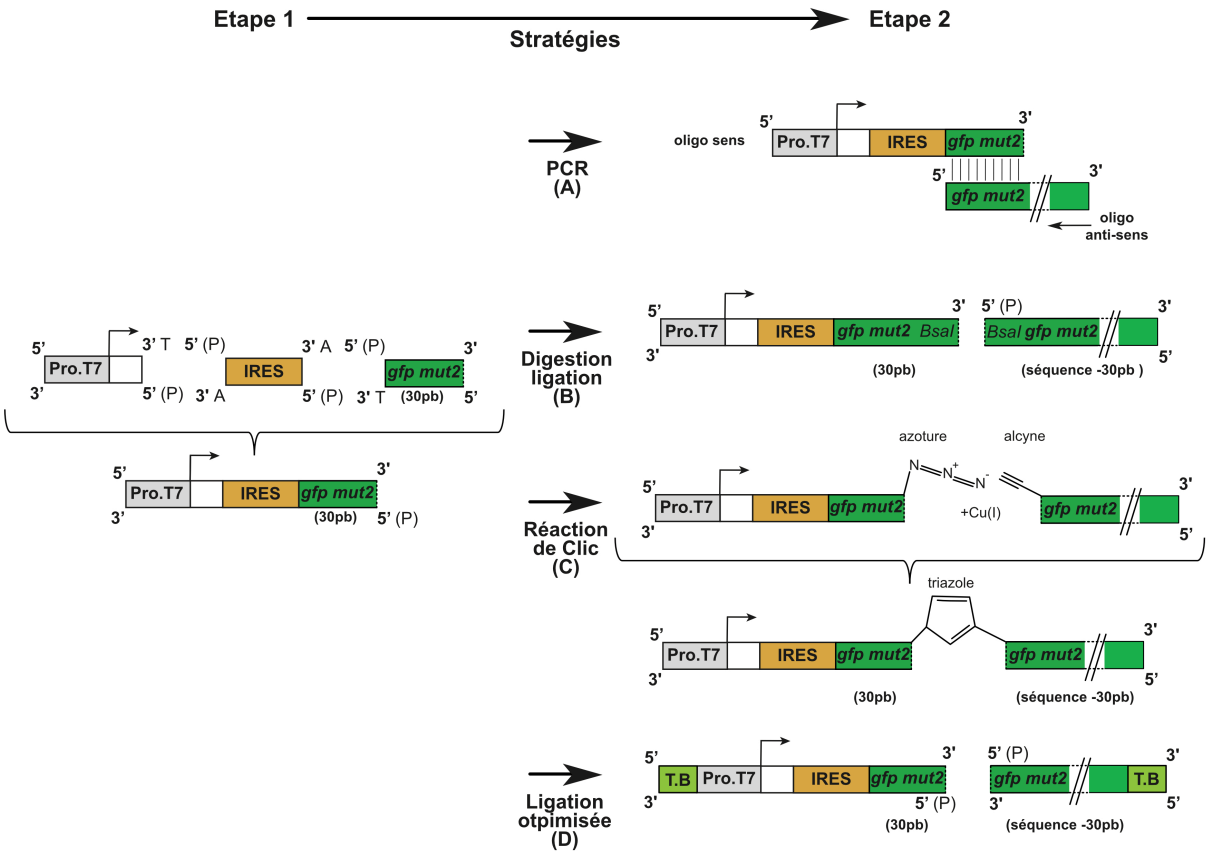


Figure 39 : Récapitulatif de l'ensemble des stratégies testées pour l'élaboration de la banque modèle. L'étape 1 représentée à gauche illustre la première étape de ligation entre l'IRES modèle en marron, le promoteur T7 en gris et le début de la séquence codante (30pb) pour la GFPmut2 en vert. Les bases sortantes (Adénine ou Thymine) en 5' ou 3' sont précisées sans parenthèse. Les stratégies employées pour l'étape 2 sont mentionnées sous les flèches noires et annotées de A à D avec respectivement : la PCR (A), la digestion/ligation (B) dont le site de restriction BsaI est précisé dans la séquence de la GFPmut2, la réaction de Clic (C), et la ligation optimisée (D) dont les extrémités comportent un PolyEthyleneGlycol (PEG) suivi d'une tige boucle, l'ensemble annoté T.B en 5' de la première ligation et en 5' du brin reverse de la GFPmut2 limitant la circularisassions des fragments sur eux-même.

Ainsi, j'ai réalisé la ligation du promoteur T7 en 5' du fragment contenant l'IRES modèle utilisée et uniquement le début de la phase codante de la GFPmut2 en 3' (Figure 39 étape 1). Par la suite la vérification du rendement est réalisée par qPCR avec les couples d'amorces 1 (IRES) et 2 (construction souhaitée) donnant des cycles seuils similaires avec ensuite une validation de la taille des fragments d'intérêts par électrophorèse sur gel d'agarose (Figure 40).

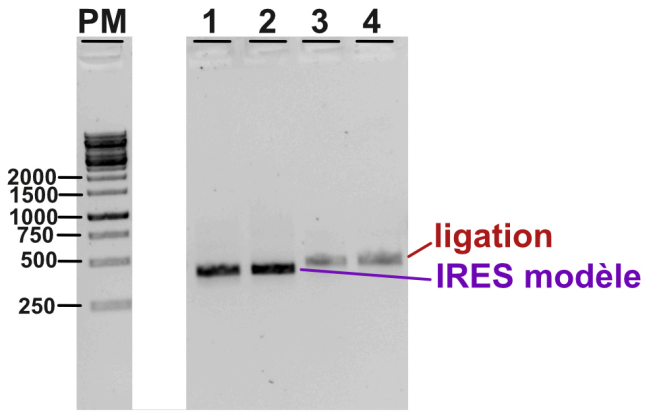


Figure 40 : Vérification sur gel d'agarose 1,5% des ré-amplifications des séquences issues de chaque tour de sélection. Le marqueur de poid moléculaire (PM) dans le premier puit est annoté en paire de base (pb). Dans les puits suivants sont déposés : 3 µL de PCR réalisés avec le couple d'amorce 1 et comme matrice la ligation (étape 1 avec un ratio de 1 pour 1 de chaque molécule (1) et un ratio de 1 IRES pour 50 de promoteur T7 et *gfp*mut2 (2)). De même, 3 µL de PCR réalisés avec le couple d'amorce 2 et comme matrice la ligation (étape 1 avec un ratio de 1 pour 1 de chaque molécule (3) et un ratio de 1 molécule d'IRES pour 50 molécules de promoteur T7 et *gfp*mut2 (4)) sont déposés. La taille attendue de l'IRES modèle seule étant de 494 pb et de la ligation de l'IRES modèle avec le promoteur T7 et le début de la séquence codantes de la *gfp*mut2 de 590 pb. Pour des raisons d'esthétisme le premier puit sans intérêt avec les données traitées a été retiré (espace entre le PM et le puit 1).

L'objectif fut ensuite de parvenir à ajouter la fin de la phase codante de la GFPmut2 en maintenant de hauts rendements de production.

Une première idée testée fut la réalisation d'une PCR pour augmenter la quantité de matériel obtenue une fois les fragments associés avec la GFPmut2 comme matrice et le produit de la première ligation comme oligo sens en tandem avec un oligo anti-sens s'hybridant à la fin de la GFPmut2 (anti-sens du couple 3 dans la Figure 37) (Figure 39 (A)). Aucune réamplification n'a eu lieu laissant penser que la première ligation, étant un double brin, a du mal à se dissocier lors de la réaction de PCR et ne peut donc pas s'hybrider sur la matrice de la GFPmut2. J'ai alors essayé d'obtenir un fragment simple brin à partir du produit de la ligation

1 (purification biotine, digestion exonucléase), mais ces procédés ont non seulement complexifié l'ensemble du procédé mais n'ont en plus de cela jamais donné de résultats.

J'ai donc décidé de changer de stratégie en introduisant un site de restriction (BsaI) en 3' de la première ligation ainsi qu'en 5' de la GFPmut2 que l'on souhaite liguer à notre première ligation (Figure 39 (B)). Ce type d'enzyme de type II clive en dehors de son site de reconnaissance permettant une fois la digestion et la ligation de ces deux fragments réalisés d'obtenir une construction finale sans l'ajout de nucléotides additionnels. Ce processus de digestion/ligation appelé aussi Golden gate nous a permis d'obtenir notre fragment complet d'intérêt avec un cycle seuil seulement de 5 cycles additionnels avec le couple d'oligo du promoteur T7 et de la fin de la GFPmut2 versus celui de la première ligation (promoteur T7 et début de la GFPmut2). Néanmoins, lorsque cette stratégie été appliquée avec le génome CrPV fragmenté à la place de l'IRES modèle, j'ai observé un biais énorme ne permettant pas d'avoir une multitude de constructions (smear sur gel) mais une bande (Figure 41). Après vérification qu'il ne s'agissait pas d'une simple contamination et que le biais était bien répétable, la principale conjecture s'est concentrée sur la favorisation d'un fragment suite à la digestion par BsaI de manière involontaire.

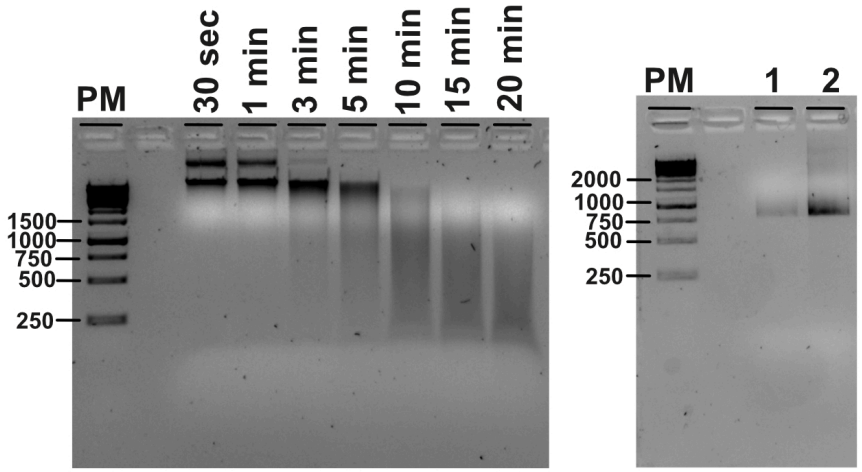


Figure 41 : Vérification sur gel d'agarose 1,5% la fragmentation du génome et essais d'élaboration de banque par la stratégie de digestion ligation avec BsaI. A. Vérification de la fragmentation du génome à différents intervalle de temps. **B.** Vérification de la ré-amplification de la banque après digestion et ligation. Le marqueur de poid moléculaire (PM) dans le premier puit est annoté en paire de base (pb). Dans les puits suivants sont déposés : 3 µL de PCR réalisés avec le couple d'amorce 3 et comme matrice la ligation (étape 2 en suivant la stratégie (B) de digestion et ligation (Figure 39)) amplifiés pendant soit 15 cycles (1) soit 20 cycles (2). Le fragment attendu correspond à un smear tel qu'en A entre 500 et 1500 pb.

Une autre stratégie fut alors envisagée avec une réaction de Clic (Réaction catalysé par du cuivre durant laquelle un azoture réagit avec un alcyne formant un triazole) à l'extrémité

3’ de la première ligation et la fin de la séquence codante pour la GFPmut2 (Figure 39 (C)). Cependant, aucune amplification ne fut observée lors du contrôle par qPCR ainsi que sur gel après la réaction. J’ai donc supposé que le triazole établît par la clic-réaction pouvait gêner la polymérase empêchant l’amplification de la construction et que la qPCR n’était pas la meilleure option pour contrôler la qualité et/ou la quantité des constructions obtenues. J’ai alors choisi une autre méthode pour suivre l’efficacité de la réaction en utilisant un oligo fluoromarqué (Alexa 488) correspondant au brin sens du promoteur T7 avec l’Alexa 488 en 5’. Ce dernier fut introduit dans la construction lors de la première ligation (étape 1), permettant ainsi de visualiser ma construction directement sur gel et ce, dès la première étape. De ce fait, j’ai pu apprécier la taille du fragment obtenu, sans biais issus d’une étape d’amplification, bien que la notion de rendement soit perdue. Par cette nouvelle méthode, j’ai rapidement visualisé que le rendement de la réaction de Clic était très faible avec majoritairement des produits de la taille de ma première ligation.

À la suite de ces nouveaux résultats, j’ai décidé de contrôler à nouveau la qualité de ma première ligation avec ce nouveau moyen de révélation, validant à nouveau cette première étape (Figure 42). Puis, pour la deuxième étape, j’ai choisi de revenir à des expérimentations plus simples avec ce nouveau contrôle potentiellement moins biaisé que la qPCR et ai testé à nouveau différentes conditions de ligation afin d’ajouter la fin de la séquence codante pour la GFPmut2 comme détaillé dans le manuscrit ci-dessous (4.2).

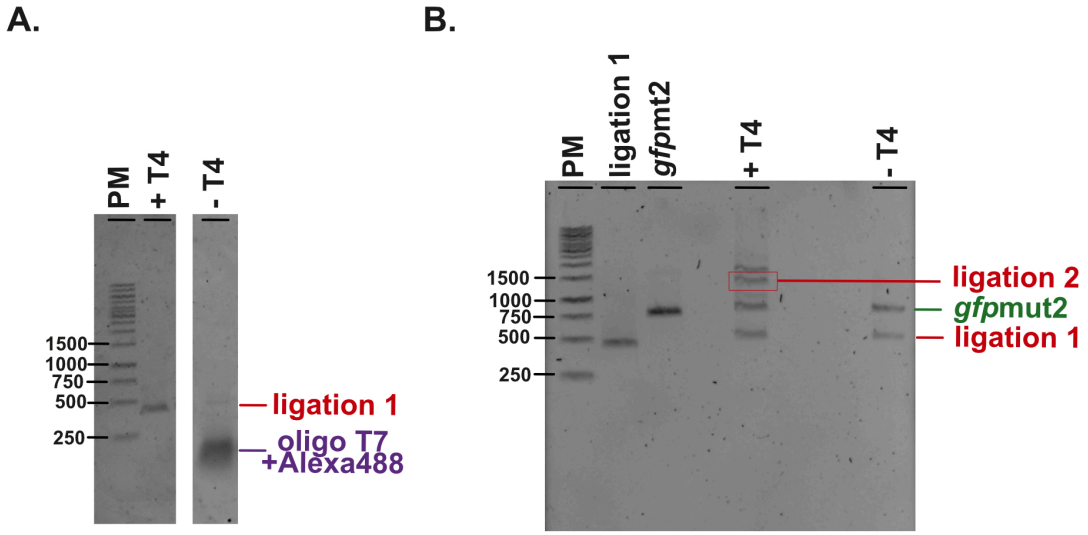


Figure 42 : Vérification sur gel d’acrylamide la ligation 1 et la ligation 2 (optimisée) (Figure 39 D). **A.** Vérification de la ligation 1 (en rouge) (590 pb) avec l’oligo T7 fluoromarqué avec un Alexa488 en 5’ (60 bases). La ligation réalisé avec la T4 ADN ligase (+T4) est déposé juste à côté du marqueur de poids moléculaire (PM) exprimé en paire de base. Pour des raisons d’esthétismes les puits sans lien avec l’expérimentation ont été retirés jusqu’au puit contenant le mélange réactionnel de ligation sans

ligase (-T4). **B.** Vérification de la ligation 2 (en rouge) (1310 pb) avec l'oligo T7 fluoromarqué avec un Alexa488 en 5' de l'ensemble de la construction. La ligation 2 est réalisé à partir de la ligation 1 et la *gfp*mut2 avec la T4 ADN ligase (+T4) ou sans la T4 ADN ligase (-T4).

Finalement, l'approche que j'ai retenue consiste en la réalisation de deux ligations consécutives dans des conditions d'expérimentation précises incluant des éléments bloquants (PEG et tige boucle) aux extrémités 5' du produit de la première ligation et en 5' du brin reverse pour la fin de la séquence codante de la GFPmut2 (Figure 39 (D)), comme présenté et détaillé dans le manuscrit ci-dessous (4.2.1.1). Le PEG ainsi que les tiges boucles vont limiter la circularisassions des fragments (produit de la première ligation ainsi que la GFPmut2) sur eux-même motivant ainsi lors de l'étape 2 la ligation entre ces derniers. Cette procédure s'est révélée concluante n'imposant aucune réelle limite de taille de fragments et son rendement important nous permet d'explorer des génomes de taille importante. Une banque ainsi préparée peut alors faire l'objet d'un criblage fonctionnel en microgouttelettes suivi d'une analyse bio-informatique comme dans mes travaux précédents et comme expliqué dans le manuscrit introduit dans la section suivante.

4.2 Identification de séquences IRES à partir d'un génome viral modèle

J'ai choisi d'utiliser ma méthode de préparation de banques et de la valider à l'aide du CrPV. Ce génome d'une taille modeste (9 896 paires de bases), contient 2 IRES (Wilson et al., 2000; Gross et al., 2017) : l'IRES IGR particulièrement bien caractérisée (Wilson *et al.*, 2000; Hertz and Thompson, 2011) qui a d'ailleurs déjà servi de modèle d'étude à l'équipe (Pernod *et al.*, 2020), et la 5'UTR caractérisée et classifiée plus récemment (Wilson et al., 2000; Gross et al., 2017) Ainsi, ce système représente un modèle intéressant pour valider l'ensemble de notre pipeline de criblage à ultra haut débit. Ces travaux font l'objet d'un manuscrit en cours de préparation et reproduit ci-dessous.

4.2.1.1 Manuscript en preparation : Genome-wide efficient discovery of functional IRES elements using microfluidic-assisted screening

Genome-wide efficient discovery of functional IRES elements using microfluidic-assisted screening

Natacha Dentz, Roger Cubì, Antonin Tidu, Franck Martin et Michael Ryckelynck

Genome-wide efficient discovery of functional IRES elements using microfluidic-assisted screening

Natacha Dentz, Roger Cubì, Antonin Tidu, Franck Martin et Michael Ryckelynck

Université de Strasbourg, CNRS, Architecture et Réactivité de l'ARN, UPR 9002, Strasbourg, F-67000, France

Introduction

Eukaryotic translation involves four main steps (initiation, elongation, termination and recycling of the ribosome (Melnikov et al., 2012)), of which, the initiation is the most targeted for regulation (Jackson et al., 2010). A eukaryotic mRNA usually affords a 5′ end-modified 7-methylguanylate (m7G) cap structure, followed by a 5' untranslated region (UTR), and it is terminated by a polyadenylated (poly-A) tail appended to its 3' end. Eukaryotic protein synthesis can be initiated *via* several mechanisms operating in a cap-dependent or a cap-independent manner. On the first hand, the canonical eukaryotic cap-dependent initiation mechanism starts with the recruitment of the 43S preinitiation complex (PIC) at the level of the 5' cap structure, a step involving the unwinding of the mRNA followed by the scanning of the 5' UTR by the PIC until the initiation codon is reached (Aitken and Lorsch, 2012). On the other hand, translation can be initiated in cap-independent manner through the direct recruitment of the ribosome at the level of highly structured region sitting in the 5' UTR of mRNA and termed Internal Ribosome Entry Site (IRES).

IRES sequences are mainly found in the mRNA of many viruses but also in some eukaryotic ones (Kwan and Thompson, 2019). All viruses rely on the cellular translation machinery to produce their proteins, and they consequently evolved mechanisms to hijack cellular protein production pathway. One of such mechanisms consists in the global suppression of host cap-dependent protein synthesis (Cao et al., 2017; Lee et al., 2017) . In those conditions, the presence of IRES-mediated initiation allows the virus to get its own mRNAs translated

which ensures viral proteins synthesis. Besides, IRES sequences can also be found in eukaryotic mRNAs related stress response such as DNA damage, amino-acid starvation, hypoxia or endoplasmic reticulum stress among others (Godet et al., 2019a). Eukaryotic IRESes appear to be less structured than the viral ones, and their mechanisms of action are less documented (Leppek et al., 2018). Yet, in both cases, IRES-mediated translation can be regulated by so-called IRES Trans Acting Factors (ITAFS) (Yang and Wang, 2019). Viral IRES sequences have been classified in four groups based on their structure and their requirement for translation factors and ITAFS. Usually, type I and II IRES need most of the initiation factors except eiF4E that binds the 5' cap in the cap-dependent translation mechanism, the main difference between both groups lying in the way the ribosome is deposited onto the mRNA. Indeed, whereas type II directly place the ribosome onto the start codon, type I IRES rather deposit the ribosome upstream the start codon and interact with all the elements needed to perform the mRNA scanning that then takes place to find the initiator codon. Type III IRES only requires eIF2 and eiF3 factors and, like type II, also perform the ribosome assembly directly onto the initiation codon. Finally, type IV IRES initiates the translation without the need of any initiator factor (Mailliot and Martin, 2018; Yang and Wang, 2019).

The first IRES were discovered in poliovirus and encephalomyocarditis virus RNA genomes (Jang et al., 1988; Pelletier and Sonenberg, 1988), using a low throughput approach in which mono- and/or bicistronic artificial mRNAs were used to identify the potential IRES contained in the 5' UTR of viral RNAs. Later, high throughput (HT) analyses were developed using a bicistronic plasmid in which either a 50-nucleotide randomized sequence or a pool of 55,000 different fully designed single-stranded 210-nucleotide oligomers were inserted between both cistrons (Venkatesan and Dasgupta, 2001; Weingarten-Gabbay et al., 2016). Other authors used the mRNA display of trillions of ~200 base pairs-long human genomic fragments (Wellensiek et al., 2013). These IRES HT screening analyses allowed the discovery of thousands of human and viral sequences with cap-independent translation activity. Yet, the complexity of these HT analysis pipelines demonstrated how challenging the identification of IRES is. *In silico* prediction of IRES elements have been also developed but they remain challenged by the fact that IRES are rather idiosyncratic elements, and those eukaryotic ones are less structured and more diverse that their viral counterparts. Nevertheless, different prediction tools (Kolekar et al., 2016; Wang and Gribskov, 2019; Zhao et al., 2018, 2020) and databases (Yang et al., 2021) have been developed to integrate all the experimentally demonstrated and the predicted IRES. Therefore, there is still a strong need for an efficient and cost-effective HT analytical method

allowing the robust and automated detection of new IRES elements. In this work, we set-up such a pipeline combining the use of cell-free expression system, droplet-based microfluidics, next generation sequencing (NGS) and bioinformatics. Using it in tandem with a new dedicated procedure to prepare genomic DNA library, we were able to analyze the genome of the Cricket Paralysis Virus (CrPV) and to precisely locate and orient the intergenic type IV IRES. This proof-of-concept experiment demonstrates the efficiency of this new technology that may serve in future experiments aiming at rapidly identify new IRES elements, especially in emerging viruses, and that could eventually serve of as new therapeutic targets, but also to rapidly characterize already known IRES.

Materials and methods

Gene library Generation

First, 3 µg of the amplicon of Cricket Paralysis Virus isolate-3 (reference Genome) (GenBank: KP974707.1) is fragmented during 15 min at 37 °C by 2 µL of DNA fragmentase (NEBNext® dsDNA Fragmentase® M0348) in fragmentase buffer v2 (NEB) diluted 10 times in a final reactional volume of 20 µL. The reaction is stopped by the addition of 2,5 µmol of ÉthylèneDiAmineTétraAcétique (EDTA). Then, fragmented genome diluted 10 times are purified with a ratio of 0,8 of SPRIselect (B23317 Beckman Coulter) according to the supplier recommendation, and eluate the matrice in 20 µL. This one is precipitated in 250 mM of NaCl in EtOH at -20 °C 12 hours. After fragmentation one µg of DNA is end-repair during 30 min at 20 °C and then 30 min at 65 °C with NEB DNA repair mix and protocol from NEB (NEBNext End Repair (E6050)). The matrice dA-tailing is purified by phenol/chloroform precipitation and again by NaCl precipitation.

0,5 pg of this fragmented genome is ligate with 0,5 pg of 5'T7 primer pre-paired by 5 min to 98 °C and 5min to 55 °C (primer T7 forward: GTCCACACGTCCACATAATACGACTCACTATAGGAGACCACAACGGTTTCCCTCC TTTAT and primer T7 reverse: (P) TAAAGGAGGGAAACCGTTGTGGTCTCCTATAGTGAGTCGTATTATGTGGACGTGT GGACG) and 0,5pg of the 3' GFPmut2 primer pre-paired as the T7 primer (primer GFPmut2 forward:(P)AAGGTGAGTAAAGGAGAAGAACTTTTCACTGG and the primer miniGFPmut2 reverse: CCAGTGAAAAGTTCTTCTCCTTTACTCACCTTT) in 10 time diluted of T4 DNA ligase buffer (NEB), 15 % of PEG 8000, 0,5 nmol of ATP (Larova), and

2000 Units of T4 DNA ligase (NEB) 12 hours à 16 °C in a final volume of 50 µL. Ligation are stopped by 20 min at 65 °C and 1 µL of ligation is used as PCR matrice. The PCR mixture use 15 pmol of each primer (Forward BlockedT7: *CACGGTGCAACTTAGCACCGTGCA (hexaethyleneglycol)AGACCACAACGGTTTCCCTCCTTTCCGTCCACATAATACGACTCAC TATAGGCTTCGTATGACTGGGGGTGTTGGG* and Reverse primer smallGFPmut2), Q5 buffer (NEB) 5 times diluted, 0.2 mM of each dNTPs (ThermoScientific 10 mM each) and 4 U of Q5® High-Fidelity DNA Polymerase (New England Biology NEB). The mixture was thermocycled starting with an initial step of denaturation of 2 min at 98 °C followed by 25 cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C to finish it. Then, PCR product was purified with Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR. To pursue, 0,5 pg of this first re-amplification of first ligation purified is ligate with 0,5 pg of the GFPmut2 truncated gene PCR, obtain with 15 pmol of each primer (Forward GFPmut2truncated: AGTTGTCCCAATTCTTGTTGAATTAGATGG and Reverse GFPmut2Blocked: TGCACGGTGCTAAGTTGCACCGTGCCCGTCTTCACCTGGCGACTTAATTTAAATCT TCTTCTGATAATAATTTTTGTTCTAATGC**)**, Q5 buffer (NEB) 5 times diluted, 0.2 mM of each dNTPs (ThermoScientific 10 mM each) and 4 U of Q5® High-Fidelity DNA Polymerase (NEB) which is thermocycled with an initial step of denaturation of 2 min at 98 °C followed by 25 cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C to finish it. Then, PCR product was purified with Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR. This second ligation is realized with 10 time diluted of T4 DNA ligase buffer (NEB), 15 % of PEG 8000, 0,5 nmol of ATP (NEB), and 2000 Units of T4 DNA ligase (NEB) 12 hours à 16 °C in a final volume of 50 µL. To finish, ligation is stopped by 20 min at 65 °C and 1 µL of ligation is used as PCR matrice. The PCR mixture use 15 pmol of each primer (Forward constant2 (Fwd cst2): AGACCACAACGGTTTCCCTCCTTTC and Reverse constant3 (Rev cst3): CCCGTCTTCACCTGGCGAC**)**, Q5 buffer (NEB) 5 times diluted, 0.2 mM of each dNTPs (ThermoScientific 10 mM each) and 4 U of Q5® High-Fidelity DNA Polymerase (Biology NEB). The mixture was thermocycled starting with an initial step of denaturation of 2 min at 98 °C followed by 25 cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C to finish it. Then, PCR product was purified with Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR and matrices are eluted in 30 µL of water nuclease free.

Extract preparation

Final mix of Reticulocytes Rabbit Lysate (RRL) extract contain 0,1125 mM of amino-acid mix, 0,1 M of Potassium acetate (KAc), 5 mM of Magnesium acetate ($MgAc_2$), 1 mM of ribonucleotides (rNTP), 5 % of T7 RNA polymerase (2,5 mg/ml in 50 % of glycerol, 100 mM of NaCl and 20 mM of Tris HCl pH 7.5, 0,5 U of RNase inhibitor and 50 % of extract is added to reactional mixture.

Microfluidics screening

Microfluidic chips were molded into polydimethylsiloxane (PDMS) and electrodes were fabricated as described in Ryckelynck et al., 2015.

For the Droplet digital PCR, the libraries were diluted in 200 µg/mL yeast total RNA solution (Ambion) to obtain the desired droplet occupancy. 1 µL of this dilution was introduced in 100 µL of PCR mixture containing 15 pmol of Fwd cst2 and Rev cst3, 0.2 mM of each dNTPs (ThermoScientific 10 mM each), 10 µM Cyanine 5 (Thermo Fisher), 0.1 % Pluronic F68 (Sigma), 2 U of Q5 DNA polymerase (NEB) and the corresponding buffer at the recommended dilution. The mixture was loaded into a PTFE tubing (Thermo) and infused into a droplet generator microfluidic chip where it was dispersed in 2.5 pL droplets carried by an HFE 7500 fluorinated oil (3 M) supplemented with 3 % of a surfactant. Droplet production frequency (∼10,000 droplets per second) was monitored in real time using an optical device and a software developed by the team (Ryckelynck et al., 2015). 2.5 pL droplets were generated by adjusting pump flow rates (MFCS, Fluigent). The emulsion was collected into 0.2 mL tube and subjected to an initial denaturation step of 2 min at 98 °C followed by 25 PCR cycles of: 15 sec at 98 °C, 20 sec at 60 °C and 1 min 45 at 72 °C, then 2 min at 72 °C.

For the addition of extract to allow *in vitro* transcription translation (IVTT) by droplet fusion, PCR droplets were reinjected into a droplet fusion device at a rate of ∼1500 droplets per second as described previously (Ryckelynck et al., 2015). PCR droplets were spaced by a stream of HFE 7500 fluorinated oil (3 M) to synchronized and paired one-to-one with a 17,5 pL RRL or S2 extract (Prepared by collaborator's team) supplemented with 0.1 % of Pluronic F68, 1 µM of Cy5, 17.5 µg/mL T7 RNA polymerase (purified in the laboratory). Extract (IVTT mixture) was loaded in a length of PTFE tubing and kept on ice during the experiment. IVTT droplets were produced using an HFE 7500 fluorinated oil (3 M) stream supplemented with 3 % (w/w) of surfactant. Flow rates (MFCS, Fluigent) were adjusted to generate 17,5 pL *IVTT* droplets and to maximize the synchronization of 1 PCR droplet with 1 *IVT* droplet. Pairs of droplets

were then fused with an AC field (500 mV at 30 kHz) and the resulting emulsion was collected in a tube and then incubated for 2 h at 30 °C for RRL extract and 2 h at 30 °C for RRL extract. For the droplet fluorescence analysis and sorting, the emulsion was finally re-injected into an analysis and sorting microfluidic device (Ryckelynck et al., 2015). Droplets were injected at a frequency of ~300 droplets per second and spaced with a stream of HFE 7500 fluorinated oil (3 M). Green (Green fluorescent protein GFPmut2) and red (Cyanine 5) fluorescence of each droplet was analyzed. In rounds of selection, all droplets with green fluorescence displaying a red fluorescence corresponding to single fused droplets (Fig) were targeted and deflected into the collection channel by applying an AC field (1200 mV 30 kHz) and collected into a 2 mL tube. Sorted droplets were recovered from the collection tubing by flushing 200 µL of HFE 7500 fluorinated oil (3 M), then 150 µL of 1H, 1H, 2H, 2H-perfluoro-1-octanol (Sigma-Aldrich) and 200 µL of 200 µg/mL yeast total RNA solution. The droplets were broken by vortexing the mixture. DNA-containing aqueous phase was finally transferred to a new tube. An aliquot of DNA-containing aqueous phase is used to do Droplet Digital PCR as describe before with the same primer.

Enrichment test

An aliquot (3 µL) of DNA-containing aqueous phase recovered at the end of each round of screening (or from the starting library R0) was amplified into 100 µL of PCR mixture containing 15 pmol of Fwd cst2 and Rev cst3, 0.2 mM of each dNTP (ThermoScientific 10 mM each), 2 U of Q5 DNA polymerase and the corresponding buffer at recommended concentration. The mixture was thermocycled with an initial denaturation step of 1 min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 60 °C for 20 sec and 72 °C for 1 min 45, then 72 °C for 2 min. PCR products were finally purified with Sera-Mag beads and quantified with a Nanodrop (Thermo Scientific). 2 µL of PCR product was introduced in an extract for *in vitro* transcription translation (IVTT) containing RRL (Team's collaborator). The green fluorescence (ex: 488 nm/em: 535 nm) of GFPmut2 was then monitored every minute for 2 h at 30 °C for RRL extract on spectrophotometer (SpectraMax® iD3 Molecular Device).

NGS library preparation

An aliquot of DNA-containing aqueous phase recovered at the end of each round of screening (or from the starting library) was amplified into 100 µL of digital PCR mixture containing 15

pmol of Forward addi2: *TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG*CTTCGTATGACTGGGGGTGTTGGG and Reverse addi1: GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCAGTGAAAAGTTCTTCTCCT TTACTCACCTT), 0.2 mM of each dNTP (ThermoScientific 10 mM each), 2 U of Q5 DNA polymerase and the corresponding buffer at recommended concentration (NEB). The PCR was thermocycled with an initial denaturation step of 1 min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 60 °C for 20 sec and 72°C for 1 min, then 72 °C for 2 min and then, were finally purified using a Sera-Mag™ (Merck) and a ratio of 2,5 volumes of bead for one volume of PCR and quantified with a Nanodrop (Thermo Scientific). Then, Illumina-index (Nextera XT DNA Library Preparation Kit Illumina) was added by PCR which contains 1 ng of PCR-i1/i2 added purified, 10 µL of both Index Illumina forward and reverse, 0.2 mM of eachs dNTPs (ThermoScientific 10 mM each), 4 U of Q5® High-Fidelity DNA Polymerase and the corresponding buffer (NEB). The mixture was thermocycled with an initial denaturation step of 1 min at 98 °C followed by repetitions of the two-step cycle: 98 °C for 15 sec, 62 °C for 20 sec and 72 °C for 1 min 15, then 72 °C for 2 min and then, were finally purified using a SPRI beads with 0,8 of ratio bead/PCR (Beckman coulter).

Bioinformatic analysis

STAR mapper was used to align the fastq reads onto the Cricket Paralysis Virus isolate-3 genome (reference Genome) (GenBank: KP974707.1). To generate the genome indices, a custom gtf file was generated. At the mapping step, the clip5pNbases parameter was set to 59 33 to eliminate the 5' and 3' adaptors and bedtools was used on the generated bam file to obtain the genome coverage.

Results and discussions

Overall experimental strategy

The goal of this work was to set-up a high-throughput (HT) analytical pipeline aimed at rapidly identifying IRES elements from viral genome. The overall idea was to fragment the genome into pieces large enough to contain a full-length functional IRES, and to insert each fragment between a transcription promoter allowing the production of a messenger RNA (mRNA) and the sequence coding for a reporter protein (Figure 1). Here, we chose to express the library from the promoter of the T7 RNA polymerase promoter, a highly efficient monomeric enzyme whom the high activity should not limit the overall assay. Moreover, the produced transcript will display a triphosphate 5' end not competent to support cap-dependent initiation. A HT screening best operates using assays with a fluorescent readout as this signal is easy to detect in a non-invasive way, offers good sensitivity and in an automatable manner. We therefore chose an improved variant of the Green Fluorescent Protein (GFPmut2, (Zhang et al., 1996)) as reporter protein. Moreover, using a fluorescent protein allows a simple and direct correlation between the amount of protein produced and the measured signal.

As model system, we chose the Cricket Paralysis Virus (CrPV), a well-characterized member of the *dicistroviridae* family made of a 9196-nucleotide long RNA genome organized in two main regions (coding for non-structural and capsid proteins) separated by a non-coding one termed the intergenic region, or IGR (Figure 2A). Moreover, two IRES elements were identified in this virus: a type III IRES located in the 5'UTR of the genomic RNA (Gross *et al*.,2017) and a type IV IRES located in the IGR (Wilson *et al*.,2000). As starting material, we chose to work with a DNA replicon, i.e., the double-stranded DNA copy of the viral genome inserted in a replicating plasmid, forming a construction spanning 12,496 base pairs. This model genome had then to be fragmented prior to being used to prepare a library of fragments later subjected to a functional screening in droplet-based microfluidic format (Figure 1). Sequencing the content of the enriched libraries and using bioinformatics to map the reads onto the genome of reference then enabled to identify those constructs carrying a functional IRES element and to precisely delimit it.

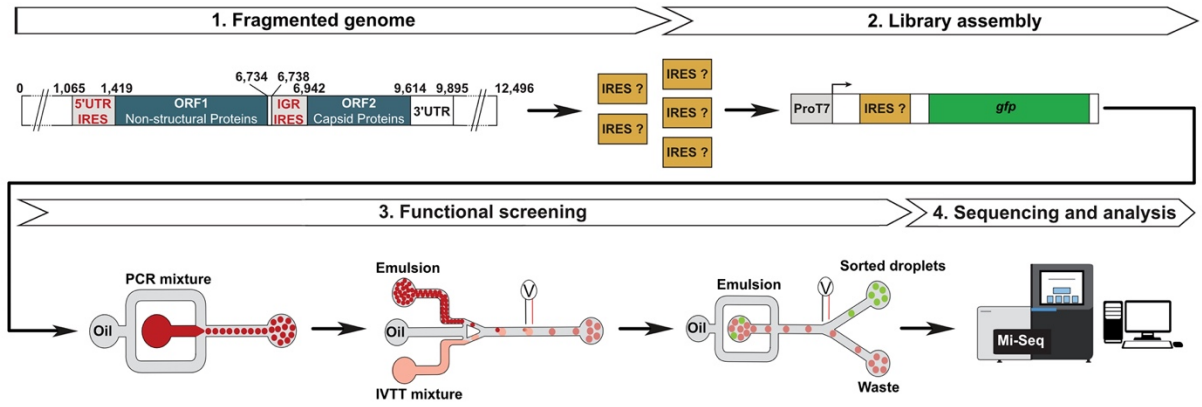


Figure 1 : Experimental screening-based strategy to search for new potential IRES sequence from viral genomes. *(1) A viral genome is fragmented into pieces (brown boxes) prior to (2) adding the T7 RNA polymerase promoter (grey box) and the gfpmut2 coding region (green box). (3) A 3-step droplet-based microfluidics workflow is then used to screen the library. Aqueous phases and water-in-oil droplets are shown in dark or light red according to their Cy5 content (high and low concentration respectively). The oil phase is represented in grey. Upon gene expression, those droplets displaying GFP fluorescence (shown in green) are sorted from the bulk of the GFP-free droplets (left in light red). (4) Finally, the content of the different library (starting and enriched) are analysed by next generation sequencing and bioinformatics.*

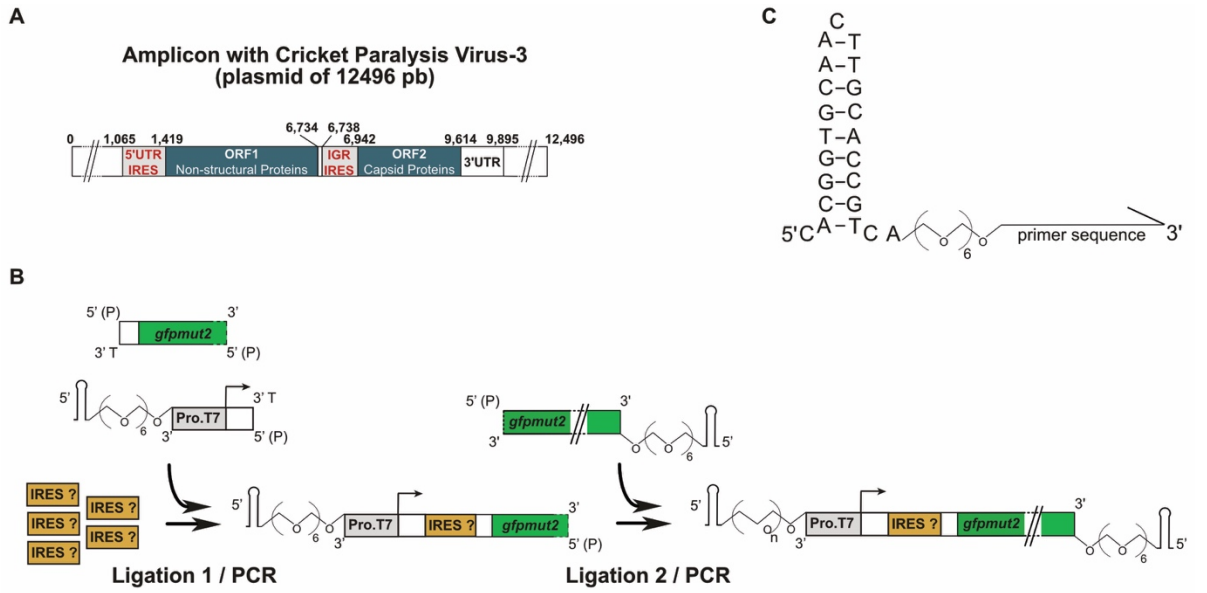


Figure 2 : Library preparation and validation. A. *Organization of the CrPV viral amplicon. Coding regions are shown in blue whereas 5' UTR and IGR containing an IRES are represented in grey and the 3'UTR is in white. Note that this genome is embedded in a plasmid represented by the white boxes at each end and that, even though shown here as a linear molecule for illustration purposes, the replicon was a circular DNA molecule.* ***B.*** *Schematic sum-up of the library preparation procedure. The pieces of fragmented genome are shown as brown boxes, the T7 RNA polymerase promoter is in grey, the gfp*

coding region is in green. The blocking hairpin is represented with its hexaethylene-glycol when present at an extremity and the 5' phosphorylated ends are indicated by a (P). ***C.*** *Typical organization of a blocked primer.*

Preparation of the DNA genomic library

A first instrumental step consisted in the development of a robust and highly efficient procedure to prepare a genomic library (Figure 2B). The CrPV replicon was first randomly fragmented into 400 to 800-bp long pieces (a range of size expected to be large enough to contain a full-length functional IRES) using dsDNA Fragmentase® (New England Biolabs). Upon end-repair, small adaptors (one containing the T7 RNA polymerase promoter and the other one the first codons of *gfpmut2* gene) were appended at each extremity of the fragment. These adaptors were made of two synthetic oligonucleotides annealed together and whom one had its 5' end phosphorylated. Moreover, T overhangs were also added to orient the proper ligation of the desired fragments. Finally, a blocking structure was introduced at the 5' end of the adaptor carrying the T7 transcription promoter. This structure (Figure 2C) corresponds to a stable non-phosphorylated stem-loop creating a steric hindrance that prevents the self-circularization of long DNA constructs to occur. Moreover, this structured element was connected to the rest of the oligonucleotide via an internal hexaethylene-glycol preventing its copy by the DNA polymerase during PCR amplification, preserving therefore is single-stranded nature and its capacity to block ligation. Ligation products were then PCR amplified using a blocked primer annealing on the T7 promoter-containing adaptor and a 5' phosphorylated oligonucleotide annealing to the *gfpmut2* fragment as second primer (Figure 2B). In parallel, a long fragment corresponding to the rest of GFPmut2-coding region was prepared by PCR using a 5' phosphorylated oligonucleotide annealing to the 5' part of the *gfpmut2* as first primer and a blocked oligonucleotide annealing to the 3' end of *gfpmut2* as second primer. Both fragments were finally ligated together prior to PCR amplifying the construct of interest (i.e., containing T7 RNA polymerase promoter, a genomic fragment and the *gfpmut2*). The blocked extremities played an instrumental role here by preventing the circularization of long DNA fragments and instead favored intermolecular ligation events. Treating CrPV replicon by this approach allowed preparing a genomic starting library (R0) mainly deprived of bias (both random cleavage and insertion were verified by NGS).

Functional screening of the library for IRES elements

The presence of an IRES in a genome fragment should, in principle, lead to ribosome recruitment and drive to the translation of the GFPmut2-coding region downstream the fragment and therefore to fluorescence emission by the synthesized GFPmut2, provided the fragment was inserted in the right orientation and the start codon of the IRES was in frame with the *gfpmut2* fragment. As the presence of a functional IRES (i.e., a sequence containing a full-length minimal IRES, in the right orientation and in frame with *gfpmut2*) was expected to be a rather rare event, the library was screened using an ultrahigh-throughput strategy (Figure 1). To do so, we adapted the μIVC-seq strategy we previously used to develop light-up RNA aptamers and Fluorogenic RNA-based Biosensors (Autour et al., 2019; Bouhedda et al., 2021).

Briefly, DNA fragments contained in the library were diluted into a PCR amplification mixture prior to disperse the solution into 2.5 pL water-in-oil droplets generated using a microfluidic device and stabilized by a surfactant. To maximize the chances of identifying rare molecules of interest, DNA dilution was adjusted such that each PCR droplet contained on average 2 DNA templates. The emulsion was collected and thermocycled to amplify each template ~ 100,000 times.

Then, the emulsion was reinjected into a second microfluidic device where PCR droplets were spaced and synchronized one-to-one with larger 16 pL droplets containing an *in vitro* transcription and translation (IVTT) mixture made of Rabbit Reticulocyte Lysate (RRL) supplemented in T7 RNA polymerase as we recently validated (Pernod et al., 2020). Pairs of droplets were then fused using an AC electric field and the resulting emulsion was incubated for 2 hours at 30°C to enable transcription and translation of each construct to take place. During this step, those constructs containing a functional IRES were expected to produce GFPmut2, so to turn the droplet they were contained in fluorescent. The emulsion was finally reinjected into a last device in which the fluorescence of each droplet was analyzed and used to sort them according to number of PCR droplets fused per IVTT droplet (red fluorescence on Figure 3A) and the amount of GFPmut2 produced per droplet (green fluorescence on Figure 3A). Those most green fluorescent droplets corresponding to an IVTT droplet fused with a single PCR one were sorted and their DNA content recovered. The round of screening was validated by amplifying the DNA and expressing it by IVTT in the presence of ^{S35}Methionine followed by a gel analysis. An increased amount of radioactive product at the size of GFPmut2 validated the round. Upon validation, the recovered DNA was used to prime a new round of screening and the whole process was repeated twice (Figure 3A). An impressive improvement in the

translation rate was observed along the different rounds (Figure 3B) indicating the rapid enrichment of the library in molecules able to efficiently initiate *gfpmut2* translation in a cap-independent manner, so likely to contain IRES elements. The different libraries were then indexed and sent off for Next Generation Sequencing on a MiSeq Illumina platform.

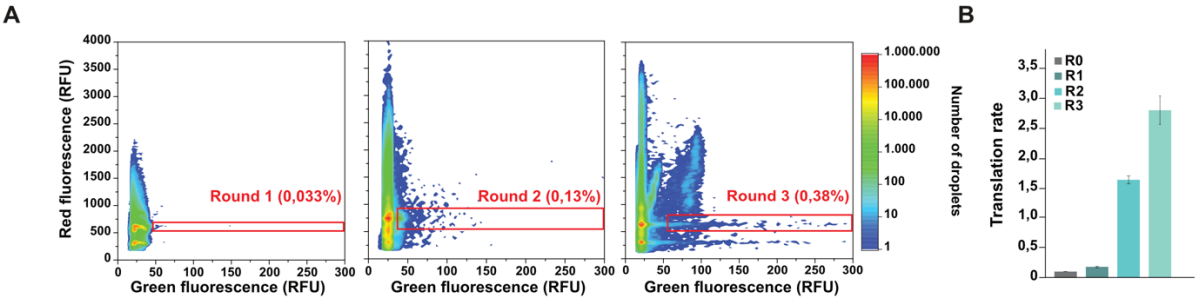


Figure 3 : Fluorescence profiles of rounds of screening and enrichment in sequences favoring translation initiation. A. Fluorescence profiles recorded during the three screening steps. The sorted droplets are boxed in red and the percentage of the population they represent is given. The red (Cy5) fluorescence informs on the number of PCR droplets fused per IVTT one, whereas the green fluorescence reports on the amount of GFPmut2 produced per droplet. **B.** Validation of the rounds of screening by enrichment assay. Each library was PCR amplified and in vitro expressed in the presence of ^{S35}Met prior to analyze the translation product by gel electrophoresis. The values are the mean of 3 independent experiments and the error-bars correspond to ± 1 standard deviation.

Analysis of the selected sequences

Upon paired-end sequencing of the different libraries (starting and enriched), the reads were mapped onto the replicon reference genome and the coverage of each nucleotide was computed (Figure 4A and B). First, we were pleased to observe that the starting library (R0) displayed a homogenous coverage all along the genome, confirming the efficiency and the absence of significant bias of the method we set-up to prepare libraries. However, while the selection process progressed, we observed a significant bias toward the region spanning nucleotide 6500 to 7150, and that gathered more than 20 % of the reads by the third round (R3) of screening (Figure 4B, inset). Excitingly, this region encompasses the IGR and contained therefore a type IV IRES (region shadowed in grey on Figure 4). Moreover, the coverage of this region constantly increased throughout the process (Figure 4B), mirroring the increase in translation efficiency we observed (Figure 3B), and confirming that libraries were enriching in functional IGR IRES. We did not detect any enrichment at the level of the 5' UTR region known to contain a second IRES (Masoumi et al., 2003). Nevertheless, the functionality of this IRES relies on

the presence of the ITAF RACK1, a protein factor absent in RRL extracts, explaining the absence of enrichment of this region.

Looking closely at the coverage profiles, we also noticed interesting features especially with the most enriched R3 library (Figure 4C). Indeed, whereas residues spanning the region 6738 - 6942 displayed an elevated coverage, as expected for nucleotides forming the functional IRES, the coverage decreased in an asymmetric manner on each side of the element. In fact, the coverage decreased in a rather gentle and progressive manner on the 5' side (red dotted box on Figure 4C) of the IRES as expected for a region not subjected to a pressure linked to a function, therefore highly tolerant to variation. On the contrary, a steep and stepped decrease was observed on the 3' side (blue dotted box on Figure 4C) suggesting that functional constraints were exerted during the selection. First, the steep decrease is likely the result of a pressure exerted by the GFPmut2 since, even though the protein tolerates N-terminal additions, its folding efficiency may be affected all the more a longer peptide is appended. Alternatively, though not exclusive of the former explanation, the addition of some viral encoded amino acids may have a stronger negative effect on GFPmut2 function independently of the absolute length of the added peptide. In both cases, the negative effect of peptide addition was mainly marked over the 30 first codons (positions 6942 to 7032) and became less pronounced after the insertion of the 31th (see the break labeled by an asterisk on Figure 4C). Second, the stepped decrease profile brought additional information on the reading frame. Computing the absolute value of the coverage difference between the nucleotide n and the nucleotide n-1 highlighted a significant bias in the distance spacing the peaks on the 3' side (Figure 4D and E). Indeed, whereas peaks were rather randomly spaced on the 5' side (Figure 4D and F) they tended to be spaced by a phased distance characterized by a 3-nucleotide period on the 3' side (Figure 4E and F), a feature consistent with an IRES in frame with the *gfpmut2* coding sequence. Consistently, most of the IGR IRES residues were in frame on the 3' side, while such a correlation was missing on the 5' side (Figure 4G). Taken together, these data indicate that analyzing the steepness and the stepping allows to functionally orient the IRES, identify its frame and eventually located the region containing the start codon (labeled by the black arrow on Figure 4C and E).

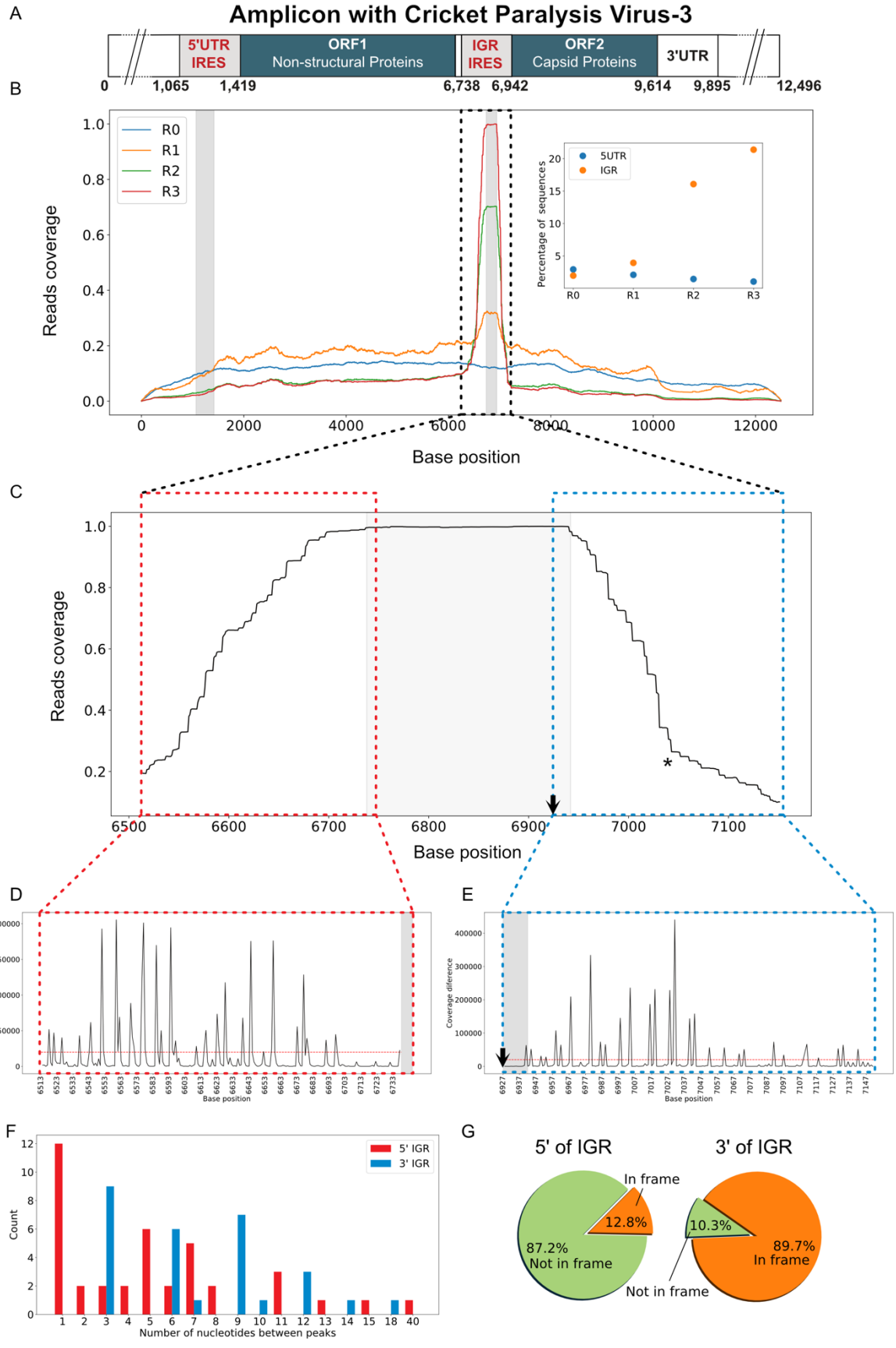

A
Amplicon with Cricket Paralysis Virus-3
5'UTR IRES
ORF1 Non-structural Proteins
IGR IRES
ORF2 Capsid Proteins
3'UTR
0
1,065
1,419
6,738
6,942
9,614
9,895
12,496
B
R0
R1
R2
R3
Reads coverage
Base position
5UTR
IGR
Percentage of sequences
C
Reads coverage
Base position
D
Coverage diference
Base position
E
Coverage diference
Base position
F
5' IGR
3' IGR
Count
Number of nucleotides between peaks
G
5' of IGR
In frame
12.8%
87.2%
Not in frame
3' of IGR
10.3%
89.7%
In frame
Not in frame

Figure 4: **Sequence analysis of libraries content. A.** Schematic representation of the CrPV replicon used to prepare the library and as reference genome. The different regions were color-coded like on Figure 2. **B.** Read coverage of the starting (R0) and the enriched libraries (R1, R2, R3) obtained after each round of screening. The regions expected to contain an IRES (5'UTR and IGR) are shadowed in grey. The enrichment of each IRES throughout the process is shown in the inset. **C.** Read coverage of the region containing the IGR. Only the data of R3 library are shown. The 5' region preceding the IRES is box-dotted in red, while the 3' region preceding the IRES is box-dotted in blue. The region containing the IRES is shadowed in grey. The position of the start codon is labelled by the black arrow and the asterisk indicates an abrupt slope change in the read coverage. **D.** Analysis of the differential coverage between consecutive nucleotides ($n - n_{-1}$) in the 5' region of the IGR IRES. Values are represented as absolute values. The horizontal red dotted line represents the peak-selection threshold used to compute the nucleotide distance between peaks. The light grey box delimits the 5' sequence of the IGR. **E.** Analysis of the differential coverage between consecutive nucleotides ($n - n_{-1}$) in the 3' region of the IGR IRES. Values are represented as absolute values. The horizontal red dotted line represents the peak-selection threshold used to compute the nucleotide distance between peaks. The light grey box delimits the 5' sequence of the IGR. The black arrow indicates the position of the start codon. **F.** Bar plot representing the occurrence of the distances spacing the coverage peaks passing the threshold line on plots D and E. The peak distance occurrences of the 5' end of the IRES are shown in red, and those of the 3' region are shown in blue. **G.** Distribution of the coverage shown on E and F according to their frame.

Conclusions

Considered for a long time as viral hallmarks, Internal Ribosome Entry Sites (IRES) are indeed common and spread among RNA viruses, but they have since then also been identified in eukaryotic genomes where they are found often associated with stress adaptation pathways and disorders as severe as cancer (Godet et al., 2019b; Lacerda et al., 2017; Sriram et al., 2018). Indeed, the current version of IRESbase gathers 1328 IRESs, including 774 eukaryotic IRESs from 11 eukaryotic organisms and 554 viral IRESs from 198 viruses (Zhao et al., 2020). These features together with the rather idiosyncratic nature of IRESs make them interesting targets for drug development. Yet the identification of a new IRES remains a long and difficult task making urgent the development of robust, rapid, efficient, and cost-effective technologies dedicated to their high-throughput discovery. In this work, we introduced such a technology that combines an efficient method to prepare genomic libraries together with the combined use of droplet microfluidic-assisted ultrahigh-through functional screening, next generation sequencing (NGS) and bioinformatics. This technology is somehow related to µIVC-seq (Autour et al., 2019; Bouhedda et al., 2021) and µIVC-Useq (Cubi et al., 2021) we recently introduced for the rapid development of light-up RNA aptamers and fluorogenic RNA-based aptamers and that proved to be highly efficient.

The approach we set-up to prepare our genomic library was directly inspired by the methods that proved to be efficient for the preparation of NGS libraries. The most important improvement was the use of the blocked oligonucleotides preventing self-ligation reactions to take place and strongly favoring the formation of wished products. Even though the method was validated with a small viral genome, it was conceived with the idea of being easily extendable to much larger genomes and the procedure should therefore be directly applicable to any source of DNA. The limitation would rather come from the microfluidic screening itself since a typical droplet-based experiment can hardly handle more than 10 million droplets per experiment, currently restricting the current use of this technology to small genomes. Yet, a great advantage of our *in vitro* screening technology is related to the use of a cell-free extract that can easily be supplemented with other factors (e.g., purified ITAF, enriched cell extract…) or even exchanged for an extract prepared from another cell type (e.g., mammalian cells, insect cells, yeast, bacteria…), making our approach highly versatile and adaptable. Among other appealing applications, analyzing the genomic RNA of emerging viruses (e.g., Zika, Dengue, SARS-Cov2…) with this technology may enable the identification of new IRES elements, becoming as many new potential targets for the development of specific antiviral drugs. Besides the discovering new IRES, this technology could also be used to characterized existing ones. For instance, mutant libraries could be prepared by saturation mutagenesis of some chosen positions of by random mutagenesis. Screening these libraries using the technology introduced in this work, would allow the rapid identification of mutations affecting or not the IRES as well as possible sequence covariations bringing highly valuable structural insights.

References

Aitken, C.E., and Lorsch, J.R. (2012). A mechanistic overview of translation initiation in eukaryotes. Nature Structural & Molecular Biology *19*, 568–576.

Autour, A., Bouhedda, F., Cubi, R., and Ryckelynck, M. (2019). Optimization of fluorogenic RNA-based biosensors using droplet-based microfluidic ultrahigh-throughput screening. Methods *161*, 46–53.

Bouhedda, F., Cubi, R., Baudrey, S., and Ryckelynck, M. (2021). µIVC-Seq: a method for ultrahigh-throughput development and functional characterization of small RNAs. Methods Mol. Biol. *2300*, 203–237.

Cao, S., Dhungel, P., and Yang, Z. (2017). Going against the Tide: Selective Cellular Protein Synthesis during Virally Induced Host Shutoff. Journal of Virology *91*.

Cubi, R., Bouhedda, F., Collot, M., Klymchenko, A.S., and Ryckelynck, M. (2021). μIVC-Useq: a microfluidic-assisted high-throughput functionnal screening in tandem with next generation sequencing and artificial neural network to rapidly characterize RNA molecules. RNA rna.077586.120.

Godet, A.-C., David, F., Hantelys, F., Tatin, F., Lacazette, E., Garmy-Susini, B., and Prats, A.-C. (2019a). IRES Trans-Acting Factors, Key Actors of the Stress Response. International Journal of Molecular Sciences *20*, 924.

Godet, A.C., David, F., Hantelys, F., Tatin, F., Lacazette, E., Garmy-Susini, B., and Prats, A.C. (2019b). IRES trans-acting factors, key actors of the stress response. International Journal of Molecular Sciences *20*.

Gross, L., Vicens, Q., Einhorn, E., Noireterre, A., Schaeffer, L., Khun, L., Imler, J., Meignin, C., Martin, F. (2017). The IRES 5'UTR of the dicistrovirus criket paralysis virus is a type III IRES containing an essential pseudoknot structure. Nucleic Acids Research *45*, 8993-9004.

Jackson, R.J., Hellen, C.U.T., and Pestova, T. v. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. Nature Reviews Molecular Cell Biology *11*, 113–127.

Jang, S.K., Kräusslich, H.G., Nicklin, M.J., Duke, G.M., Palmenberg, A.C., and Wimmer, E. (1988). A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. Journal of Virology *62*, 2636–2643.

Kolekar, P., Pataskar, A., Kulkarni-Kale, U., Pal, J., and Kulkarni, A. (2016). IRESPred: Web Server for Prediction of Cellular and Viral Internal Ribosome Entry Site (IRES). Scientific Reports *6*, 27436.

Kwan, T., and Thompson, S.R. (2019). Noncanonical translation initiation in eukaryotes. Cold Spring Harbor Perspectives in Biology *11*.

Lacerda, R., Menezes, J., and Romão, L. (2017). More than just scanning: the importance of cap-independent mRNA translation initiation for cellular stress response and cancer. Cellular and Molecular Life Sciences *74*, 1659–1680.

Lee, K.M., Chen, C.J., and Shih, S.R. (2017). Regulation Mechanisms of Viral IRES-Driven Translation. Trends in Microbiology *25*, 546–561.

Leppek, K., Das, R., and Barna, M. (2018). Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nature Reviews Molecular Cell Biology *19*, 158–174.

Mailliot, J., and Martin, F. (2018). Viral internal ribosomal entry sites: four classes for one goal: Viral internal ribosomal entry sites. Wiley Interdisciplinary Reviews: RNA *9*, e1458.

Masoumi, A., Hanzlik, T.N., and Christian, P.D. (2003). Functionality of the 5′- and intergenic IRES elements of cricket paralysis virus in a range of insect cell lines, and its relationship with viral activities. Virus Research *94*, 113–120.

Melnikov, S., Ben-Shem, A., Garreau de Loubresse, N., Jenner, L., Yusupova, G., and Yusupov, M. (2012). One core, two shells: bacterial and eukaryotic ribosomes. Nature Structural & Molecular Biology *19*, 560–567.

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

Pernod, K., Schaeffer, L., Chicher, J., Hok, E., Rick, C., Geslain, R., Eriani, G., Westhof, E., Ryckelynck, M., and 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 Research *48*, 6170–6183.

Ryckelynck, M., Baudrey, S., Rick, C., Marin, A., Coldren, F., Westhof, E., and Griffiths, A.D. (2015). Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions. RNA *21*, 458–469.

Sriram, A., Bohlen, J., and Teleman, A.A. (2018). Translation acrobatics: how cancer cells exploit alternate modes of translational initiation. EMBO Reports *19*.

Venkatesan, A., and Dasgupta, A. (2001). Novel Fluorescence-Based Screen To Identify Small Synthetic Internal Ribosome Entry Site Elements. Molecular and Cellular Biology *21*, 2826–2837.

Wang, J., and Gribskov, M. (2019). IRESpy: an XGBoost model for prediction of internal ribosome entry sites. BMC Bioinformatics *20*, 409.

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

Wellensiek, B.P., Larsen, A.C., Stephens, B., Kukurba, K., Waern, K., Briones, N., Liu, L., Snyder, M., Jacobs, B.L., Kumar, S., et al. (2013). Genome-wide profiling of human cap-independent translation-enhancing elements. Nature Methods *10*, 747–750.

Wilson, J., Powell, M., Hoover, S., Sarnow, P. (2000). Maturally occurring Dicistronic Cricket Paralysis Virus RNA is regulated by two Internal Ribosome Entry Sites *20*, 4990-4999.

Yang, Y., and Wang, Z. (2019). IRES-mediated cap-independent translation, a path leading to hidden proteome. Journal of Molecular Cell Biology *11*, 911–919.

Yang, T.-H., Wang, C.-Y., Tsai, H.-C., and Liu, C.-T. (2021). Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans. Database *2021*, baab025.

Zhang, G., Gurtu, V., and Kain, S.R. (1996). An Enhanced Green Fluorescent Protein Allows Sensitive Detection of Gene Transfer in Mammalian Cells. *227*, 707–711.

Zhao, J., Wu, J., Xu, T., Yang, Q., He, J., and Song, X. (2018). IRESfinder: Identifying RNA internal ribosome entry site in eukaryotic cell using framed k-mer features. Journal of Genetics and Genomics *45*, 403–406.

Zhao, J., Li, Y., Wang, C., Zhang, H., Zhang, H., Jiang, B., Guo, X., and Song, X. (2020). IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. Genomics, Proteomics & Bioinformatics *18*, 129–139.

Pour résumer, ma méthode de préparation de banques a pu être validée puisque les fragments de génomes fusionnés au promoteur et au gène rapporteur, constituant la banque initiale, couvrent l'ensemble du génome étudié.

En revanche, seule l'IRES IGR a pu être identifiée, l'IRES 5' restant totalement absente des banques enrichies. Il faut cependant noter que les sélections ont été réalisées dans des extraits de réticulocytes de lapins alors que le virus a pour tropisme naturel des cellules d'insectes. Ainsi, en complément des travaux décrits dans le manuscrit, de nouvelles expériences m'ont conduit à répéter les sélections dans des extraits de cellules d'insectes S2 (Schneider 2). Le premier avantage de ces extraits est l'absence d'hémoglobine et donc une détection de la fluorescence facilitée (n'étant plus absorbé par l'extrait tel que par les RRL de couleur rouge). D'autre part, ces extraits contiennent naturellement la protéine RACK1 (de l'anglais : « Receptor for Activated C Kinase 1 »), une protéine associée à la SU 40S du ribosome eucaryote impliqué dans des voies de signalisation cellulaires mais surtout indispensable au fonctionnement de l'IRES 5'UTR (Majzoub *et al.*, 2014).

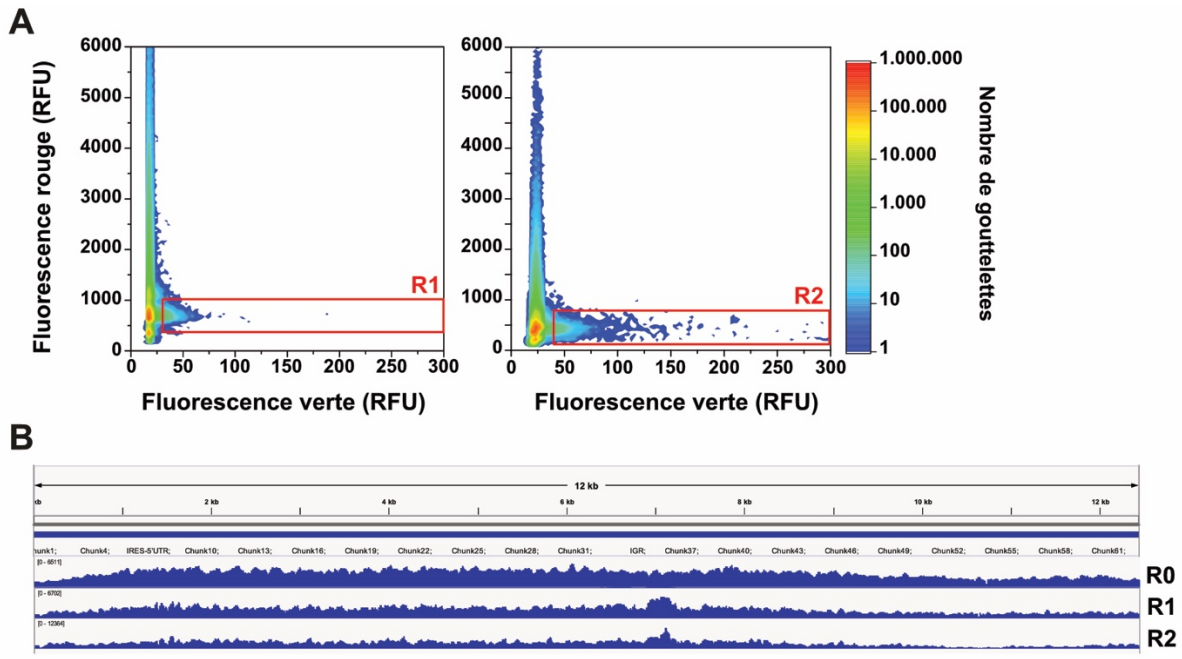


Figure 43 : Tour de sélection à l'aide d'extraits de cellule de S2. A. Tours de sélection consécutifs réalisés dans le processus microfluidique. En ordonné la fluorescence rouge illustre la taille des gouttelettes selon la concentration en Cyanine 5 et en abscisse la fluorescence verte correspond à la GFPmut2 traduite dans les gouttelettes. **B.** Représentation du placement des couples de séquence obtenus après NGS sur le génome de référence. Le génome de référence est illustré de 0 à 12 kb et les paires de séquences obtenues après séquençage au Mi-seq de chaque tour de sélection, depuis la banque initiale jusqu'au deuxième tour consécutif (R0 à R2) sont représentés en bleu selon leur localisation sur le génome.

J'ai pu réaliser deux tours de sélections en extraits S2 de la banque initialement utilisée avec les extraits RRL présentés plus haut (Figure 43). Je me suis cependant heurtée à un important problème de bruit de fond vert (Figure 44).

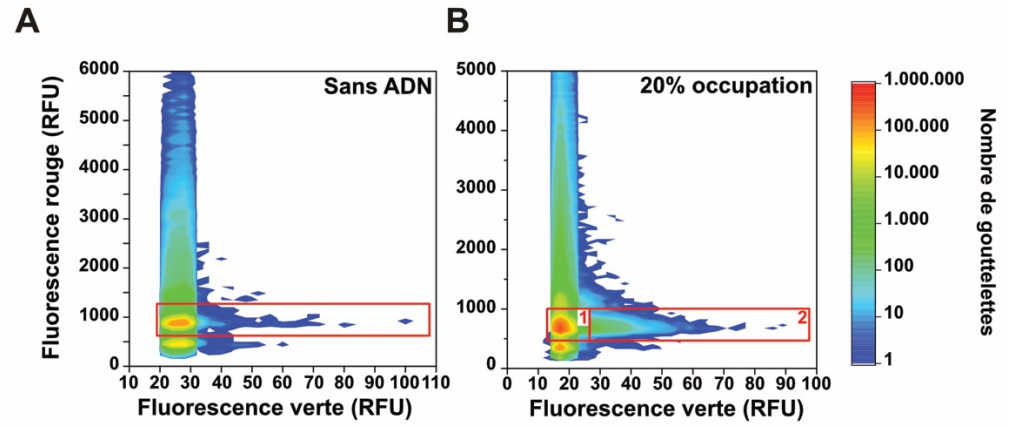


Figure 44 : Comparaison de profils microfluidiques obtenus après expression à l'aide d'extraits de S2 en absence et présence de matrice. En ordonné la fluorescence rouge illustre la taille des gouttelettes selon la concentration en Cyanine 5 et en abscisse la fluorescence verte correspond à la GFPmut2 traduite dans la gouttelette ou le bruit de fond issu des extraits de S2. **A.** Profil microfluidique obtenu en absence de matrice ADN avec de l'extrait de S2. L'encadré rouge met en avant la population de gouttelettes de TTIV (S2) fusionnées à une gouttelette de PCR ne contenant pas de matrice. **B.** Profil microfluidique obtenu avec la banque (T7/fragment/GFPmut2) utilisée et validée lors de tours de sélection avec les extraits de RRL du manuscrit présenté ci-dessus diluée de manière à occuper 20% des gouttelettes. L'encadré en rouge (1) correspond aux gouttelettes de TTIV fusionnées à une gouttelette de PCR vide ou contenant une matrice non efficace ou peu efficace pour permettre la traduction et l'encadré (2) correspond aux gouttelettes de TTIV fusionnées à une gouttelette de PCR contenant une matrice plus ou moins efficace pour permettre la traduction.

En effet, un nombre non négligeable (0,02%) de gouttelettes présentait un signal aspécifique (ne provenant pas de la production de GFPmut2 à la vue de l'absence de la matrice lors d'amplification par PCR du contenu des gouttelettes triées) probablement dû à l'extrait employé (Figure 44) les conduisant à leur sélection erronée. Malheureusement, ce signale est perçu comme un signal classique provenant d'une gouttelette et non pas tel un agrégat donnant lieu à un pic très important et peu représentatif d'une gouttelette (Figure 45).

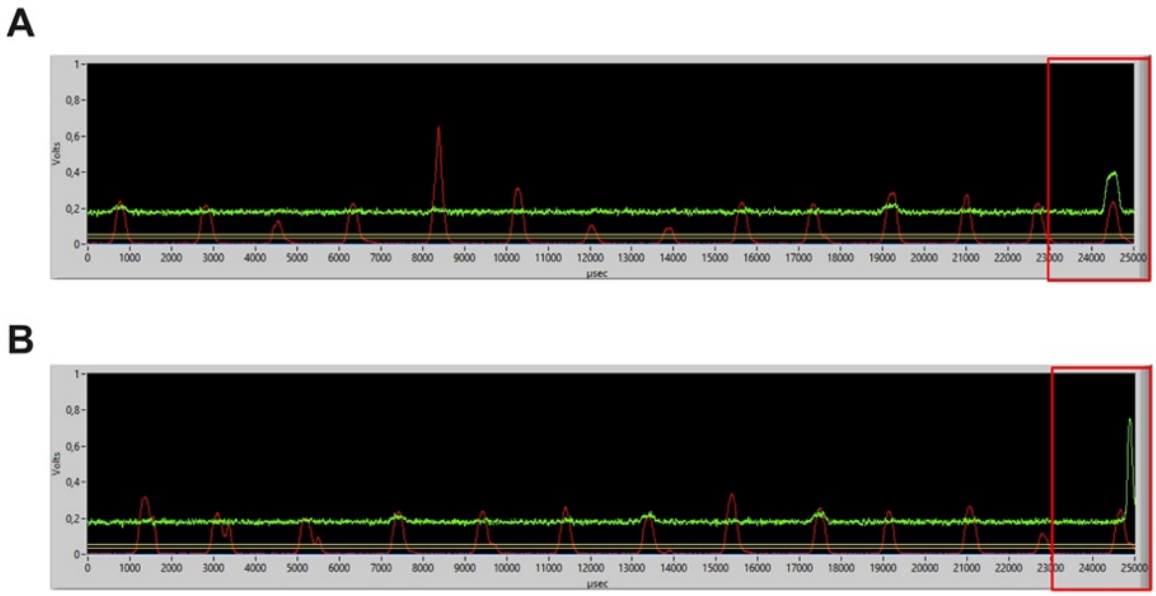


Figure 45 : Capture d'écran du profil des gouttelettes visualisées lors d'un tri sur le logiciel développé par l'équipe (Ryckelynck *et al.*, 2015). **A.** En ordonné la fluorescence est mesurée en volts avec la fluorescence rouge correspondant à la Cyanine 5 utilisée comme traceur de gouttelettes et la fluorescence verte émise par la GFPmut2 dans les gouttelettes l'ayant exprimées. Les gouttelettes observées sont issues d'un profilage après fusion avec du milieu de TTIV et expression d'une banque de variants (T7/fragments/GFPmut2). L'encadré rouge met en avant une gouttelette de TTIV fusionnée avec une gouttelette de PCR permettant la production de GFPmut2. **B.** Comme A, sauf que l'encadré rouge met en avant une gouttelette de TTIV fusionnée avec une gouttelette de PCR ne permettant pas la production de GFPmut2, le pic observé est probablement dû à un agrégat présent dans la gouttelette.

En plus de ce taux important de faux positifs, je me suis à nouveau heurtée à l'apparition et à la diffusion d'ADN parasites compromettant progressivement l'amplification des matrices d'intérêt. Je n'ai toutefois pas réussi à observer d'enrichissement notable des librairies en molécules d'intérêts contrairement ce qui avait été le cas avec les extraits RRL (voir manuscrit). Nous avons tout de même analysé le contenu de ces banques par NGS et cartographié les séquences sur le génome de référence (Figure 43 B). À l'inverse des sélections réalisées précédemment avec des extraits de RRL, aucun enrichissement notable de séquence n'a pu être observé sur une quelconque région du génome. Les difficultés rencontrées lors de ces sélections (signaux aspécifiques entrainant la sélection de faux positifs, molécules aspécifiques ré-amplifiées) rendent complexe l'analyse et l'utilisation de ce type d'extrait dans le processus microfluidique à l'heure actuelle. À l'avenir, il sera nécessaire d'améliorer la qualité ainsi que la stabilité des extraits employés afin de limiter la formation d'agrégats ou bien de modifier la protéine rapportrice employée afin que la longueur d'onde d'émission ne soit plus dans le vert. En effet, la sélection par une autre longueur d'onde d'émission de fluorescence permettra de s'affranchir des deux points limitants rencontrés

actuellement en l'occurrence le bruit de fond vert important de certains extraits (S2/HEK) et à l'inverse l'absorption de la fluorescence verte des extraits de RRL.

Cette stratégie expérimentale permettant la recherche de séquences initiatrices de la traduction est validée en extrait eucaryote RRL avec un génome viral. Comme mentionné dans le manuscrit en préparation (4.2.1.1), il sera intéressant de l'appliquer à d'autres génomes viraux comme ceux à l'origine de crises sanitaires mondiales, tel que le Zika ou le Sars-Cov 2. Notre stratégie expérimentale mettra en évidence la présence d'éléments de type IRES propres à ces virus, qui pourront constituer de nouvelles cibles thérapeutiques prometteuses pour neutraliser les infections. L'extension de ce concept aux génomes de mammifères pourrait également apporter de nouvelles pistes thérapeutiques en identifiant de nouvelles IRES, puisque certaines de ces structures ARN sont d'ores et déjà associées à des cas de développement de cancers (Sriram, Bohlen and Teleman, 2018a).

À court terme, la stratégie globale pourra aussi être appliquée à la recherche d'autres éléments de régulation tels que les riboswitches chez les procaryotes. La banque serait alors élaborée selon le nouveau protocole que je viens de valider puis, les sélections seront réalisées dans la même plateforme de criblage en présence et absence de ligand. L'intérêt étant d'identifier de potentiels riboswitches capables de moduler la transcription ou la traduction en réponse à des métabolites.

Discussions et perspectives

Le principal objectif de ma thèse a été le développement d'une plateforme de criblage universelle dans le but d'analyser et mieux comprendre les mécanismes d'initiation de la traduction que ce soit chez les procaryotes ou les eucaryotes. Durant ces trois années de thèse, j'ai pu m'intéresser à l'initiation de la traduction procaryote avec la ré-exploration du motif de liaison au ribosome (RBS), mais aussi approcher la régulation de l'expression des gènes médiée par les riboswitches (transcriptionnels), ou encore l'initiation de la traduction eucaryote en explorant d'une part les mécanismes coiffe-dépendants et d'autre part les indépendants.

1. Une plateforme d'analyse à ultrahaut-débit pour l'étude de l'initiation de la traduction procaryote

L'initiation de la traduction procaryote est, comme mentionnée tout au long de ces travaux de thèse, un point clef pour la régulation de l'expression des gènes. Cette étape limitante est finement régulée pour répondre de la manière la plus adaptée aux besoins de la cellule. Les premières caractérisations ont démarré dans les années 70 (J Shine and Dalgarno, 1974b) puis, sans interruption, les recherches se sont poursuivies dans le but d'explorer ce mécanisme d'initiation ainsi que les diverses possibilités de le réguler. Les différentes études ont alors analysé le RBS ainsi que son environnement (del Campo *et al.*, 2015; Hecht *et al.*, 2017; Cambray, Guimaraes and Arkin, 2018; Komarova *et al.*, 2020; Kuo *et al.*, 2020) les éléments régulateurs agissant en *cis* tels que les riboswitches (Husser, Dentz and Ryckelynck, 2021) ou en *trans* comme les petits ARN régulateurs (Dutta and Srivastava, 2018; Desgranges *et al.*, 2019; Chiaruttini and Guillier, 2020). Les travaux réalisés lors de cette thèse ont démontré la capacité d'observer des phénomènes de modulation de la traduction à ultrahaut-débit grâce à l'utilisation combinée de la microfluidique en gouttelettes, du séquençage à haut-débit et de la bio-informatique. Une avancée majeure qui ouvre de belles perspectives d'études des éléments de régulation complexes allant jusqu'à l'amélioration et au développement d'éléments synthétiques répondant, par exemple à de petites molécules.

1.1 Le Ribosome Binding Site 2.0

La séquence de la région RBS n'est bien évidemment pas le seul élément modulant l'efficacité d'initiation de la traduction. D'autres paramètres viennent s'ajouter, tels que la nature du codon initiateur, sa distance avec la SD ou encore la présence de structures secondaires, permettant d'appréhender l'hétérogénéité observée dans cette étape d'initiation de la traduction d'un gène à l'autre. L'impact de ces composantes du RBS a majoritairement été étudié grâce à des extraits préparés à partir de la souche modèle *E. coli* (del Campo *et al.*, 2015; Hecht *et al.*, 2017; Cambray, Guimaraes and Arkin, 2018; Komarova *et al.*, 2020; Kuo *et al.*, 2020). Une perspective attrayante serait de tester à nouveau l'ensemble de ces paramètres grâce à la plateforme de criblage que j'ai développée durant mon doctorat mais, cette fois-ci, à l'aide d'extraits plus atypiques et biologiquement plus pertinents. Ces extraits pourraient être préparés à partir d'autres bactéries à Gram négatif (par exemple du genre *Pseudomonas* également bien étudié et connu pour certains aspects pathogènes) ou à Gram positive (par exemple le genre *Bacillus* avec des applications médicales ou industrielles, ou encore les *Staphylococcus*, problème majeur de santé publique pour leur implication dans les infections nosocomiales). Enfin, une voie d'exploration pourrait cibler le cas de la protéine ribosomique S1 qui, chez *E. coli,* possède une fonction additionnelle particulière lui permettant d'arrimer le ribosome au niveau d'un codon initiateur afin de compenser la présence de structures secondaires ou d'ARN régulateurs en limitant son accessibilité (Boni *et al.*, 1990; Zheng *et al.*, 2011; Duval *et al.*, 2013; del Campo *et al.*, 2015). Il serait donc intéressant, dans le cadre de futures expériences, de comparer les variations de séquences suivant que l'extrait soit produit à partir d'une souche exprimant ou pas la protéine S1 (voire que celle-ci soit ajoutée ou non à un extrait) et ce en variant également les autres paramètres mentionnés plus haut comme par exemple les souches utilisées pour la préparation des extraits.

Un gain substantiel dans la précision des mesures pourrait être obtenu en ajoutant un aptamère fluorogène dans la région 3' UTR de nos constructions, ce qui permettrait de suivre indépendamment la transcription de la construction et la traduction de l'ARNm résultant. Cela permettrait de normaliser les processus entre eux, un paramètre important car une quantité trop importante ou trop faible d'ARN va impacter l'efficacité de la traduction par la suite (Hausser *et al.*, 2019). Par ailleurs, si l'on souhaite étendre nos travaux vers les éléments plus complexes tels que les riboswitches traductionnels, le contrôle des deux niveaux d'expression est indispensable afin de vérifier quelle(s) est (sont) la (ou les) étape(s) impactée(s) par l'action de l'ARN régulateur. De même, si à l'avenir l'idée d'utiliser des extraits obtenus à partir de cellules stressées se concrétise (e.g., conditions de cultures particulières ou infection phagique), il sera important de pouvoir différentier si l'expression est affectée dans son ensemble ou plus particulièrement à la transcription ou à la traduction. Pour ce faire, un large

éventail d'aptamères fluorogènes est disponible dans les différentes longueurs d'ondes du spectre visible (Zhou and Zhang, 2021). Durant cette thèse, j'ai pu réaliser des essais préliminaires de nos constructions « promoteur T7/GFPmut2 » associées à o-Coral que j'ai exprimé dans des extraits de S30 ou de l'extrait de PureExpress. Ces premiers essais n'ont malheureusement pas conduit à des résultats concluants à l'heure actuelle. Il est possible que le fluorogène Gemini-561 utilisé avec l'aptamère fluorogène o-Coral interagisse aussi avec des éléments issus de l'extrait employé, ne permettant pas la reproductibilité ou l'obtention d'un résultat cohérent. Ainsi, pour le moment, le choix du rapporteur de la transcription n'est pas défini. Il serait intéressant de pouvoir tester iSpinach et Mango-III (510 nm-534 nm) associé à une autre protéine fluorescente emttant à une autre longueur d'onde dans les conditions d'expression d'IVTT afin d'observer si les résultats sont cette fois plus reproductibles et pertinents.

Pour les protéines fluorescentes, l'avantage majeur de la GFPmut2 réside en son court temps de maturation (environ 5 minutes) et sa brillance élevée. Cependant, son émission dans le vert exclut son utilisation combinée avec des aptamères fluorogènes bien caractérisés tels qu'iSpinach et Mango-III. Pour pouvoir utiliser ces derniers, il faudrait recourir à des protéines fluorescant préférentiellement dans les longueurs d'ondes rouges telles que mCherry. Néanmoins, bien que cette dernière soit monomérique (la majorité des protéines rouges fluorescentes étant multimériques), celle-ci nécessite près de 20 min de maturation, réduisant de ce fait significativement la résolution temporelle de la méthode. Dans cet optique, les protéines activatrices de fluorogènes représentent une alternative très attractive. En effet, des systèmes tels que FAST (Plamont et al., 2016; Tebo et al., 2021) permettent l'obtention d'un complexe protéine/ligand fluorescent dont la longueur d'onde d'émission peut facilement être ajustée en changeant le ligand utilisé. De plus, ces systèmes permettent, comme c'est le cas avec les aptamères fluorogènes, une réponse quasi instantanée ; préservant de ce fait la résolution temporelle de la méthode. Des essais préliminaires ont d'ores et déjà été réalisés pour une émission de fluorescence dans le vert ainsi que dans le rouge à l'aide de différents extraits et semblent plutôt prometteurs. Ces expérimentations nécessiteront néanmoins d'être reproduites et optimisées dans le futur et la portabilité de la technologie FAST en gouttelettes microfluidiques reste à tester.

1.2 Les riboswitches, de la régulation à la détection de petites molécules

Les essais préliminaires de fonctionnalité des riboswitches transcriptionnels réalisés dans ces travaux de thèse ont été plutôt prometteurs notamment pour le riboswitch répondant à la FMN. Les résultats obtenus avec ce dernier montrent la possibilité de reconstituer le

mécanisme de régulation dans un système simplifié d'expression *in vitro*. Cependant, l'objectif réel de mon projet visait à exploiter des riboswitches traductionnels, approche pour le moment en suspend du fait de difficultés techniques rencontrées avec les doubles rapporteurs. Néanmoins, une fois cette difficulté technique dépassée, il suffira d'insérer la séquence du riboswitch souhaitée entre le promoteur et les rapporteurs en question pour générer des constructions, voire des banques de mutants. Ces dernières serront utilisables lors d'expériences de criblage, voire d'évolution dirigée, et permettront de répondre à des questions biologiques ciblées, d'améliorer les performances de riboswitches existants ou même de développer de nouveaux outils moléculaires.

Tout d'abord, l'évolution d'un riboswitch peut concerner uniquement la partie aptamère correspondant à la reconnaissance du ligand. Il est ainsi possible de reprogrammer une plateforme de régulation d'expression pour la placer sous le contrôle allostérique d'un ligand orthogonal comme un antibiotique (par exemple la néomycine (Etzel and Mörl, 2017; Günzel *et al.*, 2021) ou la tétracycline (Domin *et al.*, 2017) ou un ligand bio-orthogonal tel que la théophylline (Wachsmuth *et al.*, 2013, 2015). L'intérêt majeur ici est l'obtention d'un système d'expression génique contrôlable à la demande. A l'inverse, le module aptamère d'un riboswitch peut être évolué pour reconnaitre un autre ligand naturellement présent dans la cellule mais dont la concentration peut par exemple, en fonction des conditions de croissance ou de l'exposition de la cellule à un stress fluctuer (Hossain *et al.*, 2020; Piroozmand, Mohammadipanah and Faridbod, 2020; Husser, Dentz and Ryckelynck, 2021), un tel riboswitch devient alors un biosenseur codé génétiquement de la molécule cible. Bien que plus complexes et moins directs que les biosenseurs ARN fluorogènes (FRBs), ces biosenseurs dérivés de riboswitches offrent plusieurs avantages. Tout d'abord, ils peuvent contrôler l'expression d'un gène rapporteur codant pour une enzyme dont l'activité agit comme un amplificateur de signal permettant ainsi d'augmenter la sensibilité de détection théorique du système. D'autre part, de tels systèmes seraient intégralement génétiquement encodables (si le gène rapporteur est une protéine fluorescente par exemple), s'affranchissant ainsi de la nécessité d'ajouter un cofacteur (substrat d'enzyme ou fluorogène), rendant ces systèmes plus pertinents d'une utilisation avec des organismes complexes tels que des animaux ou dans des conditions dans lesquelles l'accès du cofacteur à la cellule ne peut être garanti (développement de bactéries en colonies ou biofilm par exemple).

D'autre part, l'évolution du riboswitch peut cibler la molécule dans son ensemble. L'objectif cette fois-ci serait d'améliorer le différentiel de réponse entre un état OFF (absence d'expression) et un état ON (expression active) afin d'obtenir un système d'expression sur demande optimisé, avec une expression basale minimum (voire nulle) mais maximale après ajout du ligand activateur. Ce type de molécules serait très intéressant pour le contrôle de l'expression de gènes toxiques (gène de sécurité en thérapie génique par exemple).

2. Une plateforme d'analyse à ultrahaut-débit de l'initiation de la traduction eucaryote

L'initiation de la traduction eucaryote est, comme chez les procaryotes, une étape clef et finement régulée de l'expression des gènes. Ainsi, la bonne compréhension des mécanismes impliqués est essentielle pour mieux appréhender les différents mécanismes de régulation de l'expression, l'adaptation de la cellule à son environnement et les conséquences possibles de leurs disfonctionnements. De nombreux travaux ont déjà traité du sujet en explorant la réponse cellulaire face à son environnement en adaptant l'expression de ses gènes pour répondre à ses besoins (Madlung and Comai, 2004). Parmi les points abordés au cours de cette thèse, je me suis en particulier intéressée à un exemple précis dans le contexte de chacun des deux grands modes d'initiation de la traduction eucaryote que sont l'initiation coiffe-dépendante et coiffe-indépendante (Sonenberg and Hinnebusch, 2009; Komar and Hatzoglou, 2011). D'une part, l'initiation coiffe-dépendante nécessite un grand nombre de facteurs et fait intervenir un scan de la région 5' non traduite de l'ARNm. La séquence environnant le codon initiateur a été prédite comme fonctionnellement importante pour la reconnaissance du bon codon de démarrage, bien que la nature exacte de cette séquence, dite de Kozak (Kozak, 1981), reste encore à explorer, faisant ainsi l'objet d'un premier sous-projet de ma thèse. D'autre part, l'initiation coiffe-indépendante représente un mode alternatif d'initiation et est impliqué dans une variété de cas médicalement pertinents, allant des infections virales au développement de cancers. Ces mécanismes font intervenir des structures d'ARN spécifiques dont la mise en évidence peut s'avérer parfois complexe, ce qui m'a motivé à mettre en place une technologie dédiée à leur étude dans le cadre de ma dernière partie de travail de thèse.

2.1 La séquence de Kozak 2.0

Les résultats préliminaires obtenus lors de la ré-exploration de la séquence de Kozak semblent prometteurs bien que, d'une manière générale, la procédure ait encore besoin de modifications et améliorations avant de pouvoir collecter des résultats potentiellement pertinents sur le plan biologique. A noter qu'à l'heure où ce manuscrit est rédigé, les analyses bio-informatiques sont encore en cours d'étude chez nos collaborateurs.

Pour être plus fidèle à la situation cellulaire, les ARNm utilisés auraient dû être coiffés, impliquant un lourd développement technologique qui a pu provisoirement être évité grâce à

l'idée ingénieuse de nos collaborateurs d'utiliser l'IRES de l'EMCV pour charger le ribosome à l'extrémité de l'ARNm de façon coiffe-indépendante. Cependant, une fois chargé, le ribosome entre en mode de scan pour rechercher le codon initiateur, une approche appropriée pour l'étude de la séquence de Kozak. Néanmoins les limites de ce système seront rapidement atteintes lorsque la coiffe et sa reconnaissance sont directement actrices de l'initiation de la traduction. Caractériser de tels mécanismes par notre approche de microfluidique en gouttelettes nécessiterait une lourde modification de processus microfluidique qui est à l'étude en vue d'applications futures (incluant le découplage de la transcription et la traduction, l'ajout du mélange permettant le coiffage par pico-injection et la division des gouttelettes afin de rester compatibles avec nos modules microfluidique) (3.1.3). Ainsi, une fois en place, cette nouvelle stratégie démultipliera les capacités d'analyses de la plateforme avec la possibilité de combiner un certain nombre de paramètres différents, allant du codon initiateur impliqué, à la région environnante. Il est d'ailleurs possible d'inclure ou non des structures secondaires afin d'observer l'impact sur la séquence environnante étudiée, la conservation de motif ou non selon ces éléments, le tout dans le contexte biologiquement pertinent d'un ARNm coiffé. Enfin, il sera aussi envisageable d'utiliser différentes lignées cellulaires et conditions de cultures (application de stress) pour la préparation d'extraits qui permettront d'adresser d'autant plus de nouvelles questions biologiques dans un contexte certes simplifié (les extraits), mais contrôlable et à ultrahaut-débit.

2.2 Recherche fonctionnelle de séquences initiatrices de la traduction

J'ai pu mettre en place et valider une méthode innovante de préparation de banques d'ADN génomique viral peu biaisé. Le criblage fonctionnel de cette banque m'a alors permis d'identifier, à partir du génome viral du CrPV, l'IRES IGR en bornant précisément son démarrage et sa fin de façon cohérente avec les connaissances de la littérature (Wilson *et al.*, 2000) et en réussissant même à l'orienter grâce à l'identification de sa phase de lecture (voir 4.2.1.1, manuscrit en préparation #2).

La robustesse de ce protocole étant à présent établie, il sera possible de l'appliquer à de nouveaux virus moins bien caractérisés ainsi qu'à d'autres génomes. En effet, cette approche expérimentale pourra dans un premier temps être appliquée à des virus d'intérêt sanitaire et de santé publique tels que le Zika ou le Sars-Cov2. Les IRES étant habituellement des structures idiosyncrasiques, celles-ci représentent des cibles thérapeutiques de choix

pour le développement de composés antiviraux spécifiques présentant potentiellement peu de réactions croisées avec l'hôte, donc peu d'effets secondaires.

D'autres part, les génomes de mammifères (dont notamment les cellules humaines) peuvent également renfermer des IRES (Lozano, Francisco-Velilla and Martinez-Salas, 2018; Yang *et al.*, 2021) qui restent complexes à identifier et à caractériser. En effet, les IRES sont principalement prédites par des approches *in sillico* complétées par de longues validations expérimentales. A ce jour, des banques de données rassemblent toutes sortes d'informations dont de nombreuses prédictions *in sillico* dans le but de révéler des éléments potentiellement intéressants par la mise en commun de ces données (Bonnal *et al.*, 2003; Mokrejs *et al.*, 2006; Yang *et al.*, 2021). Néanmoins, comme cela est le cas pour les riboswitches, ces prédictions se doivent d'être confirmées expérimentalement. Dans ce contexte, la nouvelle approche expérimentale développée et validée au cours de ce travail de thèse est adéquate. Un certain nombre d'IRES présentes dans le génome humain a été associé à des anomalies d'expression de gènes pouvant mener au développement de cancers comme pour le cancer inflammatoire du sein (Silvera and Schneider, 2009; Komar and Hatzoglou, 2011). Dans ce cas précis, des facteurs connus pour limiter la traduction coiffe-dépendant sont nettement augmentés (eIF4G1 et 4E-BP-1) induisant alors le switch de mécanisme d'initiation pour ceux menés par des IRES. De ce fait, l'expression de gènes impliqués dans l'angiogenèse facilite alors le développement et la progression du cancer (Silvera and Schneider, 2009; Sriram, Bohlen and Teleman, 2018a).

Allant au-delà de la découverte de nouvelles IRES, cette technologie pourra être appliquée à la découverte d'autres séquences impliquées dans la régulation de la traduction (voire transcription) en *cis*, y compris chez les procaryotes et archées. Cela est notamment le cas des riboswitches déjà discutés plus haut. Ces éléments présentent également une organisation tridimensionnelle complexe conservée d'un organisme à l'autre. Ainsi, une classe de riboswitches très répandue (répondant par exemple au TPP, à la FMN ou à la SAM) dans la phylogénie bactérienne sera aisément identifiable par génomique comparative du fait de l'important échantillonnage disponible. En revanche, les choses sont plus complexes pour les éléments plus rares qui, même lorsqu'ils sont identifiés restent souvent orphelins de leur ligand naturel (Sherlock and Breaker, 2020). De même, un repliement commun ne signifie pas une spécificité identique, rendant d'autant plus complexe l'identification du ligand (Battaglia and Ke, 2018; Sherlock, Sudarsan and Breaker, 2018). L'approche de criblage que j'ai développé durant cette thèse permettrait alors d'approcher le problème de l'identification des riboswitches et de leur ligand de façon inverse. En effet, le protocole de préparation des banques est facilement extrapolable à un génome bactérien. Il suffirait ensuite de réaliser des cycles de sélections positives et négatives, respectivement en présence et en absence d'un ligand cible,

et d'isoler les constructions montrant une modulation de la transcription et/ou de la traduction en réponse à la présence du ligand. Cette approche est fonctionnellement similaire à la stratégie utilisée en routine par l'équipe et qui a fait ses preuves pour le développement de biosenseurs ARN fluorogènes (Autour *et al.*, 2019). Cette approche serait utilisable pour identifier des molécules répondant à des ligands/métabolites pour lesquels une forte suspicion de l'existence d'un riboswitch existe, comme cela a longtemps été le cas par exemple pour l'alarmone ppGpp pour laquelle il a fallu plus de 10 ans pour identifier le riboswitch (Sherlock, Sudarsan and Breaker, 2018).

3. La microfluidique et l'étude de l'expression des gènes

Au travers de ce travail de thèse, j'ai pu mettre en place et commencer à exploiter des stratégies d'identification et de caractérisation d'éléments ARN régulateurs de l'expression des gènes à très hauts débits. Cette capacité d'analyse est rendue possible par la technologie microfluidique que ce soit avec les gouttelettes pour les criblages fonctionnels ou même les puces de séquençage à haut-débit des plateformes Illumina.

La microfluidique a donc permis d'étudier à ultrahaut-débit la traduction bactérienne offrant la possibilité de compléter les connaissances sur la souche à Gram négative modèle *E. coli* et dans un futur proche d'autres souches non modèles, notamment les bactéries à Gram positif. Une des perspectives directe étant l'identification de cibles thérapeutiques, voire la découverte de molécules pouvant interagir et moduler ces mécanismes.

Le processus microfluidique développé au cours de cette thèse aura aussi permis d'ouvrir une nouvelle voie d'étude des mécanismes eucaryotes. Par exemple, la séquence de Kozak encore peu définie à ce jour a pu être ré-explorée grâce à cette nouvelle approche expérimentale. Pour finir, cette nouvelle plateforme de criblage va permettre la découverte de nouveaux éléments régulateurs (par exemple des IRES ou tout autres éléments modulant la traduction) issus du génome humain ou de génomes de pathogènes, offrant d'autant plus de nouvelles cibles thérapeutiques potentielles et le développement de nouveaux médicaments.

L'ensemble de mes travaux ont donc permis de mettre en avant les capacités de la microfluidique en gouttelettes pour l'étude de l'expression des gènes et ce, en ouvrant la voie à des axes variés d'études aussi bien chez les procaryotes que chez les eucaryotes ; allant de la génération de nouvelles connaissances à l'ingénierie moléculaire en passant par la recherche de nouvelles cibles thérapeutiques.

Annexe I

1. Matériels et Méthodes – (complément hors manuscrits)

1.1 Élaboration des matrices

1.1.1 Création des matrices riboswitches

Riboswitches synthétiques associés à Mango-III

Les riboswitches répondant à la théophylline (teo), tétracycline (tet) et néomycine (neo) sont associés à l'aptamère fluorogène Mango-III et au promoteur T7A1 par PCR. Le mélange réactionnel de PCR contient 15 pmol de chaque oligomère (Tableau 2, avec en oligo sens : T7A1_Neo, Theo ou Tet et en oligos anti-sens Mango-III_Neo, Theo ou Tet), 1 ng de matrice envoyée par l'équipe de Mario Möerl (Tableau 2, Riboswitches synthétiques/ Matrice), 0.2 mM de chaque désoxynucléotides (Thermo Fischer) et 4 U de polymérase Q5 haute-fidélité associée

au tampon correspondant (NEB). L'amplification démarre par une étape initiale de dénaturation de 5 min à 98 °C suivi de 25 cycles de : 15 sec à 98 °C, 20 sec à 60 °C and 1 min 45 à 72 °C, puis 2 min à 72 °C. Par la suite, le produit de la PCR est purifié à l'aide de billes Sera Mag (Sera-Mag™) selon le protocole du fournisseur (Sigma aldrich) avec un ratio de 2,5.

Riboswitches synthétiques associés à o-Coral

Les riboswitches répondant à la théophylline, tétracycline et néomycine sont associés à l'aptamère fluorogène o-Coral et au promoteur T7A1 grâce à la méthode de clonage de Gibson. Ainsi chaque élément est premièrement amplifié de manière indépendante par PCR dans un mix contenant 15 pmol de chaque oligomère (Tableau 2 : Riboswitches synthétiques/ Préparation des matrices Gibson), 1 ng de matrice (Tableau 2), 0.2 mM de chaque désoxynucléotides (Thermo Fischer), 4 U de polymérase Q5 haute-fidélité associés au tampon correspondant (NEB). L'amplification démarre par une étape initiale de dénaturation de 5 min à 98 °C suivi de 25 cycles de : 15 sec à 98 °C, 20 sec à 60 °C and 1 min 45 à 72 °C, puis 2 min à 72 °C. Les produits de PCR (promoteur/ riboswitch/ o-Coral) sont ensuite mélangés tous

ensemble dans un volume final de 20 µL avec du mélange réactionnel du kit NEBuilder HiFi DNA Assembly (NEB) 1 X pendant 15 min à 50 °C. Puis, l'ensemble est transformé selon le protocole du fournisseur NEB dans des bactéries DH5α chimio-compétentes. Après clonage, les colonies sont ensuite piquées et inoculées dans du milieu 2YT (16 g/L de tryptone, 10 g/L d'extrait de levure, 5 g/L de Chlorure de sodium) à 37 °C jusqu'à saturation. Les plasmides sont extraits à l'aide des kits GenJet Plasmid Miniprep kit (Thermo-Scientific) puis les séquences sont vérifiées après séquençage Sanger (GATC Biotech). Enfin, les matrices d'intérêts sont ré-amplifiées par PCR dans les mêmes conditions que mentionnées plus haut avec leur oligomères respectifs (Tableau 2, avec en oligo sens : T7A1_Neo, Theo ou Tet et en anti-sens o-Coral_Neo ou Theo ou Tet).

Riboswitch FMN associé à o-Coral

Pour le riboswitch répondant à la FMN, le promoteur, le riboswitch ainsi que l'aptamère fluorogène o-Coral sont associés cette fois-ci par digestion/ligation. Le site de restriction de BamHI est ajouté en 3' du promoteur T7A1 et en 5' du riboswitch répondant à la FMN. Tandis que le site de restriction HindIII est ajouté en 3' du riboswitch répondant à la FMN et en 5' de l'aptamère fluorogène o-Coral. Ces ajouts sont réalisés par PCR réalisée dans les mêmes conditions que ci-dessus (riboswitches synthétiques) avec leur oligomères respectifs (Tableau 2 : Riboswitch FMN). Chaque produit de PCR est ensuite mis dans un volume final de 50 µL avec 1 X de tampon cutsmart (NEB), 20 U de HindIII (NEB) et/ou 20 U de BamHI (NEB), 1 µg de matrice (Tableau 2, T7A1/ riboswitch FMN/ o-Coral) 1 h à 37 °C puis chacun est purifié par extraction au phénol/chloroforme. Chaque produit de digestion est utilisé pour une même ligation de 50 µL avec 1 X de tampon de T4 DNA ligase, 20 µL de PCR digérées, et 400 U de T4 DNA ligase pendant 16 h à 16 °C puis 20 min à 65 °C et enfin d'une purification au phénol/chloroforme. Pour terminer, 1 µL de ligation purifiée est utilisée pour une amplification par PCR dans les mêmes conditions que mentionné ci-dessus avec l'oligo sens T7A1_FMN et anti-sens Rv_coral (Tableau 2).

	Sens	Nom	Séquence (5'-3')	Rôle
Projet Kozak	Sens	Puc_EMCV	GTCGACCTGCAGGCATGCACTAACGTTACTGGCCGAAGCCGCTTGG	Amplification de l'IRES EMCV, Ajout de séquence aléatoire autour du codon ATG et de l'extémité 5' eGFP
	Anti-sens	EMCV_N6_ATG_eGFP	CAGCTCCTCGCCCTTGCTCACTTGTTGTTGTTGTTGTTGTTC(N1:25252525)(N1)(N1)CAT(N1)(N1)(N1)(N1)(N1)(N1)CCAACTAGTTGGTAGTTG	
	Sens	T7_EMCV	CAACAAATATTAATACGACTCACTATAGGTAACGTTACTGGCCGAAGCCGCTTG	Amplification de la banque
	Anti-sens	eGFP_Tagmt2_illu1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTAATTTAAATCTTCTTCTGATAATAATTTTTGTTCTAATGC	
Riboswitch (synthétique)	Matrice	Mango-III	TACGAAGGAAGGATTGGTATGTGGTATATTCGTA	
	Matrice	R.Néomycine	AGCGGCCTAAAACATACCAGAGAAATCTGGAGAGGTGAAGAATACGACCACCTAGGCCGACAGTGGCCTAGGTGGTCGTTTTTTTTTACCCGTTTTTTGGGCTAACAGGA	-
	Matrice	R.Tétracycline	ACGAGAGGGACACGGGGAAACACCACCATAAGTGATACCAGCATCGTCTTGATGCCCTTGGCAGCACTTCAGAAATCTCTGAAGTGCTGTTTTTTTTTAGGA	-
	Matrice	R.Théophylline	AGCAAGTGATACCAGCATCGTCTTGATGCCCTTGGCAGCACTTCAGAAATCTCTGAAGTGCTGTTTTTTTTTACCCGTTTTTTGGGCTAACAGGA	-
	Sens	T7A1_Neo	CAGATTACGAATTCGGTTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCGATCCAGCGGCCTAAAACATACC	Ajout du promoteur T7A1
	Anti-sens	Mango-III_Neo	TACGAATATACCACATACCAATCCTTCCTTCGTATCCTGTTAGCCCAAAAAACGGG	Ajout de MangoIII
	Sens	T7A1_Tet	CAGATTACGAATTCGGTTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCGATCCACGAGAGGGACACGGG	Ajout du promoteur T7A1
	Anti-sens	Mango-III_Tet	TACGAATATACCACATACCAATCCTTCCTTCGTATCCTAAAAAAAACAGCACTTCAGAGATTTCTGAAG	Ajout de MangoIII
	Sens	T7A1_Theo	CATTACGAATTCGGTTATCAAAAAGAGTATTGACTTAAAGTCTAACCTATAGGATACTTACAGCCGATCCAGCAAGTGATACCAGCATCG	Ajout du promoteur T7A1
	Reverse	Mango-III_Theo	TACGAATATACCACATACCAATCCTTCCTTCGTATCCTGTTAGCCCAAAAAACGGG	Ajout de MangoIII
	Sens	T7_Theo	CGTCCACACGTCCACATAATACGACTCACTATAGGAGCAAGTGATACCAGCATCGTCTTG	Ajout du promoteur T7
	Sens	Tet_o-Coral+F30	CTGAAGTGCTGTTTTTTTTAGGACGTCCACAAAGCTTTTGCCATGTG	Préparation des matrices du Gibson
	Anti-sens	o-Coral+F30_Rtet	CACATGGCAAAAGCTTTGTGGACGTCCTAAAAAAAACAGCACTTCAGAGATTTCTG	
	Sens	T7A1_Rtet	GTTTCCCCGTGTCCCTCTCGTGGCTGTAAGTATCCTATAGGTTAGACTTTAAGTC	
	Sens	Neo/Theo_o-Coral+F30	GTTTTTTGGGCTAACAGGACGTCCACAAAGCTTTTGCCATGTG	
	Anti-sens	o-Coral+F30_RNeo/Theo	CACATGGCAAAAGCTTTGTGGACGTCCTGTTAGCCCAAAAAACGG	
	Sens	T7A1_RNeo/Theo	GACGATGCTGGTATCACTTGCTGGCTGTAAGTATCCTATAGGTTAGACTTTAAGTC	
	Anti-sens	T7A1_puc	GGTTAGACTTTAAGTCAATACTCTTTTTGATAACCCCCGGGTACCGAGCTCG	
	Sens	o-Coral+F30_puc	CAATCTAGAGACACGAGCACAGTGTACGATCCTCTAGAGTCGACCTG	

Riboswitch FMN	Matrice	R.FMN	GGATCCCAGATTACCGTCCACAGTA TAAGGACAAATGAATAAAGATTGTAT CCTTCGGGGCAGGGTGGAAATCCC GACCGGCGGTAGTAAAGCACATTTG CTTTAGAGCCCGTGACCCGTGTGCA TAAGCACGCGGTGGATTCAGTTTAA GCTGAAGCCGACAGTGAAAGTCTGG ATGGGAGAAGGATGATGAGCCGCTA TGCAAAATGTTTAAAAATGCATAGTG TTATTTCCTATTGCGTAAAATACCTA AAGCCCCGAATTTTTTATAAATTCGG GGCTTTTTTGACGGTAAATAACAAAA GAGGGGAGGGAAACAATACTGTGGA CGAAGCTT	-
	Matrice	o-Coral + F30	CGTCCACAAAGCTTTTGCCATGTGTA TGTGGGCCTGCAGGAGAACCCCGCT TCGGCGGTGATGGAGAGGCGCAAG GTTAACCGCCTCAGGTTCCGGTGAC GGGGCCTCGCTTCGGCGATGATGG AGAGGCGCAAGGTTAACCGCCTCAG GTTCTCCTGCAGGCCCACATACTCT GATGATCCTTCGGGATCATTCATGG CAATCTAGAGACACGAGCACAGTGT AC	-
	Sens	T7A1	CAGATTACGAATTCGGTTATCAAAAA GAGTATTGACTTAAAGTC	Préparatin des matrices pour la digestion et ligation
	Anti-sens	BamHI_ T7A1	GTAATCTGGGATCCGGCTGTAAGTA TCCTATAGGTTAGAC	
	Sens	BamHI_FMN	GGATCCCAGATTACCGTCCACAGTA TAAGGACAAATGAATAAAGATTGTAT CCTTCG	
	Anti-sens	HindIII_FMN	CACGACTGAAGCTTCGTCCACAGTA TTGTTTCC	
	Sens	HindIII_o-Coral+F30	CGTCCACAAAGCTTTTGCCATGTGTA TGTGGGCCTGC	
	Anti-sens	Rv_o-Coral+F30	GTACACTGTGCTCGTGTCTCTAGAT TGCCATGAATGATCCCGAAGGAT	
Projet IRES	Sens	Fwd T7Alexa488	GTCCACACGTCCACATAATACGACTC ACTATAGGAGACCACAACGGTTTCC CTCCTTTA	Révelation sur gel
	Sens	T7 sens	GTCCACACGTCCACATAATACGACTC ACTATAGGAGACCACAACGGTTTCC CTCCTTTAT	Ligation étape 1
	Anti-sens	T7 complementaire	AAAGGAGGGAAACCGTTGTGGTCTC CTATAGTGAGTCGTATTATGTGGAC GTGTGGACG	
	Sens	gfpmut2 sens	(P)AAGGTGAGTAAAGGAGAAGAACT TTTCACTGG	
	Anti-sens	gfpmut2 anti sens	CCAGTGAAAAGTTCTTCTCCTTTACT CACCTTT	
	Sens	Fwd gfpmut2-30pb	(P)AAGGTGAGTAAAGGAGAAGAACT TTTCACTGG	Amplification de GFPmut - 30pb
	Sens	Fwd T7 couple 1	GTCCACACGTCCACATAATACGACTC ACTATAGG	Vérification par QPCR
	Anti-sens	Rev gfpmut2 couple 1	CCAGTGAAAAGTTCTTCTCCTTTACT CACCTTT	
	Sens	Fwd IRESmodele couple 2	GAGCATTGCGGCTGATAAGGTTTTA G	
	Anti-sens	Rev IRESmodele couple 2	TCTTGAAATGTAGCAGGTAAATTTCT TAGGTTTTTC	
	Anti-sens	Rev gfpmut2 couple 3	TTAATTTAAATCTTCTTCTGATAATAA TTTTTGTTCTAATGC	

<u>Tableau 2 :</u> Récapitulatif des oligomères et matrices utilisés et non publiés.

1.1.2 Vérification des matrices riboswitches

Les matrices obtenues pour les riboswitches répondant à la théophylline, tétracycline et néomycine associés à MangoIII ainsi que le riboswitch répondant à la FMN ont été dilués dans un mélange réactionnelle contenant 0,2 mM de chaque dNTPs (Thermo Fisher), 2,5 U de DreamTaq^{TM} dans le tampon correspondant (NEB), chauffé à 98°C pendant 5 min puis à 72 °C pendant 20 min afin d'ajouter des adénines en 3' de nos matrices. Par la suite, ces matrices adénylées sont insérées dans le plasmide pTZ57R/T en suivant les instructions du fournisseur du kit de clonage TA (InsTAclone PCR cloning Kit, Thermo-Scientific). Le produit de ligation est purifié et récupéré par extraction au phénol/Chloroforme, puis 100 ng d'ADN sont utilisés pour être transformés dans des bactéries électro-compétentes, Electro-10 blue (Agilent), à l'aide de cuves d'éléctroporations de 2 mm (MicroPulser, Bio-Rad). Après électroporation, les bactéries sont placées une heure en culture à 37°C sous agitation, puis sont ensuite étalées sur boite 2YT et ampicilline et incubées 16 h à 37 °C.

Les matrices obtenues pour les riboswitches répondant à la théophylline, tétracycline et néomycine associés à o-Coral ont déjà été transformées lors de leur préparation à l'aide du clonage Gibson (cf titre).

Les colonies choisies pour chaque construction sont piquées et inoculées dans du milieu 2YT à 37 °C jusqu'à saturation et les plasmides sont extraits à l'aide du kit d'extraction de plasmide GenJet Plasmid Miniprep kit (Thermo-Scientific). Enfin, les séquences sont vérifiées après séquençage Sanger (GATC Biotech).

1.1.3 Création de la banque Kozak (protocole de nos collaborateurs)

Premièrement la protéine rapportrice utilisée, l'eGFP est récupérée par digestion d'un plasmide comportant cette dernière :

Plasmide contenant l'eGFP (5' - 3') (eGFP représenté en vert)

TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTG
TAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGG
CTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGC
GTAAGGAGAAAATACCGCATCAGGCGCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGT
GCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGC
CAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTCGAGCTCGGTACCCGGGGATCCTCT
AGAGTCGACCTGCAGGCATGCACCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCC
TGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCAC
CTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGA

CCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAG
TCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCG
CGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAG
GACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAA
GCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCC
GACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCA
CCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCC
GCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTAAAGAGAATTCAGAGCTCGGATCCACTCGAGATG
CATTAGAACAAAAATTATTATCAGAAGAAGATTTAAATTAATTGGCGTAATCATGGTCATAGCTGTTTCCTGTG
TGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCC
TAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGC
CAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTC
GCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATAC
GGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAAC
CGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACG
CTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCG
TGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCG
CTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCAC
GAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACA
CGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAG
AGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGCTCTGCTGAAGC
CAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTT
TTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTC
TGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAG
ATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAA
TGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCG
TGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCGCGAGACCCACGC
TCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAAC
TTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTG
CGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCC
GGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCT
CCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTA
CTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTAT
GCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAG
TGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGAT
GTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACA
GGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTT
CAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAA
CAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGACAT
TAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTC

La digestion se fait dans un volume final de 20 µL à partir de 1 ng de plasmide, du tampon Tango 1 X (Thermo Fischer) et 10 U de l'enzyme de restriction NcoI (Thermo Fischer) pendant 2 h à 30 °C puis 20 min à 65 °C.

En parallèle, 2 ng de plasmide comportant l'IRES du génome de l'EMCV (séquence du plasmide ci-dessous) sont utilisés comme matrice de PCR pour amplifier cette dernière (IRES EMCV).

Plasmide contenant l'EMCV (5'-3') (l'IRES de EMCV repésenté en bleu)

TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTG
TAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGTTGGCGGGTGTCGGGGCTGG
CTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGC
GTAAGGAGAAAATACCGCATCAGGCGCCATTCGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGT
GCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGC
CAGGGTTTTCCCAGTCACGACGTTGTAAAACGACGGCCAGTGAATTCGAGCTCGGTACCCGGGGATCCTCT
AGAGTCGACCTGCAGGCATGCACTAACGTTACTGGCCGAAGCCGCTTGGAATAAGGCCGGTGTGCGTTTGT
CTATATGTTATTTTCCACCATATTGCCGTCTTTTGGCAATGTGAGGGCCCGGAAACCTGGCCCTGTCTTCTTG
ACGAGCATTCCTAGGGGTCTTTCCCCTCTCGCCAAAGGAATGCAAGGTCTGTTGAATGTCGTGAAGGAAGCA
GTTCCTCTGGAAGCTTCTTGAAGACAAACAACGTCTGTAGCGACCCTTTGCAGGCAGCGGAACCCCCCACCT
GGCGACAGGTGCCTCTGCGGCCAAAAGCCACGTGTATAAGATACACCTGCAAAGGCGGCACAACCCCAGTG
CCACGTTGTGAGTTGGATAGTTGTGGAAAGAGTCAAATGGCTCTCCTCAAGCGTATTCAACAAGGGGCTGAA
GGATGCCCAGAAGGTACCCCATTGTATGGGATCTGATCTGGGGCCTCGGTGCACATGCTTTACATGTGTTTA
GTCGAGGTTAAAAAACGTCTAGGCCCCCCGAACCACGGGGACGTGGTTTTCCTTTGAAAAACACGACCATAA
TCAACAACAACAACAACAACAACAACAACAACAACAACAACGACACTCAGTCCTATTACTCGAGACACTAGACACA
CAGTCGCAACAACAACAACAACAACAACAACAACAACAACAACTACCAACTAGTTGGCGTAATCATGGTCATA
GCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAA
GCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGA
AACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTC
TTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAA
AGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAA
AAGGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCA
CAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTG
GAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGG
GAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGG
GCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAAC
CCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGG
CGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGTATCTGCGC
TCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAG
CGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTT
TCTACGGGGTCTGACGCTCAGTGGAACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGG
ATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCT
GACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCT
GGCTCCCCGTTGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATGATACCG
CGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAA
GTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGC
CAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGG
CTTCATTCAGCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTA
GCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCAC

TGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTC
TGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAG
CAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTG
AGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTG
GGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTC
ATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGT
ATTTAGAAAAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCA
TTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTC

En 3' de la matrice EMCV IRES une portion de séquence est dégénérée par PCR, soit les 6 nucléotides en amont du futur codon initiateur et trois nucléotides en aval. Le mix de PCR pour générer cette modification est d'un volume de 50 µL avec 0,25 mM de chaque dNTPs, du tampon haute-fidélité de la Phusion (Thermo Fischer) 1 X, 0,04 µM d'oligomère (Tableau 2, Projet Kozak) et 1% de Phusion produite et purifiée par leur soin. L'amplification PCR démarre par 1 min à 95 °C, puis 15 cycles répétant 15 sec à 95 °C, 15 sec à 60 °C et 30 sec à 72 °C puis se termine par 5 min à 72 °C.

Par la suite, on a ajouté en 3' de l'IRES EMC la séquence codante de l'eGFP. Pour cela, 3 µL du produit de PCR de l'IRES EMCV est ajouté à 1 µL de plasmide comportant l'eGFP digérée par NcoI avec le mélange réactionnel du kit NEBuilder HiFi DNA Assembly (NEB) 1 X dans un volume final de 20 µL pendant 15 min à 50 °C. Pour finir, la banque est obtenue par amplification PCR de 50 µL à partir de 1 µL de mélange réactionnel Hifi réalisé ci-dessus, avec 0,25 mM de chaque dNTPs, du tampon haute-fidélité de la Phusion (Thermo) 1 X, 0,04 µM d'oligomère (Tableau 2, Projet Kozak) et 1 % de Phusion produite et purifiée par leur soin. Le programme d'amplification démarre par 1 min à 95 °C puis 20 cycles de 15 sec à 95 °C, 15 sec à 60 °C et 1 min à 72 °C, et se termine par 5 min à 72 °C.

1.1.4 Stratégies testées pour l'élaboration de la banque IRES

Les ligations non optimisées sont réalisées avec 0,3 pmol de chaque molécule (Promoteur T7, IRES modèle et GFP mut2), avec 0,01 mM d'ATP (Larova), 1 X de tampon de la T4 DNA ligase (NEB) et 400 U de T4 DNA ligase dans un volume final de 50 µL pendant 16 h à 16°C puis 20 min à 65°C pour stopper la réaction. La ligation est ensuite purifiée au phénol/chloroforme puis à l'aide de billes Sera Mag (Sera-Mag™) selon le protocole du fournisseur (Sigma aldrich) avec un ratio de 2,5. Afin d'évaluer la qualité et le rendement du de la ligation 1 ng du produit de cette ligation purifiée est utilisé comme matrice de QPCR avec un mélange réactionnel contenant 15 pmol de chaque oligomère (Tableau 2, Stratégie banque IRES), 0.2 mM de chaque désoxynucléotides (Thermo Fischer), 1 X d'Evagreen (Biotum) et 4

U de polymérase Q5 haute-fidélité associé au tampon correspondant (NEB) dans un volume final de 50 µL. L'amplification démarre par une étape initiale de dénaturation de 5 min à 98 °C suivi de 25 cycles de : 15 sec à 98 °C, 20 sec à 60 °C and 1 min 45 à 72 °C, puis 2 min à 72 °C. Enfin les fragments amplifiés sont vérifiés sur gel d'agarose 1,5% avec du B.E.T (Roth).

La PCR utilisée pour l'une des stratégies d'élaboration de la banque IRES (4.1) se réalise dans un mélange réactionnelle contenant comme oligo sens 15 pmol (théorique) d'une première ligation optimisée et purifiée dont la réalisation est détaillée dans le manuscrit (4.2.1.1),15 pmol d'oligomère antisens (Tableau 2, Stratégie banque IRES) et), 1 ng de *gfp*mut2 comme matrice, 1 X d'Evagreen (Biotum) et 4 U de polymérase Q5 haute-fidélité associés au tampon correspondant (NEB) dans un volume final de 50 µL. L'amplification démarre par une étape initiale de dénaturation de 5 min à 98 °C suivi de 25 cycles de : 15 sec à 98 °C, 20 sec à 60 °C and 1 min 45 à 72 °C, puis 2 min à 72 °C. Enfin les fragments amplifiés sont vérifiés sur gel d'agarose 1,5% avec du B.E.T (Roth).

La Clic réaction est utilisée parmi les stratégies d'élaboration de la banque IRES (4.1) et se réalise à l'aide du kit de Clic réaction de Jena science (CuAAC Reaction Ligand Test Kit (THPTA & BTTAA based)). 10 pmol de ligation 1 étiqueté d'un azide en 3', 100 pmol de GFPmut2 étiquetée d'un hexynil en 5' avec 8 µL de tampon (1/8 de $CuSO_4$, ¼ de THTPTA et 5/8 d'ascorbate) dans un volume final de 20 µL pendant 1 h à 25°C.

La stratégie du Golden Gate à l'aide de BsaI durant laquelle la première ligation et la *gfp*mut2 sont digérés par BsaI puis ligués ensemble est réalisé dans un volume final de 50 µL avec 0,5 pmol de chaque molécule (ligation 1 et *gfp*mut2) avec 1 X de tampon Cutsmart et 10 U de BsaI pendant 1 h à 37°C et ensuite 20 min à 60°C.

1.2 Tests fonctionnels

1.2.1 Transcription *in vitro* avec l'holoenzyme de E. coli

Le même mélange réactionnel est utilisé en tubes, microplaques et gouttelettes, contenant 1 X de tampon de l'holoenzyme de *E.coli (*NEB), 12,5 mM de chaque rNTPs (Larova), 100 µM de Thiazole-Orange 1 (le ligand fluorogène de Mango-III) et/ou 500 nM de Gemini561 (le ligand fluorogène de o-Coral), 100 mM de spermidine (SIGMA), 5 ng/µL de pyrophosphatase et 2 U d'holoenzyme d'*E. coli*. Le ligand ajouté dépend du riboswitch employé, les concentrations étant de 1 mM de théophylline, 10 µM de néomycine, 1 µM de tétracycline et 1 µM de Flavine mono nucléotide. La réaction peut être incubée de 5 à 21 h à 37 °C avec la mesure en temps réel des aptamères fluorogènes ($\lambda_{ex/em}$: 485/535 nm pour

Mango-III et $\underline{\lambda}_{ex/em}$: 560/600 nm pour o-Coral) par le lecteur de microplaque SpectraMax ID3 (PMT low).

1.2.2 Transcription, traduction *in vitro*, avec de l'extrait de HEK (protocole de nos collaborateurs)

Les cellules de HEK 293FT ont été cultivées par nos collaborateurs (A. TIDU, Equipe G. Eriani) dans du milieu DMEM (Gibco™ DMEM) supplémenté avec 10 % du sérum de veau fœtal (SVF) à 37 °C en atmosphère humide avec 5 % de CO_2. Les cellules ont été passées à 80 % de confluence. Le milieu de culture est premièrement retiré et les cellules sont lavées avec 4 mL de milieu DMEM sans SVF puis détachées en ajoutant 2 mL de trypsine pendant 30 sec. 13 mL de DMEM avec 10 % de SVF sont ensuite ajoutés afin de rassembler l'ensemble des cellules dans une flasque de 250 mL.

L'ensemble de la poursuite du protocole se réalise à 4°C, avec une première centrifugation de 5 min à 1500 rpm. Le culot est re-suspendu dans 40 mL de DMEM sans SVF puis, l'opération est répétée une deuxième et une troisième fois avec la re-suspension dans des volumes respectivement de 10 puis 5 mL de tampon de re-suspension (20 mM d'HEPES-KOH à pH 7.6, 2 mM de $Mg(Ac)_2$, et 100 mM de K(Ac)). L'opération est répétée une dernière fois avec la re-suspension du culot dans 1,5 mL de tampon de re-suspension avec 1 mM de DTT, 1 X d'inhibiteur de protéase Halt™ protease inhibitor cocktail EDTA-free (Thermo) et 40 U d'inhibiteur de RNase (Promega). Les cellules sont alors lysées par cavitation nitrogène et le lysat est clarifié par centrifugation pendant 10 min à 10000 g deux fois de suite. Les surnageants collectés et clarifiés sont aliquotés puis conservés à -80°C après congélation rapide à l'aide d'azote liquide.

Le mélange réactionnel utilisé pour les TTIV en tubes, microplaques ou en gouttelettes est similaire avec 0,1125 mM d'acides aminés, 0,1 M de KAc, 5 mM de $MgAc_2$, 1 mM de rNTP, 1 X de Tampon d'activation de la traduction (20 mM pH 7.6 Hepes KOH, 0,5 mM de spermidine, 1 mM de DTT, 0,8 mM d'ATP, 0,1 mM de GTP, 8 mM de créatine phosphate et 0,1 mg/mL de créatine phosphokinase), 0,5 U/µL d'inhibiteur de RNase, 0,5 mM de l'inhibiteur de réponse au stress appelé ISRIB (de l'anglais Integrated Stress Response Inhibitor), 5 % de T7 polymérase produite et purifiée par nos collaborateurs (2,5 mg/mL dans 50 % de glycérol, 100 mM de NaCl et 20 mM de Tris HCL pH 7.5), et environ 15 % d'extrait obtenue selon le protocole en haut.

1.3 Indexage des banques pour le NGS à (protocole de nos collaborateurs)

1 µL de chaque banque (R0 dilué 50x, R1 dilué 25x, R2/R3 et R3+ non dilués) sont ajouté à un mix de PCR d'un volume de 50 µL avec 0,25 mM de chaque dNTPs (Thermo Fischer), 0,75X de tampon G/C de la Phusion (Thermo Fischer), 0,02 µM d'oligomère (Sens : CTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAATTATTTTGTAGGGTCAACAACT ACCAACTAGTTGG et anti-sens : GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCGCCCTTGCTCACTTG) et 0,015 mg/mL de Phusion produite et purifiée par leur soin. L'amplification démarre par 1 min à 95 °C puis par 20 cycles de 15 sec à 95 °C, 15 sec à 65 °C et 30 sec à 72°C puis se termine par 1 min à 72 °C. Les produits de PCR sont purifiés à l'aide de billes Sera-Mag (Sera-Mag™ Sigma aldrich) selon le protocole du fournisseur avec un ratio de 1 X.

1 µL de chaque banque purifiée est utilisé pour l'indexage par PCR. Le mix de PCR d'un volume de 100 µL contient 0,25 mM de chaque dNTPs (Thermo Fischer), 0,75 X de tampon G/C de la Phusion (Thermo), 0,015 mg/mL de Phusion produite et purifiée par leur soin et 10 µL d'oligomères issus du Kit d'indexage illumina Nextera (illumina). L'amplification démarre par 1 min à 95 °C puis par 3 cycles de 15 sec à 95 °C, 15 sec à 55 °C et 30 sec à 72 °C puis se termine avec 7 cycles de 15 sec à 95 °C, 15 sec à 60 °C et 30 sec à 72 °C. Les produits finaux sont purifiés à l'aide des billes SeraMag comme mentionné précédemment avec un ratio cette fois-ci de 1,6 X. L'ensemble des banques sont rassemblé selon les ratios suivants : R0 à 40 %, R1 et R2 à 20 % et enfin R3 et R3+ à 10 %. L'ensemble est ensuite purifié à l'aide de billes SPRI select (Beckman) avec un ratio de 1,3 X selon le protocole du fournisseur.

Annexe II

1. Résumé de la thèse (en français)

Modèle de 1ère page de résumé de Thèse de Doctorat

UNIVERSITE DE STRASBOURG

RESUME DE LA THESE DE DOCTORAT

Discipline : Science de la Vie

Présentée par : DENTZ Natacha

Titre : **Caractérisation de la machinerie traductionnelle à l'aide de criblages ultra haut-débit en gouttelettes microfluidiques**

Unité de Recherche : UPR9002 – Architecture et Réactivité de l'ARN

Directeur de Thèse : Pr. RYCKELYNCK Michaël

Localisation : Strasbourg - FRANCE

ECOLES DOCTORALES :

(cocher la case)

☐ ED - Sciences de l'Homme et des sociétés ☐ ED 99 – Humanités ☐ ED 101 – Droit, sciences politique et histoire ☐ ED 182 – Physique et chimie physique ☐ ED 221 – Augustin Cournot ☐ ED 222 - Sciences chimiques	☐ ED 269 - Mathématiques, sciences de l'information et de l'ingénieur ☐ ED 270 – Théologie et sciences religieuses ☐ ED 413 – Sciences de la terre, de l'univers et de l'environnement ☒ ED 414 – Sciences de la vie et de la santé

Caractérisation de la machinerie traductionnelle à l'aide de criblages ultra haut-débit en gouttelettes microfluidique

Les gènes codants pour des protéines s'expriment au travers d'un mécanisme en deux étapes au cours duquel l'ADN est tout d'abord transcrit en un ARN messager (ARNm) qui est par la suite traduit en protéine au niveau du ribosome. Cette seconde étape, particulièrement énergivore est finement régulée et, ce, *via* divers mécanismes ciblant principalement l'étape de l'initiation de la traduction. En effet, à l'inverse des étapes d'élongation et de terminaison de la traduction relativement similaire entre les procaryotes et les eucaryotes, l'étape d'initiation est, quant à elle, connue comme étant limitante et foncièrement différente selon l'organisme.

Chez les procaryotes, parmi un grand nombre de mécanismes permettant la modulation de l'accessibilité du site d'entrée des ribosomes (RBS pour Ribosome Binding Site), certains facteurs peuvent agir en *cis*, tels que les riboswitches et d'autres éléments en *trans*, tels que de petits ARNs régulateurs. Cependant, la séquences du RBS à proprement parlé peut affecter l'efficacité de l'initiation, de par son absence premièrement (cas de séquences « leaderless ») ou son degré d'homologie avec la séquence consensus établit par Shine and Dalgarno (SD).

Chez les eucaryotes, l'initiation de la traduction peut démarrer selon deux grands mécanismes : Cap-dépendant ou Cap-indépendant impliquant alors des sites d'entrée interne pour le Ribosome, appelé IRES (de l'anglais Internal Ribosome Entry Site). Les initiations dite Cap-dépendantes peuvent elles-mêmes être plus ou moins efficaces selon le premier codon, la similarité de la séquence en amont avec le consensus décrit par M. Kozak, ou encore le contexte structural de l'ARNm. Les IRES quant à elles sont réparties en plusieurs classes nécessitant plus ou moins de facteurs de l'initiation.

De façon générale, l'ensemble de ces mécanismes peut impliquer de longues séquences et leur compréhension au niveau moléculaire nécessite d'analyser un grand nombre de mutants. Malheureusement, l'analyse manuelle du nombre requis de mutants est chronophage et onéreuse. L'efficacité de ces analyses peut néanmoins être significativement augmentée par des approches de criblage à haut-débit. Ainsi, la cytométrie en flux par FACS (Fluorescence Activated Cell Sorter) se révèle particulièrement efficace. Elle reste néanmoins limitée par les variations entre cellules et la possibilité de pertes d'information due à la toxicité de certains mutants. Ces deux limitations peuvent être surmontées par l'utilisation de milieux de transcription, traduction *in vitro (TTIV)*. Toutefois, l'analyse d'un grand nombre (plusieurs milliers, voire millions) de mutants en parallèle au moyen de milieux de *TTIV* devient rapidement limitée par le coût engendré, justifiant la nécessité de développer des technologies permettant la miniaturisation de ces procédés.

Dans cette optique, la microfluidique en gouttelettes est très attrayante de par sa capacité à produire et manipuler des gouttelettes d'eau dans l'huile de quelques picolitres agissant comme des microréacteurs indépendants. Ainsi, cette technologie permet à partir de 100 microlitres de milieu de TTIV la production de plus de 5 millions de gouttelettes très homogènes en taille à une cadence de plusieurs centaines (voire milliers) par seconde. De plus, après expression des gènes et sélection des gouttelettes, l'utilisation du séquençage à haut-débit (NGS) permet une analyse rapide du devenir de l'ensemble des séquences. Cette utilisation de la microfluidique en gouttelettes couplée au NGS (appelée par la suite µIVC-seq, de l'anglais « microfluidic-assisted In Vitro Compartmentalization and Sequencing ») rend ainsi possible la caractérisation de l'initiation de la traduction à un niveau moléculaire de façon rapide et efficace.
Le but de ma thèse est donc de développer et d'exploiter ces nouvelles stratégies expérimentales à ultrahaut débit afin d'étudier les mécanismes de régulation de l'initiation de la traduction, plus particulièrement chez les procaryotes mais aussi chez les eucaryotes.

Lors de ma thèse, une nouvelle procédure appelée µIVC-seq a été adaptée et utilisée dans le but de caractériser des séquences modulatrices de la traduction. Comme point de départ, un premier système d'expression simple et déjà bien caractérisé a été utilisé, dans l'idée de ré-explorer l'identité des séquences capables de recruter le ribosome procaryote (*E. coli*) et de supporter l'initiation de la traduction. Pour cela, j'ai tout d'abord validé une construction comportant différents éléments : un identifiant unique de goutte (UDI pour Unique Droplet Identifier ; une séquence code-barres permettant de compter le nombre de gouttes contenant une séquence RBS donnée retenue après chaque cycle de criblage), le promoteur de l'ARN polymérase du phage T7 et la phase codante de la GFP (protéine fluorescente verte de l'anglais Green Fluorescente Protein) comme rapporteur de la traduction. Enfin, une région a été dégénérée sur neuf nucléotides et placée à 6 nucléotides en amont du codon de démarrage de la phase codant la GFP. Ainsi, cette expérience m'a conduit à tester plus de 260.000 (4^9) permutations de séquences pour leur capacité à recruter le ribosome. Après validation du pipeline, la banque de ~ 260.000 variants a été criblée pour isoler les séquences capables de recruter le ribosome d'*E. coli* et d'initier la traduction. L'ensemble du criblage a été répété une seconde fois de manière indépendante afin de valider les résultats. L'analyse de l'ensemble des banques (banque de départ et banques enrichies après chaque tour de criblage) par séquençage à haut-débit a enfin permis de suivre le devenir des différentes séquences. Les analyses bio-informatiques ont permis de mettre en avant l'enrichissement de séquences similaires au sein des deux réplicas. Ces derniers révèlent qu'un motif ((G/A)GAG(G/A)) proche du consensus de Shine et Dalgarno (SD) s'accumule au fil des tours de criblage et que la domination du pool de séquences par ce motif corrèle avec une amélioration globale des capacités de la banque à initier la traduction du gène rapporteur. Des

analyses plus approfondies n'ont pas mis en avant de positions préférentielles pour ce motif majoritaire, en accord avec les connaissances sur le placement de la séquence à une distance variable. Ainsi, le motif peut se déplacer à différentes distances du codon initiateur sans affecter de manière notable l'efficacité d'initiation de la traduction. L'ensemble des données collectées permet ainsi de valider la stratégie d'analyse par µIVC-seq et leur pertinence biologique a pu être établie par la cohérence avec les données de la littérature. Cette procédure semble donc récapituler *in vitro* dans des conditions contrôlées, le mécanisme de traduction tel qu'il est connu dans des cellules vivantes.

Par la suite, cette technologie a été exploitée pour la caractérisation et l'ingénierie d'éléments de régulation de la traduction procaryote plus complexes, notamment les riboswitches. Ces molécules peuvent contrôler en *cis* l'expression de gènes au niveau transcriptionnel ou traductionnel. Dans un premier temps, pour mettre au point cette nouvelle procédure au moyen du système techniquement le plus simple possible (coût moindre et paramètres plus simples à contrôler), j'ai débuté avec des riboswitches de contrôle transcriptionnel au travers d'une collaboration avec l'équipe de Mario Möerl. Celui-ci nous a fourni 3 riboswitches développés *in sillico* et répondant à la néomycine, la tétracycline ou à la théophylline. En absence de leur ligand, ces riboswitches forment un terminateur intrinsèque déclenchant l'arrêt précoce de la transcription. En revanche, la présence du ligand conduit à la stabilisation d'une structure anti-terminatrice permettant la poursuite de la transcription. S'agissant d'une régulation transcriptionnelle, il convient de suivre la production d'ARN, ce qui conduit à utiliser un rapporteur pertinent, ici l'aptamère fluorogène *MangoIII r*écemment développé par l'équipe. Il permet en effet de quantifier directement l'efficacité de transcription du gène cible. De plus, pour se rapprocher de la situation cellulaire (notamment en termes de vitesse de synthèse), le promoteur de l'ARN polymérase du phage T7 a été remplacé par le promoteur T7A1 recrutant l'ARN polymérase d'*E. coli*. Sa fonctionnalité (expression, suivi de transcription et réponse du riboswitch) a pu être validée en microplaques avec le riboswitch répondant à la théophylline puis des essais de transplantation de cette construction en systèmes microfluidiques ont été réalisés. L'objectif de ces expériences était d'améliorer/moduler la capacité de réponse de l'ARN à son ligand au travers de criblages fonctionnels de banques de mutants par la réalisation de tours de sélections négatives et positives. Une nouvelle technologie, qui, une fois validée, pourra être étendue aux riboswitches traductionnels.

En parallèle, une collaboration avec le Dr. F. Martin m'a également permis, d'exploiter la technologie de µIVC-seq pour la mise au point d'une plateforme d'analyse s'appliquant aux systèmes eucaryotes. En particulier, je me suis intéressé à l'identification de nouveaux sites d'entrée interne du ribosome (IRES de l'anglais Internal Ribosome Entry Site), notamment au niveau de génomes viraux. Dans un deuxième cas, les séquences régulatrices tel que le consensus décrit par M. Kozak, capable d'aider au recrutement du ribosome mais lors

d’initiations dites coiffe-dépendantes. Pour l’ensemble de ces projets les extraits d’expression bactériens ont été remplacés par des extraits eucaryotes pertinents (par exemple un lysat de cellules d’insectes (S2), de cellules de reins d’embryons humains (HEK) ou encore de réticulocytes de lapin (RLL)).

Dans le cadre du projet cherchant à identifier de nouvelles IRES, la mise au point d’une méthode robuste de préparation de banques constituées de fragments d’ADN génomique directement intégrés dans la construction rapportrice comprenant de 5’ en 3’, le promoteur T7, une séquence d’intérêt, et le gène codant pour la GFP, a été développée et validée. Une fois ce procédé au point, les premières expériences de criblage sont réalisées. Une première procédure utilisant comme modèle le réplicon du virus de la paralysie de criquet (CrpV) dont deux IRES (IRES 5’-UTR et IgR) ont été largement décrites et caractérisées dans la littérature, a été utilisé pour valider le processus de criblage. L’ensemble des sélections et analyses de séquences ont permis la localisation d’une des IRES. Cette preuve de principe sera requise dans le futur pour des expériences similaires sur d’autres virus tels que le virus Zika ou SARS-Cov2 et permettra, à termes, l’identification de nouvelles IRES pouvant servir de cibles thérapeutiques.

Enfin, l’identification de séquences impliquées lors d’une initiation de type cap-dépendant a également été réalisée. A l’heure actuelle les premiers résultats, pour une banque conçue incluant : un promoteur T7, une région variable de 6 paires de bases, le codon initiateur, une région variable de 3 paires de bases et enfin le gène codant pour l’eGFP, a un enrichissement notable qui a pu être observé en microfluidique lors de trois tours de sélection consécutifs. Les résultats obtenus ont motivé leur envoi en séquençage à haut-débit. L’une des principales perspectives sera de comparer ces résultats avec d'autres banques comportant les codons initiateurs GUG et CUG moins fréquents, la présence d’éléments structurés dans cet environnement ou encore l’utilisation d’extraits cellulaires en condition de stress.

Les objectifs de ma thèse concernant la traduction chez les eucaryotes sont en cours d’achèvement et soulèvent de nombreuses perspectives prometteuses. Pour sa part, l’ensemble de la partie concernant la traduction chez les procaryotes ouvre quant à elle la possibilité de nombreuses études que ce soit d’un point de vue plus fondamentale autours des RBS avec des possibilités d’essais incluant d’autres extraits bactériens (Gram positifs) ou d’autres environnements de séquences, que d’un point de vue d’ingénierie de riboswitches.

Bibliographie

A

Acevedo, J. M. *et al.* (2018) “Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence,” *Scientific Reports*, 8(1).

Alexander S. Mironov *et al.* (2002) “Sensing Small Molecules by Nascent RNA: A Mechanism to Control Transcription in Bacteria,” *Cell*, 111, pp. 747–756.

Arraiano, C. M. (1993) “Special Topic Review Post-transcriptional control of gene expression : bacterial mRNA degradation,” *World Journal of Microbiology and Biotechnology*, 9, pp. 421–432.

Atilho, R. M., Perkins, K. R. and Breaker, R. R. (2019) “Rare variants of the FMN riboswitch class in Clostridium difficile and other bacteria exhibit altered ligand specificity,” *RNA*, 25, pp. 23–34.

Autour, A. *et al.* (2019) “Optimization of fluorogenic RNA-based biosensors using droplet-based microfluidic ultrahigh-throughput screening,” *Methods*, 161, pp. 46–53.

Autour, A., Westhof, E. and Ryckelynck, M. (2016) “ISpinach: A fluorogenic RNA aptamer optimized for in vitro applications,” *Nucleic Acids Research*, 44(6), pp. 2491–2500.

B

Baird, S. D. *et al.* (2006) “Searching for IRES,” *RNA*, 12(10), pp. 1755–1785.

Barba-Aliaga, M., Alepuz, P. and Pérez-Ortín, J. E. (2021) “Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene,” *Frontiers in Molecular Biosciences*, 8.

Battaglia, R. A. and Ke, A. (2018) “Guanidine-sensing riboswitches: How do they work and what do they regulate?,” *Wiley Interdisciplinary Reviews: RNA*.

Benitez-Cantos, M. S. *et al.* (2020) “Translation initiation downstream from annotated start codons in human mRNAs coevolves with the Kozak context,” *Genome Research*, 30(7), pp. 974–984.

Berget, S. M., Moore, C. and Sharp, P. A. (1977) “Spliced segments at the 5’ terminus of adenovirus 2 late mRNA*,” *Biochemistry*, 74(8), pp. 3171–3175.

Blattner, F. R. *et al.* (1997) “The Complete Genome Sequence of Escherichia coli K-12,” *Sciences*, 277, pp. 1453–1462.

Boni, I. v *et al.* (1990) *Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1*, *Nucleic Acids Research*, p. 155.

Bonnal, S. *et al.* (2003) “IRESdb: The internal ribosome entry site database,” *Nucleic Acids Research*. Oxford University Press, pp. 427–428.

Borkowski, O. *et al.* (2020) “Large scale active-learning-guided exploration for in vitro protein production optimization,” *Nature Communications*, 11(1).

Bouhedda, F. *et al.* (2020) “A dimerization-based fluorogenic dye-aptamer module for RNA imaging in live cells,” *Nature Chemical Biology*, 16(1), pp. 69–76.

Bouhedda, F. *et al.* (2021) “μIVC-Seq: a method for ultrahigh-throughput development and functional characterization of small RNAs,” *Methods Molecular Biology*, 2300, pp. 203–237.

Bouhedda, F., Autour, A. and Ryckelynck, M. (2018) “Light-up RNA aptamers and their cognate fluorogens: From their development to their applications,” *International Journal of Molecular Sciences*.

Brödel, A. K. *et al.* (2013) “Functional evaluation of candidate ice structuring proteins using cell-free expression systems,” *Journal of Biotechnology*, 163(3), pp. 301–310.

Brödel, A. K., Sonnabend, A. and Kubick, S. (2014) “Cell-free protein expression based on extracts from CHO cells; Cell-free protein expression based on extracts from CHO cells,” *Biotechnology Bioengineering*, 111, pp. 25–36.

Buchner Eduard (1897) “Alkoholische Gärung ohne Hefezellen.”

C

Cambray, G., Guimaraes, J. C. and Arkin, A. P. (2018) “Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in escherichia coli,” *Nature Biotechnology*, 36(10), p. 1005.

del Campo, C. *et al.* (2015) “Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function,” *PLOS Genetics*, 11(10).

Carthew, R. W. and Sontheimer, E. J. (2009) “Origins and Mechanisms of miRNAs and siRNAs,” *Cell*, 136(4), pp. 642–655.

Caschera, F. and Noireaux, V. (2014) “Synthesis of 2.3 mg/ml of protein with an all Escherichia coli cell-free transcription-translation system,” *Biochimie*, 99(1), pp. 162–168.

Chen, H. *et al.* (2005) “Efficient production of a soluble fusion protein containing human beta-defensin-2 in E. coli cell-free system,” *Journal of Biotechnology*, 115(3), pp. 307–315.

Chiaruttini, C. and Guillier, M. (2020) “On the role of mRNA secondary structure in bacterial translation,” *Wiley Interdisciplinary Reviews: RNA*.

Cho, E. and Lu, Y. (2020) “Compartmentalizing Cell-Free Systems: Toward Creating Life-Like Artificial Cells and Beyond,” *ACS Synthetic Biology*.

Chow, L. T. *et al.* (1977) “An Amazing Sequence Arrangement at the 5’ Ends of Adenovirus 2 Messenger RNA,” *Cell*, 12, pp. 1–8.

Contreras-Llano, L. E. and Tan, C. (2018) “High-throughput screening of biomolecules using cell-free gene expression systems,” *Synthetic Biology*, 3(1).

Crick, F. (1970) “Central Dogma of Molecular Biology,” *Nature*, 227, p. 1970.

D

David B. Strauss, William A. Walter and Carol A. Gross (1987) “The heat shock response of E. coli is regulated by changes in the concentration of sigma 32,” *Nature*, 329, pp. 348–351.

Desgranges, E. *et al.* (2019) “Noncoding RNA,” *Microbiology Spectrum*, 7(2).

Dube, D. K. and Palit, S. (1981) “Differential effect of neomycin on DNA dependent DNA and RNA synthesis in vitro,” *Biochemical and Biophysical Research Communications*, 102(1), pp. 378–388.

Dubuc, E. *et al.* (2019) “Cell-free microcompartmentalised transcription–translation for the prototyping of synthetic communication networks,” *Current Opinion in Biotechnology*. Elsevier Ltd, pp. 72–80.

Dutta, T. and Srivastava, S. (2018) “Small RNA-mediated regulation in bacteria: A growing palette of diverse mechanisms,” *Gene*, 656, pp. 60–72.

Duval, M. *et al.* (2013) “Escherichia coli Ribosomal Protein S1 Unfolds Structured mRNAs Onto the Ribosome for Active Translation Initiation,” *PLOS Biology*, 11(12).

E

Etzel, M. and Mörl, M. (2017) “Synthetic Riboswitches: From Plug and Pray toward Plug and Play,” *Biochemistry*, 56(9), pp. 1181–1198.

F

Farnham, P. J. and Platt, T. (1981) “Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro,” *Nucleic Acids Research*, 9(3).

Filonov, G. S. *et al.* (2014) “Broccoli: Rapid selection of an RNA mimic of green fluorescent protein by fluorescence-based selection and directed evolution,” *Journal of the American Chemical Society*, 136(46), pp. 16299–16308.

Findeiß, S. *et al.* (2017) “Design of artificial riboswitches as biosensors,” *Sensors (Switzerland)*. MDPI AG.

Fredrick, K. and Helmann, J. D. (1997) “RNA polymerase sigma factor determines start-site selection but is not required for upstream promoter element activation on heteroduplex (bubble) templates,” *Biochemistry Communicated by Michael J. Chamberlin*, 94, pp. 4982–4987.

Furuichi, Y., Lafiandra, A. and Shatkin, A. J. (1977) “5’-Terminal structure and mRNA stability,” *Nature*, 266, pp. 235–239.

G

Garamella, J. *et al.* (2016) “The All E. coli TX-TL Toolbox 2.0: A Platform for Cell-Free Synthetic Biology,” *ACS Synthetic Biology*, 5(4), pp. 344–355.

Gross, L. *et al.* (2017) “The IRES5′UTR of the dicistrovirus cricket paralysis virus is a type III IRES containing an essential pseudoknot structure,” *Nucleic Acids Research*, 45(15), pp. 8993–9004.

Gumport, R. I. (1970) “Effects of spermidine on the RNA polymerase reaction,” *Annals of the New works academy of sciences*, 171(3), pp. 915–938.

Günzel, C. *et al.* (2021) “Beyond Plug and Pray: Context Sensitivity and in silico Design of Artificial Neomycin Riboswitches,” *RNA Biology*, 18(4), pp. 457–467.

H

Harbers, M. (2014) “Wheat germ systems for cell-free protein expression,” *FEBS Letters*, 588(17), pp. 2762–2773.

Harley, C. B. and Reynolds, R. P. (1987) “Analysis of E. coli promoter sequences,” *Nucleic Acids Research*, 15, pp. 2344–2361.

Harvey Lodish *et al.* (2000) *Molecular Cell Biology*.

Hausser, J. *et al.* (2019) “Central dogma rates and the trade-off between precision and economy in gene expression,” *Nature Communications*, 10(1).

Hecht, A. *et al.* (2017) “Measurements of translation initiation from all 64 codons in E. coli,” *Nucleic Acids Research*, 45(7), pp. 3615–3626.

Hernández, G., Osnaya, V. G. and Pérez-Martínez, X. (2019) “Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes,” *Trends in Biochemical Sciences*, 44(12), pp. 1009–1021.

Hertz, M. I. and Thompson, S. R. (2011) “In vivo functional analysis of the Dicistroviridae intergenic region internal ribosome entry sites,” *Nucleic Acids Research*, 39(16), pp. 7276–7288.

Hodgman, C. E. and Jewett, M. C. (2013) “Optimized extract preparation methods and reaction conditions for improved yeast cell-free protein synthesis,” *Biotechnology and Bioengineering*, 110(10), pp. 2643–2654.

Holstein, J. M., Gylstorff, C. and Hollfelder, F. (2021) “Cell-free Directed Evolution of a Protease in Microdroplets at Ultrahigh Throughput,” *ACS Synthetic Biology*, p. acssynbio.0c00538.

Hori, Y. *et al.* (2017) “Cell-free extract based optimization of biomolecular circuits with droplet microfluidics,” *Lab on a Chip*, 17(18), pp. 3037–3042.

Hossain, G. S. *et al.* (2020) “Genetic Biosensor Design for Natural Product Biosynthesis in Microorganisms,” *Trends in Biotechnology*, 38(7), pp. 797–810.

Husser, C., Dentz, N. and Ryckelynck, M. (2021) “Structure-Switching RNAs: From Gene Expression Regulation to Small Molecule Detection,” *Small Structures*, 2.

I

International Human Genome Sequencing Consortium (2004) “Finishing the euchromatic sequence of the human genome,” *Nature*, 431, pp. 931–945.

J

Jackson, R. J., Hellen, C. U. T. and Pestova, T. v. (2010) “The mechanism of eukaryotic translation initiation and principles of its regulation,” *Nature Reviews Molecular Cell Biology*, pp. 113–127.

Jang, S. K. *et al.* (1990) “Cap-Independent Translation of Picornavirus RNAs: Structure and Function of the Internal Ribosomal Entry,” *Enzyme*, 44, pp. 292–309.

J-C. Baret *et al.* (2009) “Fluorescence-activated droplet sorting (FADS): efficient micro fluidic cell sorting based on enzymatic activity,” *Analyst*, 134(6), pp. 1092–1098.

K

Kearse, M. G. and Wilusz, J. E. (2017) “Non-AUG translation: a new start for protein synthesis in eukaryotes,” *Gene & Development*, 31, pp. 1717–1731. doi: 10.1101/gad.305250.

Kelwick, R. *et al.* (2016) “Development of a Bacillus subtilis cell-free transcription-translation system for prototyping regulatory elements,” *Metabolic Engineering*, 38, pp. 370–381.

Kensal van Holde and Jordanka Zlatanova (1995) “Chromatin Higher Order Structure: Chasing a Mirage?,” *The Journal of Biological Chemistry*, 270(15), pp. 8373–8376.

Kim, Y. *et al.* (1994) “A Multiprotein Mediator of Transcriptional Activation and Its Interaction with the C-Terminal Repeat Domain of RNA Polymerase II,” *Cell*, 77, pp. 599–608.

Komar, A. A. and Hatzoglou, M. (2011) “Cellular IRES-mediated translation: The war of ITAFs in pathophysiological states,” *Cell Cycle*. Taylor and Francis Inc., pp. 229–240.

Komarova, E. S. *et al.* (2020) “Influence of the spacer region between the Shine–Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli,” *Microbial Biotechnology*, 13(4), pp. 1254–1261.

Kosuri, S. *et al.* (2013) “Composability of regulatory sequences controlling transcription and translation in Escherichia coli,” *Proceedings of the National Academy of Sciences of the United States of America*, 110(34), pp. 14024–14029.

Kozak, M. (1981) “Nucleic Acids Research Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes,” *Nucleic Acids Research*, 9(20).

Kozak, M. (1999) “Initiation of translation in prokaryotes and eukaryotes,” *Gene*, 234, pp. 187–208.

Krebs, J. E. *et al.* (2018) *LEWIN’S GENES XII*. Available at: www.jblearning.com.

Kuo, S. T. *et al.* (2020) “Global fitness landscapes of the Shine-Dalgarno sequence,” *Genome Research*, 30(5), pp. 711–723.

Kwan, T. and Thompson, S. R. (2019) “Noncanonical translation initiation in eukaryotes,” *Cold Spring Harbor Perspectives in Biology*, 11(4).

Kwon, Y. C. and Jewett, M. C. (2015) “High-throughput preparation methods of crude extract for robust cell-free protein synthesis,” *Scientific Reports*, 5.

L

Leppek, K., Das, R. and Barna, M. (2018) “Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them,” *Nature Reviews Molecular Cell Biology*, 19(3), pp. 158–174.

Lesley, S. A., Ann Brow, M. D. and Burgess, R. R. (1991) “Use of in Vitro Protein Synthesis from Polymerase Chain Reaction-generated Templates to Study Interaction of Escherichia coli Transcription Factors with Core RNA Polymerase and for Epitope Mapping of Monoclonal Antibodies*,” *The Journal of Biological Chemistry*, 266(4), pp. 2632–2638.

Levine, M. Z. *et al.* (2019) “Escherichia coli-Based Cell-Free Protein Synthesis: Protocols for a robust, flexible, and accessible platform technology,” *Journal of visualized experiments : JoVE*, (144).

Lewis, M. (2013) “Allostery and the lac operon,” *Journal of Molecular Biology*. Academic Press, pp. 2309–2316.

Li, J. and Zhang, Y. (2014) “Relationship between promoter sequence and its strength in gene expression,” *European Physical Journal E*, 37, pp. 1–6.

Link, D. R. *et al.* (2004) “Geometrically Mediated Breakup of Drops in Microfluidic Devices,” *Physical Review Letters*, 92(5), p. 4.

Londei, P. *et al.* (1991) “Translation and ribosome assembly in extremely thermophilic archaebacteria,” *Biochimie*, 73, pp. 1465–1472.

Lozano, G., Francisco-Velilla, R. and Martinez-Salas, E. (2018) “Deconstructing internal ribosome entry site elements: An update of structural motifs and functional divergences,” *Open Biology*, 8(11).

M

Madlung, A. and Comai, L. (2004) “The effect of stress on genome regulation and structure,” *Annals of Botany*, 94(4), pp. 481–495.

Mailliot, J. and Martin, F. (2018) “Viral internal ribosomal entry sites: four classes for one goal,” *Wiley Interdisciplinary Reviews: RNA*. Blackwell Publishing Ltd.

Majzoub, K. *et al.* (2014) “RACK1 controls IRES-mediated translation of viruses,” *Cell*, 159(5), pp. 1086–1095.

Marshall W. Nirenberg and J. Heinrich Matthaei (1960) “The Dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides,” *Biochem. Biophys. Res. Comm*, 46(2), p. 2021.

Martemyanov, K. A. *et al.* (2001) “Cell-free production of biologically active polypeptides: Application to the synthesis of antibacterial peptide cecropin,” *Protein Expression and Purification*, 21(3), pp. 456–461.

Mazutis, L. *et al.* (2009) “Droplet-based microfluidic systems for high-throughput single DNA molecule isothermal amplification and analysis,” *Analytical Chemistry*, 81(12), pp. 4813–4821.

Mazutis, L., Baret, J. C. and Griffiths, A. D. (2009) “A fast and efficient microfluidic system for highly selective one-to-one droplet fusion,” *Lab on a Chip*, 9(18), pp. 2665–2672.

Merrick, W. C. (2004) “Cap-dependent and cap-independent translation in eukaryotic systems,” *Gene*, 332(1–2), pp. 1–11.

Mikami, S. *et al.* (2006) “An efficient mammalian cell-free translation system supplemented with translation factors,” *Protein Expression and Purification*, 46(2), pp. 348–357.

Mokrejs, M. *et al.* (2006) “IRESite: the database of experimentally verified IRES structures (www.iresite.org).,” *Nucleic acids research*, 34(Database issue).

Monod, J. (1949) “The growth of bacterial cultures,” *Annual Reviews Microbiology*, 3, pp. 371–394.

Moore, S. J. *et al.* (2017) “Streptomyces venezuelae TX-TL – a next generation cell-free synthetic biology tool,” *Biotechnology Journal*, 12(4).

Murakami, K. S. (2015) “Structural biology of bacterial RNA polymerase,” *Biomolecules*, 5(2), pp. 848–864.

Mureev, S. *et al.* (2009) “Species-independent translational leaders facilitate cell-free expression,” *Nature Biotechnology*, 27(8), pp. 747–752.

O

Ozawa, Y., Mizuno, T. and Mizushima, S. (1987) “Roles of the Pribnow Box in Positive Regulation of the ompC and ompF Genes in Escherichia coli,” *Journal of Bacteriology*, 169(3), pp. 1331–1334.

P

Pardee, A. B., François Jacob and Jacques Monod (1959) “The genetic control and cytoplasmic expression of ‘Inducibility’ in the synthesis of β-galactosidase by E. coli,” *Journal of Molecular Biology*, 1(2), pp. 165–178.

Pardee, K. *et al.* (2016) “Portable, On-Demand Biomolecular Manufacturing,” *Cell*, 167(1), pp. 248-259.e12.

Pelham, H. R. B. and Jackson, R. J. (1976) “An Efficient mRNA-Dependent Translation System from Reticulocyte Lysates,” *Eur. J. Biochem*, 67, pp. 247–256.

Pernod, K. *et al.* (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 research*, 48(11), pp. 6170–6183.

Piroozmand, F., Mohammadipanah, F. and Faridbod, F. (2020) “Emerging biosensors in detection of natural products,” *Synthetic and Systems Biotechnology*, 5(4), pp. 293–303.

Plamont, M. A. *et al.* (2016) “Small fluorescence-activating and absorption-shifting tag for tunable protein imaging in vivo,” *Proceedings of the National Academy of Sciences of the United States of America*, 113(3), pp. 497–502.

Proshkin, S. *et al.* (2010) “Cooperation between translating ribosomes and RNA polymerase in transcription elongation,” *Sciences*, 328(5977), pp. 504–508.

Q

Qureshi, N. S. *et al.* (2018) “Conformational switch in the ribosomal protein S1 guides unfolding of structured RNAs for translation initiation,” *Nucleic Acids Research*, 46(20), pp. 10917–10929.

R

Richard Jackson and Tim Hunt (1983) “Preparation and Use of Nuclease-Treated Rabbit Reticulocyte Lysate for the Translation of eukaryotic Messenger RNA,” *Methods in Enzymology*, 96.

Richardson, J. P. (2013) “Rho Factor,” in *Brenner’s Encyclopedia of Genetics: Second Edition*. Elsevier Inc., pp. 241–243.

Ringquist, S. *et al.* (1992) “Translation initiation in Escherichia coii: sequences within the ribosome-binding site,” *Molecular Microbiology*, 6(9), pp. 1219–1229.

Rodnina, M. v. (2018) “Translation in prokaryotes,” *Cold Spring Harbor Perspectives in Biology*. Cold Spring Harbor Laboratory Press.

Rodriguez, E. A. *et al.* (2017) “The Growing and Glowing Toolbox of Fluorescent and Photoactive Proteins,” *Trends in Biochemical Sciences*, 42(2), pp. 111–129.

Roncarati, D. and Scarlato, V. (2017) “Regulation of heat-shock genes in bacteria: from signal sensing to gene expression output,” *FEMS microbiology reviews*, 41(4), pp. 549–574.

Rosano, G. L. and Ceccarelli, E. A. (2014) “Recombinant protein expression in Escherichia coli: Advances and challenges,” *Frontiers in Microbiology*, 5.

Ruggero, D., Creti, R. and Londei, P. (1993) “In vitro translation of archaeal natural mRNAs at high temperature,” *FEMS Microbiology Letters*, 107, pp. 378–1097.

Rui Gan and Michael C.Jewett (2014) “Cell-free protein synthesis: Search for the happy middle,” *Biotechnology Journal*, 9(5), pp. 593–594.

Ryabova, L. A. *et al.* (1997) “Functional antibody production using cell-free translation: Effects of protein disulfide isoinerase and chaperones,” *Nature Biotechnology*, 15.

Ryckelynck, M. *et al.* (2015) “Using droplet-based microfluidics to improve the catalytic properties of RNA under multiple-turnover conditions,” *RNA*, 21(3), pp. 458–469.

S

S. D. Poisson (1837) *Recherches sur la probabilitédes jugements en matièrecriminelle et en matière civile; précédées des Règlesgénérales*.

Samantara, K. *et al.* (2021) “A comprehensive review on epigenetic mechanisms and application of epigenetic modifications for crop improvement,” *Environmental and Experimental Botany*, 188.

Scharff, L. B. *et al.* (2011) “Local absence of secondary structure permits translation of mrnas that lack ribosome-binding sites,” *PLoS Genetics*, 7(6).

Schier, A. C. and Taatjes, D. J. (2020) “Structure and mechanism of the RNA polymerase II transcription machinery,” *GENES & DEVELOPMENT*, 34, pp. 465–488.

Sclavi, B. *et al.* (2005) “Real-time characterization of intermediates in the pathway to open complex formation by Escherichia coli RNA polymerase at the T7A1 promoter,” *PNAS March*, 29(13), pp. 4706–4711.

Serganov, A. and Nudler, E. (2013) “A decade of riboswitches,” *Cell*, 152(1–2), pp. 17–24.

Scherlock, M. E. and Breaker, R. R. (2020) “Former orphan riboswitches reveal unexplored areas of bacterial metabolism, signaling, and gene control processes,” *RNA*, 26, pp. 675–693.

Sherlock, M. E., Sudarsan, N. and Breaker, R. R. (2018) “Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems,” *Proceedings of the National Academy of Sciences of the United States of America*, 115(23), pp. 6052–6057.

Shimizu, Y. *et al.* (2001a) “Cell-free translation reconstituted with purified components,” *Nature Biotechnology*, 19, pp. 751–755.

Shimotohnot, K. *et al.* (1977) “Importance of 5’-terminal blocking structure to stabilize mRNA in eukaryotic protein synthesis*,” *Biochemistry*, 74(7), pp. 2734–2738.

Shin, J. and Noireaux, V. (2012) “An E. coli cell-free expression toolbox: Application to synthetic gene circuits and artificial cells,” *ACS Synthetic Biology*, 1(1), pp. 29–41.

Shine, J. and Dalgarno, L. (1974) “The 3’-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites,” *Proceedings of the National Academy of Sciences*, 71(4), pp. 1342–1346.

Shirokikh, N. E. and Preiss, T. (2018) “Translation initiation by cap-dependent ribosome recruitment: Recent insights and open questions,” *Wiley Interdisciplinary Reviews: RNA*, 9(4).

Silvera, D. and Schneider, R. J. (2009) “Inflammatory breast cancer cells are constitutively adapted to hypoxia,” *Cell Cycle*, 8(19), pp. 3091–3096.

Silverman, A. D., Karim, A. S. and Jewett, M. C. (2020) “Cell-free gene expression: an expanded repertoire of applications,” *Nature Reviews Genetics*, 21(3), pp. 151–170.

Smolskaya, S., Logashina, Y. A. and Andreev, Y. A. (2020) “Escherichia coli extract-based cell-free expression system as an alternative for difficult-to-obtain protein biosynthesis,” *International Journal of Molecular Sciences*, 21(3). doi: 10.3390/ijms21030928.

Sonenberg, N. and Hinnebusch, A. G. (2009) “Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets,” *Cell*, 136(4), pp. 731–745.

Sriram, A., Bohlen, J. and Teleman, A. A. (2018) “Translation acrobatics: how cancer cells exploit alternate modes of translational initiation,” *EMBO reports*, 19(10).

Stech, M. and Kubick, S. (2015) “Cell-free synthesis meets antibody production: A review,” *Antibodies*. MDPI AG, pp. 12–33.

Stormo, G. D., Schneider, T. D. and Gold, L. M. (1982) “Characterization of translational initiation sites in E. coli,” *Nucleic Acids Research*, 10(9), pp. 2971–2996.

T

Tawfik, D. S. and Griffiths, A. D. (1998) “Man-made cell-like compartments for molecular evolution,” *Nature Biotechnology*, 16.

Tebo, A. G. *et al.* (2021) “Orthogonal fluorescent chemogenetic reporters for multicolor imaging,” *Nature Chemical Biology*, 17(1), pp. 30–38.

Trachman, R. J. *et al.* (2019) “Structure and functional reselection of the Mango-III fluorogenic RNA aptamer,” *Nature Chemical Biology*, 15(5), pp. 472–479.

V

Vaklavas, C. *et al.* (2015) “Small molecule inhibitors of IRES-mediated translation,” *Cancer Biology and Therapy*, 16(10), pp. 1471–1485. doi: 10.1080/15384047.2015.1071729.

Valencia-Sanchez, M. A. *et al.* (2006) “Control of translation and mRNA degradation by miRNAs and siRNAs,” *Genes and Development*, 20(5), pp. 515–524.

Vitreschak, A. G. *et al.* (2002) “Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation,” *Nucleic Acids Research*, 30, pp. 3141–3151.

W

Wachsmuth, M. *et al.* (2013) “De novo design of a synthetic riboswitch that regulates transcription termination,” *Nucleic Acids Research*, 41(4), pp. 2541–2551.

Wang, H., Li, J. and Jewett, M. C. (2018) “Development of a Pseudomonas putida cell-free protein synthesis platform for rapid screening of gene regulatory elements,” *Synthetic Biology*, 3(1).

Wang, W. *et al.* (2018) “Bacteriophage T7 transcription system: an enabling tool in synthetic biology,” *Biotechnology Advances*. Elsevier Inc., pp. 2129–2137.

Weigand, J. E. and Suess, B. (2007) “Tetracycline aptamer-controlled regulation of pre-mRNA splicing in yeast,” *Nucleic Acids Research*, 35(12), pp. 4179–4185.

Wickiser, J. K. *et al.* (2005) “The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch,” *Molecular Cell*, 18(1), pp. 49–60.

Wilson, J. E. *et al.* (2000) “Naturally Occurring Dicistronic Cricket Paralysis Virus RNA Is Regulated by Two Internal Ribosome Entry Sites,” *MOLECULAR AND CELLULAR BIOLOGY*, 20(14), pp. 4990–4999.

Wilsont, K. S. and von Hippel, P. H. (1995) “Transcription termination at intrinsic terminators: The role of the RNA hairpin (Escherichia coli/RNA polymerase/rho-independent termination),” *Biochemistry*, 92, pp. 8793–8797.

Winkler, W. C., Cohen-Chalamish, S. and Breaker, R. R. (2002) “An mRNA structure that controls gene expression by binding FMN,” *PNAS*, 99(25), pp. 15908–15913.

Woan-Yuh Tarn and Joan A. STeitz (1997) “Pre-mRNA splicing: the discovery of a new spliceosome doubles the challenge,” *Trends in Biochemicals sciences, Cell*, 22, pp. 132–137.

Woodcock, C. L. and Ghosh, R. P. (2010) “Chromatin higher-order structure and dynamics.,” *Cold Spring Harbor perspectives in biology*.

Woronoff, G. *et al.* (2015) “Activity-fed translation (AFT) assay: A new high-throughput screening strategy for enzymes in droplets,” *ChemBioChem*, 16(9), pp. 1343–1349.

Y

Yang, T. H. *et al.* (2021) “Human IRES Atlas: an integrative platform for studying IRES-driven translational regulation in humans,” *Database : the journal of biological databases and curation*, 2021.

Z

Zawada, J. F. *et al.* (2011) “Microscale to manufacturing scale-up of cell-free cytokine production-a new approach for shortening protein production development timelines,” *Biotechnology and Bioengineering*, 108(7), pp. 1570–1578.

Zemella, A. *et al.* (2015) “Cell-Free Protein Synthesis: Pros and Cons of Prokaryotic and Eukaryotic Systems,” *Chembiochem*, 16, pp. 2420–2431.

Zhang, G., Gurtu, V. and Kain, S. R. (1996) “An Enhanced Green Fluorescent Protein Allows Sensitive Detection of Gene Transfer in Mammalian Cells,” 227, pp. 707–711.

Zheng, X. *et al.* (2011) “Leaderless genes in bacteria: Clue to the evolution of translation initiation mechanisms in prokaryotes,” *BMC Genomics*, 12.

Zhou, H. and Zhang, S. (2021) “Recent Development of Fluorescent Light-Up RNA Aptamers,” *Critical Reviews in Analytical Chemistry*, pp. 1–18.

Zubay, G. (1973) “IN VITRO SYNTHESIS OF PROTEIN IN MICROBIAL SYSTEMS,” *Annual Review of Genetics*, 7, pp. 267–287.

Natacha DENTZ

Caractérisation de la machinerie traductionnelle par criblage à ultrahaut-débits en gouttelettes microfluidiques

Résumé

La traduction est un procédé énergivore finement régulé et ce, majoritairement lors de son initiation. Que ce soit au travers de systèmes procaryotes ou eucaryotes, l'étude de ce mécanisme et de sa régulation nécessitent l'analyse d'un grand nombre de variants rendant ces travaux chronophages et onéreux. Les criblages *in vivo* à haut-débit étant efficaces mais limités par les variations intercellulaires ou par certaines conditions toxiques ont motivé l'emploi de milieux de transcription, traduction *in vitro* (TTIV). Toutefois, l'analyse de millions de variants par ces approches *in vitro* est rapidement limitée par le coût, d'où l'intérêt de la microfluidique en gouttelettes qui, par l'utilisation de gouttelettes d'eau dans l'huile de quelques picolitres, permet la réduction du volume, du coût mais aussi du temps de travail. L'ensemble associé au séquençage à haut-débit (NGS) et à l'analyse bio-informatique permet alors une étude rapide et exhaustive des séquences. Dans le cadre de cette thèse, une plateforme de criblage de microfluidique en gouttelettes a été utilisée pour l'analyse de l'initiation de la traduction chez les procaryotes et les eucaryotes ainsi que pour l'étude de certains mécanismes de sa régulation. Plusieurs preuves de principes clefs ont été établies et les perspectives qu'elles ouvrent sont analysées et discutées.

Mots clefs : initiation de la traduction, expression *in vitro*, criblage à ultrahaut-débit, microfluidique en gouttelettes

Summary

The translation is an energy-consuming mechanism finely regulated mainly during the initiation step. To study this mechanism in prokaryotic or eukaryotic systems, a significant number of variants needs to be tested, making such analyses time-consuming and expensive. *In vivo* high-throughput screening proved to be efficient but limited due to cell-to-cell variability and toxicity issues. As a consequence, *in vitro* transcription, translation (IVTT) can be used. However, the *in vitro* analysis of millions of variants is rapidly limited by the cost and thus motivated the development of a dedicated droplet-based microfluidic screening pipeline. The use of picoliter-sized water-in-oil droplets allows the drastic reduction of reaction volumes, the cost, the labor and the working time. Moreover, the combined use of this technology with Next Generation Sequencing and bioinformatics allows the rapid and exhaustive analysis of large sequence libraries. In this work, a droplet-based microfluidic screening pipeline was developed, validated and used to analyze the translation initiation and regulation process in prokaryotic and eukaryotic systems. Several key proof-of-concept have been achieved and the perspectives they open are analyzed and discussed.

Keywords: translation initiation, *in vitro* expression, ultrahigh-throughput screening, droplet-based microfluidics