Université du Québec

INRS Institut national de la recherche scientifique

Centre Armand-Frappier Santé Biotechnologie

DÉCOUVERTE ET CARACTÉRISATION DES ARN NON CODANTS RÉGULATEURS CHEZ *METHYLORUBRUM EXTORQUENS*, UNE BACTÉRIE AU POTENTIEL BIOTECHNOLOGIQUE

Par

Emilie Boutet

Thèse présentée pour l'obtention du grade de
Philosophiae Doctor (Ph.D.)

en Biologie

Jury d'évaluation

Président du jury et examinateur interne	**Dr. Charles Dozois** INRS – Centre Armand-Frappier Santé Biotechnologie
Examinateur externe	**Dr. Éric Massé** Université de Sherbrooke Département de biochimie et génomique fonctionnelle
Examinateur externe	**Dr. Yoann Augagneur** Université de Rennes 1 Département de biochimie pharmaceutique
Directeur de recherche	**Dr. Jonathan Perreault** INRS- Centre Armand-Frappier Santé Biotechnologie

REMERCIEMENTS

J’aimerais d’abord remercier mon directeur de recherche, Jonathan Perreault, de m’avoir accueilli au sein de son laboratoire. Merci d’avoir cru en mes capacités, et surtout de m’avoir fait confiance avec « ton » projet, le SR-PAGE, que tu avais entrepris lors de tes propres études graduées. C’était un bon défi, mais surtout une belle preuve de confiance. Merci pour les nombreuses discussions enrichissantes et tes mille et une idées. Tu as su me partager ta passion pour la science, en m’encourageant lorsque les expériences n’étaient pas fructueuses et en célébrant les réussites. Tu as réussi à me pousser à me surpasser, et j’ai énormément grandi depuis le premier jour de mon stage de baccalauréat. Merci de ton écoute et de ton ouverture, car j’ai toujours eu ton appui dans mes nombreuses implications en parallèle du doctorat. J’ai adoré mon temps au laboratoire et j’en garderai de très beaux souvenirs (les activités de Noël, les gâteaux de fêtes, les sorties à la cabane à sucre, les dîners hot-dog pour en nommer que quelques-uns).

Merci également à ma collègue Aurélie pour le travail d’équipe du tonnerre ! Je n’aurais certainement pas pu prendre en charge le projet SR-PAGE seule. Nos nombreuses sessions de « *brainstorm* » ont su développer ma pensée scientifique. Merci d’avoir répondu à mes nombreuses questions (même si souvent je connaissais la réponse, mais j’étais dans le déni). Tu as été un énorme pilier pour moi durant toutes mes études graduées. On s’est suivi dans notre parcours depuis le jour 1, alors tu étais là au travers des hauts et des bas. Je tiens aussi à remercier tous mes collègues de laboratoires, du passé et du présent, et mes amis de l’INRS : vous avez mis du bonheur dans mes journées. À travers nos petites traditions (les dîners chez Tandori pour célébrer toutes occasions, les pauses-café, les brownies de la cafétéria, les journées Pepsi, les soirées Puzzles Pint, etc.), je me suis sentie à ma place dès le premier jour grâce à vous.

J’aimerais aussi remercier Amélie Côté de m’avoir transmis sa passion pour la communication/vulgarisation scientifique. Merci de m’avoir permis de vivre de nombreuses opportunités comme les conférences dans les écoles secondaires, les visites de laboratoire et le programme Apprentis chercheurs. Ça m’a permis de développer mon esprit de synthèse et mes habiletés de communications, des compétences qui me seront très utiles dans mes prochains défis professionnels.

J’aimerais aussi remercier tous ceux que j’ai croisés lors de mes implications dans la vie étudiante de l’INRS comme ambassadrice, dans la FEINRS, l’infolettre COVID, le Congrès Armand-

Frappier et le journal La Synthèse. Je me suis beaucoup impliquée, mais j'ai eu tellement de plaisir à travailler avec vous dans chacun de ces projets que je ne pouvais pas m'arrêter.

Finalement, j'aimerais remercier mes amis (le groupe MB, mon équipe de DEK et de hockey et mes amis du secondaire) pour votre soutien inconditionnel. Même si vous n'êtes pas familier avec le parcours de recherche universitaire, vous avez célébré avec moi chacun des examens doctoraux (souvent en pensant que c'était le dernier). Je vais bientôt pouvoir répondre à LA question que vous m'avez le plus demandé « quand est-ce que tu vas terminer ? ». Le plus grand des MERCIS à ma famille (Francis, mes parents Benoit et Myriam, mes sœurs Véronique et Karina et ma belle-famille). Merci pour votre patience infinie, votre amour, votre compréhension, votre dévouement et votre soutien. Vous avez été aux premières loges d'une gamme d'émotions, et vos encouragements m'ont donné l'énergie nécessaire pour relever ce défi. Je ne vous en veux pas de ne pas encore avoir retenu le nom de la bactérie *Methylorubrum extorquens*.

Plusieurs personnes ont contribué à ma réussite. La boucle se ferme, et c'est un bonheur pour moi de vous présenter le fruit de mon travail. Ce serait mentir de dire que j'ai rédigé cette thèse seule. Je souhaite donc remercier ma co-écrivaine, mon chat Pika. Merci pour tes bonnes suggestions, je te donnerai un doctorat honorifique.

AVANT-PROPOS

Cette thèse est rédigée en respectant le guide de présentation : mémoires et thèses en sciences de la santé, sciences pures et appliquées (révision 2022) de l'INRS dans le cadre du programme de doctorat en biologie. Cette thèse est présentée en format par article où toutes les références des manuscrits et de la thèse sont regroupées au sein d'une bibliographie commune. Cette thèse commence par une revue de littérature des informations pertinentes pour la compréhension des travaux. Les chapitres de la thèse contiennent ensuite les différents manuscrits réalisés au courant de mes études graduées. Les liens entre les différents articles et les hypothèses de ma thèse sont expliqués avant chaque manuscrit.

- **Chapitre 1** : Article publié revu par les pairs (IJMS) intitulé « *Small RNAs beyond model organisms, have we only scratched the surface? »*. Le matériel supplémentaire pour cet article est accessible à l'annexe I.
- **Chapitre 2** : Article publié dans un journal non revisé par les pairs (BioRxiv) intitulé « *Analysis of non-coding RNAs in Methylorubrum extorquens reveals novel small RNAs specific to Methylobacteriaceae ».* Ce manuscrit a été soumis à un journal revu par les pairs (*RNA biology*) pour publication. Le matériel supplémentaire pour cet article peut être retrouvé à l'annexe II.
- **Chapitre 3** : Article publié dans un journal non revisé par les pairs (BioRxiv) intitulé « *Shifted-Reverse PAGE: a novel approach based on structure switching for the discovery of riboswitches and aptamers ».* Ce manuscrit a été soumis à un journal revu par les pairs (*Nature Methods*) pour publication. Le matériel supplémentaire pour cet article est disponible à l'annexe III. Ce projet a été réalisé en collaboration avec ma collègue Aurélie Devinck, nous sommes co-premières auteures.
- Les résultats obtenus pour la préparation d'un quatrième manuscrit sont abordés en discussion (section 6.1.2 et annexe IV).
- Au cours de ma thèse, j'ai aussi participé aux expériences menant à un article intitulé « *Single mutation in hammerhead ribozyme favors cleavage activity with manganese over magnesium »* publié dans un journal revu par les pairs (*Non-Coding RNA*). Cet article ne fait pas partie de cette thèse, mais les résultats sont discutés dans l'introduction. Je suis deuxième auteure, car j'ai réalisé certaines des expériences et j'ai participé à la rédaction du manuscrit.

RÉSUMÉ

Methylorubrum extorquens est une bactérie au potentiel biotechnologique capable de métaboliser le méthanol, une matière première bon marché qui peut être dérivée de déchets. C'est une méthylotrophe facultative et un organisme modèle pour étudier le métabolisme des C1. Malgré son importance d'un point de vue biotechnologique et en recherche fondamentale, les ARN non codants (ARNnc) de cette méthylotrophe sont peu connus.

Les petits ARN (*small RNA*; sRNAs) de 50 à 300 nucléotides jouent un rôle important dans une grande variété de processus cellulaires, agissant principalement par complémentarité de leur séquence avec leurs ARNm ciblés. Environ 550 familles distinctes de sRNA sont annotées dans un large éventail d'espèces bactériennes, mais ils sont plus nombreux dans les organismes modèles très étudiés par rapport au reste des bactéries séquencées, mettant l'accent sur le potentiel de découverte de nouveaux sRNAs. Par exemple, peu de sRNAs sont annotés dans le génome de *M. extorquens* et aucun n'a été confirmé expérimentalement auparavant. Dans cette étude, les sRNAs précédemment annotés BjrC1505, ffh et CC2171 ont été validés. De plus, l'analyse de données RNA-seq a permis d'établir une liste considérable de sRNAs potentiels, dont les candidats Methylo2624 et Methylo1969, spécifiques au *Methylobacteriaceae*.

Les ARNnc peuvent aussi agir en *cis* lorsqu'ils ont un impact sur l'expression de gènes avoisinants, comme c'est le cas pour les thermorégulateurs ou les *riboswitches* répondants respectivement à un changement de température ou à des métabolites. Les *riboswitches* sont constitués entre autres d'un domaine d'aptamère capable de lier un ligand, ce qui impacte l'expression du gène en aval par la formation d'une structure secondaire d'ARN, un peu à la manière d'un interrupteur. Les méthodes bioinformatiques actuelles pour leur découverte présentent diverses limitations, entre autres pour les organismes comme *M. extorquens* où le nombre de souches séquencées et les annotations génomiques sont limités. Nous avons développé une technique expérimentale appelée le SR-PAGE (*Shifted Reverse-Polyacrylamide Gel Electrophoresis*), une méthode qui tire avantage du changement de structure de la séquence après la liaison avec le ligand dans un gel de polyacrylamide natif. Nous avons d'abord optimisé et validé le SR-PAGE avec des ARN régulateurs connus. Nous avons ensuite démontré que le SR-PAGE pouvait être utilisé comme outil de sélection au sein d'un SELEX afin de sélectionner des *riboswitches* à affinité modifiés et/ou améliorés. La technique SR-PAGE permet d'effectuer un large criblage, car un grand nombre de molécules d'ARN de séquences différentes et plus d'un ligand peuvent être testés simultanément. Cette thèse met l'accent sur l'importance des

ARNnc dans la régulation génétique, en prenant exemple sur une bactérie au potentiel biotechnologique, *M. extorquens*, dont les ARN régulateurs sont peut étudiés.

Mots-clés : ARN non codant (ARNnc), *riboswitch*, ribozyme, *Methylorubrum extorquens*, régulation génétique, petit ARN (sRNA)

ABSTRACT

Methylorubrum extorquens is a biotechnologically relevant bacterium capable of metabolizing methanol, a cheap feedstock that can be derived from waste. It is a facultative methylotroph and a model organism for studying C1 metabolism. Despite its importance from a biotechnological perspective and in fundamental research, little is known about the non-coding RNAs (ncRNAs) of this methylotroph.

Small RNAs (sRNAs) between 50 and 300 nucleotides play an important role in a wide variety of cellular processes, acting by sequence complementarity with their targeted mRNAs. Approximately 550 distinct sRNA families are annotated in a wide range of bacterial species, but they are more prominent in highly studied model organisms compared to the rest of the sequenced bacteria, emphasizing the potential for discovery of new sRNAs. For example, few sRNAs are annotated in *M. extorquens* and none has ever been experimentally confirmed before. In this study, the previously annotated sRNAs BjrC1505, ffh and CC2171 were validated. In addition, analysis of RNA-seq data led to a considerable list of potential sRNAs, including candidate Methylo2624 and Methylo1969, specific to *Methylobacteriaceae.*

Non-coding RNAs can be *cis*-acting where they impact the expression of neighboring genes, as it is the case for thermoregulators or riboswitches responding to a change in temperatures or to metabolites respectively. Riboswitches consist of an aptamer domain capable of binding a ligand, which impacts the expression of the downstream gene by inducing a change in RNA secondary structure, somewhat like a switch. Current bioinformatics methods for their discovery have various limitations, especially for organisms like *M. extorquens* where the number of sequenced strains and genomic annotations are limited. We have developed an experimental technique called SR-PAGE (Shifted Reverse-Polyacrylamide Gel Electrophoresis), a method that takes advantage of the change in secondary structure upon binding of the ligand within a native polyacrylamide gel. We have optimized and validated SR-PAGE with known regulatory RNAs. We also demonstrated that SR-PAGE could be used as a selection tool within a SELEX to select for affinity modified and/or enhanced riboswitches. The SR-PAGE technique allows for broad screening, as numerous RNAs with different sequences and more than one ligand can be tested simultaneously. This thesis focuses on the importance of ncRNAs in gene regulation, using as an example a bacterium with biotechnological potential, *M. extorquens*, whose regulatory RNAs are not well characterized.

Keywords : non-coding RNA (ncRNA), riboswitch, ribozyme, *Methylorubrum extorquens*, genetic regulation, small RNA (sRNA)

TABLE DES MATIÈRES

LISTE DES FIGURES

Les titres des figures en anglais font partie des articles publiés dans cette langue.

LISTE DES TABLEAUX

Les titres des tableaux en anglais font partie des articles publiés dans cette langue.

LISTE DES ABRÉVIATIONS

ADN	Acide désoxyribonucléique
ADNc	Acide désoxyribonucléique complémentaire
ADP	Adénosine diphosphate
AdoCbl	Adénosylcobalamine
APERO	*Analysis of paired-end RNA-seq output*
AqCbl	Aquacobalamine
asRNA	*Antisense RNA*
ARN	Acide ribonucléique
ARNnc	ARN non codant
ARNm	ARN messager
ARNr	ARN ribosomal
ARNt	ARN de transfert
BSRD	*Bacterial small RNA database*
c-AMP-GMP	*Cyclic adenosine monophosphate-guanosine monophosphate*
c-di-AMP	*Cyclic di-adenosine monophosphate*
c-di-GMP	*Cyclic diguanylate monophosphate*
CDP	Cytidine diphosphate
CLASH	*Cross-linking, ligation, and sequencing of hybrids*
CRISPR	*Clustered Regularly Interspaced Short Palyndromic Repeats*
CsrA	*Carbon storage regulator A*
dADP	Désoxyadénosine diphosphate
dCas9	*Dead Cas9*
dCDP	Désoxycytidine diphosphate
DHF	Dihydrofolate
DMS	*Dimethyl sulfide*
EMSA	*Electrophoretic mobility shift assay*
FMN	Flavine mononucléotide
Formyl-THF	Formyl-tétrahydrofolate
GFP	*Green Fluorescent Protein*
GLcN6P	Glucosamine-6-phosphate
GTP	Guanosine triphosphate
Hfq	*Host factor for bacteriophage Qβ RNA replication*
HMP-PP	*4-amino-5-hydroxymethyl-2-methylpyrimidine diphosphate*
MeCbl	Méthylcobalamine
Moco	Cofacteur de molybdène
NCBI	*National Center for Biotechnology information*
PARCEL	*Parallel analysis of RNA conformations exposed to ligand binding*
PCR	*Polymerase Chain Reaction*
PHA	Polyhydroxyalcanoate
PNPase	*Polynucleotide Phosphorylase*

(p)ppGpp	*Guanosine pentaphosphate and tetraphosphate*
preQ1	Pre-queuosine 1
PRPP	*Phosphoribosylpyrophosphate*
RBS	*Ribosome binding site*
RNA	*Ribonucleic acid*
rpm	Rotation par minute
Rut	*Rho utilization site*
SAH	S-adénosylhomocystéine
SAM	S-adénosylméthionine
SELEX	*Selective Evolution of Ligands by Exponentional Enrichment*
SHAPE	*Selective 2'-hydroxyl acylation analyzed by primer extension*
sRNA	*Small RNA*
SRP	*Signal recognition particle*
TAP	*Tobacco Acid Pyrophosphatase*
TGIRT	*Thermostable group II intron reverse transcriptase*
THF	Tétrahydrofolate
TPP	Thiamine pyrophosphate
Tuco	Cofacteur de tungstène
UTR	*Untranslated region*
ZMP	*5-aminoimidazole-4-carboxamide ribonucleotide*
ZTP	*5-amino-4-imidazole carboxamide riboside 5'-triphosphate*
1M7	1-methyl-7-nitroisatoic anhydride
2'-dG	2' deoxyguanosine
3-HP	3-hydroxypropanoÏque

Les abréviations en anglais font partie des articles publiés dans cette langue.

1 INTRODUCTION

1.1 Mise en contexte

La population mondiale toujours grandissante combinée à l'augmentation de la consommation d'énergie nous porte à nous questionner sur la durabilité de ces comportements. Une manière d'économiser de l'énergie serait de profiter des ressources telles que le méthanol, un produit bon marché qui peut être dérivé à partir de sources renouvelables et durables comme des déchets municipaux (Ochsner *et al.*, 2015). Les biotechnologies sont très prometteuses pour répondre à ces questionnements de durabilité, parce que cela nous permettrait de repenser la manière dont nous créons des matériaux comme les produits chimiques en vracs ou les protéines. Mon projet se penche plus spécifiquement sur l'utilisation de la bactérie *Methylorubrum extorquens* comme outil biotechnologique. La capacité de cet organisme à consommer le méthanol peut être mise à profit pour fabriquer des produits à valeurs ajoutées tels que l'acide succinique à partir de cette substance abordable. Ces applications comme outils biotechnologiques requièrent une plus grande connaissance de la régulation génétique déjà présente chez *M. extorquens* afin de pouvoir mieux cibler l'ingénierie génétique. De plus, l'amélioration des outils génétiques disponibles pour cet organisme renforcerait le développement de cette souche comme outil biotechnologique.

Le dogme central de la biologie, un principe fondamental de la biologie moléculaire, stipule que la machinerie cellulaire transcrit fidèlement l'information génétique encodée dans l'ADN en ARN messager simple brin avant de la traduire en protéine. L'ARN a donc longtemps été considéré comme une molécule passive dans le transfert de l'information génétique de l'ADN aux protéines, jusqu'à la découverte des ARN de transfert et des ARN ribosomaux dans les années 50s (Cotter *et al.*, 1967; Hoagland *et al.*, 1958). Ce ne sont pas toutes les molécules d'ARN qui sont destinées à être traduites en protéines, et certaines jouent plutôt un rôle dans la régulation génétique. Les molécules d'ARN n'encodant pas pour une protéine sont désignées comme des ARN non codants (ARNnc), alors que celles qui sont traduites en protéines se nomment ARN messagers (ARNm). Les ARNnc sont impliqués dans un large éventail de processus cellulaires chez les bactéries, catalysent des réactions biochimiques et modulent l'activité de protéines (Lee & Moon, 2018). Il existe plusieurs types d'ARNnc chez les bactéries, comme entre autres les petits ARN (sRNAs), les ARN antisens (asRNAs), les ribozymes, les *riboswitches* et les thermorégulateurs.

Les ARNnc sont une cible intéressante pour la biologie synthétique, parce qu'ils peuvent être réorientés ou être remodelés afin de contrôler l'expression de gènes d'intérêts. Les ARNnc sont

désormais considérés comme des éléments clefs dans la régulation génétique et ils comportent quelques avantages en comparaison avec les régulateurs protéiques : leur production est plus rapide et moins coûteuse, car ils ne nécessitent pas l'étape supplémentaire de la traduction (Waters & Storz, 2009). Les ARNnc endogènes aux bactéries peuvent être ciblés afin de contrôler la croissance bactérienne ou comme biosenseurs. Comme plusieurs sRNAs jouent un rôle important dans la virulence, ils pourraient être la cible de futures thérapies antibactériennes par exemple (Waters & Storz, 2009). Les *riboswitches* quant à eux pourraient être utilisés comme biosenseurs pour détecter leurs ligands correspondants (Blount & Breaker, 2006). Des ARNnc synthétiques peuvent aussi être construits en se basant sur des principes de conceptions simples fondées sur la complémentarité des bases Watson-Crick. Par exemple, des sRNAs synthétiques, des ribozymes et le système CRISPR-Cas9 sont tous des outils de régulations génétiques basées sur l'ARN qui ont été validés chez différentes bactéries afin de réguler l'expression d'un gène d'intérêt (Kharma *et al.*, 2016; Mo *et al.*, 2020; Zhu *et al.*, 2021).

Mieux nous comprendrons les mécanismes de régulation par les ARNnc naturellement présents chez les bactéries, plus de pistes nous auront afin de développer des stratégies d'ingénierie génétique chez une bactérie d'intérêt. Cette thèse s'intéresse donc à l'étude des ARNnc chez la bactérie *M. extorquens*, une usine microbienne capable de métaboliser les carbones C1 tels que le méthanol afin de générer des produits à valeur ajoutée.

1.2 *Methylorubrum extorquens*

Methylorubrum est un nouveau genre bactérien proposé par Green et Arldey en juillet 2018 pour séparer *Methylobacterium,* un genre très varié pouvant être isolé de différents habitats, incluant les sols ou les eaux usées. C'est un groupe d'alphaprotéobactérie méthylotrophe à Gram négatif en mesure de consommer des composés à un carbone (C1) comme source d'énergie. À la suite de la comparaison de l'ARN 16S des différentes souches présentes dans le genre *Methylobacterium*, il a été observé que les séquences des 52 espèces analysées ne partageaient parfois que 92 % d'identité, ce qui est inférieur au seuil limite proposé de 94 % à l'intérieur d'un même genre. Conséquemment, le nouveau genre *Methylorubrum* a été proposé et 11 espèces qui étaient préalablement associées à *Methylobacterium* en font désormais partie (Green & Ardley, 2018). On retrouve dans ce nouveau genre toutes les espèces qui sont en mesure de croître avec du méthanol ou de la méthylamine comme seule source de carbone. En revanche, les espèces qui ne sont pas capables de croître avec seulement de la méthylamine, mais peuvent le faire avec du méthanol sont restées dans le genre *Methylobacterium.* De plus, les espèces

retrouvées dans ce genre taxonomique ne sont pas toutes de couleur rose, alors qu'elles le sont chez le nouveau genre *Methylorubrum* (Green & Ardley, 2018).

La bactérie à l'étude dans ce projet est *Methylorubrum extorquens*, une méthylotrophe facultative à la pigmentation rose qui est utilisée comme organisme modèle afin d'étudier la consommation de produit à un seul carbone. En laboratoire, *M. extorquens* est cultivée à 30 °C avec 250 rpm (rotation par minute) dans des erlenmeyers avec encoche pour permettre une meilleure aération du milieu de culture et prévenir l'agrégation bactéries. Il est possible d'effectuer des mutations dans ses gènes essentiels pour la consommation de C1 afin d'étudier leur rôle, étant donné que cette bactérie est aussi en mesure de croître sur différentes sources de carbone (Ochsner *et al.*, 2015). Cette bactérie est retrouvée dans différents habitats, incluant le sol (Doronina *et al.*, 1996), la phyllosphère (Delmotte *et al.*, 2009) et les eaux usées (Kohler-Staub *et al.*, 1986).

1.2.1 Potentiels biotechnologiques de *Methylorubrum extorquens*

L'utilisation de *M. extorquens* comme usine biologique comporte plusieurs avantages en raison premièrement de sa capacité de croître dans un milieu minimal contenant un mélange de sels et de métaux (Bourque *et al.*, 1995) (Tableau 9.6). Le bouillon de fermentation doit passer par plusieurs étapes de séparations et de purification afin d'atteindre le produit désiré. L'utilisation d'un milieu de culture minimal contenant du méthanol facilite la récupération du produit d'intérêt comparé par exemple à lorsqu'il contient de la liqueur de maïs, une source de carbone complexe souvent utilisée dans les bioprocédés classiques (Ochsner *et al.*, 2015). La plupart des bactéries ne sont pas en mesure de pousser en présence de méthanol en raison de sa toxicité et de leur incapacité à le métaboliser. Cela diminue donc les risques de contamination des fermenteurs. Le méthanol est un composé relativement abordable, ce qui nous permet de démarrer d'un produit bon marché pour produire des substances à valeur ajoutée. La production du méthanol est très flexible, incluant à partir de ressources renouvelables, comme à partir du glycérol, de bois ou de déchets municipaux (Olah, 2013). Sa production peut donc être indépendante des fluctuations du marché des combustibles fossiles, permettant ainsi d'être plus autonomes de ces ressources qui ont un impact négatif sur l'environnement.

M. extorquens a déjà montré son efficacité comme bactérie hôte pour la production de plusieurs composés (revue dans (Ochsner *et al.*, 2015). Les concentrations atteintes sont déjà concurrentielles aux méthodes normalement employées, comme avec l'utilisation de la bactérie *Escherichia coli* (*E. coli*) ou de levures. *M. extorquens* a été utilisé pour produire des composés à valeur ajoutée à partir de méthanol tel que de l'acide 3-hydroxypropanoÏque (3-HP) (Yang *et al.*,

2017), du mévalonate (Zhu *et al.*, 2016) et de l'alpha-humulène (Sonntag *et al.*, 2015) pour en nommer que quelques-uns. Le 3-HP est utilisé dans la production de plusieurs produits chimiques tels que l'acide acrylique retrouvé entre autres dans des revêtements, des adhésifs, des tissus et de la peinture (Yang *et al.*, 2017). Le mévalonate est important dans la synthèse des terpénoïdes utilisés dans les arômes et les parfums (Zhu *et al.*, 2016). Il y a un intérêt dans la production de l'alpha-humulène en raison de ses propriétés anti-inflammatoires et anticancérigènes potentielles (Sonntag *et al.*, 2015). Il est donc possible de tirer avantage des voies métaboliques présentes chez *M. extorquens* afin de créer certains intermédiaires spécifiques tels que les polyhydroxyalcanoates (PHA), ou de modifier génétiquement son génome afin que cette bactérie exprime un gène hétérologue favorisant la production d'intérêt, comme l'acide itaconique par exemple (Lim *et al.*, 2019). L'acide itaconique est utilisé dans la fabrication de polymères super absorbants, dans des détergents et des revêtements de peintures par exemple (Lim *et al.*, 2019).

Les C1 sont des gaz à effet de serre qui ont un impact sur les changements climatiques lorsqu'ils s'échappent dans l'atmosphère (Chistoserdova, 2018). Il est donc d'autant plus attrayant de les utiliser afin de produire des composés à valeur ajoutée à l'aide de bactérie-hôte telle que *M. extorquens*. Une meilleure connaissance de la régulation génétique chez cette bactérie nous permettrait d'améliorer ces procédés biotechnologiques.

1.2.2 Outils génétiques chez *Methylorubrum extorquens*

Les génomes de plusieurs souches de *M. extorquens* ont été séquencés (Vuilleumier *et al.*, 2009) et de multiples outils génétiques sont déjà mis en place afin de faciliter son utilisation comme organisme modèle, incluant des promoteurs inductibles (Carrillo *et al.*, 2019; Chubiz *et al.*, 2013; Kaczmarczyk *et al.*, 2013), une stratégie de mutagenèse basée sur le remplacement de gènes par contre-sélection (Marx & Lidstrom, 2002) et des vecteurs (Marx & Lidstrom, 2001). Par exemple, la méthode de mutation par transposon est utilisée chez *M. extorquens* pour étudier des associations gènes-phénotypes (Van Dien *et al.*, 2003b). Une limitation de cette technique est que les transposons s'insèrent dans le génome de façon aléatoire, ce qui ne permet pas l'étude de gènes ciblés. Pour pallier ce désavantage, l'interférence par système CRISPR (*Clustered Regularly Interspaced Short Palyndromic Repeats*) a été adaptée pour cette alphaprotéobactérie, en plaçant la protéine Cas9 désactivée (dCas9) de *Streptococcus pyogenes* (*S. pyogenes*) sous le contrôle d'un promoteur spécifique à *M. extorquens* (Mo *et al.*, 2020). Par contre, l'interférence par système CRISPR peut parfois mener à des effets hors cible (Vicente *et al.*, 2021). L'utilisation de sRNA (*small RNA*) synthétique afin de moduler l'expression génétique chez *M. extorquens* a aussi été récemment développée (Zhu *et al.*, 2021). Cette méthode se base sur l'hybridation d'un

sRNA synthétique complémentaire à sa cible (voir section 1.3.1 de cette thèse pour la description des sRNAs régulateurs) pour inhiber la traduction d'un gène. Les sRNAs synthétiques contiennent une portion du sRNA MicC d'*E. coli* afin de recruter la protéine Hfq, importante dans la stabilisation du sRNA lui-même et pour favoriser l'interaction entre le sRNA et son ARNm cible (Na *et al.*, 2013) (voir section 1.3.1.1 de cette thèse pour le rôle de la protéine chaperonne Hfq dans l'action des sRNAs). Bien que l'efficacité des sRNAs synthétiques dans la régulation génétique chez *M. extorquens* a été démontrée, cela nécessitait l'expression de la protéine Hfq d'*E. coli,* car la protéine Hfq endogène ne suffisait pas dans l'interaction avec la région MicC du sRNA synthétique (Zhu *et al.*, 2021).

Bien que le potentiel biotechnologique de *M. extorquens* a déjà fait ses preuves, la mise à disposition d'outils génétiques est essentielle pour contrôler finement les voies métaboliques et l'expression des gènes. Par exemple, il serait intéressant de tester l'efficacité de ribozymes *hammerheads* chez *M. extorquens.* Ce sont des molécules d'ARN ayant une activité d'autoclivage, et il est possible de les modifier afin qu'ils ciblent un gène d'intérêt (ce sujet sera discuté dans cette thèse à la section 1.3.3 et 6.1.2).

1.2.3 *Methylorubrum extorquens* et la production de caroténoïdes

M. extorquens est une bactérie à la pigmentation rose (Figure 1.1) en raison de la production de caroténoïdes. Ces molécules retrouvées dans la membrane cellulaire sont importantes dans la stabilisation de celle-ci face à divers stress comme les changements de température (Fong *et al.*, 2001). Dans son habitat naturel sur les feuilles des plantes, *M. extorquens* est exposé à des fluctuations de température entre le jour et la nuit, et les caroténoïdes pourraient être une clef pour s'y acclimater.

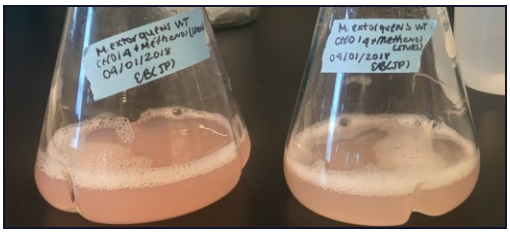

Figure 1.1 Pigmentation rose de *Methylorubrum extorquens* (Crédit photo : Emilie Boutet)

Une étude de mutagenèse par transposon chez *M. extorquens* a permis d'identifier qu'une insertion dans le gène *crtI* encodant pour une phytoène désaturase crée un mutant incolore sans aucun effet sur la croissance (Van Dien *et al.*, 2003a). Lors du développement d'outils de régulation génétique chez *M. extorquens*, d'autres études ont utilisé cela à leur avantage en

ciblant le gène *crtI* afin de vérifier le fonctionnement de leur nouvelle méthode (Mo *et al.*, 2020; Zhu *et al.*, 2021). C'est une stratégie efficace basée sur un phénotype facilement identifiable, étant donné que l'inhibition du gène cause un mutant incolore.

1.3 ARN non codants

Bien qu'il y ait un intérêt d'utiliser *M. extorquens* comme outil biotechnologique, ses ARNnc ont été très peu étudiés. Quelques ARNnc sont annotés dans son génome comme des *riboswitches* et des sRNAs, mais ils n'ont jamais été validés en laboratoire à ce jour. Il y a donc des lacunes quant à la compréhension de la régulation des voies métaboliques de cet important modèle du métabolisme des C1. Des stratégies d'ingénieries génétiques plus ciblées pourraient être développées à mesure que nous comprendrons mieux les mécanismes de régulation déjà en place, impliquant entre autres les ARNnc.

1.3.1 Petits ARN (*Small RNAs*; sRNAs)

Les sRNAs modulent l'expression génétique en s'hybridant avec leurs ARNm ciblés avec soi une complémentarité complète (actif sur le même locus, asRNA) ou avec une complémentarité partielle (actif en *trans*, sRNA). Les asRNAs sont encodés dans le brin opposé de leur cible, alors que les sRNAs sont situés à un locus différent. Ces derniers ciblent souvent plusieurs ARNm et s'appuient parfois sur l'aide de protéines chaperonnes (voir section 1.3.1.1 à 1.3.1.3 de cette thèse sur les protéines chaperonnes). Les sRNAs sont généralement entre 50 et 300 nucléotides et ils sont impliqués dans une grande variété de procédés cellulaires tels que la virulence, la formation de biofilm, l'interaction hôte-pathogène, la résistance aux antibiotiques et l'adaptation face à des changements environnementaux (Jørgensen *et al.*, 2020). Les sRNAs ont un impact au niveau de la traduction d'un ARNm, plus souvent en diminuant la synthèse de protéines qu'en la favorisant (Jørgensen *et al.*, 2020). Par exemple, la liaison d'un sRNA à sa cible peut empêcher le ribosome d'atteindre le site de reconnaissance du ribosome (RBS), soit en bloquant directement son accès en s'y hybridant (Figure 1.2, A), soit en favorisant un changement de structure secondaire qui mène à sa séquestration, empêchant ainsi la traduction de se produire.

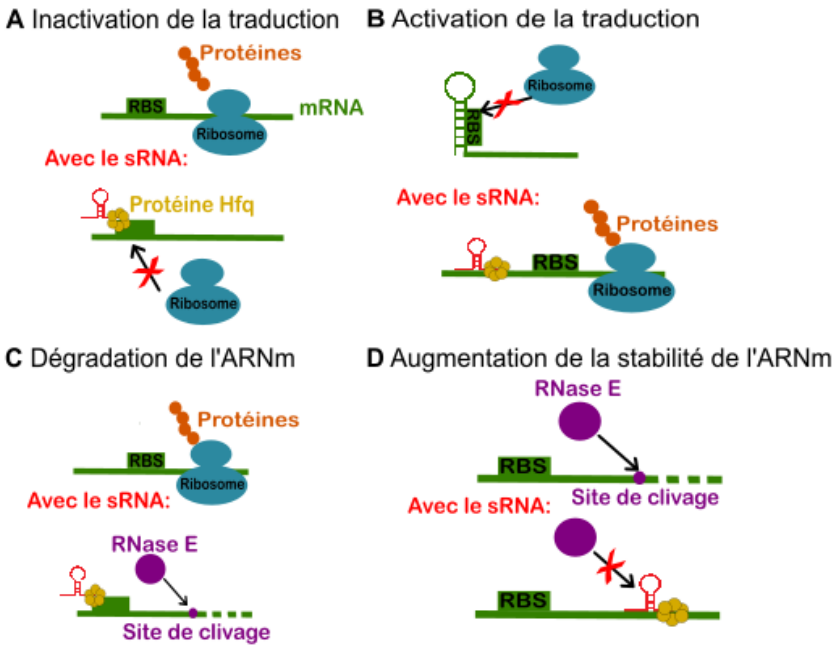

Figure 1.2 Mode d'action des sRNAs pour moduler l'expression génétique, inspiré de (Lalaouna *et al.*, 2013)

(A) La liaison du sRNA à sa cible obstrue le site de reconnaissance du ribosome (RBS), et donc inhibe la traduction. **(B)** La liaison du sRNA à l'ARNm cause un changement de structure secondaire qui libère un RBS normalement séquestré, ce qui active la traduction. **(C)** La liaison du sRNA à l'ARNm le rend plus vulnérable à l'action de la RNase E. **(D)** La liaison du sRNA à l'ARNm obstrue un site normalement reconnu par la RNase E, ce qui prévient sa dégradation.

Les étapes de la transcription et de la traduction sont normalement étroitement liées lors du traitement de l'information génétique chez les bactéries : l'ARNm fraîchement transcrit par l'ARN polymérase est rapidement lié par des ribosomes pour amorcer la traduction. La liaison d'un sRNA à sa cible découple ces étapes, ce qui rend l'ARNm plus vulnérable à la dégradation par la RNase E, une endoribonucléase (Figure 1.2, C). Alternativement, ce découplage peut permettre à la protéine Rho de lier l'ARNm et d'initier une terminaison de transcription. Bien que les sRNAs soient typiquement des répresseurs, ils peuvent aussi activer la traduction d'un ARNm ciblé. Par exemple, l'hybridation d'un sRNA donne la possibilité à une structure secondaire dans l'ARNm de se défaire et ainsi de libérer un RBS qui est normalement séquestré (Figure 1.2, B). L'hybridation d'un sRNA peut aussi augmenter la stabilité d'un ARNm en empêchant la RNase E de reconnaitre un site de clivage (Figure 1.2, D) (Lalaouna *et al.*, 2013).

1.3.1.1 Protéine chaperonne Hfq

La protéine Hfq a initialement été découverte chez *E. coli* comme un facteur essentiel pour la réplication de l'ARN du phage Qβ, ce qui explique l'origine de son nom (*Host Factor required for*

phage Qβ replication) (de Fernandez *et al.*, 1972). C'est une protéine de la superfamille SM/L-SM (*Like-SM*) dont une des caractéristiques est la formation de complexes multimériques en anneau capable de lier l'ARN. Les protéines de cette superfamille sont aussi présentes chez les eucaryotes et les archées et sont impliquées dans différentes fonctions telles que l'épissage de l'ARNm et la stabilisation de l'ARNm (Møller *et al.*, 2002a).

La protéine Hfq est présente dans les génomes d'environ 50% de toutes les bactéries (Sun *et al.*, 2002). Chez les bactéries à Gram négatif, la protéine Hfq facilite l'interaction entre un sRNA et sa cible, en plus de stabiliser l'ARN lui-même. Par exemple, la liaison de la protéine Hfq avec un sRNA peut mener à des changements dans la structure secondaire de ce dernier, favorisant ainsi son interaction avec sa cible. C'est le cas pour les sRNAs OxyS et RprA qui ciblent tous les deux le gène *rpoS* en inhibant et en favorisant sa traduction respectivement (Henderson *et al.*, 2013). Le gène *rpoS* encode pour un facteur sigma alternatif qui active la transcription de gènes en lien avec des conditions de stress chez *E. coli* (Henderson *et al.*, 2013). Les modèles démontrant l'importance de la protéine Hfq dans l'interaction d'un sRNA avec sa cible ARNm sont souvent basés sur des études *in vitro* (Lalaouna et al., 2021). Une récente étude a démontré que la protéine Hfq n'était pas toujours essentielle dans la formation de ce complexe *in vivo* (Lalaouna *et al*, 2021). En absence de la protéine Hfq, le sRNA RyhB pourrait toujours interagir avec ses cibles ARNm *sodB* et *sdhC* et inhiber leur traduction. Cependant, la présence de la protéine Hfq serait importante dans la promotion de la dégradation des ARNm ciblés et augmenterait la stabilité du sRNA RyhB (Lalaouna et al., 2021).

L'endoribonucléase RNase E et la protéine Hfq reconnaissent toutes les deux des régions d'ARN riches en A/U (Mackie & Genereaux, 1993; Møller *et al.*, 2002a). La liaison de la protéine Hfq à un sRNA peut donc le protéger du même coup de la dégradation par la RNase E (Figure 1.2, D) (Moll *et al.*, 2003). La RNase E peut aussi être impliquée dans la dégradation de l'ARNm ciblé une fois que le complexe Hfq-sRNA s'y est lié, un peu à la manière d'un mécanisme spécialisé dans la dégradation de l'ARN. Il y a un débat dans la littérature sur la question si l'interaction entre la protéine Hfq et la RNase E est indirecte et médiée par les ARN ou si celle-ci est direct. Certains modèles suggèrent que la protéine Hfq interagit directement avec l'extrémité C-terminale de la RNAse E, ce qui recrute cette endoribonucléase à la dégradation de l'ARNm ciblé par le complexe Hfq-sRNA (Figure 1.2, C). Par exemple, la dégradation des cibles des sRNAs SgrS et RyhB est médiée par ce complexe ribonucléoprotéique (Morita *et al.*, 2005). Le sRNA Sgrs contrôle l'expression du gène *ptsG* encodant pour un transporteur de glucose et est induit en réponse à une accumulation de sucres phosphates (Vanderpool & Gottesman, 2004), alors que RyhB cible

des ARNm encodant entre autres pour des protéines importantes dans le stockage du fer en réponse à une déplétion de ce métal (Massé & Gottesman, 2002; Massé *et al.*, 2005). D'un autre côté, une récente étude suggère plutôt que l'interaction entre la protéine Hfq et l'extrémité C-terminale de la RNase E serait plutôt médiée par les ARN. Une mutation dans la RNAse E qui empêche son interaction avec un ARN mène à l'inhibition de la liaison de la protéine Hfq aux sRNAs. Cela suggère donc que la RNase est recrutée après la formation du complexe Hfq-sRNA-ARNm par l'intermédiaire des ARN (Sinha & De Lay, 2022).

La protéine Hfq et les sRNAs peuvent aussi former un complexe ribonucléoprotéique avec l'exoribonucléase PNPase (*polynucleotide phosphorylase*). En absence de la protéine Hfq, les sRNAs sont rapidement dégradés par cette exoribonucléase 3′-5′ (Figure 1.3, A), jouant ainsi un rôle important dans la régulation du bassin de sRNAs disponibles. En revanche, lorsque le sRNA est lié à la protéine Hfq, l'exoribonucléase PNPase adopte plutôt une fonction de protection en formant un complexe PNPase-Hfq-sRNA, ce qui empêche d'autres ribonucléases telles que la RNase E de dégrader le sRNA (Figure 1.3, B). Lorsque le sRNA rencontre sa cible, celui-ci est relâché du complexe ribonucléoprotéique et libre d'être dégradé par des ribonucléases, incluant la PNPase elle-même ou la RNase E (Dendooven *et al.*, 2021).

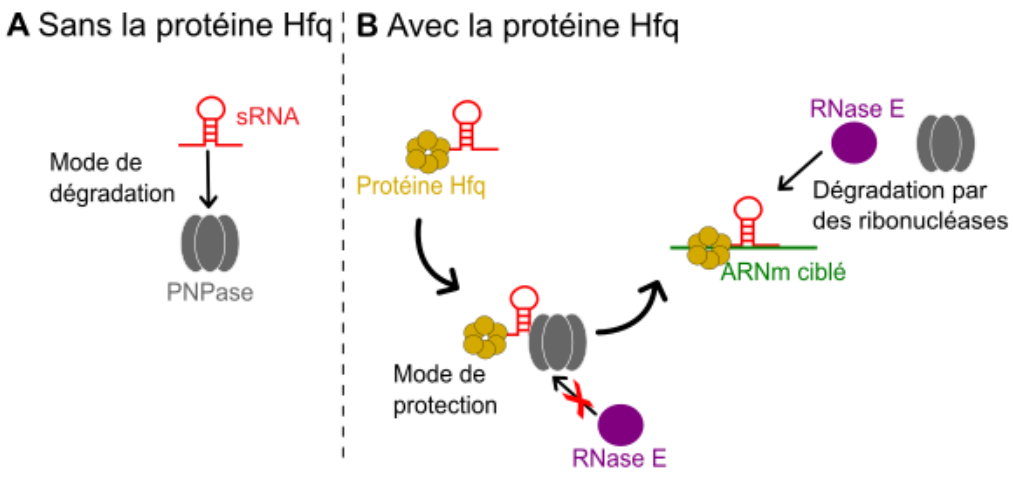

Figure 1.3 Rôle du complexe PNPase-Hfq-sRNA, inspiré de (Dendooven *et al.*, 2021)

(A) Les sRNAs peuvent être dégradés par une PNPase. **(B)** La protéine Hfq, une PNPase et un sRNA peuvent former un complexe ribonucléoprotéique. Dans cette conformation, la PNPase adopte un mode de protection, où elle empêche d'autres ribonucléases de dégrader le sRNA. Lorsqu'un sRNA rencontre sa cible ARNm, le complexe se dissocie, permettant ainsi aux ribonucléases de dégrader les ARN.

La première structure cristalline de la protéine Hfq provenant de la bactérie *Staphylococcus aureus* a permis de résoudre la structure de cette protéine, soit un hexamère sous la forme d'anneau (Schumacher *et al.*, 2002). La face proximale de cette protéine a une affinité pour les séquences riches en U (Schumacher *et al.*, 2002), alors que sa surface distale a une meilleure affinité pour celles comportant plus de A (Mikulecky *et al.*, 2004). La face latérale de la protéine quant à elle préfère lier les ARN riches en A/U (Sauer *et al.*, 2012) (Figure 1.4). Les sRNAs sont divisés en deux classes selon leur type d'interaction avec la protéine Hfq. Les sRNAs de classe I se lient à la face proximale de la protéine Hfq via leur queue poly-U en 3′ de leur terminateur Rho-indépendant ainsi qu'à la face latérale en raison de la présence d'une séquence riche en A/U suivie d'une tige-boucle (Schu *et al.*, 2015). Leur ARNm ciblé interagit avec la face distale grâce à un motif A-A-N en 5′. Les sRNAs de classe II quant à eux interagissent avec la face distale et proximale de la protéine Hfq, alors que leur cible se lie à sa face latérale (Schu *et al.*, 2015) (Figure 1.4).

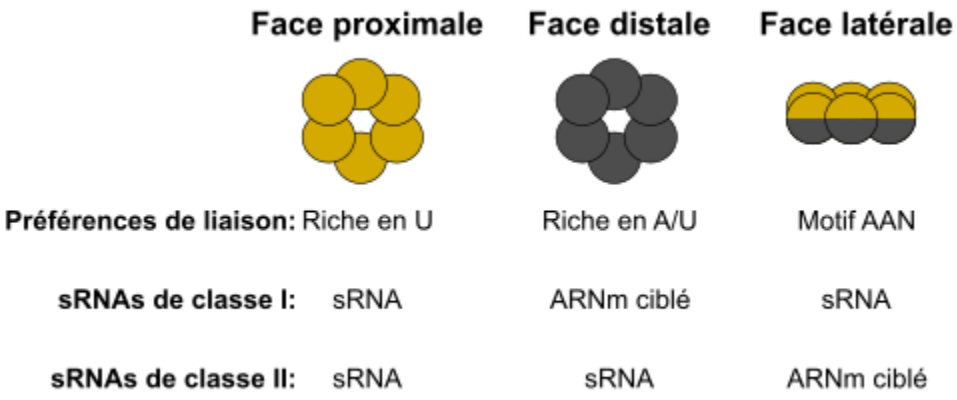

Figure 1.4 Structure et interactions de la protéine Hfq avec les sRNAs et ARNm ciblés, inspiré de (Jørgensen *et al.*, 2020)

Bien que la protéine Hfq soit aussi présente chez les bactéries à Gram positif, elle ne semble pas jouer le même rôle dans le mode d'action des sRNAs. Pour l'instant, seul le sRNA *LhrA* chez *Listeria monocytogenes* a été démontré comme dépendant de la protéine Hfq pour la liaison avec sa cible, le gène *lmo0850* (Nielsen *et al.*, 2010). Cependant, certaines études chez *Clostridium difficile* suggèrent que la protéine Hfq pourrait potentiellement jouer un rôle dans la régulation de sRNA, car la délétion de cette protéine mène à l'accumulation de ces ARN régulateurs (Boudry *et al.*, 2014; Boudry *et al.*, 2021; Fuchs *et al.*, 2021), ce qui suggérerait néanmoins un rôle très différent de celui connu chez les bactéries à Gram négatif. L'habilité de la protéine chaperonne à lier ces sRNAs a aussi été démontrée *in vitro* (Boudry *et al.*, 2014). Dans une autre étude, plusieurs sRNAs prédits chez *C. difficile* ont été sélectionnés par co-immunoprécipitation avec la

protéine Hfq, suggérant leurs interactions. La stabilité de leurs transcrits était aussi affectée par l'absence de la protéine Hfq (Fuchs *et al.*, 2021).

À part ces quelques exceptions, le rôle des sRNAs chez les bactéries à Gram positif est majoritairement indépendant de la protéine Hfq. Chez *Bacillus subtilis* par exemple, il est peu probable que la protéine Hfq joue un rôle dans la régulation génétique par les sRNAs, car leur abondance n'est pas affectée par la délétion de cette protéine (Hämmerle *et al.*, 2014). Similairement, les interactions sRNAs-ARNm ne requièrent pas la présence de la protéine Hfq chez *Staphylococcus aureus* (Bohn *et al.*, 2007). Prises ensemble, ces études indiquent que le nombre de sRNA dépendant de la protéine Hfq est bas chez les bactéries à Gram positif, suggérant que d'autres familles de protéines liant l'ARN pourraient remplir ce rôle.

1.3.1.2 Protéine chaperonne ProQ

La protéine ProQ a été étudiée pour la première fois il y a plus de 30 ans, lorsqu'il a été découvert qu'une insertion par transposon dans le gène *proQ* chez *E. coli* créait un déficit dans l'assimilation de la proline, un acide aminé offrant entre autres une source de carbone et d'azote pour la croissance cellulaire dans des environnements pauvres en nutriments et jouant un rôle dans la réponse au stress osmotique (Milner & Wood, 1989). ProQ est une protéine de la famille FinO qui est conservée chez les alpha, bêta et gamma protéobactéries (Smirnov *et al.*, 2017). L'ampleur de ses activités de liaison avec l'ARN a été tout récemment démontrée chez la bactérie *Salmonella enterica* serovar Typhimurium : des résultats de séquençage à la suite d'une immunoprécipitation de l'ARN suggèrent que cette protéine est associée à plus de 400 transcrits, où les sRNAs sont surreprésentés (Smirnov *et al.*, 2016). L'interaction de la protéine ProQ avec les sRNAs semble davantage déterminée par la structure secondaire de ces derniers que par leurs séquences. ProQ a une meilleure affinité pour les ARN double-brin que ceux simple-brin (G. Chaulk *et al.*, 2011). La protéine ProQ se lie aussi préférentiellement à des sRNAs qui sont indépendants de la présence de la protéine chaperonne Hfq (Smirnov *et al.*, 2016). Alors que cette dernière interagit avec des sRNAs qui agissent en *trans*, ProQ de son côté est plus souvent associé à ceux qui se retrouvent au même locus que leur cible (asRNA) (Holmqvist *et al.*, 2020).

Le rôle de ProQ dans la protection contre l'action de l'exoribonucléase II a été démontré pour l'ARNm *cspE* chez *S. enterica* (Holmqvist *et al.*, 2018). Il est aussi raisonnable de spéculer que ProQ bloque l'accès de la RNase E à certains ARN bien que ce ne soit pas validé en laboratoire,

étant donné que leur site de reconnaissance se chevauche à plusieurs occasions chez *S. enterica* (Holmqvist *et al.*, 2018).

Le premier exemple de régulation d'ARNm médié par un sRNA impliquant la protéine ProQ est le sRNA RaiZ initialement découvert comme un candidat liant la protéine Hfq chez *Salmonella Typhimurium* (Stnc2090) (Chao *et al.*, 2012). Bien que RaiZ interagit avec les protéines Hfq et ProQ *in vitro* et *in vivo*, seulement la délétion de cette dernière a un impact sur la stabilité de RaiZ (Smirnov *et al.*, 2017). Avec l'aide de ProQ, le sRNA se lie spécifiquement avec sa cible, l'ARNm *hupA*, interférant avec la capacité du ribosome à se lier au RBS (Chao *et al.*, 2012). L'interaction du sRNA SraL avec sa cible est aussi médiée par une protéine chaperonne, soit Hfq et/ou ProQ. La protéine Rho favorise la terminaison de la transcription et est connue pour interagir directement avec Hfq, ce qui inhibe ses fonctions (Rabhi *et al.*, 2011). En absence du sRNA SraL, la protéine Rho reconnait son site d'utilisation (*rut site, Rho utilization site*) afin d'initier une terminaison de transcription de l'ARNm *rho*. La liaison du complexe formé du sRNA SraL et de la protéine chaperonne ProQ/Hfq à la région 5' non-transcrite de l'ARNm *rho* le protège d'une terminaison de transcription causée par sa propre protéine Rho (Silva *et al.*, 2019).

Le rôle de ProQ comme protéine chaperonne dans l'interaction des sRNAs avec leur cible a récemment été élucidé et ouvrent la porte à plusieurs autres découvertes, que ce soit pour de nouveaux sRNAs ou pour la compréhension des mécanismes d'action de ceux identifiés comme interagissant avec cette protéine.

1.3.1.3 Protéine chaperonne CsrA

Même si elle est présente chez certaines bactéries à Gram positif, la protéine Hfq ne semble pas jouer un rôle important dans le mode d'action des sRNAs comme elle le fait chez les bactéries à Gram négatif, sauf chez *Listeria monocytogenes* (Nielsen *et al.*, 2010). De plus, la protéine ProQ est tout simplement absente chez les bactéries à Gram positif.

On pourrait donc supposer qu'une protéine autre que Hfq ou ProQ pourrait accomplir ce rôle chez les bactéries à Gram positif. La protéine CsrA (*Carbon Storage Regulator A*) chez *Bacillus subtilis* offre la première piste afin de résoudre cette pièce manquante du casse-tête. La protéine CsrA a d'abord été identifiée chez *E. coli* pour son rôle dans le métabolisme du carbone et dans les propriétés de surface cellulaire, telle que l'adhérence (Romeo *et al.*, 1993). Il a ensuite été démontré que CsrA était capable de lier l'ARN, avec comme premier exemple l'ARNm *glgC* encodant pour une protéine importante dans la synthèse du glycogène (Baker *et al.*, 2002). CsrA inhibe la traduction de cet ARNm en se liant à plusieurs sites dans la région 5′ non-traduite de

l'ARNm, dont une correspondant à la séquence du Shine-Dalgarno. CsrA compétitionne donc avec le ribosome pour l'accès au site de reconnaissance du ribosome (Baker *et al.*, 2002; Liu & Romeo, 1997). Bien qu'initialement découvert chez *E. coli*, ce mécanisme d'inhibition de la traduction médié par la protéine CsrA s'étend à plusieurs autres bactéries et cible plusieurs autres ARNm (Romeo *et al.*, 2013). Cette protéine joue entre autres un rôle dans la formation de biofilm, la virulence, la motilité et le *quorum-sensing*, un mécanisme de communication bactérienne via des molécules de signalisation (Romeo *et al.*, 2013). CsrA agit en homodimère, où deux régions identiques reconnaissent le motif hautement conservé GGA. La régulation par la protéine CsrA peut promouvoir la traduction d'un ARNm en le protégeant de l'action de la RNase E par exemple, comme c'est le cas pour l'opéron *flhDC* chez *E. coli* (Wei *et al.*, 2001). La liaison de CsrA à sa cible peut aussi mener à des changements de structure secondaire, menant à la traduction de l'ARNm (Patterson-Fortin *et al.*, 2013). La modification de la structure secondaire de l'ARNm causée par l'association de CsrA peut mener à la libération d'un site reconnu par la protéine Rho (site rut) et donc initier une terminaison de la transcription (Figueroa-Bossi *et al.*, 2014).

Récemment, il a été démontré que la protéine CsrA pouvait adopter un tout nouveau rôle chez *B. subtilis*, soit d'encourager la liaison d'un sRNA à son ARNm ciblé, un peu à la manière des protéines chaperonnes Hfq et ProQ chez les bactéries à Gram négatif. *In vitro*, CsrA se lie au sRNA SR1 et à sa cible, l'ARNm *ahrC* encodant pour une protéine importante dans le catabolisme de l'arginine. Le sRNA SR1 inhibe sa cible, freinant ainsi le métabolisme de cet acide aminé. Lorsque le sRNA SR1 est surexprimé, la bactérie ne peut pas utiliser l'arginine comme seule source de carbone. Par contre, en absence de la protéine CsrA, la bactérie retrouve sa croissance normale malgré la surexpression de SR1, démontrant son rôle essentiel dans l'interaction du sRNA avec sa cible (Yakhnin *et al.*, 2007). En plus de la découverte de cette nouvelle fonction comme protéine chaperonne, CsrA agit aussi comme régulateur traductionnel chez *B. subtilis* en bloquant par exemple l'accès du RBS du gène *hag* encodant pour une flagelline (Yakhnin *et al.*, 2007).

Il n'existe pour l'instant qu'un seul exemple du rôle de CsrA comme protéine chaperonne dans l'interaction d'un sRNA avec sa cible. Ces nouveaux résultats ouvrent la porte à de nouvelles découvertes : plus de 1500 bactéries encodent aussi pour cette protéine, donc il est fort possible qu'il existe d'autres exemples de régulation génétique médiée par un complexe sRNA-CsrA (Van Assche *et al.*, 2015).

L'endoribonucléase RNase III pourrait aussi combler le rôle de chaperonne chez les bactéries à Gram positif selon une récente étude chez *S. aureus* (Mediati *et al.*, 2022). La technique CLASH

(*UV cross-linking, ligation, and sequencing of hybrids*) a été utilisée afin d'étudier les interactions ARN-ARN associées avec la RNase III. L'endoribonucléase sert d'appât et permet de capturer les duplexes ARN-ARN qui y sont associés *in vivo* (McKellar *et al*, 2022; Mediati *et al*., 2022). Comme avec l'endoribonucléase RNase E, les duplexes sRNA-ARNm sont ciblés par la RNase III, menant à la dégradation de l'ARNm. Cependant, il a été suggéré que la RNase III pouvait aussi avoir une fonction non catalytique (Calin-Jageman & Nicholson, 2003). Par exemple, les résultats CLASH chez *S. aureus* ont démontré que les sRNAs RsaI et RsaE formaient un complexe avec la RNase III. Bien que ce complexe stimule la dégradation du sRNA RsaE *in vitro*, sa stabilité n'est pas affectée *in vivo*. Dans cet exemple, la RNase III agirait de façon non catalytique et sa fonction première serait de stabiliser l'interaction entre les deux sRNAs (McKellar *et al.*, 2022).

1.3.1.4 Méthode pour la découverte de sRNAs candidats

Plusieurs approches peuvent être utilisées afin de découvrir de nouveaux sRNAs chez un organisme d'intérêt, soit des méthodes expérimentales basées sur leur expression, ou des stratégies basées sur leur séquence ou leur contexte génomique par voie bioinformatique.

Les premiers sRNAs ont été découverts entre autres grâce au marquage radioactif d'ARN abondants. Le phosphate radiomarqué (^{32}P) est incorporé dans les acides nucléiques des bactéries. L'ARN total radiomarqué est isolé et séparé sur un gel de polyacrylamide. Les bandes d'intérêt sont ensuite sélectionnées et identifiées grâce à la digestion par des nucléases (Vogel & Sharma, 2005). Le sRNA Spot 42 (initialement nommé *spf*) important dans le métabolisme de carbohydrates a été identifié par cette méthode (Ikemura & Dahlberg, 1973), mais son rôle dans la régulation génétique n'a été élucidé que plusieurs années après, lorsqu'il a été démontré qu'il pouvait inhiber l'opéron galactose (Møller *et al.*, 2002b). L'avantage de cette technique est que l'intensité des bandes radioactives est reliée à l'abondance de l'ARN dans la cellule bactérienne.

L'identification de potentiels sRNAs peut aussi être basée sur le profil d'expression lors de la transcription. La technique des *microarrays* permet d'analyser l'expression de séquences intergéniques à l'aide de sondes. Cette méthode a été utilisée afin d'identifier 34 nouveaux sRNAs candidats chez *E. coli* par exemple (Wassarman *et al.*, 2001). Cependant, cette stratégie est coûteuse à mettre en place, car elle nécessite la création d'un grand nombre de sondes et a pratiquement été rendue obsolète par l'arrivée des technologies de séquençage à haut-débit.

Avant le développement du séquençage à l'échelle transcriptomique, les premières méthodes expérimentales nécessitaient la transcription inverse de l'ARN en ADN complémentaire (ADNc) suivi d'étapes de clonage afin d'être en mesure de les séquencer (Vogel et al., 2003). Cette

technique a entre autres été utilisée chez *E. coli* afin d'identifier de nouveaux sRNAs à l'échelle du génome : l'ARN total des bactéries est isolé dans différentes conditions de croissances à différent temps et des ADNc sont ensuite créés par transcription inverse de l'ARN. Ceux-ci sont clonés dans des plasmides et ils sont ensuite séquencés. Les séquences obtenues sont comparées avec le génome déjà annoté de la bactérie d'intérêt (Vogel et al., 2003).

Les sRNAs interagissent souvent avec des protéines chaperonnes pour le bon fonctionnement de leur rôle de régulation génétique (comme discuté dans les sections 1.3.1.1 à 1.3.1.3 de cette thèse). Ces protéines peuvent être importantes entre autres pour stabiliser le sRNA lui-même, pour favoriser son hybridation avec sa cible ou pour le protéger de l'action de nucléases. La co-purification des sRNAs avec une protéine est donc un bon moyen de tirer profit de cette interaction afin d'identifier de nouveaux sRNAs. La protéine Hfq est souvent utilisée comme appât, car elle a été démontrée comme détenant un rôle important dans la fonction de plus d'un tiers des sRNAs connus chez *E. coli* par exemple (Zhang et al., 2003).

Il est aussi possible de tirer profit de l'interaction des sRNAs avec des protéines pour découvrir de nouvelles molécules régulatrices en utilisant le procédé du SELEX (*Selective Evolution of Ligands by Exponential Enrichment*). L'idée est de créer une librairie génomique de l'ADN de la bactérie d'intérêt contenant des séquences de tailles désirées. La librairie est constituée de fragments contenant des séquences connues aux extrémités, ce qui s'avère utile pour amplifier les séquences par PCR (*polymerase chain reaction*) (Lorenz *et al.*, 2006). La librairie génomique est transcrite en ARN à l'aide du promoteur T7 avant d'être incubée avec l'appât protéique. Seulement les séquences en mesure d'interagir avec la protéine d'intérêt sont retenues. Celles-ci sont amplifiées par PCR à la suite d'une transcription inverse, avant d'être à nouveau transcrites en ARN. Un autre tour de sélection est ensuite effectué où les molécules sont de nouveau mises en contact avec l'appât protéique. Chacun de ces tours de sélection permet d'avoir un enrichissement des séquences capables de s'hybrider avec la protéine d'intérêt. Après quelques tours de sélection, les séquences sont clonées et séquencées afin d'identifier les potentiels sRNAs (Figure 1.5) (Lorenz *et al.*, 2006).

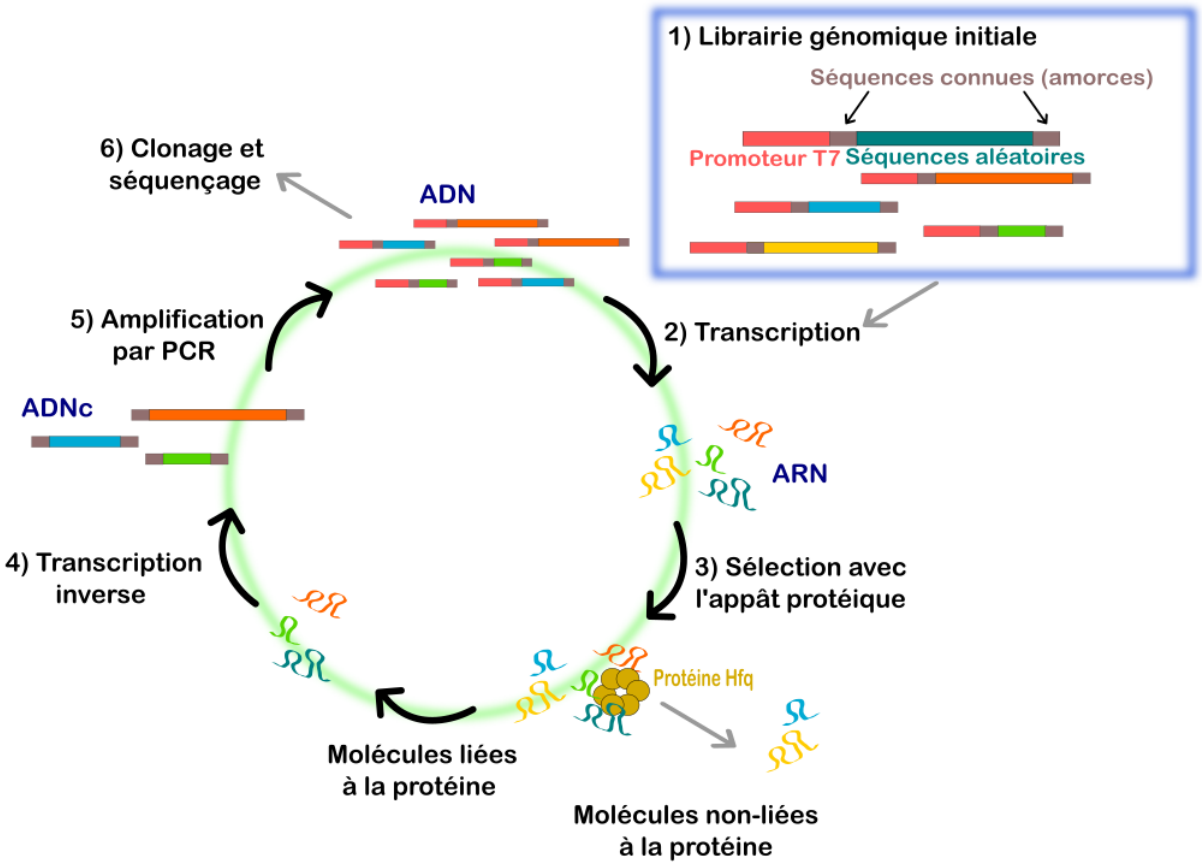

Figure 1.5 Vue d'ensemble du procédé de SELEX inspiré de (Lorenz *et al.*, 2006)

Un des grands inconvénients de ces techniques basées sur l'interaction entre un sRNA et un appât protéique est que cela nécessite que cette interaction demeure tout au long des procédés de sélection par co-purification ou par SELEX.

Le sRNA-seq (*sRNA-sequencing*) est une autre approche expérimentale non biaisée pour la découverte de sRNA. L'ARN total est extrait et une gamme de taille prédéterminée est sélectionnée. Les ARN abondants qui sont surreprésentés comme l'ARN ribosomal (ARNr) 5S ou les ARN de transfert (ARNt) sont retirés du bassin de séquence à l'aide d'oligonucléotides spécifiques. Des séquences connues sont ajoutées en 5′ et 3′ des transcrits afin d'être en mesure de les amplifier par PCR après une transcription inverse. Les séquences sont ensuite clonées et séquencées. La comparaison des candidats avec le génome d'intérêt permet l'identification de potentiels sRNAs (Liu & Camilli, 2011). Cette méthode a été utilisée pour l'identification de sRNA chez le pathogène humain *Vibrio cholerae*, ce qui a mené à la détection de 500 sRNAs et 127 asRNAs potentiels. De plus, les 20 sRNAs connus au préalable chez *V. cholerae* ont aussi été sélectionnés, validant ainsi cette méthode (Liu *et al.*, 2009). Un avantage de la technique *sRNA-*

seq est qu'elle nous fournit les renseignements pour la taille exacte du sRNA, étant donné que les extrémités 5′ et 3′ sont identifiées en même temps.

Des candidats peuvent être identifiés grâce à la présence d'un site d'initiation de transcription avec la technique de 5′ RACE (*Rapid Amplification of cDNA ends*). Pour les identifier, une librairie d'ARN est traitée avec l'enzyme TAP (*Tobacco Acid Pyrophosphatase*). Cet enzyme convertit le groupe triphosphate présent en 5′ des séquences d'ARN en monophosphate, ce qui permet la liaison d'un adaptateur utilisé dans l'amplification des fragments à la suite de la transcription inverse. Les amplicons de la librairie d'ARN qui ont subi le traitement à l'enzyme TAP est comparé à la librairie contrôle. Les produits d'amplifications qui sont seulement présents à la suite du traitement avec l'enzyme TAP indiquent la présence d'un site d'initiation de la transcription. Similairement, le 3′ RACE peut être utilisé afin de distinguer les sites de terminaison de la transcription. Un adaptateur en mesure de lier le groupe hydroxyle en 3′ des ARN est ligué aux ARN d'un transcriptome (Argaman *et al.*, 2001).

Des approches computationnelles peuvent aussi être utilisées afin d'identifier des sRNAs potentiels en regardant par exemple la présence d'un promoteur et d'un terminateur dans une région intergénique, la conservation entre les espèces ou la formation de structures secondaires d'ARN (Vogel & Sharma, 2005). Cependant, les approches bioinformatiques s'appuient sur des caractéristiques recherchées, et les ARN qui ne répondent à ces critères pourraient nous échapper. Un exemple d'une telle étude inclut les travaux d'Argaman *et al.*, 2001. Afin d'identifier de nouveaux sRNAs chez *E. coli*, les régions intergéniques non annotées (les régions entre deux séquences codant pour des protéines) contenant des sites d'initiation de transcription et des terminateurs de transcriptions Rho-indépendant ont été sélectionnées. Afin de limiter le nombre de candidats, seulement ceux entre 50 et 400 nucléotides qui sont conservés dans d'autres espèces ont été sélectionnés, ce qui a mené à l'identification de 24 sRNAs potentiels chez *E. coli* (Argaman *et al.*, 2001). La conservation de séquences intergéniques est une bonne indication que ces régions ont une fonction importante. Ce paramètre a aussi été utilisé auparavant afin d'identifier de potentiels sRNAs chez *E. coli* en comparant avec les séquences intergéniques des espèces *Salmonella* et *Klebsiella* (Wassarman, 2001).

1.3.1.5 Découverte de sRNAs candidats à l'aide de *RNA-seq*

La découverte de nouveaux sRNAs a été bouleversée par l'analyse de données transcriptomique par *RNA-seq*. Ces jeux de données peuvent être analysés par inspection manuelle, par le développement d'outil informatique ou par un mélange de ces deux approches.

Les différences transcriptionnelles entre les conditions de croissance peuvent être un point de départ afin de découvrir de nouveaux sRNAs. Par exemple, plusieurs sRNAs ont été identifiés chez *Agrobacterium tumefaciens* en comparant des conditions de virulence et de non-virulence induite à l'aide de l'acétosyringone, un produit chimique qui induit la virulence bactérienne de cette bactérie (Wilms *et al.*, 2012).

Plusieurs méthodes computationnelles ont été créées afin d'analyser des données transcriptomiques pour identifier des candidats de sRNAs. Ces potentielles séquences régulatrices sont reconnues soit en comparant les transcrits à des génomes préalablement annotés ou en se basant sur différents critères de sélection. Par exemple, RNA-eXpress prend en considération la taille d'un ARN ainsi que le niveau de couverture minimale pour identifier des sRNAs, et ce indépendamment des annotations génomiques déjà existantes (Forster *et al.*, 2013). Avec cette méthode, aucun biais n'est créé avec les informations déjà connues. D'un autre côté, DETR'PROK (*Detection of noncoding RNA in prokaryotes*) compare tout d'abord les données transcriptomiques avec des annotations génomiques afin de classifier chaque lecture en trois catégories distinctes, soit une région codante, un sRNA ou un ARN antisens (asRNA) (Figure 1.6) (Toffano-Nioche *et al.*, 2013).

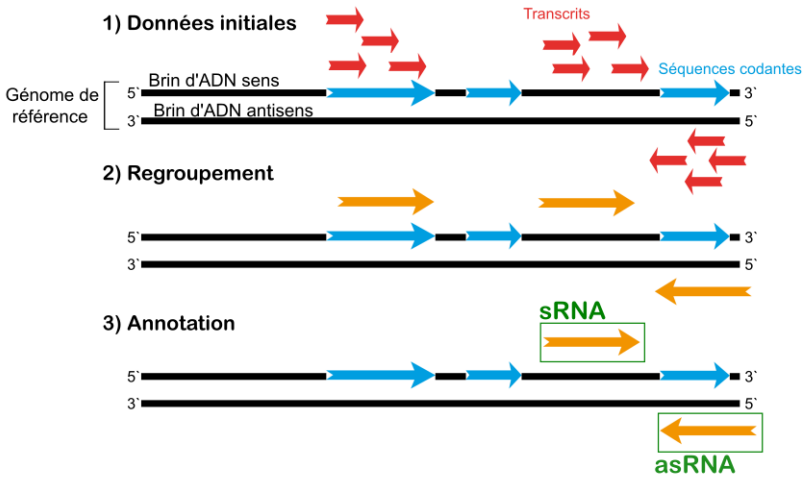

Figure 1.6 Représentation graphique des paramètres utilisés par DETR'PROK inspiré de (Toffano-Nioche *et al.*, 2013)

L'outil DETR'PROK compare les différents transcrits (représentés en rouge) avec les séquences codantes (représentées en bleu) dans le génome de référence. Les transcrits sont tout d'abord regroupés (représentés en jaune) avant d'être classifiés en région codante, sRNA ou asRNA selon la comparaison avec le génome de référence (Toffano-Nioche *et al.*, 2013).

DETR'PROK n'utilise pas le séquençage en paire, ce qui lui a valu le plus faible taux de détection de sRNA déjà connus lors d'une étude de comparaison des outils d'identification de sRNA à partir

de données transcriptomiques (Leonard *et al.*, 2019). Une autre approche informatique appelée *sRNA-Detect* peut être utilisée afin d'identifier des sRNAs à partir de résultats d'une étude transcriptomique. Cet outil ne se base pas sur les annotations génétiques préalablement établies, mais il prend plutôt en considération la taille des séquences et la couverture des lectures à la suite de l'étude transcriptomique. Pour être considéré comme un sRNA par cet outil, un fragment doit être plus petit que 250 nucléotides et être transcrit relativement uniformément tout au long de sa séquence (Peña-Castillo *et al.*, 2016). Lorsqu'on parle de couverture dans une étude transcriptomique, on fait référence au nombre de lectures correspondant à chaque nucléotide d'une séquence d'intérêt. Par exemple, si les lectures recouvrent une séquence de manière égale, on parle d'une couverture uniforme (Figure 1.7).

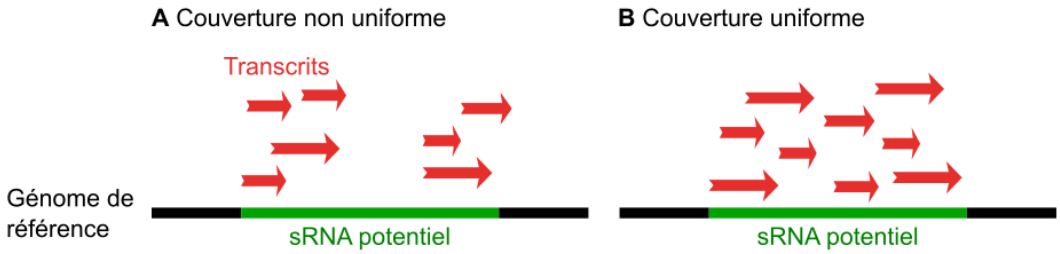

Figure 1.7 Couvertures non uniformes et uniformes des potentiels sRNAs par les lectures du RNA-seq (transcrits)

ANNOgesic est un outil qui combine plusieurs autres algorithmes afin de faire des prédictions plus robustes pour annoter les données transcriptomiques. Ce dernier n'est pas seulement conçu pour identifier des sRNAs, mais détecte aussi les régions intergéniques, les sites d'initiation de transcription et les opérons (Yu *et al.*, 2018). Afin d'identifier un potentiel sRNA, ANNOgesic regarde d'abord l'uniformité de la couverture des transcrits. Les candidats sont ensuite comparés avec BSRD (*Bacterial small RNA database*), une base de données qui compile les sRNAs validés expérimentalement (Li *et al.*, 2013) et avec NCBI (*National Center for Biotechnology information*) (Wheeler *et al.*, 2007) afin d'identifier de potentielles séquences homologues (à noter que BSRD n'est plus en fonction). Un transcrit doit aussi respecter plusieurs autres critères avant d'être classé comme un potentiel sRNA, soit contenir une prédiction pour un site d'initiation de transcription, former une structure secondaire stable prédite par RNAfold (Gruber *et al.*, 2008) et être entre 30 et 500 nucléotides de long (Yu *et al.*, 2018). Finalement, cet outil se démarque, car il permet aussi la prédiction d'interaction entre le potentiel sRNA et des ARNm grâce à RNAplex (Tafer & Hofacker, 2008), RNAup (Mückstein *et al.*, 2006) et IntaRNA (Mann *et al.*, 2017).

Un nouvel outil d'analyse de données transcriptomiques pour l'identification de sRNA a récemment été développé ne se basant pas sur l'uniformité de la couverture. Il a été démontré que les algorithmes basés sur l'uniformité de la couverture tels que *sRNA-Detect* ou ANNOgesic avaient tendance à créer un grand nombre de petits transcrits (Leonard *et al.*, 2019). Cependant, ANNOgesic permet de modifier le seuil de tolérance pour l'uniformité de la couverture des transcrits, et donc de limiter le potentiel impact que ce critère pourrait avoir sur la taille des sRNAs prédits (Yu *et al.*, 2018). APERO (*Analysis of Paired-end RNA-seq Output*) ne tient pas en compte la couverture des transcrits pour identifier de potentiels sRNAs, ce qui lui permet d'identifier le 5′ et le 3′ des candidats avec plus de précision (Leonard *et al.*, 2019).

1.3.1.6 Validation expérimentale des sRNAs candidats

Il existe plusieurs outils afin d'identifier de potentiels sRNAs, mais ultimement, il faut les valider expérimentalement. Les potentiels sRNAs peuvent être confirmés à l'aide de la technique de buvardage par *Northern* (l'appellation anglaise de cette méthode, soit le *Northern blot*, sera utilisée dans cette thèse par simplicité) (Thomas, 1983). Cette technique qui permet de détecter l'ARN sur une membrane est couramment utilisée pour valider et étudier l'expression de sRNAs, étant donné que c'est une méthode accessible, relativement peu coûteuse et quantitative qui peut fournir de l'information à propos de la taille, de l'abondance et le profil d'expressions des ARN (López-Gomollón, 2011). L'ARN est séparé selon sa taille dans un gel de polyacrylamide dénaturant. Des sondes complémentaires aux candidats sRNAs sont créées. La visualisation des sondes marquées à la radioactivité donne une indication des conditions de croissance dans lesquelles les sRNAs candidats sont exprimés (Figure 1.8).

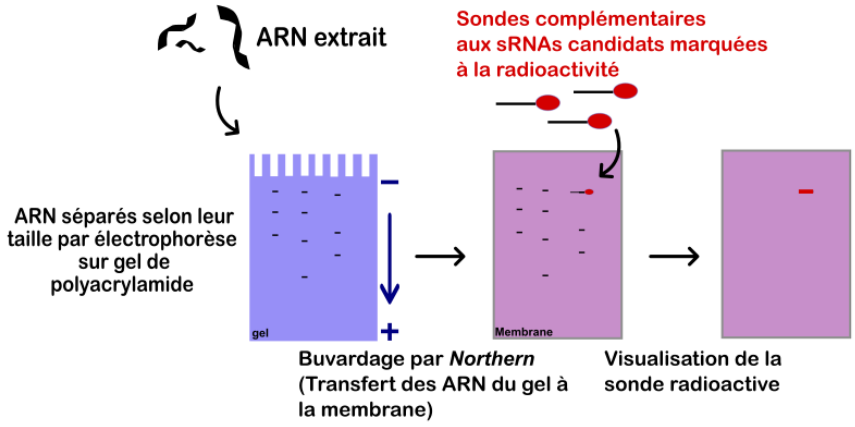

Figure 1.8 Schéma représentant la procédure pour détecter des ARN par *Northern blot*

1.3.2 *Riboswitches*

L'existence d'ARN régulateurs en *trans* tels que les sRNAs peut être confirmée seulement en détectant la présence de l'ARN : la transcription d'une région d'ARN qui n'encode pas pour une protéine suggère qu'elle joue un rôle dans la régulation génétique. Il en est autrement dans le cas des ARN régulateurs en *cis* comme les *riboswitches*, car ceux-ci font partie intégrante des ARNm qu'ils régulent. Même s'ils sont transcrits, il faut aussi évaluer leur capacité à jouer un rôle dans la régulation de l'expression du gène en aval. Une technique telle que le *Northern blot* ne peut donc pas être utilisée pour valider un *riboswitch*.

Un *riboswitch* est composé de deux régions, soit le domaine aptamère et la plateforme d'expression. La liaison d'un ligand au domaine aptamère entraîne un changement dans la structure secondaire de l'ARN, ce qui déclenche ou arrête l'expression du gène en aval, un peu à la manière d'un interrupteur. Les deux conformations possibles d'un *riboswitch* sont mutuellement exclusives en raison d'un chevauchement entre le domaine aptamère et la plateforme d'expression (Serganov & Nudler, 2013) (représentée en bleu, Figure 1.9). Le métabolite est reconnu par le domaine de l'aptamère, une région à la séquence fortement conservée (Serganov & Nudler, 2013). Alternativement, la plateforme d'expression est constituée d'une séquence beaucoup plus variable (Breaker, 2011). Les *riboswitches* sont retrouvés dans la région en 5′ UTR des ARN messager (ARNm) qu'ils contrôlent. Cet arrangement permet au *riboswitch* d'être transcrit en premier, ce qui lui donne le temps de lier son métabolite et ainsi modifier l'expression génétique au besoin (Breaker, 2011).

1.3.2.1 Découverte des *riboswitches*

L'expression des gènes dans une cellule est rigoureusement régulée pour assurer des concentrations appropriées de protéines, d'ARN et de métabolites en tout temps. Les bactéries doivent donc être en mesure de détecter des signaux environnementaux pour adapter leurs réponses cellulaires. Ces riborégulateurs, d'abord appelés « box », ont été découverts sous la forme de régions conservées d'ARN capable de lier des dérivés de vitamines sans avoir besoin de cofacteurs protéiques. Ces premiers exemples de *riboswitches* nommés B_{12} *box*, *thi box* et l'élément RFN peuvent lier spécifiquement l'adénosylcobalamine (AdoCbl) (Nahvi *et al.*, 2002; Vitreschak *et al.*, 2003), la thiamine pyrophosphate (TPP) (Miranda-Ríos *et al.*, 2001; Mironov *et al.*, 2002; Winkler *et al.*, 2002a) et la flavine mononucléotide (FMN) (Gelfand *et al.*, 1999; Mironov *et al.*, 2002; Vitreschak *et al.*, 2002; Winkler *et al.*, 2002b), respectivement. Ces « boîtes » d'ARN

hautement conservées sont situées en amont de gènes importants dans le métabolisme de leurs molécules cibles respectives. Par exemple, le B_{12} *box* retrouvé dans la région en 5′ de l'ARN non traduit (UTR) du gène *btuB* encodant pour un transporteur de la vitamine B12 chez *E. coli* peut lier sélectivement le métabolite AdoCbl, menant à la séquestration du site de reconnaissance du ribosome (RBS) et empêchant ainsi la traduction de se dérouler. Le B12 *box* est donc directement impliqué dans la régulation de la synthèse de la protéine de transport de cobalamine (BtuB) en détectant les niveaux de cette coenzyme (Nahvi *et al.*, 2002).

1.3.2.2 Mode d'action des *riboswitche*s

Les *riboswitches* peuvent avoir un impact autant au niveau de la transcription que de la traduction et ceci pour activer ou réprimer un gène (représenté avec les cadres verts ou rouges respectivement sur la Figure 1.9) (Serganov & Nudler, 2013). La liaison du métabolite peut favoriser l'apparition d'un terminateur de transcription indépendant de la protéine Rho (Figure 1.9*,* A). La structure alternative formée peut aussi libérer un site de liaison de la protéine Rho et ainsi induire une terminaison de transcription (Figure 1.9*,* B) (Serganov & Nudler, 2013). Alternativement, la liaison d'un métabolite peut stabiliser une conformation d'ARN contenant une structure anti-terminatrice, permettant ainsi à la transcription d'avoir lieu (Figure 1.9, C). Un *riboswitch* peut aussi agir au niveau de la traduction en piégeant le site de liaison du ribosome (RBS) dans une structure secondaire (Figure 1.9*,* D). Alternativement, la liaison d'un métabolite peut stabiliser une structure secondaire qui donne accès à la séquence de Shine-Dalgarno (Figure 1.9*,* E) (Serganov & Nudler, 2013).

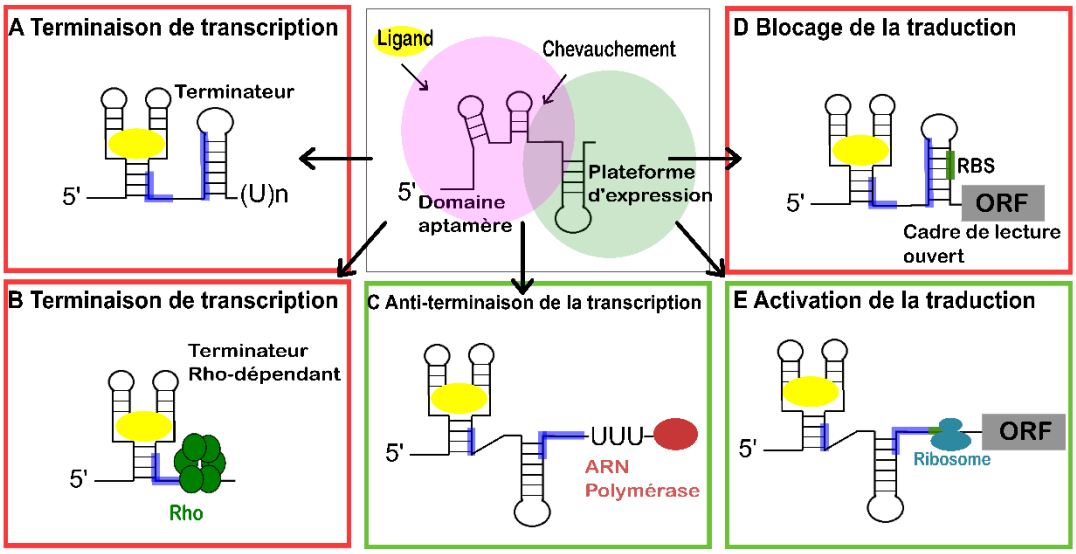

Figure 1.9 Diversité du mode de régulation des *riboswitches* chez les bactéries, inspiré de (Serganov & Nudler, 2013)

La liaison du ligand au domaine aptamère entraine la terminaison de la transcription par un terminateur intrinsèque **(A)** ou par la formation d'un terminateur Rho-dépendant **(B)**. **(C)** En absence du ligand, un terminateur de transcription est formé (en bleu). La liaison du ligand au domaine aptamère permet à la structure alternative de se former (anti-terminateur) permettant à la transcription d'avoir lieu. Les *riboswitches* peuvent aussi avoir un impact au niveau de la traduction. L'association d'un métabolite avec la région de l'aptamère peut engendrer la séquestration ou la libération d'un site de reconnaissance du ribosome (RBS) dans une structure secondaire, menant au blocage **(D)** ou à l'activation de la traduction **(E)**, respectivement. Les régions pouvant s'hybrider pour former les structures alternatives sont représentées en bleu.

D'autres *riboswitches* utilisent des modes d'action plus sophistiqués pour réguler l'expression du gène en aval, comme des réactions d'autoclivage, des aptamères en tandem ou des liaisons coopératives (Serganov & Nudler, 2013). Par exemple, le 5′ UTR de l'ARNm encodant pour la protéine glucosamine-6-phosphate (GlcN6P) synthase a d'abord été identifié comme un potentiel *riboswitch*. Cependant, on n'observe pas un changement de structure secondaire à la suite de la liaison avec son métabolite, mais on remarque plutôt une réaction d'autoclivage qui est accélérée par la présence du ligand. L'ARNm *glmS* est un des deux seuls exemples de *riboswitch* ayant une action catalytique spécifique à la liaison d'un ligand (Klein & Ferré-D'Amaré, 2006). En présence élevée de GlcN6P, le ribozyme *glmS* est activé, menant à une réaction d'autoclivage qui dégrade l'ARNm encodant pour une protéine impliquée dans la synthèse de GlcN6P. D'un autre côté, en absence de GlcN6P, le ribozyme glmS est inactif, permettant à la synthèse de GlcN6P d'avoir lieu (Klein & Ferré-D'Amaré, 2006). La liaison de GlcN6P ne mène pas à un changement de structure comme pour les autres *riboswitches* : GlcN6P agit plutôt comme un cofacteur où le groupement amine est important dans l'activité catalytique du ribozyme (McCarthy *et al.*, 2005).

Le *riboswitch* c-di-GMP-II chez *Clostridium difficile* collabore aussi avec un ribozyme (intron de groupe I) adjacent, où la liaison du ligand au domaine aptamère mène à une réaction d'autoépissage du ribozyme (Figure 1.10) (Lee *et al.*, 2010).

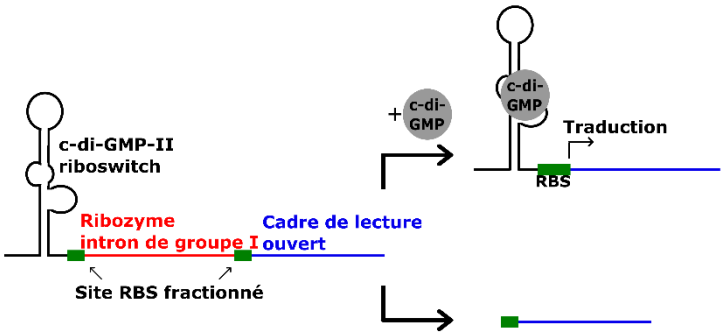

Figure 1.10 Collaboration entre le *riboswitch* c-di-GMP-II et le ribozyme intron de groupe I adjacent, inspiré de (Lee *et al.*, 2010)

La liaison de c-di-GMP au *riboswitch* mène à une réaction d'autoépissage du ribozyme intron de group I (en rouge), ce qui permet la formation d'un site de reconnaissance du ribosome (RBS) (en vert) et la traduction du gène en aval. En absence de c-di-GMP, la réaction d'autoépissage mène à un RBS tronqué.

Certains *riboswitches* ont un arrangement en tandem, où deux *riboswitches* se retrouvent un à la suite de l'autre pour assurer un meilleur contrôle d'un ou des métabolites d'intérêt. Par exemple, le *riboswitch* glycine se retrouve souvent en tandem, avec deux régions aptamères suivies d'une seule plateforme d'expression (Mandal *et al.*, 2004). Les *riboswitches* en tandem peuvent aussi être constitués de deux *riboswitches* distincts, permettant de moduler l'expression en réponse à différents métabolites. C'est le cas pour le tandem SAM-AdoCbl retrouvé dans la région 5′ non-traduite du gène *metE* chez *Bacillus clausii*. Ces *riboswitches* peuvent lier indépendamment les molécules SAM (S-adénosylméthionine) ou la coenzyme B12 respectivement (Sudarsan *et al.*, 2006).

La liaison de métabolites au domaine aptamère d'un *riboswitch* peut aussi se faire de façon coopérative. Par exemple, le *riboswitch* NiCo lie sélectivement les cations nickel (Ni^{2+}) et cobalt (Co^{2+}) de façon coopérative (Furukawa *et al.*, 2015). Il existe aussi des cas où deux ligands identiques lient le même domaine d'aptamère, comme le *riboswitch* liant le tétrahydrofolate (THF) ou le dihydrofolate (DHF), des formes réduites de l'acide folique. Majoritairement retrouvé chez les Firmicutes, il se situe entre autres en amont de gènes qui contrôlent le transport du folate, comme *folT* (Trausch et al., 2011).

1.3.2.3 Méthodes pour la découverte de *riboswitches*

La plupart des *riboswitches* connus ont été découverts à l'aide d'outils bioinformatiques basés sur la génomique comparative, car le domaine d'aptamère de ces molécules régulatrices est très conservé. Les potentiels *riboswitche*s sont identifiés en raison de la présence de conservation des séquences ainsi que de covariation. La covariation représente des changements dans la séquence des nucléotides pour les deux positions impliquées dans une paire de bases qui n'affectent pas la formation de structure secondaire, renforçant ainsi l'importance de la formation de cette structure (Weinberg *et al.*, 2010).

Certaines recherches exploratoires pour la découverte de nouveaux ARNnc sont basées sur la présence d'un promoteur et d'un terminateur dans une région intergénique (Argaman *et al.*, 2001). En revanche, cette technique ne peut être utilisée que chez des bactéries pour lesquelles les mécanismes de transcriptions sont bien compris, comme chez *E. coli*. D'autres études ont simplement regardé la conservation des régions intergéniques pour analyser si les changements dans la séquence étaient plus cohérents avec un ARNnc qu'une région d'ARN encodant pour une protéine (Rivas *et al.*, 2001). Une autre stratégie intitulée GC-IGR est de rechercher pour des segments d'ARN dans les régions intergéniques qui sont riches en G-C dans un génome qui est normalement riche en A-T, car dans ce contexte, les structures secondaires d'ARN ont un pourcentage plus élevé en guanine et cytosine (Klein *et al.*, 2002; Meyer *et al.*, 2009; Schattner, 2002; Stav *et al.*, 2019).

Cette approche basée sur des séquences homologues est limitante, étant donné que l'on pourrait ignorer des séquences ayant trop de conservation. On pourrait aussi passer à côté de *riboswitches* ayant une structure moins complexe ou de séquences régulatrices étant plus rares. On ne peut pas non plus utiliser des stratégies tels que le GC-IGR chez des bactéries comme *Methylorubrum extorquens* qui a un pourcentage en G-C élevé de 68%. Les démarches bioinformatiques pour l'identification de *riboswitches* sont aussi limitées par l'annotation des gènes. Finalement, ce type d'approche ne permet pas d'identifier directement le ligand associé au *riboswitch*, et peut mener à un *riboswitch* orphelin pour lequel il est difficile de déterminer le métabolite correspondant (Corbino *et al.*, 2005; Greenlee *et al.*, 2018; Stav *et al.*, 2019; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2017b; Weinberg *et al.*, 2010) (les *riboswitches* orphelins sont discutés dans la section 1.3.2.6 de cette thèse).

La majorité des *riboswitches* ont été découverts par méthode bioinformatique et annotés grâce à l'homologie de séquences et de structures avec des motifs déjà connus ou par l'alignement de séquences conservées. À l'inverse, la technique *Term-seq* est une approche expérimentale

alternative sans les biais de la bioinformatique afin de découvrir des régulateurs de la terminaison de la transcription chez les bactéries. À la suite de la liaison d'un ligand, un *riboswitch* change de structure secondaire, ce qui entraîne la formation d'un terminateur de transcription, dépendant ou indépendant de la protéine Rho. Un séquençage à haut débit est utilisé afin d'identifier les terminateurs de transcription (*Term-seq*). Pour découvrir un régulateur répondant à un métabolite, il suffit d'incuber la bactérie d'intérêt avec ce ligand et de comparer la longueur des produits de transcription dans ces deux conditions. Un plus grand ratio de petits transcrits à la suite de l'interaction avec le métabolite serait une bonne indication du changement de conformation d'une molécule régulatrice et de l'apparition de terminateurs de transcription (Figure 1.11) (Dar et *al.*, 2016).

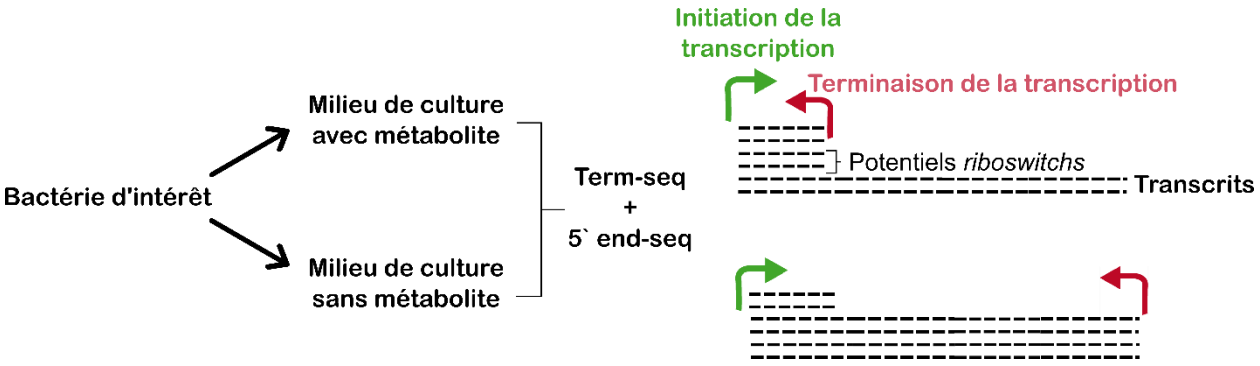

Figure 1.11 Méthode *Term-seq*, inspiré de (Dar *et al.*, 2016)

Approche expérimentale pour détecter les régulateurs basés sur la terminaison de la transcription en réponse à un métabolite. La bactérie d'intérêt est cultivée avec ou sans métabolite. L'ARN est ensuite extrait et séquencé afin d'identifier les sites d'initiation et de terminaison de la transcription. La taille des transcrits entre les deux conditions de croissance est comparée afin d'identifier de potentiels *riboswitches*.

Cependant, cette technique ne peut pas différencier les terminaisons de transcription causée par un *riboswitch* de celle causée par une protéine. On pourrait aussi passer à côté de *riboswitch* ayant plutôt un effet sur l'étape de la traduction ou encore de *riboswitches* dont le promoteur réprimerait l'expression du gène dans les conditions de l'étude.

La méthode PARCEL (*Parallel Analysis of RNA Conformations Exposed to Ligand binding*) quant à elle révèle l'interaction d'un métabolite avec l'ARN en comparant les sites de coupures d'une RNase en présence et en absence du ligand. L'ARN total est extrait des organismes sous différentes conditions de croissance avec ou sans le métabolite. La RNase V1 coupe l'ARN dans les régions où il y a des paires de bases (Figure 1.12). La nucléase S1 peut aussi être utilisée alternativement, mais celle-ci coupe l'ARN simple brin. Les ARN digérés avec la RNase sont utilisés afin de créer deux librairies qui sont comparées pour identifier des différences dans les sites de coupure de l'enzyme en absence et en présence du ligand d'intérêt. Des ADN

complémentaires sont produits à partir des sites de clivages et ceux-ci sont séquencés. Les produits de séquençages sont comparés au génome afin d'identifier les régions qui ont subi un changement de structure à la suite de la liaison du ligand (Tapsin *et al.*, 2018).

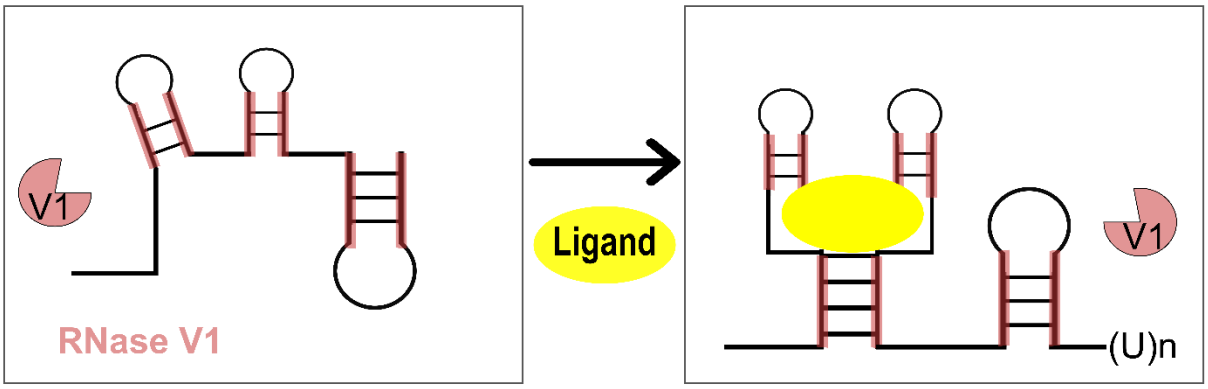

Figure 1.12 Méthode PARCEL (*Parallel Analysis of RNA Conformations Exposed to Ligand binding*), inspiré de (Tapsin et *al*., 2018).

L'endoribonucléase RNase V1 coupe les régions doubles brins de l'ARN. Les produits de dégradation de l'ARN par la RNase V1 en présence ou en absence d'un métabolite diffèrent, dû à un changement de structure secondaire.

Les désavantages de la technique PARCEL sont que nous sommes limités aux régions qui sont hautement exprimées et qui sont accessibles pour les endoribonucléases. De plus, on ne peut chercher que dans un génome à la fois en utilisant les méthodes de PARCEL et de *Term-seq*, alors la recherche de nouveaux *riboswitches* est très ciblée en utilisant qu'un seul ligand à la fois.

La méthode SHAPE (*Selective 2'-hydroxyl acylation analyzed by primer extension)* peut aussi être utilisée afin de détecter des ARN en mesure de lier un ligand (Zeller *et al.*, 2022). L'ARN est traité avec un réactif SHAPE tel que le 1M7 (1-methyl-7-nitroisatoic anhydride) qui modifie l'ARN selon sa structure (Figure 1.13). Les nucléotides dans une région d'ARN simple brin sont modifiées chimiquement par le réactif SHAPE (Watters *et al.*, 2016). Les modifications chimiques de l'ARN sont détectées par séquençage, car cela crée une mutation dans la séquence lors de la transcription inverse (Figure 1.13). L'avantage de cette méthode est donc qu'en plus de détecter la liaison d'un ligand avec l'ARN, elle nous fournit l'information quant à la structure secondaire au nucléotide près (Zeller *et al.*, 2022). Plusieurs séquences peuvent être testées en même temps, étant donné que chacune d'entre elles est identifiée lors du séquençage (Zeller *et al.*, 2022).

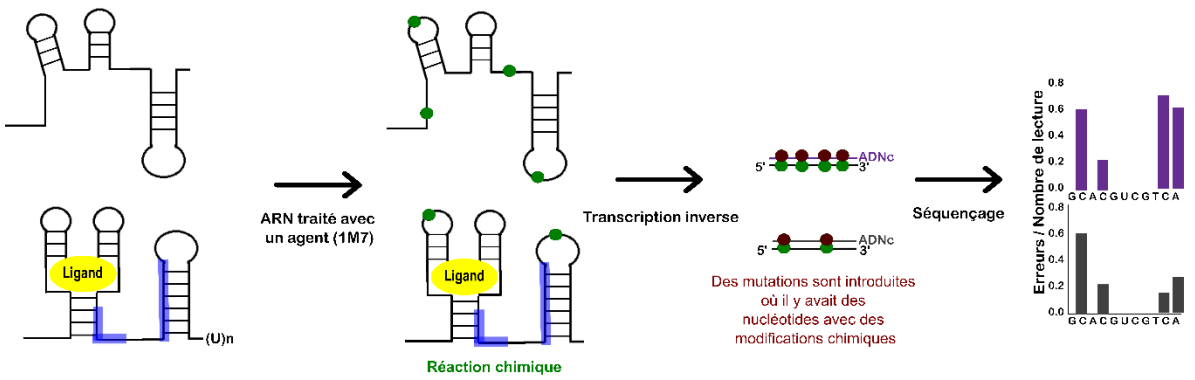

Figure 1.13 Méthode SHAPE pour détecter des ARN en mesure de lier un ligand, inspiré de (Zeller *et al.,* 2022)

1.3.2.4 Caractérisation de *riboswitches*

Lorsqu'un candidat *riboswitch* est identifié, on peut tester l'impact du ligand sur la structure secondaire de l'ARN avec des techniques basées sur des réactions chimiques telles que SHAPE (*Selective 2'-hydroxyl acylation analyzed by primer extension*) (Watters *et al.*, 2016), le DMS-MapSeq (*dimethyl sulfate mutational profiling with sequencing*) (Zubradt *et al.*, 2017) ou le *in-line probing* (Regulski & Breaker, 2008).

Les méthodes SHAPE et DMS-MapSeq suivent le même ordre idée : l'ARN en présence ou en absence du ligand est soumis à des réactions chimiques. L'ARN modifié chimiquement est ensuite converti en ADN complémentaire par transcription inverse. Les différences entre les conditions sont détectées pour discerner l'impact du ligand sur la structure secondaire de l'ARN. La méthode SHAPE a été discutée dans la section 1.3.2.3. Le DMS-MapSeq est une approche basée sur les modifications chimiques de l'ARN par le sulfure de diméthyle (DMS). Le DMS modifie les nucléotides adénosine triphosphate et cytidine triphosphate qui ne sont pas liées. Des changements dans le profil mutationnel entre les conditions en absence ou en présence d'un ligand suggèrent la formation de différentes structures secondaires à la suite de la liaison du métabolite au domaine d'aptamère. Les modifications sont détectées lors de la transcription inverse en utilisant l'enzyme *thermostable group II intron reverse transcriptase* (TGIRT) et des amorces spécifiques à l'ARN d'intérêt. Cet enzyme est en mesure de lire les nucléotides modifiés chimiquement par le DMS et d'introduire dans la séquence de l'ADN complémentaire (ADNc) produite une erreur à la position exacte des adénosines triphosphate et des cytidines triphosphates endogènes (Zubradt *et al.*, 2017). Un changement de structure secondaire entre les deux conditions est représenté par une différence dans le taux d'incorporation d'erreur dans

la séquence des ADNc produits à partir de la transcription inverse des ARNr à la suite du traitement au DMS.

Ces deux techniques basées sur le séquençage sont complexes comparé au procédé basé sur un gel électrophorèse comme le *in-line probing.* Elles nécessitent aussi la transcription inverse, une étape qui pourrait être entravée par la présence d'un ligand. La méthode du *in-line probing* est donc plus couramment utilisée afin d'identifier l'affinité du ligand avec la séquence candidate. Cette méthode est basée sur la dégradation spontanée de l'ARN en condition légèrement alcaline (pH 8.3). Un ARN simple-brin est beaucoup plus vulnérable à la dégradation qu'une région contenant des structures secondaires. La comparaison des profils de dégradation en présence et en absence du ligand peut donc être utilisée afin d'élucider l'interaction d'un ligand avec l'ARN (Figure 1.14). Pour ce faire, il suffit d'incuber le ligand d'intérêt avec l'ARN marqué à la radioactivité. Les molécules sont ensuite migrées sur un gel d'électrophorèse de polyacrylamide dénaturant pour séparer les produits de clivage (Regulski & Breaker, 2008). En exposant les ARN à différentes concentrations du ligand d'intérêt, il est aussi possible d'estimer leur affinité.

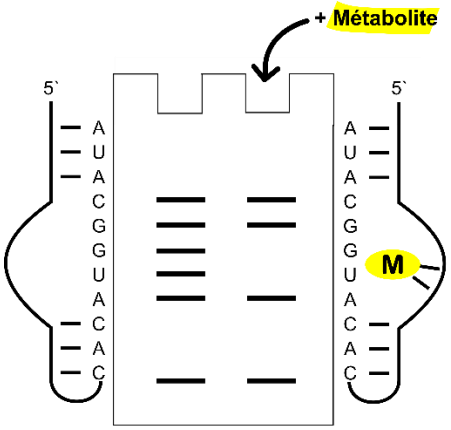

Figure 1.14 Principe de la technique du *in-line probing,* inspiré de (Regulski & Breaker, 2008)

Le profil de dégradation à la suite de la réaction *in-line* ne sera pas le même en présence du ligand, ce qui peut nous donner des indications sur la structure secondaire formée à la suite de la liaison du métabolite.

1.3.2.5 Métabolites et ions régulés par les *riboswitches*

Jusqu'à ce jour, plus de 50 classes de *riboswitches* ont été découvertes, regroupant une liste de ligands variés allant des métabolites aux ions. La majorité des *riboswitches* les plus répandus lie des molécules dérivées d'ARN ou de leurs précurseurs comme les coenzymes TPP (thiamine

pyrophosphate) et SAM (S-adénosylméthionine) et des messagers secondaires comme le di-GMP cyclique et di-AMP cyclique (McCown *et al.*, 2017). Cette observation supporte l'idée que les *riboswitches* pourraient avoir émergé à l'époque où le monde était basé sur l'ARN. Il n'y avait donc pas de protéines qui auraient pu agir comme détecteur de ces métabolites. Les *riboswitches* les plus communs, comme ceux de TPP et de SAM, sont ceux qui ont été découverts parmi les premiers, simplement parce qu'il y avait une plus haute probabilité qu'ils soient découverts dû à leur grande distribution. À l'inverse, plusieurs classes de *riboswitch* sont très rares, représentant peut-être une nouvelle apparition due à la sélection naturelle ou, inversement, désignant d'anciens *riboswitches* qui sont dirigés vers la limite de l'extinction. Le *riboswitch* preQ1-III liant la molécule preQ1 (pre-queuosine 1) est un bon exemple de structures régulatrices rares, où seulement 86 représentants de cet ARN capable de lier le précurseur du nucléoside queuosine sont connus (McCown *et al.*, 2014). La vaste majorité des classes de *riboswitch* sont en mesure de lier des métabolites, soit des cofacteurs d'enzyme, des précurseurs de nucléotide, des acides aminés ou des molécules signales. Seulement quatre classes sont connues pour avoir des cations divalents comme ligand, soit Mn^{2+}, Mg^{2+} (deux classes), Ni^{2+} et Co^{2+} (où Ni^{2+} et Co^{2+} représente une seule classe), et une seule classe reconnaît un anion (F^-) (Tableau 1.1). Par contre, une récente étude a démontré que le *riboswitch czcD* (NiCo) subissait un changement de structure secondaire à la suite d'une liaison avec le Fe^{2+} (Xu & Cotruvo Jr, 2022). Le motif DUF1646 a aussi récemment été démontré comme un *riboswitch* reconnaissant le Na^+ (White et al., 2022). Les métabolites qui sont reconnus par plusieurs classes de *riboswitches* sont représentés en gras dans le tableau ci-dessous (Tableau 1.1).

Tableau 1.1 Ligands reconnus par des *riboswitches*

Cofacteur d'enzyme	Précurseur de nucléotide	Acide aminé	Éléments	Molécules signales	Autres
AdoCbl, AqCbl, DHF, FMN, MeCbl, Riboflavine, Moco, Tuco, **NAD^+**, **SAH**, **SAM**, **THF**, TPP	Adénine, ADP, dADP, CDP, dCDP, Guanine, HMP-PP, **PreQ1,** PRPP, **Xanthine**, **2′-dG**, Acide urique, PRA	**Glutamine**, Gycine, Lysine	Co^{2+}, F^-, Mn^{2+}, **Mg^{2+}**, Ni^{2+}, **Na^{2+}**, Fe^{2+}	c-AMP-GMP, c-di-AMP, **c-di-GMP**, ppGpp, ZMP, ZTP	Antibiotique Aminoglycoside, Azaaromatique, GlcN6P, **Guanidine**

Les métabolites qui sont reconnus par plusieurs classes de *riboswitches* sont indiqué en gras. Références : Acide urique (Yu & Breaker, 2020); adénosine diphosphate/ADP; désoxyadénosine diphosphate / dADP; Cytidine diphosphate / CDP ; désoxycytidine diphosphate / dCDP (Sherlock *et al.*, 2018a); Adénine (Mandal & Breaker, 2004); AdoCbl / adénosylcobalamine (Nahvi et al., 2002); antibiotique aminoglycoside (Jia *et al.*, 2013); AqCbl / aquocobalamine (Johnson Jr et al., 2012); composés azaaromatiques (Li *et al.*, 2016; Weinberg *et al.*, 2010); *Cyclic guanosine monophosphate–adenosine monophosphate* / c-AMP-GMP (Weinberg et al., 2007); Cyclic Di-adenosine Monophosphate */c*-di-AMP (Nelson *et al.*, 2013); *Cyclic diguanylate monophosphate* / c-di-GMP (Sudarsan et al., 2008); cobalt / Co^{2+}(Furukawa *et al.*, 2015); cofacteur tungstène / Tuco (Regulski *et al.*, 2008; Weinberg *et al.*, 2007); dihydrofolate / DHF (Trausch *et al.*, 2011); Fer/Fe^{2+}(Xu & Cotruvo Jr, 2022); fluor / F^-(Baker *et al.*, 2012; Ren *et al.*, 2012); flavine mononucléotide / FMN (Gelfand et al., 1999; Mironov et al., 2002; Vitreschak et al., 2002; Winkler et al., 2002b); glucosamine-6-phosphate / GlcN6P (Hampel & Tinsley, 2006; Jones & Ferré-D'Amaré, 2015; Weinberg *et al.*, 2010); glutamine (Ames & Breaker, 2011; Weinberg *et al.*, 2010); glycine (Mandal et al., 2004); guanidine (Nelson *et al.*, 2017; Sherlock & Breaker, 2017; Sherlock *et al.*, 2017); guanine (Mandal *et al.*, 2003); 4-amino-5-hydroxymethyl-2-methylpyrimidine diphosphate / HMP-PP (Atilho *et al.*, 2019a); lysine (Grundy *et al.*, 2003; Mandal *et al.*, 2003; Rodionov *et al.*, 2003; Sudarsan *et al.*, 2003); manganèse / Mn^{2+} (Argaman *et al.*, 2001; Barrick *et al.*, 2004; Dambach *et al.*, 2015; Price *et al.*, 2015); Magnésium / Mg^{2+} (Barrick *et al.*, 2004; Ramesh & Winkler, 2010); Méthylcobalamine / MeCbl (Johnson Jr et al., 2012); cofacteur de molybdène / Moco (Regulski *et al.*, 2008; Weinberg *et al.*, 2007); NAD+ (Malkowski *et al.*, 2019; Weinberg *et al.*, 2017a; Panchapakesan *et al*, 2021); nickel / Ni^{2+} (Furukawa *et al.*, 2015); phosphoribosylamine / PRA (Malkowski *et al.*, 2020); pré-queuosine1 / preQ1 (Kang *et al.*, 2014; Liberman *et al.*, 2015; McCown *et al.*, 2014; Reader *et al.*, 2004; Weinberg *et al.*, 2007); guanosine tétraphosphate / ppGpp (Sherlock *et al.*, 2018b); phosphoribosylpyrophosphate / PRPP (Sherlock *et al.*, 2018c); riboflavine (Atilho *et al.*, 2019b); S-adénosylhomocystéine / SAH (Wang *et al.*, 2008); S-adénosylméthionine / SAM (Winkler et al., 2003); sodium/Na^+ (White *et al.*, 2022); Tétrahydrofolate / THF (Trausch *et al.*, 2011); thiamine pyrophosphate / TPP (Miranda-Ríos et al., 2001; Mironov et al., 2002; Winkler et al., 2002a); xanthine(Hamal Dhakal *et al.*, 2022; Yu & Breaker, 2020); *5-amino-4-imidazole carboxamide riboside 5′-triphosphate* / ZTP ; *5-aminoimidazole-4-carboxamide ribonucleotide* / ZMP (Jones & Ferré-D'Amaré, 2015; Kim *et al.*, 2015; Weinberg *et al.*, 2010); 2'-désoxyguanosine / 2'-dG (Kim *et al.*, 2007; Weinberg et al., 2017b).

Étant donné que l'ARN est chargé négativement, les cations peuvent neutraliser la charge négative de ligands afin de limiter la répulsion avec l'ARN et favoriser leurs liaisons. Cette fonction est souvent assumée par le cation Mg^{2+} en raison de sa grande concentration cellulaire, comme c'est entre autres le cas pour les *riboswitche*s liant le TPP (Serganov *et al.*, 2006) et le FMN (Serganov *et al.*, 2009). Par contre, la liaison de l'acide aminé lysine à son *riboswitch* correspondant est médiée par la présence du cation K^+ (Serganov *et al.*, 2008).

1.3.2.6 Les *riboswitches* orphelins

La majorité des *riboswitches* portent le même nom que leur ligand correspondant. Certains sont cependant nommés selon le gène retrouvé en aval, parce que le métabolite reconnu par le domaine d'aptamère est inconnu. On les appelle des *riboswitches* orphelins et ils sont découverts

grâce à des méthodes bioinformatiques nécessitant la comparaison de séquence. Par exemple, la présence de covariation dans la séquence de nucléotides est un excellent indice de l'importance de la structure secondaire, parce que malgré un changement dans l'identité d'un nucléotide à une certaine position, la structure secondaire est conservée.

L'identification d'un ligand pour un candidat *riboswitch* est souvent facilitée par la nature du gène en aval. Cependant, l'annotation des gènes ne fournit pas toujours l'information nécessaire afin d'identifier le métabolite, et elle est même parfois manquante ou trompeuse. Il y a actuellement une longue liste de *riboswitches* orphelins pour lesquelles un ligand est recherché (Corbino *et al.*, 2005; Greenlee *et al.*, 2018; Stav *et al.*, 2019; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2017b; Weinberg *et al.*, 2010). Le nombre de *riboswitches* orphelins risque d'augmenter au fur et à mesure que des recherches pour de nouveau candidat sont effectuées par méthode bioinformatique.

Certains candidats se retrouvent en amont de divers gènes impliqués dans une grande variété de processus qui peuvent, à première vue, ne pas sembler être liés les uns avec les autres, ce qui complique l'identification d'un ligand. Lorsque le contexte génomique n'est d'aucune aide, une large gamme de métabolites peuvent être testés avec une approche non biaisée afin de découvrir le ligand potentiel. Cependant, ce type d'approche n'est pas toujours fructueuse, car il faut déjà que le ligand reconnu par le domaine aptamère se retrouve dans la banque de candidats de départ. De plus, ce ne sont pas tous ces motifs d'ARN orphelins qui sont réellement des *riboswitches* : certains pourraient être un autre type d'élément de régulation en *cis* comme des thermorégulateurs d'ARN qui répondent à des changements de températures.

Au fils des années, les ligands de certains anciens *riboswitches* orphelins ont été découverts (Tableau 1.2), mais quelques-uns sont toujours identifiés avec leur nom d'origine au lieu d'être désignés par le nom de leur métabolite respectif.

Tableau 1.2 Ancienne classe de *riboswitches* orphelins, inspiré de (Sherlock & Breaker, 2020)

Découverte du motif d'ARN	Nom du motif d'ARN	Ligand	Validation du ligand
(Barrick *et al.*, 2004)	*gcvT*	Glycine	(Mandal *et al.*, 2004)
	ydaO	c-di-AMP	(Nelson *et al.*, 2013)
	ykkC sous-type 1	Guanidine	(Nelson *et al.*, 2017)
	ykkC sous-type 2a	ppGpp	(Sherlock *et al.*, 2018b)
	ykkC sous-type 2b	PRPP	(Sherlock *et al.*, 2018c)
	ykkC sous-type 2c	(d)ADP/(d)CDP	(Sherlock *et al.*, 2018a)
	ykoK	Mg^{2+}	(Dann III *et al.*, 2007)
	ykvJ	$preQ_1$	(Roth *et al.*, 2007)
	yybP	Mn^{2+}	(Dambach *et al.*, 2015; Price *et al.*, 2015)
(Stav *et al.*, 2019)	*mepA*	Guanidine	(Salvail *et al.*, 2020)
	thiS	HMP-PP	(Atilho *et al.*, 2019a)
(Weinberg *et al.*, 2007)	GEMM (*Gene for the Environment, Membranes and Motility*)	c-di-GMP	(Sudarsan *et al.*, 2008)
	Mini-*ykkC*	Guanidine	(Sherlock *et al.*, 2017)
(Weinberg *et al.*, 2010)	*crcB*	F^-	(Baker *et al.*, 2012)
	pfl	ZMP/ZTP	(Kim *et al.*, 2015)
	yjdF	Composés azaaromatiques	(Li *et al.*, 2016)
	ykkC-III	Guanidine	(Sherlock & Breaker, 2017)
(Weinberg *et al.*, 2017a)	*folE*	THF	(Chen *et al.*, 2019)
	nadA	NAD+	(Malkowski *et al.*, 2019)
	NMT1	Xanthine	(Yu & Breaker, 2020)
	Motif DUF1646	Na^+	(White *et al.*, 2022)

1.3.3 Ribozymes

La versatilité des capacités biochimiques de l'ARN ne s'arrête pas à des fonctions de récepteurs comme c'est le cas chez les *riboswitches :* l'ARN peut aussi catalyser des réactions chimiques, une découverte effectuée de façon indépendante par les laboratoires de Thomas Cech et de Sidney Altman (Guerrier-Takada *et al.*, 1983; Kruger *et al.*, 1982), ce qui leur a valu le prix Nobel de chimie en 1989. Depuis cette avancée, plusieurs types de ribozymes ont été identifiés, incluant l'ARNase P (ribonucléase P) (Guerrier-Takada *et al.*, 1983), les introns de groupe I (Kruger *et al.*, 1982) et II (Michel & Ferat, 1995), les ribozymes en tête de marteau (*hammerhead ribozymes*) (Buzayan *et al.*, 1986; Prody *et al.*, 1986), les ARN du spliceosome (Staley & Guthrie, 1998), les

ARN ribosomaux (Cech, 2000), le ribozyme du virus HDV (*Hepatitis Delta Virus*) (Sharmeen *et al.*, 1988), le ribozyme VS (*Varkud satellite*) (Saville & Collins, 1990), les ribozymes *twister* (Roth et al., 2014), les ribozymes *pistol (Roth et al., 2014),* les ribozymes *hatchet (Weinberg et al., 2015),* les ribozymes *hairpin* (Butcher & Burke, 1994) et le ribozyme *glmS* (Winkler et al., 2004). Depuis leurs découvertes chez les viroïdes de plantes à la fin des années 80 (Buzayan *et al.*, 1986; Prody *et al.*, 1986), les ribozymes *hammerheads* sont les ARN capables d'autoclivage les plus étudiés (De la Peña *et al.*, 2017) (afin de faciliter la lecture, le nom en anglais des ribozymes à tête de marteau sera utilisé dans cette thèse, soit ribozyme *hammerhead*).

1.3.3.1 Structure des ribozymes *hammerheads*

La forme minimale du ribozyme *hammerhead* est composée d'un cœur catalytique comportant 15 nucléotides hautement conservés entourés de trois tiges (I, II et III) (Figure 1.15) (De la Peña *et al.*, 2017). Leur structure ressemble à la forme de la tête des requins-marteaux, d'où leur nom est inspiré. Il existe trois types de ribozymes *hammerhead*, tout dépendamment quelle tige est ouverte (Figure 1.15 A, B and C).

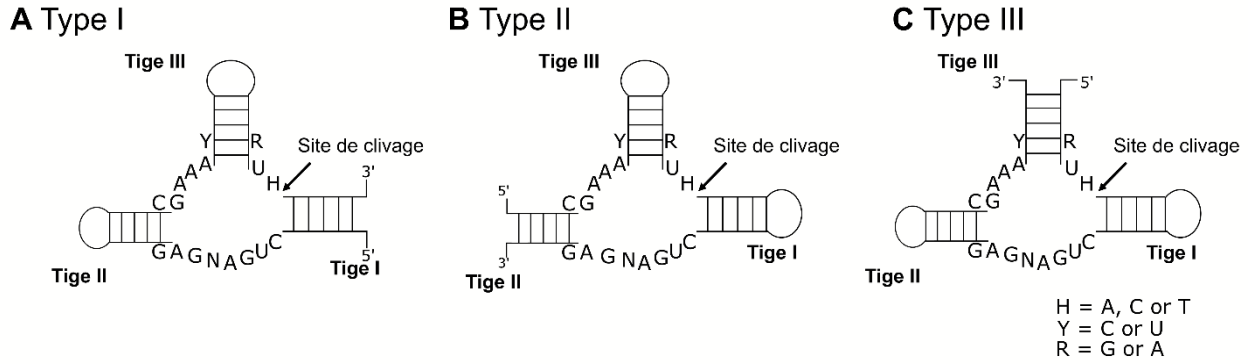

Figure 1.15 Structure des ribozymes *hammerheads*, inspiré de (De la Peña *et al.*, 2017)

Les nucléotides conservés du cœur catalytique et le site de clivage du ribozyme *hammerhead* sont illustrés. Les trois types de ribozymes *hammerheads* dépendent de quelle tige est ouverte. Le nombre de paires de bases dans les tiges est un schéma et ne représente pas la taille exacte de ces tiges.

Cette forme minimale est essentielle à l'action catalytique des ribozymes *hammerheads*. Cependant, le ribozyme complet comprend une interaction entre la tige I et II, ce qui augmente la vitesse de réaction d'un facteur de près de 1000 (Martick & Scott, 2006). La présence de cation est cruciale pour l'action des ribozymes. En raison de sa grande concentration dans la cellule, les réactions de clivage sont souvent dépendantes des ions Mg^{2+} pour stabiliser les structures secondaires et tertiaires du ribozyme. Cependant, les réactions d'autoclivage peuvent aussi avoir

lieu en présence de différents cations, comme le Mn^{2+}. Dans une récente étude de notre laboratoire, nous avons démontré qu'une mutation naturelle dans le cœur catalytique normalement hautement conservé d'un ribozyme *hammerhead* (A6C) favorise la réaction en présence du Mn^{2+} au lieu du Mg^{2+} (Naghdi, Boutet *et al*, 2020). De plus, si cette mutation est reproduite dans un autre ribozyme *hammerhead*, le même changement de spécificité est observé du magnésium au manganèse (Naghdi, Boutet *et al*, 2020).

1.3.3.2 Ribozyme *hammerhead* fonctionnel en *trans*

Il est possible de modifier la forme minimale des ribozymes *hammerhead*s afin qu'ils agissent en *trans* et coupent un ARN d'intérêt : il suffit de modifier la séquence de la tige I et III d'un ribozyme *hammerhead* de type I ou III afin qu'elles soient complémentaires à un ARN ciblé (Figure 1.16) (Najeh, 2017). De plus, il faut que l'ARN ciblé contienne la séquence GUC (ou à tout le moins NUH), soit le site de clivage reconnu par les ribozymes *hammerheads* (Figure 1.16).

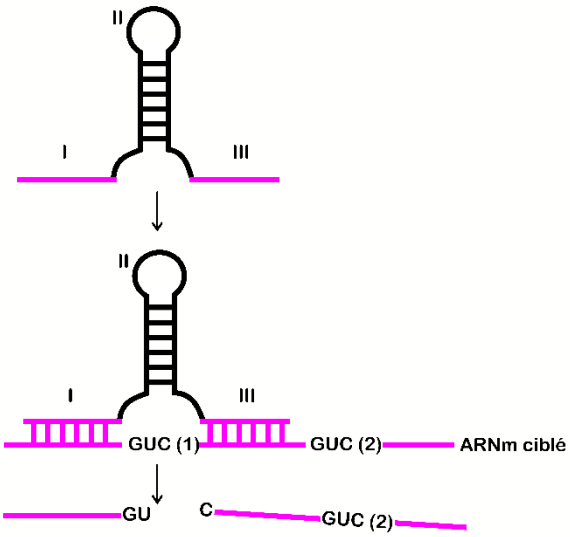

Figure 1.16 Structure d'un ribozyme *hammerhead* agissant en *trans,* inspiré de (Najeh, 2017)

Ribosoft 2.0 est un service web accessible au public qui peut être utilisé pour concevoir entre autres des ribozymes *hammerhead* ciblant un ARN d'intérêt. Une liste de ribozymes potentiels est créée à partir de la séquence d'ARN d'intérêt, prenant en compte de multiples critères tels que la structure du ribozyme, l'interaction tertiaire et l'accessibilité de la cible pour n'en citer que quelques-uns (Kharma *et al.*, 2016). L'efficacité des ribozymes créés par l'outil Ribosoft 2.0 a été validé *in vitro* et *in vivo* : les ribozymes pouvaient inhiber l'expression de la protéine RFP chez *E. coli* (Najeh, 2017). Cet outil pourrait donc être utilisé afin de créer des ribozymes ciblant un gène d'intérêt chez d'autres bactéries comme *M. extorquens* (cette approche sera discutée en

perspective de cette thèse à la section 6.1.2). Comme mentionné auparavant, une mutation dans le cœur catalytique du ribozyme *hammerhead* (A6C) change son activité de clivage en présence d'un ion, soit du magnésium au manganèse dans le cas démontré, ce qui témoigne de la versatilité des ribozymes comme outils de régulation génétique (Naghdi, Boutet *et al*, 2020). Le cœur catalytique d'un ribozyme synthétique pourrait par exemple être ciblé afin de modifier dans quelle condition l'activité de clivage aura lieu.

2 PROBLÉMATIQUE, HYPOTHÈSES ET OBJECTIFS

Il y a un intérêt d'utiliser *Methylorubrum extorquens* comme outil biotechnologique pour produire des produits à valeur ajoutée à partir de méthanol. Il serait important de comprendre les systèmes de régulation génétique déjà présents chez cet organisme modèle. Le rôle des ARNnc chez cette méthylotrophe en mesure de consommer les C1 est peu étudié. L'approfondissement de nos connaissances des ARNnc chez *M. extorquens* pourrait révéler des aspects jusqu'alors insoupçonnés de la régulation du métabolisme des C1, avec des conséquences intéressantes sur l'utilisation biotechnologique de cette bactérie. Nous nous sommes particulièrement intéressées aux sRNAs et aux *riboswitches* chez *M. extorquens*, car plusieurs étaient annotés dans son génome, sans jamais avoir été validés en laboratoire auparavant.

Les sRNAs jouent un rôle important dans la cellule et ils sont souvent le lien manquant afin de comprendre la croissance des bactéries dans différentes conditions ou le métabolisme de certains composés. Nous avions comme hypothèse préliminaire qu'il reste de nombreux sRNAs à découvrir chez *M. extorquens.* Cependant, avant de nous lancer dans la validation de sRNA chez cet organisme modèle, nous voulions tout d'abord établir plus rigoureusement une représentation de l'occurrence des sRNAs dans les génomes bactériens dans le but de souligner le potentiel de découverte de nouveaux sRNAs, ce qui justifierait la recherche de ce type d'ARN régulateur chez *M. extorquens*. ***Bien que des avancées majeures dans la recherche associée aux sRNAs aient eu lieu au cours des dernières années, nous n'avons peut-être qu'effleuré la surface de l'étendue des sRNAs retrouvés dans les génomes bactériens.*** (**Hypothèse 1** étudiée par l'objectif 1 discuté dans le chapitre 1 et article 1 de cette thèse).

Après avoir démontré qu'il y a potentiellement encore un grand nombre de sRNAs qui restent à découvrir chez les bactéries, d'autant plus si on se concentre sur des organismes qui sont moins souvent sous la loupe des chercheurs, nous partirons à la recherche de sRNAs candidats chez *M. extorquens*. La protéine Hfq a été démontrée comme importante chez approximativement un tiers des sRNAs connus chez *E. coli* (Zhang *et al.*, 2003). Étant donné que *M. extorquens* est aussi une protéobactérie comme *E. coli* et qu'elle encode également pour la protéine Hfq, ***nous émettons l'hypothèse qu'il y a plusieurs sRNAs naturellement présents dans son génome, mais que ceux-ci n'ont simplement pas été découverts à ce jour*** (**Hypothèse 2** étudiée par l'objectif 2, discutée dans le chapitre 2 et article 2 de cette thèse).

Tout comme pour les sRNAs, plusieurs indices suggèrent que plusieurs *riboswitches* demeurent à découvrir. Ces ARN sont d'importants régulateurs des voies métaboliques et ils sont aussi des

médiateurs des changements physiologiques dans la cellule. Plusieurs *riboswitches* sont annotés dans le génome de *M. extorquens*, tels que ceux régulant le fluor, la cobalamine, la guanidine, la glycine et le TPP. La plupart des *riboswitches* ont été découverts grâce à des outils basés sur la génomique comparative (Barrick *et al.*, 2004; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2017b). Ceux déjà annotés dans le génome de *M. extorquens* sont le résultat de ces études et correspondent à des *riboswitches* largement répandus. Les recherches exploratoires pour la découverte de nouveaux *riboswitches*, potentiellement plus nichés (qui pourraient, par exemple, être spécifiques au métabolisme des C1), sont entre autres basées sur l'annotation des gènes, de même que sur la présence d'un promoteur et/ou d'un terminateur, recherches qui ne sont pas adéquates chez une bactérie où ces derniers sont peu annotés comme c'est le cas pour *M. extorquens.* De plus, d'autres approches bioinformatiques récemment utilisées avec succès pour la découverte de *riboswitches* ne peuvent pas être utilisées chez les bactéries riches en G-C tels que *M. extorquens*. En effet, ces stratégies recherchent habituellement des régions d'ARN riches en G-C dans un génome qui est normalement riche en A-T, afin d'identifier des structures secondaires (Klein *et al.*, 2002; Meyer *et al.*, 2009; Schattner, 2002; Stav *et al.*, 2019). Les méthodes expérimentales existantes, comme PARCEL (Tapsin *et al.*, 2018) ou Term-seq (Dar *et al.*, 2016) par exemple, ont quant à elles leurs propres limitations : la recherche est très délimitée, en ciblant un génome à la fois pour un métabolite d'intérêt, le tout avec une forte dépendance pour les niveaux d'expression des gènes dans les conditions de croissance choisies.

Afin de pallier les désavantages des méthodes mentionnées, nous avons développé la technique du SR-PAGE qui identifie une molécule régulatrice en tirant parti du changement de structure secondaire à la suite de la liaison du ligand. ***Nous émettons l'hypothèse que le changement de structure à la suite de l'ajout de ligand dans un gel natif permettra d'identifier des riboswitches pour une variété de ligands*** (***Hypothèse* 3** étudiée par l'objectif 3 discutée dans le chapitre 3 et article 3 de cette thèse). Avant d'être en mesure d'utiliser la technique du SR-PAGE dans la recherche de nouveaux *riboswitches*, nous optimiserons et validerons d'abord cette nouvelle méthode avec des *riboswitches* connus. Ensuite, nous démontrerons que le SR-PAGE peut être utilisé comme outil de sélection au sein d'un SELEX, ce qui confirmerait que cette méthode peut être utilisée pour la recherche de nouveaux *riboswitches* à partir d'une librairie de séquences d'ARN.

2.1 Objectifs

L'objectif général de mon projet de doctorat est d'avoir une meilleure compréhension du rôle des ARN régulateurs chez *M. extorquens*, plus spécifiquement les sRNAs et les *riboswitches*. Afin de répondre aux hypothèses établies, les objectifs suivants ont été émis :

Objectif 1 : Prévalence des sRNAs chez les bactéries

- Établir une représentation plus précise de l'occurrence des sRNAs chez les bactéries, en soulignant le potentiel de découverte de nouveaux sRNAs à l'aide d'une approche bioinformatique

Objectif 2 : Identification de sRNAs potentiels chez *M. extorquens*

- Analyser les résultats d'une étude transcriptomique chez *M. extorquens* afin de définir une liste de candidats potentiels
- Confirmer expérimentalement les sRNAs prédits
- Caractériser les candidats (prédiction de structures secondaires conservées et d'ARNm ciblés, évaluation de leur contexte génomique et de leur expression dans différentes conditions de croissances)

Objectif 3 : Développement de la technique du SR-PAGE

- Valider et optimiser la technique du SR-PAGE
- Sélectionner des *riboswitches* à affinité modifiés (démontrer que le SR-PAGE peut être utilisé comme méthode de sélection au sein d'un SELEX)

3 CHAPITRE 1: SMALL RNAS BEYOND MODEL ORGANISMS: HAVE WE ONLY SCRATCHED THE SURFACE?

Les sRNAs au-delà des organismes modèles: avons-nous qu'effleuré la surface? (Traduction française)

Auteurs :

Emilie Boutet [1], Samia Djerroud [1], Jonathan Perreault [1]

[1] INRS- Centre Armand-Frappier Santé Biotechnologie, Laval, QC, H7V 1B7, Canada

Journal : *International Journal of Molecular Sciences* (IJMS) (journal révisé par les pairs)

Soumission : 21 mars 2022

Acceptation : 15 avril 2022

Publication : 18 avril 2022

DOI : https://doi.org/10.3390/ijms23084448

Contribution des auteurs :

Emilie Boutet : Réalisation de la recherche bioinformatique, analyse des résultats, préparation des figures et tableaux, révision de la littérature, rédaction du manuscrit

Samia Djerroud : Mise à jour de la base de données nécessaire à la recherche bioinformatique, aide dans l'analyse des résultats

Jonathan Perreault : Supervision et révision du manuscrit

3.1 Liens entre les objectifs et ce chapitre de la thèse

Avant de nous lancer dans la recherche de nouveaux sRNAs chez *M. extorquens*, nous voulions d'abord justifier cette exploration en démontrant le potentiel de découvertes de sRNAs chez les bactéries en général. L'article sous forme de perspective retrouvé dans ce chapitre répond à l'objectif 1 de ma thèse. Le choix du type « perspective » dans ce cas-ci se trouve à être à mi-chemin entre un article de recherche et un article de revue de littérature. En effet, bien que certains aspects de cet article s'apparentent à une revue de littérature, les analyses génomiques assez poussées que nous y avons incluses l'en distinguaient.

Ce chapitre ne se centralise pas sur notre organisme d'intérêt *M. extorquens*, car nous voulions d'abord établir la pertinence de s'intéresser aux ARNnc des bactéries. L'objectif de ce premier chapitre est donc de mieux supporter notre problématique que les ARNnc chez les bactéries en général sont peu étudiés et que l'on gagnerait à s'y attarder. Avec les progrès des technologies de séquençages transcriptomiques, des nouveaux ARNnc sont découverts, révélant des fonctions uniques. L'étude des ARNnc a permis entre autres d'élucider plusieurs mécanismes de virulences et de pathogénicité bactériens. Par exemple, la bactérie *Streptococcus pneumoniae* responsable de la pneumonie utilise des thermosenseurs d'ARN afin de détecter des changements de températures, ce qui lui permet de contrôler l'expression des gènes pour survivre à ce nouvel environnement (Eichner *et al.*, 2021). Une meilleure compréhension du rôle des ARNnc dans les maladies causées par des bactéries pourrait donc permettre le développement de nouveaux agents antibactériens ciblant l'ARN par exemple (Eichner *et al.*, 2022).

Pour des bactéries au potentiel biotechnologique tel que *M. extorquens*, il serait important de comprendre les moyens que ces bactéries utilisent déjà pour être en mesure de modifier l'expression génétique à notre avantage en étudiant les ARNnc, généralement très impliqués dans le contrôle de l'expression. Ceux-ci pourraient être utilisés comme outils de régulation génétique ou comme cible de l'ingénierie génétique pour arriver à la production d'un composé d'intérêt. À d'autres occasions, l'identification de plusieurs cibles pour un même sRNA pourrait aider à comprendre des liens jusqu'alors insoupçonnés entre différentes voies métaboliques.

Ce premier chapitre permet donc de justifier l'importance de réaliser des projets de recherches en lien avec l'exploration et la découverte des ARNnc comme ceux effectués dans le cadre de cette thèse (chapitre 2 et 3). L'étendue de nos connaissances sur les ARNnc des bactéries, plus précisément les sRNAs, y est exposée, mettant l'emphase sur le fait que la majorité de celles-ci proviennent de recherches associées aux organismes modèles.

3.2 Résumé (traduction française)

Les petits ARN (sRNAs) sont des régulateurs clés dans l'adaptation des bactéries aux changements environnementaux et agissent en se liant aux ARNm ciblés par complémentarité de base. Environ 550 familles distinctes de sRNA ont été identifiées depuis leur première caractérisation dans les années 1980, accélérée par l'émergence du séquençage des ARN. Les sRNAs sont présents dans un large éventail de phyla bactériens, mais ils sont plus apparents dans les organismes modèles très étudiés par rapport au reste des bactéries séquencées. En effet, *Escherichia coli* et *Salmonella enterica* contiennent le plus grand nombre de sRNAs annotés, avec 98 et 118, respectivement. Les *Enterobacteriaceae* encodent pour 145 sRNAs distincts, alors que les autres familles de bactéries n'ont que sept sRNAs annotés en moyenne. Bien que les dernières années aient été marquées par des avancées majeures dans la recherche sur les sRNAs, nous n'avons peut-être fait qu'effleurer la surface, d'autant plus que les annotations des ARN sont à la traîne par rapport aux annotations des gènes. Une tendance distinctive peut être observée pour les gènes, où leur nombre augmente avec la taille du génome, mais cela n'est pas observable pour les ARN, bien qu'on s'attende à ce qu'ils suivent la même tendance. Dans cette perspective, nous avons cherché à établir une représentation plus précise de l'occurrence des sRNAs chez les bactéries, en soulignant le potentiel de découverte de nouveaux sRNAs.

3.3 Abstract

Small RNAs (sRNAs) are essential regulators in the adaptation of bacteria to environmental changes and act by binding targeted mRNAs through base complementarity. Approximately 550 distinct families of sRNAs have been identified since their initial characterization in the 1980s, accelerated by the emergence of RNA-sequencing. Small RNAs are found in a wide range of bacterial phyla, but they are more prominent in highly researched model organisms compared to the rest of the sequenced bacteria. Indeed, *Escherichia coli* and *Salmonella enterica* contain the highest number of sRNAs, with 98 and 118, respectively, with *Enterobacteriaceae* encoding 145 distinct sRNAs, while other bacterial families have only seven sRNAs on average. Although the past years brought major advances in research on sRNAs, we have perhaps only scratched the surface, even more so considering RNA annotations trail behind gene annotations. A distinctive trend can be observed for genes, whereby their number increases with genome size, but this is not observable for RNAs, although they would be expected to follow the same trend. In this

perspective, we aimed at establishing a more accurate representation of the occurrence of sRNAs in bacteria, emphasizing the potential for novel sRNAs discoveries.

3.4 Introduction

Small RNAs (sRNAs) are important post-transcriptional regulators involved in many cellular mechanisms such as biofilm formation, adaptation to environmental changes and virulence (Jørgensen *et al.*, 2020). They modulate gene expression by base pairing with their target mRNA either with perfect (*cis*-acting) or partial (*trans*-acting) complementarity. *Cis*-acting sRNAs (better known as antisense RNAs; asRNAs) are encoded in the opposing strand of their target mRNAs, whereas *trans*-acting sRNAs are in a different locus. The latter tend to target multiple mRNAs and often rely on the help of chaperone proteins such as Hfq or ProQ in Gram-negative bacteria (Watters *et al.*, 2016). Here, we focused on *trans*-acting sRNAs, though a similar analysis dedicated to asRNAs is available in the Supplementary Material (Supplementary Material Figure 8.1, Table 8.1, Table 8.2 and Table 8.3).

The effects of sRNA binding to its mRNA target are manifold. Small RNAs are between 50 and 300 nucleotides, and they have an impact on the translation of their target mRNA, more often via downregulation of protein synthesis than upregulation (Storz *et al.*, 2011). The binding of an sRNA to its target can prevent the ribosome from reaching the ribosome binding site (RBS) either by directly obstructing its access or by promoting a structural change that leads to its sequestration, therefore preventing translation from occurring (Adams & Storz, 2020; Heidrich *et al.*, 2007). Inversely, this binding could result in changes in the secondary structure of an mRNA, releasing an RBS that would otherwise be sequestered (Majdalani *et al.*, 2001). An sRNA-Hfq complex can also promote RNA degradation by the recruitment of ribonuclease E (RNase E) (Morita *et al.*, 2005). Small RNA binding can also lead to ribosome stalling, which can reveal downstream RNase E sites and promote target mRNA degradation (Pfeiffer *et al.*, 2009). All this to say, sRNAs' modes of action are diverse and rely on regulatory mechanisms that affect mRNA stability, degradation, or accessibility to the ribosome and RNA-binding proteins.

In Gram-negative bacteria, sRNA regulation is often facilitated by chaperone proteins Hfq and ProQ. Homologs of the protein Hfq are found in approximately 50% of all sequenced bacteria (Sun *et al.*, 2002), whereas ProQ is specific to Gram-negative microorganisms (Christopoulou & Granneman, 2021). We hypothesized that sRNAs could be found in all Gram-negative bacteria encoding for either chaperone proteins. Even if it is present in Gram-positive bacteria, Hfq does not seem to operate in the same manner as in Gram-negative bacteria (Christopoulou &

Granneman, 2021). The identification of RNA-binding proteins in Gram-positive bacteria with a similar impact on gene regulation as Hfq and ProQ is an important missing factor in paving the way to novel sRNA discovery. It was suggested that the protein CsrA could fulfill this function in Gram-positive bacteria, but research is lacking. In fact, it was only demonstrated that CsrA could promote the interaction between the sRNA SR1 and its target in *B. subtilis* (Müller *et al.*, 2019).

The first characterized sRNA, MicF, was described approximately 40 years ago. Initially identified as a “repressor RNA”, MicF is an sRNA that regulates an important outer membrane protein in *Escherichia coli*, OmpF (Andersen *et al.*, 1989; Cohen *et al.*, 1988; Mizuno *et al.*, 1984). Since this first breakthrough, numerous sRNAs have been identified; the rate of these discoveries has increased since the advent of next-generation sequencing, which permitted RNA-sequencing. However, their discovery mainly focused on model organisms such as *Escherichia* and *Salmonella* species, overlooking other bacteria that also have the potential to encode numerous sRNAs. We wanted to estimate whether we are far from the true number of sRNAs by getting an overview outside these common models. By demonstrating the biases toward model organisms and pathogens, we hope to pique the interest of other non-coding RNA enthusiasts and pave the way for new sRNAs discoveries.

3.5 Prevalence of sRNAs in Bacteria

Information about sRNAs annotated in bacterial genomes compiled for this article was procured from RiboGap (Naghdi *et al.*, 2017) (queries are available in Supplementary Material, Table 8.4). This database facilitates the inspection of non-coding regions in prokaryotes. The compilation of annotated sRNAs in RiboGap comes from Rfam, a database compiling sequences from structural RNA families (Kalvari *et al.*, 2021), and is limited to available annotations. However, additional sRNAs are predicted within RiboGap compared to Rfam since homology searches were executed on all prokaryotic genomes available in NCBI (Sayers *et al.*, 2010) from covariance models of the entire sRNA collection in Rfam.

Rfam allowed us to examine the prevalence of sRNAs in a wide range of bacteria, but other organism-specific databases exist. To name a few, sRNAMap is a web-based application for Gram-negative bacteria only (Huang *et al.*, 2009), whereas sRNAdb (Pischimarov *et al.*, 2012) is specific to Gram-positive bacteria. RegulonDB (Santos-Zavaleta *et al.*, 2019) and Ecocyc (Keseler *et al.*, 2009) compile sRNAs from *E. coli*, while published data on sRNAs in Staphylococci with a focus on *Staphylococcus aureus* are gathered in the SRD database (Sassi *et al.*, 2015). BSRD also contains a repertoire of small bacterial RNA, but most of its data are homologs found

in Rfam (Li *et al.*, 2013). We, therefore, chose to work with Rfam to obtain a sense of the extent of sRNAs in bacteria, but it is worth mentioning that other databases are available when the research is more focused on a particular organism, although this is generally limited to model organisms. This article also focuses on sRNAs with an E-value lower than 0.0005, to remove any sRNAs with poor homology prediction.

Since the characterization of the first sRNA in the 1980s, numerous sRNAs have been discovered in a wide range of bacterial phyla, including 549 distinct sRNA families listed in Rfam. Proteobacteria and Terrabacteria groups encode the highest number of distinct sRNAs (Table 3.1).

Table 3.1 Number of distinct annotated sRNAs in different phyla

Phylum Group	sRNAs
Acidobacteria	4
Aquificae	1
Calditrichaeota	1
Dictyoglomi	1
FCB group [1]	16
Fusobacteria	2
Nitrospirae	3
PVC group [2]	8
Proteobacteria	345
Spirochaetes	6
Synergistetes	1
Terrabacteria group	210
Thermodesulfobacteria	1
Thermotogae	1

[1] FCB group stands for Fibrobacteres, Chlorobi, and Bacteroidetes, whereas [2] PVC group represents Planctomycetes, Verrucomicrobia, and Chlamydiae.

Bacteria from the phylum Proteobacteria and the Terrabacteria phylum group both encode many distinct sRNAs (345 and 210, respectively). It comes as no surprise that the Terrabacteria super-phylum group stands out from others in terms of the number of annotated sRNAs since it encompasses approximately two-thirds of all identified species, including all Gram-positive bacteria and most spore-producing bacteria (Battistuzzi & Hedges, 2009). It also includes human pathogens such as *Clostridium*, *Staphylococcus* and food and waterborne pathogens such as *Listeria* and *Campylobacter* (Battistuzzi & Hedges, 2009). Proteobacteria is a well-studied phylum, since it is predominant in the human gut microbiome and often associated with multiple intestinal and extraintestinal diseases and includes many human pathogens, such as those from the genera *Bordetella*, *Brucella*, *Burkholderia*, *Francisella*, *Helicobacter*, *Neisseria*, *Rickettsia*, *Salmonella* and *Yersinia (Rizzatti et al., 2017)*, which would explain incentives to study them.

Most species have a relatively small number of distinct sRNAs annotated within their genomes (Supplementary Material Figure 8.2, A), whereas those with the highest sRNA occurrences are within the phylum Proteobacteria and Terrabacteria group (Table 3.1). If we disregard those overrepresented phyla, the remaining bacteria have an average of only 1 to 2 sRNAs encoded in their genome (Supplementary Material Figure 8.2, A). From that list, most are non-pathogenic and are not considered model organisms. However, there are a few exceptions, including those responsible for the sexually transmitted infections (STI), chlamydia and syphilis (*Chlamydia trachomatis* (Taylor-Robinson, 1994) and *Treponema pallidum* (Weinstock et al., 1998), respectively), bacteria associated with dog bite infections (*Capnocytophaga* sp. (Le Moal *et al.*, 2003)) as well as plant (*Liberibacter* sp. (Sena-Vélez *et al.*, 2019)), poultry (*Riemerella* sp. (Hess *et al.*, 2013)) and fish (*Tenacibaculum* sp. (Saad & Atallah, 2014)) pathogens. This list also includes model organisms in specific fields of research, such as *Chlorobaculum* sp., which is used to study sulfur metabolism and photosynthesis (Marnocha *et al.*, 2016), as well as *Porphyromonas* sp., used to study the interaction of anaerobic bacteria with host cells (Wunsch & Lewis, 2015). Despite their relevance as pathogens and in fundamental research, the presence of sRNAs has not been examined in these species. A genome-wide transcriptomic study was realized in *Chlamydia trachomatis,* identifying 43 candidate sRNAs (Albrecht *et al.*, 2010), but only one is referenced within the Rfam database, IhtA (Tattersall *et al.*, 2012). It would be interesting to dedicate future sRNA studies to these bacteria since they have a very small number of annotated sRNAs. Conversely, numerous bacterial strains from the major Gram-negative phylum Proteobacteria encode for large numbers of sRNAs (Figure 3.1).

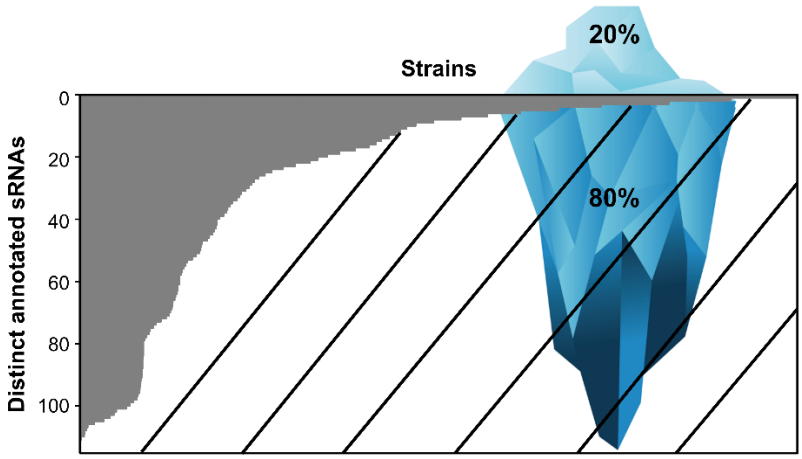

Figure 3.1 Number of distinct annotated sRNAs per bacterial strain in Proteobacteria.

The iceberg is intended to be a graphical representation of the knowledge we have about the prevalence of sRNAs in Proteobacteria (gray section) as opposed to what we could be missing (hatched section). The ratio of the surface versus underwater portions of the iceberg is proportional to results represented in the graph, where the gray region is what is known (i.e., the visible part of the iceberg), and the hatched area under that region is what could be left to discover (that is, the underwater section of the iceberg). Percentages also represent this ratio. This figure represents a compilation of 2629 strains. Only sRNAs with an E-value lower than 0.0005 were considered.

Figure 3.1 represents the potential to discover novel sRNAs, where the underwater portion of the iceberg depicts the sRNAs that remain to be found if all strains contain similar quantities of sRNAs as the most well-studied bacteria. Given chaperone proteins ProQ and Hfq are highly conserved in Gram-negative bacteria (de Fernandez *et al.*, 1972; Olejniczak & Storz, 2017), we feel comfortable making this extrapolation since the occurrence of either or both chaperone proteins in the genome of bacteria could be a good indication of the presence of sRNAs. We also represented the potential for sRNA discovery in bacteria from other phyla (Supplementary Material Figure 8.2). However, given the chaperone protein ProQ is absent in Gram-positive bacteria (Olejniczak & Storz, 2017), this extrapolation is less reliable. Despite the fact that an Hfq homolog is present in Gram-positive bacteria, it does not seem to act as a matchmaker for sRNAs and their targets, which is its most prominent role in Gram-negative bacteria (Jørgensen *et al.*, 2020).

3.6 Species Encoding for sRNAs

The model organisms *Salmonella enterica* and *Escherichia coli* contain the most distinct sRNAs annotated in their genomes, with 118 and 98, respectively, if you consider all strains for each species (Figure 3.2).

For Proteobacteria, it is hardly surprising that *Escherichia coli* is at the top of the list, since it is the microbiologist's bacteria of choice in the laboratory due to its ease of handling and the availability of associated tools. It is the most studied and best understood bacteria (Blount, 2015), and much of our fundamental understanding of biology has come from this model organism, including the genetic code (Crick *et al.*, 1961) and the characterization of the first sRNA (Andersen *et al.*, 1989; Cohen *et al.*, 1988; Mizuno *et al.*, 1984). As a very close relative of *E. coli*, *Salmonella enterica* is expected to contain similar sRNAs, although many other species-specific sRNAs were found, presumably due to extensive research on host-pathogen interactions, which made use of this model organism. *Salmonella* sp. are attractive model organisms because they can target a wide range of hosts with multiple evasion strategies giving an idea of major tactics adopted by other pathogens (Garai *et al.*, 2012). For example, the sRNA IsrJ in *Salmonella* sp. was demonstrated to contribute to the invasion of epithelial cells, and knockout strains for this sRNA lead to less invasive mutants (Padalon-Brauch *et al.*, 2008). For Proteobacteria, all the bacteria from the Figure 3.2 belong to the family *Enterobacteriaceae*, which encode for 145 distinct sRNAs compared to an average of seven for all other bacterial families.

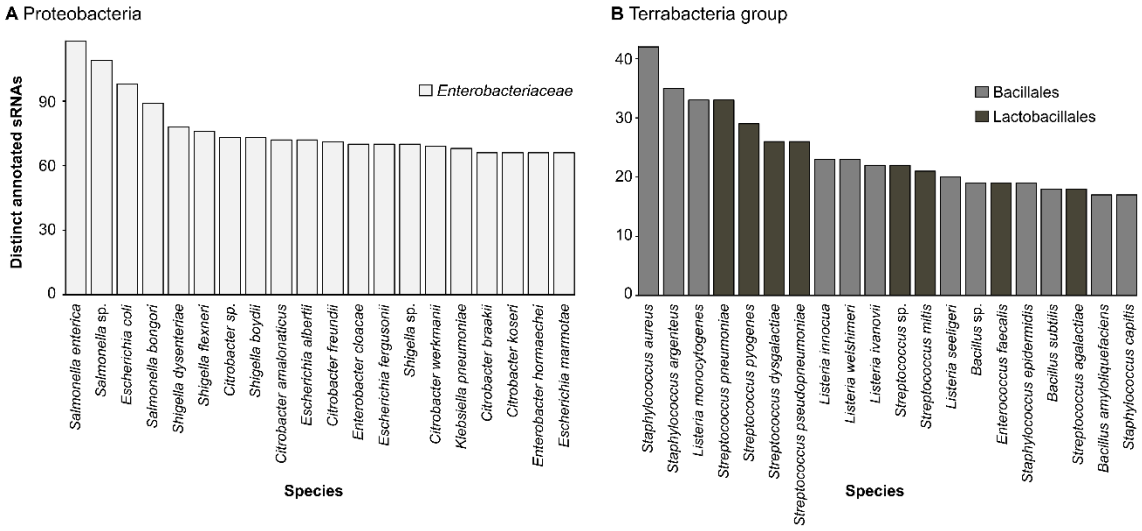

Figure 3.2 Top 20 bacterial species with the highest number of distinct annotated sRNAs

(A) Proteobacteria and in **(B)** bacteria from the Terrabacteria group. Species denoted with "sp." represent instances where only the genus of the bacteria was noted. It can be observed that in **(A)**, all species are from the same family, *Enterobacteriaceae*. In **(B)**, species from different orders are emphasized by their own color. The number of distinct sRNAs considers all strains for each species. Only sRNAs with an E-value lower than 0.0005 were considered.

In the case of bacteria from the Terrabacteria group, human pathogens *Staphylococcus* and *Listeria* have the highest number of distinct annotated sRNAs (Archer, 1998; Drevets & Bronze, 2008). As for *Streptococcus* sp., some of its species are considered part of the normal human microbiome, but others, such as *Streptococcus pneumoniae,* are responsible for most cases of

pneumonia worldwide (Krzyściak *et al.*, 2013). The model organism *Bacillus subtilis* also has a high number of annotated sRNAs, perhaps because it is a common Gram-positive bacteria to investigate biofilm formation (Errington & van der Aart, 2020), among other processes. As we can observe, the species with the most annotated sRNAs are those associated with high research intensity, either because they are a threat to human health or due to their attractiveness as model organisms (Table 3.2). By digging past this bacterial all-star list, we hypothesized that multiple novel sRNAs are left to be discovered. By focusing on less standard organisms, we could potentially extend the role of sRNAs to unexpected new functions. Moreover, sRNAs discovered in understudied bacteria could be the missing puzzle piece to solve an incomplete regulatory mechanism in a model organism.

Table 3.2 Description of genus encoding for the most distinct sRNAs

Genus	Nb of distinct sRNAs [1]	Description	Ref
		Proteobacteria	
Salmonella	119	Model organism to study host-pathogen interactions	(Garai *et al.*, 2012)
Escherichia	99	Most well-understood bacteria	(Blount, 2015)
Citrobacter	88	Third most common cause of UTIs in hospitalized patients	(Ranjan & Ranjan, 2013)
Shigella	85	Causative pathogen of shigellosis	(Killackey *et al.*, 2016)
Enterobacter	78	Responsible for nosocomial infections	(Sanders Jr & Sanders, 1997)
Klebsiella	74	Nosocomial pathogen, model organism to study drug resistance	(Bi *et al.*, 2015)
		Terrabacteria group	
Streptococcus	55	Responsible for most cases of pneumonia worldwide	(Krzyściak *et al.*, 2013)
Staphylococcus	46	Most prevalent cause of infection in hospitalized patient	(Archer, 1998)
Listeria	35	Foodborne human pathogens causing central nervous system infections	(Drevets & Bronze, 2008)
Bacillus	26	Most-studied Gram-positive bacteria, model organisms for cellular development	(Errington & van der Aart, 2020)
Enterococcus	25	Principal cause of the healthcare-associated death worldwide	(Bi *et al.*, 2015; García-Solache & Rice, 2019; Sanders Jr & Sanders, 1997)

[1] The number represents the quantity of distinct annotated sRNAs in all bacterial strains within this genus. Only sRNAs with a E-value lower than 0.0005 were considered.

3.7 Most Abundant Small RNAs

We were then interested to know which sRNAs were the most present throughout all bacterial genomes. If an sRNA was annotated multiple times within the same strain, we counted all individual instances (Figure 3.3).

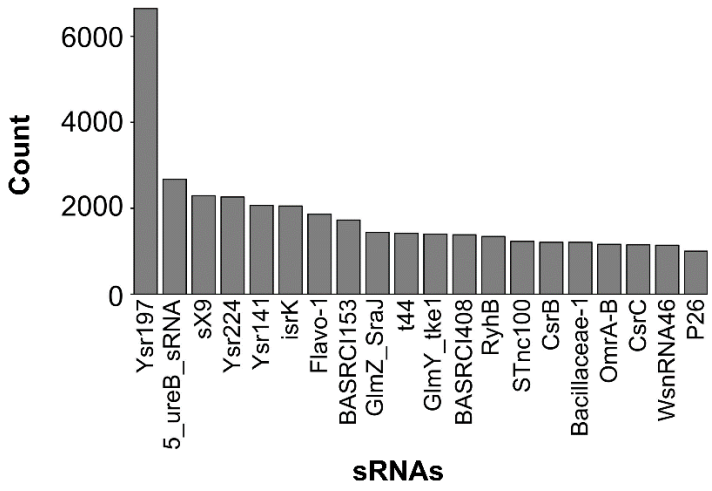

Figure 3.3 Top 20 sRNAs annotated in bacteria.

Each individual occurrence of sRNAs were counted, even if some were found multiple times within the same genome. Only sRNAs with an E-value lower than 0.0005 were taken into consideration.

From the top lists of sRNAs, most were discovered in human pathogenic bacteria (Ysr197 (Achtman *et al.*, 1999; Nuss *et al.*, 2015), 5_ureB_sRNA (Arnvig & Young, 2009), Ysr224 (Nuss *et al.*, 2015; Righetti *et al.*, 2016), Ysr141 (Schiano *et al.*, 2014), isrK (Padalon-Brauch *et al.*, 2008; Sittka *et al.*, 2008), BASRCI153 (Dong *et al.*, 2014), BASRCI408 (Dong *et al.*, 2014) and STnc100 (Sittka *et al.*, 2008)), in causative agent of plant infection (sX9 (Schmidtke *et al.*, 2012)) or in parasitic microbes (WsnRNA-46 (Mayoral *et al.*, 2014)). The latter was found in *Wolbachia* sp., the most prevailing vertically transmitted endosymbiont around the world, impacting more than 40% of arthropods (Mayoral *et al.*, 2014). The remaining were found in the model organism *E. coli* (GlmZ_SraJ (Reichenbach *et al.*, 2008; Righetti *et al.*, 2016; Rivas *et al.*, 2001; Urban & Vogel, 2008), t44 (Tjaden *et al.*, 2002), GlmY_tke1 (Reichenbach *et al.*, 2008; Righetti *et al.*, 2016; Rivas *et al.*, 2001; Urban & Vogel, 2008), RyhB (Argaman *et al.*, 2001; Davis *et al.*, 2005; Gottesman, 2005; Massé & Gottesman, 2002; Neuhaus *et al.*, 2017; Porcheron *et al.*, 2014; Zhang *et al.*, 2018), CsrB (Cui *et al.*, 1995; Heroven *et al.*, 2012; Liu *et al.*, 1997; Mei *et al.*, 2017; Yang *et al.*, 2008), OmrA-B (Argaman *et al.*, 2001; Guillier & Gottesman, 2006; Holmqvist *et al.*, 2010; Wassarman *et al.*, 2001) and CsrC (Argaman *et al.*, 2001; Weilbacher *et al.*, 2003)) or by computational homology searches (Flavo-1 (Weinberg *et al.*, 2010), Bacillaceae-1 (Weinberg *et al.*, 2010) and P26 (Livny *et al.*, 2006)) (Table 3.3).

Table 3.3 Description of top 20 most prevalent sRNAs in bacteria

sRNA	Description	Rfam ID	sRNA Expression	Discovered in	Ref
Ysr197	*Yersinia* sRNA 197	RF02849	Expressed in exponential phase	*Yersinia pseudotuberculosis*	(Nuss *et al.*, 2015)
5_ureB_sRNA	-	RF02514	Downregulate expression of operon *ureAB*	*Helicobacter pylori*	(Wen *et al.*, 2013)
sX9	*Xanthomonas* sRNA sX9	RF02228	-	*Xanthomonas campestris pv.* vesicatoria *(*Xcv*)*	(Schmidtke *et al.*, 2012)
Ysr224	*Yersinia* sRNA 224	RF02770	Temperature-responsive	*Yersinia pseudotuberculosis*	(Nuss *et al.*, 2015; Righetti *et al.*, 2016)
Ysr141	*Yersinia* sRNA 141	RF02675	Influence the expression of Yop-Ysc type III secretion system (T3SS) (critical system for virulence)	*Yersinia pestis*	(Schiano *et al.*, 2014)
isrK	isrK Hfq binding RNA	RF01394	Stationary phase, low oxygen, low magnesium	*Salmonella typhimurium*	(Padalon-Brauch *et al.*, 2008; Sittka *et al.*, 2008)
Flavo-1	-	RF01705	-	Bacteroidetes	(Weinberg *et al.*, 2010)
BASRCI153	*Brucella* sRNA CI153	RF02604	Putative target: BAB1_1361	*Brucella abortus*	(Dong *et al.*, 2014)
GlmZ_SraJ	GlmZ RNA activator of *glmS* mRNA	RF00083	activator of *glmS* mRNA	*Escherichia coli*	(Reichenbach et al., 2008; Righetti et al., 2016; Rivas et al., 2001; Urban & Vogel, 2008)
t44	-	RF00127	-	*Escherichia coli*	(Tjaden *et al.*, 2002)
GlmY_tke1	GlmZ RNA activator of *glmS* mRNA	RF00128	activator of *glmS* mRNA	*Escherichia coli*	(Reichenbach et al., 2008; Righetti et al., 2016; Rivas et al., 2001; Urban & Vogel, 2008)
BASRCI408	*Brucella* sRNA CI408	RF02599	Putative target: BAB1_2002	*Brucella abortus*	(Dong *et al.*, 2014)
RyhB	-	RF00057	Iron metabolism, regulates siderophore production and virulence, persistence regulation	*Escherichia coli*	(Argaman et al., 2001; Davis et al., 2005; Gottesman, 2005; Massé & Gottesman, 2002; Neuhaus et al., 2017; Porcheron et al., 2014; Zhang et al., 2018)
STnc100	Gammaproteobacterial sRNA STnc100	RF02076	-	*Salmonella* sp.	(Sittka *et al.*, 2008)
CsrB	CsrB/RsmB RNA family	RF00018	Binds the CrsA protein	*Escherichia coli*	(Cui et al., 1995; Heroven et al., 2012; Liu et al., 1997; Mei et al., 2017; Yang et al., 2008)
Bacillaceae-1	-	RF01690	-	Bacteroidetes	(Weinberg *et al.*, 2010)
OmrA-B	-	RF00079	Target several genes encoding outer membrane proteins	*Escherichia coli*	(Argaman et al., 2001; Guillier & Gottesman, 2006; Holmqvist et al., 2010; Wassarman et al., 2001)
CsrC	-	RF00084	Binds the CrsA protein	*Escherichia coli*	(Argaman et al., 2001; Weilbacher et al., 2003)
Ysr276	*Yersinia* sRNA 276	RF02850	-	*Yersinia pseudotuberculosis*	(Nuss *et al.*, 2015)
WsnRNA46	*Wolbachia* sRNA 46	RF02625	Expressed in cells infected by parasitic microbe *Wolbachia*	*Wolbachia* sp.	(Mayoral *et al.*, 2014)
P26	*Pseudomonas* sRNA P26	RF00630	-	Pseudomonas aeruginosa	(Livny et al., 2006)

In other words, not only are we missing numerous sRNA instances in various bacteria, as underscored by Figure 3.1, but the diversity of sRNA families is also expected to be much greater. Indeed, most sRNAs are unique to limited taxonomic groups, which means that each exploratory sRNA study in an underrepresented taxon will likely lead to the discovery of novel sRNA families. Then, by homology searches, they could be related to other bacteria of interest and further deepen our knowledge of gene regulation mediated by sRNAs.

3.8 Biases towards model organisms and pathogens

In order to demonstrate that research intensity is biased toward model organisms, pathogens, and closely related species, we looked at the number of annotated genes and RNAs in bacteria (Figure 3.4).

Information about genome size and the number of annotated genes and RNA comes from RiboGap (Naghdi *et al.*, 2017), which extracts data from the NCBI FTP site (Sayers *et al.*, 2010). For gene annotations, sizes are based on complete genomes, which include all plasmids and chromosomes of a given strain if applicable. However, RNAs are compiled per “DNA fragment” (chromosome or plasmid) since it is not accessible per genome within the RiboGap database. The size of each fragment was taken from all available Genbank files from the NCBI FTP site (Sayers *et al.*, 2010). Results were limited by the available annotations. For example, some strains did not have annotated genes in NCBI and were removed from Figure 3.4. Moreover, some entries were mislabeled as complete genomes but were, in fact, WGS (Whole Genome Shotguns) projects with incomplete genomes, leading to a miscalculation in the number of genes (values doubled up). These erroneous data were removed from Figure 3.4 (shown for transparency purposes in Supplementary Material Figure 8.3).

Expectedly, the number of annotated genes increases proportionally with the genome size, with on average one gene per kb and a relative standard deviation (RSD) of 7% (Figure 3.4, A).The top species with the most annotated sRNAs (Figure 3.2) from Proteobacteria and Terrabacteria groups (blue and black dot, respectively, (Figure 3.4, A) tend to have slightly higher numbers of genes for a given genome length. We also graphed the number of annotated RNAs compared to the fragment size, which highlights the disparity in the annotation of RNA versus protein-coding genes. There is, on average, one annotated RNA every 25 kb with a relative standard deviation of 47%, emphasizing how spread out the values are from the average number, ranging from ~1/10 kb to ~1/100 kb (Figure 3.4, B). Information about RNA families comes from RiboGap (Naghdi *et*

al., 2017) and is derived from Rfam (except for terminators, which can be found in RiboGap but were not included in these results). In principle, we should expect a similar trend for RNAs (Figure 3.4, B) as for genes (Figure 3.4, A), i.e., the number of annotated RNAs should increase proportionally with fragment size. However, it is clearly not the case here, emphasizing how RNA annotations trail behind gene annotations.

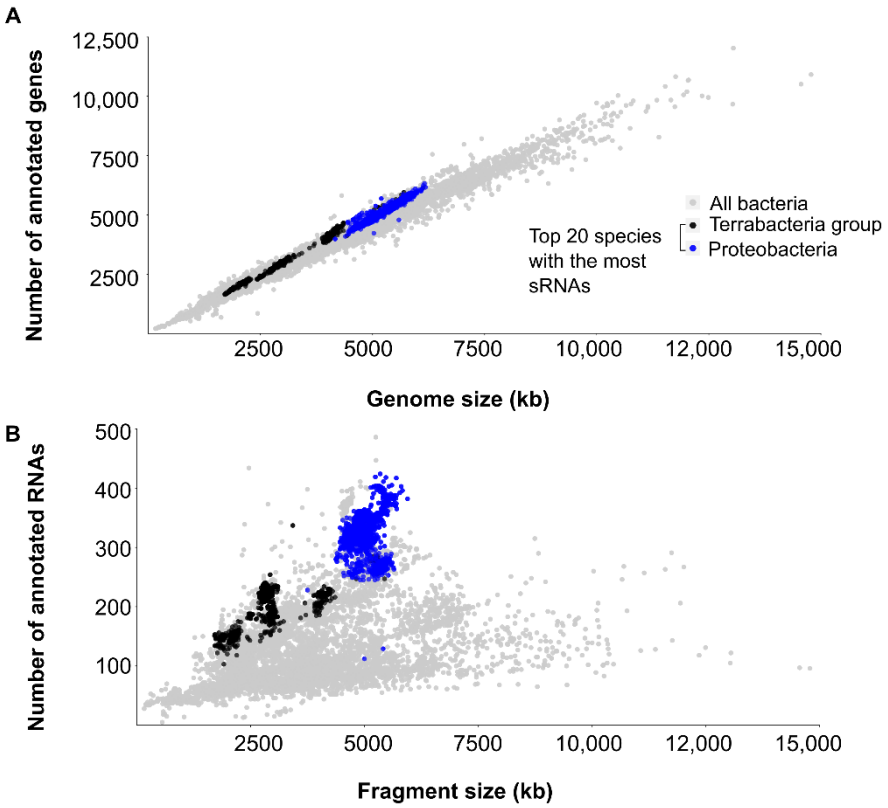

Figure 3.4 Number of annotated genes and RNAs in bacteria.

Data required for the creation of this graph were taken from RiboGap (Naghdi *et al.*, 2017). **(A)** The number of annotated genes is graphed according to the genome sizes, which comprises all chromosomes and plasmids of each individual strain if applicable. **(B)** The number of annotated RNAs is graphed according to the "fragment size", which considers chromosomes and plasmids separately for each individual strain. RNAs are not limited only to sRNAs but also include CRISPR RNAs, antisense RNAs, sRNAs, long non-coding RNAs (lncRNAs), rRNAs, ribozymes, tRNAs and cis-regulatory elements. Species from Terrabacteria group and Proteobacteria that were found to have the most annotated sRNAs (Figure 3.2) are represented by black and blue dots, respectively; all other strains are shown in gray.

Annotations are dependent on the research intensity associated with each strain: the fact that some RNAs are not annotated does not mean that they are not present, but simply that they have yet to be identified. When we emphasize the species with the most annotated sRNAs in their genome (black and blue dots, Figure 3.4, B), they also tend to be those that have the highest number of RNAs in general for a given fragment length. Therefore, their large number of annotated sRNAs likely results from high research intensity. We recreated Figure 3.4, this time

emphasizing bacteria labeled as human pathogens in RiboGap (Naghdi *et al.*, 2017) (Supplementary Material Figure 8.4). Even if there are large incentives to study human pathogenic bacteria, only a handful of model organisms were well characterized. There is still room for novel RNA discovery even among numerous pathogens, as suggested by the fact that the number of annotated RNAs does not necessarily increase as expected with fragment size.

3.9 Conclusions and Perspectives

Small RNAs are important for gene regulation and modulation of responses to environmental changes. They are found in numerous bacterial phyla, especially Proteobacteria and Terrabacteria groups. However, we underestimate their prevalence because of the focus on model organisms and pathogens. Genera encoding for the highest number of sRNAs are human pathogens (*Salmonella*, *Escherichia*, *Citrobacter*, *Shigella*, *Enterobacter*, *Klebsiella*, *Streptococcus*, *Staphylococcus* and *Listeria,* amongst others) or model organisms (*Bacillus*, *Escherichia*, *Salmonella* and others). Only a small fraction of all bacteria encode for numerous sRNAs, but it would be surprising that others would not have the same variety of regulatory RNAs, especially if they encode for the RNA chaperone proteins Hfq and/or ProQ. Moreover, the diversity of sRNAs is anticipated to be much greater, since most sRNAs are unique to limited taxonomic groups. For instance, the species that encode the most distinct sRNAs within the phylum Proteobacteria are all from the same family, *Enterobacteriaceae*.

Expectedly, species associated with high research intensity are also those with the largest number of annotated genes (relative to genome size) and even more so of RNAs, but what was less obvious before is how much RNA annotations fall behind gene annotations. By increasing RNA studies of infrequently studied bacteria, we could improve our capacity to annotate sRNAs and our knowledge of the extent of RNA families in bacteria, including sRNAs.

Even if there is still much to learn on sRNAs in major experimental models, our goal was to highlight the potential to discover novel sRNAs by stressing that current findings are focused on model organisms and pathogens. It was also an opportunity to take stock of the extent of our knowledge. Although there are fewer incentives to study bacteria that are neither models nor pathogens nor of direct industrial interest, new sRNA discoveries could deepen our comprehension of genetic regulation and perhaps lead to new and fascinating mechanisms. Furthermore, beyond the *E. coli* and *B. subtilis* models, there are numerous organisms that provide important models for specific biological processes. A few examples include

Methylorubrum extorquens for the metabolism of 1-carbon compounds (Vuilleumier *et al.*, 2009), *Myxococcus xanthus* for bacterial social behavior (Saïdi *et al.*, 2020), *Azotobacter vinelandii* for nitrogen fixation (Setubal *et al.*, 2009) or *Mycoplasma genitalium* for minimal organisms (Hutchison III *et al.*, 2016). RNA-seq and sRNA discovery methodologies permitted transcriptome-wide evaluation of potential sRNAs, even if further experimental validation requires a significant amount of work. Small RNAs should still be in the spotlight of research in relation to non-coding RNA-mediated genetic regulation because we have just scratched the surface of their full potential and likely have an underappreciation of the true complexity of the regulation of gene expression by sRNAs in bacteria.

3.10 Funding

E.B. was supported by Natural Sciences and Engineering Research Council (NSERC), Fonds de Recherche du Québec Natures and Technologies (FRQNT) and Foundation Armand-Frappier. J.P. is a junior 2 FRQS research scholar. This work was supported by NSERC [RGPIN-2019-06403].

3.11 Acknowledgments

This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Digital Research Alliance of Canada (www.alliancecan.ca). The authors would like to thank Jessie Muir for revising the manuscript.

Supplementary material is available in annexe I (le matériel supplémentaire pour cet article est disponible en annexe I).

4 CHAPITRE 2: ANALYSIS OF NON-CODING RNAS IN *METHYLORUBRUM EXTORQUENS* REVEALS NOVEL SMALL RNAS SPECIFIC TO *METHYLOBACTERIACEAE*

L'étude des ARN non codants chez *Methylorubrum extorquens* révèle de nouveaux sRNA spécifiques aux *Methylobacteriaceae* (Traduction française)

Auteurs :

Emilie Boutet[1], Samia Djerroud[1], Kadidia Dite Selly N'Diaye[1], Katia Smail[1], Marie-Josée Lorain[2], Meiqun Wu[2], Martin Lamarche[1,2], Roqaya Imane[1], Carlos Miguez[2] and Jonathan Perreault[1]

[1] INRS- Centre Armand-Frappier Santé Biotechnologie, Laval, QC, H7V 1B7, Canada

[2] National Research Council Canada, 6100 Royalmount, Montréal, Québec, Canada, H4P-2R2.

Journal : BioRxiv (journal non révisé par les pairs)

Publié : 24 janvier 2022

DOI : https://doi.org/10.1101/2022.01.24.477521

Journal : *RNA biology* (journal révisé par les pairs)

Soumission : 20 octobre 2022

Contribution des auteurs :

EB a réalisé et conçu la plupart des expériences et a rédigé le manuscrit. SD et KT ont participé à l'analyse bioinformatique. KDSN a contribué aux essais sur les conditions de stress. MJL et MW ont participé à la mise en place des expériences de fermenteurs pour le RNA-seq. MGL a effectué les extractions d'ARN pour le RNA-seq. RI, MGL et CBM ont contribué à la conception des expériences et ont initié une partie du projet de recherche. JP a supervisé le projet et a révisé le manuscrit.

4.1 Liens entre les objectifs et ce chapitre de la thèse

Dans le chapitre 1 de cette thèse, nous avons montré à quel point les sRNAs ont été peu étudiés chez les bactéries (si on exclut les « top modèles ») et, par le fait même, le grand potentiel de découverte de nouveaux sRNAs si on se penche sur la question chez ces organismes. Les Figure 3.1 et Figure 3.4 de cette thèse ont été retouchées, mais cette fois l'accent est mis sur notre bactérie d'intérêt, *M. extorquens* (avec une flèche rouge et des points verts, respectivement) (Figure 4.1 et 4.2).

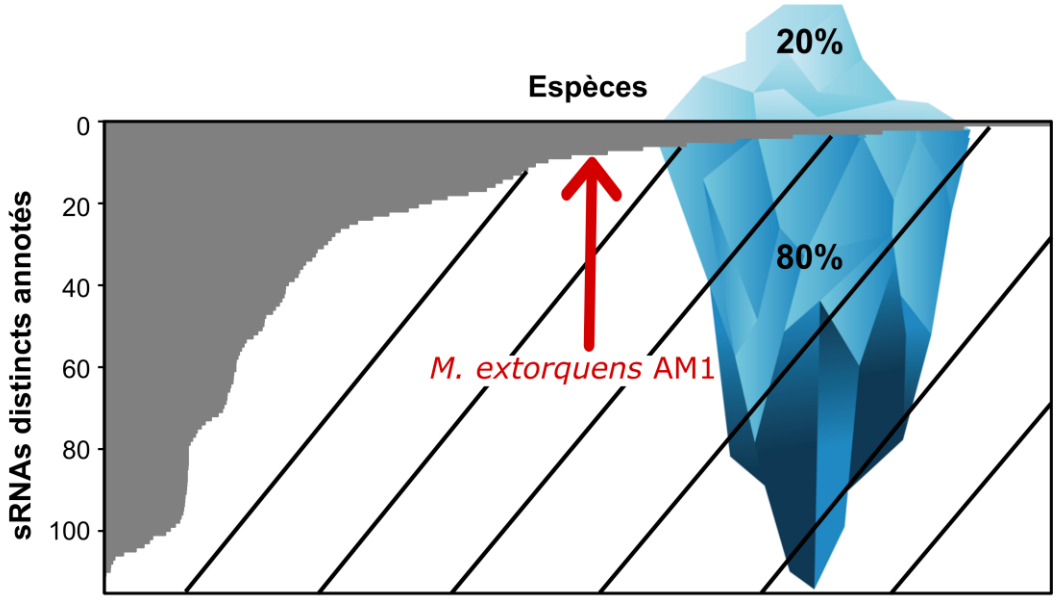

Figure 4.1 Nombre de sRNAs distincts annotés par souche bactérienne chez les Protéobactéries avec une emphase sur M. extorquens, adapté de la Figure 3.1

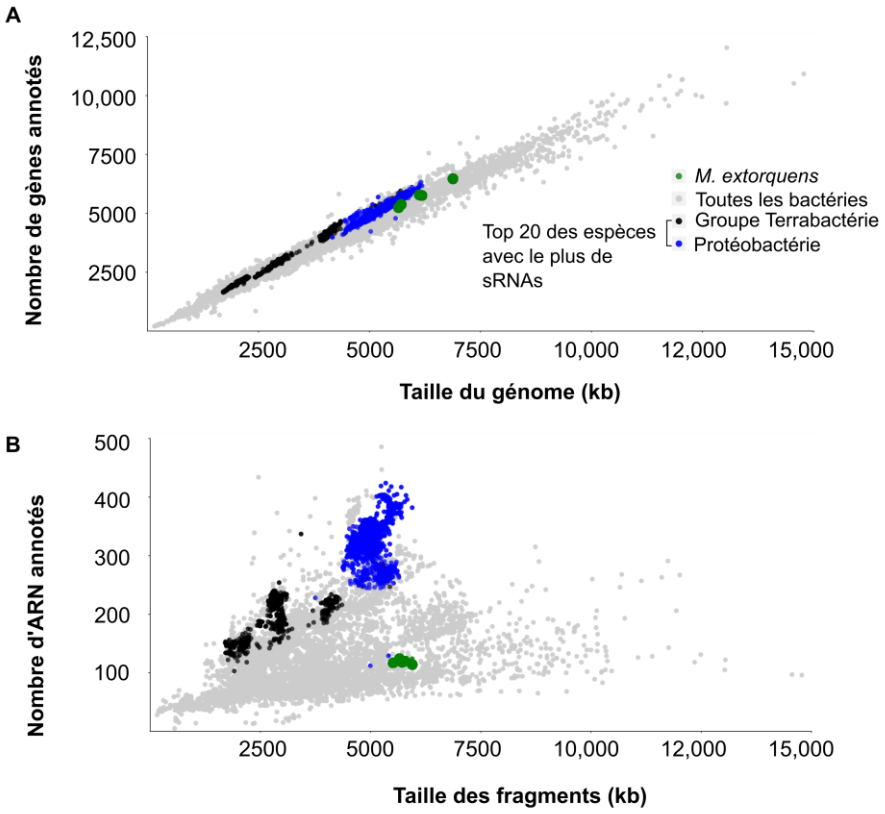

Figure 4.2 Nombre de gènes et d'ARN annotés chez les bactéries avec une emphase sur *M. extorquens,* adapté de la Figure 3.4

Ce deuxième chapitre est donc un exemple avec notre organisme d'intérêt *M. extorquens* qui illustre bien cet état de fait. *M. extorquens* a beaucoup moins de sRNAs distincts annotés dans son génome lorsque l'on compare avec d'autres Protéobactéries (Figure 4.1) Le génome de *M. extorquens* correspond à la tendance attendue, c'est-à-dire où le nombre de gènes encodant des protéines annotées augmente proportionnellement avec la taille du génome (Figure 4.2, A). Cependant, il y a un retard dans le nombre d'ARN annotés (Figure 4.2, B).

Une étude sur les ARN de cette bactérie pourrait améliorer notre capacité à annoter ses ARN. Dans cette étude, les sRNAs ffh, CC2171 et BjrC1505 précédemment annotés dans le génome de cette méthylotrophe ont été validés expérimentalement par *Northern blot*, confirmant ainsi l'expression de ces sRNAs chez cet organisme (et dans l'ordre des *Hyphomicrobiales*). Ces sRNAs étaient préalablement annotés dans le génome de *M. extorquens* dans la base de données Rfam (Kalvari *et al.*, 2021) par recherche d'homologie, mais leurs expressions n'avaient jamais été validées auparavant. Une étude *RNA-seq* a aussi permis d'établir une liste

considérable de potentiels sRNAs, révélant deux nouveaux sRNAs spécifiques aux *Methylobacteriaceae*, Methylo2624 et Methylo1969. Ce chapitre est donc un premier pas dans l'exploration des sRNAs de *M. extorquens*, et par extension des *Methylobacteriaceae.*

4.2 Résumé (traduction française)

Methylorubrum extorquens, une méthylotrophe facultative, est un organisme modèle utilisé pour étudier le métabolisme du C1. Il y a un intérêt considérable d'utiliser cette bactérie comme outil biotechnologique, en raison de sa capacité à métaboliser le méthanol, une matière première bon marché qui peut être dérivée de déchets. Malgré l'intérêt pour *M. extorquens* dans une bioéconomie basée sur le méthanol, ses petits ARN non codants (sRNAs) sont peu étudiés, des régulateurs essentiels de l'expression génétique bactérienne à la suite de changements environnementaux. De nombreux sRNAs sont probablement retrouvés dans le génome de *M. extorquens*, d'autant plus qu'elle encode également pour la protéine Hfq, une protéine chaperonne importante dans l'interaction entre les sRNAs et leur cible, mais aussi dans la stabilisation des sRNAs eux-mêmes. Dans cette étude, les sRNAs ffh, CC2171, BjrC1505 précédemment annotés ont été confirmés par *Northern blot*, validant ainsi pour la première fois l'expression des sRNAs dans *M. extorquens*. De plus, l'analyse de données de séquençage de l'ARN a permis d'établir une liste considérable de sRNAs potentiels. Plusieurs candidats d'intérêt ont été testés par *Northern blot*, révélant de nouveaux sRNAs spécifiques aux *Methylobacteriaceae,* Methylo2624 et Methylo1969. Cette recherche fournit la première validation expérimentale des sRNAs chez *M. extorquens* et ouvre la voie à d'autres découvertes.

4.3 Abstract

Methylorubrum extorquens, a facultative methylotroph, is a model organism used to study C1 metabolism. There is considerable interest to exploit this microorganism as a biotechnological tool since it metabolizes methanol, a cheap raw material that can be derived from waste. Despite the appeal of using *M. extorquens* in a methanol-based bioeconomy, little is known about its non-coding small RNAs (sRNAs), essential regulators of bacterial gene expression following environmental changes. *M. extorquens* is expected to contain many sRNAs, especially since it also encodes for the protein Hfq, a chaperone protein important for the interaction between sRNAs and their target and critical for the stabilization of sRNAs themselves. In this study, formerly annotated sRNAs ffh, CC2171, BjrC1505 were confirmed by Northern blot, validating the

expression of sRNAs in *M. extorquens.* Moreover, analysis of RNA-sequencing data established a considerable list of potential sRNAs. Candidates of interest were tested by Northern blot, revealing novel sRNAs specific to *Methylobacteriaceae*, Methylo2624 and Methylo1969. This work provides the first experimental validation of sRNAs in *M. extorquens* and paves the way for other sRNA discoveries.

4.4 Introduction

Methylorubrum extorquens (formerly *Methylobacterium extorquens*) (Green & Ardley, 2018) has potential in a future C1-carbon based bioeconomy for its ability to produce value-added product on a large scale from an inexpensive alternative substrate derived from waste: methanol. This Alphaproteobacteria has already been engineered to generate numerous value-added products from methanol including itaconic acid (Lim *et al.*, 2019) and violacein (Le *et al.*, 2022), among others. Despite its importance at the bioindustrial level and in fundamental research, little is known about its non-coding small RNAs (sRNA).

Usually between 50 and 400 nucleotides, sRNAs are either *cis* or *trans*-encoded and act by base pair complimentarity with their mRNA targets (Waters & Storz, 2009). This pairing affects the translation of the regulated mRNA by diverse modes of action, either by affecting the mRNA stability, by blocking the accessibility for the ribosome or RNA-binding proteins or by promoting its degradation by RNAse E (Adams & Storz, 2020; Heidrich *et al.*, 2007; Majdalani *et al.*, 2001; Morita *et al.*, 2005; Pfeiffer *et al.*, 2009). Since the discovery of the first sRNA MicF, which targets the *ompF* mRNA in *Escherichia coli (E. coli),* approximately 40 years ago (Andersen *et al.*, 1989; Delihas & Forst, 2001), sRNAs have become the most important group of bacterial post-transcriptional regulators (Papenfort & Vogel, 2009). They play a crucial role in regulating gene expression under various conditions including quorum sensing, virulence, and stress response among others (Papenfort & Vogel, 2009; Wagner & Romby, 2015). The chaperone protein Hfq is often associated with sRNAs in Gram-negative bacteria, because it can not only stabilize the interaction between the sRNA and its target, but it can also stabilize the sRNA itself (Møller *et al.*, 2002a; Vogel & Luisi, 2011).

The serine cycle is indispensable for C1 metabolism, whereas multicarbon compounds are processed through the TCA cycle (Peyraud *et al.*, 2012). The expression of enzymes necessary for the specific use of these carbon metabolism pathways are tightly regulated (Šmejkalová *et al.*, 2010), potentially with help of sRNAs. To develop *M. extorquens* as an even more powerful biotechnological tool, it is also important to identify sRNAs, an efficient set of regulatory elements

for gene control that we predict is contained in this bacterium, and that could be used to our advantage in the future. The Hfq protein is encoded within the genome of *M. extorquens* (WP_003600267.1). Jointly with this chaperone protein, sRNAs play an important role in genetic regulation in other Alphaproteobacteria of the same order (*Hyphomicrobiales*) such as *Brucella melitensis*, where 24 distinct sRNAs are annotated (with an E-value lower than 0.0005; information extracted from RiboGap (Naghdi *et al.*, 2017)). *E. coli*, another Proteobacteria, has approximately 100 sRNAs that were experimentally validated by Northern blot (Bak *et al.*, 2015). Since *M. extorquens* is also a Proteobacteria that encodes for the Hfq protein, we hypothesized that it may have a similar number of sRNAs. Indeed, several sRNAs were predicted in its genome, but were never experimentally validated. A better understanding of the role of sRNAs in the genetic regulation of *M. extorquens* could provide insight on how to maximize industrial processes. In this study, previously annotated sRNAs were experimentally validated, confirming their expression for the first time in *M. extorquens*. Moreover, analysis of the transcriptomic data led to the creation of a considerable list of putative sRNAs. Further experimental analysis highlighted novel sRNAs specific to *Methylobacteriaceae*: Methylo2624 and Methylo1969.

4.5 Results and discussion

4.5.1 Annotated sRNAs and intergenic regions size distribution

To get a sense of the likelihood to discover sRNAs in *M. extorquens*, we first looked at the number of annotated sRNAs within the genomes of Alphaproteobacteria (Figure 4.3). This article considers only sRNAs with an E-value lower than 0.0005, to disregard any sRNAs with poor homology prediction. This information was extracted from the database RiboGap (version 2) (Naghdi *et al.*, 2017), since it facilitates the examination of intergenic regions (IGR) from prokaryotes. Evidence about sRNAs in RiboGap comes from the Rfam database (Kalvari *et al.*, 2021) and is limited to available covariance models used for homology searches and annotations.

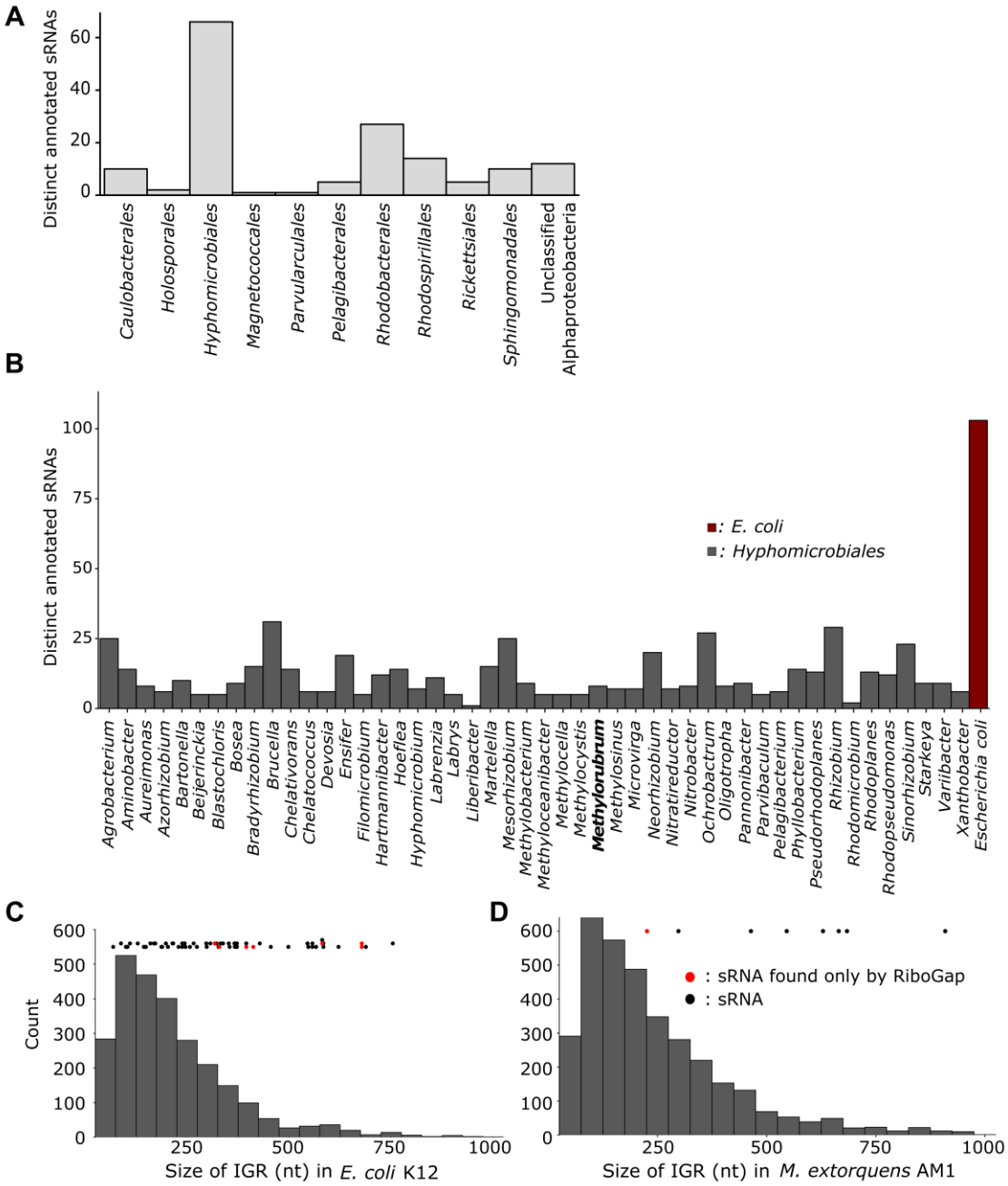

Figure 4.3 Annotated sRNAs in Alphaproteobacteria

Distinct sRNAs with an E-value smaller than 0.0005 annotated in the genome of different orders within the Alphaproteobacteria class in **(A)** and within different genera of the *Hyphomicrobiales* order in **(B)** in grey. The different sRNAs found in the genome of the model Gammaproteobacteria *E. coli* are depicted in red in **(B)**. The size distribution of intergenic gaps that are in between 50 and 1000 nucleotides for the strain *E. coli* K-12 substr. MG1655 and *M. extorquens* AM1 are shown in **(C)** and **(D)** respectively. IGRs containing a sRNA have been identified with dots, where red ones represent sRNAs identified only by RiboGap. Black dots illustrate sRNAs found by both Rfam and RiboGap databases. The histograms are grouped in bins of 50.

A total of 87 distinct sRNAs were annotated in Alphaproteobacteria and spread throughout all orders of this bacterial class (Figure 4.3, A). The order with the highest number of distinct annotated sRNAs (66 sRNAs) is *Hyphomicrobiales*, which comprises *M. extorquens* (Figure 4.3, A). Numerous sRNAs are predicted in different genera of the order *Hyphomicrobiales* (Figure 4.3, B) where eight distinct potential sRNAs (Corbino *et al.*, 2005; del Val *et al.*, 2012; Dong *et al.*, 2014; Landt *et al.*, 2008; Madhugiri *et al.*, 2012; Meyer *et al.*, 2009; Schmidtke *et al.*, 2012; Wen *et al.*, 2013) are annotated in the genome of all available strains of the genus *Methylorubrum*, but were never confirmed in laboratory conditions (Table 4.1). This is still very few compared to the 103 distinct sRNAs annotated within the genome of *E. coli* alone (Figure 4.3, B), a Gammaproteobacteria model organism. All *Hyphomicrobiales* bacteria represented in (Figure 4.3, B) also encode for the chaperone protein Hfq.

Table 4.1 Annotated sRNAs and housekeeping RNAs

sRNAs	Description	Size (nt)	Rfam	Probe	Ref
Annotated sRNAs					
ar45	Alphaproteobacterial sRNA ar45	229	RF02347	Not tested	(del Val *et al.*, 2012)
BASRCI27	Brucella sRNA CI27	157-165	RF02600	Not tested	(Dong *et al.*, 2014)
BjrC1505	Alphaproteobacterial sRNA BjrC1505	147-148	RF02356	TGGATCCTGATTGGGATCTCTTTCCAGT	(Madhugiri *et al.*, 2012)
CC2171	caulobacter sRNA CC2171	170	RF01867	CTCCGGCGTGTGCGCCTAACGCACCCG	(Landt *et al.*, 2008)
ffh		53	RF01793	CCGACAGCGCGTGTGCGCCCTCGG	(Meyer *et al.*, 2009)
suhB	Makes More Granules Regulator RNA (mmgR)	75-80	RF00519	Not tested	(Corbino *et al.*, 2005)
sX4	Proteobacterial sRNA sX4	144-153	RF02223	Not tested	(Schmidtke *et al.*, 2012)
5_ureB_sRNA		285-302	RF02514	Not tested	(Wen *et al.*, 2013)
Housekeeping RNAs					
tRNA	Leucine tRNA	85	RF00005	TGCCCAGGAAAGGACTCGAACCTTCACCTCTTGCAAGACTGGTACCTGAA	(Hou, 1993)
5S RNA	Ribosomal RNA 5S	115	RF00001	CTGGCGGCGACCGACTCTCCCGTGTCTTG	(Szymanski *et al.*, 2002)
tmRNA	Alphaproteobacteria transfer-messenger RNA	378	RF01849	TTGTCGTTGGCAACTATTGCAAAGGCCCGA	(Felden *et al.*, 1997)

Probes were designed for sRNAs BjrC1505, C2171 and ffh to be tested experimentally by Northern blot analysis. Probes for housekeeping RNAs served as positive and normalization controls as well as size guidelines. Sizes of annotated RNAs represent the ranges within all *Methylorubrum*, whereas the sizes for control RNAs depict those within *M. extorquens* AM1 specifically.

This survey of sRNAs using the RiboGap database in Alphaproteobacteria, more importantly in *Hyphomicrobiales*, supports our hypothesis that *M. extorquens* most likely contains more sRNAs than what has been annotated.

The identification of sRNAs in bacteria is biased towards model organisms like *E. coli* simply because they are more heavily studied (Boutet *et al.*, 2022). We decided to compare the size distribution of IGRs in *E. coli* K-12 substr. MG1655 (NC_000913.3) with those in *M. extorquens* AM1 (NC_012808.1) focusing on gaps between 50 and 1000 nucleotides (Figure 4.3, C-D). This specific *Methylorubrum* strain was chosen since it is the most closely related sequenced strain to the one we used (ATCC55366). Seven distinct sRNAs are annotated in the genome of *M. extorquens* AM1 (ffh, ar45, CC2171, BjrC1505, BASRCI27, suhB and 5_ureB_sRNA). Small RNA sX4 (Table 4.1) was annotated within the genome of species from the genus *Methylorubrum*, but it was not found in our strain of interest (*M. extorquens* AM1). Small RNAs 5_ureB_sRNA, BASRCI27 and suhB are present more than once in the genome (with two, three and four copies respectively). Homology searches from Rfam covariance models from the entire sRNAs repertoire was performed on all available prokaryotic genomes in NCBI (Wheeler *et al.*, 2007), leading to more sRNA predictions in RiboGap than in Rfam (Figure 4.3, C-D), most likely due to Rfam thresholds generally more stringent than 0.0005. The size distribution of IGRs in the range of 50 and 1000 nucleotides for both proteobacteria has a similar pattern. In *E. coli*, approximately 65% of the annotated sRNAs are concentrated in the intergenic gaps between 50 and 400 nucleotides. Remarkably, only two *trans*-regulatory elements are annotated in the IGRs of that size in *M. extorquens*, reinforcing the idea that there is more to be discovered.

4.5.2 Expression of annotated sRNAs

We first wanted to confirm the expression of several annotated sRNAs in *M. extorquens* by Northern blot analysis of bacteria grown in 1% methanol. Three of the annotated sRNAs in *M. extorquens* AM1 were selected to be experimentally validated (ffh, CC2171, BjrC1505) (Table 4.1). To use as positive controls for Northern blots, probes for 5S RNA, transfer-messenger RNA (tmRNA) and the leucine tRNA were created as well, all of which were expected to be highly transcribed (Table 4.1). These would also act as a size guideline for the predicted sRNAs.

Hybridization was observed for all positive controls (5S RNA, tRNA-leu and tmRNA), confirming the proper transfer of the extracted RNA on the nitrocellulose membrane (supplementary material, Figure 9.1). Bands were also detected for all three sRNAs that were annotated in the genome of *M. extorquens* (ffh, CC2171, BjrC1505), validating for the first time the presence of sRNAs in this biotechnologically relevant bacteria (Figure 4.4, A). All hybridization experiments were done in triplicates (data not shown).

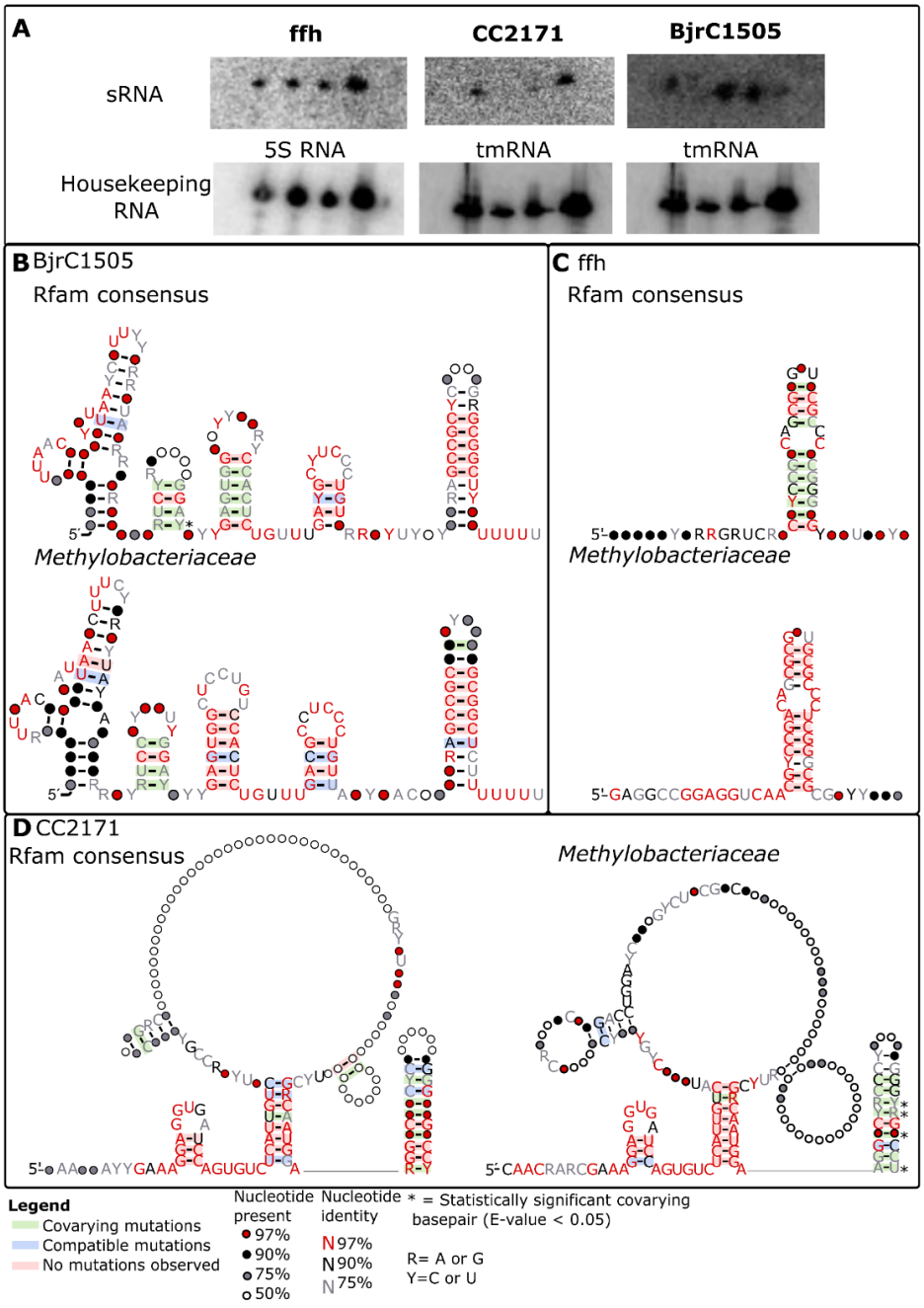

Figure 4.4 Structure of sRNAs validated by Northern Blot analysis

(A) Hybridization of radioactively labeled probes for annotated sRNAs (ffh, CC2171 and BjrC1505). The four bands correspond to different time points (24h, 48h, 72h, 96h). A housekeeping RNA was also hybridized on the same membrane as the sRNAs for control (5S RNA or tmRNA). The secondary structure for the sRNA BjrC1505, ffh, and CC2171 are depicted in **(B)**, **(C)** and **(D)** respectively. For each RNA, the consensus structure from all bacteria (Rfam consensus) and that from Methylobacteriaceae are shown. Base pairs with statistically significant covariation are represented by * (E-value <0.05) for Methylobacteriaceae secondary structure. Their significance was assessed with R-scape (Rivas *et al.*, 2001).

For every validated sRNA, the secondary structure from the *Methylobacteriaceae* family corresponds to the Rfam consensus, which is the taxonomy classification of *M. extorquens*. However, more sequence conservation and less covariation are observed since the species are more closely related (Figure 4.4, B-C-D). Statistically significant covarying base pairs in *Methylobacteriaceae* can be observed only for the secondary structure of CC2171 (Figure 4.4, D).

The sRNA CC2171 was first discovered by microarray analysis in the bacteria *Caulobacter crescentus*, but no condition affecting expression of this sRNA was identified (Landt *et al.*, 2008). This is the first instance where the expression of sRNA CC2171 was validated in an Alphaproteobacteria other than *C. crescentus* by Northern blot. The sRNA ffh was identified as a well-conserved motif found upstream of the *ffh* gene encoding for the cytoplasmic protein of the bacterial signal recognition particle (SRP) (Meyer *et al.*, 2009). It is widespread among Alphaproteobacteria with over 600 representatives. This study is the first to detect it by Northern blot, although its transcript was found in metatranscriptomic data from ocean water samples (Frias-Lopez *et al.*, 2008; Shi *et al.*, 2009). The transcripts of sRNA ffh and CC2171 were also detected in transcriptomic data from *Mesorhizobiym huakuii* 7653R, a plant-associated *Hyphomicrobiales* (Fuli et al., 2017). Finally, the sRNA BjrC1505 was shown to accumulate in the stationary phase of plant-associated Alphaproteobacteria from the family *Bradyrhizobiaceae* and *Rhizobiaceae* using both Northern blot and microarrays (Madhugiri *et al.*, 2012), though it was never detected in other bacteria.

4.5.3 Prediction of sRNA Candidates

4.5.3.1 sRNA-Detect

Beyond these newly confirmed, sRNAs, we were interested in discovering potential novel sRNAs. For this, we analyzed the transcriptome with sRNA-Detect (Peña-Castillo *et al.*, 2016). This data came from a transcriptomic study in another project. Briefly, *M. extorquens* strain ATCC55366 was genetically engineered to allow the accumulation of the tricarboxylic acid cycle (TCA) metabolite, succinic acid, using a Δ*sdhA gap20*::145 Δ*phaC*::Km^R triple mutant (Lamarche *et al.*, 2018). To investigate the impact of these mutations at the transcriptional level, RNA-seq data were acquired for the WT strain and the mutant at pH 6.5 and without pH control. We used the RNA-seq data from these three samples (each in triplicates) for the sRNA-Detect analysis, but without further focus on mutants vs. WT strains, unless otherwise mentioned in the text.

Inspection of the transcriptome using sRNA-Detect resulted in a list of 10,267 detected candidates from all three conditions. Most of these were repetitive among growth conditions, with approximately 3,500 potential sRNAs for each of them. This includes multiples hits for a single ncRNA (e.g., 15 candidates are found within the 23S rRNA sequence). Sequences with a predicted length of 50 to 250 nucleotides by sRNA-Detect within the main chromosome (NC_012808.1) for the WT strain were kept for further analysis (2,079 candidates).

4.5.3.2 Annotated RNAs among sRNA-Detect Candidates

Candidates were first inspected for the presence of already annotated RNAs. The list of all annotated RNAs within the genome of *M. extorquens* AM1 was obtained using RiboGap (Naghdi *et al.*, 2017). Among our list of presumptive regulatory elements, ribosomal RNAs (rRNAs), sRNAs, transfer RNAs (tRNAs) and *cis*-regulatory elements were found (Table 4.2). Importantly, we were able to recover six of the seven distinct sRNAs that are annotated in the genome of *M. extorquens* AM1 (all except 5_ureB_sRNA), confirming that sRNA-Detect is a reliable tool to detect sRNAs.

Table 4.2 Annotated RNAs within sRNA-Detect candidates

RNAs	Description	Strand	Rfam	sRNA-Detect ID		
				WT	Mutant (pH 6.5)	Mutant
Housekeeping RNAs						
tmRNA	transfer-messenger RNA	1	RF00023	1637	1596	1668
5S_rRNA	5S ribosomal RNA	-1	RF00001	2129,2165,2166, 2239 to 2241, 2353, 2354, 2366, 2367	2055, 2087, 2088, 2263, 2264	2207, 2250, 2251, 2449, 2450
Bacteria_small_SRP	Bacterial small signal recognition particle RNA	-1	RF00169	2572	(-)	2640
SSU_rRNA_bacteria	Bacterial small subunit ribosomal RNA	-1	RF00177	2137, 2173, 2360	2095, 2259, 2271, 2272	2258 to 2261, 2329 to 2332, 2440 to 2443, 2457

beta_tmRNA	Betaproteobacteria transfer-messenger RNA	1	RF01850	1289	1262	1306
LSU_rRNA_bacteria	Bacterial large subunit ribosomal RNA	-1	RF02541	2130, 2167, 2355, 2368	2056, 2089, 2091, 2158, 2255, 2265	2208 to 2210, 2252 to 2254, 2323 to 2325, 2434 to 2436, 2451 to 2453
alpha_tmRNA	Alphaproteobacteria transfer-messenger RNA	1	RF01849	1289	(-)	1307
sRNAs						
ffh	ffh sRNA	1	RF01793	556	(-)	(-)
CC2171	caulobacter sRNA CC2171	1	RF01867	553	552	559, 560
BjrC1505	Alphaproteobacterial sRNA BjrC1505	1	RF02356	1169	1143	1178
ar45	Alphaproteobacterial sRNA ar45	-1	RF02347	(-)	(-)	2682
suhB	Makes More Granules Regulator RNA (mmgR)	-1	RF00519	(-)	(-)	2849
BASRCI27	Brucella sRNA CI27	1	RF02600	(-)	(-)	1738
***cis*-regulatory element**						
cspA	cspA thermoregulator	1	RF01766	721	708	729
Cobalamin	Cobalamin riboswitch	-1	RF00174	(-)	2202	2377
Transfer RNAs						
tRNA-Ala	Alanine	1		1199, 2133, 2357, 2169	1170, 2057, 2092, 2159, 2256, 2268, 1596	1208, 1668, 2211, 2255, 2326, 2437, 2454
tRNA-Ala	Alanine	1		1199, 2133, 2357, 2169	1170, 2057, 2092, 2159, 2256, 2268, 1596	1208, 1668, 2211, 2255, 2326, 2437, 2454
tRNA-Arg	Arginine	1		831	803, 1629, 2227, 2228	1712, 2407
tRNA-Asn	Asparagine	1		1943	(-)	(-)

tRNA-Asp	Aspartic acid		689	(-)	686
tRNA-Gln	Glutamine	1	1844, 1854	1791	1912
tRNA-Glu	Glutamine acid	1	429, 654	(-)	438, 2438
tRNA-Gly	Glycine	1	1514	886, 1476	1535
tRNA-His	Histidine		(-)	2706	(-)
tRNA-Ile	Isoleucine	-1 / 1	2133, 2357, 2169	2057, 2092, 2159, 2256, 2268	1208, 2211, 2255, 2326, 2437, 2454
tRNA-Leu	Leucine	-1 / 1	2651, 1845	(-)	(-)
tRNA-Lys	Lysine	1	1399, 1997	1376, 1931	2064
tRNA-Met	Methionine	-1	3002	2879, 2880	3075
tRNA-Phe	Phenylalanine	-1	2785	(-)	2844
tRNA-Pro	Proline	1	1678, 1897	1629	1712
tRNA-Ser	Serine	-1	2063	1998	2130
tRNA-Thr	Threonine	-1	3399	3282	3524
tRNA-Trp	Tryptophan	1	1719	(-)	(-)
tRNA-Tyr	Tyrosine	1	1312	(-)	(-)
tRNA-Val	Valine	1	1037, 1192	1011, 1166	1040, 1203
tRNA-fMet	N-Formylmethionine	-1	2164, 2128, 2352	2054, 2086, 2262	2206, 2248, 2448
pseudo-tRNA		1	763	743	775, 1061, 2128

Numerous sRNA candidates were proposed along the genome of *M. extorquens* AM1 for the WT strain. The program sRNA-Detect provided a score for each candidate representing the average read depth coverage which is the sum of the reads mapped to each nucleotide of small transcripts divided by the length of such transcripts. It could therefore be interpreted as the level of expression of that RNA region. A cut-off of 1000 was determined to be acceptable since the mean score of all annotated RNAs was higher than this value (Figure 4.5).

Among potential sRNAs with a sRNA-Detect score higher than 1000 and a length between 50 to 250 nucleotides (388 candidates), 22 were arbitrarily selected to be tested experimentally (supplementary material, Table 9.1).

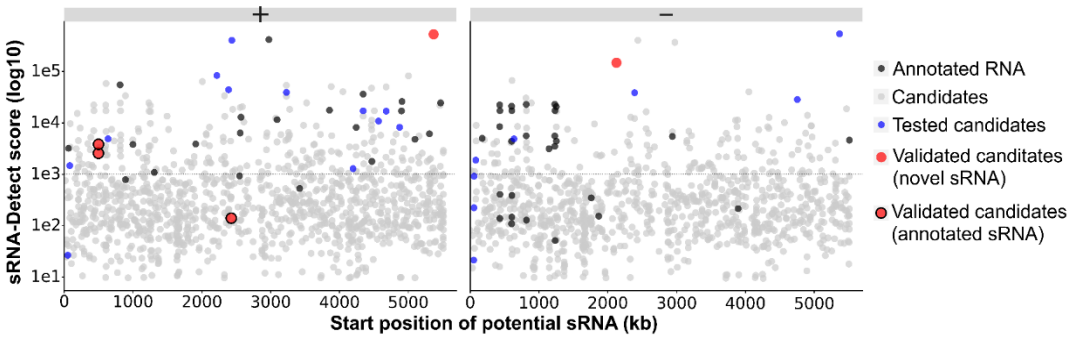

Figure 4.5 Putative sRNA candidates identified by sRNA-Detect

The x-axis represents the starting position of the putative RNA elements on the chromosome in kilobase pairs and the y-axis represents the score given by sRNA-Detect (log10). A cut-off of 1000 for this score is depicted by the dotted line. Already annotated RNAs (housekeeping RNAs, *cis*-regulatory elements and tRNAs from Table 4.2) with an E-value lower than 0.0005 are denoted in black. Blue dots illustrate candidates that could not be detected by Northern blot analysis. Red points show candidates where bands were observed with Northern blot analysis. Red dots with a black outline are sRNAs that were previously annotated, whereas red dots without an outline are novel sRNAs specific to this study. Information is divided into positive (left panel) and negative strand (right panel).

4.5.3.3 Detection of candidates sRNA2624 and sRNA1969 by Northern blot

Transcripts for sRNA1969 and sRNA2624 were detected by Northern blot analysis (Figure 4.6). Hybridization of the probes for both candidates was observed in triplicates (data not shown). When comparing their migration profile on a membrane with RNAs of known sizes, sRNA2624 is in between 250-300 nucleotides (supplementary material, Figure 9.1). Small RNA1969 is slightly larger than sRNA2624 (Figure 4.6).

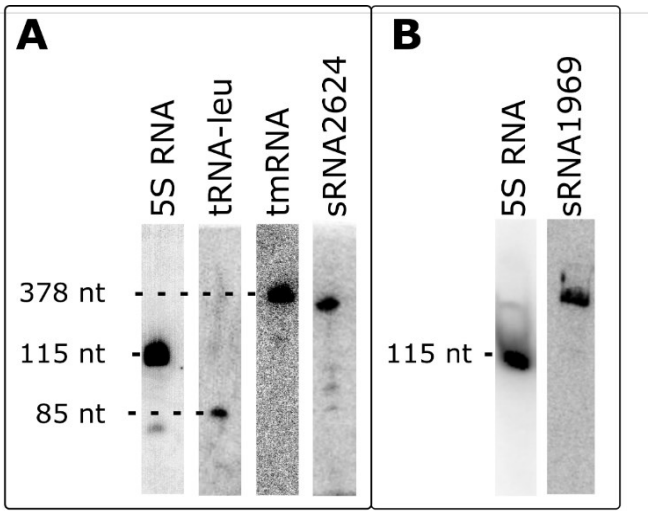

Figure 4.6 Expression of candidates sRNA2624 and sRNA1969 by Northern blot analysis

(A) Hybridization of the probe complementary to sRNA2624. **(B)** Hybridization of the probe complementary to sRNA1969. To evaluate sizes of sRNAs, probes for RNAs of known size were also hybridized on the same membranes (5S RNA, 115 nt; tRNA-leu, 85 nt and tmRNA, 378 nt).

4.5.3.4 Conserved genomic context of candidates sRNA1969 and sRNA2624 in *Methylobacteriaceae*

Using the BLASTn tool from NCBI (Altschul *et al.*, 1990) on the Reference Sequence (RefSeq) representative genome database (O'Leary *et al.*, 2016), we find that both sRNAs are only found in *Methylobacteriaceae* within *Methylorubrum* or *Methylobacterium* species (intergenic regions containing sRNA2624 and sRNA1969 can be found in supplementary material, Table 9.2). Candidates were therefore renamed as Methylo2624 and Methylo1969. In all *Methylorubrum* species encoding for Methylo2624, the sRNA was preceded by an aminotransferase and followed by a hypothetical protein (supplementary material, Figure 9.2). Methylo2624 was also found in *Methylobacterium* species, where it was preceded by either an aminotransferase or NAD (P)-dependent oxidoreductase and followed by a different hypothetical protein than that observed for *Methylorubrum* species (supplementary material, Figure 9.2).

Methylo1969 was detected in genomes of both *Methylorubrum* and *Methylobacterium* species following a BLASTn homology search in NCBI. However, only hits within *Methylorubrum* species were found (supplementary material, Table 9.2) after comparison with the RefSeq representative genome database. A review of the genus *Methylobacterium* was recently proposed (Green & Ardley, 2018) and unclassified *Methylobacterium* species remain (*Methylobacterium* sp. AMS5, DM1, NI91, CLZ) in which Methylo1969 is detected. Moreover, this sRNA is found in *Methylobacterium populi*, which should be denoted *Methylorubrum populi* following the

nomenclature suggestion given by Green and Ardley (Green & Ardley, 2018). Methylo1969 is most likely specific to *Methylorubrum* species where it is preceded by either a glutathione S-transferase N-terminal domain-containing protein or a DUF3280 domain-containing protein. In all cases, no proteins are annotated downstream of Methylo1969 within 300 nucleotides.

4.5.3.5 Predicted transcription start sites and terminators of sRNAs candidates

Transcription start sites (TSSs) had not been extensively studied in *Methylobacteriaceae* until Maucourt *et al* performed (Maucourt *et al.*, 2022) a genome-wide TSSs mapping in *Methylorubrum extorquens* DM4, a reference strain for the consumption of dichloromethane as a sole source of carbon (Maucourt *et al.*, 2022). Following a differential RNA-seq (dRNA-seq) approach, numerous TSSs were identified. Some were described as orphans when no genes were annotated within 250 nucleotides of the identified start site. Orphan TSSs can also be associated with non-coding or undetected genes (Maucourt *et al.*, 2022). Position of orphan TSSs determined by Maucourt *et al* (Maucourt *et al.*, 2022) agreed with our predicted location for Methylo1969 (shown in red, supplementary material, Table 9.2). All IGRs containing either sRNA were aligned with Clustal Omega (Sievers *et al.*, 2011) to identify the more conserved regions (sequences are underlined in supplementary material, Table 9.2). These conserved regions were then used to establish a covariance model with Graphclust (Heyne *et al.*, 2012). Given that Methylo2624 and Methylo1969 were only found in the family of *Methylobacteriaceae*, sequence conservation is relatively high, which limits covariation. Still, some covariations and compatible mutations could be observed within the predicted conserved structures for Methylo2624 and Methylo1969 (Figure 4.7).

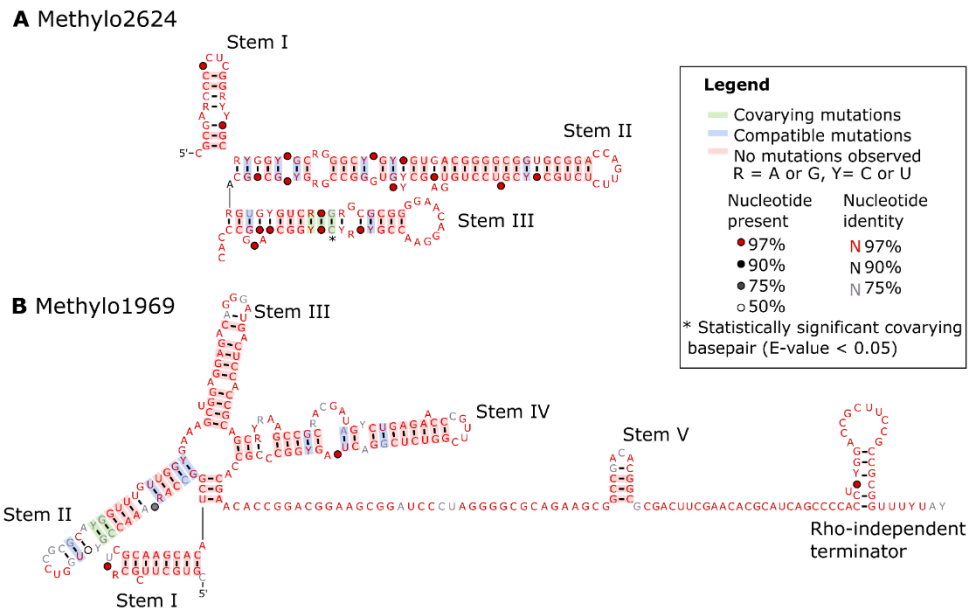

Figure 4.7 Secondary structure prediction of Methylo2624 and Methylo1969

The secondary structures of Methylo2624 **(A)** and Methylo1969 **(B)** were drawn by the program R2R (Weinberg & Breaker, 2011) using intergenic sequences from supplementary material, Table 9.2. Taken individually, covariation of a base pair in stem II of Methylo2624 is considered statistically significant according to R-scape (Rivas *et al.*, 2001) **(A)**, whereas none of the indicated covarying base pairs are considered statistically significant according to R-scape (Rivas *et al.*, 2001) for Methylo1969 **(B)**. A Rho-independent terminator is predicted in Methylo1969 using ARNold (Naville *et al.*, 2011).

A Rho-independent terminator was predicted in Methylo1969 with ARNold (Naville *et al.*, 2011). Since no terminators were predicted in Methylo2624, its size was validated using probe towards the 5′ and 3′ ends (supplementary material, Figure 9.3). Probes at 5′ and 3′ ends led to the same band pattern as for the probe used to detect Methylo2624, suggesting they were all hybridizing the same RNA (supplementary material, Figure 9.3 and Table 9.3).

4.5.3.6 Methylo2624 and Methylo1969: Coding or Regulatory RNAs?

The program RNAcode (Washietl *et al.*, 2011) was used to verify if Methylo2624 and Methylo1969 were putative coding sequences. Small open reading frames (ORF) are often missed by gene annotations for proteins, posing a challenge to identifying novel sRNAs. Software like RNAcode assess the coding potential of conserved regions to discriminate between protein coding and functional RNAs. A multiple sequence alignment containing all IGRs where Methylo2624 or Methylo1969 is found was created with ClustalW (Thompson *et al.*, 2003) and submitted to RNAcode (Washietl *et al.*, 2011). Results are represented in a list of hypothetical coding sequences with their associated score and P-value (supplementary material, Table 9.4). All identified potential coding sequence in a negative frame can be disregarded since we know the

transcribed strand coding for both sRNAs. RNAcode assigned a score to each presumed small ORF, where random non-coding regions generally do not have a scoring factor higher than 15 (Washietl *et al.*, 2011). For both sRNAs, none of the candidate ORFs had a scoring element higher than 15 or a P-value lower than the suggested limit of 0.01, suggesting they were not encoding for small proteins (supplementary material, Table 9.4).

To strengthen our claim, the secondary structures of Methylo2624 and Methylo1969 were analyzed with RNAz (Gruber *et al.*, 2010) to corroborate the idea that it is a functional RNA (supplementary material, Table 9.5). The same multiple alignment file submitted to RNAcode was provided to the RNAz program, leading to an "RNA-class probability" of 0.61 for both sRNAs, indicating that it is most likely to be a functional RNA. To make such forecast, RNAz considers the structure conservation index (SCI) and the thermodynamic stability (negative z-score). Functional RNAs are associated with a high SCI and thermodynamic stability, which were 0.93 and -1.28 respectively in the case of Methylo2624. For Methylo1969, these scores were 0.88 and -1.41 respectively. A SCI value close to 1 represents a perfectly conserved structure, whereas a measure close to 0 indicates that there is no consensus structure (Gruber *et al.*, 2010). A negative thermodynamic stability score (negative z-score) shows that the predicted secondary structure is more stable than a random one (Gruber *et al.*, 2010). RNAz analysis strengthen our hypothesis that Methylo2624 and Methylo1969 are functional RNAs.

4.5.3.7 Identification of potential sRNA targets

The tool CopraRNA (Wright *et al.*, 2014) was applied to detect mRNA targets of Methylo1969 (Table 4.3) using homologs of this sRNA in *Methylobacteriaceae* as an input (*Methylorubrum extorquens* DM4, *Methylorubrum extorquens* strain PSBB040, *Methylorubrum zatmanii* strain PSBB041 and *Methylorubrum populi* strain YC-XJ1). Since there are a limited number of *Methylobacteriaceae* genomes available in the Refseq database in NCBI, it can restrict the capacity of CopraRNA to identify potential targets because all homologs are only identified in *Methylobacteriaceae*.

Table 4.3 CopraRNA target predictions for Methylo1969

Rank	P-value	Locus tag / Gene	Annotation
1	$6.6 \cdot 10^{-5}$	metd_rs08835	DUF1775 domain-containing protein
2	$8.7 \cdot 10^{-5}$	metd_rs23635	YdiU family protein
3	$2.6 \cdot 10^{-4}$	metd_rs07695 / *ureG*	urease accessory protein UreG
4	$4.1 \cdot 10^{-4}$	metd_rs09310	F0F1 ATP synthase subunit delta
5	$5.2 \cdot 10^{-4}$	metd_rs27020 / *pgl*	6-phosphogluconolactonase
6	$6.7 \cdot 10^{-4}$	metd_rs22085	Hypothetical protein
7	$7.1 \cdot 10^{-4}$	metd_rs17685	Hypothetical protein
8	$1.0 \cdot 10^{-3}$	metd_rs19555	ABC transporter substrate-binding protein
9	$1.1 \cdot 10^{-3}$	metd_rs27120	WcbI family polysaccharide biosynthesis putative acetyltransferase
10	$6.4 \cdot 10^{-4}$	metd_rs26475	GcrA cell cycle regulator

The top 10 targets with the best P-values included multiple imprecisely identified mRNA targets of unknown function (DUF1775 domain-containing protein and hypothetical proteins). Methylo1969 also potentially targets a protein from the YdiU family and the gene *pgl*, known to be related to stress conditions in bacteria (Al Mamun *et al*., 2012; Yang *et al*., 2020). CopraRNA also suggested that Methylo1969 was linked to ATPase activity coupled with transmembrane movement of substances, such as for the ABC (ATP-binding cassette) transporter substrate-binding protein (Locher, 2009) and the FOF1 ATP synthase subunit delta (Deckers-Hebestreit & Altendorf, 1996). Other possible targets include *ureG*, a nickel-binding enzyme important for the hydrolysis of urea (Fong *et al.*, 2013) and *GcrA*, a cell cycle regulator. All interactions between Methylo1969 and the top 10 targets are shown in supplementary material, Figure 9.4. No targets were predicted for Methylo2624 using CopraRNA.

4.5.3.8 sRNA expression in different stress conditions

M. extorquens AM1 was cultivated under different stress conditions to assay the expression of

Methylo2624 and Methylo1969, and to potentially give insights into their function. As a control, 5S RNA was also assessed under the same condition (Figure 4.8). For example, 2% ethanol and 100 mM NaCl were added to the growth medium to evaluate the change in sRNA transcripts under ethanol tolerance and osmotic stress respectively (Gourion *et al.*, 2008). *M. extorquens* is often exposed to temperature changes due to fluctuations between day and night in its natural habitat on plant leaves and could be using sRNAs to regulate its gene expression in the face of changing weather (Green & Ardley, 2018). Therefore, different temperatures were tested (20 °C, 30 °C and 37 °C), with 30 °C being its optimal temperature. Finally, we tested the addition of different metabolites and molecules (1 mM fluoride, 30 mM urea and guanidine) and evaluated the impact of cobalt removal, which is normally part of its growth medium, since they are associated with riboswitches annotated within the genome of *M. extorquens*. Their concentration was determined following a toxicity assay, where we selected those that influenced the growth of *M. extorquens* without preventing it (data not shown). All stress conditions were tested with 1% methanol as a source of carbon, except for the ethanol tolerance condition.

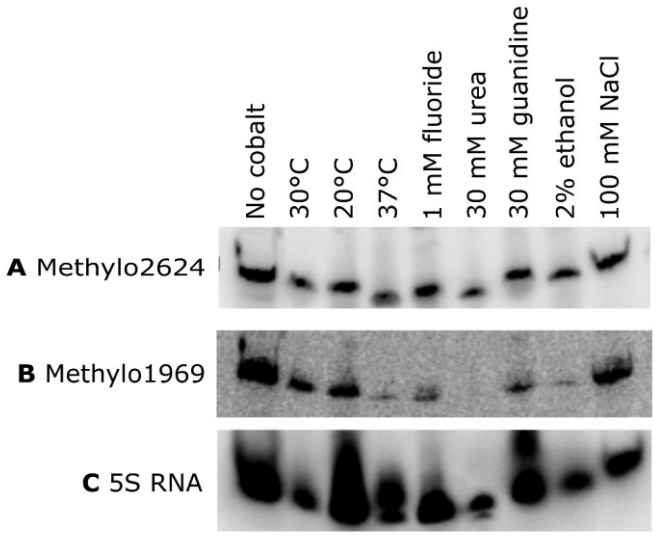

Figure 4.8 Expression of Methylo2624, Methylo1969 and 5S RNA in different growth conditions

Northern blot results in **(A)**, **(B)** and **(C)** were all produced on the same membrane. Each stress conditions contain 15 µg of RNA collected after 48 hours of growth. These results are representative of three Northern membranes (data not shown). Full membranes can be found in supplementary material, Figure 9.5.

Methylo2624's transcript remained the same in all tested conditions, suggesting it is expressed constitutively (Figure 4.8, A). *M. extorquens* was also grown with succinic acid as a source of carbon, and it did not impact Methylo2624's expression either (supplementary material, Figure 9.6). RNAs that are constitutively expressed are often associated with an organism's central metabolism, like DNA repair or protein synthesis (Hoopes, 2008), and are not typical

characteristic of sRNAs. Methylo2624 also distinguishes itself from sRNAs by its lack of A/U rich regions that could interact with the Hfq protein (Møller *et al.*, 2002a) or a predicted Rho-independent terminator (Livny & Waldor, 2007). Thus, it might not be surprising that CopraRNA does not predict any target for the sRNA Methylo2624. In contrast, Methylo1969's expression was greatly impacted by several stress conditions, namely growth at 37 °C, as well as growth with ethanol and urea (Figure 4.8, B). Interestingly, CopraRNA predicted that Methylo1969 may target *ureG*, a urease catalyzing the hydrolysis of urea to ammonia (Fong *et al.*, 2013). One could hypothesize that Methylo1969 prevents *ureG* mRNA translation. Indeed, in presence of urea Methylo1969 is not transcribed, allowing the hydrolysis reaction performed by UreG to occur. Some variations could be observed for the 5S RNA (Figure 4.8, C). Even if ribosomal RNAs are typically stable, conditions hindering cell growth have been linked to rRNA degradation (Basturea *et al.*, 2011; Okamura *et al.*, 1973; Sulthana *et al.*, 2016; Zundel *et al.*, 2009), so it is not surprising that some variations in the 5S RNA transcript could be observed under stress conditions.

4.6 Conclusion

This research is the first to demonstrate the expression of sRNAs in the C1 metabolism model organism *M. extorquens*. A list of potential sRNAs in *M. extorquens* was created which could be further analyzed to identify new and interesting candidates for future research. Importantly, this study highlights Methylo2624 and Methylo1969 as sRNAs specific to *Methylobacteriaceae*. Methylo2624 expression appears to be constitutive over several stress conditions. In contrast, Methylo1969 is affected in different growth conditions, most notably in presence of urea where its transcript could not be detected. Since it was predicted to target an mRNA encoding a urease, it would be interesting to further test this hypothesis with the deletion or overexpression of this sRNA to help elucidate its function and putative regulatory role. It is the first characterization of sRNAs in *M. extorquens*, paving the way for future discovery of other novel sRNAs within *M. extorquens*, contributing to the potential of *M. extorquens* as a biotechnological tool. This first assessment of sRNAs in the biotechnologically relevant bacterium *M. extorquens* could help lead to a better understanding of mechanisms of regulation within this bacterium and provide insight on how to promote production yield. For example, recently implemented genetic tools like synthetic sRNAs (Zhu *et al.*, 2021) or CRISPR-Cas9 (Mo *et al.*, 2020) could target newly discovered sRNAs and impact the output of a desirable product or the metabolism of a carbon source.

4.7 Material and Methods

4.7.1 Bioinformatics selection of candidates

4.7.1.1 RNA-sequencing data

A DASGIP® parallel bioreactor system (Eppendorf) equipped with 1.5 L reactor vessels was used to grow the *M. extorquens* wild-type strain ATCC55366 at pH 6.5 and its isogenic ΔsdhA gap20::145 ΔphaC::KmR triple mutant at pH 6.5 and without pH control (Lamarche *et al.*, 2018), each in biological triplicates, for a total of nine fermentation runs. Precultures were prepared as follows: two 3 L baffled Erlenmeyer containing 400 mL of CHOI Medium 4 (Bourque *et al.*, 1995), supplemented with 0.3% malic acid, were inoculated with cells harvested from freshly grown agar plates. Malic acid was supplemented to the growth culture to compensate for the ΔsdhA mutation that is interrupting the TCA cycle. Kanamycin was exclusively added to triple mutant precultures (40 µg/mL). Precultures were incubated overnight at 30 °C under an agitation of 250 rpm. Then, precultures were used to seed reactors to obtain initial optical densities (600 nm) of approximately 0.25. Each reactor contained 1 L of CHOI Medium 4 (Bourque *et al.*, 1995) supplemented with 0.3% malic acid. Antifoam xiameter 0.07% were used during fermentations, without antibiotics. Air flow rate was set at 35 sL/h whereas dissolved oxygen was kept at 30% using solely an agitation cascade as no pure oxygen supply was needed. The temperature was set 30 °C and the pH was maintained using phosphoric acid (1M) and ammonium hydroxide (28%). Methanol concentration was also kept constant at 0.2% (v/v) using a methanol sensing and reading system (Intempco; Montréal, QC) coordinated with DASGIP feeding pumps. Commercial methanol was used (J.T. Baker®, HPLC grade). Reactors were run for 22-24 hours and achieved similar optical densities ($2.7 \leq x \leq 3.4$). Then, RNA extractions were performed using MasterPure™ RNA Purification Kit (Epicentre®) according to the manufacturer protocol. Absence of DNA contamination was confirmed by PCR. Furthermore, to test RNA quality, all samples were submitted to Agilent 2100 Bioanalyser using Prokaryote Total RNA Nano Chips. Then, samples were sent to the Centre d'innovation Génome Québec et Université McGill (Montréal, QC) for the preparation of KAPA rRNA depleted libraries and Illumina® sequencing (HiSeq V4 – PE 125 pb sequencing lane).

4.7.1.2 sRNA-Detect

The input for the sRNA-Detect (Peña-Castillo *et al.*, 2016) is RNA-seq data aligned to a reference genome (SAM file). The genome of *M. extorquens* AM1 was taken from the NCBI database as a reference (NC_012808.1). The output is a list of potential sRNAs in a gene transfer format (GTF). To identify potential sRNAs, this method highlights RNA sequences that have a minimum depth coverage within a given range. Selected features also demonstrate low depth variation across their whole sequence. Previously annotated regions were not classified as candidates. The complete protocol is described by Peña-Castillo *et al.* 2016. The sRNA-Detect workflow is accessible at http://www.cs.mun.ca/~lourdes/site/Welcome.html. A list of potential candidates was obtained for all three growth conditions of our RNA-seq experiment: WT, mutant and mutant with controlled pH (6.5). For all candidates, the following information was available: start and end positions, expected length, strands and sRNA-Detect score. As with any method, there is a risk of false positives, so it is important to consider other criteria. In this article for example, candidates Methylo2624 and Methylo1969 were further supported by the analysis of its genomic context and conservation, as well as predictions of transcription start sites. Moreover, its coding potential was evaluated with bioinformatic programs like RNAcode (Washietl *et al.*, 2011) and RNAz (Gruber *et al.*, 2010) to support that it is a functional RNA.

4.7.1.3 RiboGap

This database can be accessed via a web server (www.ribogap.iaf.inrs.ca) (Naghdi *et al.*, 2017). It allows for the easy retrieval of information regarding intergenic sequences of prokaryotes. Information on RNAs in RiboGap is extracted from the Rfam database (Kalvari *et al.*, 2021). Results are therefore limited to annotations within the Rfam database. A complete list of all annotated RNAs (including rRNAs, sRNAs and tRNAs) found in *M. extorquens* AM1 was created using RiboGap (Version 2). The RNA of interest was specified on the interface in the section ‘’type’’ of RNA family, whereas the organism of concern was selected using their corresponding accession number. Only annotated RNAs with an E-value lower than 0.0005 (as determined by Infernal (Nawrocki & Eddy, 2013)) were kept for comparison with sRNA-Detect candidates. This database was also used to retrieve all intergenic regions from *E. coli* K12 and *M. extorquens* AM1 to determine size distribution and presence of sRNAs. Finally, a list of all Alphaproteobacteria with annotated sRNAs was obtained from RiboGap.

4.7.2 Bioinformatic Analysis of Candidates

Several bioinformatic tools were used to further characterize Methylo2624 and Methylo1969. Intergenic sequences containing the sRNA sequences were obtained from BLASTn tool results from NCBI (Altschul *et al.*, 1990) on the Reference Sequence (RefSeq) representative genome database (O'Leary *et al.*, 2016) (supplementary material, Table 9.2). The representative genome database gives hits from genomes each representing a clade cluster (O'Leary *et al.*, 2016). Information on proteins encoded upstream and downstream of both candidates were also extracted from NCBI (Wheeler *et al.*, 2007) (supplementary material, Figure 9.2). All sequences were aligned using the Clustal Omega interface (https://www.ebi.ac.uk/Tools/msa/clustalo/) (Sievers *et al.*, 2011) to identify conserved regions (underlined sequences in supplementary material, Table 9.2).

To assess for the coding potential of our candidates sRNAs, the conserved region containing all intergenic regions with Methylo2624 and Methylo1969 was submitted to RNAcode (version 0.3) (Washietl *et al.*, 2011). The input of this program is an alignment file in the Clustal Omega format (Sievers *et al.*, 2011). To validate it was a functional RNA, the same alignment file was submitted to RNAz (Gruber *et al.*, 2010), which is accessible as a tool in the Galaxy suite (Afgan *et al.*, 2018). Information regarding the Rho-independent terminator was obtained with ARNold web interface (http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/) (Naville *et al.*, 2011).

To predict the secondary structure of Methylo2624 and Methylo1969, a covariance model was created with the Graphclust (Heyne *et al.*, 2012) workflow on the Galaxy platform (Afgan *et al.*, 2018). The consensus structure drawing was generated with R2R (Weinberg & Breaker, 2011) using the alignment of the corresponding covariance model generated by Graphclust (Heyne *et al.*, 2012) with all intergenic regions containing the regulatory RNA (from supplementary material, Table 9.2). The alignment was generated with the cmalign tool from infernal (Nawrocki & Eddy, 2013). For sRNAs CC2171, ffh, BjrC1505, the Rfam consensus structure was created using the Stockholm file from Rfam. The *Methylobacteriaceae* consensus was produced from the alignment of their covariance model with all sequences from this family containing the desired regulatory element from a FASTA file generated by RiboGap and aligned with cmalign (Nawrocki & Eddy, 2013). To evaluate whether the covarying basepairs were significant, the RNA multiple sequence alignment in a Stockholm format for each structure was submitted in R-scape (http://eddylab.org/R-scape/) (Rivas *et al.*, 2001).

Several figures in this article are represented with the help of ggplot2 (Wickham, 2016) package within the Jupyter notebook (Kluyver *et al.*, 2016). The programs RNAcode (Washietl *et al.*, 2011), Infernal (Nawrocki & Eddy, 2013) and R2R (Weinberg & Breaker, 2011) were ran in the Graham server from Compute Canada. RNAz (Gruber *et al.*, 2010) and Graphclust (Heyne *et al.*, 2012) were accessed via the Galaxy suite (Afgan *et al.*, 2018). All other programs are accessible online via the provided links.

4.7.3 Northern blot analysis

4.7.3.1 Growth conditions of *M. extorquens*

Cultures of *M. extorquens* ATCC55366 were grown in 250 mL baffled Erlenmeyer flasks at 30 °C and 200 rotations per minutes (rpm) in CHOI Medium 4 as described in Bourque *et al*, 1995. The CHOI growth medium corresponded to 1/5 of the volume of the baffled Erlenmeyer flask to allow for proper oxygenation. Methanol (1%) was added to the growth medium as a source of carbon and supplied every 24 hours (0.5%). The optical density at 600 nm was taken by spectrophotometry with Eppendorf BioSpectrometer ®. The bacterial pellets were stored at -80 °C before RNA extraction.

For *M. extorquens* grown under stress conditions, cobalt chloride was removed from the CHOI Medium 4 (Bourque *et al.*, 1995) for the condition without this metal (supplementary material, Table 9.5).Ethanol (1%) was added as a sole source of carbon instead of methanol (2%) for the condition under ethanol tolerance. For osmotic stress, sodium chloride (100 mM) was supplemented to the growth medium (Gourion *et al.*, 2008). Other stress conditions included the addition of 1 mM potassium fluoride dihydrate (FH_4KO_2), 30 mM urea and 30 mM guanidine. To assess the impact of temperature, cultures were placed at 20 °C or 37 °C instead of the normal growth conditions at 30 °C. Bacteria were grown for 48 hours, and the optical density was measured before storing the bacteria pellet at -80 °C prior to RNA extraction.

4.7.3.2 RNA Extraction

The bacterial pellets previously stored at -80 °C were lysed with 100 µL of a solution of 400 µg/mL of lysozyme in TE buffer (0.5 M EDTA and 1 M Tris-HCl, adjusted pH of 8.0) for 5 minutes at room temperature. RNA was extracted using TRIzol reagent (Invitrogen) as described in Rio *et al.*, 2010 (Rio *et al.*, 2010). The RNA contained in the supernatant was precipitated by adding 2 volumes of 100% chilled ethanol and 0.1 volumes of 3 M sodium acetate (pH 5.2) and cooled at -80 °C for at least 2 hours. The RNA was centrifuged at 4° C for 30 minutes at 14,000 rpm. The supernatant

was discarded and 500 µL of 70% chilled ethanol was added to rinse the pellet and then centrifuged at 4° C for 5 minutes at 14,000 rpm. The supernatant was removed, and the pellet was left to dry for at least 15 minutes before being resuspended in RNAse-free water. The extracted RNA was quantified using a Nanodrop (Thermoscientific Nanodrop 2000).

4.7.3.3 Northern Blot

Fifteen micrograms of total RNA for each sample were migrated into a 10% denaturing 8 M urea polyacrylamide gel (PAGE) to separate the RNA according to its size. Samples were loaded with 2 X gel loading dye (0.05% bromophenol blue, 0.05% xylene cyanol, 10 mM EDTA pH 8, 95% formamide) with TBE 1 X (0.09 M Tris-base, 1 mM EDTA, 0.09 M boric acid) as running buffer. The gel was migrated at 15 W for 1 hour. The RNA in the gel was transferred overnight to a nitrocellulose membrane with a positive charge (GE healthcare Amersham^{TM} Hybond^{TM} -N+) by capillary transfer using an assembly of Whatman® filter paper. The capillary transfer was set up as follows: 10X SSC (1.5 M sodium chloride and 0.15 M sodium citrate dihydrate) was poured into a container. Four Whatman® filter paper and a nitrocellulose membrane were cut to the length and width of the polyacrylamide gel. Another Whatman® filter paper was cut the same width, but its length was long enough to touch the buffer when placed on a support. They were all pre-soaked into 10X SSC buffer for 30 minutes prior to the assembly. The Whatman® filter papers, the nitrocellulose membrane and the polyacrylamide gel were all stacked one on top of the other.

The RNA was left to transfer from the polyacrylamide gel to the nitrocellulose membrane overnight. The next morning, the membrane was dried for a few minutes. To fix the RNA unto the membrane, shortwave UV light was used (UV stratalinker 2400 Stratagene). The membrane was stained with a methylene blue solution (0.02% methylene blue and 0.3 M sodium acetate pH 5.5) for 10 minutes with agitation to verify proper transfer of the RNA. The membrane was rinsed with distilled water for at least one hour. As the excess coloration was washed from the membrane, the bands corresponding to the highly abundant transferred RNA were revealed (data not shown).

4.7.4 Hybridization of probes corresponding to candidate sRNAs

4.7.4.1 Radiolabelling of the DNA probe

For each candidate, a 50-nucleotides sequence complementary to the potential sRNA was selected in the middle of the sequence. Probes ordered from Integrated DNA technologies (IDT) were radiolabelled at the 5′ end with γ-^{32}P ATP. A reaction was prepared with 0.5 µM of DNA probe, 1 X kinase buffer PNK, 20 µCi γ-^{32}P ATP in a final volume of 20 µL. The labeling reaction

was left to incubate for 1 hour at 37 °C. The labeled probes were purified on a 6% denaturing gel (8 M urea PAGE, polyacrylamide gel electrophoresis). Loading dye 2 X and 1 X TBE was used as described before. The gel was exposed with phosphor imaging screens for 5 minutes before being scanned with a Typhoon™ FLA9500 (GE Healthcare Life Sciences). The bands corresponding to the probes were cut out of the gel and conserved at -20 °C for future work. Intensity of radioactive bands was quantified with ImageJ (Abràmoff *et al.*, 2004).

The nitrocellulose membrane with the transferred RNA was pre-incubated with 15 mL hybridization buffer (20 X SSC, 50 X Denhardt's solution 2% Bovine Serum Albumin (BSA), 2% Ficoll 400, 2% Polyvinylpyrrolidone (PVP), 10% sodium dodecyl sulfate (SDS), 100 µg/mL salmon sperm Invitrogen, Thermofisher scientific) in a rotating hybridization oven at 42 °C for 1 hour in a flask. After pre-incubation, the gel fragment containing a radiolabelled DNA probe was added inside the flask and left to incubate overnight in a rotating oven. The next day, the gel fragment and the hybridization buffer were recovered from the flasks and stored at -20 °C for future work. The same probe can be used for many experiments if the DNA probe is still radioactive. The nitrocellulose membrane was washed in four steps as follows: 1 minute with washing solution I (2 X SSC buffer, 0.1% SDS), 5 minutes with washing solution I, and twice for 10 minutes with washing solution II (0.2 X SSC buffer, 0.1% SDS). All washing steps were performed in the rotation oven at 42 °C. The washed membranes were wrapped into Saran plastic wrap and exposed on phosphor imaging screens overnight before being scanned by a Typhoon™ FLA9500 (GE Healthcare Life Sciences). Intensity of radioactive bands were quantified with ImageJ (Abràmoff *et al.*, 2004). Membranes containing RNA can be used several times with different probes, if they are washed between each candidate with washing solution III (0.1X SSC buffer, 0.1% SDS) for 2 hours at 80 °C. To ensure that the membranes were cleaned, they were exposed in phosphor imaging screens as before. If radioactivity was still present, the last washing step was repeated. Membrane were stored in Saran wrap plastic between uses.

4.8 Data availability statement

Data available on request from the authors

4.9 Competing interests

The authors declare there are no competing interests.

4.10 Funding

EB was supported by Fondation Armand-Frappier, Natural Sciences and Engineering Research Council (NSERC) and Fonds de Recherche du Québec Natures and Technologies (FRQNT). This research project was supported by grants from CRIBIQ, Mitacs and NSERC (418240-2012-RGPIN and RGPIN-2019-06403) to JP.

Author's Contributions

EB carried and conceived most of the experiments and wrote the manuscript. SD and KT helped with the bioinformatic analysis. KDSN helped with stress conditions assays. MJL and MW set up fermentation experiments for RNA-seq. MGL performed RNA extractions for RNA-seq. RI, MGL and CBM contributed with the experiment design and initiated the research project. JP supervised the project and revised the manuscript.

Acknowledgments

The authors would like to thank Jessie Muir for language proofreading the manuscript and the laboratory of Charles Greer for assistance and sharing of equipment. The authors would also like to acknowledge Aurélie Devinck, Quetia Joseph and Philip Loranger for the revision of this manuscript. This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and Compute Canada (www.computecanada.ca).

Supplementary material is available in annexe II (le matériel supplémentaire pour cet article est disponible en annexe II).

5 CHAPITRE 3: SHIFTED-REVERSE PAGE: A NOVEL APPROACH BASED ON STRUCTURE SWITCHING FOR THE DISCOVERY OF RIBOSWITCHES AND APTAMERS

***Shifted-Reverse* PAGE : une nouvelle approche basée sur les changements de structure pour la découverte de *riboswitches* et d'aptamères** (Traduction française)

Auteurs :

Aurélie Devinck[1*], **Emilie Boutet**[1*], Jonathan Ouellet[1,2], Rihab Rouag[1], Balasubramanian Sellamuthu[1], Jonathan Perreault[1]

* co-premières auteures

[1] INRS – Centre Armand-Frappier Santé Biotechnologie, Laval, Qc, H7V 1B7, Canada, [2] Maintenant à l'Université Monmouth, NJ, USA.

Journal : BioRxiv (journal non révisé par les pairs)

Soumission : 26 juillet 2022

DOI : https://doi.org/10.1101/2022.07.26.501614

Journal : *Nature Methods* (journal révisé par les pairs)

Soumission : 5 août 2022

Contribution des auteurs :

AD et **EB** ont contribué de façon égale à ce travail. JP a conçu la méthode SR-PAGE. AD, **EB**, JO, RR, BS et JP ont optimisé la méthode. JO a réalisé la majorité des cycles de SELEX avec les séquences dérivées du *riboswitch* TPP. AD et **EB** ont réalisé toutes les autres expériences présentées dans l'article. Les figures ont été créées par JP, JO, AD et **EB**. Le manuscrit a été rédigé par JP, AD et **EB**. Le projet a été supervisé par JP. Le manuscrit a été révisé par tous les auteurs.

5.1 Liens entre les objectifs et ce chapitre de la thèse

L'annotation des ARNnc chez *M. extorquens* a un grand retard par rapport à celle des gènes (Figure 4.2). Les *riboswitches* annotés dans le génome de *M. extorquens* proviennent d'analyses bioinformatiques basées sur la génomique comparative. Bien que ce soit une approche très puissante, elle est limitée par la disponibilité des séquences et les annotations génomiques, une restriction notable lorsqu'on travaille avec des organismes avec une intensité de recherche moins importante comme *M. extorquens*. Ainsi, les *riboswitches* annotés chez *M. extorquens* sont essentiellement des *riboswitches* avec une très large distribution, soit ceux liant la vitamine B12, le fluor, la guanidine, la glycine et le TPP (information tirée de RiboGap (Naghdi *et al.*, 2017)). Plus de 2000 exemples du *riboswitch* guanidine-I sont annotés dans quatre phylums bactériens (Reiss *et al.*, 2017). Le *riboswitch* liant la vitamine B12 est la famille la plus répandue chez les bactéries (Barrick & Breaker, 2007). Les *riboswitches* TPP et fluor sont aussi répandus chez les bactéries et ils sont même retrouvés en dehors du domaine bactérien chez les archées et/ou les eucaryotes (Barrick & Breaker, 2007; Speed *et al.*, 2018). Plusieurs *riboswitches* fluor sont annotés chez *M. extorquens* par RiboGap, mais ils ont en moyenne un mauvais E-value (0.2) et ils ne sont pas dans Rfam, ce qui pourrait suggérer une nouvelle sous-classe à découvrir et caractériser. Le *riboswitch* glycine est aussi largement distribué chez les bactéries et il a été retrouvé chez les bactéries à Gram négatif et positif (Barrick & Breaker, 2007). Les méthodes basées sur la génomique comparative ont permis d'annoter les *riboswitches* les plus répandus, mais ceux qui restent à découvrir sont probablement rares et donc difficiles à identifier avec ces techniques bioinformatiques. Il pourrait être possible d'identifier de nouvelles familles de *riboswitches* plus rares et spécifiques à certains genres, comme le *riboswitch* SAM-VI retrouvé seulement chez les *Bifidobacterium* (Mirihana Arachchilage et al., 2018) ou le *riboswitch* NAD-I qui, comme plusieurs autres *riboswitches*, est trouvé dans un seul ordre, celui des *Acidobacteriales* (Malkowski et al. 2019).

Étant donné que les annotations d'ARNnc chez *M. extorquens* sont apparemment incomplètes, la recherche de nouveaux *riboswitches* chez cette bactérie apparaît justifiée. *M. extorquens* peut utiliser comme source d'énergie des composés à un seul carbone comme le méthanol (C1) ou des substrats multi-carbones comme l'éthanol (C2), le pyruvate (C3) ou le succinate (C4) (Šmejkalová *et al.*, 2010). Lors de la croissance avec des composés multi-carbones, le cycle de Krebs est utilisé pour le métabolisme du carbone, alors que c'est plutôt le cycle de la sérine qui est employé en présence de C1 (Peyraud *et al.*, 2012). L'expression des enzymes nécessaires à

l'utilisation spécifique de ces sources de carbones est régulée strictement (Šmejkalová et al., 2010), potentiellement avec l'aide de *riboswitches*. Par exemple, le *riboswitch* ZMP est associé au contrôle de gènes en lien avec le métabolisme des C1 (Kim *et al.*, 2015). Comme il a été mentionné dans la problématique, les outils disponibles pour la découverte de *riboswitches* ont plusieurs limites. Nous avons donc développé une nouvelle méthode expérimentale pour la recherche de *riboswitches* appelée le SR-PAGE. Cet article présente la validation de cette nouvelle technique qui pourra être utilisée pour découvrir de potentiels *riboswitches* chez *M. extorquens.*

5.2 Résumé (traduction française)

Les *riboswitches* sont des ARNnc composées d'un domaine aptamère capable de lier un ligand et d'une plateforme d'expression. La liaison d'un ligand au domaine d'aptamère cause un changement de structure secondaire de l'ARN, ce qui impacte l'expression du gène en aval. Les méthodes bioinformatiques actuelles pour leur découverte ont diverses limitations. Pour les contourner, nous avons développé une méthode expérimentale pour découvrir de nouveaux *riboswitches* appelée SR-PAGE (*Shifted Reverse Polyacrylamide Gel Electrophoresis*). Cette nouvelle technique tire avantage du changement de structure secondaire de l'ARN à la suite de l'interaction avec un ligand dans un gel de polyacrylamide natif afin d'identifier un potentiel *riboswitch*. Des *riboswitches* connus ont été testés avec leurs ligands correspondants afin de valider cette méthode. De plus, le SR-PAGE a été utilisé comme méthode de sélection au sein d'un SELEX pour sélectionner des séquences dégénérées du *riboswitch* TPP dont la préférence de liaison a été modifiée pour favoriser la thiamine plutôt que le TPP. La technique du SR-PAGE permet de réaliser un large criblage en cherchant dans plusieurs organismes simultanément et en testant plus d'un ligand à la fois.

5.3 Abstract

Riboswitches are regulatory sequences composed of an aptamer domain capable of binding a ligand and an expression platform that allows the control of the downstream gene expression based on a conformational change. Current bioinformatic methods for their discovery have various limitations. To circumvent this, we developed an experimental technique to discover new riboswitches called SR-PAGE (Shifted Reverse Polyacrylamide Gel Electrophoresis). A ligand-based regulatory molecule is recognized by exploiting the conformational change of the sequence

following binding with the ligand within a native polyacrylamide gel. Known riboswitches were tested with their corresponding ligands to validate our method. SR-PAGE was imbricated within an SELEX to enrich switching RNAs from a TPP riboswitch-based degenerate library to change its binding preference from TPP to thiamine. The SR-PAGE technique allows for performing a large screening for riboswitches, searching in several organisms and testing more than one ligand simultaneously.

5.4 Introduction

Microbial gene expression has to be tightly regulated to adapt to constant changing environments (Hennecke, 1990). Cells are able to import molecules from the extracellular environment and their concentration can be crucial to modulate enzymatic pathways (Hennecke, 1990). A specific element of the mRNA mostly found in the 5′ untranslated region (UTR) called riboswitch can bind these molecules (ligand) which leads to a change in their secondary structure. This change will affect transcription or translation of the downstream gene. The affected genes often encode for the synthesis or the transportation of the bound molecule (Serganov & Nudler, 2013). These riboswitches thus comprise an aptamer domain and an expression platform, the structure of which depends on the binding of the aptamer's cognate ligand. To date, more than 50 riboswitches specifically binding with different metabolites and ligands have been reported in bacteria (Barrick *et al.*, 2004; Mandal *et al.*, 2003; Mandal *et al.*, 2004; McCown *et al.*, 2017; Ramesh & Winkler, 2010; Sherlock *et al.*, 2018b; Winkler *et al.*, 2002b). Interestingly, the thiamine pyrophosphate (TPP) riboswitches were also observed in archaea, plants and fungi (McCown *et al.*, 2017) .

The discovery of most riboswitches was made possible by the availability of sequenced genomes for numerous bacteria combined with powerful bioinformatic approaches (Stav *et al.*, 2019; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2017b; Weinberg *et al.*, 2010). Potential riboswitches are identified based on the presence of sequence covariation and conservation since the aptamer domain of these regulatory molecules is highly conserved. Covariation occurs when changes in the nucleotide sequence does not affect the RNA secondary structure, reinforcing the importance of its formation (Weinberg *et al.*, 2010). However, bioinformatic tools are limited by sequence availability and their annotations within databases. Also, an alignment of three sequences is theoretically enough to provide covariation that will support a secondary structure prediction and sufficient to warrant experimental evaluation of a putative riboswitch. However, in practice, alignments leading to discovering new riboswitches typically had dozens to hundreds of sequences. The identification of the ligand recognized by the

aptamer domain is often facilitated by the nature of the downstream gene. Problematically, gene annotation is sometimes missing or misleading and does not always provide the necessary information to identify the metabolite recognized by the riboswitch. A list of orphan riboswitches for which the associated ligands are currently unknown resulted from bioinformatics-based discovery of numerous putative regulatory RNA structures (Greenlee *et al.*, 2018; Stav *et al.*, 2019; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2017b; Weinberg *et al.*, 2010). From the current approach for riboswitch discoveries, it is expected that a very small number of all riboswitch classes have been unveiled (Breaker, 2011). The remaining riboswitches would unfortunately be scarcer, making them much harder to be discovered by bioinformatic tools. Thus, a method which does not rely on sequences available in public databases might be the best way to circumvent this problem.

Some experimental methods have been developed by other groups in recent years to discover new riboswitches. Term-seq enables to screen all the transcription termination within chosen organisms using a high-throughput sequencing approach (Dar *et al.*, 2016). This method can be used to identify regulations involving a premature stop of transcription as can be found with some riboswitches. However, this method omits all riboswitches that have an impact on translation. Also, Term-seq cannot differentiate transcription termination caused by a riboswitch from transcription termination caused by a protein, which could be derived from indirect regulation. Another experimental method that can be used to discover new riboswitches is Parallel Analysis of RNA Confirmations Exposed to Ligand Binding (PARCEL) (Tapsin *et al.*, 2018). It is an *in vitro* method to detect structural changes of natural RNA in presence of a ligand by comparing the targeted sites of the RNAse V1, an enzyme that cleaves base-paired regions. A change in the degradation pattern of the RNAse V1 between tested conditions is a good indicator that the ligand induces a change of conformation in the RNA, which is characteristic of riboswitch (Tapsin *et al.*, 2018). However, this technique is limited to regions that are accessible by RNase and only one genome can be analyzed at a time. Another technique derived from SHAPE-MaP (selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling) (Siegfried *et al.*, 2014) was developed to investigate RNA molecules that interact with ligands. It was recently used to screen hundreds of ligands for their ability to bind RNA molecules (Zeller *et al.*, 2022). Nevertheless, it could also be used to screen for riboswitches as with PARCEL, but with similar limitations, except perhaps that changes in RNA secondary structure upon binding is resolved to the nucleotide level.

Aptamers for various ligands have successfully been identified based on SELEX (Selective Evolution of Ligands by Exponential Enrichment) (Irvine *et al.*, 1991), the first ones being RNA aptamers selected against T4 DNA polymerase (Tuerk & Gold, 1990) and various organic dyes (Ellington & Szostak, 1990). The general idea is to start with a mix of an extremely diverse library of sequences (~10^{16} sequences) and a target of interest. Unbound sequences are separated from those showing an affinity for the target by a selection method. Selected sequences are amplified for the next round of enrichment. Although the general idea remains the same, many selection methods have been used within a SELEX to isolate aptamers, including capillary electrophoresis (CE-SELEX) (Zhu *et al.*, 2019), capture-SELEX (Boussebayle *et al.*, 2019) and magnetic bead SELEX (Espelund *et al.*, 1990) to name a few (reviewed in (Bayat *et al.*, 2018; Darmostuk *et al.*, 2015)). We hypothesized that a SELEX-based approach could also be applied to discover riboswitches. Nonetheless, there are certain restrictions preventing the use of typical SELEX experiments to find natural riboswitches. These regulatory sequences are often engulfing their ligand, guaranteeing that any attempt to immobilize the ligand on a column would cause steric hindrance. Even when riboswitches do not envelop their ligand, the chemical groups chosen for immobilization are likely to be among the groups bound by the potential riboswitch. As for methods that immobilize the library instead of the ligand (e.g., capture SELEX (Boussebayle *et al.*, 2019)), it would be difficult to use them to find riboswitches because users need to define a capture sequence within the random region, which is not possible for a library containing unknown natural riboswitches.

We have developed a novel method capable of getting over these obstacles, addressing at the same time the disadvantages of other experimental techniques. This method relies only on the two main characteristics of riboswitches: their ability to bind a ligand and their conformational change upon such binding. These characteristics are evaluated by electrophoresis approach analogous to 2D gels, but instead of running through a second dimension, electrodes are reversed. During the second migration, the RNA runs backwards in presence of a ligand. This allows the detection of a conformational change of the RNA upon binding with a ligand compared to RNA that does not interact with one. We named our technique Shifted Reverse PolyAcrylamide Gel Electrophoresis (SR-PAGE). Our method was validated with known riboswitches by verifying the shifting ability of several constructions. The power of this method was also demonstrated by selecting for a thiamine aptamer from a degenerate library of TPP riboswitches with the SR-PAGE imbricated within a SELEX as a selection method.

5.5 Results

5.5.1 Validating SR-PAGE with known riboswitches

SR-PAGE can identify riboswitches by taking advantage of the change in the conformation of the RNA sequence following binding with its cognate ligand within a native polyacrylamide gel. Briefly, the method is composed of two native gel migrations: during the first one, the potential regulatory RNA sequences migrate from the wells towards the positive electrode getting separated based on their size and their structure (Figure 5.1a). After 24 hours of migration, the gel is unmolded, and the solution of ligand(s) is sprayed on the gel to allow interaction with the RNA molecules (Figure 5.1b). The second migration is then started with the same condition as the first migration, but the polarity of the gel is reversed, forcing all RNA sequences to migrate back to the starting point (Figure 5.1c). Upon binding with the ligand, the RNA is subjected to a conformational change, either assuming a more compact form that would allow it to migrate past the well line or creating a more relaxed conformation that would slow down its migration in the gel (Figure 5.1c). SR-PAGE can recognize the ligand-based regulatory molecule by selecting those that had a change in migration after the addition of the ligand.

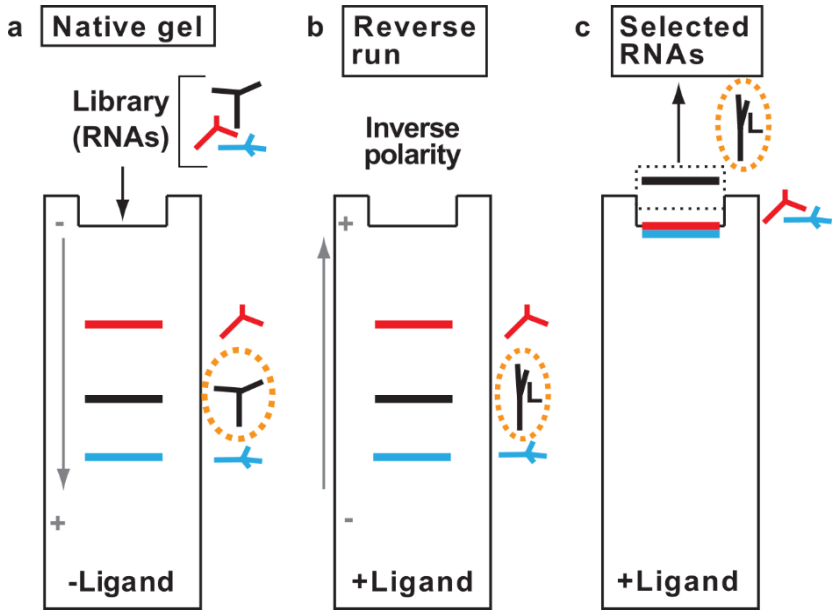

Figure 5.1 Overview of SR-PAGE

(A) An RNA library is loaded on a native gel and migrated in absence of ligand. RNAs are separated based on structure and size. **(B)** The top glass plate is taken off and the ligand is sprayed on the gel, which only changes the conformation of corresponding riboswitches (within the orange dotted circle) upon binding. The plate is put back on and migration is started with inverted polarity. **(C)** The gel is run until the RNA comes back to its starting point. RNAs that interact with the ligand will have a change of migration due to a change in their secondary structure upon ligand binding. A more detailed schematic of the SR-PAGE method is available in the supplementary material (Supplementary Figure 10.1).

As a proof of concept, we first wanted to validate that a change in migration upon binding of the ligand could be observed using SR-PAGE with known riboswitches. Up to five different constructions of six known riboswitches were designed, namely the known riboswitches for the following metabolites: flavin mononucleotide (FMN) (Pedrolli *et al.*, 2015), fluoride (Breaker, 2012), c-di-GMP (type I) (Sudarsan *et al.*, 2008) , nickel-cobalt (NiCo) (Furukawa *et al.*, 2015), thiamine pyrophosphate (TPP) (Rentmeister *et al.*, 2007) and glycine (Kwon & Strobel, 2008) . The sequence for the glycine riboswitch was based on results by Kwon et al. where a shift was observed in a native gel during migration with glycine (Kwon & Strobel, 2008). Their constructions varied in length, ranging from only the aptamer to incorporating different portions of the expression platform (Figure 5.2abc; Supplementary Table 10.1).

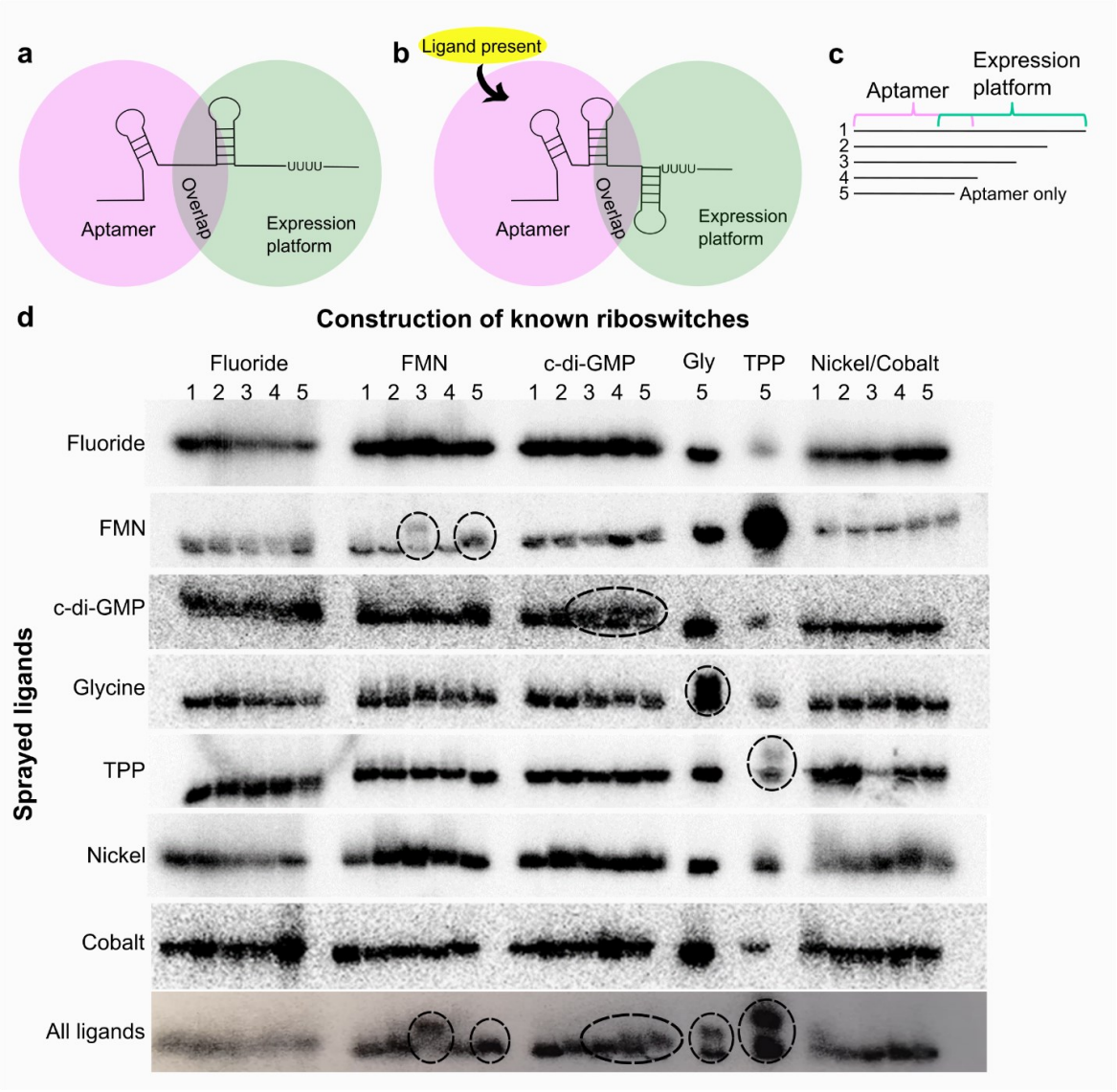

Figure 5.2 Validation of SR-PAGE with known riboswitches

(A) Schematic representation of a riboswitch in its unbound conformation, with the aptamer domain and expression platform emphasized in pink and green respectively. The overlap between those two regions is shown. **(B)** Schematic of a riboswitch in its bound conformation, where a ligand is bound to the aptamer domain, leading to a change in the secondary and tertiary structure. As an example, the change caused the formation of a Rho-independent transcription terminator in this case. **(C)** Representation of how the different constructions (1 to 5) of the known riboswitches were created. **(D)** Results from all SR-PAGE. Each group of horizontal bands corresponds to a different gel, where distinct ligands were sprayed into the gel before the second migration (only fluoride, FMN, c-di-GMP, glycine, TPP, nickel, cobalt or all ligands). Five different radiolabeled constructions for each known riboswitch were tested, except for the TPP and glycine riboswitch, where only the aptamer construction was assayed. Observable shifts in migration are circled.

Multiple SR-PAGE experiments were performed, where only one ligand at a time was sprayed at a concentration 100-fold higher than their respective known dissociation constant (K_D) (Figure 5.2d). Only the riboswitch constructions corresponding to the sprayed cognate ligands interacted

with the metabolites, resulting in a shift of migration (circled in Figure 5.2d). RNA molecules that did not interact with the sprayed ligands returned to the starting point in a straight line.

When FMN was sprayed, a significant change in migration for FMN_3 and a slight shift for FMN_5 constructs were observed (Figure 5.2d). Small changes in migration for c-diGMP_3, c-di-GMP_4 and c-di-GMP_5 constructs were also detected when the corresponding ligand was sprayed (Figure 5.2d). A change in aptamer migration was also observed for the glycine and TPP aptamers (Figure 5.2d) when the corresponding ligands were sprayed. Interestingly, the identical shift in migrations was observed when all ligands were sprayed at the same time on the gel, indicating that SR-PAGE can be used to select for multiple riboswitches with different ligand affinity at the same time (Figure 5.2d).

5.5.2 SR-PAGE to investigate structure changes and expression platforms

We also hypothesized that SR-PAGE could be a means to investigate the expression platform of a riboswitch. It could be used to select for the construction that results in a bigger change in migration, implying it has a bigger conformational change. FMN_3 and FMN_5 were further analyzed to understand why different shift levels were observed between constructions of the same riboswitch: they technically both have similar ability to interact with the metabolite, since they are constituted of the same aptamer domain (Figure 5.3). The Mfold web service (Zuker, 2003) was employed to analyze the free energy of both constructions in their bound and unbound forms. To mimic the bound conformation, sequences were forced to adopt the FMN aptamer domain structure (Serganov *et al.*, 2009) using constraints (constraints used to force the formation of the aptamer can be found in Supplementary Table 10.2). The equilibrium constant was then measured from the difference in free energies between the two conformations.

For FMN_3, the RNA is only slightly more stable in its unbound conformation (unconstrained) compared to when constraints are used to force the aptamer conformation, with free energies of -68.9 kcal/mol and -68 kcal/mol respectively (Figure 5.3a). Our interpretation is that during the first migration, the FMN_3 assumes its more stable unbound conformation (Figure 5.3a, left) which can potentially interconvert with the aptamer containing structure (Figure 5.3a, right), with a preference of 5:1 for the former compared to the latter, according to the calculated equilibrium constant. However, when the FMN ligand is sprayed unto the gel, not only does the equilibrium change, but the “aptamer state” of the RNA (capable of interacting with the ligand) also likely has changes in its tertiary structure. The change in structure from the more stable one (on the left) to the aptamer state results in the shift of migration observed with SR-PAGE (Figure 5.3a). Even if

the so-called unbound state should be favored, since the free energy of both structures is similar, the presence of the ligand suffices to permit the switch. The same situation seems to apply for FMN_5 (Figure 5.2b), with free energies of -50.70 kcal/mol and -48.70 kcal/mol for the unbound and bound conformations, respectively. However, the two structures appear closely related, which may explain the less prominent shift. Previous work also demonstrated that the aptamer of the FMN riboswitch is pre-formed in the absence of the metabolite at physiological concentrations of magnesium (Vicens *et al.*, 2011). In this case, in spite of a Keq suggesting a ratio of ~40:1 of the most stable vs. the aptamer state, the closeness of both structures may permit an easier switch in equilibrium, but also a more complete one, as suggested by the quasi-absence of a non-shifted band for FMN_5.

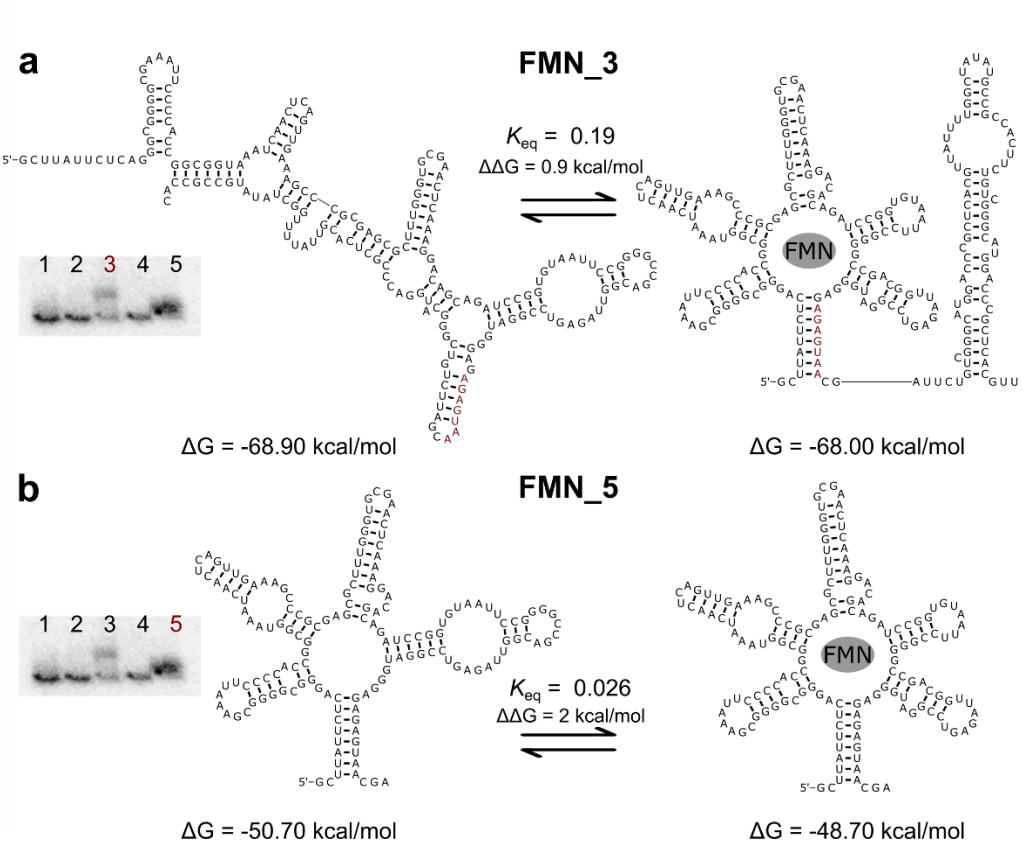

Figure 5.3 Shifting constructions 3 and 5 of the FMN riboswitch have similar free energies for their bound and unbound states.

(A, B) Predicted structures, free energies (ΔG) and equilibrium constant (Keq) of the unbound RNA and ligand-bound state of FMN_3 and FMN_5, respectively. Nucleotides emphasized in red can basepair to form the basal stem of the aptamer, leading to a mutually exclusive conformation. Shifts observed by SR-PAGE are shown, with corresponding construction number highlighted in red for **(A)** and **(B).**

Both FMN_3 and FMN_5 structures are more stable in their unbound state, so the equilibrium constant favors in each case this conformation (in absence of ligand), represented by a Keq value smaller than 1. Some of the RNA molecules were still able to adopt the ligand-binding conformation, since the Keq value is close to 1. A similar analysis of the free energies and equilibrium constants of ligand-bound state compared to free RNA for all constructions of riboswitches FMN, fluoride, c-di-GMP and nickel-cobalt is available in the supplementary material (Supplementary Figure 10.2-Figure 10.5). The larger differences in free energies observed for all the constructions of fluoride and nickel-cobalt riboswitches may explain that none of the tested constructs shifted (Supplementary Figure 10.2-Figure 10.5). Moreover, the equilibrium constant for most of these cases had very small values, meaning that the “aptamer state” is likely almost absent. As for c-di-GMP_3-4-5, FMN_3 and FMN_5, most free energies are very similar with Keq closer to 1, which explains that the constructs all seem to shift, but the shifts are more subtle given the similarity between the predicted bound and unbound states.

5.5.3 Selection of thiamine aptamer from degenerate libraries of TPP riboswitch

As a proof of concept that the SR-PAGE can be used as a method of selection in a SELEX, the aptamer sequence from the TPP riboswitch was partially randomized to create three libraries possessing each 4,096, 1,048,576 and 68,719,476,736 possible sequence combinations (Figure 5.4a). Libraries 1, 2 and 3 contained 6, 10 and 18 degenerated nucleotides respectively. Briefly, RNA libraries were run on native polyacrylamide gel electrophoresis during SR-PAGE. Thiamine was added to the second run (reversed polarity) of SR-PAGE, first at 2 mM (generations 1 to 4) and in later rounds of selection, at a lower concentration of 20 µM (generation 5) and 200 nM (generations 7 to 10) to increase the stringency. Shifted RNAs were purified and amplified, then loaded on gel together with TPP (as a negative selection ligand) for the next round of selection, where thiamine is again added only for the reverse migration. We therefore selected for RNA molecules that preferentially bind thiamine rather than its native TPP. After four, five, seven and ten rounds of selection, selected sequences were cloned and also analyzed by Illumina sequencing (Bentley *et al.*, 2008) to identify the enriched sequences. RNAs were selected based on their shifting abilities during the SR-PAGE (Figure 5.4b). A smaller dissociation constant (K_D) represents a higher affinity of a molecule to its ligand. While the wildtype aptamer has a K_D smaller than 10 nM for TPP and of 4 µM ± 1.9 µM for thiamine in our tested conditions (Figure 5.4c,d), selected sequences using SR-PAGE as a selection method kept their binding affinity for thiamine and lost affinity for TPP through rounds of SELEX (Figure 5.4f). K_D of candidates were determined

by in-line probing (Weinberg *et al.*, 2010) . One of the best clones (clone 13) showed a slightly improved binding for thiamine molecules, with a K_D of 3 µM ± 1.3 µM for thiamine and 7.8 µM ± 3.5 µM for TPP (Figure 5.4c,d,e). Other clones (1, 2, 9, 13, 23, 24, 25, 41, 42, 43, 44, 45, 46 and 47) also showed that the aptamer recognize thiamine with a better affinity than TPP molecules (Figure 5.4f and Supplementary Table 10.3). Only clones where a K_D could be estimated from in-line probing are shown in the Figure 5.4. Sequences of all selected clones as well as their K_D with thiamine and TPP when measurable can be found in Supplementary Table 10.3 (nomenclature is explained in Supplementary Figure 10.6).

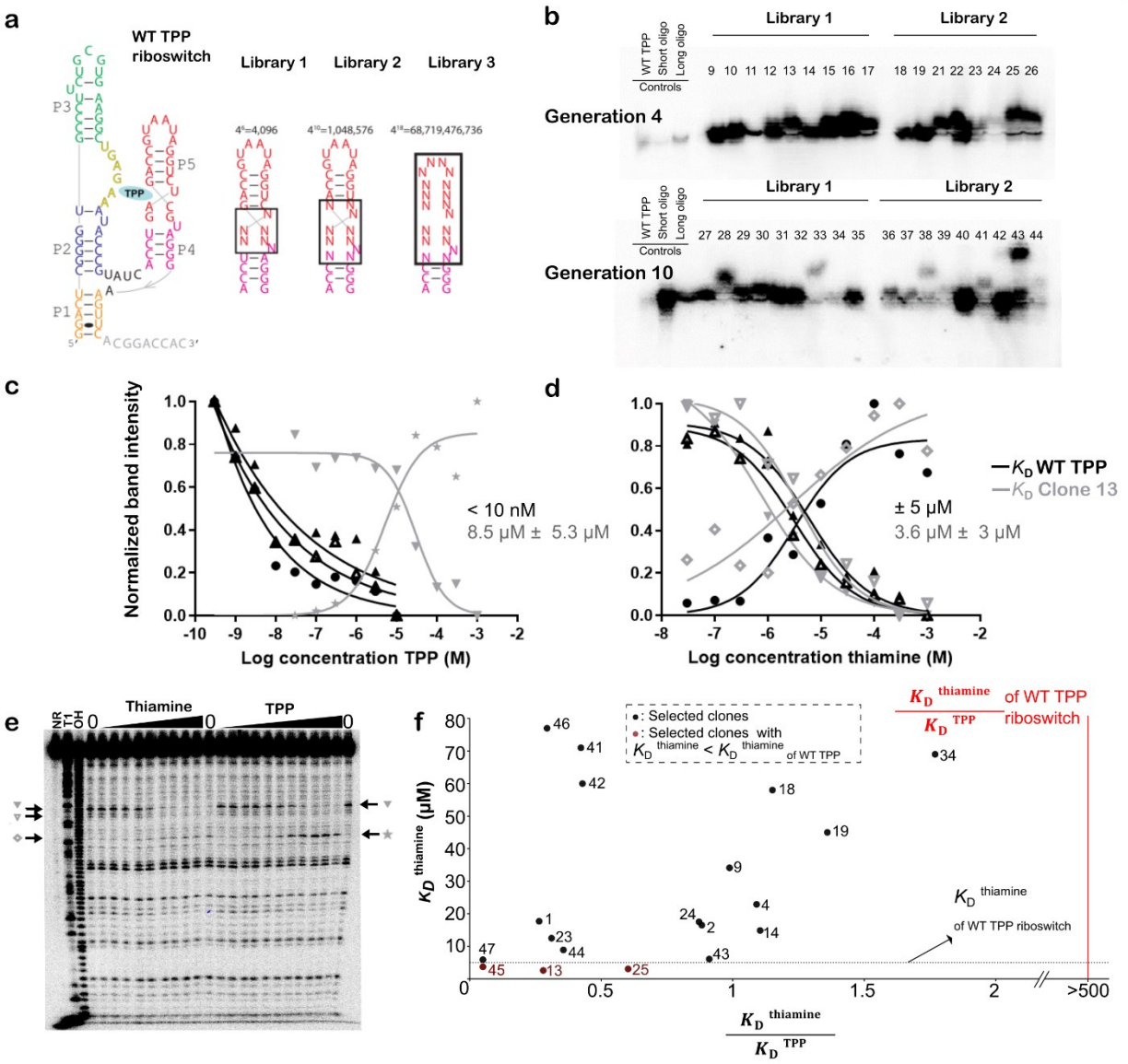

Figure 5.4 Selection of a thiamine aptamer from degenerated libraries of the TPP riboswitch using the SR-PAGE as a selection tool within a SELEX

(A) Degenerate libraries of the TPP riboswitch were created. Libraries 1, 2 and 3 had 6, 10 and 18 degenerated nucleotides respectively within stem P5. **(B)** Shifting abilities as a result of SR-PAGE from libraries 1 and 2 with clones selected during generations 4 and 10 of the SELEX. TPP was present during the first migration within the running buffer (10 µM), and thiamine (20 µM) was sprayed on the gel before the second migration. RNA molecules can be observed since they are radioactively labeled. As controls, a short and long oligos as well as the WT TPP riboswitch were migrated on the same gel. **(C)** K_D curve for WT TPP riboswitch (in black) and selected sequence 13 from library 1 after four generations of SELEX (in gray) with TPP. **(D)** K_D curve for WT TPP riboswitch (in black) and selected sequence 13 from library 1 after four generations of SELEX (in gray) with thiamine. **(E)** Inline probing gel of clone 13. The arrows represent the bands used for K_D calculation (graphed in **C**). Tested concentrations for thiamine and TPP ranged from 30 nM to 1 mM. **(F)** Affinity of all selected sequences with thiamine and TPP. All dots represent individual selected sequences, where those emphasized in red are clones with a better affinity for thiamine than the WT TPP riboswitch. The dotted horizontal line depicts the K_D of the WT TPP riboswitch for thiamine. The vertical red line represents the ratio of K_D for thiamine over K_D for TPP of the WT TPP riboswitch.

In library 3, the P5 stem of the WT TPP riboswitch was initially replaced by random nucleotides, therefore only a small proportion of sequences are expected to form a similar stem, and none are expected to have the WT internal loop, since we do not select for TPP binding (Figure 5.4a). After four and seven rounds of SELEX, selected libraries were analyzed by Illumina sequencing. Two of the clones (45 and 47) that we tested from this initial library show no binding to the TPP molecule up to a concentration of 1 mM (Figure 5.4f). In spite of the fact that we completely degenerated stem P4-P5 (except for two bp), we apparently still selected for a stem (Supplementary Figure 10.7a). Overall, secondary structure prediction for enriched sequences (i.e., more than three reads/sequence) showed that we selected for the formation of a stem in P5, as opposed to random sequences that do not typically form a stem (Supplementary Figure 10.7 and Supplementary Table 10.4). Because the thiamine binding is not greatly improved (as suggested by K_D similar to that of WT), this rather suggests that the stem was selected because it improved the shift between unbound and thiamine-bound states. Indeed, angles between stems are known to critically impact how RNA runs in a native PAGE (Lafontaine *et al.*, 2001). One concern of using the SR-PAGE within a SELEX was the impact of the adapters necessary for the PCR amplification of the selected sequences in a SELEX round. Validation of SR-PAGE with known riboswitches was done without the addition of adapters at the extremities of each sequences to allow for PCR amplification (Figure 5.2). Our results show that the adapters can indeed prevent riboswitch shifting using our positive controls, perhaps by preventing the proper formation of the adapters (Supplementary Figure 10.8). The addition of oligonucleotides complementary to the adapters helps to overcome this problem (Supplementary Figure 10.8). All sequences for the aptamers and oligonucleotides can be found in the supplementary materials (Supplementary Table 10.1).

5.6 Discussion

In this paper, we presented the Shifted Reverse PAGE (SR-PAGE), a novel method to study riboswitches and aptamers going beyond the limits of bioinformatics and other existing experimental methods. This technique was validated with several known riboswitches, namely c-di-GMP-I, FMN, TPP and glycine, including when multiple ligands were used at the same time. We also demonstrated that the SR-PAGE could be imbricated within an SELEX as a method of selection, allowing us to identify degenerated TPP riboswitch sequences that lost their affinity for TPP and improved their affinity for thiamine, as confirmed by in-line probing.

A shift in migration for all known riboswitches was not always detectable, like for the fluoride and nickel-cobalt riboswitches for example (Figure 5.2d). This can be explained by the tested constructions. Indeed, as demonstrated with the FMN_3 and FMN_5 constructions (Figure 5.3), a few nucleotides make the difference between an RNA that shifts and an RNA that does not shift, likely due to a subtle equilibrium of structures between the conditions with and without ligands. Thus, a slight difference in sequence can result in a big change in terms of secondary structure. Even if all tested constructions technically contained the aptamer domain, it does not necessarily mean that the aptamer was forming. We hypothesized that a shift was not observed for all known riboswitches simply because the RNA sequence resulting in a bigger change of secondary structure upon ligand binding was not tested, or likely because the difference in free energy between the unbound and bound structure was too large to be overcome in other cases. The difference in predicted free energies was in average of 1.78 kcal/mol in constructions where a shift in migration was observed (FMN_3, FMN_5, c-di-GMP_3-4-5; Figure 5.2d), whereas it was on average 12.1 kcal/mol for those where no change in migration was seen (FMN_1-2-4, c-di-GMP_1- 2, NiCo_1-2-3-4-5, fluoride_1-2-3-4-5; Figure 5.2d). In these latter cases, the free energy difference between the unbound state and the ligand-bound state is too large to be overcome by the ligand. The equilibrium constant values in cases where no shift was observed were also small (typically < 10^{-3}), suggesting that the secondary structures probably could not adopt the ligand-bound conformation and/or that the energy of the ligand-binding was insufficient to overcome the gap between ΔGs. A shift was observed in cases where the equilibrium constant was closer to 1, suggesting the RNA structure could overcome the difference in free energy to be able to bind the ligand. This is the induced-fit model (Heppell *et al.*, 2011), where the addition of the ligand favors a change in secondary structure of a given RNA molecule. On another end, the addition of the ligand could favor RNA molecules that are already in the proper secondary structure, which is the conformational selection model (Haller *et al.*, 2011). In this case, the dynamic interchange of structures within the gel would be locked into the “aptamer state” upon addition of ligand to explain the change in migration. It is likely that combinations of both models occur in many cases, including also more subtle changes of the tertiary structure which could nevertheless lead to detectable shifts in the gel.

The amplitude of the shift (the height of the shift) appears to correlate with the level of structural difference between the unbound and the bound structure, where similar secondary structure result in a smaller change in migration. This was observed for example in FMN_5, where the shift was less pronounced than that of FMN_3, since it was already in a secondary structure closely related to that of the aptamer (Figure 5.3), as well as c-di-GMP constructions, which were similar for the

most part. Therefore, one of the limitations of SR-PAGE is that we might miss riboswitches that are able to bind a ligand, but the conformational change is not drastic enough to result in a detectable change in migration. However, this limitation can be used to our advantage, since the SR-PAGE can be used to identify constructions that will produce larger shifts and gain insight on the expression platform, further improving our knowledge of existent riboswitches and aptamers, as suggested with the analysis of the FMN riboswitch construction (Figure 5.3).

We have also used SR-PAGE within a SELEX strategy to select riboswitches with a modified affinity for their ligands. The SR-PAGE selected RNA molecules based on two properties: first, for their ability to preferentially bind thiamine rather than TPP molecules, and for a change in the RNA conformation involving a shift by SR-PAGE upon that binding. More than forty sequences identified from that SELEX were tested separately by in-line probing. Out of all these sequences, 14 showed a better affinity for thiamine compared to that of the TPP molecule (K_D thiamine / K_D TPP 500 for WT). Moreover, three clones (13, 25 and 45) had a better affinity for thiamine than the WT TPP riboswitch (5 µM). This experiment successfully demonstrated that SR-PAGE could be used as a selection step within a SELEX to select for affinity modified and/or enhanced aptamers.

Now that the SR-PAGE method has been validated, it could be employed to identify potential novel riboswitches starting with genomic libraries of intergenic sequences of bacteria, for example. The validation of the SR-PAGE technique presents a new approach to discover riboswitches, which has advantages over the normally used bioinformatics methods. Only intergenic sequences are studied with bioinformatics analysis, whereas SR-PAGE allows researchers to look at the entire genomic sequences. Less complex RNA secondary structure can be overlooked when using bioinformatics, which SR-PAGE would not disregard. For example, mini-ykkC (aka guanidine-II) was deemed too simple to be a riboswitch until it was assayed with guanidine (Sherlock *et al.*, 2017). Moreover, novel regulatory structures will be directly linked to their ligand, contrary to bioinformatics, which often leads to orphan riboswitches, where the specificity of the ligand is not known (Greenlee *et al.*, 2018; Meyer *et al.*, 2011; Weinberg *et al.*, 2007; Weinberg *et al.*, 2017a; Weinberg *et al.*, 2010). The SR-PAGE selects for sequences that change conformation upon binding of a ligand. Therefore, potential riboswitches still need to be validated in-vivo using for example riboswitch-reporter fusion assay to confirm it has an impact on the downstream gene expression. Nevertheless, the use of SR-PAGE is not intended to overshadow bioinformatics techniques, but rather to complement them, since both have their advantages. This technique can be applied to several types of ligands such as coenzymes, amino

acids and metal ions. Each new discovery will provide new insights into the biochemical capacity of RNA and open opportunities for technological advances.

5.7 Methods

5.7.1 PCR construction of riboswitches

Forward and reverse primers were designed to amplify the different constructions of the FMN and the fluoride riboswitches from the genomes of *Escherichia coli* (*E. coli*) and *Burkholderia thailandensis* (*B. thailandensis*) respectively, with the addition of the T7 promoter. For the glycine, TPP, nickel-cobalt and c-di-GMP class I riboswitches, PCR assembly was used to create the template using Primerize (Tian & Das, 2017). The different constructions of the c-di-GMP riboswitch were then created in the same manner as for the FMN and the fluoride riboswitches. All oligonucleotides used are listed in Supplementary Table 10.1. The PCR reactions were done with 2.5 U Hot Start Taq Polymerase (Qiagen/203203) with Hot Start buffer (1X), dNTPs (200 µM), forward and reverse primers (1 µM) completed with Milli-Q water. For assembly PCR, the end primers have a concentration of 1 µM and the internal oligonucleotides have a concentration of 0.1 µM.

Three libraries derived from *E. coli thiM* TPP riboswitch were designed to obtain an aptamer that binds thiamine much better than TPP. To do so, the pyrophosphate binding pocket within P5 was randomized to different degrees with degenerated primers (Figure 5.4a) to create the libraries 1, 2 and 3. These oligonucleotides are also listed in the Table 10.1.

5.7.2 *In vitro* transcription

For SR-PAGE, all PCR products of the known riboswitches as well as those from the clones derived from the SELEX with degenerated TPP riboswitches were radiolabeled with [α^{32}P] UTP (Perkin Elmer) during *in vitro* transcription using T7 RNA polymerase for 3 hours at 37 °C (20 µL PCR products [~400 ng], 20 µL 5X transcription buffer [400 mM HEPES-KOH, pH 7.5, 120 mM MgCl2, 10 mM spermidine, 200 mM DTT], 0.001 U pyrophosphatase (Roche), 0.5 µL RNAse inhibitor Ribolock (Thermo Fischer/EO0382) [40 U/ µL], T7 polymerase [100 U], 5 mM ATP, 5 mM GTP, 5 mM CTP, 1 mM UTP and 5 µCi [α^{32}P] UTP). The volume was completed to 100 µL with Milli-Q water. The transcribed RNA was precipitated at -80°C with 0.1 volume of sodium acetate 3 M (pH 5.2) and 2 volumes of 100% chilled ethanol for at least 2 hours. The precipitated RNA was resuspended in RNAse-free Milli-Q water and purified on a 6% denaturing polyacrylamide

gel electrophoresis (8 M urea PAGE). Samples were loaded on the gel with 2X denaturing loading buffer (0.05% bromophenol blue, 0.05% xylene cyanol, 10 mM EDTA [pH 8], 95% formamide). The expected size bands corresponding to our RNA were eluted in the elution buffer (0.3 M NaCl) and the eluate was precipitated as described before. Radiolabeled RNAs were resuspended in 250 µL of Milli-Q sterilized water.

For in-line probing, RNA molecules were radiolabeled at their 5' end. *In vitro* transcription was performed as before, but with equal concentrations of all rNTPs, since radioactive [α^{32}P] UTP were not used. After an *in vitro* transcription, 5' ends were dephosphorylated using 5 U of Antarctic phosphatase (New England BioLabs/M0289S). A mix containing the RNA sample (~10 pmol), Antarctic phosphatase buffer 1X and 0.5 µL RNase inhibitor Ribolock (Thermo Fischer/EO0382). The volume was adjusted to 20 µL with sterilized Milli-Q water. The reaction was incubated at 37 °C for 20 minutes and deactivated at 65°C for 5 minutes. RNA molecules were radiolabeled at their 5' end in a reaction with ~10 pmol dephosphorylated RNA samples, 0.2 µL T4 Polynucleotide Kinase (PNK) enzyme (New England BioLabs/M0201S), 1 µL 10X T4 PNK buffer and 2 µL [γ^{32}P] ATP (Perkin Elmer). The volume was adjusted to 10 µL with water and incubated at 37°C for 1 hour. Labeled RNA samples were purified on a 6% denaturing PAGE, imaged and purified as described before.

5.7.3 SR-PAGE preparation

A slab gel system, typical of sequencing gels, was used. To mold the gel, we used glass plates of dimensions (38 cm x 45 cm) for the larger one and of (38 x 43 cm) for the smaller one. The smaller glass plate was treated with a solution of Rain-X ®, a hydrophobic solution, whereas the larger glass plate was treated with a solution of potassium chloride (KOH) and methanol (5 g of KOH in 100 mL of methanol), a hydrophilic solution. These steps become important later when it is time to disassemble the gel to ensure that the gel sticks to the larger hydrophilic plate rather than on the smaller hydrophobic one. The plates are then assembled leaving a 2 cm gap at the bottom of the gel between the small and large plate (Supplementary Figure 10.1a). Spacers are placed between the two plates, representing the thickness of the gel (approximately 0.8 mm). A 250 mL solution of native polyacrylamide gel (29:1) 10% was prepared with a final TBMg concentration of 1X (0.09 M Tris base, 0.09 M boric acid and 5 mM $Mg(CH_3COO)_2$ pH 8.0 at ambient temperature). The volume was completed with sterilized Milli-Q water. A volume of 50 mL of this solution was kept at 4°C for later use to maintain the exact same percentage of acrylamide for the solution used to cover the void left by the well removal. To polymerize the gel, 2 mL of

10% ammonium persulfate (APS, Bio-Rad) and 70 µL of tetramethylethylenediamine (TEMED, Bioshop) was added to the remaining 200 mL preparation of the native polyacrylamide gel. Once the gel was poured, the gel was left to polymerize for at least 30 minutes. The assembly of the SR-PAGE was made as shown in Supplementary Figure 10.1b. To allow a good circulation of the buffer throughout the gel, a peristaltic pump system was installed to carry the buffer from the bottom tank to top tanks. The buffer was also allowed to flow from the top tank to the bottom one via a tube (Supplementary Figure 10.1b). Pre-migration was carried out overnight at 450 V with TBMg 1X as running buffer. From this point on, all the steps were carried in a 4°C room.

RNA samples were prepared in TBMg 1X, 3,3 µL native blue 6X loading dye (40% sucrose, 0.05 % bromophenol blue, 0.05% xylene cyanol). Different controls were used to monitor proper migration. We used radioactive RNAs of different sizes (short and long) to show that regardless of the size, the RNA comes back to the starting point if it did not interact with the ligand. Another control is a Cy3 fluorescent oligonucleotide to visualize the migration within the gel with the naked eye (Supplementary Table 10.1). Finally, depending on the SR-PAGE assay being run, a riboswitch construct known to shift in native gels was often used as a positive control (validated by our preliminary results). These controls were prepared in the same way as the RNA samples. The wells were cleaned with a syringe to remove all potential bubbles and unpolymerized acrylamide in the wells before loading the samples and controls. RNA molecules were separated according to their size and structure for 24 hours in the 10% native polyacrylamide gel at 450 V.

5.7.4 SR-PAGE reverse migration

After the first migration of 24 hours, the glass plates were carefully separated. A 50 mL solution containing the ligands of interest was prepared. The final vaporized concentration of each known riboswitch's ligand is equivalent to hundred times their respective known K_D. The sprayed concentrations for each ligand were as follows: 6 mM FH_4KO_2 (Baker *et al.*, 2012), 5 µM FMN sodium salt, hydrate (Howe *et al.*, 2016) (Cayman chemical company), 3 µM c-di-GMP sodium salt (Smith *et al.*, 2009) (Sigma-Aldrich), 10 mM glycine (Huang *et al.*, 2010) (Bioshop) , 3 µM TPP chloride (Rentmeister *et al.*, 2007) (Sigma-Aldrich), 3 mM $NiCl_2$ (Furukawa *et al.*, 2015) (Fischer), 3 mM $CoCl_2$ (Furukawa *et al.*, 2015) (Fisher) in TBMg 1X. When we worked with metal ions, 10 mM L-glutathione ($C_{10}H_{17}N_3O_6S$) (Sigma-Aldrich) was also added to the sprayed solution as a reducing agent. The volume was adjusted to 50 mL with sterilized Milli-Q water. This solution was then sprayed directly onto the gel and left in contact for ten minutes (Supplementary Figure 10.1c). It is important to spray the solution across all the surface of the gel as uniformly as

possible. The wells were cut out and the space left by the removal of these wells was carefully dried with Kimwipes® paper (Kimtech). TEMED solution (70 µL) was added to the base of the wells to help seal the junction between the old and newly added acrylamide solution. The glass plates were reassembled, this time perfectly aligning the bottom of the two plates. The previously prepared native polyacrylamide solution was degassed and used to fill the space left by the wells, taking care to add the ligands sprayed into the solution at a final concentration 10 times lower than that present in the sprayed solution, since we do not rely on passive diffusion in the gel in this case. To allow the gel to polymerize, 700 µL of 10% APS and 70 µL of TEMED were added to the 50 mL native polyacrylamide gel preparation. Note that these steps were carried at 4°C, so higher concentrations of APS and TEMED are required for the gel to polymerize. The gel was left to polymerize for at least 30 minutes. The SR-PAGE was then reassembled as before, but the polarity of the gel was reversed, with the negative electrode now at the bottom and the positive at the top. The migration was restarted at 450 V (Supplementary Figure 10.1c). Using a marker, the previous well line was delineated on the glass plates to use as a reference point to know when to stop the second migration.

Migration was stopped once the fluorescent oligonucleotide reached the wells. Due to a decreased resistance, the fluorescent oligonucleotide reached the baseline in a little less time than the first migration. This time could range anywhere between 19h to 23h, so it was important to monitor closely the SR-PAGE when we reached the end of the second migration. When it was not necessary to purify RNA from the gel, as for radioactively labeled riboswitches from Figure 5.2d and Figure 5.4b, the top of the gel was cut off and then dried using a gel dryer (model 583 Biorad, with a vacuum pump CVC3000 Vacuubrand). The gel was then exposed overnight on a phosphor screen and visualized with a Typhoon FLA9500 (GE Healthcare Life Sciences). For SELEX assays, gel strips were cut. Band 0 corresponds to the well baseline, band 1 corresponds to ~2 mm above the well up to 1 cm above the wells and band 2 corresponds to all the rest of the top of the gel; then the RNAs were eluted as described before.

5.7.5 SELEX of thiamine-binding TPP-derived riboswitches

When SR-PAGE was imbricated in an SELEX, like for the selection of a modified aptamer starting from degenerate libraries of TPP riboswitches, gel regions corresponding to “shifted-RNA” were cut out. RNAs were eluted in elution buffer overnight and ethanol precipitated like described before. Reverse transcription (RT) was performed with M-MuLV reverse transcriptase (NEB) according to manufacturer instructions with a short oligonucleotide to prevent annealing of the

oligonucleotides on RNAs that lost a few nucleotides at the 3′ end during the migration (and would thus run faster in the reverse run, co-migrating with “shifted” RNA). Half of the RT was used as a PCR template for a 100 µL reaction with conditions like library generation, but with 30 cycles for the first generation, and then progressively less for subsequent generations. RNAs were transcribed *in vitro* to start a new round of selection with an SR-PAGE like described before. During selection, 10 µM of TPP was added to the running buffer of the SR-PAGE. For the second migration, thiamine was sprayed on the gel, with a decreasing concentration as the SELEX generation increased to improve stringency. Therefore, 2 mM of thiamine was sprayed on the gel for generation 1 through 4. Then, 20 µM of thiamine was added for generation 5. Finally, 200 nM of thiamine was sprayed for the last cycles (generations 7 to 10).

5.7.6 In-line probing and dissociation constant determination

Sequences from the initial library or from generation 4, 5, 7 and 10 were either cloned with the pGEM®-T kit (Promega) to assay individual sequences and/or sequenced by Illumina (Centre d’expertises et de services Genome Québec). After analysis of Illumina sequencing results, sequences that were enriched during the cycles were selected. Individual sequences were amplified by PCR, transcribed to RNA by *in vitro* transcription, purified on 6% 8 M urea PAGE, eluted and precipitated as described before.

In-line probing reactions were performed with different concentrations of ligands ranging from 30 nM to 1 mM in a final volume of 20 µL containing 20 mM $MgCl_2$, 50 mM Tris pH 8.3 (at 25°C) and 100 mM KCl at room temperature for approximately 40 hours. For the WT TPP riboswitch, the concentrations of ligands were ranging from 30 nM to 1 mM for thiamine and from 1 nM to 30 µM for TPP. The volume of the reaction was adjusted with Milli-Q water. The resulting reaction was quenched with 2X loading buffer. Three ladders (NR: No Reaction, i.e., RNA in water; T1: RNAse T1 which cleaves at guanine bases; and OH: alkaline reaction which cleaves all positions) were made to map the RNA sequence. For the T1 ladder, 10 µL of the radioactively labeled RNA was incubated in 50 mM of sodium citrate, 3 µL of formamide 100% and 4 µL of Milli-Q water. The solution was incubated at 56°C for 2 minutes before 1.5 µL of T1 RNAse 1 U/µL was added to the T1 solution. It was left to react at 56°C for 5 minutes. For the alkaline digestion (OH), 10 µL of the radioactively labeled RNA was incubated with 8 µL of Milli-Q water and 50 mM sodium carbonate at 90°C for 90 seconds. The non-reacted ladder contained only 10 µL of the radioactively labeled RNA with Milli-Q water to a final volume of 20 µL. All reactions for the creation of the ladders were stopped with 20 µL of 2X loading buffer. Samples and ladders were

loaded on an 8% urea-PAGE for approximately 3 hours at 50 W (Regulski & Breaker, 2008). The gel was then dried and exposed on a phosphor screen overnight before being visualized on a Typhoon FLA9500 (GE Healthcare Life Sciences).

For dissociation constant determination, the intensity of the lanes and modulating bands were quantified using the ImageQuant TL software (GE Healthcare Life Sciences). These bands were normalized to give values between 0 and 1 according to minimum and maximum intensity of the bands. The K_D representation and determination were done with logarithmic curves on GraphPad 7 software (Sherlock & Breaker, 2017).

5.7.7 Free energy calculation and RNA secondary structure

Free energy of all RNA sequences were assessed using the website service of Mfold accessible at www.unafold.org (Zuker, 2003). RNAs were forced in the aptamer state using constraint based on consensus of secondary structure in Rfam (Kalvari *et al.*, 2021). All constraints applied to Mfold are available in supplementary material (Supplementary Table 10.2). RNA secondary structures were designed using the Stockholm file generated by Mfold (Zuker, 2003) as input for R2R (Weinberg & Breaker, 2011).

5.7.8 Equilibrium constant calculation

The equilibrium constants between the bound and unbound conformation of the tested RNA molecules were calculated using the following formula:

$$K_{eq} = e^{\frac{-\Delta\Delta G}{RT}}$$

The ΔΔG (in joules) is the difference in free energy (ΔG) of the bound and unbound conformation. The gas constant (R) is 8.314 J/mol·K. The temperature is 277 K, since all SR-PAGE experiments have been carried out at 4 ° C.

5.8 Acknowledgments

Authors wish to thank Gaël Montagne for technical help. Early work on SR-PAGE was started within the Breaker lab, JP wishes to thank Ronald Breaker for this, as well as members of the Breaker lab, including Rüdiger Welz who had performed gel shifts with the glycine riboswitch, providing the first positive control for SR-PAGE. Authors wish to thank Jessie Muir for the revision of the manuscript.

5.9 Authors contributions

A.D. and E.B. contributed equally to this work. J.P. conceived the SR-PAGE method. A.D., E.B., J.O., R.R., B.S. and J.P. optimized the method. J.O. realized the SELEX of thiamine binding TPP-derived riboswitches. A.D. and E.B. performed all other experiments presented in the article. Figures were created by J.P., J.O., A.D. and E.B. Manuscript was written by J.P., A.D. and E.B. Project was supervised by J.P. Manuscript was revised by all authors.

Supplementary material is available in annexe III (le matériel supplémentaire pour cet article est disponible en annexe III).

6 DISCUSSION GÉNÉRALE

Des solutions basées sur la microbiologie ont été proposées pour relever de nombreux défis mondiaux, comme la bioproduction, les changements climatiques et la santé humaine pour en nommer quelques-uns. Pour répondre à ces défis, la recherche pourrait se tourner vers des bactéries qui sont parfois moins étudiées, mais qui sont d'intéressantes pistes pour la production d'antibiotique et de protéines via l'ingénierie métabolique par exemple. Lorsque nous nous éloignons des modèles dits plus classiques de la recherche tels qu'*Escherichia coli* ou *Bacillus subtilis*, nous faisons face à des défis multiples : la communauté de chercheurs étudiant ces organismes d'intérêt est plus restreinte, limitant ainsi les ressources disponibles telles que les outils génétiques, l'annotation du génome et les souches disponibles. Bien qu'en continuelle progression, le niveau de connaissance accumulée au fil des années à propos de ces organismes est nettement inférieur. Nous sommes donc loin des organismes plus couramment étudiés qui : sont généralement plus faciles à cultiver en laboratoire avec un temps de génération souvent plus rapide; peuvent être commodément manipulés à l'aide d'outils de régulations génétiques; et proposent des protocoles minutieusement optimisés. La communauté de chercheurs travaillant avec ces organismes modèles est très grande, favorisant ainsi le partage des connaissances et des ressources.

Methylorubrum extorquens est un organisme modèle pour l'étude du métabolisme des C1 (Anthony 2011) et a un potentiel comme outil biotechnologique pour la production de composés à partir d'une matière première bon marché, le méthanol. Cette alphaprotéobactérie a déjà été modifiée pour produire de nombreux produits à valeur ajoutée (discuté dans la section 1.2.1 de cette thèse). Malgré tout, les ARNnc de cette méthylotrophe ont été peu étudiés avant les travaux présentés dans cette thèse. L'objectif de mon projet de doctorat était de développer une meilleure compréhension du rôle des ARN régulateurs chez *M. extorquens*. Les ARNnc comme les *riboswitches* et les sRNAs jouent un rôle important dans la régulation génétique, un rôle bien documenté pour de nombreuses autres bactéries.

Nous voulions d'abord obtenir un portrait des connaissances sur les sRNAs, car nous suspections que ***nous n'avons peut-être qu'effleuré la surface de l'étendue des sRNAs retrouvés dans les génomes bactériens, et ce malgré les avancées majeures dans la recherche associée aux sRNAs qui ont eu lieu au cours des dernières*** **années** (hypothèse 1). Le premier chapitre de cette thèse répond donc à cette première hypothèse à l'aide d'une approche bioinformatique.

Nous avons compilé toutes les informations sur les sRNAs annotés (selon les familles définies dans Rfam) chez les bactéries à partir de la base de données RiboGap (Naghdi *et al.*, 2017). Nous avons démontré que le nombre de sRNAs est probablement sous-estimé en raison de l'accent mis sur les organismes modèles et les pathogènes. Les bactéries encodant pour le plus grand nombre de sRNAs distincts sont celles associées à une grande intensité de recherche, comme les espèces *Salmonella*, *Escherichia* et *Staphylococcus* par exemple. Seule une infime fraction de bactéries encodent pour un grand nombre de sRNAs. Il ne serait pas surprenant que les autres espèces possèdent la même variété d'ARN régulateurs, d'autant plus s'ils encodent pour une des protéines chaperonnes Hfq et/ou ProQ, importantes dans le mode d'action des sRNAs (Figure 3.1). La diversité des sRNAs est aussi probablement sous représentée, car plusieurs sRNAs sont uniques à un groupe taxonomique limité. En effet, les vingt espèces qui encodent pour le plus grand nombre de sRNA chez les Protéobactéries appartiennent toutes à la famille des *Enterobacteriaceae*, qui inclue notamment l'organisme modèle *Escherichia coli* et le pathogène bien étudié *Salmonella enterica* (Figure 3.2). Les espèces associées à une forte intensité de recherche sont aussi celles ayant le plus grand nombre d'ARNnc par rapport à la taille de leur génome (Figure 3.4). L'annotation des ARN (*riboswitch*, ribozyme, ARN CRISPR et thermorégulateur) a du retard par rapport à l'annotation génomique, à l'exception de l'annotation des ARN ribosomaux et des ARN de transfert. En augmentant les études dédiées à l'identification d'ARNnc chez les bactéries peu étudiées, on pourrait pallier ce retard et avoir une meilleure représentation de l'étendue des différentes familles d'ARN, y compris pour les sRNAs.

Notre portrait global de l'étendue des connaissances des ARNnc chez les bactéries réalisées dans le troisième chapitre de cette thèse est précis et prend en compte la littérature récemment publiée, mais il reste toutefois imparfait. Une des limitations de cette approche bioinformatique dans le recensement des sRNAs chez les bactéries est que nous étions restreints aux annotations génomiques disponibles. Par exemple, le sRNA MicF est encodé dans un locus différent de sa cible, la protéine de la membrane externe OmpF chez *E. coli* (Delihas & Forst, 2001), répondant ainsi au critère pour être un sRNA agissant en *trans*. Cependant, le sRNA MicF est identifié comme un ARN antisens (agissant au niveau du même locus) dans la base de données Rfam (Kalvari *et al.*, 2021), car la distinction entre un ARN antisens et un sRNA n'était pas préalablement effectuée. Le sRNA MicF n'est donc pas pris en compte dans notre étude sur la prévalence des sRNAs chez les bactéries, mais se retrouve plutôt dans le travail d'analyse distinct que nous avons effectué pour les asRNA en matériel supplémentaire (section 8.1.1 de cette thèse).

De plus, notre analyse est basée sur l'outil RiboGap (Naghdi *et al.*, 2017), qui extrait les informations de la base de données Rfam (Kalvari *et al.*, 2021). Les familles d'ARN retrouvés sur Rfam s'appuient sur des modèles de covariation et les séquences d'ARN sont annotées basées sur ceux-ci (Burge *et al.*, 2013). Une des limitations de cette approche est que les modèles de covariation ne prennent pas en compte les interactions tertiaires telles que les pseudonoeuds, soit lorsque des nucléotides au sein d'une boucle interagissent avec des bases à l'extérieur de cette boucle au sein d'une même séquence. Ce type d'interaction tertiaire peut jouer un rôle dans le mode d'action de sRNAs, comme c'est le cas pour le sRNA RydC qui affecte la sensibilité thermique de la bactérie *E. coli* par exemple (Antal *et al.*, 2005). De plus, il est sans intérêt d'utiliser des modèles de covariation pour des ARN où la structure secondaire n'est pas essentielle à leur fonction. Par exemple, les ARN antisens contrôlent l'expression de leur cible en s'y hybridant avec une complémentarité parfaite, ne laissant pas de place à la formation de structure secondaire dans leur mode d'action (Nawrocki & Eddy, 2013).

Finalement, notre analyse est aussi limitée par le choix de la base de données utilisée. Avec les nouvelles technologiques de séquençage à haut-débit de l'ARN, le nombre de sRNAs prédits a augmenté considérablement, menant à la création de plusieurs bases de données sans nomenclature unifiée. Il est donc difficile de comparer différentes bases de données entre elles, en raison de la présence de redondances dans les séquences, où une même région est nommée avec des identifiants différents. Il y a aussi la présence d'ARN potentiellement mal annotés, qui sont en fait des ARNnc agissant en *cis* comme des *riboswitches*, des ARN ribosomaux ou des régions codantes. Nous avons donc choisi de travailler avec Rfam, car cela nous permettait d'étudier la présence de sRNAs dans une large gamme de bactéries. Nous avons aussi sélectionné que les sRNAs avec un E-value plus petit que 0.0005 pour ne pas tenir compte de tous les sRNAs dont la prédiction d'homologie était mauvaise. Les bases de données spécifiques à un organisme sont souvent en lien avec des organismes modèles, comme SRD qui compile toutes les données publiées à propos des ARN régulateurs chez *Staphylococcus*, un pathogène opportuniste (Sassi *et al.*, 2015). En tenir compte en plus des données retrouvées dans Rfam n'aurait que supporté, voire amplifié, notre argumentaire que les sRNAs sont plus nombreux dans les organismes modèles que dans le reste des bactéries séquencées. Bien qu'il existe plusieurs autres sRNAs lorsqu'on regarde les bases de données spécifiques à certains organismes, nous avons établi une vue d'ensemble représentative avec une tendance selon laquelle les sRNAs sont plus nombreux dans les bactéries à haute intensité de recherche.

Avec ce manuscrit, nous avons mis en évidence le potentiel de découvrir de nouveaux sRNAs, notamment si nous nous concentrons sur des bactéries qui sont moins dans la mire de la recherche. *M. extorquens* est un excellent exemple d'une bactérie pour laquelle il serait intéressant d'approfondir nos connaissances à propos de ces ARNnc. *M. extorquens* encode pour la protéine Hfq, une protéine chaperonne importante entre autres pour stabiliser l'interaction entre un sRNA et sa cible et le sRNA lui-même. ***Nous avons donc émis l'hypothèse qu'il y avait plusieurs sRNAs naturellement présents dans son génome, mais que ceux-ci n'ont simplement pas été découverts à ce jour*** (hypothèse 2). Le deuxième chapitre de cette thèse répond à cette hypothèse. Tout d'abord, l'expression des sRNAs préalablement annotés ffh, CC2171 et BjrC1505 a été confirmée chez *M. extorquens* par *Northern blot*, validant ainsi pour la première fois la présence de ces ARN régulateurs chez notre organisme modèle. Les génomes des bactéries *Bradyrhizobium japonicum (B. japonicum)* USDA 110 et *Rhodopseudomonas palustris* BisB5 ont été comparés afin d'identifier des candidats sRNAs, dont BjrC1505. Ce dernier a ensuite été validé chez *B. japonicum* par *Northern blot*, mais sa séquence est conservée chez d'autres Rhizobiales, incluant *M. extorquens* (Madhugiri et al., 2012). À notre connaissance, c'est la première fois que ce sRNA est confirmé chez une bactérie autre que celle où il a été découvert. Le sRNA ffh a été retrouvé en 5' du gène portant le même nom qui encode pour une protéine de la SRP (*signal recognition particle*) à la suite d'une analyse des régions intergéniques riches en G-C chez la bactérie marine *Candidatus* Pelagibacter ubique HTCC 1062 (Meyer *et al.*, 2009). Ce motif est répandu chez les alphaprotéobactéries, mais cette étude est la première à le valider expérimentalement, à l'exception de transcrits de ffh détectés dans des études métatranscriptomiques d'échantillons provenant de l'océan avant son identification chez *Candidatus* P. ubique (Frias-Lopez *et al.*, 2008; Shi *et al.*, 2009). Le sRNA CC2171 a été identifié chez *Caulobacter crescentus* à la suite d'une analyse par *RNA-seq* (Landt *et al.*, 2008). Il a ensuite été détecté chez *Mesorhizobium huakuii* 7653R par la même méthode (Fuli *et al.*, 2017). Cette étude est donc la première à en valider l'expression par *Northern blot* à notre connaissance.

Une étude transcriptomique a été effectuée dans le cadre d'un autre projet dans notre laboratoire (Lamarche *et al.*, 2018) et nous avons analysé les résultats avec l'outil sRNA-Detect afin d'obtenir une liste de candidats de sRNAs potentiels chez *M. extorquens* (Peña-Castillo et al., 2016). L'outil sRNA-Detect prédit les sRNAs potentiels selon la couverture des lectures et la taille de ceux-ci. Il existe plusieurs méthodes bioinformatiques pour traiter les données *RNA-seq* afin d'identifier des ARNnc. Récemment, un article a été publié afin de comparer les différents outils disponibles (Yu *et al.*, 2018). Ils ont démontré que les algorithmes basés sur la couverture des transcrits sont portés à suggérer des candidats potentiels plus petits que leur taille réelle (Yu *et al.*, 2018), ce

qui avait déjà été soulevé pour l'outil de sRNA-Detect dans l'article décrivant cet outil (Peña-Castillo *et al.*, 2016). Nous avons aussi observé ce phénomène avec notre candidat Methylo2624 : sa taille suggérée par sRNA-Detect était de 105, alors que la comparaison avec nos marqueurs de taille suggérait plutôt une taille d'environ 300 nucléotides (Figure 4.6). Similairement, Methylo1969 avait une taille prédite de 155 nucléotides par sRNA-Detect, alors que les résultats expérimentaux suggéraient qu'il est aussi d'environ 300 nucléotides, mais un peu plus grand que Methylo2624 (Figure 4.6).

Au départ, nous avons testé les candidats sRNAs dans des conditions de croissance optimales, ce qui n'est pas une stratégie idéale considérant que les sRNAs sont souvent associés à des conditions de stress (Waters & Storz, 2009). Plusieurs candidats testés par *Northern blot* n'ont pas mené à des résultats concluants, peut-être simplement parce qu'ils n'étaient pas exprimés dans les conditions testées. Nous avons donc par la suite testé les candidats sur des membranes contenant de l'ARN extrait dans une large gamme de conditions de croissance.

Les candidats étaient sélectionnés selon leur score sRNA-Detect, où un chiffre élevé signifie que l'ARN est hautement transcrit, mais pas annoté (Peña-Castillo *et al.*, 2016). Plusieurs candidats ont été testés, sans mener à des résultats concluants. Une présélection des candidats par analyse bioinformatique, comme celle faite pour Methylo2624 et Methylo1969, nous aurait permis de mieux cibler ceux à tester en laboratoire, en maximisant nos chances de succès. Par exemple, on aurait pu regarder l'homologie des séquences avec d'autres bactéries, la structure secondaire des candidats, le contexte génomique, la proximité avec des promoteurs et la présence de terminateur Rho-indépendant. Ces analyses ont toutefois été effectuées une fois qu'un résultat positif a été observé par *Northern blot*, mais l'approche inverse aurait probablement été plus efficace pour confirmer un plus grand nombre de sRNAs. Il aurait cependant fallu que ces analyses soient automatisées au travers d'un programme, car cela a représenté beaucoup de travail de les faire pour Methylo2624 et Methylo1969. De plus, les régions promotrices chez *M. extorquens* n'ont pas encore fait l'objet d'étude approfondie (Maucourt *et al.*, 2022), donc on aurait pu passer à côté de bons candidats de cette façon. De plus, les données transcriptomiques ont été obtenues dans le cadre d'une autre étude (Lamarche *et al.*, 2018), où *M. extorquens* a été cultivé dans des conditions de fermenteurs. Nous n'avons pas pu reproduire ces conditions pour les *Northern blot*, donc les candidats n'étaient pas testés dans les mêmes conditions dans lesquelles ils ont été identifiés.

À la suite de ce projet, nous avons identifié les candidats Methylo2624 et Methylo1969. Methylo2624 semble avoir une fonction de type constitutive. En effet, nous avons fait croître

M. extorquens avec différentes sources de carbone (acide succinique, méthanol et éthanol) et nous l'avons soumis à différents stress, comme un choc osmotique avec l'ajout de sel (100 mM NaCl) et l'ajout ou l'élimination de certains composés (sans cobalt, avec 1 mM d'urée et 1 mM de guanidine) ou encore différentes températures de croissance (20°C, 37°C), et aucune condition n'a affecté l'expression de Methylo2624 (Figure 4.8). Methylo2624 est donc exprimé de façon constitutive, ou du moins, son expression ne changeait pas dans les conditions testées. Nous avons même observé de plus grandes variations dans l'expression de l'ARNr 5S. Les ARN ribosomaux sont considérés comme stables, mais des conditions de stress affectant la croissance cellulaire peuvent mener à la dégradation de l'ARNr (Basturea *et al.*, 2011; Okamura *et al.*, 1973; Sulthana *et al.*, 2016; Zundel *et al.*, 2009). Une recherche d'homologie avec des sRNAs connus ne donne pas de résultats et donc ne nous aide pas à suggérer une fonction à ce sRNA. Cependant, le fait que Methylo2624 soit si stable d'une condition à l'autre suggère qu'il a une fonction essentielle, comme dans la synthèse de protéines (tel que les ARNt, ARNr et, dans une moindre mesure, l'ARNtm), la transcription (tel que l'ARN 6S), la réplication de l'ADN ou la réparation de l'ADN.

D'un autre côté, un lien entre Methylo1969 et le métabolisme de l'urée peut être établi grâce aux résultats expérimentaux et aux outils de prédiction. L'expression de Methylo1969 a été impactée par plusieurs conditions de croissance, notamment par l'ajout de l'urée où le sRNA n'était plus détecté par *Northern blot* (Figure 4.8). De plus, l'outil de prédiction CopraRNA suggère que Methylo1969 serait en mesure de cibler l'ARNm *ureG* qui encode pour une protéine accessoire d'une uréase qui catalyse la conversion de l'urée en ammoniac (Fong *et al.*, 2013). En présence d'urée, Methylo1969 n'est pas exprimé (Figure 4.6). Une hypothèse serait que Methylo1969 réprime l'action de la protéine accessoire UreG. En absence d'urée, la protéine accessoire d'une uréase n'a pas besoin d'être activée, donc Methylo1969 est transcrit afin d'inhiber la traduction de l'ARNm *ureG* en bloquant l'accès au site de reconnaissance du ribosome. Lorsque l'urée est ajoutée au milieu de culture, l'action de l'uréase devient importante, donc Methylo1969 n'est pas transcrit (Figure 6.1).

Figure 6.1 Schéma du mode d'action de Methylo1969 pour l'inactivation de la traduction de l'ARNm *ureG* (modèle proposé)

(A) En présence d'urée, l'ARNm *ureG* serait traduit. En absence d'urée, Methylo1969 inhiberait la traduction de l'ARNm *ureG* en bloquant l'accès au site de reconnaissance du ribosome (RBS). (B) L'interaction entre l'ARNm ureG et Methylo1969 prédite par CopraRNA (Wright *et al*., 2014).

Dans de futures recherches, il serait intéressant de valider cette hypothèse en effectuant par exemple des EMSA (*electrophoretic mobility shift assay*), où l'interaction entre Methylo1969 et sa cible *ureG* serait évaluée par un changement de migration sur gel. Il serait aussi intéressant de vérifier le rôle de régulation, par exemple avec un gène rapporteur qui aurait la séquence complémentaire à Methylo1969.

L'outil CopraRNA analyse les 100 cibles potentielles d'un sRNA afin de vérifier s'il y a un lien possible entre leurs fonctions. Plusieurs ARNm potentiellement ciblés par Methylo1969 sont associés avec une activité ATPase couplée à un mouvement transmembranaire de composés. Bien que cela offre une piste potentielle à la fonction de Methylo1969, l'enrichissement de cette fonction parmi la liste des 100 ARNm ciblés n'est pas significatif (le score d'enrichissement est de 1.12, mais une valeur significative pour ce score est de 1.3 ou plus grand) (Wright *et al.*, 2014).

Dans tous les cas, cette recherche est la première à démontrer l'expression des sRNAs chez *M. extorquens*. Une longue liste de candidats potentiels a été créer qui pourrait être analysé plus en détail afin d'identifier de nouveaux sRNAs.

Le dernier chapitre de cette thèse porte sur l'optimisation et la validation d'une nouvelle méthode pour la découverte de *riboswitches* : le SR-PAGE. Ce travail d'envergure a été effectué en collaboration avec ma collègue Aurélie Devinck, avec qui je suis co-première auteure pour l'article portant sur cette méthode. La technique du SR-PAGE tire avantage du changement de structure secondaire d'un ARN à la suite de la liaison d'un ligand. ***Nous avons émis l'hypothèse que le changement de structure à la suite de l'ajout de ligand dans un gel natif permettra d'identifier des riboswitches pour une variété de ligands*** (hypothèse 3). Avant d'utiliser cette

méthode dans la recherche de nouveaux *riboswitches*, il fallait d'abord l'optimiser et la valider en utilisant des *riboswitches* connus comme témoins positifs, soit ceux liant le fluor, le FMN, le c-di-GMP, la glycine, le TPP, le nickel et le cobalt. Nous avons testé plusieurs constructions de ces ARN régulateurs, contenant le domaine d'aptamère et différentes tailles de leur plateforme d'expression. Des changements de migration ont été observés pour certaines des constructions des *riboswitches* FMN, c-di-GMP, glycine et TPP lorsque les ligands correspondants étaient vaporisés avant la deuxième migration du SR-PAGE (Figure 5.2). Il est intéressant de noter que les mêmes changements de migration étaient observables lorsque nous vaporisions tous les ligands en même temps, démontrant que le SR-PAGE peut être utilisé pour augmenter le débit de criblage en permettant la recherche simultanée de *riboswitches* pour plusieurs ligands différents (Figure 5.2).

Une des limites de la méthode du SR-PAGE est que nous sélectionnons pour des ARN qui ont un changement de structure secondaire menant à un changement de migration. Lorsque nous avons testé différentes constructions des *riboswitches* connus, nous avons remarqué que ce ne sont pas toutes les constructions qui avaient un changement de migration observable à la suite du SR-PAGE, et ce même si chacune d'entre elles comprenait le domaine d'aptamère. Chacune des séquences avait théoriquement la capacité de lier leur ligand respectif. Nous avons utilisé l'outil de prédiction de structure secondaire Mfold (Zuker, 2003) pour mieux comprendre les conditions pour lesquelles nous observions des changements de migration à la suite du SR-PAGE. Nous avons comparé les différences en énergie libre de toutes les constructions des *riboswitches* choisis pour les tests : un changement de migration a été observé pour les constructions où cette différence en énergie libre était minime (matériel supplémentaire, Figure 10.2-Figure 10.5). D'un point de vue qualitatif, les molécules d'ARN où les changements de structure étaient plus importants entre la structure secondaire contenant l'aptamère et celle repliée sans contrainte ont aussi causé des plus grands changements de migration: il y avait une plus grande distance entre la ligne des puits et l'ARN en haut de cette ligne de départ. Nous avons donc démontré que le SR-PAGE pouvait être utilisé pour sélectionner des changements de structures secondaires à la suite de la liaison avec un ligand, une caractéristique clef des *riboswitches*. Nous pourrions cependant passer à côté de *riboswitches* pour lesquels le changement de structure secondaire est plus subtil.

On peut toutefois utiliser cette limitation à notre avantage comme outil pour mieux étudier les plateformes d'expression, car celles-ci sont mal caractérisées (Ceres *et al.*, 2013). Contrairement au domaine d'aptamère qui est hautement conservé, les plateformes d'expression sont très

variables, donc on ne peut pas les identifier facilement en utilisant simplement la comparaison de séquences. La plateforme d'expression chevauche le domaine d'aptamère, et elle contient la séquence qui se réorganise afin d'affecter l'expression du gène en aval (Ceres *et al.*, 2013). Il y a une grande variété au niveau des plateformes d'expression, et elles peuvent adopter diverses structures pour avoir un impact au niveau de la transcription, de la traduction ou de la dégradation par exemple (Bédard *et al.*, 2020). La taille des plateformes d'expression de *riboswitches* liant la même molécule varie aussi énormément : le *riboswitch* TPP devant le gène *thiC* chez *E. coli* a une grande plateforme d'expression pour permettre la liaison de la protéine Rho ou de la RNAse E, alors que celle du *riboswitch* TPP devant le gène *thiM* chez la même bactérie est beaucoup plus petite, car elle ne nécessite pas l'action de ces protéines (Bédard *et al.*, 2020). On pourrait utiliser le SR-PAGE pour étudier les plateformes d'expression des *riboswitches*, en testant le domaine d'aptamère suivi de différentes longueurs de séquences. Les constructions menant à un changement de structure secondaire à la suite de l'ajout du ligand pourraient ensuite être étudiées plus en détail pour déterminer le mode d'action, permettant au moins de mieux délimiter la plateforme d'expression.

Nous voulions aussi démontrer que le SR-PAGE pouvait être utilisé comme outil de sélection au sein d'un SELEX. Comme preuve de concept, une librairie du *riboswitch* liant le TPP a été créée, où différents nombres de nucléotides dégénérés étaient inclus dans la séquence de départ. L'idée était de soumettre ces librairies à un SELEX afin de sélectionner pour un *riboswitch* à affinité modifié, liant la thiamine plutôt que le TPP. L'affinité de plusieurs clones a été mesurée pour la thiamine et le TPP à l'aide de la technique d'*in-line probing*, et ce pour des séquences sélectionnées à différents cycles du SELEX. Des 47 clones testés, 18 ont démontré une meilleure affinité pour la thiamine que pour le TPP, dont trois parmi ceux-ci ont révélé avoir une meilleure affinité pour la thiamine que le *riboswitch* sauvage (Figure 5.4d). Il y avait aussi deux clones pour lesquels l'affinité de la séquence pour le TPP n'a pas pu être mesurée, car aucune modulation a été observée à la suite de la réaction *in-line* dans les conditions testées. Nous avons donc démontré que le SR-PAGE pouvait être utilisé dans le cadre d'un SELEX comme outil de sélection pour un *riboswitch* à affinité modifié et/ou amélioré.

Chacune des réactions d'*in-line probing* a été répétée afin de valider les résultats. L'ARN utilisé lors du *in-line probing* était marqué à la radioactivité. Bien que nous connaissions la quantité d'ARN utilisé lors de la réaction de marquage, nous ne pouvions pas mesurer quelle quantité d'ARN (en picomoles) était obtenue à la suite des différentes étapes du marquage et de purification. Durant la réaction d'*in-line*, le facteur limitant n'est pas la concentration du ligand

testé, mais plutôt la concentration d'ARN marqué à la radioactivité : les métabolites doivent être en excès. Lorsqu'on teste des petites concentrations de métabolites, il faut donc travailler avec des concentrations encore plus petites de l'ARN radiomarqué. L'efficacité du marquage radioactif est donc limitante, car l'ARN radiomarqué doit être présent en assez faible concentration pour que les métabolites soient en excès, tout en gardant un signal radioactif assez fort afin qu'il soit détecté (Regulski & Breaker, 2008). Avec la durée de vie limitée de la radioactivité, il a aussi fallu préparer les échantillons radioactifs à plusieurs reprises pour effectuer tous les tests nécessaires, car il fallait s'assurer de garder un signal radioactif suffisant à la détection. Nous aurions pu travailler avec des techniques de marquage basé sur la fluorescence, car il a été démontré que le marqueur radioactif en 5' pouvait être changé pour un fluorescent, sans affecter les résultats : la comparaison des réactions à la suite des deux méthodes de marquage a démontré des résultats pratiquement identiques lorsque le test a été effectué avec le *riboswitch* lysine (Strauss *et al.*, 2012). Le *riboswitch* lysine a une affinité de liaison (K_D) d'environ 1 µM (Strauss *et al.*, 2012). Cette technique pourrait donc fonctionner pour des *riboswitches* avec des K_D dans les mêmes ordres de grandeur, comme celui pour la glycine (± 10 µM) (Ruff & Strobel, 2014), mais il n'a pas été démontré si la fluorescence était assez sensible pour détecter des interactions dans la gamme des bas nanomolaires, comme c'est le cas pour le *riboswitch* TPP (± 100 nM) (Haller *et al.*, 2013).

Nous considérons que la technique du SR-PAGE peut maintenant être appliquée dans d'autres laboratoires en utilisant du matériel que des groupes de recherches en biochimie ont déjà sous la main. Cependant, l'optimisation de cette méthode a été un grand défi, et nous avons dû penser à plusieurs détails dans la méthode pour éviter des soucis techniques. Par exemple, cela prend une certaine dextérité pour démouler le gel avant de vaporiser les ligands, c'est pourquoi nous traitons les vitres avec un produit hydrofuge d'un côté et un produit hydrophile de l'autre pour favoriser le démoulage. Nous dégazions aussi la solution d'acrylamide en extra que l'on ajoutait au gel démoulé afin de limiter la formation de bulles. Nous avons aussi vérifié la concentration de ligand absorbé par le gel en quantifiant l'absorbance à 393 nM de celui-ci après avoir vaporisé du sulfate de nickel (Mathpal & Kandpal, 2009). La comparaison de l'absorbance de la solution vaporisée par rapport à celle absorbée par le gel nous a permis d'établir qu'il fallait asperger pour chaque ligand une concentration équivalente à cent fois leur K_D respectif. Certains défis techniques existent encore. Une chose qui était en partie hors de notre contrôle était de s'assurer que le SR-PAGE soit ininterrompu toute la nuit. En effet, le SR-PAGE prend une semaine à compléter, incluant plusieurs étapes où l'ARN migre toute la nuit jusqu'au lendemain. Il fallait donc s'assurer que le système d'électrophorèse ne fuyait pas et un système de pompe permettait

le bon roulement du tampon. Dans tous les cas, nous n'avons pas été à l'abri de fuites et de coupures électriques, qui nous ont parfois retardés dans nos expériences. De plus, pour illustrer certains défis lors de la mise au point, le SR-PAGE n'a pas fonctionné pendant plusieurs semaines, et aucun changement de migration d'ARN ne pouvait être observé pour les témoins positifs. Nous avons finalement réalisé que nous avions changé le fournisseur de l'APS. On utilise ce produit pour polymériser l'acrylamide avec du TEMED afin de remplacer le trou laissé par le retrait des puits avant la deuxième migration. En retournant vers le fournisseur initial, l'APS de Biorad, nos expériences fonctionnaient à nouveau. Bien que nous nous expliquions mal cette différence, nous n'avons utilisé que l'APS de Biorad par la suite. Nous pensons tout de même que d'autres laboratoires seront en mesure de reproduire le SR-PAGE dans leur propre installation.

6.1 Perspective

6.1.1 Recherche de nouveaux *riboswitches* par SR-PAGE

Maintenant que nous avions validé la technique du SR-PAGE, l'idée était d'utiliser cette nouvelle méthode afin de découvrir de nouveaux *riboswitches* chez notre organisme modèle *M. extorquens* et chez d'autres bactéries d'intérêts. En effet, le SR-PAGE nous permet de démarrer notre sélection à partir d'une librairie de séquences, donc nous ne nous sommes pas limités à une seule bactérie. Quelques *riboswitches* sont annotés dans le génome de *M. extorquens* AM1, incluant ceux liant le fluor, la guanidine, la cobalamine, le TPP et la glycine. Un ligand pour notre organisme modèle était par exemple le méthanol, car cette méthylotrophe facultative est en mesure de détecter la présence de cette source de carbone pour initier le cycle de la sérine important dans le métabolisme des C1 (Peyraud *et al.*, 2012). Les *riboswitches* ZMP/ZTP sont aussi en lien avec le métabolisme des C1 (Kim *et al.*, 2015) et ils ne sont pas annotés dans notre organisme modèle, mais ils sont retrouvés chez d'autres *Methylobacteriaceae* (*Methylobacterium radiotolerans* et *Methylobacterium tarhaniae*). Le THF est aussi un cofacteur impliqué dans le métabolisme des carbones C1 (Schirch, 1998) et au moins deux familles de *riboswitches* interagissent avec ce ligand, mais aucun n'est retrouvé chez *M. extorquens* pour l'instant.

Nous avons créé des librairies génomiques contenant des séquences dérivées du génome de bactéries flanquées de régions fixes selon la méthode décrite par Singer *et al*, 1997. Nous avons utilisé les bactéries que nous avions à portée de main à notre laboratoire et de nos collègues (*Bacillus thuringiensis*, *Vibrio harveyi*, *Pseudomonas aeruginosa*, *Streptococcus agalactiae*,

Escherichia coli , *Hyphomicrobium nitrativorans*, *Streptomyces avermitilis*, *Methylophaga nitratireducenticrescens* , *Helicobacter pylori*, *Mycobacterium tuberculosis*, *Yersinia enterocolitica*, *Neisseria elongata*, Burkholderia thailandensis, *Halomonas elongata*, Staphylococcus aureus, Myxococcus xanthus, *Methylorubrum extorquens* et *Microcystis aeruginosa*). Après quelques générations de SELEX, nous avons remarqué à la suite d'un séquençage à haut-débit qu'il y avait un grand biais pour la séquence connue des amorces aléatoires. En effet, nous avons observé que les régions du génome qui avaient plusieurs bases complémentaires avec la séquence connue des amorces (plutôt que les régions aléatoires de neuf bases) créaient un biais majeur pour la réaction d'amplification. Notre librairie d'ARN de départ comportait donc un biais important qui a affecté les cycles de sélection.

Nous avons donc choisi de recommencer le SELEX en commandant cette fois notre librairie de séquences chez la compagnie Genscript. Un travail bioinformatique nous a permis de créer une liste des régions intergéniques de toutes les bactéries qui contenaient un terminateur de transcription prédit, incluant notre organisme d'intérêt *M. extorquens*. La compagnie nous permettait de commander des séquences de maximum 170 nucléotides. En considérant les amorces, les séquences intergéniques testées pouvaient donc être au maximum de 130 nucléotides (Figure 6.2).

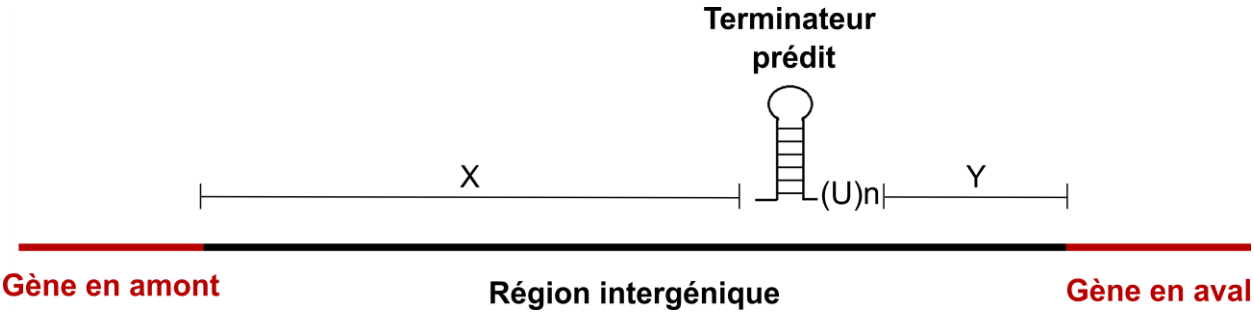

Figure 6.2 Design des librairies des régions intergéniques

Nous avons sélectionné les régions intergéniques qui contenaient un terminateur prédit plus proche du gène en avant que celui en amont (où X nucléotides est plus grand que Y).

Nous avons sélectionné des séquences de 130 nucléotides en amont du milieu des terminateurs prédits. Les *riboswitches* ont cependant une taille entre 50 et 250 nucléotides, donc les ARN régulateurs de plus grandes tailles seraient ignorés. Le plus petit *riboswitch* connu est celui liant le preQ1, avec un domaine aptamère d'approximativement 40 nucléotides (Rieder *et al.*, 2010)

Nous avons effectué trois tours de sélection en vaporisant différents métaux (nickel, cobalt, cuivre, zinc, manganèse, calcium, cadmium, plomb, lithium, argent et molybdène) ainsi que divers métabolites (méthanol, ppGpp, pRpp, c-di-GMP, cAMP, TPP et glycine). Les séquences

sélectionnées ainsi que les librairies de départ ont été analysées par séquençage de nouvelle génération Illumina. Malheureusement, nous avons observé plusieurs concatémères d'amorces dans la librairie initiale qui n'avait pas encore été sélectionnée, ce qui a affecté le processus complet de sélection : ces courts fragments étaient surreprésentés à la suite des trois tours de sélection. En parcourant la littérature, nous avons réalisé que c'était un problème commun lorsque l'on commande une librairie d'ADN d'un manufacturier dans le but de la transcrire *in vitro* en ARN (Takahashi *et al.*, 2016). Lors de l'amplification initiale de la librairie, plusieurs biais sont introduits, incluant des séquences contenant des chimères d'amorces (Figure 6.3).

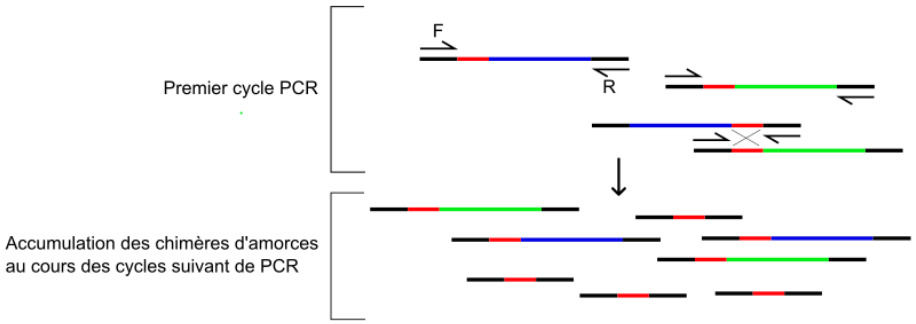

Figure 6.3 Chimères d'amorces introduites par PCR, inspiré de (Williams *et al.*, 2006)

Les différents fragments d'ADN à amplifier sont représentés en vert et en bleu. Les régions en noir sont les amorces afin de permettre l'amplification par PCR. Des événements de recombinaison peuvent se produire entre les régions homologues représentées en rouge entre les séquences (des sites de restriction), ce qui mène à la formation de produits chimériques. Les petits fragments sont amplifiés plus efficacement, menant à une accumulation de cours produits chimériques au fils des cycles PCR.

Les chimères d'amorces sont produites à la suite d'une recombinaison entre les régions homologues de deux séquences, comme des sites reconnus par des enzymes de restrictions qui sont présents sur toutes les séquences. Nous avions ajouté ces sites de restrictions au cas où nous allions faire du clonage avec ces séquences. Les petits fragments créés sont amplifiés plus efficacement, c'est pourquoi on observe leur accumulation au fils des cycles PCR. C'est exactement ce que nous avons observé dans nos résultats séquençage, où plusieurs fragments correspondaient à nos deux amorces pratiquement l'une en face de l'autre. Si cette expérience était à refaire, il aurait été préférable d'analyser notre librairie initiale par séquençage de nouvelle génération dès le départ. Nous aurions donc remarqué ce biais, au lieu de soumettre ces séquences à plusieurs cycles de sélection, ce qui représente plusieurs mois de travail. Une façon de régler ce problème aurait été de faire des PCR par émulsion, où les fragments d'ADN sont divisés dans des gouttelettes contenant une seule ou très peu de molécules d'ADN par réaction

PCR, limitant ainsi les chances d'avoir des recombinaisons entre les séquences (Williams *et al.*, 2006). En contrepartie, cela complique et prolonge l'étape d'amplification, en plus de réduire la taille des librairies amplifiables (le nombre de séquence).

La technique du SR-PAGE a été validée au cours de cette thèse, car il a été démontré que cette méthode peut sélectionner des séquences d'ARN qui ont un changement de structure secondaire menant à un changement de migration dans un gel de polyacrylamide natif à la suite de la liaison d'un ligand. Malheureusement, nous n'avons pas pu appliquer cette nouvelle méthode afin de découvrir de nouveaux *riboswitches* chez notre organisme modèle, en raison des problèmes dans la réalisation des librairies d'ARN de départ. Le SR-PAGE combiné au SELEX est la seule technique qui permettrait le criblage de librairies contenant un grand nombre de séquences naturelles. Une technique comparable est le capture-SELEX qui permet de sélectionner des aptamères d'ARN ou d'ADN pour différents ligands (Lyu *et al.*, 2021). Le capture-SELEX nécessite cependant que les séquences d'ARN ou d'ADN soient immobilisées sur une matrice, alors qu'aucune immobilisation n'est nécessaire dans le cas du SR-PAGE, que ce soit pour les ligands ou les séquences d'ARN. Le ligand est souvent entièrement entouré par le domaine d'aptamère d'un *riboswitch*, donc l'immobilisation des séquences d'intérêt n'est pas idéale pour la sélection pour ce type d'ARNnc régulateur (Matylla-Kulinska *et al.*, 2012). Dans le cas du SR-PAGE, la séquence d'ARN interagit librement avec le ligand en solution, sans aucune immobilisation. De plus, le capture-SELEX débute avec une librairie synthétique, alors que le SR-PAGE utilise des séquences retrouvées naturellement. La méthode du capture-SELEX n'est pas idéale lorsqu'on travaille avec des séquences retrouvées naturellement, parce qu'il faudrait sélectionner quelle région immobilisée sur la matrice. On pourrait donc immobiliser une section importante dans la liaison du domaine d'aptamère au ligand d'intérêt, et donc manquer un potentiel *riboswitch*.

La validation de la technique du SR-PAGE permet tout de même de présenter une nouvelle approche afin de découvrir des *riboswitches*, qui présente des avantages par rapport à l'approche normalement utilisée de la bioinformatique. La technique du SR-PAGE pourra aussi être appliquée pour découvrir des *riboswitches* pour une variété de ligands comme des acides aminés ou des coenzymes. La recherche de *riboswitches* peut aussi s'étendre sur un grand nombre d'organismes, étant donné qu'il n'est pas plus fastidieux d'étudier plusieurs organismes en même temps en utilisant une librairie métagénomique. Chaque nouvelle découverte offrira de nouveaux

points de vue sur la capacité biochimique de l'ARN et la régulation des gènes chez les bactéries, en plus d'ouvrir des opportunités pour des avancées technologiques.

6.1.2 Développement d'outils de régulation génétique basés sur les ribozymes synthétiques chez *Methylorubrum extorquens*

L'étude des ARNnc est un domaine de recherche fleurissant : ils sont une piste intéressante pour le développement de biotechnologie, car ils peuvent être programmés ou ciblés pour rediriger les voies métaboliques pour surproduire une molécule d'intérêt. Les récentes avancées en biologie synthétique se sont tournées vers l'ingénierie de différents types d'ARNnc afin d'agrandir la liste d'outils de régulation génétique disponibles chez des organismes au potentiel biotechnologique. La disponibilité d'outils de régulation génétique est essentielle pour poursuivre le développement de *M. extorquens* d'un point de vue biotechnologique. Récemment, l'interférence par le système CRISPR a été optimisée pour la bactérie *M. extorquens* (Mo et al., 2020). La protéine dCas9 de *Streptococcus pyogenes* (protéine Cas9 mutée où l'activité endonucléase est inactivée) est recrutée par un ARN guide sur le gène d'intérêt, ce qui inhibe l'expression du gène. Même si cette méthode s'est avérée efficace, cette technique présente également un risque élevé d'induire des effets hors cibles, puisqu'il a été démontré que l'expression des gènes peut être inhibée à des sites ayant à peine neuf nucléotides d'homologie avec l'ARN guide (Cui et al. 2018). Des petits ARN (sRNAs) synthétiques ont également récemment été implémentés chez *M. extorquens*. Ils contiennent une région du sRNA MicC retrouvé chez *E. coli* afin de promouvoir le recrutement de la protéine Hfq, une protéine chaperonne nécessaire pour stabiliser l'interaction entre le sRNA et sa cible, mais aussi le sRNA lui-même (Zhu *et al.*, 2021). Les sRNAs synthétiques se lient à leurs cibles en raison de la complémentarité de leur séquence. Cet appariement affecte l'expression de l'ARNm ciblé. Les sRNAs synthétiques ont été démontrés comme fonctionnels chez *M. extorquens*, mais ils nécessitaient la surexpression de la protéine Hfq d'*E. coli*, malgré le fait que *M. extorquens* encode aussi pour cette protéine chaperonne (Zhu *et al.*, 2021). Le gène *crtI* a été ciblé dans la preuve de concept de ces deux outils de régulation génétique, car l'efficacité des outils peut rapidement être évaluée par un changement dans la couleur des bactéries, en raison de l'effet sur la production de caroténoïdes (Mo *et al.*, 2020; Zhu *et al.*, 2021).

Ici, nous proposons le développement d'un nouvel outil de régulation génétique chez *M. extorquens* basé sur l'ARN et qui ne nécessite aucun cofacteur protéique, les ribozymes *hammerheads.* Il est possible de modifier les ribozymes *hammerheads*, soit des molécules ayant une activité d'autoclivage, afin qu'ils ciblent un gène d'intérêt à l'aide de l'outil Ribosoft 2.0 (Kharma *et al.*, 2016) (voir section 1.3.3.2). Les ribozymes synthétiques ont déjà été démontrés

comme étant efficaces chez d'autres bactéries comme *E. coli* (Najeh, 2017) ou *Lactococcus lactis* par exemple (Fiola et al., 2006), mais leurs rendements n'ont toujours pas été testés chez *M. extorquens*. Mon objectif était donc de tester cet outil de régulation génétique en ciblant le gène *crtI*, parce qu'une mutation dans ce gène crée une souche de *M. extorquens* sans la production de caroténoïdes qui lui confèrent normalement une pigmentation rose. Une délétion de ce gène n'affecte pas la croissance de la bactérie et ce changement de couleur est mesurable (Van Dien *et al.*, 2003a). Nous avions donc émis l'hypothèse que les ribozymes synthétiques pourraient être utilisés comme outil de régulation génétique chez *M. extorquens*. L'implémentation de nouveaux outils de régulation génétique basés sur l'ARN chez cet organisme modèle comme les ribozymes synthétiques pourrait faciliter l'optimisation de procédés industriels. La description des protocoles se trouve à l'annexe IV de cette thèse.

Ribosoft 2.0, un service web accessible au public (https://ribosoft2.fungalgenomics.ca/), a été utilisé pour concevoir des ribozymes *hammerheads* ciblant l'ARN d'intérêt. Lorsqu'on soumet la séquence de l'ARNm ciblé, *crtI*, cette plateforme web crée une liste de ribozymes *hammerheads* avec un bon potentiel prédit de coupure de la cible. Ribosoft 2.0 prend en compte de multiples critères tels que la structure du ribozyme, l'interaction ribozyme-cible et l'accessibilité de la cible, pour n'en citer que quelques-uns (Kharma et al. 2016). Comme preuve de concept, huit ribozymes ciblant l'expression du gène *crtI* ont été sélectionnés de la liste fournie par Ribosoft 2.0 (Tableau 6.1).

Tableau 6.1 Séquences des ribozymes *hammerheads* ciblant le gène *crtI*

Ribozymes	Séquences
1	caggccggcuaaucgaag**cugauga**gucgcugaaaugcgac**gaaa**cuucuucucguag
2	ucuucuaaucguag**cugauga**gucgcugaaaugcgac**gaaa**cuugccgaccu
3	gaccucgaauccgcc**cugauga**gucgcugaaaugcgac**gaaa**cgagc
4	ccgaccucgaauccgcc**cugauga**gucgcugaaaugcgac**gaaa**cgagcucgcg
5	aggccggcuaaucgaag**cugauga**gucgcugaaaugcgac**gaaa**cuuc
6	acuucuucuaaucguag**cugauga**gucgcugaaaugcgac**gaaa**cuug
7	ucguaggacaauuugcc**cugauga**gucgcugaaaugcgac**gaaa**ccucgc
8	caggccggcuaaucgaag**cugauga**gucgcugaaaugcgac**gaaa**cuucuucucguagg

Le cœur catalytique des ribozymes hammerhead est représenté en gras alors que les parties qui s'hybrident à la cible sont soulignées.

L'activité de clivage des ribozymes *hammerheads* contre l'ARNm *crtI* marqué à la radioactivité a d'abord été démontrée *in vitro* (Figure 6.4).

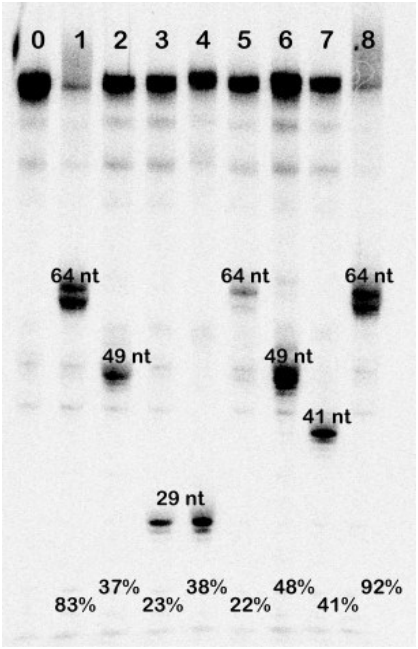

Figure 6.4 Efficacités de clivage (%) *in vitro* des ribozymes *hammerheads* ciblant l'ARNm du gène *crtI*.

L'intensité des bandes est quantifiée avec ImageJ. La colonne 0 représente l'ARNm *crtI* seul, alors que celles de 1 à 8 correspondent à l'ARNm *crtI* incubé avec le ribozyme *hammerhead* correspondant.

Des plasmides exprimant les ribozymes 1, 4, 7 et 8 ont été construits en utilisant le plasmide pCM110-GFP comme matrice (Marx & Lidstrom, 2001). Nous voulions tester *in vivo* des ribozymes qui ciblaient différents locus sur la séquence du gène *crtI*. Les ribozymes sont sous le contrôle du promoteur $P_{mxaF,}$ un promoteur fort inductible par le méthanol chez *M. extorquens* (Fitzgerald & Lidstrom, 2003) et d'un double terminateur T1/TE. Étant donné que les cultures sont cultivées en présence de méthanol, le promoteur agit donc de façon constitutive. En guise de contrôle, un plasmide exprimant un ribozyme ne ciblant pas le génome de *M. extorquens* a été construit de la même façon. Ce ribozyme cible plutôt la protéine fluorescente rouge (RFP), et son efficacité *in vivo* chez *E. coli* a été validée dans le cadre d'une autre étude (Najeh, 2017). La construction des différents plasmides a été validée par séquençage Sanger (Centre d'expertise et de services Génome Québec) (les séquences sont disponibles dans le Tableau 11.4). Tous les plasmides ont été transformés dans *M. extorquens*.

Afin d'évaluer l'efficacité des ribozymes *in vivo*, la production de caroténoïdes a été mesurée en comparant les souches de *M. extorquens* exprimant les différents plasmides ciblant le gène *crtI* (ribozyme 1, 4, 7 et 8) avec la souche sauvage ainsi que celle exprimant un ribozyme contrôle.

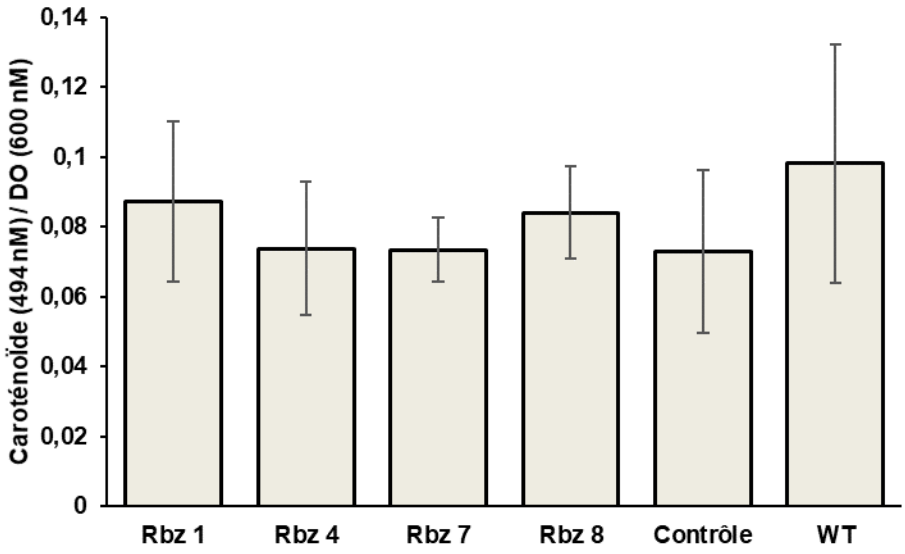

Figure 6.5 Effets des ribozymes sur la production de caroténoïdes chez *M. extorquens*

Les souches Rbz 1, Rbz 4, Rbz 7 et Rbz 8 contiennent les plasmides exprimant les ribozymes ciblant le gène *crtI*. La souche contrôle contient le plasmide qui exprime un ribozyme ciblant plutôt la protéine RFP, et ne devrait donc pas avoir d'impact sur la production de caroténoïdes. La souche sauvage est représentée par l'appellation « WT ». Chacune des conditions est reproduite en plusieurs réplicas.

En principe, si les ribozymes *hammerhead* avaient fonctionné *in vivo*, nous aurions observé une diminution de la production de caroténoïdes chez les souches contenant les ribozymes ciblant le gène *crtI*. Cependant, nous n'avons pas observé de différence significative entre nos conditions (Figure 6.5). L'utilisation de ribozyme synthétique chez *M. extorquens* devra donc être optimisée.

Quelques explications peuvent aider à interpréter ces résultats. Les cultures de *M. extorquens* varient souvent d'une expérience à l'autre et il est difficile de prévoir la croissance bactérienne. En effet, le milieu de culture de cet organisme contient plusieurs métaux, et des préparations différentes du même milieu produisent des taux de croissance très différents (Chou *et al.*, 2009). De plus, *M. extorquens* est en mesure de former des biofilms et s'agglutine pendant sa croissance, ce qui complique la lecture de la densité optique (Delaney *et al.*, 2013). Certains laboratoires ont choisi de travailler avec une souche dont les gènes pour la synthèse de la cellulose (Δ*celABC*) sont supprimés pour pallier cet inconvénient. Étant donné que la croissance bactérienne a un impact sur le nombre de bactéries et donc de la quantité de caroténoïdes extraits, il est important d'être en mesure de mesurer la densité optique de façon reproductible. Pour éviter ce problème sans avoir à travailler avec la souche mutante (Δ*celABC*) qui ne forme

pas de biofilm, on pourrait évaluer l'impact des ribozymes par RT-qPCR. Nous aurions aussi pu envisager de tester de nouveaux candidats. L'outil Ribosoft 2.0 fournit une liste de ribozymes selon le gène ciblé et génère un score selon différents critères : l'accessibilité de la cible, la température d'hybridation et la formation de structure secondaire (Kharma *et al.*, 2016). Idéalement, nous aurions pu nous fier aux scores générés par Ribosoft 2.0 pour sélectionner les ribozymes à tester *in vitro* et *in vivo*. Par contre, il n'y a pas de corrélation claire entre les scores donnés pour ces différents critères et l'efficacité des ribozymes (Najeh, 2017). Ribosoft 2.0 avait créé une longue liste de plus de 400 candidats, alors plusieurs pourraient être testés afin d'en trouver un ayant une meilleure efficacité. Finalement, des constructions de plusieurs ribozymes ciblant différentes régions du même gène pourraient être construites afin de maximiser leur impact, une telle approche a déjà permis d'augmenter l'efficacité de coupure au-delà de 90% (Kharma et al., 2016).

6.2 Conclusion

L'objectif général de mon projet de doctorat est d'avoir une meilleure compréhension du rôle des ARN régulateurs chez *M. extorquens*, plus spécifiquement les sRNAs et les *riboswitches.* Au terme de cette thèse, j'espère avoir réussi à démontrer l'importance d'étudier les ARNnc bactériens chez d'autres organismes : une première étude des sRNAs chez *M. extorquens* a offert des résultats prometteurs et met la table à plusieurs autres découvertes. L'annotation des ARN traîne derrière celle des ARNm encodant pour des protéines, mais leur fonction est tout aussi importante. Les technologies pour étudier les ARN se sont nettement améliorées, notamment avec le séquençage transcriptomique. Une meilleure compréhension des ARN noncodants bactériens nous permettrait de mieux cibler l'ingénierie génétique pour la production d'un métabolite d'intérêt, tout en faisant avancer nos connaissances. Le développement de la nouvelle technique du SR-PAGE est aussi un pas important afin d'approfondir nos connaissances sur les *riboswitches*, car c'est un outil qui favorise l'identification de nouveaux ARN régulateurs à grande échelle, soit à partir d'une librairie de séquences issues de plusieurs organismes en testant de nombreux métabolites d'intérêts en même temps.

7 BIBLIOGRAPHIE

Abràmoff MD, Magalhães PJ & Ram SJ (2004) Image processing with ImageJ. *Biophotonics international* 11(7):36-42.

Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A & Carniel E (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences* 96(24):14043-14048.

Adams PP & Storz G (2020) Prevalence of small base-pairing RNAs derived from diverse genomic loci. *Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863(7):194524.

Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N & Grüning BA (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic acids research* 46(W1):W537-W544.

Ahmad H, Masroor T, Parmar SA & Panigrahi D (2021) Urinary tract infection by a rare pathogen Cedecea neteri in a pregnant female with Polyhydramnios: rare case report from UAE. *BMC Infectious Diseases* 21(1):1-6.

Aiba H (2007) Mechanism of RNA silencing by Hfq-binding small RNAs. *Current opinion in microbiology* 10(2):134-139.

Albrecht M, Sharma CM, Reinhardt R, Vogel J & Rudel T (2010) Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic acids research* 38(3):868-877.

Al Mamun AAM, Lombardo M-J, Shee C, Lisewski AM, Gonzalez C, Lin D, Nehring RB, Saint-Ruf C, Gibson JL & Frisch RL (2012) Identity and function of a large gene network underlying mutagenic repair of DNA breaks. *Science* 338(6112):1344-1348.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.

Ames TD & Breaker RR (2011) Bacterial aptamers that selectively bind glutamine. *RNA biology* 8(1):82-89.

Andersen J, Forst S, Zhao K, Inouye M & Delihas N (1989) The function of micF RNA: micF RNA is a major factor in the thermal regulation of OmpF protein in *Escherichia coli*. *Journal of Biological Chemistry* 264(30):17961-17970.

Antal M, Bordeau V, Douchin V & Felden B (2005) A small bacterial RNA regulates a putative ABC transporter. *Journal of Biological Chemistry* 280(9):7901-7908.

Archer GL (1998) *Staphylococcus aureus:* a well-armed pathogen. *Reviews of Infectious Diseases* 26(5):1179-1181.

Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EGH, Margalit H & Altuvia S (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Current Biology* 11(12):941-950.

Arnvig KB & Young DB (2009) Identification of small RNAs in *Mycobacterium tuberculosis*. *Molecular microbiology* 73(3):397-408.

Atilho RM, Arachchilage GM, Greenlee EB, Knecht KM & Breaker RR (2019a) A bacterial riboswitch class for the thiamin precursor HMP-PP employs a terminator-embedded aptamer. *Elife* 8:e45210.

Atilho RM, Perkins KR & Breaker RR (2019b) Rare variants of the FMN riboswitch class in *Clostridium difficile* and other bacteria exhibit altered ligand specificity. *RNA* 25(1):23-34.

Bak G, Lee J, Suk S, Kim D, Young Lee J, Kim K-s, Choi B-S & Lee Y (2015) Identification of novel sRNAs involved in biofilm formation, motility and fimbriae formation in *Escherichia coli*. *Scientific reports* 5(1):1-19.

Baker CS, Morozov I, Suzuki K, Romeo T & Babitzke P (2002) CsrA regulates glycogen biosynthesis by preventing translation of glgC in *Escherichia coli*. *Molecular microbiology* 44(6):1599-1610.

Baker JL, Sudarsan N, Weinberg Z, Roth A, Stockbridge RB & Breaker RR (2012) Widespread genetic switches and toxicity resistance proteins for fluoride. *Science* 335(6065):233-235.

Barrick JE & Breaker RR (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology* 8(11):1-19.

Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N & Jona I (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proceedings of the National Academy of Sciences* 101(17):6421-6426.

Basturea GN, Zundel MA & Deutscher MP (2011) Degradation of ribosomal RNA during starvation: comparison to quality control during steady-state growth and a role for RNase PH. *Rna* 17(2):338-345.

Battistuzzi FU & Hedges SB (2009) A major clade of prokaryotes with ancient adaptations to life on land. *Molecular biology and evolution* 26(2):335-343.

Bayat P, Nosrati R, Alibolandi M, Rafatpanah H, Abnous K, Khedri M & Ramezani M (2018) SELEX methods on the road to protein targeting with nucleic acid aptamers. *Biochimie* 154:132-155.

Bédard A-SV, Hien ED & Lafontaine DA (2020) Riboswitch regulation mechanisms: RNA, metabolites and regulatory proteins. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863(3):194501.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL & Bignell HR (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *nature* 456(7218):53-59.

Bi D, Jiang X, Sheng Z-K, Ngmenterebo D, Tai C, Wang M, Deng Z, Rajakumar K & Ou H-Y (2015) Mapping the resistance-associated mobilome of a carbapenem-resistant *Klebsiella pneumoniae* strain reveals insights into factors shaping these regions and facilitates generation of a 'resistance-disarmed'model organism. *Journal of Antimicrobial Chemotherapy* 70(10):2770-2774.

Blount KF & Breaker RR (2006) Riboswitches as antibacterial drug targets. *Nature biotechnology* 24(12):1558-1564.

Blount ZD (2015) The natural history of model organisms: The unexhausted potential of *E. coli*. *Elife* 4:e05826.

Bohn C, Rigoulay C & Bouloc P (2007) No detectable effect of RNA-binding protein Hfq absence in *Staphylococcus aureus*. *BMC microbiology* 7(1):1-9.

Bouché F & Bouché JP (1989) Genetic evidence that DicF, a second division inhibitor encoded by the *Escherichia coli* dicB operon, is probably RNA. *Molecular microbiology* 3(7):991-994.

Boudry P, Gracia C, Monot M, Caillet J, Saujet L, Hajnsdorf E, Dupuy B, Martin-Verstraete I & Soutourina O (2014) Pleiotropic role of the RNA chaperone protein Hfq in the human pathogen *Clostridium difficile*. *Journal of bacteriology* 196(18):3234-3248.

Boudry P, Piattelli E, Drouineau E, Peltier J, Boutserin A, Lejars M, Hajnsdorf E, Monot M, Dupuy B & Martin-Verstraete I (2021) Identification of RNAs bound by Hfq reveals widespread RNA partners and a sporulation regulator in the human pathogen *Clostridioides difficile*. *RNA biology* 18(11):1931-1952.

Bourque D, Pomerleau Y & Groleau D (1995) High-cell-density production of poly-β-hydroxybutyrate (PHB) from methanol by *Methylobacterium extorquens*: production of high-molecular-mass PHB. *Applied microbiology and biotechnology* 44(3):367-376.

Boussebayle A, Groher F & Suess B (2019) RNA-based capture-SELEX for the selection of small molecule-binding aptamers. *Methods* 161:10-15.

Boutet E, Djerroud S & Perreault J (2022) Small RNAs beyond Model Organisms: Have We Only Scratched the Surface? *International journal of molecular sciences* 23(8):4448.

Breaker R (2012) New insight on the response of bacteria to fluoride. *Caries research* 46(1):78-81.

Breaker RR (2011) Prospects for riboswitch discovery and analysis. *Molecular cell* 43(6):867-879.

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP & Bateman A (2013) Rfam 11.0: 10 years of RNA families. *Nucleic acids research* 41(D1):D226-D232.

Butcher SE & Burke JM (1994) Structure-mapping of the hairpin ribozyme: magnesium-dependent folding and evidence for tertiary interactions within the ribozyme-substrate complex. *Journal of molecular biology* 244(1):52-63.

Buzayan JM, Gerlach WL & Bruening G (1986) Satellite tobacco ringspot virus RNA: A subset of the RNA sequence is sufficient for autolytic processing. *Proceedings of the National Academy of Sciences* 83(23):8859-8862.

Calin-Jageman I & Nicholson AW (2003) RNA structure-dependent uncoupling of substrate recognition and cleavage by *Escherichia coli* ribonuclease III. *Nucleic acids research* 31(9):2381-2392.

Carrillo M, Wagner M, Petit F, Dransfeld A, Becker A & Erb TJ (2019) Design and control of extrachromosomal elements in *Methylorubrum extorquens* AM1. *ACS synthetic biology* 8(11):2451-2456.

Cech TR (2000) The ribosome is a ribozyme. *Science* 289(5481):878-879.

Ceres P, Garst AD, Marcano-Velázquez JG & Batey RT (2013) Modularity of select riboswitch expression platforms enables facile engineering of novel genetic regulatory devices. *ACS synthetic biology* 2(8):463-472.

Chao Y, Papenfort K, Reinhardt R, Sharma CM & Vogel J (2012) An atlas of Hfq-bound transcripts reveals 3′ UTRs as a genomic reservoir of regulatory small RNAs. *The EMBO journal* 31(20):4005-4019.

Chen X, Arachchilage GM & Breaker RR (2019) Biochemical validation of a second class of tetrahydrofolate riboswitches in bacteria. *RNA* 25(9):1091-1097.

Chistoserdova L (2018) Applications of methylotrophs: can single carbon be harnessed for biotechnology? *Current opinion in biotechnology* 50:189-194.

Chou H-H, Berthet J & Marx CJ (2009) Fast growth increases the selective advantage of a mutation arising recurrently during evolution under metal limitation. *PLoS genetics* 5(9):e1000652.

Christopoulou N & Granneman S (2021) The role of RNA-binding proteins in mediating adaptive responses in Gram-positive bacteria. *The FEBS Journal*.

Chubiz LM, Purswani J, Carroll SM & Marx CJ (2013) A novel pair of inducible expression vectors for use in *Methylobacterium extorquens*. *BMC research notes* 6(1):1-8.

Cohen S, McMurry L & Levy S (1988) marA locus causes decreased expression of OmpF porin in multiple-antibiotic-resistant (Mar) mutants of *Escherichia coli*. *Journal of bacteriology* 170(12):5416-5422.

Corbino KA, Barrick JE, Lim J, Welz R, Tucker BJ, Puskarz I, Mandal M, Rudnick ND & Breaker RR (2005) Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome biology* 6(8):1-10.

Cotter RI, McPhie P & Gratzer W (1967) Internal organization of the ribosome. *Nature* 216(5118):864-868.

Crick F, Barnett L, Brenner S & Watts-Tobin RJ (1961) General nature of the genetic code for proteins.

Cui Y, Chatterjee A, Liu Y, Dumenyo CK & Chatterjee AK (1995) Identification of a global repressor gene, rsmA, of *Erwinia carotovora* subsp. carotovora that controls extracellular enzymes, N-(3-oxohexanoyl)-L-homoserine lactone, and pathogenicity in soft-rotting *Erwinia* spp. *Journal of bacteriology* 177(17):5108-5115.

Dambach M, Sandoval M, Updegrove TB, Anantharaman V, Aravind L, Waters LS & Storz G (2015) The ubiquitous yybP-ykoY riboswitch is a manganese-responsive regulatory element. *Molecular cell* 57(6):1099-1109.

Dann III CE, Wakeman CA, Sieling CL, Baker SC, Irnov I & Winkler WC (2007) Structure and mechanism of a metal-sensing regulatory RNA. *Cell* 130(5):878-892.

Dar D, Shamir M, Mellin J, Koutero M, Stern-Ginossar N, Cossart P & Sorek R (2016) Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria. *Science* 352(6282):aad9822.

Darmostuk M, Rimpelova S, Gbelcova H & Ruml T (2015) Current approaches in SELEX: An update to aptamer selection technology. *Biotechnology advances* 33(6):1141-1161.

Davis BM, Quinones M, Pratt J, Ding Y & Waldor MK (2005) Characterization of the small untranslated RNA RyhB and its regulon in *Vibrio cholerae*. *Journal of bacteriology* 187(12):4005-4014.

Deckers-Hebestreit G & Altendorf K (1996) The F0F1-type ATP synthases of bacteria: structure and function of the F0 complex. *Annual review of microbiology* 50:791-826.

de Fernandez MTF, Hayward WS & August JT (1972) Bacterial proteins required for replication of phage Qβ ribonucleic acid: purification and properties of host factor I, a ribonucleic acid-binding protein. *Journal of Biological Chemistry* 247(3):824-831.

De la Peña M, García-Robles I & Cervera A (2017) The hammerhead ribozyme: a long history for a short RNA. *Molecules* 22(1):78.

del Val C, Romero-Zaliz R, Torres-Quesada O, Peregrina A, Toro N & Jiménez-Zurdo JI (2012) A survey of sRNA families in α-proteobacteria. *RNA biology* 9(2):119-129.

Delaney NF, Kaczmarek ME, Ward LM, Swanson PK, Lee M-C & Marx CJ (2013) Development of an optimized medium, strain and high-throughput culturing methods for *Methylobacterium extorquens*. *PLoS One* 8(4):e62957.

Delihas N & Forst S (2001) MicF: an antisense RNA gene involved in response of *Escherichia coli* to global stress factors. *Journal of molecular biology* 313(1):1-12.

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C & Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences* 106(38):16428-16433.

Dendooven T, Sinha D, Roeselová A, Cameron TA, De Lay NR, Luisi BF & Bandyra KJ (2021) A cooperative PNPase-Hfq-RNA carrier complex facilitates bacterial riboregulation. *Molecular Cell* 81(14):2901-2913. e2905.

Dong H, Peng X, Wang N & Wu Q (2014) Identification of novel sRNAs in *Brucella abortus* 2308. *FEMS microbiology letters* 354(2):119-125.

Doronina N, Sokolov A & Trotsenko YA (1996) Isolation and initial characterization of aerobic chloromethane-utilizing bacteria. *FEMS microbiology letters* 142(2-3):179-183.

Drevets DA & Bronze MS (2008) *Listeria monocytogenes*: epidemiology, human disease, and mechanisms of brain invasion. *FEMS Immunology & Medical Microbiology* 53(2):151-165.

Dühring U, Axmann IM, Hess WR & Wilde A (2006) An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proceedings of the National Academy of Sciences* 103(18):7054-7058.

Eichner H, Karlsson J & Loh E (2022) The emerging role of bacterial regulatory RNAs in disease. *Trends in Microbiology*.

Eichner H, Karlsson J, Spelmink L, Pathak A, Sham L-T, Henriques-Normark B & Loh E (2021) RNA thermosensors facilitate *Streptococcus pneumoniae* and *Haemophilus influenzae* immune evasion. *PLoS Pathogens* 17(4):e1009513.

Ellington AD & Szostak JW (1990) In vitro selection of RNA molecules that bind specific ligands. *nature* 346(6287):818-822.

Errington J & van der Aart LT (2020) Microbe Profile: *Bacillus subtilis*: model organism for cellular development, and industrial workhorse. *Microbiology* 166(5):425.

Espelund M, Stacy R & Jakobsen K (1990) A simple method for generating single-stranded DNA probes labeled to high activities. *Nucleic acids research* 18(20):6157.

Faubladier M & Bouché J-P (1994) Division inhibition gene *dicF* of *Escherichia coli* reveals a widespread group of prophage sequences in bacterial genomes. *Journal of bacteriology* 176(4):1150-1156.

Felden B, Himeno H, Muto A, McCUTCHEON JP, Atkins JF & Gesteland RF (1997) Probing the structure of the *Escherichia coli* 10Sa RNA (tmRNA). *Rna* 3(1):89-103.

Figueira MM, Laramée L, Murrell JC, Groleau D & Miguez CB (2000) Production of green fluorescent protein by the methylotrophic bacterium *Methylobacterium extorquens*. *FEMS Microbiology letters* 193(2):195-200.

Figueroa-Bossi N, Schwartz A, Guillemardet B, D'Heygère F, Bossi L & Boudvillain M (2014) RNA remodeling by bacterial global regulator CsrA promotes Rho-dependent transcription termination. *Genes & development* 28(11):1239-1251.

Fiola K, Perreault J-P & Cousineau B (2006) Gene targeting in the Gram-Positive bacterium *Lactococcus lactis*, using various delta ribozymes. *Applied and Environmental Microbiology* 72(1):869-879.

Fitzgerald KA & Lidstrom ME (2003) Overexpression of a heterologous protein, haloalkane dehalogenase, in a poly-β-hydroxybutyrate-deficient strain of the facultative methylotroph *Methylobacterium extorquens* AM1. *Biotechnology and bioengineering* 81(3):263-268.

Fong N, Burgess M, Barrow K & Glenn D (2001) Carotenoid accumulation in the psychrotrophic bacterium Arthrobacter agilis in response to thermal and salt stress. *Applied microbiology and biotechnology* 56(5):750-756.

Fong YH, Wong HC, Yuen MH, Lau PH, Chen YW & Wong K-B (2013) Structure of UreG/UreF/UreH complex reveals how urease accessory proteins facilitate maturation of *Helicobacter pylori* urease. *PLoS biology* 11(10):e1001678.

Forster SC, Finkel AM, Gould JA & Hertzog PJ (2013) RNA-eXpress annotates novel transcript features in RNA-seq data. *Bioinformatics* 29(6):810-812.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW & DeLong EF (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences* 105(10):3805-3810.

Fuchs M, Lamm-Schmidt V, Sulzer J, Ponath F, Jenniches L, Kirk JA, Fagan RP, Barquist L, Vogel J & Faber F (2021) An RNA-centric global view of Clostridioides difficile reveals broad activity of Hfq in a clinically important gram-positive bacterium. *Proceedings of the National Academy of Sciences* 118(25).

Fuli X, Wenlong Z, Xiao W, Jing Z, Baohai H, Zhengzheng Z, Bin-Guang M & Youguo L (2017) A genome-wide prediction and identification of intergenic small RNAs by comparative analysis in *Mesorhizobium huakuii* 7653R. *Frontiers in microbiology* 8:1730.

Furukawa K, Ramesh A, Zhou Z, Weinberg Z, Vallery T, Winkler WC & Breaker RR (2015) Bacterial riboswitches cooperatively bind Ni2+ or Co2+ ions and control expression of heavy metal transporters. *Molecular cell* 57(6):1088-1098.

G. Chaulk S, Smith− Frieday MN, Arthur DC, Culham DE, Edwards RA, Soo P, Frost LS, Keates RA, Glover JM & Wood JM (2011) ProQ is an RNA chaperone that controls ProP levels in *Escherichia coli*. *Biochemistry* 50(15):3095-3106.

Garai P, Gnanadhas DP & Chakravortty D (2012) *Salmonella enterica* serovars Typhimurium and Typhi as model organisms: revealing paradigm of host-pathogen interactions. *Virulence* 3(4):377-388.

García-Solache M & Rice LB (2019) The *Enterococcus*: a model of adaptability to its environment. *Clinical microbiology reviews* 32(2):e00058-00018.

Gelfand MS, Mironov AA, Jomantas J, Kozlov YI & Perumov DA (1999) A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends in Genetics* 15(11):439-442.

Georg J & Hess WR (2011) cis-antisense RNA, another level of gene regulation in bacteria. *Microbiology and Molecular Biology Reviews* 75(2):286-300.

Gerhart E, Wagner H & Nordström K (1986) Structural analysis of an RNA molecule involved in replication control of plasmid RI. *Nucleic acids research* 14(6):2523-2538.

Gottesman S (2005) Micros for microbes: non-coding regulatory RNAs in bacteria. *TRENDS in Genetics* 21(7):399-404.

Gourion B, Francez-Charlot A & Vorholt JA (2008) PhyR is involved in the general stress response of *Methylobacterium extorquens* AM1. *Journal of bacteriology* 190(3):1027-1035.

Green PN & Ardley JK (2018) Review of the genus *Methylobacterium* and closely related organisms: a proposal that some Methylobacterium species be reclassified into a new genus, *Methylorubrum* gen. nov. *International journal of systematic and evolutionary microbiology*.

Greenlee EB, Stav S, Atilho RM, Brewer KI, Harris KA, Malkowski SN, Mirihana Arachchilage G, Perkins KR, Sherlock ME & Breaker RR (2018) Challenges of ligand identification for the second wave of orphan riboswitch candidates. *RNA biology* 15(3):377-390.

Gruber AR, Findeiß S, Washietl S, Hofacker IL & Stadler PF (2010) RNAz 2.0: improved noncoding RNA detection. *Biocomputing 2010,* World Scientific. p 69-79.

Gruber AR, Lorenz R, Bernhart SH, Neuböck R & Hofacker IL (2008) The vienna RNA websuite. *Nucleic acids research* 36(suppl_2):W70-W74.

Grundy FJ, Lehman SC & Henkin TM (2003) The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proceedings of the National Academy of Sciences* 100(21):12057-12062.

Guerrier-Takada C, Gardiner K, Marsh T, Pace N & Altman S (1983) The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* 35(3):849-857.

Guillier M & Gottesman S (2006) Remodelling of the *Escherichia coli* outer membrane by two small regulatory RNAs. *Molecular microbiology* 59(1):231-247.

Hajjar R, Ambaraghassi G, Sebajang H, Schwenter F & Su S-H (2020) *Raoultella ornithinolytica*: emergence and resistance. *Infection and Drug Resistance* 13:1091.

Haller A, Altman RB, Soulière MF, Blanchard SC & Micura R (2013) Folding and ligand recognition of the TPP riboswitch aptamer at single-molecule resolution. *Proceedings of the National Academy of Sciences* 110(11):4188-4193.

Haller A, Rieder U, Aigner M, Blanchard SC & Micura R (2011) Conformational capture of the SAM-II riboswitch. *Nature chemical biology* 7(6):393-400.

Hamal Dhakal S, Panchapakesan SS, Slattery P, Roth A & Breaker RR (2022) Variants of the guanine riboswitch class exhibit altered ligand specificities for xanthine, guanine, or 2′-deoxyguanosine. *Proceedings of the National Academy of Sciences* 119(22):e2120246119.

Hämmerle H, Amman F, Večerek B, Stülke J, Hofacker I & Blaesi U (2014) Impact of Hfq on the *Bacillus subtilis* transcriptome. *PloS one* 9(6):e98661.

Hampel KJ & Tinsley MM (2006) Evidence for preorganization of the glmS ribozyme ligand binding pocket. *Biochemistry* 45(25):7861-7871.

Heidrich N, Moll I & Brantl S (2007) In vitro analysis of the interaction between the small RNA SR1 and its primary target *ahrC* mRNA. *Nucleic acids research* 35(13):4331-4346.

Henderson CA, Vincent HA, Casamento A, Stone CM, Phillips JO, Cary PD, Sobott F, Gowers DM, Taylor JE & Callaghan AJ (2013) Hfq binding changes the structure of *Escherichia coli* small noncoding RNAs OxyS and RprA, which are involved in the riboregulation of *rpoS*. *Rna* 19(8):1089-1104.

Hennecke H (1990) Regulation of bacterial gene expression by metal–protein complexes. *Molecular microbiology* 4(10):1621-1628.

Heppell B, Blouin S, Dussault A-M, Mulhbacher J, Ennifar E, Penedo JC & Lafontaine DA (2011) Molecular insights into the ligand-controlled organization of the SAM-I riboswitch. *Nature chemical biology* 7(6):384-392.

Heroven A, Sest M, Pisano F, Scheb-Wetzel M, Böhme K, Klein J, Münch R, Schomburg D & Dersch P (2012) Crp induces switching of the CsrB and CsrC RNAs in *Yersinia pseudotuberculosis* and links nutritional status to virulence. *Frontiers in cellular and infection microbiology* 2:158.

Hess C, Enichlmayr H, Jandreski-Cvetkovic D, Liebhart D, Bilic I & Hess M (2013) Riemerella anatipestifer outbreaks in commercial goose flocks and identification of isolates by MALDI-TOF mass spectrometry. *Avian Pathology* 42(2):151-156.

Heyne S, Costa F, Rose D & Backofen R (2012) GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* 28(12):i224-i232.

Hoagland MB, Stephenson ML, Scott JF, Hecht LI & Zamecnik PC (1958) A soluble ribonucleic acid intermediate in protein synthesis. *J. biol. Chem* 231(1):241-257.

Holmqvist E, Berggren S & Rizvanovic A (2020) RNA-binding activity and regulatory functions of the emerging sRNA-binding protein ProQ. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863(9):194596.

Holmqvist E, Li L, Bischler T, Barquist L & Vogel J (2018) Global maps of ProQ binding in vivo reveal target recognition via RNA structure and stability control at mRNA 3′ ends. *Molecular cell* 70(5):971-982. e976.

Holmqvist E, Reimegård J, Sterk M, Grantcharova N, Römling U & Wagner EGH (2010) Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *The EMBO journal* 29(11):1840-1850.

Hoopes L (2008) Introduction to the gene expression and regulation topic room. *Nature Education* 1(1):160.

Hou Y-M (1993) The tertiary structure of tRNA and the development of the genetic code. *Trends in biochemical sciences* 18(10):362-364.

Howe JA, Xiao L, Fischmann TO, Wang H, Tang H, Villafania A, Zhang R, Barbieri CM & Roemer T (2016) Atomic resolution mechanistic studies of ribocil: a highly selective unnatural ligand mimic of the *E. coli* FMN riboswitch. *RNA biology* 13(10):946-954.

Huang H-Y, Chang H-Y, Chou C-H, Tseng C-P, Ho S-Y, Yang C-D, Ju Y-W & Huang H-D (2009) sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic acids research* 37(suppl_1):D150-D154.

Huang L, Serganov A & Patel DJ (2010) Structural insights into ligand recognition by a sensing domain of the cooperative glycine riboswitch. *Molecular cell* 40(5):774-786.

Hutchison III CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ & Ma L (2016) Design and synthesis of a minimal bacterial genome. *Science* 351(6280):aad6253.

Ikemura T & Dahlberg JE (1973) Small ribonucleic acids of *Escherichia coli*: I. Characterization by polyacrylamide gel electrophoresis and fingerprint analysis. *Journal of Biological Chemistry* 248(14):5024-5032.

Imane R (2016) *Optimisation de la production d'acide succinique chez Methylbacterium extorquens par le biais de petits arn régulateurs*. (Université du Québec, Institut National de la Recherche Scientifique).

Irvine D, Tuerk C & Gold L (1991) Selexion: Systematic evolution of ligands by exponential enrichment with integrated optimization by non-linear analysis. *Journal of molecular biology* 222(3):739-761.

Jia X, Zhang J, Sun W, He W, Jiang H, Chen D & Murchie AI (2013) Riboswitch control of aminoglycoside antibiotic resistance. *Cell* 152(1-2):68-81.

Jiang X, Liu X, Law CO, Wang Y, Lo WU, Weng X, Chan TF, Ho P & Lau TC (2017) The CTX-M-14 plasmid pHK01 encodes novel small RNAs and influences host growth and motility. *FEMS Microbiology Ecology* 93(7).

Johnson Jr JE, Reyes FE, Polaski JT & Batey RT (2012) B12 cofactors directly stabilize an mRNA regulatory switch. *Nature* 492(7427):133-137.

Jones CP & Ferré-D'Amaré AR (2015) Recognition of the bacterial alarmone ZMP through long-distance association of two RNA subdomains. *Nature structural & molecular biology* 22(9):679-685.

Jørgensen MG, Pettersen JS & Kallipolitis BH (2020) sRNA-mediated control in bacteria: An increasing diversity of regulatory mechanisms. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1863(5):194504.

Kaczmarczyk A, Vorholt JA & Francez-Charlot A (2013) Cumate-inducible gene expression system for sphingomonads and other Alphaproteobacteria. *Applied and environmental microbiology* 79(21):6795-6802.

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D & Weinberg Z (2021) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* 49(D1):D192-D200.

Kang M, Eichhorn CD & Feigon J (2014) Structural determinants for ligand capture by a class II preQ1 riboswitch. *Proceedings of the National Academy of Sciences* 111(6):E663-E671.

Keseler IM, Bonavides-Martínez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S & Paulsen IT (2009) EcoCyc: a comprehensive view of *Escherichia coli* biology. *Nucleic acids research* 37(suppl_1):D464-D470.

Khanna A, Khanna M & Aggarwal A (2013) *Serratia marcescens*-a rare opportunistic nosocomial pathogen and measures to limit its spread in hospitalized patients. *Journal of clinical and diagnostic research: JCDR* 7(2):243.

Kharma N, Varin L, Abu-Baker A, Ouellet J, Najeh S, Ehdaeivand M-R, Belmonte G, Ambri A, Rouleau G & Perreault J (2016) Automated design of hammerhead ribozymes and validation by targeting the PABPN1 gene transcript. *Nucleic acids research* 44(4):e39-e39.

Killackey SA, Sorbara MT & Girardin SE (2016) Cellular aspects of *Shigella pathogenesis*: focus on the manipulation of host cell processes. *Frontiers in cellular and infection microbiology* 6:38.

Kim JN, Roth A & Breaker RR (2007) Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2′-deoxyguanosine. *Proceedings of the National Academy of Sciences* 104(41):16092-16097.

Kim PB, Nelson JW & Breaker RR (2015) An ancient riboswitch class in bacteria regulates purine biosynthesis and one-carbon metabolism. *Molecular cell* 57(2):317-328.

Kittle J, Simons RW, Lee J & Kleckner N (1989) Insertion sequence IS10 anti-sense pairing initiates by an interaction between the 5′ end of the target RNA and a loop in the anti-sense RNA. *Journal of molecular biology* 210(3):561-572.

Klein DJ & Ferré-D'Amaré AR (2006) Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science* 313(5794):1752-1756.

Klein RJ, Misulovin Z & Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proceedings of the National Academy of Sciences* 99(11):7542-7547.

Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick JB, Grout J & Corlay S (2016) *Jupyter Notebooks-a publishing format for reproducible computational workflows.*

Kohler-Staub D, Hartmans S, Gälli R, Suter F & Leisinger T (1986) Evidence for identical dichloromethane dehalogenases in different methylotrophic bacteria. *Microbiology* 132(10):2837-2843.

Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE & Cech TR (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *cell* 31(1):147-157.

Krzyściak W, Pluskwa K, Jurczak A & Kościelniak D (2013) The pathogenicity of the *Streptococcus genus*. *European Journal of Clinical Microbiology & Infectious Diseases* 32(11):1361-1376.

Kwon M & Strobel SA (2008) Chemical basis of glycine riboswitch cooperativity. *Rna* 14(1):25-34.

Lafontaine DA, Norman DG & Lilley DM (2001) Structure, folding and activity of the VS ribozyme: importance of the 2-3-6 helical junction. *The EMBO Journal* 20(6):1415-1424.

Lalaouna D, Prévost K, Park S, Chénard T, Bouchard M-P, Caron M-P, Vanderpool CK, Fei J & Massé E (2021) Binding of the RNA Chaperone Hfq on Target mRNAs Promotes the Small RNA RyhB-Induced Degradation in *Escherichia coli*. *Non-coding RNA* 7(4):64.

Lalaouna D, Simoneau-Roy M, Lafontaine D & Massé E (2013) Regulatory RNAs and target mRNA decay in prokaryotes. *BBA - Gene Regulatory Mechanisms* 1829(6-7):742-747.

Lamarche MG, Perreault J, Miguez C, Arbour M & Choi YJ (2018) Genetically engineered c1-utilizing microorganisms and processes for their production and use. WO/2016/165025

Landt SG, Abeliuk E, McGrath PT, Lesley JA, McAdams HH & Shapiro L (2008) Small non-coding RNAs in *Caulobacter crescentus*. *Molecular microbiology* 68(3):600-614.

Le HTQ, Mai DHA, Na J-G & Lee EY (2022) Development of *Methylorubrum extorquens* AM1 as a promising platform strain for enhanced violacein production from co-utilization of methanol and acetate. *Metab. Eng.* 72:150-160.

Le Moal G, Landron C, Grollier G, Robert R & Burucoa C (2003) Meningitis due to *Capnocytophaga canimorsus* after receipt of a dog bite: case report and review of the literature. *Clinical infectious diseases* 36(3):e42-e46.

Lee ER, Baker JL, Weinberg Z, Sudarsan N & Breaker RR (2010) An allosteric self-splicing ribozyme triggered by a bacterial second messenger. *science* 329(5993):845-848.

Lee YJ & Moon TS (2018) Design rules of synthetic non-coding RNAs in bacteria. *Methods* 143:58-69.

Leonard S, Meyer S, Lacour S, Nasser W, Hommais F & Reverchon S (2019) APERO: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data. *Nucleic acids research* 47(15):e88-e88.

Li L, Huang D, Cheung MK, Nong W, Huang Q & Kwan HS (2013) BSRD: a repository for bacterial small regulatory RNA. *Nucleic acids research* 41(D1):D233-D238.

Li S, Hwang XY, Stav S & Breaker RR (2016) The yjdF riboswitch candidate regulates gene expression by binding diverse azaaromatic compounds. *Rna* 22(4):530-541.

Liberman JA, Suddala KC, Aytenfisu A, Chan D, Belashov IA, Salim M, Mathews DH, Spitale RC, Walter NG & Wedekind JE (2015) Structural analysis of a class III preQ1 riboswitch reveals an aptamer distant from a ribosome-binding site regulated by fast dynamics. *Proceedings of the National Academy of Sciences* 112(27):E3485-E3494.

Light J & Molin S (1983) Post-transcriptional control of expression of the repA gene of plasmid R1 mediated by a small RNA molecule. *The EMBO journal* 2(1):93-98.

Lim CK, Villada JC, Chalifour A, Duran MF, Lu H & Lee PK (2019) Designing and engineering *Methylorubrum extorquens* AM1 for itaconic acid production. *Frontiers in microbiology* 10:1027.

Liu JM & Camilli A (2011) Discovery of bacterial sRNAs by high-throughput sequencing. *High-Throughput Next Generation Sequencing,* Springer. p 63-79.

Liu JM, Livny J, Lawrence MS, Kimball MD, Waldor MK & Camilli A (2009) Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic acids research* 37(6):e46-e46.

Liu MY, Gui G, Wei B, Preston JF, Oakford L, Yüksel Um, Giedroc DP & Romeo T (1997) The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. *Journal of Biological Chemistry* 272(28):17502-17510.

Liu MY & Romeo T (1997) The global regulator CsrA of *Escherichia coli* is a specific mRNA-binding protein. *Journal of bacteriology* 179(14):4639-4642.

Livny J, Brencic A, Lory S & Waldor MK (2006) Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic acids research* 34(12):3484-3493.

Livny J & Waldor MK (2007) Identification of small RNAs in diverse bacterial species. *Current opinion in microbiology* 10(2):96-101.

Locher KP (2009) Structure and mechanism of ATP-binding cassette transporters. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364(1514):239-245.

López-Gomollón S (2011) Detecting sRNAs by Northern blotting. *MicroRNAs in Development,* Springer. p 25-38.

Lorenz C, Von Pelchrzim F & Schroeder R (2006) Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nature Protocols* 1(5):2204-2212.

Lyu C, Khan IM & Wang Z (2021) Capture-SELEX for aptamer selection: A short review. *Talanta* 229:122274.

Mackie GA & Genereaux JL (1993) The role of RNA structure in determining RNase E-dependent cleavage sites in the mRNA for ribosomal protein S20 in vitro. *Journal of molecular biology* 234(4):998-1012.

Madhugiri R, Pessi G, Voss B, Hahn J, Sharma CM, Reinhardt R, Vogel J, Hess WR, Fischer H-M & Evguenieva-Hackenberg E (2012) Small RNAs of the *Bradyrhizobium/Rhodopseudomonas* lineage and their analysis. *RNA biology* 9(1):47-58.

Majdalani N, Chen S, Murrow J, St John K & Gottesman S (2001) Regulation of RpoS by a novel small RNA: the characterization of RprA. *Molecular microbiology* 39(5):1382-1394.

Malkowski SN, Atilho RM, Greenlee EB, Weinberg CE & Breaker RR (2020) A rare bacterial RNA motif is implicated in the regulation of the *purF* gene whose encoded enzyme synthesizes phosphoribosylamine. *RNA* 26(12):1838-1846.

Malkowski SN, Spencer TC & Breaker RR (2019) Evidence that the nadA motif is a bacterial riboswitch for the ubiquitous enzyme cofactor NAD+. *Rna* 25(12):1616-1627.

Mandal M, Boese B, Barrick JE, Winkler WC & Breaker RR (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* 113(5):577-586.

Mandal M & Breaker RR (2004) Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature structural & molecular biology* 11(1):29-35.

Mandal M, Lee M, Barrick JE, Weinberg Z, Emilsson GM, Ruzzo WL & Breaker RR (2004) A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* 306(5694):275-279.

Mandin P, Repoila F, Vergassola M, Geissmann T & Cossart P (2007) Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. *Nucleic acids research* 35(3):962-974.

Mann M, Wright PR & Backofen R (2017) IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic acids research* 45(W1):W435-W439.

Marnocha C, Levy A, Powell D, Hanson T & Chan C (2016) Mechanisms of extracellular S0 globule production and degradation in *Chlorobaculum tepidum* via dynamic cell–globule interactions. *Microbiology* 162(7):1125.

Martick M & Scott WG (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* 126(2):309-320.

Marx CJ & Lidstrom ME (2001) Development of improved versatile broad-host-range vectors for use in methylotrophs and other Gram-negative bacteria. *Microbiology* (Reading, England) 147(Pt 8):2065-2075.

Marx CJ & Lidstrom ME (2002) Broad-host-range cre-lox system for antibiotic marker recycling in gram-negative bacteria. *Biotechniques* 33(5):1062-1067.

Massé E & Gottesman S (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proceedings of the National Academy of Sciences* 99(7):4620-4625.

Massé E, Vanderpool CK & Gottesman S (2005) Effect of RyhB small RNA on global iron use in *Escherichia coli*. *Journal of bacteriology* 187(20):6962-6971.

Mathpal S & Kandpal N (2009) Colorimetric estimation of Ni (II) Ions in aqueous solution. *E-Journal of Chemistry* 6(2):445-448.

Matylla-Kulinska K, Boots JL, Zimmermann B & Schroeder R (2012) Finding aptamers and small ribozymes in unexpected places. *Wiley Interdisciplinary Reviews: RNA* 3(1):73-91.

Maucourt B, Roche D, Chaignaud P, Vuilleumier S & Bringel F (2022) Genome-wide transcription start sites mapping in *Methylorubrum* grown with dichloromethane and methanol. *Microorganisms* 10(7):1301.

Mayoral JG, Hussain M, Joubert DA, Iturbe-Ormaetxe I, O'Neill SL & Asgari S (2014) Wolbachia small noncoding RNAs and their role in cross-kingdom communications. *Proceedings of the National Academy of Sciences* 111(52):18721-18726.

McCarthy TJ, Plog MA, Floy SA, Jansen JA, Soukup JK & Soukup GA (2005) Ligand requirements for glmS ribozyme self-cleavage. *Chemistry & biology* 12(11):1221-1226.

McCown PJ, Corbino KA, Stav S, Sherlock ME & Breaker RR (2017) Riboswitch diversity and distribution. *Rna* 23(7):995-1011.

McCown PJ, Liang JJ, Weinberg Z & Breaker RR (2014) Structural, functional, and taxonomic diversity of three preQ1 riboswitch classes. *Chemistry & biology* 21(7):880-889.

McKellar SW, Ivanova I, Arede P, Zapf RL, Mercier N, Chu L-C, Mediati DG, Pickering AC, Briaud P & Foster RG (2022) RNase III CLASH in MRSA uncovers sRNA regulatory networks coupling metabolism to toxin expression. *Nature Communications* 13(1):1-20

Mediati DG, Wong JL, Gao W, McKellar S, Pang CNI, Wu S, Wu W, Sy B, Monk IR & Biazik JM (2022) RNase III-CLASH of multi-drug resistant *Staphylococcus aureus* reveals a regulatory mRNA 3′ UTR required for intermediate vancomycin resistance. *Nature Communications* 13(1):1-15.

Mei L, Xu S, Lu P, Lin H, Guo Y & Wang Y (2017) CsrB, a noncoding regulatory RNA, is required for BarA-dependent expression of biocontrol traits in *Rahnella aquatilis* HX2. *Plos one* 12(11):e0187492.

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ & Breaker RR (2009) Identification of candidate structured RNAs in the marine organism'*Candidatus* Pelagibacter ubique'. *BMC genomics* 10(1):1-16.

Meyer MM, Hammond MC, Salinas Y, Roth A, Sudarsan N & Breaker RR (2011) Challenges of ligand identification for riboswitch candidates. *RNA biology* 8(1):5-10.

Michel F & Ferat J-L (1995) Structure and activities of group II introns. *Annual review of biochemistry* 64(1):435-461.

Mikulecky PJ, Kaw MK, Brescia CC, Takach JC, Sledjeski DD & Feig AL (2004) *Escherichia coli* Hfq has distinct interaction surfaces for DsrA, rpoS and poly (A) RNAs. *Nature structural & molecular biology* 11(12):1206-1214.

Milner JL & Wood JM (1989) Insertion proQ220:: Tn5 alters regulation of proline porter II, a transporter of proline and glycine betaine in *Escherichia coli*. *Journal of bacteriology* 171(2):947-951.

Miranda-Ríos J, Navarro M & Soberón M (2001) A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proceedings of the National Academy of Sciences* 98(17):9736-9741.

Mirihana Arachchilage G, Sherlock ME, Weinberg Z & Breaker RR (2018) SAM-VI RNAs selectively bind S-adenosylmethionine and exhibit similarities to SAM-III riboswitches. *RNA biology* 15(3):371-378.

Mironov AS, Gusarov I, Rafikov R, Lopez LE, Shatalin K, Kreneva RA, Perumov DA & Nudler E (2002) Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* 111(5):747-756.

Mizuno T, Chou M-Y & Inouye M (1984) A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proceedings of the National Academy of Sciences* 81(7):1966-1970.

Mo X-H, Zhang H, Wang T-M, Zhang C, Zhang C, Xing X-H & Yang S (2020) Establishment of CRISPR interference in *Methylorubrum extorquens* and application of rapidly mining a new phytoene desaturase involved in carotenoid biosynthesis. *Applied Microbiology and Biotechnology* 104(10):4515-4532.

Moll I, Afonyushkin T, Vytvytska O, Kaberdin VR & Bläsi U (2003) Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs. *Rna* 9(11):1308-1314.

Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG & Valentin-Hansen P (2002a) Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Molecular cell* 9(1):23-30.

Møller T, Franch T, Udesen C, Gerdes K & Valentin-Hansen P (2002b) Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes & development* 16(13):1696-1706.

Morita T, Maki K & Aiba H (2005) RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs. *Genes & development* 19(18):2176-2186.

Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF & Hofacker IL (2006) Thermodynamics of RNA–RNA binding. *Bioinformatics* 22(10):1177-1182.

Müller P, Gimpel M, Wildenhain T & Brantl S (2019) A new role for CsrA: promotion of complex formation between an sRNA and its mRNA target in *Bacillus subtilis*. *RNA biology* 16(7):972-987.

Murashko ON & Lin-Chao S (2017) *Escherichia coli* responds to environmental changes using enolasic degradosomes and stabilized DicF sRNA to alter cellular morphology. *Proceedings of the National Academy of Sciences* 114(38):E8025-E8034.

Na D, Yoo SM, Chung H, Park H, Park JH & Lee SY (2013) Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. *Nature biotechnology* 31(2):170-174.

Naghdi MR, Boutet E, Mucha C, Ouellet J & Perreault J (2020) Single mutation in hammerhead ribozyme favors cleavage activity with manganese over magnesium. *Non-coding RNA* 6(1):14.

Naghdi MR, Smail K, Wang JX, Wade F, Breaker RR & Perreault J (2017) Search for 5′-leader regulatory RNA structures based on gene annotation aided by the RiboGap database. *Methods* 117:3-13.

Nahvi A, Sudarsan N, Ebert MS, Zou X, Brown KL & Breaker RR (2002) Genetic control by a metabolite binding mRNA. *Chemistry & biology* 9(9):1043-1049.

Najeh S (2017) *Développement d'un circuit logique basé sur les ARN non codants dans les bactéries*. (Université du Québec, Institut national de la recherche scientifique).

Naville M, Ghuillot-Gaudeffroy A, Marchais A & Gautheret D (2011) ARNold: a web tool for the prediction of Rho-independent transcription terminators. *RNA biology* 8(1):11-13.

Nawrocki EP & Eddy SR (2013) Computational identification of functional RNA homologs in metagenomic data. *RNA biology* 10(7):1170-1179.

Nelson JW, Atilho RM, Sherlock ME, Stockbridge RB & Breaker RR (2017) Metabolism of free guanidine in bacteria is regulated by a widespread riboswitch class. *Molecular cell* 65(2):220-230.

Nelson JW, Sudarsan N, Furukawa K, Weinberg Z, Wang JX & Breaker RR (2013) Riboswitches in eubacteria sense the second messenger c-di-AMP. *Nature chemical biology* 9(12):834-839.

Neuhaus K, Landstorfer R, Simon S, Schober S, Wright PR, Smith C, Backofen R, Wecko R, Keim DA & Scherer S (2017) Differentiation of ncRNAs from small mRNAs in *Escherichia coli* O157: H7 EDL933 (EHEC) by combined RNAseq and RIBOseq–ryhB encodes the regulatory RNA RyhB and a peptide, RyhP. *BMC genomics* 18(1):1-24.

Nielsen JS, Lei LK, Ebersbach T, Olsen AS, Klitgaard JK, Valentin-Hansen P & Kallipolitis BH (2010) Defining a role for Hfq in Gram-positive bacteria: evidence for Hfq-dependent antisense regulation in *Listeria monocytogenes*. *Nucleic acids research* 38(3):907-919.

Nordgren S, Slagter-Jäger JG & Wagner EGH (2001) Real time kinetic studies of the interaction between folded antisense and target RNAs using surface plasmon resonance. *Journal of molecular biology* 310(5):1125-1134.

Nuss AM, Heroven AK, Waldmann B, Reinkensmeier J, Jarek M, Beckstette M & Dersch P (2015) Transcriptomic profiling of *Yersinia pseudotuberculosis* reveals reprogramming of the Crp regulon by temperature and uncovers Crp as a master regulator of small RNAs. *PLoS genetics* 11(3):e1005087.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B & Ako-Adjei D (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* 44(D1):D733-D745.

Ochsner AM, Sonntag F, Buchhaupt M, Schrader J & Vorholt JA (2015) *Methylobacterium extorquens*: methylotrophy and biotechnological applications. *Applied Microbiology and Biotechnology* 99(2):517-534.

Okamura S, MARUYAMA HB & YANAGITA T (1973) Ribosome Degradation and Degradation Products in Starved *Escherichia coli*: VI. Prolonged Culture during Glucose Starvation. *The Journal of Biochemistry* 73(5):915-922.

Olah GA (2013) Towards Oil Independence Through Renewable Methanol Chemistry. *Angewandte Chemie International Edition* 52(1):104-107.

Olejniczak M & Storz G (2017) ProQ/FinO-domain proteins: another ubiquitous family of RNA matchmakers? *Molecular microbiology* 104(6):905-915.

Padalon-Brauch G, Hershberg R, Elgrably-Weiss M, Baruch K, Rosenshine I, Margalit H & Altuvia S (2008) Small RNAs encoded within genetic islands of *Salmonella typhimurium* show host-induced expression and role in virulence. *Nucleic acids research* 36(6):1913-1927.

Panchapakesan SS, Corey L, Malkowski SN, Higgs G & Breaker RR (2021) A second riboswitch class for the enzyme cofactor NAD+. *Rna* 27(1):99-105.

Papenfort K & Vogel J (2009) Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level. *Research in microbiology* 160(4):278-287.

Patterson-Fortin LM, Vakulskas CA, Yakhnin H, Babitzke P & Romeo T (2013) Dual posttranscriptional regulation via a cofactor-responsive mRNA leader. *Journal of molecular biology* 425(19):3662-3677.

Pedrolli D, Langer S, Hobl B, Schwarz J, Hashimoto M & Mack M (2015) The ribB FMN riboswitch from *Escherichia coli* operates at the transcriptional and translational level and regulates riboflavin biosynthesis. *The FEBS journal* 282(16):3230-3242.

Peña-Castillo L, Grüell M, Mulligan ME & Lang AS (2016) Detection of bacterial small transcripts from RNA-SEQ data: a comparative assessment. *Biocomputing 2016: Proceedings of the Pacific Symposium.* World Scientific, p 456-467.

Peyraud R, Kiefer P, Christen P, Portais J-C & Vorholt JA (2012) Co-consumption of methanol and succinate by *Methylobacterium extorquens* AM1. *PloS one* 7(11):e48271.

Pfeiffer V, Papenfort K, Lucchini S, Hinton JC & Vogel J (2009) Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nature structural & molecular biology* 16(8):840-846.

Pischimarov J, Kuenne C, Billion A, Hemberger J, Cemič F, Chakraborty T & Hain T (2012) sRNAdb: a small non-coding RNA database for gram-positive bacteria. *BMC genomics* 13(1):1-8.

Porcheron G, Habib R, Houle S, Caza M, Lépine F, Daigle F, Massé E & Dozois CM (2014) The small RNA RyhB contributes to siderophore production and virulence of uropathogenic *Escherichia coli*. *Infection and immunity* 82(12):5056-5068.

Price IR, Gaballa A, Ding F, Helmann JD & Ke A (2015) Mn2+-sensing mechanisms of yybP-ykoY orphan riboswitches. *Molecular cell* 57(6):1110-1123.

Prody GA, Bakos JT, Buzayan JM, Schneider IR & Bruening G (1986) Autolytic processing of dimeric plant virus satellite RNA. *Science* 231(4745):1577-1580.

Rabhi M, Espéli O, Schwartz A, Cayrol B, Rahmouni AR, Arluison V & Boudvillain M (2011) The Sm-like RNA chaperone Hfq mediates transcription antitermination at Rho-dependent terminators. *The EMBO journal* 30(14):2805-2816.

Ramesh A & Winkler WC (2010) Magnesium-sensing riboswitches in bacteria. *RNA biology* 7(1):77-83.

Ranjan K & Ranjan N (2013) Citrobacter: An emerging health care associated urinary pathogen. *Urology annals* 5(4):313.

Reader JS, Metzgar D, Schimmel P & de Crécy-Lagard V (2004) Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *Journal of Biological Chemistry* 279(8):6280-6285.

Regulski EE & Breaker RR (2008) In-line probing analysis of riboswitches. *Post-transcriptional gene regulation,* Springer. p 53-67.

Regulski EE, Moy RH, Weinberg Z, Barrick JE, Yao Z, Ruzzo WL & Breaker RR (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Molecular microbiology* 68(4):918-932.

Reichenbach B, Maes A, Kalamorz F, Hajnsdorf E & Görke B (2008) The small RNA GlmY acts upstream of the sRNA GlmZ in the activation of glmS expression and is subject to regulation by polyadenylation in *Escherichia coli*. *Nucleic acids research* 36(8):2570-2580.

Reiss CW, Xiong Y & Strobel SA (2017) Structural basis for ligand binding to the guanidine-I riboswitch. *Structure* 25(1):195-202.

Ren A, Rajashankar KR & Patel DJ (2012) Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature* 486(7401):85-89.

Rentmeister A, Mayer G, Kuhn N & Famulok M (2007) Conformational changes in the expression domain of the *Escherichia coli* thiM riboswitch. *Nucleic acids research* 35(11):3713-3722.

Rieder U, Kreutz C & Micura R (2010) Folding of a transcriptionally acting preQ1 riboswitch. *Proceedings of the National Academy of Sciences* 107(24):10804-10809.

Righetti F, Nuss AM, Twittenhoff C, Beele S, Urban K, Will S, Bernhart SH, Stadler PF, Dersch P & Narberhaus F (2016) Temperature-responsive in vitro RNA structurome of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences* 113(26):7237-7242.

Rio DC, Ares M, Hannon GJ & Nilsen TW (2010) Purification of RNA using TRIzol (TRI reagent). *Cold Spring Harbor Protocols* 2010(6):pdb. prot5439.

Rivas E, Klein RJ, Jones TA & Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Current biology* 11(17):1369-1373.

Rizzatti G, Lopetuso L, Gibiino G, Binda C & Gasbarrini A (2017) Proteobacteria: a common factor in human diseases. *BioMed research international* 2017.

Rodionov DA, Vitreschak AG, Mironov AA & Gelfand MS (2003) Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic acids research* 31(23):6748-6757.

Romeo T, Gong M, Liu MY & Brun-Zinkernagel A-M (1993) Identification and molecular characterization of csrA, a pleiotropic gene from *Escherichia coli* that affects glycogen biosynthesis, gluconeogenesis, cell size, and surface properties. *Journal of bacteriology* 175(15):4744-4755.

Romeo T, Vakulskas CA & Babitzke P (2013) Post-transcriptional regulation on a global scale: form and function of Csr/Rsm systems. *Environmental microbiology* 15(2):313-324.

Roth A, Weinberg Z, Chen AG, Kim PB, Ames TD & Breaker RR (2014) A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nature chemical biology* 10(1):56-60.

Roth A, Winkler WC, Regulski EE, Lee BW, Lim J, Jona I, Barrick JE, Ritwik A, Kim JN & Welz R (2007) A riboswitch selective for the queuosine precursor preQ 1 contains an unusually small aptamer domain. *Nature structural & molecular biology* 14(4):308-317.

Ruff KM & Strobel SA (2014) Ligand binding by the tandem glycine riboswitch depends on aptamer dimerization but not double ligand occupancy. *Rna* 20(11):1775-1788.

Saad T & Atallah S (2014) Studies on bacterial infection in marine fish. *Journal of the Arabian Aquaculture Society* 374(3354):1-20.

Saïdi F, Jolivet NY, Lemon DJ, Nakamura A, Belgrave AM, Garza AG, Veyrier FJ & Islam ST (2021) Bacterial glycocalyx integrity drives multicellular swarm biofilm dynamism. *Molecular Microbiology* 116(4):1151-1172.

Salvail H, Balaji A, Yu D, Roth A & Breaker RR (2020) Biochemical validation of a fourth guanidine riboswitch class in bacteria. *Biochemistry* 59(49):4654-4662.

Sanders Jr WE & Sanders CC (1997) Enterobacter spp.: pathogens poised to flourish at the turn of the century. *Clinical microbiology reviews* 10(2):220-241.

Santos-Zavaleta A, Salgado H, Gama-Castro S, Sánchez-Pérez M, Gómez-Romero L, Ledezma-Tejeida D, García-Sotelo JS, Alquicira-Hernández K, Muñiz-Rascado LJ & Peña-Loredo P (2019) RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic acids research* 47(D1):D212-D220.

Sassi M, Augagneur Y, Mauro T, Ivain L, Chabelskaya S, Hallier M, Sallou O & Felden B (2015) SRD: a *Staphylococcus* regulatory RNA database. *Rna* 21(5):1005-1017.

Sauer E, Schmidt S & Weichenrieder O (2012) Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proceedings of the National Academy of Sciences* 109(24):9396-9401.

Saville BJ & Collins RA (1990) A site-specific self-cleavage reaction performed by a novel RNA in *Neurospora mitochondria*. *Cell* 61(4):685-696.

Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M & Federhen S (2010) Database resources of the national center for biotechnology information. *Nucleic acids research* 39(suppl_1):D38-D51.

Schattner P (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Research* 30(9):2076-2082.

Schiano CA, Koo JT, Schipma MJ, Caulfield AJ, Jafari N & Lathem WW (2014) Genome-wide analysis of small RNAs expressed by *Yersinia pestis* identifies a regulator of the Yop-Ysc type III secretion system. *Journal of bacteriology* 196(9):1659-1670.

Schirch V (1998) Mechanism of folate-requiring enzymes in one-carbon metabolism. *Comprehensive biological catalysis* 1:211-252

Schmidtke C, Findeiß S, Sharma CM, Kuhfuß J, Hoffmann S, Vogel J, Stadler PF & Bonas U (2012) Genome-wide transcriptome analysis of the plant pathogen *Xanthomonas* identifies sRNAs with putative virulence functions. *Nucleic acids research* 40(5):2020-2031.

Schu DJ, Zhang A, Gottesman S & Storz G (2015) Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *The EMBO journal* 34(20):2557-2573.

Schumacher MA, Pearson RF, Møller T, Valentin-Hansen P & Brennan RG (2002) Structures of the pleiotropic translational regulator Hfq and an Hfq–RNA complex: a bacterial Sm-like protein. *The EMBO journal* 21(13):3546-3556.

Sena-Vélez M, Holland SD, Aggarwal M, Cogan NG, Jain M, Gabriel DW & Jones KM (2019) Growth dynamics and survival of *Liberibacter crescens* BT-1, an important model organism for the citrus Huanglongbing pathogen "*Candidatus* Liberibacter asiaticus". *Applied and environmental microbiology* 85(21):e01656-01619.

Serganov A, Huang L & Patel DJ (2008) Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* 455(7217):1263-1267.

Serganov A, Huang L & Patel DJ (2009) Coenzyme recognition and gene regulation by a flavin mononucleotide riboswitch. *nature* 458(7235):233-237.

Serganov A & Nudler E (2013) A decade of riboswitches. *Cell* 152(1-2):17-24.

Serganov A, Polonskaia A, Phan AT, Breaker RR & Patel DJ (2006) Structural basis for gene regulation by a thiamine pyrophosphate-sensing riboswitch. *Nature* 441(7097):1167-1171.

Setubal JC, Dos Santos P, Goldman BS, Ertesvåg H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D & Cromes L (2009) Genome sequence of *Azotobacter vinelandii*, an obligate aerobe specialized to support diverse anaerobic metabolic processes. *Journal of bacteriology* 191(14):4534-4545.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J & Reinhardt R (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* 464(7286):250-255.

Sharmeen L, Kuo M, Dinter-Gottlieb G & Taylor J (1988) Antigenomic RNA of human hepatitis delta virus can undergo self-cleavage. *Journal of virology* 62(8):2674-2679.

Sherlock ME & Breaker RR (2017) Biochemical validation of a third guanidine riboswitch class in bacteria. *Biochemistry* 56(2):359-363.

Sherlock ME & Breaker RR (2020) Former orphan riboswitches reveal unexplored areas of bacterial metabolism, signaling, and gene control processes. *RNA* 26(6):675-693.

Sherlock ME, Malkowski SN & Breaker RR (2017) Biochemical validation of a second guanidine riboswitch class in bacteria. *Biochemistry* 56(2):352-358.

Sherlock ME, Sadeeshkumar H & Breaker RR (2018a) Variant bacterial riboswitches associated with nucleotide hydrolase genes sense nucleoside diphosphates. *Biochemistry* 58(5):401-410.

Sherlock ME, Sudarsan N & Breaker RR (2018b) Riboswitches for the alarmone ppGpp expand the collection of RNA-based signaling systems. *Proceedings of the National Academy of Sciences* 115(23):6052-6057.

Sherlock ME, Sudarsan N, Stav S & Breaker RR (2018c) Tandem riboswitches form a natural Boolean logic gate to control purine metabolism in bacteria. *Elife* 7:e33908.

Shi Y, Tyson GW & DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* 459(7244):266-269.

Siegfried NA, Busan S, Rice GM, Nelson JA & Weeks KM (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods* 11(9):959-965.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M & Söding J (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology* 7(1):539.

Silva IJ, Barahona S, Eyraud A, Lalaouna D, Figueroa-Bossi N, Massé E & Arraiano CM (2019) SraL sRNA interaction regulates the terminator by preventing premature transcription termination of *rho* mRNA. *Proceedings of the National Academy of Sciences* 116(8):3042-3051.

Sinha D & De Lay NR (2022) Target recognition by RNase E RNA-binding domain AR2 drives sRNA decay in the absence of PNPase. *Proceedings of the National Academy of Sciences* 119(48):e2208022119.

Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC & Vogel J (2008) Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS genetics* 4(8):e1000163.

Šmejkalová H, Erb TJ & Fuchs G (2010) Methanol assimilation in *Methylobacterium extorquens* AM1: demonstration of all enzymes and their regulation. *PLoS One* 5(10):e13001.

Smirnov A, Förstner KU, Holmqvist E, Otto A, Günster R, Becher D, Reinhardt R & Vogel J (2016) Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. *Proceedings of the National Academy of Sciences* 113(41):11591-11596.

Smirnov A, Wang C, Drewry LL & Vogel J (2017) Molecular mechanism of mRNA repression in trans by a ProQ-dependent small RNA. *The EMBO journal* 36(8):1029-1045.

Smith KD, Lipchock SV, Ames TD, Wang J, Breaker RR & Strobel SA (2009) Structural basis of ligand binding by a c-di-GMP riboswitch. *Nature structural & molecular biology* 16(12):1218-1223.

Sonntag F, Kroner C, Lubuta P, Peyraud R, Horst A, Buchhaupt M & Schrader J (2015) Engineering *Methylobacterium extorquens* for de novo synthesis of the sesquiterpenoid α-humulene from methanol. *Metab. Eng.* 32:82-94.

Speed MC, Burkhart BW, Picking JW & Santangelo TJ (2018) An archaeal fluoride-responsive riboswitch provides an inducible expression system for hyperthermophiles. *Applied and environmental microbiology* 84(7):e02306-02317.

Staley JP & Guthrie C (1998) Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* 92(3):315-326.

Stav S, Atilho RM, Arachchilage GM, Nguyen G, Higgs G & Breaker RR (2019) Genome-wide discovery of structured noncoding RNAs in bacteria. *BMC microbiology* 19(1):1-18.

Storz G, Vogel J & Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell* 43(6):880-891.

Strauss B, Nierth A, Singer M & Jäschke A (2012) Direct structural analysis of modified RNA by fluorescent in-line probing. *Nucleic acids research* 40(2):861-870.

Sudarsan N, Hammond MC, Block KF, Welz R, Barrick JE, Roth A & Breaker RR (2006) Tandem riboswitch architectures exhibit complex gene control functions. *Science* 314(5797):300-304.

Sudarsan N, Lee E, Weinberg Z, Moy R, Kim J, Link K & Breaker R (2008) Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science* 321(5887):411-413.

Sudarsan N, Wickiser JK, Nakamura S, Ebert MS & Breaker RR (2003) An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes & development* 17(21):2688-2697.

Sulthana S, Basturea GN & Deutscher MP (2016) Elucidation of pathways of ribosomal RNA degradation: an essential role for RNase E. *Rna* 22(8):1163-1171.

Sun X, Zhulin I & Wartell RM (2002) Predicted structure and phyletic distribution of the RNA-binding protein Hfq. *Nucleic acids research* 30(17):3662-3671.

Szymanski M, Barciszewska MZ, Erdmann VA & Barciszewski J (2002) 5S ribosomal RNA database. *Nucleic Acids Research* 30(1):176-178.

Tafer H & Hofacker IL (2008) RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics* 24(22):2657-2663.

Takahashi M, Wu X, Ho M, Chomchan P, Rossi JJ, Burnett JC & Zhou J (2016) High throughput sequencing analysis of RNA libraries reveals the influences of initial library and PCR methods on SELEX efficiency. *Scientific reports* 6(1):1-14.

Tapsin S, Sun M, Shen Y, Zhang H, Lim XN, Susanto TT, Yang SL, Zeng GS, Lee J & Lezhava A (2018) Genome-wide identification of natural RNA aptamers in prokaryotes and eukaryotes. *Nature communications* 9(1):1289.

Tattersall J, Rao GV, Runac J, Hackstadt T, Grieshaber SS & Grieshaber NA (2012) Translation inhibition of the developmental cycle protein HctA by the small RNA IhtA is conserved across *Chlamydia*. *PloS one* 7(10):e47439.

Taylor-Robinson D (1994) *Chlamydia trachomatis* and sexually transmitted disease. *BMJ* 308(6922):150-151.

Tétart F & Bouché JP (1992) Regulation of the expression of the cell-cycle gene ftsZ by DicF antisense RNA. Division does not require a fixed number of FtsZ molecules. *Molecular microbiology* 6(5):615-620.

Thomas PS (1983) [18] Hybridization of denatured RNA transferred or dotted to nitrocellulose paper. *Methods in enzymology,* Elsevier, Vol 100. p 255-266.

Thomason MK & Storz G (2010) Bacterial antisense RNAs: how many are there, and what are they doing? *Annual review of genetics* 44:167-188.

Thompson DK & Sharkady SM (2020) Expanding spectrum of opportunistic *Cedecea* infections: Current clinical status and multidrug resistance. *International Journal of Infectious Diseases* 100:461-469.

Thompson JD, Gibson TJ & Higgins DG (2003) Multiple sequence alignment using ClustalW and ClustalX. *Current protocols in bioinformatics* (1):2.3. 1-2.3. 22.

Tian S & Das R (2017) Primerize-2D: automated primer design for RNA multidimensional chemical mapping. *Bioinformatics* 33(9):1405-1406.

Tian S, Yesselman JD, Cordero P & Das R (2015) Primerize: automated primer assembly for transcribing non-coding RNA domains. *Nucleic acids research* 43(W1):W522-W526.

Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E & Rosenow C (2002) Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic acids research* 30(17):3732-3738.

Toffano-Nioche C, Luo Y, Kuchly C, Wallon C, Steinbach D, Zytnicki M, Jacq A & Gautheret D (2013) Detection of non-coding RNA in bacteria and archaea using the DETR'PROK Galaxy pipeline. *Methods* 63(1):60-65.

Trausch JJ, Ceres P, Reyes FE & Batey RT (2011) The structure of a tetrahydrofolate-sensing riboswitch reveals two ligand binding sites in a single aptamer. *Structure* 19(10):1413-1423.

Tuerk C & Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *science* 249(4968):505-510.

Urban JH & Vogel J (2008) Two seemingly homologous noncoding RNAs act hierarchically to activate *glmS* mRNA translation. *PLoS biology* 6(3):e64.

Van Assche E, Van Puyvelde S, Vanderleyden J & Steenackers HP (2015) RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in microbiology* 6:141.

Van Dien SJ, Marx CJ, O'Brien BN & Lidstrom ME (2003a) Genetic characterization of the carotenoid biosynthetic pathway in *Methylobacterium extorquens* AM1 and isolation of a colorless mutant. *Applied and Environmental Microbiology* 69(12):7563-7566.

Van Dien SJ, Okubo Y, Hough MT, Korotkova N, Taitano T & Lidstrom ME (2003b) Reconstruction of C3 and C4 metabolism in *Methylobacterium extorquens* AM1 using transposon mutagenesis. *Microbiology* 149(3):601-609.

Vanderpool CK (2007) Physiological consequences of small RNA-mediated regulation of glucose-phosphate stress. *Current opinion in microbiology* 10(2):146-151.

Vanderpool CK & Gottesman S (2004) Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Molecular microbiology* 54(4):1076-1089.

Vanderpool CK & Gottesman S (2007) The novel transcription factor SgrR coordinates the response to glucose-phosphate stress. *Journal of bacteriology* 189(6):2238-2248.

Vicens Q, Mondragón E & Batey RT (2011) Molecular sensing by the aptamer domain of the FMN riboswitch: a general model for ligand binding by conformational selection. *Nucleic acids research* 39(19):8586-8598.

Vicente MM, Chaves-Ferreira M, Jorge JM, Proença JT & Barreto VM (2021) The Off-Targets of Clustered Regularly Interspaced Short Palindromic Repeats Gene Editing. *Frontiers in Cell and Developmental Biology* :2392.

Vitreschak AG, Rodionov DA, Mironov AA & Gelfand MS (2002) Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic acids research* 30(14):3141-3151.

Vitreschak AG, Rodionov DA, Mironov AA & Gelfand MS (2003) Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *Rna* 9(9):1084-1097.

Vogel J, Bartels V, Tang TH, Churakov G, Slagter-Jäger JG, Hüttenhofer A & Wagner EGH (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic acids research* 31(22):6435-6443.

Vogel J & Luisi BF (2011) Hfq and its constellation of RNA. *Nature Reviews Microbiology* 9(8):578-589.

Vogel J & Sharma CM (2005) How to find small non-coding RNAs in bacteria.

Vuilleumier S, Chistoserdova L, Lee M-C, Bringel F, Lajus A, Zhou Y, Gourion B, Barbe V, Chang J & Cruveiller S (2009) Methylobacterium genome sequences: a reference blueprint to investigate microbial metabolism of C1 compounds from natural and industrial sources. *PloS one* 4(5):e5584.

Wagner EGH & Romby P (2015) Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Advances in genetics* 90:133-208.

Wang JX, Lee ER, Morales DR, Lim J & Breaker RR (2008) Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. *Molecular cell* 29(6):691-702.

Washietl S, Findeiß S, Müller SA, Kalkhof S, Von Bergen M, Hofacker IL, Stadler PF & Goldman N (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *Rna* 17(4):578-594.

Wassarman KM (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes & Development* 15(13):1637-1651.

Wassarman KM, Repoila F, Rosenow C, Storz G & Gottesman S (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes & development* 15(13):1637-1651.

Waters LS & Storz G (2009) Regulatory RNAs in bacteria. *Cell* 136(4):615-628.

Watters KE, Strobel EJ, Angela MY, Lis JT & Lucks JB (2016) Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature structural & molecular biology* 23(12):1124-1131.

Wei BL, Brun-Zinkernagel AM, Simecka JW, Prüß BM, Babitzke P & Romeo T (2001) Positive regulation of motility and flhDC expression by the RNA-binding protein CsrA of *Escherichia coli*. *Molecular microbiology* 40(1):245-256.

Weilbacher T, Suzuki K, Dubey AK, Wang X, Gudapaty S, Morozov I, Baker CS, Georgellis D, Babitzke P & Romeo T (2003) A novel sRNA component of the carbon storage regulatory system of *Escherichia coli*. *Molecular microbiology* 48(3):657-670.

Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF & Sudarsan N (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic acids research* 35(14):4809-4819.

Weinberg Z & Breaker RR (2011) R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC bioinformatics* 12(1):1-9.

Weinberg Z, Kim PB, Chen TH, Li S, Harris KA, Lünse CE & Breaker RR (2015) New classes of self-cleaving ribozymes revealed by comparative genomics analysis. *Nature chemical biology* 11(8):606-610.

Weinberg Z, Lünse CE, Corbino KA, Ames TD, Nelson JW, Roth A, Perkins KR, Sherlock ME & Breaker RR (2017a) Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. *Nucleic acids research* 45(18):10811-10823.

Weinberg Z, Nelson JW, Lünse CE, Sherlock ME & Breaker RR (2017b) Bioinformatic analysis of riboswitch structures uncovers variant classes with altered ligand specificity. *Proceedings of the National Academy of Sciences* 114(11):E2077-E2085.

Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH & Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome biology* 11(3):1-17.

Weinstock GM, Hardham JM, McLeod MP, Sodergren EJ & Norris SJ (1998) The genome of *Treponema pallidum*: new light on the agent of syphilis. *FEMS Microbiology Reviews* 22(4):323-332.

Wen Y, Feng J & Sachs G (2013*) Helicobacter pylori* 5′ ureB-sRNA, a cis-encoded antisense small RNA, negatively regulates ureAB expression by transcription termination. *Journal of bacteriology* 195(3):444-452.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R & Federhen S (2007) Database resources of the national center for biotechnology information. *Nucleic acids research* 36(suppl_1):D13-D21.

White N, Sadeeshkumar H, Sun A, Sudarsan N & Breaker RR (2022) Na+ riboswitches regulate genes for diverse physiological processes in bacteria. *Nature Chemical Biology* :1-8.

Wickham H & Sievert C (2016) *Ggplot2 : elegant graphics for data analysis.* Springer International Publishing, Cham, Second edition.

Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS & Griffiths AD (2006) Amplification of complex gene libraries by emulsion PCR. *Nature methods* 3(7):545-550.

Wilms I, Overlöper A, Nowrousian M, Sharma CM & Narberhaus F (2012) Deep sequencing uncovers numerous small RNAs on all four replicons of the plant pathogen *Agrobacterium tumefaciens. RNA biology* 9(4):446-457.

Winkler ME & Ramos-Montañez S (2009) Biosynthesis of histidine. *EcoSal Plus* 3(2).

Winkler W, Nahvi A & Breaker RR (2002a) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 419(6910):952-956.

Winkler WC, Cohen-Chalamish S & Breaker RR (2002b) An mRNA structure that controls gene expression by binding FMN. *Proceedings of the National Academy of Sciences* 99(25):15908-15913.

Winkler WC, Nahvi A, Roth A, Collins JA & Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* 428(6980):281-286.

Winkler WC, Nahvi A, Sudarsan N, Barrick JE & Breaker RR (2003) An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nature Structural & Molecular Biology* 10(9):701-707.

Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, Kleinkauf R, Hess WR & Backofen R (2014) CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic acids research* 42(W1):W119-W123.

Wunsch CM & Lewis JP (2015) Porphyromonas gingivalis as a model organism for assessing interaction of anaerobic Bacteria with host cells. *JoVE (Journal of Visualized Experiments)* (106):e53408.

Xu J & Cotruvo Jr JA (2022) Reconsidering the czcD (NiCo) Riboswitch as an Iron Riboswitch. *ACS Bio & Med Chem Au*.

Yakhnin H, Pandit P, Petty TJ, Baker CS, Romeo T & Babitzke P (2007) CsrA of *Bacillus subtilis* regulates translation initiation of the gene encoding the flagellin protein (hag) by blocking ribosome binding. *Molecular microbiology* 64(6):1605-1620.

Yang S, Peng Q, Zhang Q, Yi X, Choi CJ, Reedy RM, Charkowski AO & Yang C-H (2008) Dynamic regulation of GacA in type III secretion, pectinase gene expression, pellicle formation, and pathogenicity of *Dickeya dadantii* (*Erwinia chrysanthemi* 3937). *Molecular plant-microbe interactions* 21(1):133-142.

Yang Y-M, Chen W-J, Yang J, Zhou Y-M, Hu B, Zhang M, Zhu L-P, Wang G-Y & Yang S (2017) Production of 3-hydroxypropionic acid in engineered *Methylobacterium extorquens* AM1 and its reassimilation through a reductive route. *Microbial cell factories* 16(1):1-17.

Yang Y, Yue Y, Song N, Li C, Yuan Z, Wang Y, Ma Y, Li H, Zhang F & Wang W (2020) The YdiU domain modulates bacterial stress signaling through Mn2+-dependent UMPylation. *Cell Reports* 32(12):108161.

Yu D & Breaker RR (2020) A bacterial riboswitch class senses xanthine and uric acid to regulate genes associated with purine oxidation. *RNA* 26(8):960-968.

Yu S-H, Vogel J & Förstner KU (2018) ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes. *Gigascience* 7(9):giy096.

Zeller MJ, Favorov O, Li K, Nuthanakanti A, Hussein D, Michaud A, Lafontaine DA, Busan S, Serganov A & Aubé J (2022) SHAPE-enabled fragment-based ligand discovery for RNA. *Proceedings of the National Academy of Sciences* 119(20):e2122660119.

Zhang A, Wassarman KM, Rosenow C, Tjaden BC, Storz G & Gottesman S (2003) Global analysis of small RNA and mRNA targets of Hfq. *Molecular microbiology* 50(4):1111-1124.

Zhang Q, Zhang Y, Zhang X, Zhan L, Zhao X, Xu S, Sheng X & Huang X (2015) The novel cis-encoded antisense RNA AsrC positively regulates the expression of rpoE-rseABC operon and thus enhances the motility of *Salmonella enterica* serovar typhi. *Frontiers in microbiology* 6:990.

Zhang S, Liu S, Wu N, Yuan Y, Zhang W & Zhang Y (2018) Small Non-coding RNA RyhB mediates persistence to multiple antibiotics and stresses in uropathogenic *Escherichia coli* by reducing cellular metabolism. *Frontiers in microbiology* 9:136.

Zhu C, Yang G, Ghulam M, Li L & Qu F (2019) Evolution of multi-functional capillary electrophoresis for high-efficiency selection of aptamers. *Biotechnology advances* 37(8):107432.

Zhu LP, Song SZ & Yang S (2021) Gene repression using synthetic small regulatory RNA in *Methylorubrum extorquens*. *Journal of Applied Microbiology* 131(6):2861-2875.

Zhu W-L, Cui J-Y, Cui L-Y, Liang W-F, Yang S, Zhang C & Xing X-H (2016) Bioconversion of methanol to value-added mevalonate by engineered *Methylobacterium extorquens* AM1 containing an optimized mevalonate pathway. *Applied microbiology and biotechnology* 100(5):2171-2182.

Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS & Rouskin S (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature methods* 14(1):75-82.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* 31(13):3406-3415.

Zundel MA, Basturea GN & Deutscher MP (2009) Initiation of ribosome degradation during starvation in *Escherichia coli*. *Rna* 15(5):977-983.

8 ANNEXE I: SUPPLEMENTARY MATERIAL: SMALL RNAS BEYOND MODEL ORGANISMS: HAVE WE ONLY SCRATCHED THE SURFACE?

8.1.1 Antisense RNA

Antisense RNAs play an essential role in genetic regulation and are associated in the tight regulation of transposases, toxic proteins, transcription regulators and virulence proteins to name a few (reviewed in (Thomason & Storz, 2010)). They bind with their targeted mRNA with perfect complementarity since they are encoded in the opposite strand. Their dimension varies greatly, from a few hundred nucleotides (100 to 300 nt) to larger sizes (700 to 3500 nt) (Georg & Hess, 2011). The formation of an asRNA-mRNA complex could impact the secondary structure of both RNAs, leading to a change in their stability and perhaps degradation (Dühring *et al.*, 2006). The binding of asRNA to its target could prevent the ribosome from reaching the RBS. Antisense sRNA could also impact the mRNA in the opposite strand due to transcription interference without directly binding with one another. Divergently transcribing promoters could interfere with each other, resulting in the collision of RNA polymerase complexes for example (Georg & Hess, 2011). An elongating RNA polymerase (RNAP) on the antisense strand could impede the formation of an initiation complex on the sense strand through transcription occlusion or dislodged an already formed one through sitting duck interference (Georg & Hess, 2011).

Information about asRNAs comes from RiboGap (Naghdi *et al.*, 2017), which extracts data from Rfam to facilitate the analysis of non-coding RNA (Kalvari *et al.*, 2021). Only asRNAs with an E-value lower than 0.0005 were taken into consideration. Rfam is a database on RNA families bases on secondary structures and covariance model. Since asRNAs do not rely on secondary structures, they may be underrepresented in Rfam. Nevertheless, it still gives us a good estimation of the extent of knowledge of asRNAs in bacteria. Moreover, we are limited by the available annotations in Rfam. For example, the sRNA MicF is known to be encoded in a different locus than its target, the outer membrane protein OmpF in *Escherichia coli* (Delihas & Forst, 2001). It would therefore meet criteria to be classified as a trans-acting sRNA rather than an asRNA. However, early research did not make the same distinction between asRNA and sRNA, so it was classified as an asRNA in Rfam, a categorization which remains. It would be tedious to go through the list of all asRNAs family in Rfam to verify their classification, and we are confident that it would not change the conclusion of this perspective article.

8.1.1.1 Prevalence of asRNAs in bacteria

Forty distinct asRNAs were annotated in bacterial genomes based on the Rfam database. Like for sRNAs, the phyla Proteobacteria and those from the Terrabacteria group also encode for the most distinct asRNAs (Table 8.1) with 29 and 17 respectively. Interestingly, the proportion of asRNAs in the two most studied phyla relative to the sum of all phyla from (Table 8.1) is similar to that of sRNAs from Table 3.1 (60% vs. 58% for Proteobacteria and 35% vs. 35% for Terrabacteria), even though asRNAs act through very different mechanisms, also suggesting that these numbers strongly correlate with the “intensity of research” within these phyla.

Table 8.1 Number of distinct annotated asRNAs encoded in different phyla

Phylum group	asRNAs
FCB group [1]	2
Proteobacteria	29
Terrabacteria group	17

[1] FCB group stands for Fibrobacteres, Chlorobi, and Bacteroidetes, whereas [2] PVC group represents Planctomycetes, Verrucomicrobia, and Chlamydiae.

Genus that encodes for the most distinct asRNAs are all from the family *Enterobacteriaceae*, apart for *Serratia marcescens* that is a *Yersiniaceae* (Figure 8.1, A). We also in examined the top 10 most annotated asRNAs in all bacteria (Figure 8.1, B).

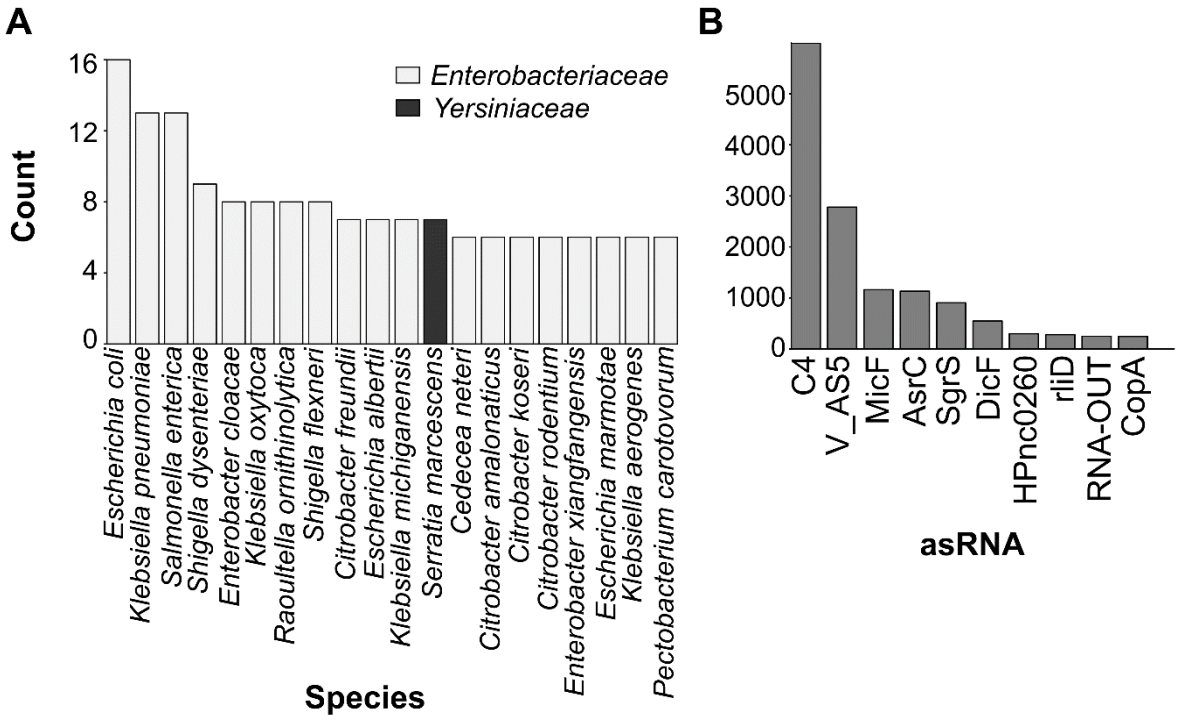

Figure 8.1 Prevalence of asRNAs in bacterial genomes

(A) Top 20 species that encodes for the most distinct asRNAs. All strains from the same species are considered. **(B)** Top 10 asRNAs that are most found within bacterial genomes. Each individual occurrence of an asRNA was taken into consideration. Only asRNAs with E-value lower than 0.0005 were kept.

Some of the genera that encode for the highest number of distinct asRNAs were also those that contain the most sRNAs. However, three genera were not discussed before: *Raoultella*, *Serratia* and *Cedecea* (Table 8.2, in bold) and they are all human pathogens. *Cedecea neteri* for example was isolated at the CDC (Centers for Disease Control and Prevention), from where its name originates. Even if the incidence of Cedecea infections is infrequent, increasing occurrences and its antibiotic resistance warrant more research interest. It is considered an opportunistic pathogen, since it is isolated in immunocompromised patients (Thompson & Sharkady, 2020).

Table 8.2. Description of genus encoding for the most distinct asRNAs.

Genus	Nb of distinct asRNAs[1]	Description	Ref
Escherichia	16	Most well-understood bacteria	(Blount, 2015)
Klebsiella	13	Nosocomial pathogen, model organism to study drug resistance	(Bi *et al.*, 2015)
Salmonella	13	Model organism to study host-pathogen interactions	(Garai *et al.*, 2012)
Shigella	11	Causative pathogen of shigellosis	(Killackey *et al.*, 2016)
Enterobacter	9	Responsible for nosocomial infections	(Sanders Jr & Sanders, 1997)
Citrobacter	9	Third most common urinary pathogen	(Ranjan & Ranjan, 2013)
Raoultella	8	Associated with histamine poisoning in human	(Hajjar *et al.*, 2020)
Serratia	8	Opportunistic nosocomial pathogen	(Khanna *et al.*, 2013)
Cedecea	6	Rare pathogen associated with urinary tract infections; antibiotic resistance	(Ahmad *et al.*, 2021)

[1]The number represents the quantity of distinct sRNAs in all bacterial strains within this genus.

The most annotated asRNAs in bacterial genomes were also discovered in the model organism *E. coli (*MicF (Delihas & Forst, 2001), SgrS (Aiba, 2007; Vanderpool, 2007; Vanderpool & Gottesman, 2007), DicF (Bouché & Bouché, 1989; Faubladier & Bouché, 1994; Murashko & Lin-Chao, 2017; Tétart & Bouché, 1992) and RNA-OUT (Kittle *et al.*, 1989)), in pathogens (V_AS5 (Liu *et al.*, 2009), AsrC (Zhang *et al.*, 2015) ,HPnc0260 (Sharma *et al.*, 2010), rliD (Mandin *et al.*, 2007) and CopA (Gerhart *et al.*, 1986; Jiang *et al.*, 2017; Light & Molin, 1983; Nordgren *et al.*, 2001)) or as part of computational homology searches (C4 (Weinberg *et al.*, 2010)) (Table 8.3).

Table 8.3. Description of top 10 most prevalent asRNAs in bacteria

asRNA	Description	RFAM	asRNA expression	Discovered in	Ref
C4	C4 antisense RNA	RF01695	-	Proteobacteria, Phages	(Weinberg *et al.*, 2010)
V_AS5	*Vibrio* RNA AS5	RF02818	-	*Vibrio cholerae*	(Liu *et al.*, 2009)
MicF	-	RF00033	Regulates outer membrane protein OmpF	*Escherichia coli*	(Delihas & Forst, 2001)
AsrC	antisense RNA of *rseC* mRNA	RF02746	Target *rseC*; promote bacterial motility	*Salmonella enterica* serovar typhi	(Zhang *et al.*, 2015)
SgrS	-	RF00534	Coordinate response to glucose-phosphate stress	*Escherichia coli*	(Aiba, 2007; Vanderpool, 2007; Vanderpool & Gottesman, 2007)
DicF	-	RF00039	Inhibitor of gene *ftsZ* involved in cell division	*Escherichia coli*	(Bouché & Bouché, 1989; Faubladier & Bouché, 1994; Murashko & Lin-Chao, 2017; Tétart & Bouché, 1992)
HPnc0260	Bacterial antisense RNA HPnc0260	RF02194	-	*Helicobacter pylori*	(Sharma *et al.*, 2010)
rliD	*Listeria* sRNA *rliD*	RF01494	Antisense of the gene *pnpA,* a Polynucleotide phosphorylase	*Listeria monocytogenes*	(Mandin *et al.*, 2007)
RNA-OUT	-	RF00240	Tn10/IS10 antisense system	*Escherichia coli*	(Kittle *et al.*, 1989)
CopA	CopA-like RNA	RF00042	Regulate copy number of plasmid R1	Plasmid R1 (first isolated from *Salmonella* sp. *[34]*	(Gerhart *et al.*, 1986; Jiang *et al.*, 2017; Light & Molin, 1983; Nordgren *et al.*, 2001)

As demonstrated, most of the knowledge we have for asRNAs comes from study on research-intensive pathogens and model organisms. The most prevalent asRNAs are also found in closely related species of the same order, almost exclusively from the *Enterobacteriaceae* family, apart from *Yersiniaceae*. By extending our research to other bacteria, we could improve our understanding of the role of asRNAs in genetic regulation.

8.1.2 Materials and Methods

8.1.2.1 RiboGap

Information about sRNA, coding sequence and annotations was extracted from RiboGap (Naghdi *et al.*, 2017). This database is accessible via a web interface: http://ribogap.iaf.inrs.ca/ribo_gap_advanced_version_ribogap_v2.pl (version 2). Queries can be selected with a user-friendly interface or be typed in SQL directly in the appropriate box. Queries

used for this article are found in Table 8.4. All genetic information (RNAs and genes) compiled for each species are naturally found in their genome.

Table 8.4. RiboGap queries

	Query
sRNAs annotated in bacteria	select distinct fragment.taxonomy,fragment.description as description_of_fragment,rna_family.fam_id,rna_family.fam_name,rna_family.description as description_of_rna_family,rna_family.type,rna_known.evalue from fragment inner join rna_known on fragment.fragment = rna_known.fragment inner join rna_family on rna_known.fam_id = rna_family.fam_id where rna_family.description Like '%sRNA%' OR rna_family.type Like '%sRNA%' LIMIT 0, 50
Bacterial genome size and number of annotated genes	select distinct fragment.fragment,fragment.length,fragment.chromosome,fragment.gene_num,fragment.description from fragment where fragment.chromosome Like '%chromosome%' LIMIT 0, 50
Number of annotated RNA	select * from fragment inner join rna_known on fragment.fragment=rna_known.fragment where fragment.chromosome like '%chromosome%' and fam_id like '%RF%';
Number of annotated tRNA	select * from fragment inner join rna_known on fragment.fragment=rna_known.fragment where fragment.chromosome like '%chromosome%' and fam_id like '%tRNA%';
AsRNAs annotated in bacteria	select distinct fragment.taxonomy,fragment.description as description_of_fragment,gap5.accession,rna_family.fam_id,rna_family.fam_name,rna_family.description as description_of_rna_family,rna_family.type,rna_known.evalue from fragment inner join gap5 on fragment.fragment = gap5.fragment inner join rna_gap5 on gap5.num_cle = rna_gap5.num_cle inner join rna_known on rna_gap5.rna_id = rna_known.rna_id inner join rna_family on rna_known.fam_id = rna_family.fam_id where rna_family.type Like '%antisense%' OR rna_family.type Like '%asRNA%' LIMIT 0, 50
Human pathogenic bacteria	select distinct fragment.fragment,fragment.description,organism.pathogenic_in from organism INNER JOIN fragment on organism.organism_id=fragment.organism_id where organism.pathogenic_in Like '%human%' LIMIT 0, 50

Only RNAs with an E-value lower than 0.0005 were kept for this article. We made similar queries that would consider E-values up to 100, which increased the number of sRNAs by ~20% in some well-studied classes and almost doubled number of sRNAs in less studied classes, which does not fundamentally change our conclusions (even if it had significantly changed our figures), but would, however, have resulted in a much less reliable set of data to prepare figures and tables presented in this article. Size of fragments (chromosome or plasmid) is not available yet on RiboGap. This information was therefore extracted from all available Genbank files in the FTP of NCBI (Sayers et al., 2010).

8.1.2.2 Graphical representation

Graphic representations were created with the ggplot2 package (Wickham, 2016) within Jupyter notebook (Kluyver *et al.*, 2016).

8.1.3 Supplementary Figures

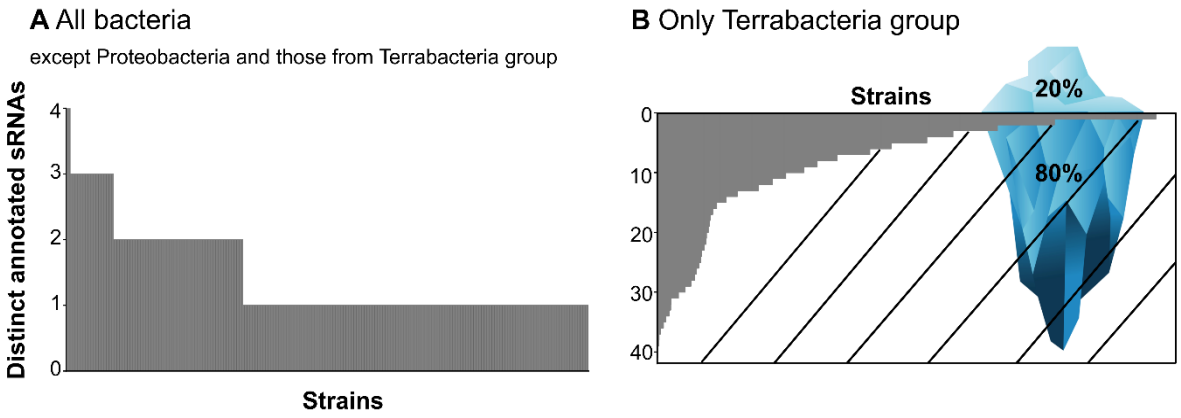

Figure 8.2 Number of distinct annotated sRNAs

In (**A)** all bacteria except Proteobacteria and those from the Terrabacteria group and in **(B)** only bacteria from the latter. The iceberg is a representation of the number of sRNAs that we could be missing in the Terrabacteria group, where the above water portion of the iceberg portrays the already known sRNAs (gray section), and the underwater section depicts what could be left to be discovered (hatched section) if all bacteria contained as many sRNAs as those with the highest number of them. Percentages also represent this ratio of what is known versus what could be left to discover. This figure represents a compilation of 398 and 1604 strains in **(A)** and **(B)** respectively.

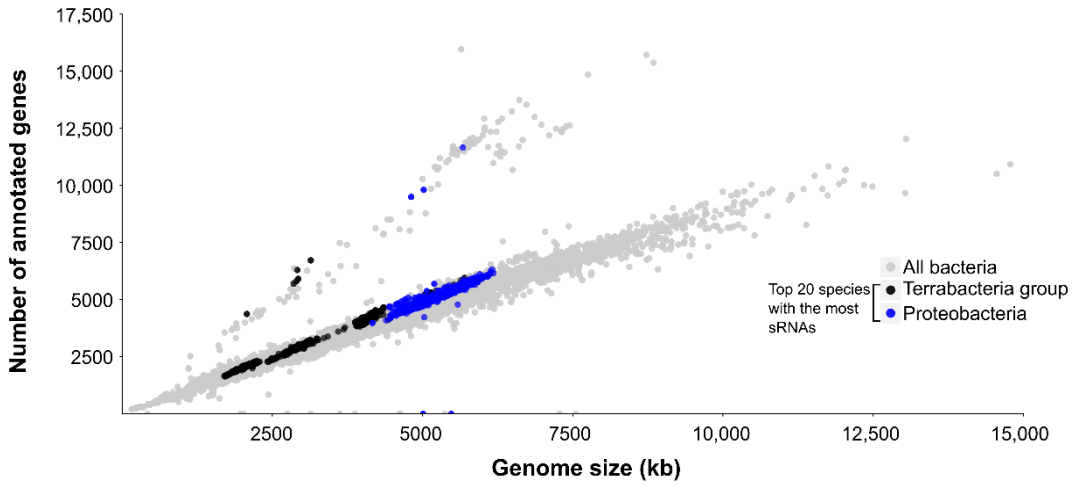

Figure 8.3 Number of annotated genes compared to genome size.

The top 20 species containing the most annotated sRNAs from Terrabacteria group and Proteobacteria are emphasized with black and blue dots respectively. This figure also includes outliers from bacterial strains with no annotations available in NCBI (Sayers *et al.*, 2010) and some mislabeled as complete genomes.

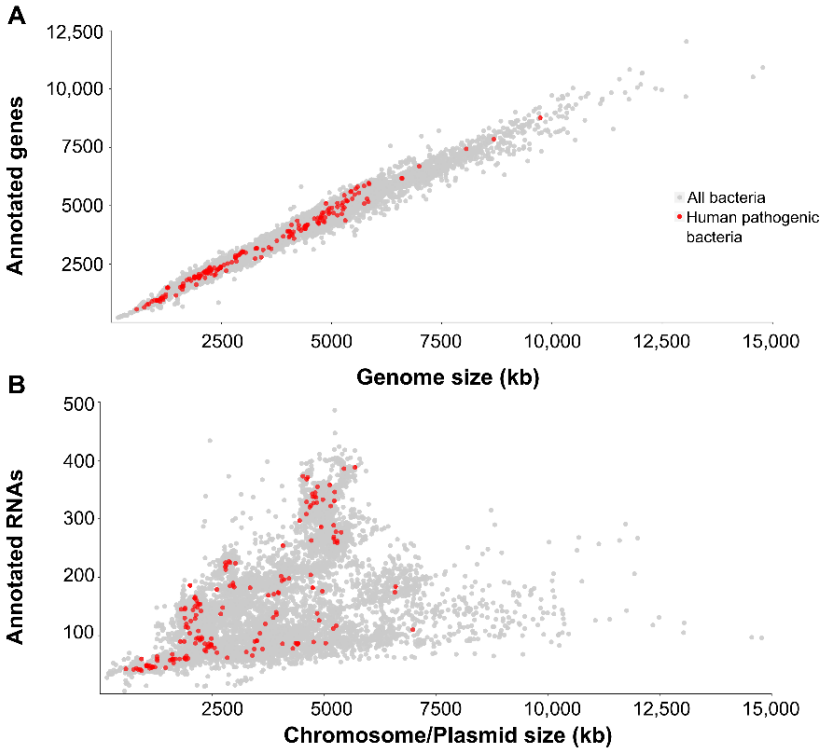

Figure 8.4 Number of annotated genes and RNA, where human pathogenic bacteria are emphasized in red.

(A) Number of annotated genes compared to genome size (all chromosomes and plasmids for each bacterial strain are considered when applicable). **(B)** Number of annotated RNAs compared to fragment size (chromosomes or plasmids). RNAs include CRISPR RNAs, antisense RNAs, sRNAs, tRNAs, long non-coding RNAs, ribozymes and cis-regulatory elements. Bacteria were considered as human pathogens when they were labeled as such within the RiboGap database (Naghdi et al., 2017) and are derived from former tables that NCBI does not update anymore and thus do not include all pathogens (it includes 217 pathogens from a sample of 1023 bacteria and can thus still be considered a substantial sample).

From the analysis shown in Figure 8.4, it appears clear that even if most of the species with the highest number of annotated sRNAs (and ncRNAs in general) are pathogens, even among human pathogens there are still numerous bacteria that are understudied from the point of view of their ncRNAs.

9 ANNEXE II: ANALYSIS OF NON-CODING RNAS IN *METHYLORUBRUM EXTORQUENS* REVEALS NOVEL SMALL RNAS SPECIFIC TO *METHYLOBACTERIACEAE* (SUPPLEMENTARY MATERIAL)

9.1 Supplementary Tables

Table 9.1 Probes for candidates tested by Nothern blot analysis

ID	Start	End	Size	Strand	Probe
426	51119	51301	182	+	GGCAGCCCAGCCTGCTGCTCCTTCAGGATGCCGATGATCTGCTCTTCGCT
432	80737	80895	158	+	CGGGCTCACGGTCTTTGGCCGTGAACCCTAAAAGTGCGGCGATATGAACC
609	636257	636362	105	+	GAAGGATCCTCCAGGGATCGCGCGGGATCTGGACGATCCTTCGTGGCCGC
1113	2218259	2218369	110	+	TGTGCTGATGTTTCCGCTTCTGCGGATCACCGTCACGAAAAACCCGAGCGG
1153	2388747	2388799	52	+	TCTGGAGCCCTCCTTCGAGGCCTCCGCTTCGCTCCGGCACCTCAGGATGA
1175	2436125	2436237	112	+	AGCTCGTCAGGCTCATAACCTGAAGGTCGCTGGTTCAAATCCAGCCCCCG
1348	3229332	3229428	96	+	AAGCTCGGCCGTATGTGAACCTTGGCCGTTCGGCAGCGTTTCCCGAATGG
1627	4198109	4198315	206	+	CACGCGGGGAGAGGGTCGCGACGACCACGG
1679	4343857	4343949	92	+	CGGAACTTGCAAACCATCTGCAAGCCGTAACCCCGACGCCGAAGGCATCA
1752	4567059	4567168	109	+	CGCGACGACCGGATGCGCCCTCGACGGGCGCTTCCGCCGCCGGCACGATC
1776	4678233	4678297	64	+	GCCTGCGAGCGCCCTGCCGGGCACTTCACGTGGGCCATCCGCCGGAACGG
1819	4874156	4874225	69	+	GTGGGATCGCGTTGGCCGCTATCCCGCGCT
1969	5367912	5368067	155	+	CGTCCGGTGTTCGTGGCGGGCCACTCAGTCCAAGACCGAACGGGTTCTCA
2036	51121	51275	154	-	AGACGGACGATGAAGAAGAGCCGGTTTAGCGAAGAGCAGATCATCGGCAT
2038	54339	54419	80	-	TCTGATCGTTGCCTCGCGTGCTAGACCCATGCGCGGGTCGCCGCAATGCG
2039	57158	57258	100	-	TAGCGCCGCCCGAGGCGCGCGGGCAAGAGCCCGTTCCGGCCACGCCGGTA
2043	80757	80895	138	-	AATGCCTTCGGGCTCACGGTCTTTGGCCGTGAACCCTAAAAGTGCGGCGA
2185	636269	636362	93	-	AGCGGCCACGAAGGATCGTCCAGATCCCGCGCGATCCCTGGAGGATCCTTC
2624	2125529	2125634	105	-	AACGGCGCAGAGAACTGGTCCGCACCGCCCCGTCACTGCCGGCCCCGCCA
2674	2388747	2388799	52	-	CCTCATCCTGAGGTGCCGGAGCGAAGCGGAGGCCTCGAAGGAGGGCTCCA
3211	4753719	4753781	62	-	GACACCTTCAACTAGAAGGCGTCGCCGGACTCGATCCGGTGAGAACG GG
3355	5367928	5368087	159	-	TGAGAACCCGTTCGGTCTTGGACTGAGTGGCCCGCCACGAACACCGGACG

Table 9.2 Intergenic regions containing Methylo2624 and Methylo1969

Methylo2624

> *Methylorubrum extorquens* DM4

cacggcctcgcatcaaagctcccgatcaccgatcgctgtccctcggggcacggccgcgtggcgggagttctgtgagaagcaggtcagccgagggccggcga
aggctggaacgtacatcggtggcgggggagcggcgtcacggcacggttgccgtcggcggcgcgggcctctacatcggccctccccgcggccgatgcgcctg
acggtttcgcgagccctctcgggcctgcgtggt**tggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtgacgctcagtgg
gccgagcggctgcaggtgcgtcgggagcgcggggaacaggaaccgcgaccccggcgcctatccgaattccctcgaccgcattccctaagccccagaccgg
cacgcatccggtccggggccgggcctgggtacagcgcgagaagcgttaacggtctttctcaagggaaagcgaaccggtggcaggcgatttatcgcagattgtc
gccgaaaggtcgcacc

> *Methylorubrum rhodesianum* strain DSM 5687

gcggcgctccgatcgttttggcaatgcattgccggtagggcagcggcgtcagggcgcggttgcccgcggcggcgcgggcctctacatcggccctccccgcggc
cgatgcgcccgatggtttcgcgagccctctcgggcctgcgtggc**ggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtg
acgctcagtgggccgagcggctgccggtgcgtcgggagcgcggggaacaggaaccgcgatcccggcgcaaggcccacctattcggctcccgtcaggccga
agaggtcacccggccctgtgccggagtgccgttaacggtctttctcaagggaaagcgaaccactcgcgggcaatatttcgccagttgtccaacagaggtcgctcc

> *Methylorubrum podarium* strain DSM 15083

gcggaagcgaacctgacgtcggaggagcggcgtcacggcgcggttgcccgcggcggcgcggacctctacatcggccctccccgcggccgatgcgcctgac
ggtttcgcgagccctctcgggcctgcgtggc**ggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtgacgctcagtgggc
cgagcggctgcaggtgcgtcggcagcgcggggaacaggaaccgcgatgtcggcgcatggcccaccgactcccgcaccccgatccgcggccgagcccgg
atcggggacggcggtccaggatgcgggaaacgttaacggtctttcgcaagggaaaacgaaccagcagcatgcgatccctcgacctgttgtcccagagaggtc
gcacc

> *Methylorubrum aminovorans* strain NBRC 15686

gccgcacgtcggccttgcgtcgccggaagggcagcggcgtcacggcgcggttgctcgcggcggcgcggacctctacatcggccctccccgcggccgatgcg
cctgatggtttcgcgagccctctcgggcctgcgtggc**ggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtgacgctcag
tgggccgagcggctgcaggtgtgtcggaagcgcggggaacaggaaccgcgatttcggcgcatggcccaccgattcccagcctcgccgatccgggctccag
cccggatcgggaacggtgttcccgggttggcgccagatccgttaacggtctttcgcaagggaaaaccaaccagacgcatgcaaccattttgcgctttgtctcaga
aaggttgcaca

> *Methylorubrum populi* BJ001

gcggcgccacggggcaaagcgtcgccgaaggggcagcggcgtcacggcgcggttgcccgcggcggcacggacctctacatcggccctccccgcggccga
tgcgcctgacggtttcgcgagccctctcgggcctgcgtggc**ggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtgacg
ctcagtgggccgagcggctgcaggtgcgtcgggagcgcggggaacaggaaccgcgaccccggcgcaaggcccacacgattcctgtgctctccgatccggc
cttaagcccggatcgggggctgccttccgggcgccgcgcgagaaccgttaagggtctttcgcaagggaaaacaaaccggctccatgcaaccttcgcacgttgt
ctcgaaaaggttgcaca

> *Methylorubrum zatmanii* strain LMG 6087

agacaacgcgggaccaagcgtcgcaaccgcgaagcagagccgctgcgcggttgcccagggcagcccggacctctacatcggcccctccccgcggccgat
gcgcccgatggtttcgcgagccctctcgggcctgcgtggc**ggcggggcctgcagtgacggggcggtgcggaccagttctctgcgccg**atcctgtgacgc
tcagtgggccgagcggctgcaggcgcgtcgggagcgcggggaacaggaaccgcgatcccggcgcaaggcccaccacgttttcgcgatccagagatcgtta
acggtctttctcaaacaaaaacgaacggagcagagggtagccgtcgagcctgaccgctcgaaggttgcaacc

> *Methylobacterium adhaesivum* strain DSM 17169

gtttcccgtcaacgatggccacgccgtgggacgatgtcctgcggcgtggcggccagttgccggcttccggtccgacccctacaccgatcctccctgctgccgatg
cgctcgatggtttcgcgaacccgctcgggctttgcatggcggcggggcctgtc**gtgacggggcggtgcggaccagttctctgct**ccgtt**cctgtgaggcctag**
tgggccgggtcgcggcaggtgtgtcacgggcgcggggaacaggaaccgtcgccgtggcgcaaggcccaccttctcgaaacgtttcaagatgagtcccggtg
acctgaaccgggatggtcacgcgccgcggatgggagcggtcctttaacgatcctgcgcgagagtcgtatcaggccggggaaccggttatgtgccaccgcgtag
cttcgggccagaccaca

> *Methylobacterium bullatum* strain DSM 21893

tgcagccgggttggttgccggctcaagtcaggatgtctacatccaccatccccgcagccgatgcgtccgatggtttcgcgaaccctctcggactggcgcggttgca
gggcttgtc**gtgacggggcggtgcggaccagttctctgcgtcg**atcctgtgaggcctcgtgggccgggtcgcggcaagtgtgtcacgggcgcggggaaca

ggaaccgtcgccgtggctgatggcccacgctcatcgtagtagcggatgggtcgcgccatttcaggacaggctccgggccggccagcggccaaatggtttaccg
acttctcgcacaagtgagacgggccggggaaccggtcgcgttcctccgcgtagtctcgggctttccagc

> *Methylorubrum thiocyanatum* strain JCM 10893

gcggcgccacggggcaaagcgtcgccgatggggcagcggcgtcacggcgcggttgcccgcggcggcacggacctctacatcggccctccccgcggccgat
gcgcctgacggtttcgcgagccctctcgggcctgcgtggc**ggcggggccggcagtgacggggcggtgcggaccagttctctgcgccgtt**cctgtgacgc
tcagtgggccgagcggctgcaggtgcgtcgggagcgcggggaacaggaaccgcgaccccggcgcaaggcccacacgattccttcgctctccgatccggcct
cgagcccggatcgggggctgccttccgggcgccgcgcgagaaccgttaagggtctttcgcaagggaaaacaaaccggctccatgcaacctttcgcacgttgtct
cgaaaaggttgcaca

Methylo1969

> *Methylorubrum extorquens* DM4
cgcgggggcgtccgaccggtccgagaccggacgtccaagacgtgtttgtccgcgggccggcccacctgccggggagcttcaggccggcccgcgggcccgtc
ctggcgccggattagcggcgaggacaggctccccaatctcaccgacgattgatgctgctttgtgtccgccgacggccgcgctccgccgtgatgggggcgtctgc
cgtcgcatgcggtgtcctcgtcggtgcgctgaggggtcttccccgcggggacgcgcagcgtcgggaaccgacctcggaggggtgctgccagcgctcgcgtcgt
atccggcgctgatgtcccgcccgatgcggacagccctttcgcatcgccggtcgtaagccgttgccgacgaacgattttactcatcaagcattcgtcagcacgaaa
atctgaaggcttggcggcaaattcccgcttgacggtctgccaatacggacgcagtctacattcgtgcttcgcgatcgcaagcacatcggccagtaaccgcgtggtc
tttgcgtggtttgttggtgaaagctggaggagacagggatgactccaccgcagcatgaagccgcaacgatagtc**tgagaacccgttcggtct**c**ggactgagt
ggcccgccacgaacaccggacg**gaagcggatccttagggcgcagaagcggccgacacggcacgacttcgaacacgcatcagccccacctgcggac
cgcttccgccgcggtttttattatcagtgccaaggttttcaccgtgccagccgccgaggctcgcgcggatttcagcgcgggctgtgctgttgaagcgaagccgttcgg
ctcgcaacgacggtgtgggttggccgctccgtcattgcgagacttcgtccgcgctctgtgcgcttccggccggcttacttgcccgacaacagcggcgaccaagtcg
ggtccgaggcga
> *Methylorubrum rhodesianum* strain DSM 5687 Ga0373205_13
gaccgggagcccggcgcctcgcgaagaggcacggagtgttttctcccaggacgcacaggactttatccctgggccgcagtaagccgttgccggcgaacgatttt
gccggtcagtaattcgtcgagataaaaaacagtcggcttggcagcagatttcttcttgacgacctgccaataccgacgcagtctctctgcgtgcttcgcattcgcaa
gcacatcggccaagaaaccgctggtccgcgcacggtttgttggcgaaagctggaggagacaggaatgactccaccgcagcgcaaagccgcgacgatagcc
tgagaacccgttcggtctc**ggactt****agtggcccgccacgaacaccggacg**gaagcggctccctaggggcgcagaagcggccgacacggcgcgactt
cgaacacgcatcagccccacctgcggaccgcttccgccgcggtttctattgcgagagagacgattgcaccgtgccagccgttgaggctcgcgcggatttcgtcat
tgctgcggagtcctagctcgacggaatgacagtccggccgaagtcggcagcgccggaacgcgtcgggggacggagattcccgctaccgtccccgtccgcatc
gtccgggaagcattcgtcaccggcctcgcggggcggggccggtcgtc

> *Methylorubrum podarium* strain DSM 15083 entry415

tcgaccgcgagccggcacgcctgctgggtgcgtcgtgccggcccgcgagcctgtccggttgccgggggggggagcggccaggacaggcaggacagtctcgc
cgacgattgatgctggtttgtgtcgatgcctcgacggcgcgctcgcgccggacctccgccgtcctgccgaaaaccgtcgttttcggcccgctgcacatcggacttttc
catcgtgcgccgtaagccgttgccggcgaacgattttgtcgatcaagcattcgctaggacgagaaacagtcggcttggcagcaaattcccgcttgacggcctgcc
aatacggacgcagtctctatccgtgcttcgcggtcgcaagcacatcggccaggaaaccgggtggtccgcgcacggtttgttggcgaaagctggaggagacagg
gatgactccaccgcagcgcaaagccgctacgatagcc**tgagaacccgttcggtctcggact****agcggcccgccacgaacaccggacg**gaagcggat
ccctaggggcgcagaagcggccgacacggcgcgacttcgaacacgcatcagccccacctctggaccgcttccgccgcggtttttatcgcgaaccagacgagtt
tctgccgtgccggccgtcgaggcccgcgcggatttcgtcgttggagcgcggccggaagccggccttcgcggcgcggtcgagtgcgctcggaggagagcgccc
ctcgacggctgcgcctccgaccgtcggcctacttccccgataggagcggcgaccaggtcgggtccgaggcgaaggacggggtcggc

> *Methylorubrum aminovorans* strain NBRC 15686 entry16

tccgactttgccgggaccgggacggggcctgctgcgggccggcgcgctgccgggggacaaggtgccgacccgcgcgcctgccccggggccgggggatca
gcagcgacgacgagcctctctgtttttcctgacaaatgatgctgctttgtgtcgggcccctgccctaccgttgctcggggcatcagcgagatccgcctctaggaggc
gcctcgacgcgctgcagtttcggcagggcctgtctcttgccgggtgcgctggggactttgctgccgcccgacgtaagccgttgccggcgaacgattttgccgttca
agcattcgtgcgggcaaaaaacagtcggcttggcagcagatttccgcttgacgacctcccaatacggacgcagtctctctgcgtgcttcgcattcgcaagcacatc
ggccaagaaaccgctggtccgcgcacggtttgttggcgaaagctggaggagacagggatgactccaccgcagcgtcaagccgcaacgatagcc**tgagaac
ccgttcggtct**c**ggactt****agc****ggcccgccacgaacaccggacg**gaagcggatccccaggggcgcagaagcggcctatacggcgcgacttcgaacac
gcatcagccccacctttggaccgcttccgccgcggtttctactgcaagcgggaagactttcaccgtgccggccgtcggggcccgcgcggatttcgccatgggtga
cagggcgatcgcgctggcgaagaggggtcggagcctaccgaccgacccgttgcggctttcacgacgcttctcatgctgcagccgctgccgcgaagtacgccg
cgacgcttcccccgagcgatcgacaccacgggatcaccgt

> *Methylorubrum thiocyanatum* strain JCM 10893 entry5

ggcggatccgcgctgcggcccggcttcgaccgcgggtcggcctatcggccggggagcttcgcgccgacccgcgggcccgtcctgtcgctgggggtgaaggc
gaggacaggcgcctcatccttgcggaggattgatgctgctttgtgtcggcgcccggccgcgaacctgcgaaatcggtccgccttccgtccggccgcgtggagctg
aacaggtcgccgtcttcgcttcccggttctccgcgcgacgccctcacgctgtcttctgcccgcctcgcggtcgaagcggacaggacttttcccctcgtgccgcgtaa
gccgttgccggcgaacgagtttgctgatcaagcattcgtgggggcgaaaaacagacggcttggcggcaaattcccgcttgacgacctgccaatacggacgcag
tctctatccgtgcttcgcagccgcaagcacatcggccaagaaaccgcgtggtccgcgcacggtttgttggtgaaagctggaggagacagggatgactccaccg
cagcatgaagccgcaacgataggc**tgagaacccgttcggtct**c**ggactt****agc****ggcccgccacgaacaccggacg**gaagcggatccctaggggcgca
gaagcggccgacacggcgcgacttcgaacacgcatcagccccacctgtggaccgcttccgccgcggtttctatcgcaagcgagaggcttttgaccgttccggcc

cgtcgaggcccgcgcggatttcgccgttggcgcgcggctggcgtcctatcgctgaaggcgttctcgagatcctccagccgatccgttccggcgatcgtaacgatcc
tcaggctgaggtgctgcgtcggcagcctcgacgcacgccgcaacacctgtctcagccgtgagggcggc

> *Methylorubrum populi* BJ001

ggcggatccgcgctgcggcccggcttcgaccgcgggtcggcctatcggccggggagcttcgcgccgacccgcgggcccgtcctgtcgctgggggtgaaggc
gaggacaggcgcctcatccttgcggaggattgatgctgctttgtgtcggcgcccggccgcgggcctgcgagaatcggtccgccttccgtccggccgcgtggacct
gaactgatcgccgtcttcgcttcccggttctccgcgcgacgccctcacgctgtcttctgcccgctccgcggttgaagcggacaggacttttcccctcgtgccgcgtaa
cccgttgccggcgaacgagtttgctgatcaagcattcgtgggggcgaaaaacagacggcttggcggcaaattcccgcttgacgacatgccaatacggacgcag
tctctatccgtgcttcgcagccgcaagcacatcggccaagaaaccgcgtggtccgcgcacggtttgttggcgaaagctggaggagacagggatgactccaccg
cagcatgaagccgcgacgataggc**tgagaacccgttcggtctcggact**agc**ggcccgccacgaacaccggacg**gaagcggatccctaggggcgca
gaagcggccgacacggcgcgacttcgaacacgcatcagccccacctgtggaccgcttccgccgcggtttctatcgcaagcgagagggttttgaccgtgccggc
ccgtcgaggcccgcgcggatttcgccgttggcgcgcggctggcgtcctatcgctgaaggcgttttcaagatccttcagccgatccgttcggcgatcagagcgatcc
gcatgctgaagtgatgcgtcggcagcctcgaagcacgccgcaacgccgatcccagccgtgagggcggcg

> *Methylorubrum zatmanii* strain LMG 6087

acccttggcgccggggaatttccacgcccgcaaacacgctttcgcgcggctcgacttaagtcattgccgacaaacgattttggtggcgaagatcgatgtcgggcg
aaaaacagtcggcttggcggcaaattcccgcttgacgacccgccaatagggtcgcagtctatcccggtgcttcgcagtcgcaagcacatcggtcaacaaactct
gttgtccgcacagagtttgttggtgaaagctggaggagactgggatgactccaccgcagcacgaagccgcgaggatggcc**tgagaaccgttcggtct**cagg
ctgagtggcccgccacgaacaccggacggaagcggatccctaggggcgcagaagcggccgacacggcgcgacttcgaacacgcatcagccccacct
gcggaccgcttccgccgcggtttcttgtgatcggtaaaggttttgcgccgtgcaagcccgtcgggctcgcgcgggtttgtcgttctgtgctccaaaggcgtgccggc
ggcgcgcttttggcggagagagtagcatccgttgagaatgtgccgggcgccgagggaaagccggagaaaagcttgacctgacgggcgagcaggctgctgga
acagacgatgtcgcgctgccaag

Sequences were obtained using the BLASTn tool (Altschul *et al.*, 1990) on RefSeq (O'Leary *et al.*, 2016) representative genomes from NCBI with the probe sequence for Methylo2624 and Methylo1969 (Table 9.1), which gives hit from one genome representing a clade cluster. The probe for each sRNA is shown in bold, whereas the conserved region used for RNAcode (Washietl *et al.*, 2011) and RNAz (Gruber *et al.*, 2010) underlined. Transcription starts sites identified by Maucourt *et al*. are emphasized in red (Maucourt *et al.*, 2022).

Table 9.3 Probes to test size of Methylo2624

Probe	Sequence
5` end, probe A	ggccgatgtagaggcccgcgccgccgacggcaaccgtgccgtgacgccgc
5` end, probe B	cgcaggcccgagagggctcgcgaaaccgtcaggcgcatcggccgcgggga
3` end, probe A	gtgggccttgcgccggggtcgcggttcctgttccccgcgctcccgacgca
3` end, probe B	atgcgtgccggtctggggctaggggaatgcggtcgagggaattcggatag

Table 9.4 RNAcode analysis of Methylo2624 and Methylo1969

	Amino acids (aa)				Nucleotides			
Frame	**Length**	**From**	**To**	**Name**	**Start**	**End**	**Score**	**P**
				Methylo2624				
+1	12	36	47	*Methylorubrum*	106	141	9.34	0.609
+2	9	30	38	*podarium* strain	89	115	3.32	1
+3	15	14	28	DSM 15083	42	86	2.15	1
				Methylo1969				
+3	44	47	90		141	272	11.32	0.676
+2	15	22	36		65	109	8.32	0.931
+2	38	50	87		149	262	4.07	1
+3	3	95	97	*Methylorubrum*	285	293	4.06	1
+3	8	100	107	*rhodesianum* strain	300	323	3.87	1
+3	8	18	25	DSM 5687	54	77	2.45	1
+1	4	102	105		305	316	2.05	1
+1	5	29	33		85	99	1.93	1

Regions used for this analysis are underlined in supplementary material, Table 9.2. RNAcode randomly selected a sequence from the alignment to determine the start and end positions of potential coding regions in amino acids and in nucleotides. Positions correspond to the intergenic region of *Methylorubrum podarium* strain DSM 15083 for Methylo2624, whereas it is from *Methylorubrum rhodesianum* strain DSM 5687 for Methylo1969.

Table 9.5 RNAz analysis of Methylo2624 and Methylo1969

	Methylo2624	**Methylo1969**
Mean pairwise identity	83.81	88.77
Mean single sequence minimum free energy (MFE)	-87.72	-47.18
Consensus MFE	-81.38	-41.68
Energy contribution	-77.38	-41.43
Covariance contribution	-4.00	-0.25
Mean z-score	-1.28	-1.41
Structure conservation index (SCI)	0.93	0.88
RNA-class probability	0.611487	0.608725
Prediction	RNA	RNA

Table 9.6 Composition of the CHOI culture medium (Bourque *et al.*, 1995)

	Compounds	Chemical formula	Volume to take from stock solutions [1]
Salts (1M stock solutions)	Ammonium sulfate	$(NH_4)_2SO_4$	6.81 mL
	Monopotassium phosphate	KH_2PO_4	5.754 mL
	Disodium phosphate	$Na_2HPO_4 \cdot 7H_2O$	8.998 mL
	Calcium chloride	$CaCl_2 \cdot 2H_2O$	24.6 µL
Metals (0.5M stock solutions)	Boric acid	H_3BO_3	34.93 µL
	Manganese sulfate	$MnSO_4 \cdot H_2O$	104 µL
	Zinc sulfate	$ZnSO_4 \cdot 7H_2O$	32.5 µL
	Copper sulfate	$CuSO_4 \cdot 5H_2O$	11.53 µL
	Sodium molybdate	$Na_2MoO_4 \cdot 2H_2O$	11.90 µL
	Cobalt chloride	$CoCl_2 \cdot 6H_2O$	12.10 µL
	Magnesium sulfate	$MgSO_4 \cdot 7H_2O$	3.288 mL
	Iron sulfate	$FeSO_4 \cdot 7H_2O$	259 µL

[1] For a final volume of 600 mL

9.2 Supplementary Figures

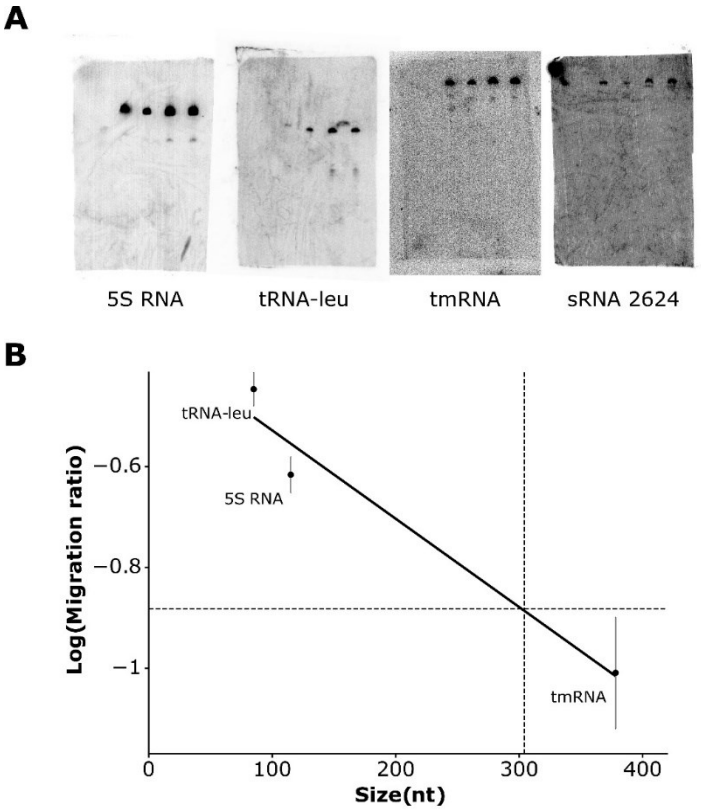

Figure 9.1 Size estimation of candidate sRNA2624 based on known RNAs

Hybridization of probes for controlled RNA is observed with radioactivity labelling. Since the size of these controlled RNAs are known (5S RNA, 115 nt; tRNA-leu, 85 nt and tmRNA; 378 nt), their migration ratio in a membrane can be measured and graphed to estimate the size of candidate RNAs such as sRNA2624 **(B)**. Probes in **(A)** are all on the same membrane, but the experiment was repeated on two other membranes (data not shown). The expected size of sRNA2624 differs between the three replicate (273 nt, 296 nt and 364 nt). RNAs are separated based on size within a gel before being transferred on a nitrocellulose membrane for Northern blot. Longer migration time results in better separation of larger RNAs, therefore facilitating the estimation of RNA size. These results guided us to determine the size of sRNA2624, but it was validated with probes towards both extremities (supplementary material, Table 9.3).

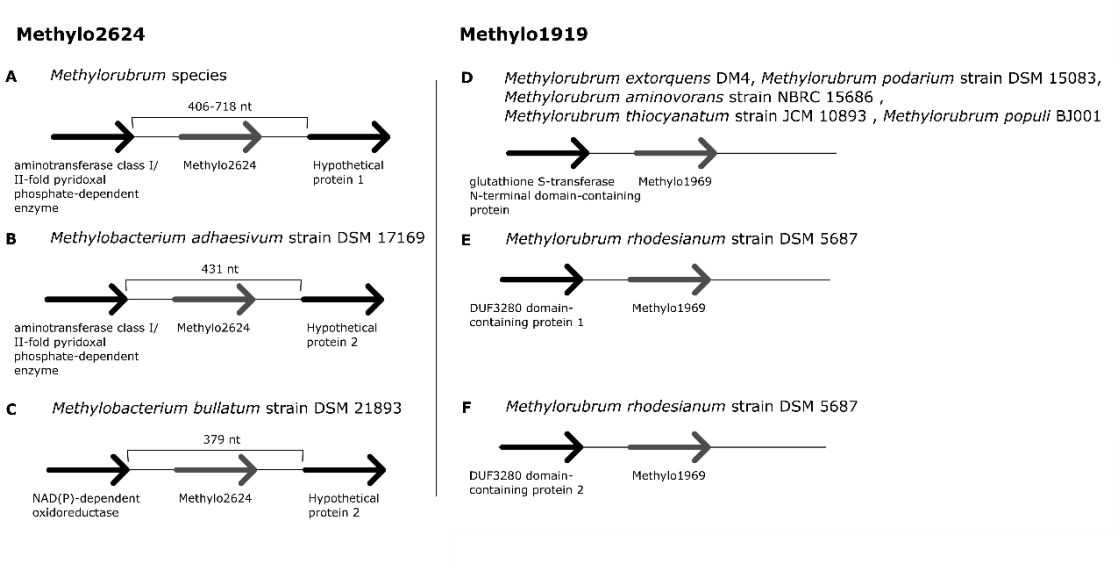

Figure 9.2 Genomic context of Methylo2624 and Methylo1969

Results are based on NCBI BLASTn (Altschul *et al.*, 1990) of the probe sequence for Methylo2624 and Methylo1969 (supplementary material Table 9.1) on RefSeq representatives genomes (O'Leary *et al.*, 2016). Hypothetical protein in **(A)** is different than that in **(B)** and **(C)**. DUF3280 domain-containing protein in **(E)** is different than that in **(F).**

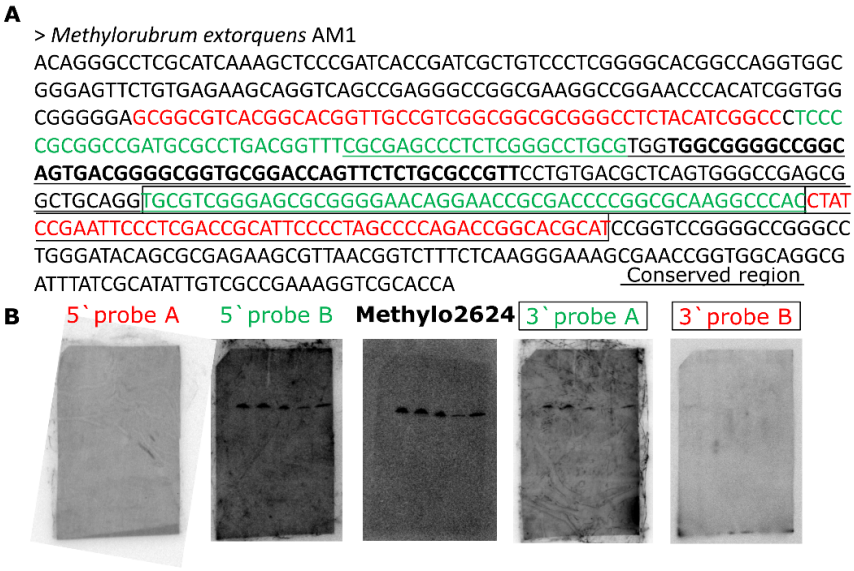

Figure 9.3 Probing Methylo2624 to delimitate its size

(A) Intergenic region from *Methylorubrum extorquens* AM1 containing Methylo2624. The conserved region is underlined. The probe region for Methylo2624 is emphasized in bold. The probes that resulted in a band compared to those that did not are shown in green and red respectively. Probes for the 3' end of Methylo2624 are boxed. The sequences of the probe 5' and 3' A and B can be found in supplementary material Table 9.3. The probe for Methylo2624 is available in supplementary material Table 9.1 **(B)** Northern blot results for probes delimiting Methylo2624. Probes for 5'B and 3'A hybridized on the membrane on the same bands for Methylo2624, whereas probe 5'A and 3'B led to no results. All probes were tested on the same membrane. The lanes correspond to different growth conditions (with methanol, succinic acid, hot and cold thermal shock and in a minimal media). These results also agree with estimated size based on Northern blot results (supplementary material, Figure 9.1). Even if the size of the region from the 5' and the 3' end probes in green is 186 nucleotides, there could be some extra nucleotides from both extremities included in sRNA2624, resulting in an estimated size between 250 and 350. Note also that the TSS predicted by Maucourt et al. (Maucourt *et al.*, 2022) (see Table 9.2) corresponds to the strain DM4 and that, given the significant sequence variation of IGRs outside of the conserved region, it is possible that the TSS of our strain is different (and upstream, which could hypothetically mean that Methylo2624 would be larger in our strain than in DM4), which would explain that Northern-based estimates of size range between ~250 and 350, while the conserved region is only ~150.

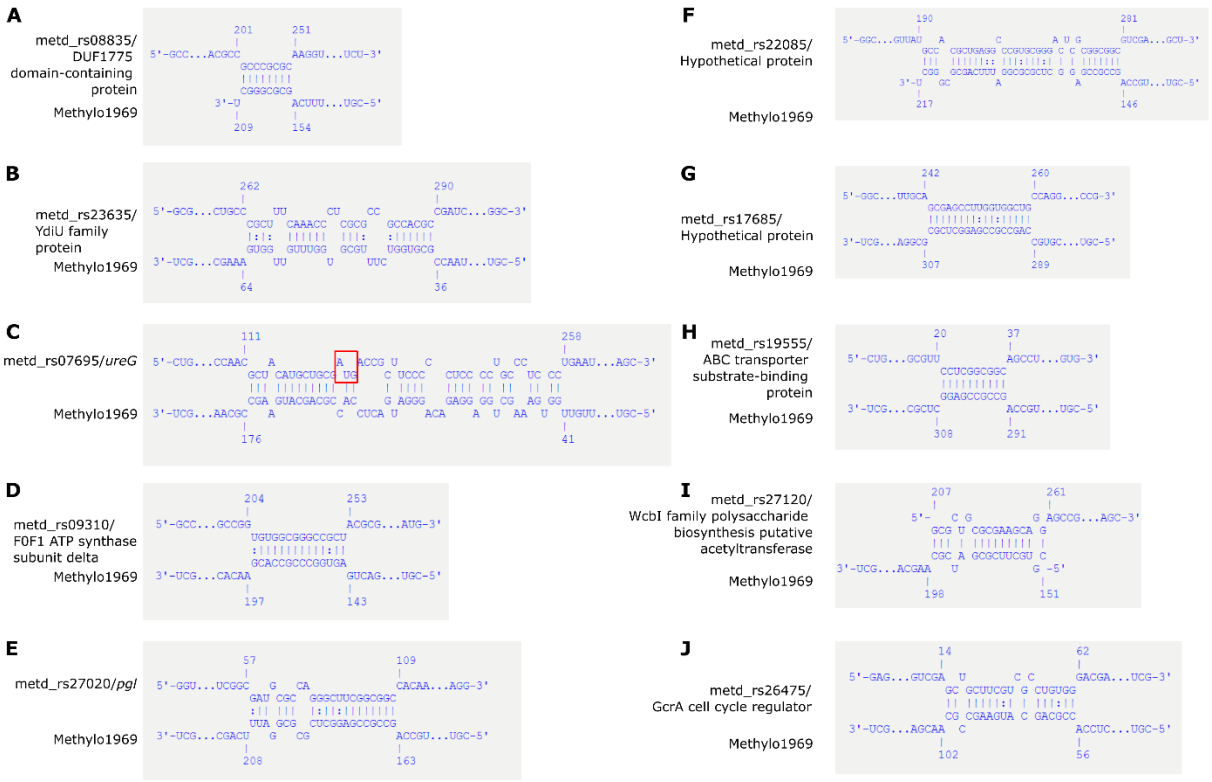

Figure 9.4 Predicted interactions between Methylo1969 and targets

Interactions are predicted by the tool IntaRNA (Mann *et al*., 2017) as part of CopraRNA (Wright *et al*., 2014). The start codon for gene *ureG* is framed in red.

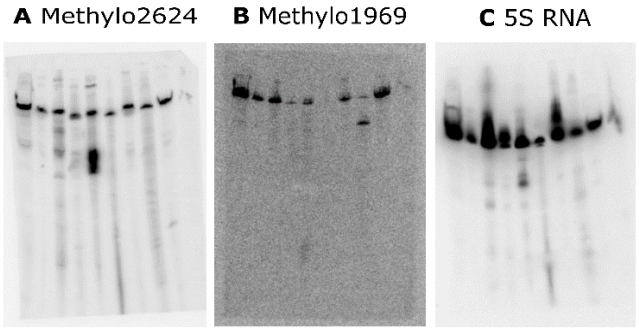

Figure 9.5 Methylo2624, Methylo1969 and 5S RNA in multiple growth conditions (full membranes)

Growth conditions from left to right: No cobalt, 30°C, 20°C, 37°C, 1 mM fluor, 30 mM urea, 30 mM guanidine, 2% ethanol and 100 mM NaCl. Results from A, B and C are taken from the same membrane, with different probe (Methylo2624, Methylo1969 and 5S RNA respectively). All conditions were complemented with 1% methanol as a source of carbon, except for growth with 2% ethanol.

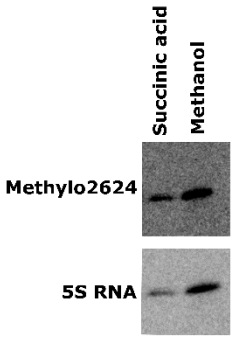

Figure 9.6 Expression of Methylo2624 when grown with succinic acid (20 mM) and methanol (1%)

10 ANNEXE III: SHIFTED-REVERSE PAGE: A NOVEL APPROACH BASED ON STRUCTURE SWITCHING FOR THE DISCOVERY OF RIBOSWITCHES AND APTAMERS (SUPPLEMENTARY MATERIAL)

10.1 Supplementary figures

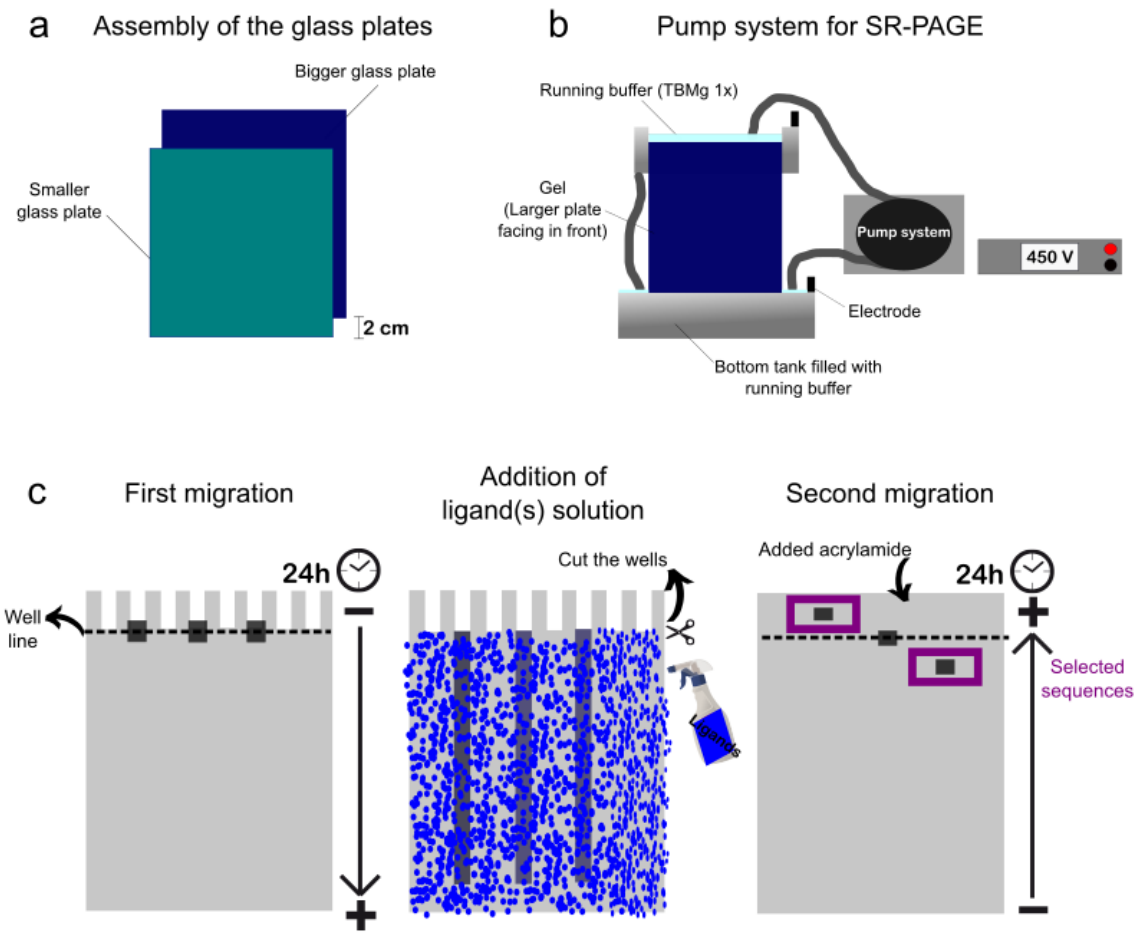

Figure 10.1 SR-PAGE method

(A) When the gel is poured between the two glass plates, the smaller glass plate is placed 2 cm below the bigger glass plates as shown. **(B)** Schematic representation of a SR-PAGE assembly with the pump system. **(C)** First migration of the RNA sequences within a native polyacrylamide gel for 24 hours. Secondly, the gel is small plate is removed and the ligand solution is sprayed unto the gel. The wells are cut. The space left by the removal of the wells is filled by the leftover native polyacrylamide so that the reverse run could cover above the wells. Finally, the second migration of the RNA sequence where the polarity of the electrode is inverted. The sequences that had a change in migration (that did not end at the well line after the second migration) are selected.

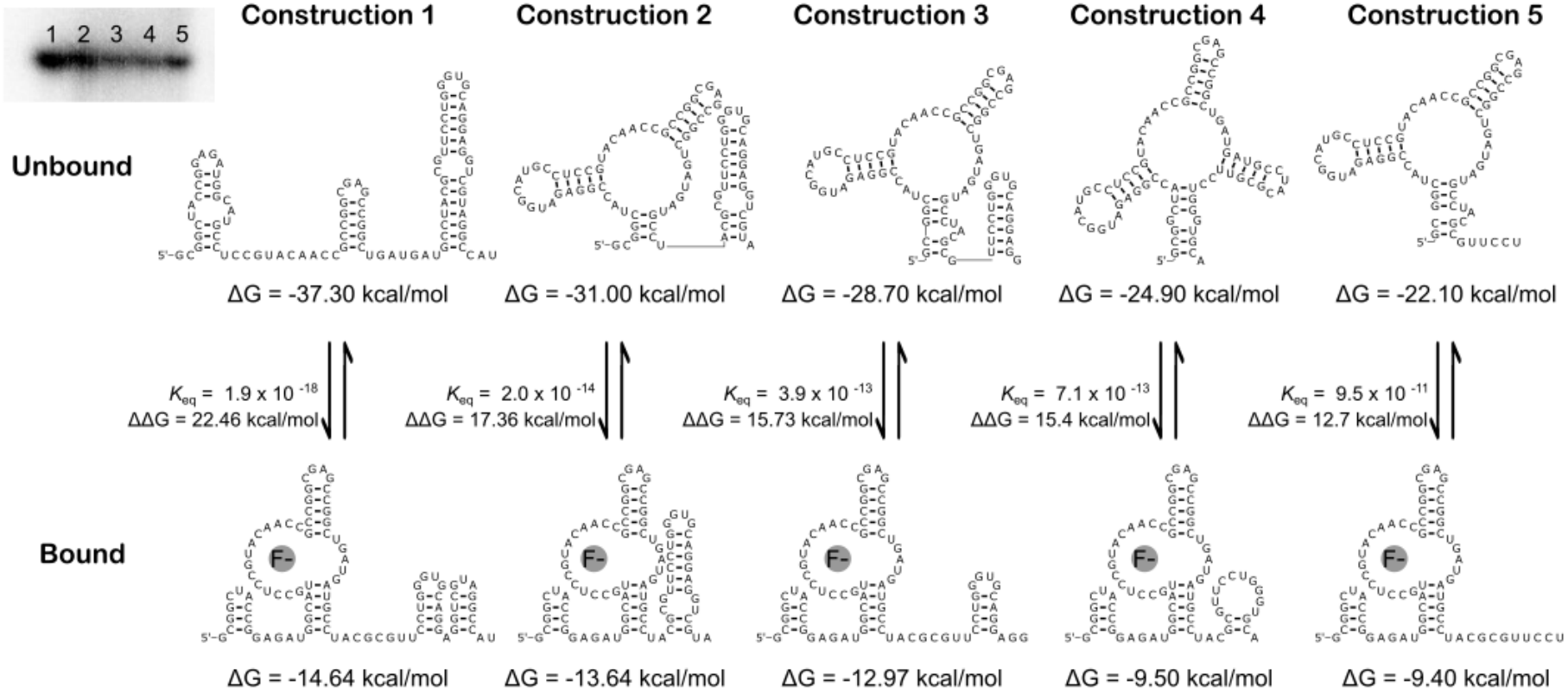

Figure 10.2 Free energy of all different constructions of the fluoride riboswitch used in the SR-PAGE experiment in their bound (constrained) and unbound (unconstrained) conformations

For Supplementary figures 10.2-10.5, the structures were obtained with Mfold (Zuker, 2003) using constraints described in Supplementary Table 10.2 and corresponding to the known folding of the aptamer bound-conformation according to available atomic resolution data of the given riboswitch aptamer domain (as the structure of the expression platform was not solved for these examples), within the boundaries of Mfold's capacity (i.e. precluding non-canonical base pairs and other similar tertiary interactions). The unbound version simply corresponds to the unconstrained Minimum Free Energy (MFE) structure predicted by Mfold. Secondary structures were generated with R2R (Weinberg & Breaker, 2011).

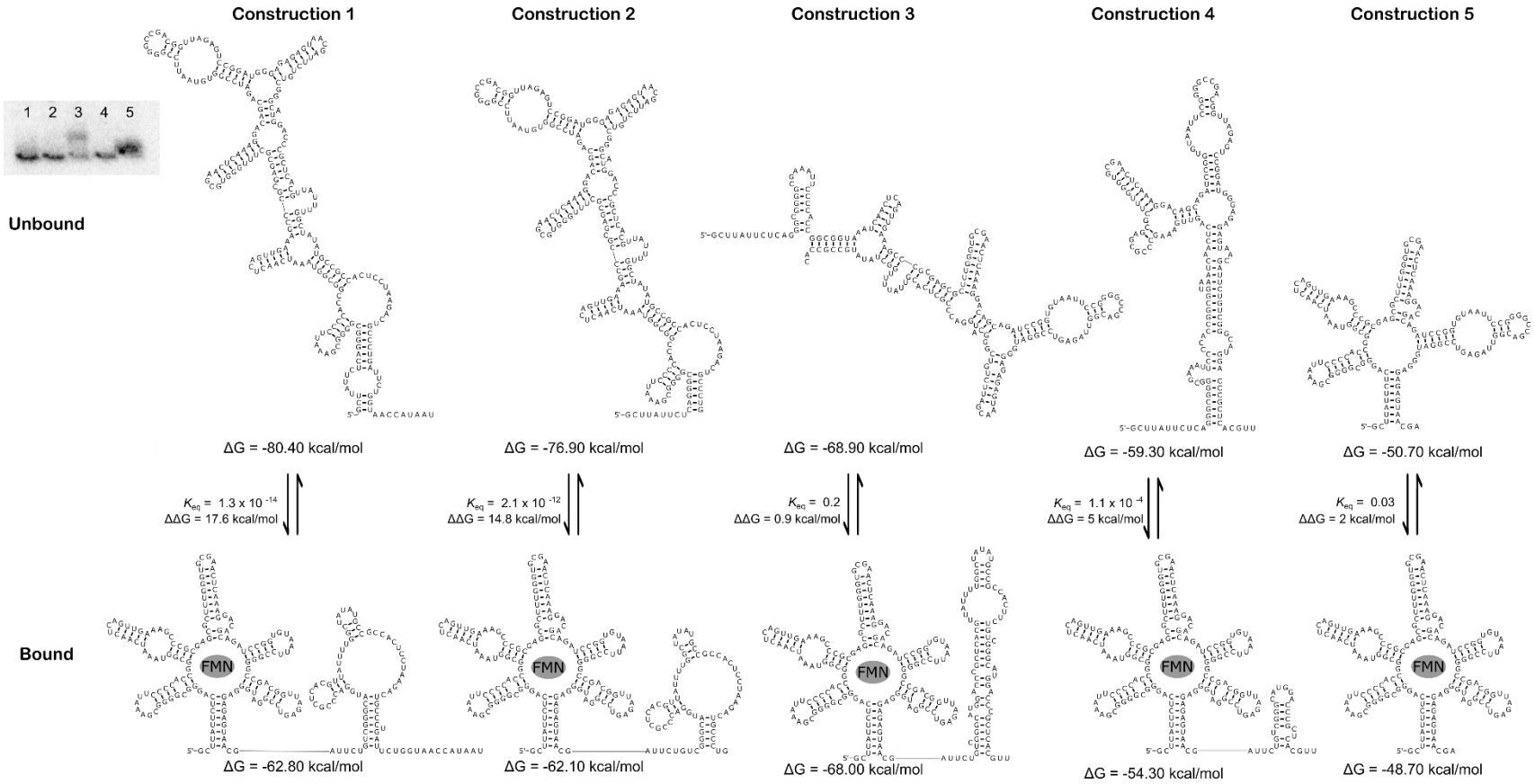

Figure 10.3 Free energy of all different constructions of the FMN riboswitch used in SR-PAGE experiment in their bound (constrained) and unbound (unconstrained) conformations.

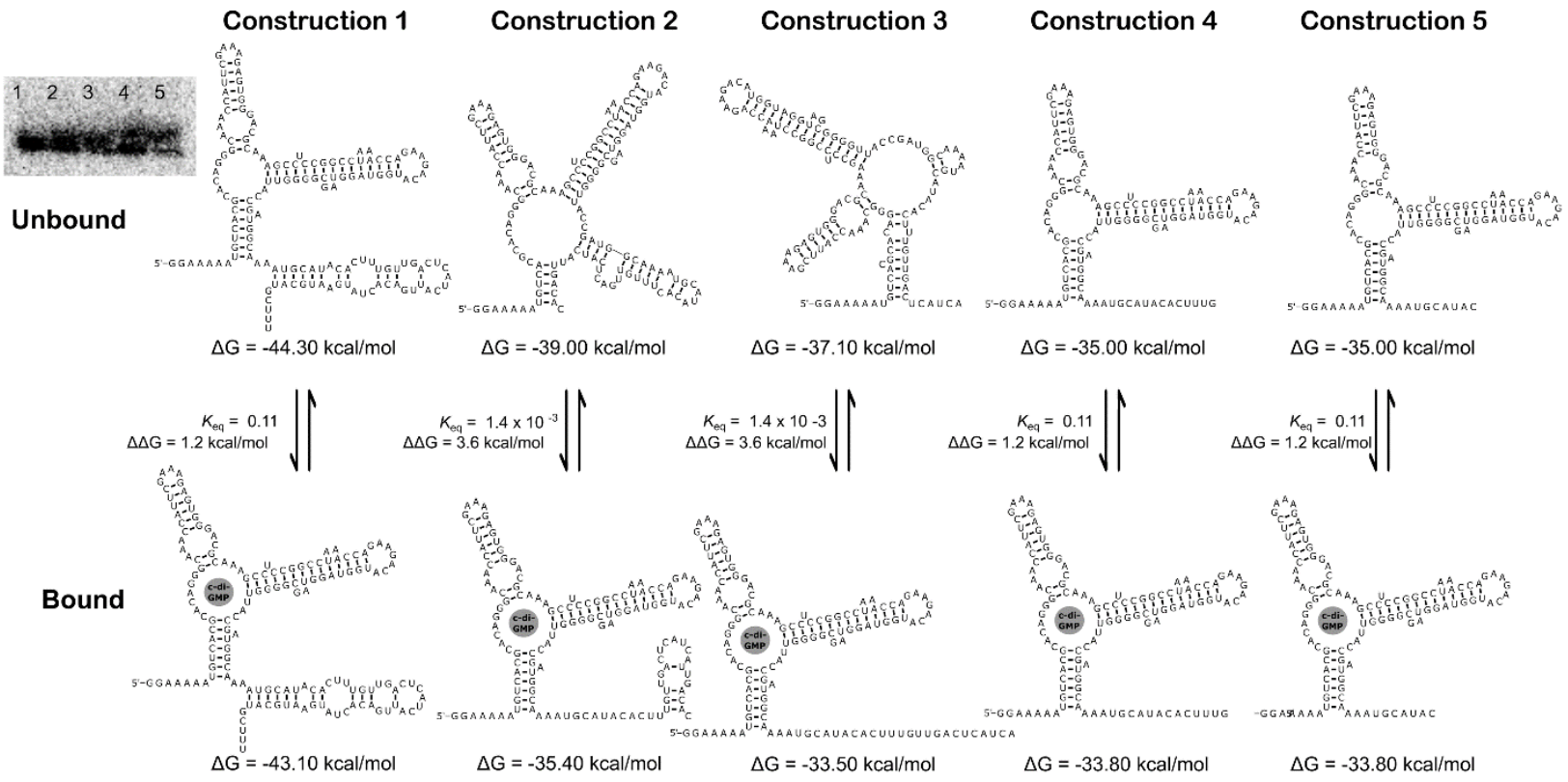

Figure 10.4 Free energy of all different constructions of the c-di-GMP I riboswitch used in the SR-PAGE experiment in their bound (constrained) and unbound (unconstrained) conformations

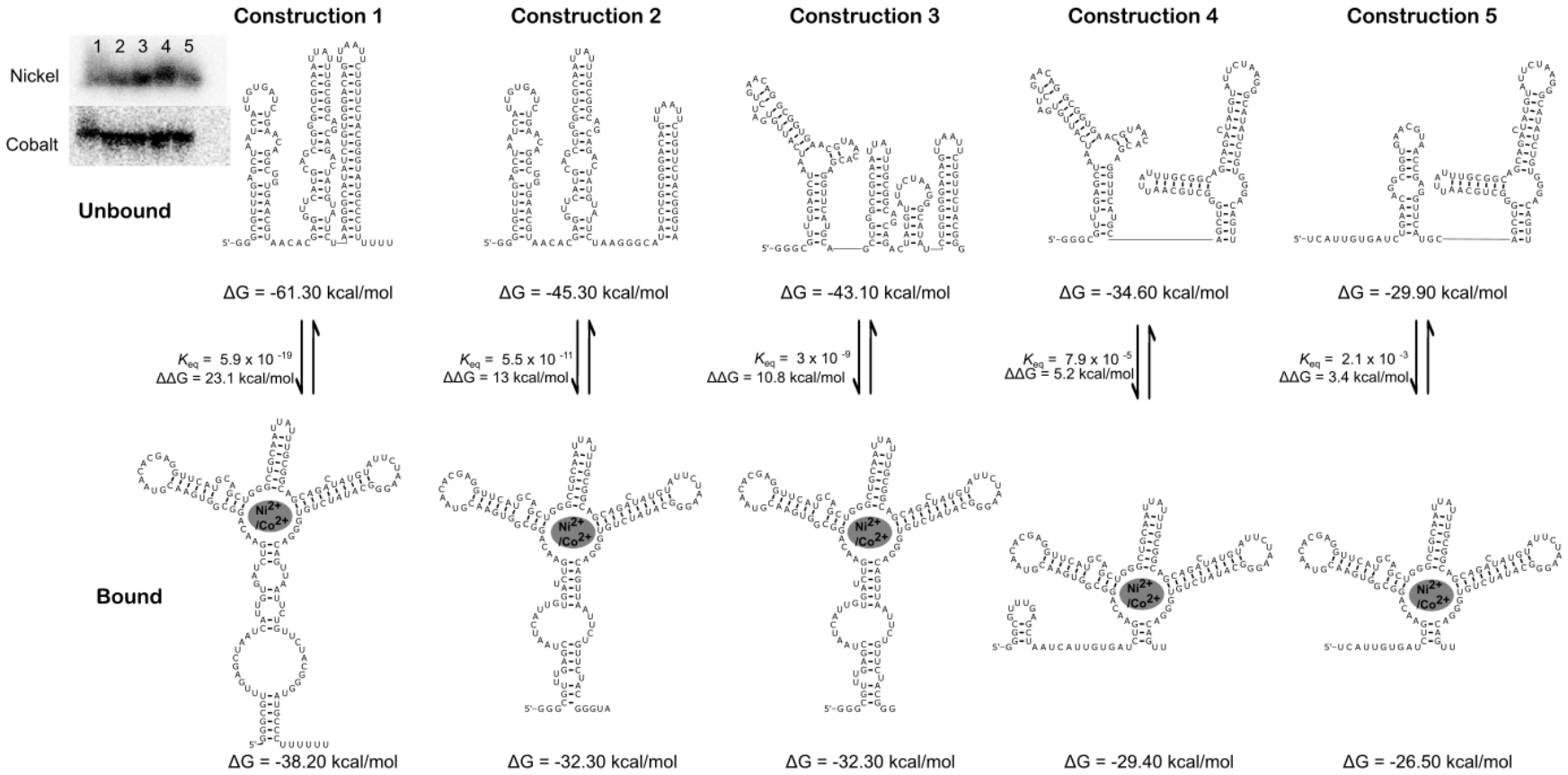

Figure 10.5 Free energy of all different constructions of the nickel-cobalt riboswitch used in the SR-PAGE experiment in their bound (constrained) and unbound (unconstrained) conformations.

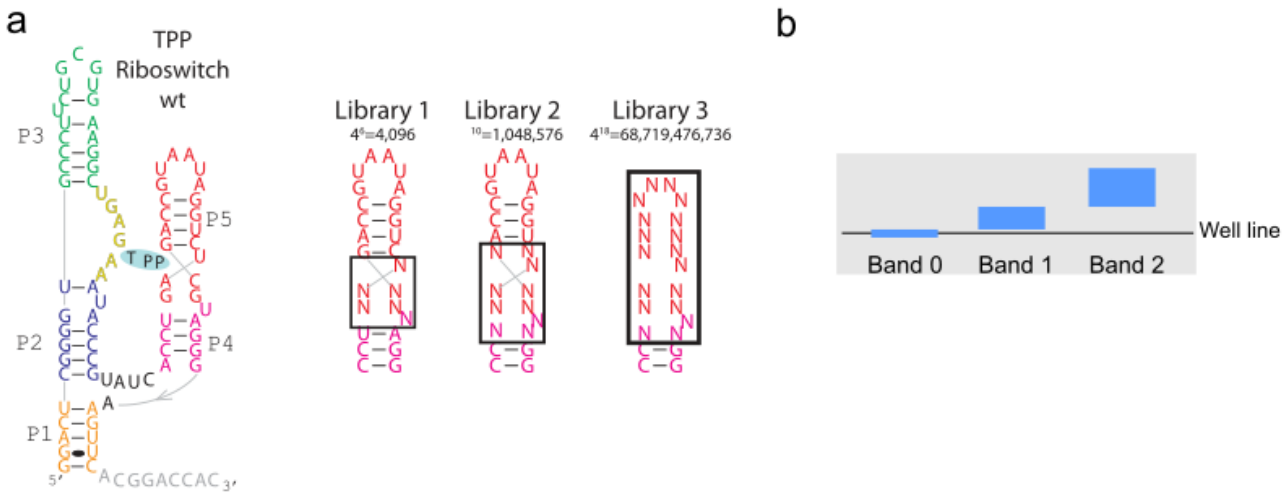

Figure 10.6 Degenerated libraries of TPP riboswitch

(A) Librairies 1, 2 and 3 had 6, 10 and 18 degenerated nucleotides respectively within stem P5. **(B)** Schematic representation of the bands that were cut-out after the second migration of the SR-PAGE (Supplementary Figure 10.1c). The cut 0 is only the wells. The cut 1 is approximately from 2 mm to 1 cm above the wells and the cut 2 is approximately 1 cm above the wells and above.

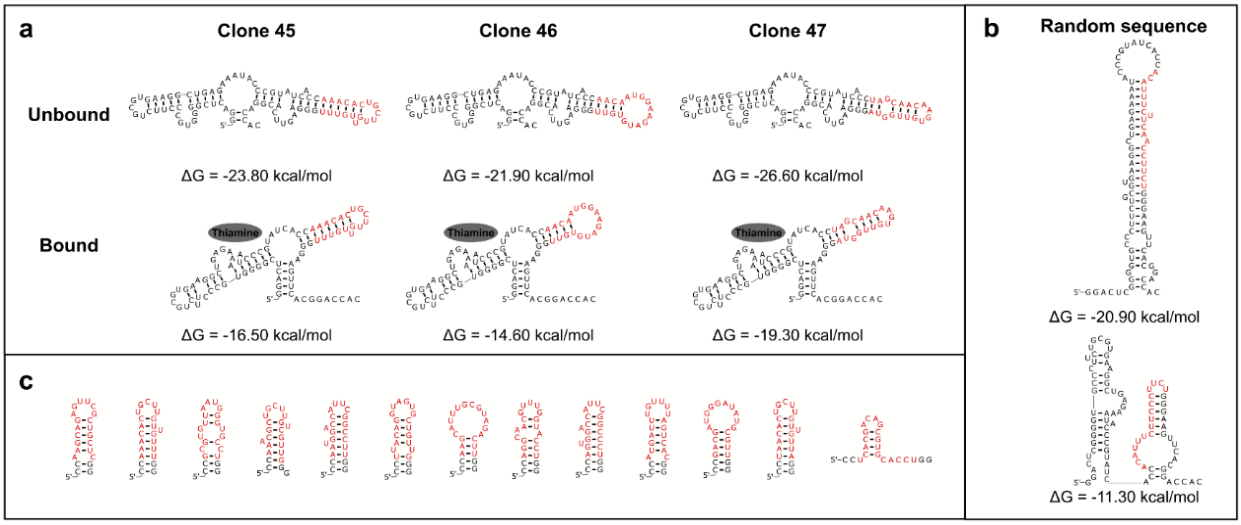

Figure 10.7 Library 3 TPP-derived thiamine switches have a stem that replaces P4-P5.

(A) Secondary structures were prepared as for Supplementary Figure 10.2-Figure 10.5 by constraining only P2 and P3, between which the thiamine binding pocket is found. It is noteworthy that the three validated sequences derived from the selected library 3 have similar overall bound/unbound structures, i.e. where the randomized regions (in red) can form a stem present in both the predicted bound and unbound states. **(B)** Secondary structure of a randomly generated sequence expected to be found in the starting pool, with the same constraints as in A. **(C)**The representativity of these examples was further evaluated by looking at the sequences with four or more reads in our deep sequencing. Most selected sequences can form stems. Conversely, only a minority of non-selected random sequences can form stems (Supplementary Table 10.4).

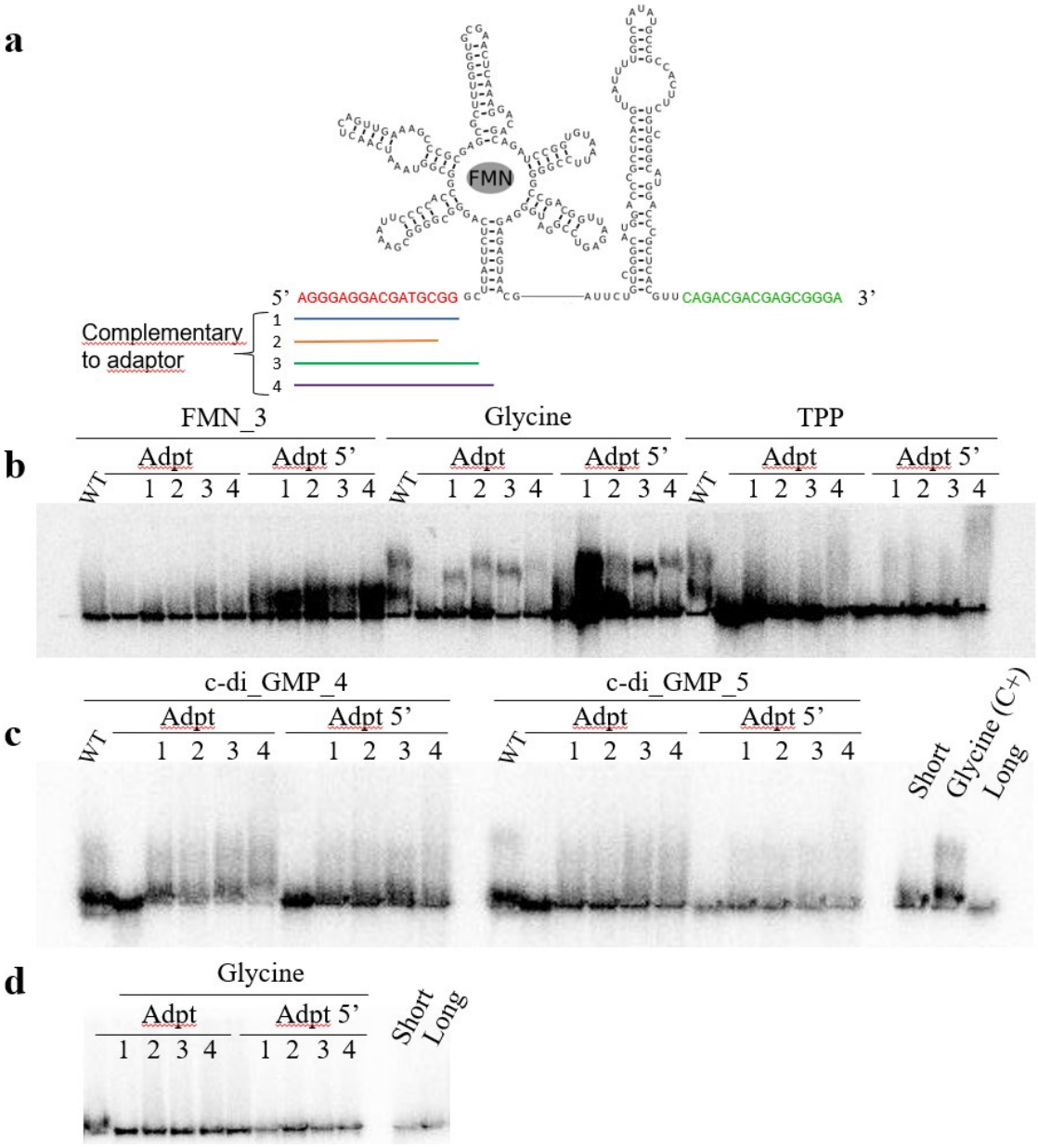

Figure 10.8 The presence of oligonucleotides complementary to the adapters makes it possible to restore the shift of the riboswitches by the SR-PAGE method.

(A) Schematic of the design of complementary oligonucleotides with an example on the FMN riboswitch. The sequences in red and green are the adapters for PCR amplification. The colored lines represent the different complementary oligonucleotides tested: 1 is fully complementary to the adapter; 2 is complementary to the adapter minus two nucleotides; 3 is fully complementary to the adapter with two additional nucleotides complementary to the riboswitch sequence; and 4 is fully complementary to the adapter with four additional nucleotides complementary to the riboswitch sequence. **(B)** Result of SR-PAGE with FMN_3, Glycine and TPP riboswitches by spraying the corresponding ligand. “Adpt’’ corresponds to the 5' and 3' complementary adapters. “Adpt 5’’ corresponds to the 5' complementary adapter only. “WT’’ is the riboswitch without adapters. The numbers correspond to the combination of the oligonucleotides as described before. **(C)** Result of SR-PAGE with c-di-GMP_4 and _5 with a solution of c-di-GMP sprayed. ‘’Short’’, ‘’Long ‘’ represents two controls to verify proper gel migration. Glycine riboswitch is used as a positive control, so the ligand glycine was also sprayed. **(D)** Result of SR-PAGE with glycine riboswitch without spraying any ligand.

10.2 Supplementary tables

Table 10.1 List of all the oligonucleotides

Degenerated libraries of the TPP riboswitch
Librairie 1 _SELEX_TPP - Lib1_For: TAATACGACTCACTATAggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacc - Lib1_Rev: gtggtccgtgaacttccctNNNctggcattatccagNNNaggtgatacgggtatttctc
Librairie 2_SELEX_TPP - Lib2_For: TAATACGACTCACTATAggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacc - Lib2_Rev: gtggtccgtgaacttcccNNNNNtggcattatccaNNNNNggtgatacgggtatttctc
Librairie 3_SELEX_TPP - Lib3_For: TAATACGACTCACTATAggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacc - Lib3_Rev: gtggtccgtgaacttcccnnnnnnnnnnnnnnnnnnnggtgatacgggtatttctc
Amplification from clones pGEMT (amplified from miniprep) - Cloning_For: TAATACGACTCACTATAggactcggggtgccctt - Cloning_Rev: gtggtccgtgaacttccc (note: full sequence of these clones available in supplementary Table 10.3)
Enriched clones (Assembly PCR) - TPP_For: TAATACGACTCACTATAggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacc - TPP_Rev: o TPP1_Rev: gtggtccgtgaacttccctaggctggcattatccaggcgaggtgatacgggtatttctc o TPP2_Rev: gtggtccgtgaacttcccttgtctggcattaccaggtcaggtgatacgggtatttctc o TPP3_Rev: gtggtccgtgaacttcccgagcatggcattatccaaacaaggtgatacgggtatttctc o TPP4_Rev: gtggtccgtgaacttcccaagcatggcattatccaaacaaggtgatacgggtatttctc o TPP5_Rev: gtggtccgtgaacttcccagcattggcattatccaaaaccggtgatacgggtatttctc o TPP6_Rev: gtggtccgtgaacttcccaaccatggcattatccaaagcaggtgatacgggtatttctc o TPP7_Rev: gtggtccgtgaacttcccttcgctggcattatccagacgaggtgatacgggtatttctc o TPP8_Rev: gtggtccgtgaacttccctgctctggcattatccaactggggtgatacgggtatttctc o TPP45_Rev: gtggtccgtgaacttcccaaacaacaagcagtgtttggtgatacgggtatttctc o TPP46_Rev: gtggtccgtgaacttcccaacacatcttccattgttggtgatacgggtatttctc o TPP47_Rev: gtggtccgtgaacttccctaccaacacttgttgctaggtgatacgggtatttctc o TPP48_Rev: gtggtccgtgaacttcccgagcgtccgatctgcttcggtgatacgggtatttctc
Fluorescent oligonucleotide (control for the SR-PAGE migration)
Cy3-tccccttgcatacatgtcgcgtttatgcttcgaccctctagccgcacccttgg as well as Cy3-ccatcggcactgacggcctcacttgcagcagaacgtgggcagcttgctgaatcgttctgcaagagtgagcccgtaacataatgg
Primers for constructs of FMN riboswitch (Amplified from *Escherichia coli*)
FMN_For: TAATACGACTCACTATAGGgcttattctcagggcgg

FMN_1

- FMN_1_Rev: attatggttaccagaatcagggca
- FMN_1_full sequence:
  TAATACGACTCACTATAGGgcttattctcagggcggggcgaaattccccaccggcggtaaatcaactcagttgaaagcccgcg
  Agcgctttgggtgcgaactcaaaggacagcagatccggtgtaattccggggccgacggttagagtccggatgggagagagtaacgattct
  gtcgggcatggacccgctcacgttattttggctatatgccgccactcctaagactgccctgattctggtaaccataat

FMN_2

- FMN_2_Rev: cagggcagtcttaggagt
- FMN_2_full sequence:
  TAATACGACTCACTATAGGgcttattctcagggcggggcgaaattccccaccggcggtaaatcaactcagttgaaagcccgcg
  agcgctttgggtgcgaactcaaaggacagcagatccggtgtaattccggggccgacggttagagtccggatgggagagagtaacgattct
  gtcgggcatggacccgctcacgttattttggctatatgccgccactcctaagactgccctg

FMN_3

- FMN_3_Rev: gtggcggcatatagccaa
- FMN_3_full sequence:
  TAATACGACTCACTATAGGgcttattctcagggcggggcgaaattccccaccggcggtaaatcaactcagttgaaagcccgcg
  agcgctttgggtgcgaactcaaaggacagcagatccggtgtaattccggggccgacggttagagtccggatgggagagagtaacgattct
  gtcgggcatggacccgctcacgttattttggctatatgccgccac

FMN_4

- FMN_4_Rev: aacgtgagcgggtccat
- FMN_4_full sequence:
  TAATACGACTCACTATAGGgcttattctcagggcggggcgaaattccccaccggcggtaaatcaactcagttgaaagcccgcg
  agcgctttgggtgcgaactcaaaggacagcagatccggtgtaattccggggccgacggttagagtccggatgggagagagtaacgattct
  gtcgggcatggacccgctcacgtt

FMN_5

- FMN_5_Rev: tcgttactctctcccatcc
- FMN_5_full sequence:
  TAATACGACTCACTATAGGgcttattctcagggcggggcgaaattccccaccggcggtaaatcaactcagttgaaagcccgcg
  agcgctttgggtgcgaactcaaaggacagcagatccggtgtaattccggggccgacggttagagtccggatgggagagagtaacga

Primers for constructs of Fluoride riboswitch (Amplified from *Burkholderia thailandensis*)

Fluor_For: TAATACGACTCACTATAGGcggctaccggagatgg

Fluor_1

- Fluor_1_Rev: atggcctacgacctcctg
- Fluor_1_full sequence:
  TAATACGACTCACTATAGGgcggctaccggagatggcatgcctccgtacaaccgccggcgagccggctgatgatgcctacgc
  gttcctgggtgcaggaggtcgtaggccat

Fluor_2

- Fluor_2_Rev: tacgacctcctgcaccc
- Fluor_2_full sequence:
  TAATACGACTCACTATAGGgcggctaccggagatggcatgcctccgtacaaccgccggcgagccggctgatgatgcctacgc
  gttcctgggtgcaggaggtcgta

Fluor_3

- Fluor_3_Rev: cctcctgcacccagg
- Fluor_3_full sequence:
  TAATACGACTCACTATAGGgcggctaccggagatggcatgcctccgtacaaccgccggcgagccggctgatgatgcctacgc

gttcctgggtgcaggagg

Fluor_4

- Fluor_4_Rev: tgcacccaggaacgc
- Fluor_4_full sequence:
  TAATACGACTCACTATAGGgcggctaccggagatggcatgcctccgtacaaccgccggcgagccggctgatgatgcctacgc
  gttcctgggtgca

Fluor_5

- Fluor_5_Rev: aggaacgcgtaggcatca
- Fluor_5_full sequence:
  TAATACGACTCACTATAGGgcggctaccggagatggcatgcctccgtacaaccgccggcgagccggctgatgatgcctacgc
  gttcct

Primers for construction of c-di-GMP I riboswitch (PCR assembly of riboswitch from *Vibrio cholerae*)

c-di-GMP_1 (created by PCR assembly and used as template for other GMP constructs)

- c-di-GMP_1_For_A: TAATACGACTCACTATAGGgaaaaatgtcacgcacagggc
- c-di-GMP_1_Rev_B: ggtttaggccggaggctttgcgtcccactctttcgaatggtttgccctgtgcgtg
- c-di-GMP_1_For_C: ctccggcctaaaccagaagacatggtaggtagcggggttaccgatggcaaaatgcataca
- c-di-GMP_1_Rev_D: aaagcatgcattcatagtgtcaatgatgagtcaacaaagtgtatgcattttgccatcg
- c-di-GMP_1_full sequence:
  TAATACGACTCACTATAGGggaaaaatgtcacgcacagggcaaaccattcgaaagagtgggacgcaaagcctccggccta
  aaccagaagacatggtaggtagcggggttaccgatggcaaaatgcatacactttgttgactcatcattgacactatgaatgcatgcttt

c-di-GMP_2

- c-di-GMP_for: TAATACGACTCACTATAGGgaaaaatgtcacgcacagggc
- c-di-GMP_2_Rev: gtgtcaatgatgagtcaacaaagt
- c-di-GMP_2_full sequence:
  TAATACGACTCACTATAGGggaaaaatgtcacgcacagggcaaaccattcgaaagagtgggacgcaaagcctccggccta
  aaccagaagacatggtaggtagcggggttaccgatggcaaaatgcatacactttgttgactcatcattgacac

c-di-GMP_3

- c-di-GMP_3_Rev: tgatgagtcaacaaagtgtatgc
- c-di-GMP_3_full sequence:
- TAATACGACTCACTATAGGggaaaaatgtcacgcacagggcaaaccattcgaaagagtgggacgcaaagcctccggccta
  aaccagaagacatggtaggtagcggggttaccgatggcaaaatgcatacactttgttgactcatca

c-di-GMP_4

- c-di-GMP_4_Rev: aaagtgtatgcattttgccatc
- c-di-GMP_4_full sequence:
  TAATACGACTCACTATAGGggaaaaatgtcacgcacagggcaaaccattcgaaagagtgggacgcaaagcctccggccta
  aaccagaagacatggtaggtagcggggttaccgatggcaaaatgcatacactttg

c-di-GMP_5

- c-di-GMP_5_Rev: gtatgcattttgccatcggta
- c-di-GMP5_full sequence:
- TAATACGACTCACTATAGGggaaaaatgtcacgcacagggcaaaccattcgaaagagtgggacgcaaagcctccggccta
  aaccagaagacatggtaggtagcggggttaccgatggcaaaatgcatac

Primers for construction of glycine riboswitch (PCR assembly of riboswitch from *Vibrio cholerae*)

- Gly_For_A: TAATACGACTATAGGgttgaagactgcaggagagtggttgttaaccagattttaacatctgagccaaataacccg
- Gly_Rev_B: tcgcttattcgttgccaatatatggctaagaataatgcacctgaaagatttacttcttcggcgggttatttggctcagatg
- Gly_For_C: ttggcaacgaataagcgaggactgtagttggaggaacctctggagagaaccgtttaatcggtcgccgaaggagcaag
- Gly_Rev_D: tcctctgtccttttgcctgagagtttcactctgcatatgcgcagagcttgctccttcggcgaccgat
- Gly_full sequence:
  TAATACGACTATAGGgttgaagactgcaggagagtggttgttaaccagattttaacatctgagccaaataacccgccgaagaagta
  Aatctttcaggtgcattattcttagccatatattggcaacgaataagcgaggactgtagttggaggaacctctggagagaaccgtttaatcggtc
  gccgaaggagcaagctctgcgcatatgcagagtgaaactctcaggcaaaaggacagagga

Primers for construction of TPP riboswitch (Assembly PCR of TPP riboswitch from *Escherichia coli*)

- TPP_For: TAATACGACTCACTATAggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacc
- TPP_Rev: gtggtccgtgaacttccctacggctggcattatccagatcaggtgatacgggtatttctc
- TPP_full sequence: TAATACGACTCACTATAGGggactcggggtgcccttctgcgtgaaggctgagaaatacccgtatcacctgatctggataatgcca
  gcgtagggaagttcacggaccac

Primers for construction of nickel-cobalt riboswitch (PCR assembly of riboswitch from *Listeria monocytogenes*)

NiCo_1 (template NiCo_4)

- NiCo_1_F: TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcg
- NiCo_1_R: aaaaaagggcatacccgtagaacagaattaac
- NiCo_1_full_sequence:
  TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcggtgaacgtaacacgaggttcatgcagct
  gggctgcaattatttgcggcagcagactatgtattctaagggcatatctgtgggacagttaattctgttctacgggtatgccctttttt

NiCo_2 (template NiCo_4)

- NiCo_2_F: TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcg
- NiCo_2_R: tacccgtagaacagaattaactgtcccacagatatgc
- NiCo_2_full_sequence:
  TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcggtgaacgtaacacgaggttcatgcagct
  gggctgcaattatttgcggcagcagactatgtattctaagggcatatctgtgggacagttaattctgttctacgggta

NiCo_3 (template NiCo_4)

- NiCo_3_F: TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcg
- NiCo_3_R: cccgtagaacagaattaactgtcccacagatatgc
- NiCo_3_full_sequence:
  TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcggtgaacgtaacacgaggttcatgcagct
  gggctgcaattatttgcggcagcagactatgtattctaagggcatatctgtgggacagttaattctgttctacggg

NiCo_4 (template Nico_5)

- NiCo_4_F: TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcg
- NiCo_4_R: aactgtcccacagatatgccctt
- NiCo_4_full_sequence:
  TTCTAATACGACTCACTATAGGgcgtttgagctaatcattgtgatctgaacaggcggtgaacgtaacacgaggttcatgcagct
  gggctgcaattatttgcggcagcagactatgtattctaagggcatatctgtgggacagtt

NiCo_5 (created by PCR assembly and used as template for other nickel-cobalt constructs)

- NiCo_1_For_A: TTCTAATACGACTCACTATAGGtcattgtgatctgaacaggcggtgaacgtaa
- NiCo_1_Rev_B: tgcagcccagctgcatgaacctcgtgttacgttcaccgcctg
- NiCo_1_For_C: cagctgggctgcaattatttgcggcagcagactatgtattctaagggcatatctgtggga
- NiCo_1_Rev_D: aactgtcccacagatatgccctt
- NiCo_1_full_sequence:
  TTCTAATACGACTCACTATAGGtcattgtgatctgaacaggcggtgaacgtaacacgaggttcatgcagctgggctgcaattat

ttgcggcagcagactatgtattctaagggcatatctgtgggacagtt

Test of the influence of adapters

c-di-GMP adapter (to synthesize the riboswitch with few combinations of adapters)

Template control riboswitch:

- c-di-GMP_For_noT7: ggaaaaatgtcacgcacaggg
- c-di-GMP4_Rev: caaagtgtatgcattttgccatc
- c-di-GMP5_Rev: gtatgcattttgccatcggta

c-di-GMP4_adpt 5'_3':

- c-di-GMP4_adapt 5'_3'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggaaaaatgtcacgcacaggg
- c-di-GMP4_adapt 5'_3'_Rev : *tcccgctcgtcgtctgc*aaagtgtatgcattttgccatc

c-di-GMP4_adpt 5':

- c-di-GMP4_adapt 5'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggaaaaatgtcacgcacaggg
- c-di-GMP4_adapt 5'_Rev : caaagtgtatgcattttgccatc

c-di-GMP4_adpt 3':

- c-di-GMP4_adapt 3'_For: TTAATACGACTCACTATAgggaaaaatgtcacgcacagggc
- c-di-GMP4_adapt 3'_Rev : *tcccgctcgtcgtctgc*aaagtgtatgcattttgccatc

c-di-GMP5_adpt 5'_3':

- c-di-GMP5_adapt 5'_3'_For : gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggaaaaatgtcacgcacaggg
- c-di-GMP5_adapt 5'_3'_Rev : *tcccgctcgtcgtctg*gtatgcattttgccatcggta

c-di-GMP5_adpt 5':

- c-di-GMP5_adapt 5'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggaaaaatgtcacgcacaggg
- c-di-GMP4_adapt 5'_Rev : gtatgcattttgccatcggta

c-di-GMP5_adpt 3':

- c-di-GMP5_adapt 3'_For: TTAATACGACTCACTATAgggaaaaatgtcacgcacagggc
- c-di-GMP5_adapt 3'_Rev : *tcccgctcgtcgtctgc*gtatgcattttgccatcggta

Glycine adapter (to synthesize the riboswitch with few combinations of adapters)

Template:

- Gly_For_noT7: gttgaagactgcaggagag
- Gly_Rev: tcctctgtcctttgcctga

Gly_adpt 5'_3':

- Gly_adapt 5'_3'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*gttgaagactgcaggagag
- Gly_adapt 5'_3'_Rev : *tcccgctcgtcgtctg*tcctctgtcctttgcctga

Gly_adpt 5':

- Gly_adapt 5'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*gttgaagactgcaggagag
- Gly_adapt 5'_Rev : tcctctgtccttttgcctga

Gly_adpt 3':

- Gly_adapt 3'_For: TTAATACGACTCACTATAgg
- Gly_adapt 3'_Rev : *tcccgctcgtcgtctg*tcctctgtccttttgcctga

FMN adapter (to synthesize the riboswitch with few combinations of adapters)

Template:

- FMN_For_noT7: gcttattctcagggcgg
- FMN_Rev: gtggcggcatatagccaa

FMN_adpt 5'_3':

- FMN_adapt 5'_3'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*gcttattctcagggcgg
- FMN_adapt 5'_3'_Rev : *tcccgctcgtcgtctg*gtggcggcatatagccaa

FMN_adpt 5':

- FMN_adapt 5'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*gcttattctcagggcgg
- FMN_adapt 5'_Rev : gtggcggcatatagccaa

FMN_adpt 3':

- FMN_adapt 3'_For: TAATACGACTCACTATAgggcttattctcagggcgg
- FMN_adapt 5'_Rev : *tcccgctcgtcgtct*ggtggcggcatatagccaa

TPP adapter (to synthesize the riboswitch with few combinations of adapters)

Template:

- TPP_For_noT7: ggactcggggtgccctt
- TPP_Rev: gtggtccgtcaagtccct

TPP_adpt 5'_3':

- TPP_adapt 5'_3'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggactcggggtgccctt
- TPP_adapt 5'_3'_Rev : *tcccgctcgtcgtctg*gtggtccgtcaagttccct

TPP_adpt 5':

- TPP_adapt 5'_For: gaaaTTAATACGACTCACTATAGGGA*gggaggacgatgcgg*ggactcggggtgccctt
- TPP_adapt 5'_Rev : gtggtccgtcaagtccct

TPP_adpt 3':

- TPP_adapt 3'_For: TTAATACGACTCACTATAGgg
- TPP_adapt 3'_Rev : *tcccgctcgtcgtctg*gtggtccgtcaagttccct

Primers complementary to adapter sequences for all tested riboswitches

- all_adpt_5' (1): ccgcatcgtcctccc
- all_adpt_3' (1): tcccgctcgtcgtctg

- Adpt_2_5' (2): gcatcgtcctccc - Adpt-2_3' (2): tcccgctcgtcgtc
Primers complementary to adapter sequences for c-di-GMP riboswitch
- Adpt_c-di-GMP 4/5 + 2nts_5'(3): ccccgcatcgtcctccc - Adpt_c-di-GMP 4/5 + 4nts_5'(3): ttccccgcatcgtcctccc - Adpt_c-di-GMP4 + 4nts_3' (3 & 4): tcccgctcgtcgtctgcaaa - Adpt_c-di-GMP5 + 4nts_3' (3 & 4): tcccgctcgtcgtctggtat
Primers complementary to adapter sequences for glycine riboswitch
- Adpt_Gly + 2nts_5' (3): acccgcatcgtcctccc - Adpt_Gly + 4nts_5'(4): caacccgcatcgtcctccc - Adpt_Gly + 4nts_3' (3 & 4): tcccgctcgtcgtctgtcct
Primers complementary to adapter sequences for FMN riboswitch
- Adpt_FMN+ 2nts_5' (3): gcccgcatcgtcctccc - Adpt_FMN + 4nts_5'(4): aagcccgcatcgtcctccc - Adpt_FMN + 4nts_3' (3 & 4): tcccgctcgtcgtctggtgg
Primers complementary to adapter sequences for glycine riboswitch
- Adpt_TPP + 2nts_5' (3): cccgcatcgtcctccc - Adpt_TPP + 4nts_5' (4): gtccccgcatcgtcctccc - Adpt_TPP + 4nts_3' (3 & 4): tcccgctcgtcgtctggtgg

The capital letters correspond to the sequence of the T7 promoter. The letter ‘’N’’ represents degenerated nucleotides. The abbreviation ‘’For’’ corresponds to forward primers, whereas ‘’Rev’’ stands for reverse primers. The underlined nucleotides correspond to the aptamer of each riboswitch. The italicized nucleotides correspond to the adapters, either in 5’ or 3’ end of the riboswitch

Table 10.2 List of constraints applied to Mfold software

Riboswitch	Constraints
Fluoride	.....xx(((((.....xxxx(((((....)))))xxxxx)))))xx.....
FMN	..(((((((..((.((((.....)))).)).((((....(((((...)))))....)))))..((.(((((((....)))))))....))...(((((.....))))) .(((.(((.....)))..)))..)))))))...
c-di_GMP I	.....((((xxxxxx((...((((((....))))))...))xxx(((.(((((((...(((.....))).)))))..))))))xxx)).)).....
Nickel-cobalt	.....(((.....((...(((.....))).....))...((.((.....)).)).)..((((.....)))).....)))..
Stem P2-P3, Library 3 TPP	.....(((((((((((((.....))))))......))))))).....

For the realization of Figure 5.3 and Supplementary Figures 10.2-10.5 and 10.7. Dots represent nucleotides with no constraints. "X" shows nucleotides that are forced to remain unbound, whereas parentheses indicate base pairs.

Table 10.3 Clones selected with the SELEX of the degenerated TPP riboswitch

Clone from	Clone number	Lib.(see Fig. 5a)	Select in band (see Fig. 5b)	K_D Thiamine (μM)	K_D TPP (μM)	Ratio K_D^T/ K_D^{TPP}
Illumina sequencing	1	1	2	17.7	67.1	0.26
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU CGCCUGGAUAAUGCCAGCCUAGGGAAGUUCACGGACCAC					
	2	1	2	16.5	18.5	0.88
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GACCUGGAUAAUGCCAGACAAGGGAAGUUCACGGACCAC					
	3	2	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUU GUUUGGAUAAUGCCAUGCUCGGGAAGUUCACGGACCAC					
	4	2	2	22.9	21	1.09
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU UGUUUGGAUAAUGCCAUGCUUGGGAAGUUCACGGACCAC					
	5	2	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCG GUUUUGGAUAAUGCCAAUGCUGGGAAGUUCACGGACCAC					
	6	2	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GCUUUGGAUAAUGCCAUGGUUGGGAAGUUCACGGACCAC					
	7	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU CGUCUGGAUAAUGCCAGCGAAGGGAAGUUCACGGACCAC					
	8	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCC GAGUUGGAUAAUGCCAGAGCAGGGAAGUUCACGGACCAC					
	9	1	0	34.12	34.55	0.99

Generation 4	RNA sequence: GGACUCGGGGUGCCCUUCUGCUGAAGGCUGAGAAAUACCCGUAUCACCU GUACUGGAUAUGCCAGAAUAGGGAAGUUCACGGACCAC					
	10	1	0	/	/	/
	RNA sequence : GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU ACACUGGAAAUGCCAGACAAGGGAAGUUCACGGACCAC					
	11	1	0	/	/	/
	RNA sequence : GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GUUCUGGAUAAUGCAGUAUAGGGAAGUUCACGGACCAC					
	12	1	1	/	/	/
	RNA sequence : GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU AGUCUGGAAAUGCCAGAUGAGGGAAGUUCACGGACCAC					
	13	1	1	2.59	9.27	0.28
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GUACUGGAUAAUGCCAGCGAGGGAAGUUCACGGACCAC					
	14	1	1	14.86	13.46	1.1
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU CUGCUGGAUAAUGCCAGGAUAGGGAAGUUCACGGACCAC					
	15	1	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUACCU GUGCUGGUAAUGCCAGUUCAGGGAAGUUCACGGACCAC					
	16	1	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU AUUCUGGAUAAUGCCAGUAAGGGAAGUUCACGGACCAC					
	17	1	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GACUGGUAAUGCCAGUAUAGGGAAGUUCACGGACCAC					
	18	2	0	58	50.4	1.15
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU					

	CGCUUGGAUAUGCCACAUGAGGGAAGUUCACGGACCAC					
	19	2	0	45	33.1	1.36
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU CGCUUGGAUAUGCCACAUGAGGGAAGUUCACGGACCAC					
	20	2	0	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GGAUUGGAUAAUGCCAAGUUGGGGAAGUUCACGGACCAC					
	21	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GUGCUGGAUAAUGCCAGAGUAGGGAAGUUCACGGACCAC					
	22	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGCUGAGAAAUACCCGUAUCACCCAU GAUGGAUAAUGCCAAUAAGGGGAAGUUCACGGACCAC					
	23	2	1	12.5	40.2	0.31
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUA AAUUGGAUAAUGCCACUAGAGGGAAGUUCACGGACCAC					
	24	2	2	17.53	20.11	0.87
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU AUAUUGGAUAAUGCCACCUAAGGGAAGUUCACGGACCAC					
	25	2	2	3.07	5.1	0.6
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACC AAGGAUGGAUAAUGCCACUAAGGGGAAGUUCACGGACCAC					
	26	2	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACC UAAACUGGAUAAUGCCAUAAAGGGGAAGUUCACGGACCAC					
	27	1	0	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUCGUGAAAGAAGAGAAAUCCCGUAUCACCUU					

Generation 10	UGCUGGAAAUGUCAGCUUAGGGAAGUUCACGGACCAC					
	28	1	0	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCCGAAGGCUGAGAAAUACCCGUAUCACCUA GACUGGAUAAUCCAGCCCAGGGAAGUUCACGGACCAC					
	29	1	0	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGUGCUGAGAAAGGAAUACCCGUAUACCUAU GCUGGAAAUGCCAGUGCAGGGAAGUUCACGGAC					
	30	1	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCUGAAGGCUGAGAAAUACCCGUAUCACCU UGUCUGUAAUGCAGUCGAGGGAAGUUCACGGACCAC					
	31	1	1	/	/	/
	RNA sequence: GGCUCGGGGUGCCCUUCUCGUGAAGCUGAGAAAUACCCGUAUCACCUAAA CUGGUAAUGCCAGUCUAGGGAAGUUCACGGACCAC					
	32	1	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGGUGAAGGUGAGAAAUCCUGUAUCACCUGAUCUGGAUAAUGC CGCCCAGGGAAGUUCACGGACCAC					
	33	1	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUACACCUACACUGGAUAAU GCCAGCUAAGGGAAGUUCACGGACCAC					
	34	1	2	69	39	1.77
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCU GAAGAUGUUUCUGCUCGGGGAAGUUCACGGACCAC					
	35	1	2	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGGUGAAGGUGAGAAAUCCUGUAUCACCUGAUCUGGAUAAUGC CGCCCAGGGAAGUUCACGGACCAC					
	36	2	0	/	/	/
	RNA sequence: GGACUCGGGGUGUCUUCCUCGUGAGGCUGAGAAUAUCCGAUACCAGACUCC					

	37	2	0	/	/	/
	RNA sequence: GGACUCGGGGUGGCGNCCUGCGUGAAGCUGAGAAUCUCGUUCUCUAAAGCCCN					
	38	2	0	/	/	/
	RNA sequence: GGACUCGGGGUGUCCUUCUGCGUGAGGCUGAGAAUCNUCGUUCACUCCCGACCCNN					
	39	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGACUUGGAUAA UGCCACUUGGGGGAAGUUCACGGACCAC					
	40	2	1	/	/	/
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGCAUCACCAGCACUGGUAAU GCCAAUACUGGGAAGUUCACGGACCAC					
	41	2	1	71	168	0.42
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUUACCG CGUGUGGAUAAUCCAGCGUCGGGAAGUUCAC					
	42	2	2	60	140	0.43
	RNA sequence :GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCA ACGAACGCUAGAACGUUGGGAAGUUCACGGACCAC					
	43	2	2	6.1	6.7	0.91
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCAC GAAUGGAUAAUGCCAGGCUGGGAAGUUCACGGACCAC					
	44	2	2	8.9	25	0.36
	RNA sequence: GGACUCGGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCC AAACUGGAUAAUGCCAUCCGGGGGAAGUUCACGGACCACAAUCAC					
Illumina sequencing	45	3	2	3.7	1000*	0.004
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCAA ACACUGCUUGUUGUUUGGGAAGUUCACGGACCAC					
	46	3	2	77	262	0.29
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCAA CAAUGGAAGAUGUGUUGGGAAGUUCACGGACCAC					
	47	3	2	5.89	1000*	0.005
	RNA sequence: GGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUA GCAACAAGUGUUGGUAGGGAAGUUCACGGACCAC					

Many sequences were assayed by in-line probing without providing conclusive results, in such cases, the K_D is indicated as "/", this is presumably due to lack of binding to the ligand or a modulation too weak to be quantified above background, as estimated by in-line probing. Also, we have assayed sequences from the initial libraries (prior to selection), but none showed modulation. For the affinity with TPP of clones 45 and 47 (represented with an asterisk, we could not measure the K_D, since no modulation was observed. We put a value of 1 mM, which was the largest tested concentration.

Table 10.4 Predicted stem formation in the random region of library 3

Seq ID	5′-3′	dG value	GC% (within 18 bp)	Final GC pairing occurs or not (position)
>22_1	CCAAGCAGAGTTCGCTGCTCGG	-7.80	56%	Yes (1,22)
>6_2	CCAAACACTGCTTGTTGTTTGG	-6.30	33%	Yes (1,22)
>6_3	CCGCGTGTTAATGGGTGCCTGG	-2.60	56%	Yes (1,22)
>5_5	CCAAACGCTGCTTGTCGTTGGG	-4.40	50%	Yes (1,21)
>5_6	CCAATGGACCATTCGGCCTTGG	-6.70	50%	Yes (1,22)
>5_4	CCTTACAGGTAGTGCTGTTGGG	-6.10	44%	Yes (1,22)
>4_11	CCAAGCATTTGCGTAGACTTGG	-2.80	39%	Yes (1,22)
>4_9	CCAGGCAACGTTTGGTACCTGG	-6.10	50%	Yes (1,22)
>4_7	CCAGTGGACCATTCGGCCCTGG	-9.10	61%	Yes (1,22)
>4_8	CCATGATTTGTTTTAGTCACGG	-4.10	28%	Yes (1,22)
>4_10	CCGACGATGGGATATGCGTTGG	-5.40	50%	Yes (1,22)
>4_13	CCTAACACTGCTTGTTGTTAGG	-6.70	33%	Yes (1,22)
>4_12	CCTCACGACAGCGTGCACCTGG	-2.90	61%	No
>3_17	CCTATGATTGCTCAGTCGTAGG	-9.50	39%	Yes (1,22)
>3_22	CCTATGCATCCGAGACACAAGG	0.30	44%	Yes (1,22)
>3_14	CCTCCATCAGACGGGATGGCGG	-8.20	61%	Yes (1,22)
>3_23	CCTCTGATGAGTGTCAGAAGGG	-8.40	44%	No
>3_28	CCTGATAGACTAGTGCTAAGGG	-1.70	39%	No
>3_18	CCTGATTAGCACTGCGAAATGG	0.30	39%	Yes (1,22)
>3_33	CCTTATGTTGGAGTAGCTAGGG	-3.50	39%	Yes (1,22)
>3_29	CCTTTTGTTGCGCAGTTTTAGG	0.00	33%	Yes (1,22)
>3_69	CCAACAATGGAAGATGTGTTGG	-4.40	33%	Yes (1,22)
>3_71	CCAACACGTGTCTGATTGTTGG	-4.30	39%	Yes (1,22)
>3_73	CCAACAGTGTCTCGCTGCTTGG	-7.10	50%	Yes (1,22)
>3_75	CCAACGAATTGATGTCGGTTGG	-3.90	39%	Yes (1,22)
>3_77	CCAAGATTCGCACCGATCTTGG	-7.90	44%	Yes (1,22)
>3_79	CCAATCATGGAAGATGGGTTGG	-4.10	39%	Yes (1,22)

>3_81	CCAATGCTCTTGACGGTGTTGG	-6.70	44%	Yes (1,22)
>3_83	CCACCAGTTTACGGCTGGCGGG	-10.00	61%	No
>3_85	CCACCTCATGGGTATGAGGGGG	-9.70	56%	Yes (1,22)
>3_87	CCACGAAGAACAGTGTCGAGGG	-2.90	50%	No
>3_89	CCACGCACTGTTGGGGTGCTGG	-8.20	61%	Yes (1,22)
>3_91	CCACGGAATGGCTAACCGCAGG	-1.60	56%	Yes (1,22)
>3_93	CCACTCATGGCTCATGTGTTGG	-3.70	44%	Yes (1,22)
>3_95	CCACTGATTCTTGTTGAAAGGG	-1.10	53%	No
>3_97	CCAGACAGGATGCTGTCATGGG	-9.60	61%	No
>3_99	CCAGACTGCTGTTGCCGCCTGG	-5.30	70%	Yes (1,22)
>3_101	CCAGATCTACCGTTGGATTTGG	-8.50	57%	Yes (1,22)
>3_103	CCAGCCCTACTTTGTCATATGG	0.50	56%	Yes (1,22)
>3_105	CCAGCGGCTGCTGCATGCCTGG	-6.60	74%	Yes (1,22)
>3_107	CCAGGCCACTTCTGTTACCTGG	-7.50	50%	Yes (1,22)
>3_109	CCAGGCTGGAGCCCTTGCTTGG	-8.40	61%	Yes (1,22)
>3_111	CCAGGTGCGGTCTAAAACCAGG	-2.60	50%	Yes (1,22)
>3_113	CCCAAATGCCAATTGGCCAGGG	-6.50	50%	Yes (1,22)
>3_115	CCCCATGCGTTACGTTTGGAGG	-5.70	50%	Yes (1,22)
>3_117	CCCCTACAAGAGACTGCTGGGG	-6.80	56%	Yes (1,22)
>3_119	CCCTACCATCTGCTTGCGAGGG	-4.70	56%	Yes (1,22)
>3_121	CCCTACCTGAGCTAGTTAAGGG	-2.60	44%	Yes (1,22)
>3_123	CCCTTGCCGCTGGTATGTCGGG	-3.70	61%	Yes (1,22)
>3_125	CCCTTGGACACTTTCCGGGGGG	-12.20	61%	No
>3_127	CCGAAGGCTTGTGCTTAGTAGG	-3.20	44%	Yes (1,22)
>3_129	CCGCTTGATCCGTATATGCCGG	-0.90	50%	No
>3_131	CCGGACAGCACTCGCTACCCGG	-8.60	67%	Yes (1,22)
>3_133	CCGGAGCATTCCTGTGCCCCGG	-10.80	67%	Yes (1,22)
>3_135	CCGGCTGTTCTGTACCGCCCGG	-5.60	67%	Yes (1,22)
>3_137	CCGGGATCGCTCGACGCCCCGG	-8.30	78%	Yes (1,22)
>3_139	CCGTCAGCCACGTCGCTTAAGG	-2.70	56%	Yes (1,22)
>3_141	CCGTTGATGCTTTCGCAACCGG	-3.30	50%	Yes (1,22)
>3_143	CCTAAGCCCATCCAACCTTAGG	-3.90	44%	Yes (1,22)
>3_145	CCTAAGCGTTGTCGCGCCTGGG	-8.10	61%	Yes (1,22)
>3_147	CCTAGATGCTCTTAGCATGGGG	-5.80	44%	Yes (1,22)
>3_149	CCTAGCGGTTTGCAAAGTCGGG	-2.60	50%	Yes (1,22)

>3_151	CCTAGTTCTGGATCGCACTGGG	-4.30	50%	Yes (1,22)
>3_153	CCTATCGCTGAGTGCGCCTGGG	-5.60	61%	Yes (1,22)
>3_155	CCTCCACGCTCGGATTACCCGG	-3.00	61%	No
>3_157	CCTCTTATTTCACCGAAGGAGG	-4.50	39%	Yes (1,22)
>3_159	CCTGAACCAGACGCGAATTTGG	-2.50	44%	No
>3_161	CCTGAAGGTAGGAATACTAGGG	-3.60	39%	No
>3_163	CCTGAATGCACGTAGCTCATGG	-2.10	44%	Yes (1,22)
>3_165	CCTGACCATCTGGGTTGTTGGG	-5.40	50%	Yes (1,22)
>3_167	CCTGACTTGCGTGTACCTTGGG	-1.50	50%	Yes (1,22)
>3_169	CCTGAGATTTGCCGTCTCGAGG	-6.50	50%	Yes (1,22)
>3_171	CCTGCGAATGCGCGTTCATGGG	-4.90	56%	Yes (1,22)
>3_173	CCTGCTCGTTGGTAGTGCCAGG	-3.10	56%	No
>3_175	CCTGGACGTTTTACACGCTGGG	-6.20	50%	Yes (1,22)
>3_177	CCTGGTGTTTGGCTATGCTGGG	-6.50	50%	Yes (1,22)
>3_179	CCTGTCGATGTTCGTCGATGGG	-10.70	50%	Yes (1,22)
>3_181	CCTGTTAGATGTGCTCTACAGG	-6.40	39%	Yes (1,22)
>3_183	CCTGTTCACGGGCGTGCGTAGG	-6.60	61%	Yes (1,22)
>3_185	CCTTCACTATTGCTGGGATGGG	-3.20	44%	No
>3_187	CCTTCCACACTGCCGCACCTGG	0.50	61%	No
>3_189	CCTTCCGGAGAGGAGCCATCGG	-3.00	61%	No
>3_191	CCTTGCACAATGCTGTGCGGGG	-12.00	56%	Yes (1,22)
>3_193	CCTTGTCATCTGCTGTGCTGGG	-1.80	50%	Yes (1,22)
>3_195	CCTTGTTGCAATTGGCGCAGGG	-7.40	50%	Yes (1,22)
>3_197	CCTTTGCGCGTTGGGTGCCTGG	-5.00	61%	Yes (1,22)

Random Seq No	Random Seq (5'-3')	dG	GC% (within 18 bp)	Final GC pairing occurs or not (position)
1.	ccTTCCCTTGCATATATGTTgg	-0.90	33%	No
2.	ccACATTTCTTCAACCTTCTgg	-0.20	33%	No
3.	ccAATTGCACCCTTAGGACGgg	-2.60	50%	No
4.	ccAAGACAGATATGTTCTTAgg	-3.00	28%	Yes (1, 22)
5.	ccCCTATATTTCATCATTGGgg	-5.00	33%	Yes (1,22)
6.	ccCAACGGGATCGCATGTCCgg	-5.90	61%	No
7.	ccCACGTAAAACATTGTTAAgg	1.00	28%	No
8.	ccACCCTCAGGTTTTTGAGCgg	-4.90	50%	No
9.	ccGACAAAAACTTTAAAAAGgg	0.30	22%	No

10.	ccAAATTCGCGCTCATAACTgg	1.10	39%	Yes (1,22)
11.	ccGTTAGGCCACGATTGCGTgg	-5.90	56%	No
12.	ccGAGTTTCGGCCCTGTGCTgg	-3.50	61%	No
13.	ccGCGCTGTATAGCCGATTCgg	-3.40	56%	No
14.	ccTCATTCGGGCCTTATATCgg	-0.60	44%	No
15.	ccTGGAAACCCCAACCTATTgg	-1.70	44%	No
16.	ccTAGACAGCATCATTGGCCgg	0.10	50%	No
17.	ccGAAGTTATTGGGCATATTgg	-2.00	33%	No
18.	ccCACCGTAAAGTCCTCCTCgg	-1.00	56%	No
19.	ccGGGCGTCCCTCCTTTAAAgg	-3.30	56%	No
20.	ccAGATGATAAGCTCCGGCAgg	-1.50	50%	No
21.	ccAAGGATCGGTGATATTAAgg	0.50	33%	No
22.	ccCAAAGATTCGGCACATTAgg	0.10	39%	No
23.	ccCTCTTGTTGGTGTGGTATgg	-0.30	44%	No
24.	ccCGCTTAACTGCGTGGCGGgg	-10.20	67%	No
25.	ccAGCCTTATGGCAAAATCGgg	-5.00	44%	No
26.	ccTTCGGGAATGATTCTGGTgg	-5.80	44%	Yes (1,22)
27.	ccAACGCTAAAGGTCCATAGgg	0.40	44%	No
28.	ccCACATACATCGCAACCTGgg	-2.00	50%	Yes (1,22)
29.	ccGCATGCGTTCAATTTGACgg	-2.00	44%	No
30.	ccGATCGCTTGGCGCTAAGAgg	-1.10	56%	No
31.	ccTTAAAGCGGCTGCACTGCgg	-4.30	56%	No
32.	ccTGTAAGGACGATTACGGAgg	-5.00	44%	No
33.	ccGTGGGCGGCCTGGGGGGAgg	-5.00	83%	No
34.	ccGCACTACCCCATCGACCTgg	-0.60	61%	No
35.	ccGTACAGGAACACTCTATAgg	-1.00	39%	No
36.	ccTTGCTCTCAGACGAACAAgg	-5.70	44%	Yes (1,22)
37.	ccATTACTAGAGTGCCGCTTgg	-0.50	44%	No
38.	ccTCAGCCCCCCTGTCGTCGgg	-3.00	72%	No
39.	ccCGTTGTTGTGATTCGACTgg	-3.70	44%	Yes (1,22)
40.	ccCTATTGAGGCATCAACTGgg	-6.60	44%	Yes (1,22)
41.	ccAATGAATCGGCCTATGTCgg	-3.00	44%	No
42.	ccCCCGATGTCGTTAGTGAAgg	0.20	50%	No
43.	ccGGTTCCGACGCATACCTCgg	-2.80	61%	No
44.	ccCTTCGTTGAGAACCCACAgg	0.00	50%	No

45.	ccATCATACAACTGGGGACAgg	-5.80	44%	No
46.	ccTAATCCCTACGCCCATCAgg	0.40	50%	No
47.	ccTCTACACGCGTCTCTGTGgg	-3.40	56%	No
48.	ccGCTCCAGTTCATGTGCTGgg	-6.80	56%	No
49.	ccGGAGAGCACCCTCCACAAgg	-5.90	61%	No
50.	ccGGTCTAGTGGTATGGTGGgg	-3.10	56%	No
51.	ccTGATACACGCGGCAGGGGgg	-3.60	67%	No
52.	ccTAGGACCATCGGTAGTAGgg	-6.80	50%	No
53.	ccCTGACCACTGCCTATAGGgg	-4.20	56%	No
54.	ccAGAGTGTCAGCCAGTGTAgg	0.40	50%	No
55.	ccACCCCACGAGGATCCGAGgg	-5.00	67%	No
56.	ccAAGGCGAACCGGGCCAGAgg	-4.30	67%	No
57.	ccCTCAAAGCCGCGGCGCGAgg	-7.60	72%	No
58.	ccAGTAGCCCCGGGGTGAACgg	-4.70	67%	Yes (1,22)
59.	ccACCTATGGGGCTGGATAAgg	-4.40	50%	No
60.	ccAACTGCCCTGGTGAGCGCgg	-4.10	67%	No
61.	ccTCTGCTGCTCGAGGCCGTgg	-4.00	67%	No
62.	ccTCGCCGATGCTTGCTGCGgg	-3.90	67%	No
63.	ccTCCCCAGCCGCTACATCTgg	-1.90	61%	No
64.	ccGTCTCTTTGCCGACTAATgg	-1.90	44%	No
65.	ccGCGAACAACCACACCATAgg	0.30	50%	No
66.	ccGCGATTCGTCGGGGCGCCgg	-5.20	78%	No
67.	ccTCGGAATACGGTATGGGCgg	-1.60	56%	No
68.	ccTCGCGGACGCCAGGCATCgg	-4.50	72%	No
69.	ccGTGCAGGTAGCGGAGGCCgg	-4.20	72%	No
70.	ccCGCACGCGAGACGAACTGgg	-2.90	67%	Yes (1,22)

We observed that selected sequences could form stems within the randomized region. To evaluate whether it occurred much more often than it would for random sequences, we evaluated the average delta-G of structures for that portion of sequence for random sequences (as would be found in the initial library, prior to selection with SRPAGE) vs our selected sequences. Random sequences (18nt in size, with additional cc in 5' and gg in 3') were created using the "Random DNA Sequence Generator" http://www.faculty.ucr.edu/~mmaduro/random.htm with an input of seven different values of GC percentage in the sequences (0.35, 0.40, 0.45, 0.5, 0.55, 0.60 and 0.65) and to evaluate its folding nature through structure formation, whether dG value and terminal GC base pairs occurred or not. Folding was performed with Mfold. It is noteworthy that some of our sequenced samples selected from library 3 were actually contaminations from library 2 (sequences which started dominating the population of library 2 earlier, given its smaller complexity than library 3). These sequences were excluded from the above analysis (given their different size and larger number of fixed nucleotides, they were easily recognizable, a total of 625,731 reads corresponding to the correct size for library 3 were obtained from selected pools), but the fact that we selected them as contaminants from library 2 actually supports our interpretation that using SR-PAGE, we selected for the presence of a stem at a position roughly equivalent to P4 and P5.

11 ANNEXE IV: DÉVELOPPEMENT D'OUTILS DE RÉGULATION GÉNÉTIQUE BASÉS SUR LES RIBOZYMES SYNTHÉTIQUES CHEZ METHYLORUBRUM EXTORQUENS (MATÉRIELS ET MÉTHODES)

11.1 Préparation de la cible (gène *crtI*)

La séquence d'ADN cible (gène *crtI*) a été construite par assemblage PCR avec les amorces *crtI*_F1, *crtI*_R1, *crtI*_F2 et *crtI*_R2 (Tableau 11.1). La cible de 186 nucléotides (incluant le promoteur T7) correspond à une portion de la séquence du gène *crtI* de la bactérie *M. extorquens* AM1 (Tableau 11.1). L'outil Primerize a été utilisé pour la conception des amorces pour l'assemblage PCR (Tian *et al.*, 2015). Primerize est accessible sur ce site Web : https://primerize.stanford.edu/.

Tableau 11.1 Séquences des oligonucléotides utilisés pour l'assemblage PCR du fragment du gène *crtI*

Oligonucléotides	Séquences
*crtI*_F1	**taatacgactcactatag**gattcggtgcgcacctaccgcgagctcgtcggcggcga
*crtI*_R1	ccggagcaggccggctcgaaggacttcttctcgtaggacttgccgacctcgccgccgacg
*crtI*_F2	cggcctgctccggcgtggtgctctatctcggcctgaacaagcgctacgagcacctgaacc
*crtI*_R2	ctcctccgggtcgcgggagaacacgaaatcgtggtggttcaggtgctcgtagcgctt
Séquence du gène *crtI*	**taatacgactcactatag**gattcggtgcgcacctaccgcgagctcgtcggcggcgaggtcggcaagtcctacgag aagaagtccttcgagccggcctgctccggcgtggtgctctatctcggcctgaacaagcgctacgagcacctgaacc accacgatttcgtgttctcccgcgacccggaggag

* Le promoteur T7 est représenté en gras.

Un ARN cible marqué à la radioactivité a été synthétisé par transcription *in vitro* en utilisant le fragment du gène *crtI* assemblé par PCR comme matrice. La réaction de 100 µL contenant 20 uL d'ADN, 80 mM HEPES-KOH (pH 7,5) 24 mM $MgCl_2$, 0,5 U de pyrophosphatase de levure (Roche Diagnostic), 4 µL de polymérase T7 (au moins 25 U/µL), 5 mM de NTPs (GTP, CTP et ATP), 0,02 mM d'UTP, 5 µCi α-UT ^{32}P (Perkin Elmer) et 2 mM de spermidine a été incubée à 37 °C pendant 2 h 30. Afin de dégrader les fragments d'ADN non transcrit, 2 U de DNase I (NEB) sont ajoutés à la réaction pendant 30 min à 37 °C (Milligan et al. 1987). L'échantillon d'ARN radioactif a été précipité avec de l'éthanol 100 % (2 v/v) et de l'acétate de sodium 3M (0,1 v/v) pendant au

moins 2 heures à -20 °C. Après le lavage avec de l'éthanol à 70 %, l'ARN transcrit a été purifié sur un gel de polyacrylamide dénaturant 6 % (8 M urée) (PAGE, rapport 19:1 de l'acrylamide au bisacrylamide) dans du tampon TBE (45 mM Tris-borate pH 7,5 et 1 mM EDTA). L'échantillon a migré sur gel avec un volume égal de colorant bleu dénaturant (formamide 95 %, EDTA 10 mM, 0,05 % de xylène de cyanol et 0,05 % de bleu de bromophénol). Le gel a été exposé sur un écran phosphoré pendant 5 min et visualisé avec l'instrument Typhoon FLA9500 (GE Healthcare Life Sciences). Le transcrit radiomarqué a été découpé dans le gel et élué pendant la nuit à 4 °C sur un rotor avec un tampon d'élution (0,3 M NaCl).

11.2 Préparation des ribozymes ciblant le gène *crtI*

Ribosoft 2.0, un service web accessible au public a été utilisé pour concevoir des ribozymes hammerheads ciblant le gène *crtI* (Kharma *et al.*, 2016). Il suffit d'insérer la séquence ciblée dans l'interface web (Tableau 11.1) afin de générer une liste de ribozymes *hammerheads*. Ribosoft 2.0 est accessible sur ce site : https://Ribosoft 2.02.fungalgenomics.ca/.

Huit ribozymes *hammerheads* ont été sélectionnés afin d'être testés d'abord *in vitro*. La séquence complémentaire de chaque ribozyme contenant le promoteur T7 a été commandée sous forme d'ADN simple brin (IDT) (Table 11.2).

Tableau 11.2 Séquences des ribozymes ciblant le gène *crtI*

Oligonucléotides	Séquences
Rbz1_*crtI*	ctacgagaagaagtttcgtcgcatttcagcgactcatcagcttcgattagccggcctg**ctatagtgagtcgtatta**
Rbz2_*crtI*	aggtcggcaagtttcgtcgcatttcagcgactcatcagctacgattagaaga**ctatagtgagtcgtatta**
Rbz3_*crtI*	gctcgtttcgtcgcatttcagcgactcatcagggcggattcgaggtc**ctatagtgagtcgtatta**
Rbz4_*crtI*	cgcgagctcgtttcgtcgcatttcagcgactcatcagggcggattcgaggtcgg**ctatagtgagtcgtatta**
Rbz5_*crtI*	gaagtttcgtcgcatttcagcgactcatcagcttcgattagccggcct**ctatagtgagtcgtatta**
Rbz6_*crtI*	caagtttcgtcgcatttcagcgactcatcagctacgattagaagaagt**ctatagtgagtcgtatta**
Rbz7_*crtI*	gcgaggtttcgtcgcatttcagcgactcatcagggcaaattgtc**ctacgactatagtgagtcgtatta**
Rbz8_*crtI*	cctacgagaagaagtttcgtcgcatttcagcgactcatcagcttcgattagccggcctg**ctatagtgagtcgtatta**

* Le promoteur T7 est représenté en gras.

Les oligonucléotides simple-brin ont été utilisés pour créer des séquences d'ADN double brin pour chaque ribozyme en effectuant cinq cycles de PCR avec 1 µM de la séquence du promoteur T7 comme amorce. La polymérase TAQ avec son tampon commercial et 200 µM de dNTPs ont été utilisés à une température d'hybridation de 55 °C. La taille de chaque ribozyme a été confirmée sur un gel d'agarose 2 % avec du bleu natif 6X (40 % sucrose, 0,05 % bleu de bromophénol et 0,05 % xylène cyanol) pour la migration. Chaque ribozyme a été transcrit *in vitro* dans une réaction de 100 uL comme décrit plus haut, mais cette fois avec la même concentration d'UTP utilisée que pour les autres nucléotides, car il n'y a pas de radioactivité. Les réactions de transcription ont été précipitées pendant la nuit et purifiées sur un PAGE dénaturant comme décrit précédemment. Les transcriptions ont été visualisées sur gel avec une lumière UV. Les bandes correspondantes ont été coupées sur le gel et éluées comme mentionné précédemment.

11.3 Essai *in vitro* de l'activité catalytique des ribozymes *hammerheads*

Chaque ribozyme (5 pmol) a été pré-incubé avec le tampon de clivage (100 mM NaCl, 25 mM KCl, 50 mM Tris-HCl pH 7) à 85 °C pendant 5 minutes, avant d'être refroidi sur glace. Les ribozymes ont été incubés pendant 1 h à 37 °C avec 2 uL du transcrit du gène *crtI* marqué à la radioactivité et 10 mM de $MgCl_2$. La réaction de clivage a été arrêtée avec l'ajout d'un volume égal de colorant bleu dénaturant. Comme témoin, un échantillon ne contenant aucun ribozyme a également été réalisé. Toutes les réactions de clivage ont été purifiées sur un PAGE 10 % dénaturant 8 M urée. Le gel a été exposé sur un écran phosphoré toute la nuit et visualisé avec le Typhoon FLA9500 comme mentionné précédemment. L'efficacité de clivage de chaque ribozyme a été calculée en quantifiant l'intensité des bandes radioactives à l'aide du logiciel ImageJ. Le pourcentage de clivage a été mesuré en divisant l'intensité de la bande correspondant au produit de clivage par le total de toutes les bandes du même puit, y compris la cible pleine longueur.

11.4 Constructions des plasmides contenant les ribozymes *hammerheads* ciblant le gène *crtI*

Le plasmide pCM110-GFP (Marx & Lidstrom, 2001) contenant une résistance à la tétracycline a été utilisé comme base pour la construction de nos plasmides contenant les ribozymes ciblant le gène *crtI*. Les quatre ribozymes présentant la meilleure efficacité de clivage *in vitro* ont été sélectionnés pour être testés *in vivo*, soit les ribozymes 1, 4, 7 et 8 (Figure 6.5). Un ribozyme ciblant la protéine RFP (protéine fluorescente rouge) a également été utilisé comme témoin négatif ne ciblant pas *crtI*, car son efficacité a déjà été démontrée *in vivo* chez *E. coli* (Najeh,

2017). Les ribozymes ont été conçus pour être sous le contrôle du promoteur P_{mxaF} et du terminateur transcriptionnel T1/TE. Le promoteur P_{mxaF} a été amplifié à partir du plasmide pCM110-GFP à l'aide des amorces AflII- P_{mxaF}_F et P_{mxaF}_R (Tableau 11.3) avec la Hot Start Taq polymérase (NEB) et son tampon correspondant à une température d'hybridation de 47 °C. Un site de restriction reconnu par l'endoribonucléase AflII (NEB) a été ajouté à la séquence. Chacun des ribozymes (1, 4, 7, 8 et le contrôle RFP) a été amplifié avec leur paire d'amorces correspondante (Tableau 11.3) à une température d'hybridation de 50 °C avec la Hot Start Taq polymérase (NEB). Les amorces sens contiennent un chevauchement avec l'extrémité 3' du promoteur P_{mxaF}, tandis que les amorces antisens incluent un chevauchement avec l'extrémité 5' du terminateur T1/TE. Les produits PCR d'ADN double brin pour chaque ribozyme obtenu précédemment (décrit dans la section 11.2) ont été utilisés comme matrice. Le terminateur T1/TE a été amplifié à partir du plasmide pRI02 construit dans notre laboratoire lors d'une étude précédente (Imane, 2016) avec les amorces T1/TE_F et T1/TE_R (Tableau 11.3) à une température d'hybridation de 51°C avec la polymérase Hot Start Taq (NEB). Un site de restriction reconnu par l'endoribonucléase BlpI (NEB) a été ajouté à la séquence. La taille de toutes les séquences d'ADN a été vérifiée sur un gel d'agarose à 2%.

Tableau 11.3 Oligonucléotides utilisées pour la construction des plasmides exprimant les ribozymes ciblant le gène *crtI*

Amplicon	Oligonucléotides	Séquences
Promoteur P_{mxaF}	AflII-P_{mxaF}-F	cttaagcttcccgcttggtcg
	P_{mxaF}-R	ctatattttctaggctttgattgga
Ribozyme 1	RBz1_P_{mxaF}_F	**tcgcgtccaatcaaagcctagaaaatatag**caggccggctaatcga
	RBz1_Term_R	atttgatgcctggctcgagcctacgagaagaagtttcgtcg
Ribozyme 4	RBz4_P_{mxaF}_F	**tcgcgtccaatcaaagcctagaaaatatag**ccgacctcgaatccg
	RBz4_Term_R	atttgatgcctggctcgagccgcgagctcgtttcgt
Ribozyme 7	RBz7_P_{mxaF}_F	**tcgcgtccaatcaaagcctagaaaatatag**tcgtaggacaatttgccc
	RBz7_Term_R	atttgatgcctggctcgagcgcgaggtttcgtcgc
Ribozyme 8	RBz8_P_{mxaF}_F	**tcgcgtccaatcaaagcctagaaaatatag**caggccggctaatcg
	RBz8_Term_R	atttgatgcctggctcgagccctacgagaagaagtttcgt
Ribozyme contrôle	RBzCtrl_P_{mxaF}_F	**tcgcgtccaatcaaagcctagaaaatatag**ttctgcattaaaaccgctga
	RBzCtrl_Term_R	atttgatgcctggctcgagccgtccgacggtttcg
Terminateur T1/TE	T1/TE_F	gctcgagccaggcat
	T1/TE_R	gctcagctataaacgcagaaaggccca
Construction finale	AflII-P_{mxaF}_extra_F	aaaaaacttaagcttcccgcttggtcg
	T1/TE_BlpI_extra_R	tttttgctcagctataaacgcagaaaggccca

Séquence chevauchant le promoteur P_{mxaF}; Séquence chevauchant le terminateur T1/TE; Site de restriction AflII et BlpI

La méthode d'assemblage par Gibson a été utilisée afin de construire les fragments d'ADN contenant le promoteur P_{mxaF}, les ribozymes et le terminateur T1/TE. Les trois amplicons (0,5 pmol chacun) ont été incubés à 50 °C pendant 1 h avec un volume égal du mix d'assemblage Gibson (NEB Builder Hifi DNA Assembly). Pour amplifier les constructions finales de chaque insert, les amorces AflII-P_{mxaF}_extra_F et T1 /TE_BlpI_extra_R (Tableau 11.3) ont été utilisées à une température d'hybridation de 56 °C avec la Hot Start Taq polymérase (NEB). Ces amorces ajoutent six nucléotides (AAAAAA) aux deux extrémités de l'insert pour permettre un meilleur clivage par les endoribonucléases. Chacun des inserts et le plasmide pCM110-GFP ont été digérés avec les enzymes de restriction AflII (NEB) et BlpI (NEB) selon le protocole du fabricant avec le tampon *cut smart* (NEB). Le plasmide digéré a été purifié sur gel d'agarose 1 % avant d'être ligué avec les inserts en utilisant l'ADN ligase T4 (NEB). Un ratio 3:1 insert/plasmide a été utilisé pour la liaison. Les plasmides formés ont été transformés dans des *E. coli* DH5α compétentes par choc thermique. Les clones ont été sélectionnés sur des géloses de LB avec comme outil de sélection de la tétracycline (15 µg/mL). Tous les plasmides ont été vérifiés par séquençage Sanger (Centre d'expertise et de services Génome Québec) (Tableau 11.4).

Tableau 11.4 Séquences des plasmides confirmées par séquençage Sanger

Plasmides	Séquences
pRbz1	cttaag**cttcccgcttggtcgggccgcttcgcgagggcccgttgacgacaacggtgcgatgggtcccggccccggtcaagacg atgccaatacgttgcgacactacgccttggcacttttagaattgccttatcgtcctgataagaaatgtccgaccagctaaagacatc gcgtccaatcaaagcctagaaaatatag**caggccggctaatcgaagctgatgagtcgctgaaatgcgacgaaacttcttctcgtaggct cgagccaggca*tcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgtttttgtcggtgaacgctctc*tactagagtca cactggctcaccttcgggtgggcctttctgcgtttatagctgagc
pRbz4	cttaag**cttcccgcttggtcgggccgcttcgcgagggcccgttgacgacaacggtgcgatgggtcccggccccggtcaagacg atgccaatacgttgcgacactacgccttggcacttttagaattgccttatcgtcctgataagaaatgtccgaccagctaaagacatc gcgtccaatcaaagcctagaaaatatag**ccgacctcgaatccgccctgatgagtcgctgaaatgcgacgaaacgagctcgcggctcg agccaggcatcaaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgtttttgtcggtgaacgctctctactagagtcaca ctggctcaccttcgggtgggcctttctgcgtttatagctgagc
pRbz7	cttaag**cttcccgcttggtcgggccgcttcgcgagggcccgttgacgacaacggtgcgatgggtcccggccccggtcaagacg atgccaatacgttgcgacactacgccttggcacttttagaattgccttatcgtcctgataagaaatgtccgaccagctaaagacatc gcgtccaatcaaagcctagaaaatatag**tcgtaggacaatttgccctgatgagtcgctgaaatgcgacgaaacctcgcgctcgagcca ggcat*caaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgtttttgtcggtgaacgctctc*tactagagtcacactggct caccttcgggtgggcctttctgcgtttatagctgagc
pRbz8	cttaag**cttcccgcttggtcgggccgcttcgcgagggcccgttgacgacaacggtgcgatgggtcccggccccggtcaagacg atgccaatacgttgcgacactacgccttggcacttttagaattgccttatcgtcctgataagaaatgtccgaccagctaaagacatc gcgtccaatcaaagcctagaaaatatag**caggccggctaatcgaagctgatgagtcgctgaaatgcgacgaaacttcttctcgtagggc tcgagccaggcat*caaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgtttttgtcggtgaacgctctc*tactagagtca cactggctcaccttcgggtgggcctttctgcgtttatagctgagc
pRbzCtrl	cttaag**cttcccgcttggtcgggccgcttcgcgagggcccgttgacgacaacggtgcgatgggtcccggccccggtcaagacg atgccaatacgttgcgacactacgccttggcacttttagaattgccttatcgtcctgataagaaatgtccgaccagctaaagacatc gcgtccaatcaaagcctagaaaatatag**ttctgcattaaaaccgctgatgagtcgctgaaatgcgacgaaaccgtcggacggctcgag ccaggcat*caaataaaacgaaaggctcagtcgaaagactgggcctttcgttttatctgtttttgtcggtgaacgctctct*actagagtcacactg gctcaccttcgggtgggcctttctgcgtttatagctgagc

Site de restriction AflII et BlpI, **promoteur P_{mxaF}**, séquence du ribozyme, Terminateur *T1*/TE

11.5 Souches bactériennes et conditions de culture

Methylorubrum extorquens ATCC 55366 (Bourque et al. 1992) est cultivé à 30 °C et 250 rpm (rotation par minute) dans un erlenmeyer avec encoche de 250 mL dans 35 mL de milieu CHOI 4, comme décrit dans (Bourque *et al.*, 1995). Une concentration de 1 % (v/v) de méthanol a été ajoutée au milieu de culture comme source de carbone. La bactérie *Escherichia coli* DH5α utilisée pour la construction des plasmides a été cultivée dans le milieu Luria Broth (LB) à 37 °C avec

agitation de 250 rpm. L'antibiotique tétracycline (15 µg/mL) a été ajouté au milieu pour sélectionner les colonies dans lesquelles les plasmides étaient transformés avec succès.

11.6 Électroporation de plasmides dans *M. extorquens*

Des bactéries *M. extorquens* électro-compétentes sont préparées comme décrit dans (Figueira *et al.*, 2000). Un volume de 100 µL de cellules électro-compétentes est mélangé avec 500 ng de plasmides dans une cuvette d'électroporation (Fisherbrand ™ Electroporation Cuvette plus ™, 1 mm gap) et incubé sur glace pendant 5 min. L'électroporation est réalisée avec un Gene Pulser Xcell ™ (Biorad) avec les paramètres suivants : 2400 V, capacité de 25 µF et résistance de 200 Ω. Après l'électroporation, un mL de milieu CHOI 4 contenant 1 % (v/v) de méthanol est rapidement ajouté aux cellules et celles-ci sont laissées en incubation pendant une nuit à 30 °C sans sélection par antibiotique. Des géloses de CHOI 4 agar avec tétracycline (15 µg/mL) et 1 % de méthanol sont ensuite utilisées afin de sélectionner les colonies dans lesquelles le plasmide a été transformé avec succès.

11.7 Extraction des caroténoïdes

Les caroténoïdes sont extraits en suivant le protocole décrit par (Zhu *et al.*, 2021). La croissance bactérienne est mesurée par densité optique à l'aide su spectrophotomètre afin de normaliser les caroténoïdes extraits (densité optique de 600 nm). Les bactéries sont précipitées et le culot bactérien est resuspendu dans 1 mL de méthanol à 65 °C avant d'être mélangé par vortex pendant une minute. Ensuite, 400 µL d'eau et 500 µL sont ajoutés aux échantillons et mélangés par vortex pendant 10 minutes. Ceux-ci sont ensuite centrifugés pendant 10 minutes à la vitesse maximale. La phase organique inférieure qui contient les caroténoïdes est extraite vers un nouveau tube où 2 mL d'acétone y sont ajoutés. Les échantillons sont laissés à -20 °C pour la nuit. Ils sont ensuite laissés à évaporer. Les caroténoïdes sont finalement resuspendus dans 1 mL de méthanol et quantifiés à l'aide d'un spectrophotomètre à 494 nm.