RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à l'Université Toulouse III - Paul Sabatier

Régulation de la double activité de dégradation de l'ARN de la RNase J par les modifications épi-transcriptomiques et les structures secondaires

---

Thèse présentée et soutenue, le 4 octobre 2024 par

Tomas CAETANO

École doctorale
BSB - Biologie, Santé, Biotechnologies

Spécialité
Bio-informatique, Génomique et Biologie des systèmes

Unité de recherche
LMGM - Laboratoire de Microbiologie et Génétique Moléculaires

Thèse dirigée par
Gwennaele FICHANT et Roland BARRIOT

Composition du jury
Mme Anna NIARAKI, Présidente, Université Toulouse III - Paul Sabatier
Mme Marie-France SAGOT, Rapporteure, INRIA Lyon
Mme Anaïs LE RHUN, Rapporteure, INSERM Nouvelle-Aquitaine
M. Pierre NICOLAS, Rapporteur, INRAE Île-de-France Jouy-en-Josas
Mme Christine GASPIN, Examinatrice, INRAE Occitanie-Toulouse
M. Philippe BOULOC, Examinateur, CNRS Île-de-France Gif-sur-Yvette
Mme Gwennaele FICHANT, Directrice de thèse, Université Toulouse III - Paul Sabatier
M. Roland BARRIOT, Co-directeur de thèse, Université Toulouse III - Paul Sabatier

Remerciements

Je tiens à exprimer ma plus profonde gratitude à toutes les personnes qui ont contribué à la réalisation de ce travail. En premier lieu, je remercie chaleureusement mes directeurs de thèse, Roland et Gwennaele, pour leur soutien indéfectible, leurs conseils éclairés, et leur patience tout au long de cette aventure scientifique. Leur expertise et leurs encouragements ont été essentiels pour mener à bien ce projet.

En particulier, merci à Gwennaele pour les innombrables recommandations littéraires et livres prêtés et pour les discussions/critiques pendant les pauses café. Un merci spécial à mon co-directeur Roland, qui a l'immense mérite de m'avoir supporté pendant ces 3 ans de thèse mais aussi les 6 mois de stage, et les 2 ans de master. Sans lui je n'aurais jamais été aussi loin, j'ai énormément appris à ses côtés et j'ai eu beaucoup de chances de l'avoir comme encadrant. Je souhaite également remercier les autres membres de mon équipe de recherche, Yves et Petra, pour leur collaboration précieuse, leurs discussions stimulantes et leur esprit d'équipe, qui ont rendu ce travail encore plus enrichissant.

Je tiens aussi à exprimer ma gratitude à l'équipe Redder qui m'accompagne depuis le tout début et qui a fourni les données essentielles à ce travail. En particulier, je souhaite remercier Peter pour son soutien continu, sa disponibilité et ses précieux conseils tout au long de ce projet.

Mes remerciements vont également aux rapporteurs de cette thèse, Marie-France Sagot, Anaïs Le Rhun et Pierre Nicolas, pour le temps qu'ils ont consacré à l'évaluation de mon travail, ainsi que pour leurs remarques constructives qui ont permis d'améliorer ce manuscrit. Merci également aux autres membres de mon jury, Christine Gaspin, Anna Niarakis et Philippe Bouloc, qui à l'heure où j'écris ces mots n'ont pas encore lu le manuscrit mais qui, j'en suis convaincu, contribueront à enrichir les discussions par leurs questions intéressantes et bienveillantes.

Merci également aux membres de mon comité de thèse, Cecile Albenne, Matthias Zytnicki et Olivier Cuvier pour leur temps, leur expertise et leurs conseils avisés. Leurs remarques constructives et leur soutien ont été d'une grande aide pour orienter et améliorer mon travail.

Je souhaite désormais exprimer ma reconnaissance à ma famille, dont le soutien et les encouragements ont été une source de force durant l'entièreté de mon parcours. En premier lieu, merci à ma mère pour être la femme la plus inspirante que je connaisse. Sa curiosité sans limite et sa soif de mouvement l'ont poussé à faire des dizaines de métiers et à avoir des milliers de vies. Elle m'a transmis des richesses inestimables, et son sourire, son optimisme et son amour (et bien d'autres qualités qui me font défaut), m'encouragent tous les jours à être une meilleure personne. Elle m'a toujours soutenu et guidé dans mes choix et mes aspirations alors merci pour tout, tout ce que tu as dit, tout ce que tu fais et tout ce que tu es. Merci ensuite à Cécile pour m'avoir accueilli dans sa vie de manière si bienveillante et si naturelle. Merci aussi pour toutes les activités que nous avons faites ensemble, qui ont éveillé ma curiosité et ensoleillé mes week-ends.

Merci à mon père, qui n’a jamais, ne serait-ce qu’une seconde, douté de ma capacité à accomplir tout ce que je souhaite entreprendre à condition que ça contribue à mon bonheur. C’est sans doute lui qui m'a enseigné l'art de prendre du recul et d'aborder la vie avec légèreté, soulignant que la valeur des choses réside dans l'importance qu'on leur accorde.

Merci à mes grands-parents qui ont toujours été fier de moi. Ma soif de connaissance n’est pas sans lien avec la volonté de surpasser le savoir de mon grand-père auprès duquel j’ai également compris la force du vécu. Par ailleurs, pour quelqu’un comme moi, pour qui il est difficile de décrypter les gens, je me suis toujours senti à l’aise avec l’honnêteté (souvent abusive) de ma grand-mère en qui je me suis toujours retrouvé et c’est pourquoi je suis fier de porter son prénom alors merci pour ça. Merci aussi à toutes mes sœurs et frères (vrais, faux, demi, quasi, cachés ?). Même s’il est parfois difficile de préserver les liens, chacun d’eux, à leur manière et par leur existence, a apporté une pierre à l’édifice de mon être (Jaja TMTC).

Merci à Lucas, la personne qui m'a sauvé de ce que j'aurais pu devenir. Je mesure pleinement ma chance d’avoir trouvé ma quinte flush royale. Merci pour la patience, la compréhension et l’amour, qui m'ont permis de traverser les moments difficiles avec sérénité. L'un de mes plus grands plaisirs est de cuisiner à tes côtés, d'explorer de nouvelles recettes, parfois délicieuses, parfois moins, mais toujours remplies de complicité, de rires et de Nino Ferrer. Tu apportes saveur et chaleur à ma vie, et pour cela, je t'en suis infiniment reconnaissant.

Je tiens également à remercier mes deux sista besta, Mathilde (a.k.a Bouldoswèèèère) et Thomas (a.k.a Mata Hari), pour leur amitié inconditionnelle, leur folie contagieuse et la moitié de mes cicatrices. Je ne suis jamais vraiment moi-même qu’avec vous, vous transformez chaque instant en un véritable spectacle de grand n’importe quoi, et c’est précisément ce qui rend la vie si délicieuse à vos côtés. Entre les fous rires, les délires insensés et les moments où la réalité semble complètement s’effacer, vous êtes les piliers de cette aventure délirante qu’est ma vie. Bisous D.va.

Enfin, je souhaite remercier chaleureusement les autres doctorants du laboratoire et plus généralement toutes les personnes que j’ai côtoyées pendant ces environ 4.16666666667 années pour leur camaraderie, leurs échanges enrichissants et leur soutien mutuel, qui ont grandement contribué à rendre cette expérience plus agréable.

Pour terminer mes remerciements, j’emprunterai une citation de Jean d’Ormesson qui, à mon sens, capture parfaitement la dualité de tout parcours y compris le mien :

"Merci pour les roses et merci pour les épines. La vie (ou la thèse) n'est pas une fête perpétuelle, c'est une vallée de larmes. Mais c'est aussi une vallée de roses. Si vous parlez des larmes, il ne faut pas oublier les roses et si vous parlez des roses, il ne faut pas oublier les larmes".

Table des matières

1 Introduction 11
1.1 Métabolisme ARN chez les bactéries 11
1.1.1 Modèle(s) d'étude 11
1.1.2 Machineries moléculaire impliquées dans l'expression des gènes bactériens 14
1.1.2.1 La transcription 15
1.1.2.1.1 L'ARN polymérase 15
1.1.2.1.2 Les promoteurs bactériens 16
1.1.2.1.3 L'initiation 17
1.1.2.1.4 L'élongation 20
1.1.2.1.5 La terminaison 21
1.1.2.1.6 ARN bactériens 23
1.1.2.2 La traduction 24
1.1.2.2.1 Le ribosome 24
1.1.2.2.2 Les ARN de transfert 26
1.1.2.2.3 L'initiation 27
1.1.2.2.4 L'élongation 29
1.1.2.2.5 La terminaison 30
1.1.2.3 Régulateurs de l'expression génique 31
1.1.2.3.1 Facteur de transcription 32
1.1.2.3.2 Les ARN régulateurs de l'expression génique 34
1.1.2.3.3 Protéines de liaison à l'ARN 39
1.1.2.3.4 Processing des ARN 42
1.1.2.3.5 Mécanismes de dégradation des ARN 49
1.2 Analyses transcriptomiques 73
1.2.1 Puces à ADN 73
1.2.2 Séquençage de l'ARN 75
1.2.2.1 Les débuts et le séquençage Sanger (1ère génération) 75
1.2.2.2 2ème génération et apparition du séquençage haut débit 76
1.2.2.2.1 « 454 Sequencing » (2005) 76
1.2.2.2.2 Illumina (2007) 77
1.2.2.2.3 SOLid (Sequencing by Oligonucleotide Ligation and Detection) (2008) 78
1.2.2.2.4 Ion torrent (2010) 80
1.2.2.3 Séquençage de 3ème génération 80
1.2.2.4 Préparation des librairies RNA-Seq 82
1.2.2.5 Applications et variantes du RNA-Seq 85

1.2.2.5.1 Catalogue de gènes exprimés et comparaison entre conditions.....85
1.2.2.5.2 Annotation de génome.....86
1.2.2.5.3 Détermination de la demi-vie des ARN.....86
1.2.2.5.4 Etude de l'interactome ARN.....87
1.2.2.5.5 Étude des extrémités des ARN.....91
1.2.2.6 Analyse de données RNA-Seq.....97
1.2.2.6.1 Pré-traitement.....98
1.2.2.6.2 Alignement des lectures.....98
1.2.2.6.3 Quantification et duplicats PCR.....103
1.2.2.6.4 Post-traitement (normalisation et filtrage).....107
1.2.2.6.5 Bilan.....112
1.3 Projet de thèse.....113
2 Résultats.....115
2.1 Chapitre 1 : La RNase J1 est la principale architecte du paysage transcriptomique de *S. aureus*.....115
2.1.1 Dégradation des ARN par la RNase J chez *S. aureus*.....115
2.1.2 Stabilité des ARN et impact de la RNase J.....116
2.1.2.1 Plus de 80 % des ARN ont une demi-vie impactée par la délétion de la RNase J 117
2.1.2.2 La demi-vie des gènes est corrélée avec leur niveau d'expression.....125
2.1.2.3 Près de la moitié des gènes ont la dégradation de leurs transcrits qui est dépendante de la RNase J.....128
2.1.2.4 Les gènes non impactés par la délétion de la RNase J1 sont dégradés par d'autres ribonucléases.....132
2.1.3 ARN ciblés directement par la RNase J1.....135
2.1.3.1 ARN avec lesquels la RNase J1 interagit.....135
2.1.3.2 La RNase J1 impacte et cible moins souvent le premier gène d'un opéron que les suivants.....140
2.1.3.3 Aucune caractéristique commune n'a été identifiée entre les extrémités 5' des ARN impactés ou celles des ARN non impactés.....146
2.1.3.3.1 Les états de phosphorylation des extrémités 5' des ARN des gènes ciblés sont similaires aux non ciblés.....151
2.1.3.3.2 Les séquences primaires, les pourcentages en GC et les énergies libres des extrémités ne sont pas différentes entre les ARN ciblés impactés et les non ciblés non impactés.....153
2.1.4 La RNase J est la principale actrice de la dégradation co-traductionnelle chez *S. aureus* 156
2.1.4.1 Analyse de la périodicité par sous ensemble de gènes.....163
2.1.4.1.1 Ciblés impactés *vs* non ciblés non impactés.....163

2.1.4.1.2 Périodicité par gène 166
2.1.5 Activité endoribonucléasique de la RNase J1 175
2.1.5.1 125 potentiels sites de clivage endoribonucléolytique de la RNase 175
2.1.5.2 Les sites potentiels de clivage de la RNase J1 sont situés juste après une guanine et sont souvent proches d'une région structurée 179
2.1.6 Discussion 184
2.2 Chapitre 2 : États de phosphorylation des extrémités 5' des ARN et identification de la NUDIX chez *S. aureus* 189
2.2.1 Candidats NUDIX et quantification des états de phosphorylation *in vivo* des extrémités 5' des ARN 189
2.2.1.1 Traitement statistique des données de PQ-EMOTE 194
2.2.1.1.1 Développement d'une chaîne de traitement de données PQ-EMOTE 194
2.2.1.1.2 Étape de Pré-traitement des données brutes 195
2.2.1.1.3 Alignement des lectures et quantification des extrémités 5' 200
2.2.1.1.4 Post-traitement des données de la table de comptage 204
2.2.1.1.5 Analyses des données PQ-EMOTE : Comparaison des proportions d'extrémités 5'P 232
2.2.2 Traitement des données publiées sur la RppH de *H. pylori* 258
2.2.3 Spécificité de la ou des RppH de *S. aureus* 266
2.2.4 Discussion 272
2.3 Chapitre 3 : EMOTE toolkit : une librairie R pour l'analyse de données de séquençage des extrémités des ARN 276
2.3.1 Analyse 277
2.3.2 Conception 284
2.3.2.1 Pré-traitement d'un fichier FASTQ avant l'alignement des lectures sur un génome 284
2.3.2.2 Alignement des lectures sur un génome de référence 287
2.3.2.3 Quantification des extrémités 5' à partir des alignements 288
2.3.2.4 Post-traitements des tables de comptages 290
2.3.2.4.1 Désaturation 290
2.3.2.4.2 Filtrage et normalisation des comptages dans des échantillons appariés 293
2.3.2.4.3 Normalisation des comptages sur une fenêtre génomique 294
2.3.3 Réalisation 296
2.3.4 Discussion 305
3 Matériels et méthodes 308
3.1 Analyse de gènes différentiellement exprimés (Differential gene expression - DGE) 308
3.2 Alignement des lectures sur un génome de référence 309
3.3 Représentation de la conservation de séquences nucléotidiques par Weblogos 310

3.4 Comparaison de distribution entre deux échantillons 311
3.5 Ajustement des p-valeurs 311
3.6 Analyse en composantes principales (ACP) 312
3.7 Modèle linéaire 312
3.8 Prédiction de structure secondaire d'ARN 313
4 Discussions et perspectives 315
4.1 La RNase J1 : l'actrice majeure de la dégradation des ARN chez *S. aureus*. 315
4.2 Etats de phosphorylation des extrémités 5' des ARN et impact de 4 NUDIX sur la régulation de ces états 318
4.3 EMOTE-tk librairie R pour le traitement et l'analyse de données end RNA-Seq 321
4.4 Le mot de la fin 322
5 Références 323

INTRODUCTION

Métabolisme ARN chez les bactéries

et

Analyses transcriptomiques

1 Introduction

1.1 Métabolisme ARN chez les bactéries

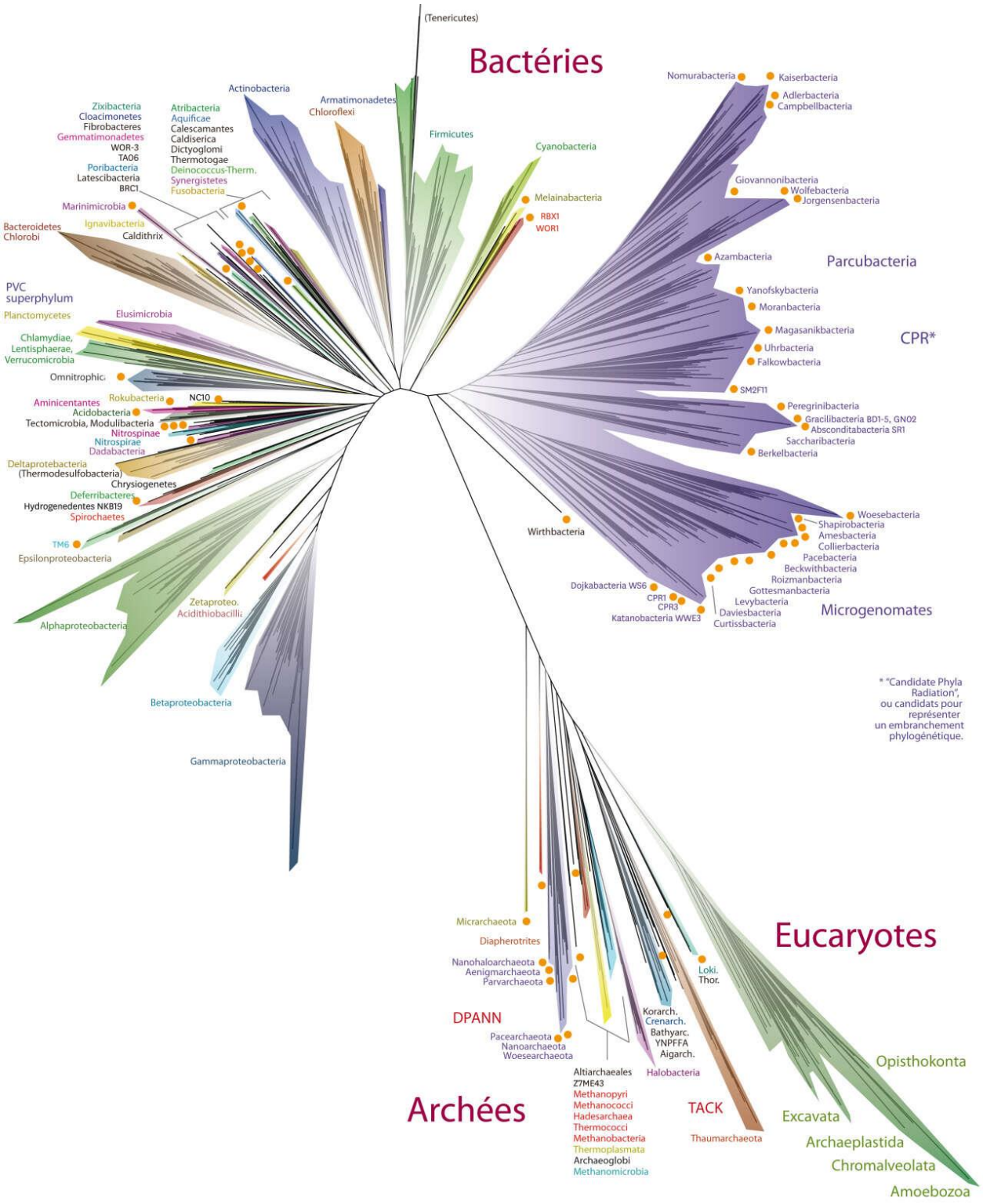
1.1.1 Modèle(s) d'étude

Cette étude étant consacrée au métabolisme ARN chez les bactéries, nous ne détaillerons pas le modèle eucaryote et nous nous concentrerons exclusivement sur le modèle bactérien (voir Fig. 1). Plus particulièrement, nous détaillerons le groupe taxonomique des Bacillota, anciennement connu sous le nom de Firmicutes, et plus précisément les Bacilli, des bactéries à Gram positif (voir Fig. 2). Bien que mon travail de thèse porte sur le métabolisme de l'ARN chez *Staphylococcus aureus*, une espèce bactérienne appartenant au phylum Bacillota, dans cette introduction, nous ferons régulièrement des comparaisons avec le modèle bien étudié de la bactérie *Escherichia coli*, une Gammaproteobactérie à Gram négatif. Ces comparaisons permettront de mettre en lumière les similarités et les divergences dans le métabolisme de l'ARN entre différents modèles bactériens.

S. aureus est une bactérie en forme de coque à gram positive, et, tout comme la bactérie modèle *Bacillus subtilis*, elle appartient au phylum des Bacillota (anciennement appelés Firmicutes). Il s'agit d'une espèce commensale à l'humain faisant partie de la flore naturelle d'environ 30% des individus. On la retrouve principalement au niveau de la peau et des voies respiratoires et, bien qu'elle soit généralement inoffensive, elle peut déclencher son pouvoir pathogène dans certaines conditions faisant du staphylocoque doré l'un des pathogène humain les plus dangereux.

Sa transmission peut se faire de façon directe (peau à peau) et, bien que l'épiderme soit une excellente barrière, si *S. aureus* entre dans le flux sanguin via des lésions, elle peut provoquer des infections suppuratives pouvant aller dans certains cas jusqu'à la septicémie. Chez les personnes ayant un système immunitaire diminué ou voire immunodéprimée, elle peut également être la source de multiples maladies telles que des pneumonies, des endocardites, des pyélonéphrites ou encore des ostéomyélites (Tong 2015, s. d.).

En raison de la concentration élevée de patients avec des plaies ouvertes, des incisions chirurgicales et des dispositifs médicaux invasifs, les hôpitaux rendent les infections plus faciles et offrent ainsi un environnement favorable au développement du staphylocoque doré. Par ailleurs, *S. aureus* est aussi connue pour acquérir des résistances aux antibiotiques et notamment à la vancomycine et méticilline, faisant d'elle un des « agents pathogènes prioritaires » de l'organisation mondiale de la santé (OMS/WHO) **(Rapport de 2017).**



L'arbre représente les trois grands domaines du vivant, qui ont évolué à partir d'un ancêtre commun (non mentionné ici) :

Eucaryotes
Du grec eu et karyon, "noyau vrai", c'est un groupe auquel appartiennent aussi les animaux (parmi lesquels l'homme) et les champignons.

Archées
Comprend les microbes qui vivent dans des conditions extrêmes.

Bactéries
Comprend les micro-organismes unicellulaires et procaryotes dont les dimensions sont de l'ordre de quelques micromètres.

En italique
Les lignées évolutives dont les organismes ont été isolés.

● Avec une pastille orange
Les lignées évolutives dont les organismes n'ont pas été isolés

Figure 1 *:* ***Vue actuelle de l'arbre de la vie, englobant la diversité totale représentée par les génomes séquencés.*** *L'arbre comprend 92 embranchements bactériens, 26 embranchements archéens et les cinq supergroupes eucaryotes. Les principales lignées sont associées à des couleurs arbitraires et nommées, avec des noms de lignées bien caractérisées. Les lignées qui n'ont pas de représentant isolé sont mises en évidence par des points rouges. Les noms Tenericutes et Thermodesulfobacteria sont mis entre parenthèses pour indiquer que ces lignées se ramifient au sein des Firmicutes et des Deltaproteobacteria, respectivement. Les supergroupes eucaryotes sont indiqués, mais ne sont pas délimités en raison de la faible résolution de ces lignées. Les embranchements du CPR (Candidate phyla radiation) sont affectés d'une couleur unique car ils sont entièrement composés d'organismes sans représentants isolés et sont encore en cours de définition aux niveaux taxonomiques inférieurs. Extrait de* (L. A. Hug et al., 2016)

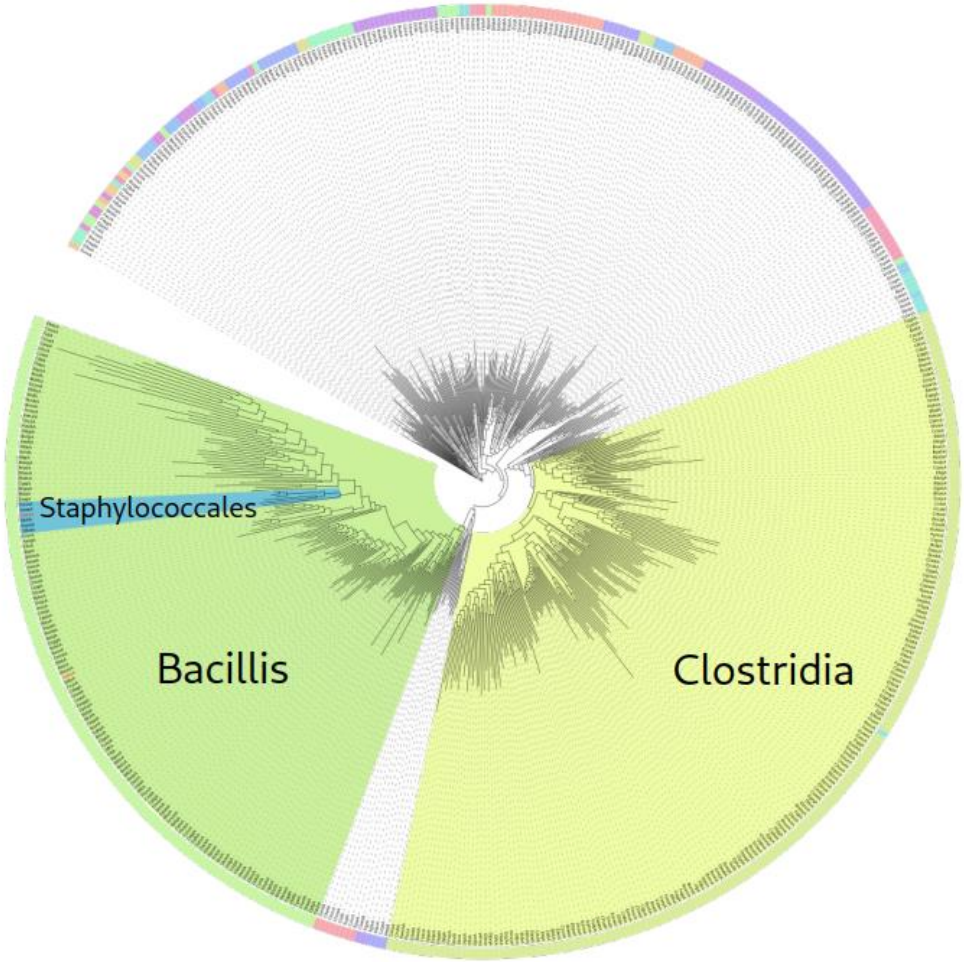


Figure 2 *:* ***Phylogénie des Bacillota.*** *Les organismes écrits en rouge correspondent à Staphylococcus aureus pour celui présent dans les Staphylococcales et à Bacillus subtilis pour l'autre. (Communication personnelle)*

Afin de réguler ses processus cellulaires et donc par exemple de déclencher son pouvoir pathogène ou encore de s'adapter face un stress environnemental certaines bactéries dont *S. aureus* ont besoin d'avoir un contrôle rapide et précis de son expression génique passant par divers mécanismes régulant la quantité d'ARN (Broglia et al., 2018)**,** (Kaito et al., 2005)**,** (Marincola & Wolz, 2017)**,** (Obana et al., 2017).

Le génome de *S. aureus* est constitué d'un chromosome circulaire d'environ 2,8 millions de paires de bases (Mb). Il s'agit d'un génome compact, c'est-à-dire qu'il est caractérisé par une forte densité de gènes, avec peu d'espaces intergéniques. Il contient environ 2 700 gènes codant pour des protéines. En plus du chromosome principal, *S. aureus* possède souvent des éléments génétiques mobiles, tels que des plasmides, des transposons et des îlots génomiques, qui jouent un rôle crucial dans la plasticité génomique de la bactérie.

1.1.2 Machineries moléculaire impliquées dans l'expression des gènes bactériens

L'expression génique est définie comme l'ensemble des processus par lesquels l'information héréditaire codée par un gène est transformée en un produit fonctionnel, ce produit pouvant être une protéine ou un ARN. Une première étape est la transcription qui permet de produire une copie ARN à partir d'un des brins ADN d'un gène. Si le gène code pour une protéine, une deuxième étape, la traduction permet la synthèse du produit protéique à partir de l'ARN, appelé alors ARN messager (ARNm). Toutes les étapes de ces processus peuvent être régulées. La régulation de l'expression génétique permet de contrôler quand le produit d'un gène (protéines ou ARN non codant (ARNnc)) doit être synthétisé et en quelle quantité. Ce contrôle est vital pour permettre à la cellule de synthétiser les produits dont elle a besoin au moment où elle en a besoin, donc de mettre en place les réponses nécessaires pour s'adapter à un environnement variable, à des signaux externes, à des dommages subis par la cellule, *etc.* Le processus d'expression des gènes varie donc considérablement en fonction des conditions environnementales et cellulaires. Cependant, il existe des gènes qui sont exprimés quelles que soient les conditions. Ils sont appelés « gènes constitutifs » ou encore « gènes de ménages » et leurs produits sont nécessaires au maintien des fonctions cellulaires de base.

1.1.2.1 La transcription

La synthèse des ARN est un processus essentiel et ubiquitaire chez tous les organismes vivants. Elle correspond à l'étape de transcription de l'ADN en ARN. Ce processus se déroule en trois étapes principales : l'initiation, l'élongation et la terminaison et résulte en un brin d'ARN complémentaire au brin anti-sens d'ADN. Ce processus de transcription est réalisé par l'ARN polymérase (ARNP).

1.1.2.1.1 L'ARN polymérase

Contrairement aux eucaryotes qui codent pour plusieurs ARNP, les bactéries n'en possèdent qu'une seule. L'ARNP est un complexe enzymatique dont le cœur se compose de cinq sous-unités : les grandes sous-unité β et β′, deux exemplaires de la sous-unités α et la petite sous-unité ω (Fig. 3). Chaque sous-unité α est constituée de domaines amino-terminaux (αNTD) et carboxy-terminaux (αCTD) repliés indépendamment et reliés par un linker flexible. Le domaine αNTD contient des déterminants pour l'assemblage de la RNAP alors que le domaine αCTD contient les déterminants permettant des interactions avec l'élément UP (région riche en A et T) faisant partie de certains promoteurs bactérien (Ross et al., 1993) et des déterminants pour les interactions avec les protéines activatrices de la transcription au niveau des promoteurs régulés positivement (Ebright & Busby, 1995). La sous-unité β et la sous-unité β′ (Blatter et al., 1994)**,** (Aboshkiwa et al., 1992) sont assemblées en se liant aux domaines N-terminaux des sous-unités α, et forment une « pince de crabe » qui contient, à sa base, le site actif. La sous-unité ω participe à la liaison de l'enzyme sur la matrice d'ADN. Elle est principalement une chaperonne pour la sous-unité β′ en permettant de maintenir celle-ci dans sa bonne conformation et en la recrutant pour former une enzyme fonctionnelle (Keller et al., 2014)**,** (Minakhin et al., 2001)**,** (Murakami & Darst, 2003)**,** (Weiss et al., 2014). L'enzyme cœur peut ensuite se fixer à une autre sous-unité, le facteur sigma, formant ainsi le complexe connu sous le nom d'ARNP holoenzyme (Fig. 3). Le facteur sigma est nécessaire à l'étape d'initiation pour que l'ARNP se fixe sur le premier nucléotide à transcrire. Pour cela, le facteur sigma se fixe à l'ADN en reconnaissant de manière spécifique une séquence, le promoteur qui est situé en amont du site d'initiation de la transcription.

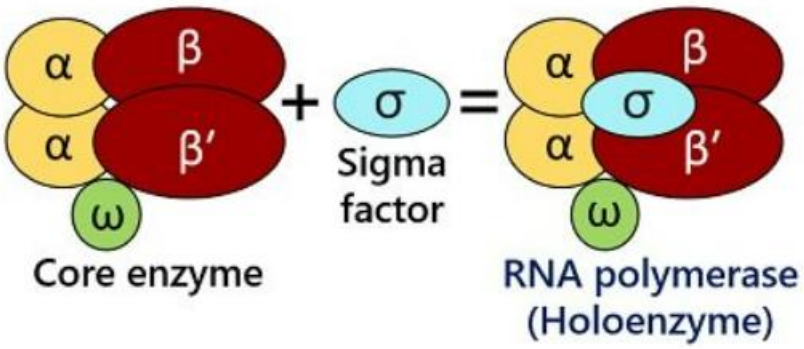


Figure 3 *:* ***Sous unités de l'ARN polymérase.*** *ARN polymérase cœur qui devient l'ARN polymérase holoenzyme par son assemblage avec le facteur sigma. Extraite du site web de « biology reader »*

1.1.2.1.2 Les promoteurs bactériens

Les promoteurs bactériens possèdent deux séquences caractéristiques de 6 pb chacune, séparées par une séquence non spécifique de 17 à 19 nucléotides. Ces deux séquences sont appelées régions ou « boîtes » -35 et -10 en référence à la distance moyenne qui les séparent du site de début de transcription (Fig. 4A). Les séquences de ces deux régions ne sont pas strictement identiques mais une séquence consensus a cependant pu être déterminée lors de la comparaison d'un ensemble de promoteurs. Ces séquences consensus sont TTGACA pour l'élément -35 et TATAAT pour l'élément -10 séparés par une distance optimale de 17 pb (Harley & Reynolds, 1987). Les promoteurs ayant le plus de résidus en commun avec ces deux séquences consensus sont considérés comme des promoteurs « fort », c'est-à-dire avec un rythme d'initiation de la transcription plus élevé. Ce rythme est influencé par l'efficacité de l'ARNP, *via* le facteur sigma, à se lier au promoteur, à s'isomériser et à se détacher. La différence de séquence des promoteurs permet donc une régulation fine du taux d'expression des gènes, les gènes dotés d'un promoteur fort sont les plus exprimés.

L'élément UP, une région riche en A/T, et liant également l'ARNP est trouvé localisé en amont de la région -35 de certains promoteurs, notamment les plus forts, comme ceux dirigeant par exemple l'expression des gènes codant les ARN ribosomiques (ARNr). Il augmente l'initiation de la transcription de 10 à 100 fois (Ross et al., 1993) en permettant une liaison spécifique supplémentaire de l'ARNP via le domaine αCTD (Fig. 4B)

Certains promoteurs ne possèdent pas de régions -35 et présentent alors une région -10 étendue qui correspond à une boîte -10 standard contenant, en amont, une courte séquence supplémentaire permettant des contacts additionnels avec l'ARNP, compensant ainsi l'absence de la boîte -35 (Fig. 4C) (Mitchell et al., 2003).

Finalement un dernier élément, appelé le discriminateur, a été identifié juste en aval de l'élément -10 (Fig. 4D). La force de liaison entre cet élément et l'ARNP influence la stabilité du complexe enzyme-promoteur.

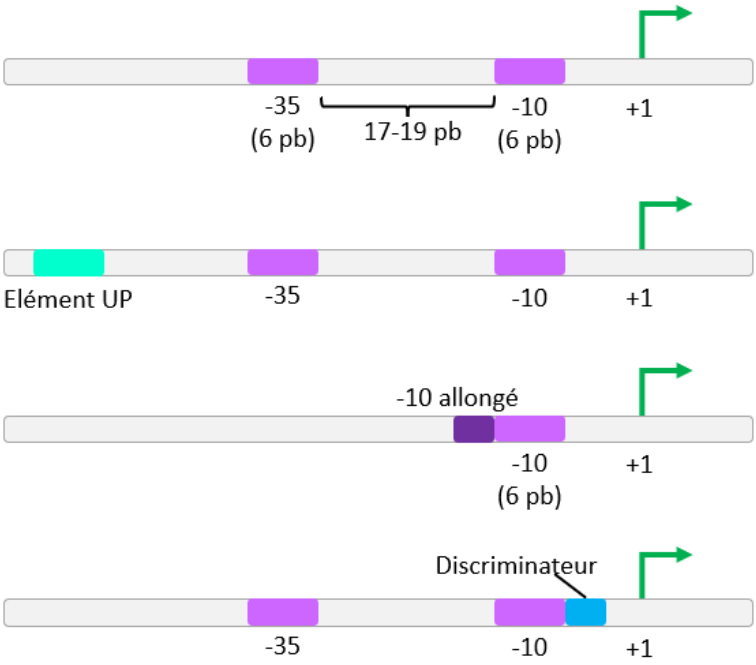


Figure 4 :* Description des éléments de la région promotrice**. *Adaptée de* ***Watson et al. *(2009)*

1.1.2.1.3 L'initiation

Toutes les bactéries possèdent un facteur sigma essentiel prédominant, connu sous le nom de facteur sigma de ménage ou domestique (tel que σ^{70} chez *Escherichia coli* ou σ^{A} chez *Bacillus subtilis* et *Staphylococcus aureus* ; également connu sous le nom de RpoD), qui est responsable de la reconnaissance de la plupart des promoteurs. Les facteurs sigma de la famille σ^{70} sont composés de quatre domaines structuraux reliés par des linkers flexibles. Les domaines 2 et 4 établissent des contacts spécifiques avec les régions -10 et -35, respectivement, tandis que le domaine 3 interagit avec la région -10 étendue présente dans certains promoteurs.

L’élément discriminateur est reconnu par le domaine 1 et comme dit précédemment, l’élément UP est reconnu par le domaine αCTD de la sous-unité α (Fig. 5) (Watson et al. 2009).

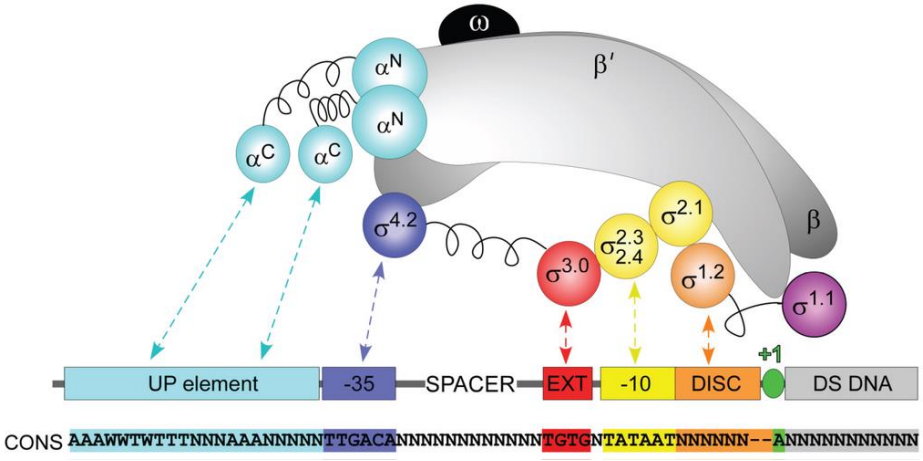


Figure 5 : Interactions spécifiques de la séquence entre l’ARNP-σ[70] et les régions du promoteur. *Représentations schématiques des sous-unités de l’ARNP cœur, σ[70], et de l'ADN promoteur. RNAP : α_2 : cyan ; β et β' : gris ; ω : noir. Régions σ : comme indiqué. Promoteur : élément UP : cyan ; élément -35 : bleu ; élément -10 étendu : rouge ; élément -10 : jaune ; discriminateur : orange ; site de début de transcription : vert ; ADN en aval du site de début de transcription : gris. Les régions de liaison des sous-unités α et σ sont représentées par des ressorts. Extrait de* (Ruff et al., 2015)

La liaison initiale de l’ARNP au promoteur, forme un complexe fermé. La transition vers un complexe ouvert implique une série de changements de conformation dans les deux biomolécules (souvent appelée isomérisation) qui ouvre ~13 pb à partir de l'élément -10 jusqu'au -delà du site de démarrage de la transcription, créant la « bulle » d'initiation. L’isomérisation est irréversible et garantit que la transcription va ensuite démarrer. Par contre, la formation du complexe fermé est réversible, la polymérase pouvant facilement se dissocier du promoteur ou bien transiter vers la forme ouverte.

L'ajout de nucléosides triphosphates (NTP) permet une nouvelle transition vers le complexe d'initiation, qui synthétise la transcription de l'ARN. Au départ, le brin matrice de l'ADN est attiré dans le complexe d'initiation et replié à l’intérieur de l’enzyme, un processus connu sous le nom de compression (scrunching). Ce complexe peut être maintenu au niveau du promoteur, ce qui entraîne des cycles d'initiation avortés qui ne produisent que de petits fragments d'ARN de moins de 10 nucléotides (synthèse abortive) (Fig. 6). Quand l’'ARNP s'échappe du promoteur, elle entre alors dans la phase d’élongation (Watson et al. 2009).

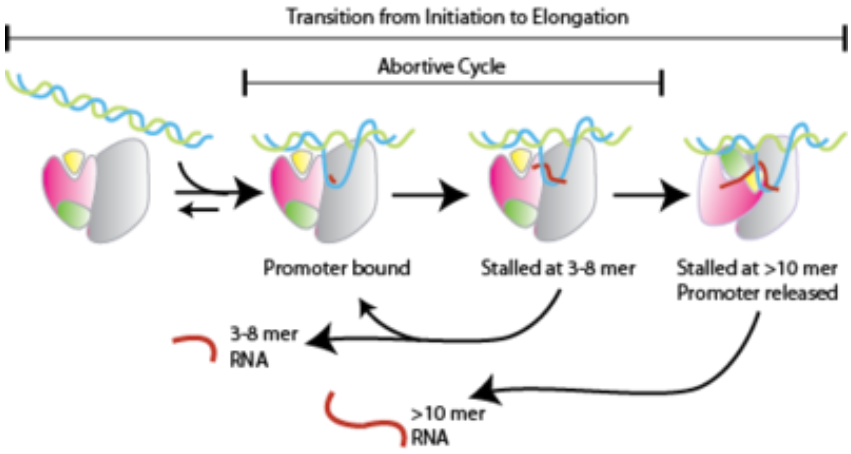


***Figure 6 :* La synthèse abortive se produit avant la libération du facteur sigma.**

1) L'ARN polymérase se lie à l'ADN promoteur pour former un complexe fermé ARN polymérase-promoteur. 2) L'ARN polymérase déroule ensuite un tour d'ADN autour du site de démarrage de la transcription pour former un complexe ouvert ARN polymérase-promoteur. 3) L'ARN polymérase entre dans des cycles de synthèse avortés et libère des produits d'ARN courts (contenant moins de 10 nucléotides). 4) L'ARN polymérase s'échappe du promoteur et entre dans l'étape d'élongation de la transcription. Extrait de ***Wikipedia***

1.1.2.1.3.1 Facteurs σ alternatifs

Cependant, outre le facteur σ domestique, qui est essentiel à la viabilité dans des conditions normales de croissance et qui recrute l'ARNP sur la majorité des promoteurs, presque toutes les bactéries possèdent un ou plusieurs facteurs σ supplémentaires, appelés facteurs σ alternatifs. Ils forment deux classes distinctes : la famille des facteurs σ^{70}, majoritaire appelé ainsi car apparenté phylogénétiquement au facteur σ^{70} de *E. coli* (Paget & Helmann, 2003) et la famille des facteurs σ^{54}. Le nombre de membres de la famille σ^{70} est très variable d'une espèce à l'autre, avec par exemple sept membres codés dans le génome de *E. coli* (Cho et al., 2014), 17 dans celui de *B. subtilis* (Haldenwang, 1995; Helmann, 2002) et jusqu'à 65 dans celui de *Streptomyces coelicolor* A3 (Kormanec, 2022). Ces facteurs reconnaissent des motifs de séquence d'ADN distincts du facteur σ^{70} et guident l'ARNP vers différents promoteurs, ce qui fournit un mécanisme simple pour reprogrammer la transcription vers un ensemble différent de gènes. Ils permettent l'expression de régulons spécifiques en réponse au stress, à la densité cellulaire, aux transitions développementales et aux signaux nutritionnels. D'après les données disponibles, les deux types de facteurs sigma (domestiques et alternatifs) se fixeraient sur les mêmes sites de l'ARNP, mais l'abondance supérieure du facteur σ ferait qu'il pourrait supplanter le facteur alternatif. Cependant,

quand la quantité de facteur alternatif augmente, celui peut alors rentrer en compétition avec le facteur domestique et permettre la reprogrammation d'un sous-ensemble de molécules d'ARNP (Browning & Busby, 2016; Feklístov et al., 2014; Gruber & Gross, 2003).

Les facteurs de la famille σ^{70} ont été divisés en quatre groupes phylogénétiques en fonction de leur structure en domaines et de la fonction des gènes. Le groupe 1 correspond aux facteurs σ domestiques étroitement apparentés au facteur σ^{70} de *E. coli*. Le groupe 2 correspond à un groupe de paralogues étroitement apparentés au facteur σ domestique mais on essentiel à la croissance bactérienne. Les membres du groupe 3 sont plus distants phylogénétiquement de σ^{70} domestique et activent des régulons en fonctions de signaux spécifiques. Par exemple, ils régulent l'expression des gènes impliqués dans la sporulation, la biosynthèse du flagelle, la chimiotaxie, la réponse à des chocs thermiques. Le dernier groupe, numériquement le plus important, renferme les facteurs σ de fonction extracytoplasmique (ECF) qui forment la sous-famille la plus distante de σ^{70} et dont la structure ne compte que les deux domaines 2 et 4 (Paget et al., 2003). Aux côtés des systèmes à un ou deux composants, ils constituent l'un des principaux mécanismes de transduction du signal chez les bactéries permettant ainsi à l'organisme de mettre en place une réponse adaptative au stress environnemental détecté. Une analyse phylogénomique récente a permis d'identifier 157 groupes d'ECF sur la base d'un voisinage génétique conservé (Casas-Pastor et al., 2021).

La famille des facteurs σ^{54} est moins abondante dans les génomes bactériens. Ces facteurs reconnaissent des promoteurs différents de ceux reconnus par σ^{70}, avec des positions conservées autour des positions -24 et -12 du site de début de transcription (Barrios et al., 1999) . Une fois que la RNAP- σ^{54} est liée à son promoteur, elle ne peut initier la transcription que si elle est activée par des activateurs transcriptionnels (Siegel & Wemmer, 2016). Vu historiquement comme un system de régulation spécialisée pour la synthèse des gène impliqués dans l'assimilation ou la fixation de l'azote, les nombreux travaux réalisés depuis ont conduit à considérer de système comme important pour la transcription de différents types de gènes, comme le métabolisme des acides aminés, la formation de biofilm, la virulence et pourrait jouer un rôle dans la modulation de l'expression des gènes des produits naturels de *Myxococcus xanthus* (Ma et al., 2021).

1.1.2.1.4 L'élongation

Le facteur σ est donc essentiel pour initier la transcription mais n'est pas nécessaire pour la suite de la transcription. Excepté dans de rares cas, il va donc se dissocier du complexe enzymatique, signal de la fin de la phase d'initiation. Pendant l'élongation, l'ARNP glisse le long de l'ADN double brin, le déroulant et transcrivant la séquence d'ADN nucléotidiques en ribonucléotides (ajout de

nucléotides complémentaires au brin matrice d'ADN en 3' du brin d'ARN). Le brin d'ARN nouvellement formé est pratiquement identique au brin codant de l'ADN (brin sens ou brin non matrice), à l'exception de l'uracile qui remplace la thymine et d'un squelette de sucre ribose au lieu d'un squelette de sucre désoxyribose. Plusieurs ARNP sont généralement actives en même temps, ce qui signifie que de nombreux brins d'ARN peuvent être produits très rapidement.

1.1.2.1.5 La terminaison

Deux mécanismes de terminaison sont connus chez les bactéries : la terminaison Rho-indépendante et la terminaison Rho-dépendante.

1.1.2.1.5.1 Rho indépendante

La terminaison intrinsèque, également connue sous le nom de terminaison Rho-indépendante, implique des séquences spécifiques d'ADN qui provoquent l'arrêt de l'ARN polymérase. Ces séquences comprennent deux éléments, une courte séquence répétée inversée riche en GC suivie d'une région de 7 à 8 pb riche en T (en U sur l'ARN transcrit). Lors de sa transcription en ARN, la séquence répétée inversée adopte une structure en tige et boucle qui provoque une pause de l'ARN polymérase. La chaîne d'ARN naissante est alors associée à la matrice dans le centre actif uniquement par des paires de bases A : U qui forment des interactions faibles. L'ARN synthétisé peut donc se détacher de sa matrice (Fig. 7) (Farnham & Platt, 1981), (Peters et al., 2011).

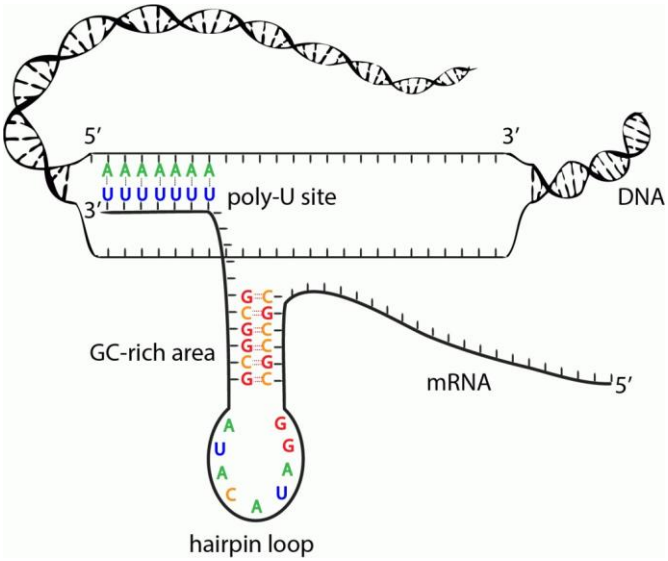


Figure 7 : Mécanisme de terminaison Rho-indépendante*. Extrait d'online Biology notes*

1.1.2.1.5.2 Rho dépendante

La terminaison Rho-dépendante se produit lorsque le facteur Rho, une protéine hexamèrique avec une activité hélicase, se lie au brin d'ARN nouvellement synthétisé. Ce facteur parcourt l'ARN de l'extrémité 5' vers l'extrémité 3'. Lorsque l'ARNP approche de la fin du gène et rencontre une série de nucléotides riches en G, elle ralentit considérablement. Ce ralentissement permet au facteur Rho de rattraper l'ARNP qui, grâce à son activité hélicase ATP-dépendante, va dissocier le duplex ARN-ADN de l'ARNP, libérant ainsi l'ARN nouvellement synthétisé et mettant fin à la transcription (Fig. 8) (Peters et al., 2011).

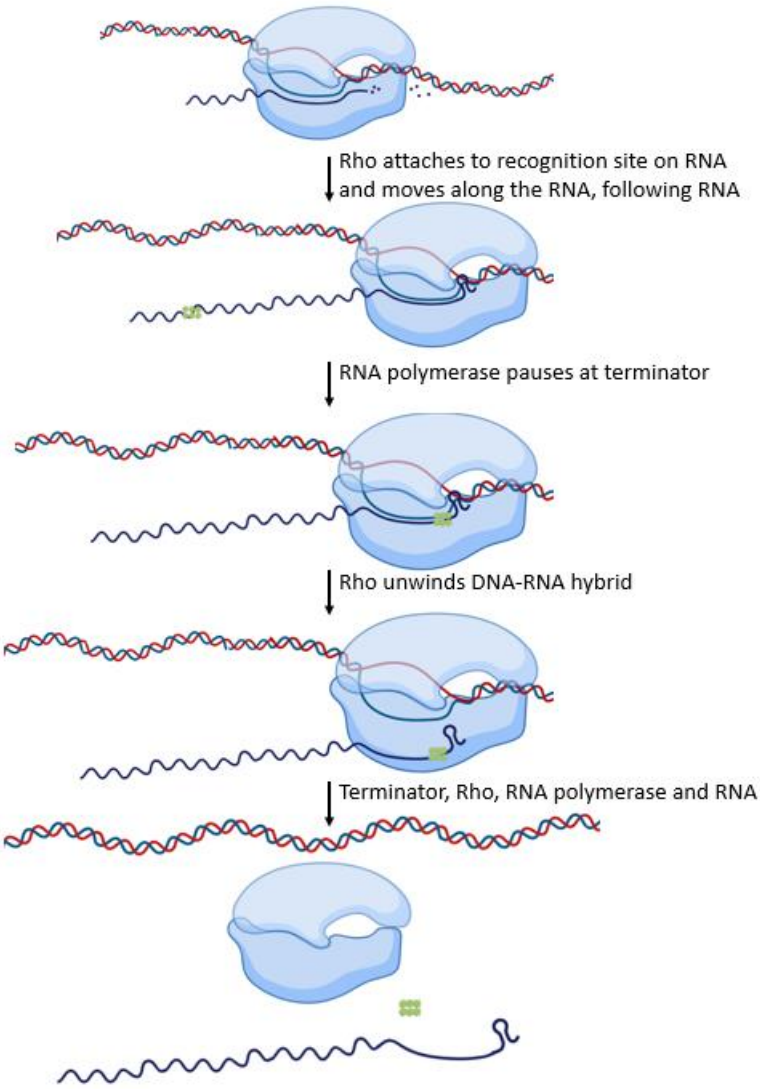


Figure 8 : Mécanisme de terminaison Rho-dépendante*. Adaptée* ***de « Online Biology notes »***

1.1.2.1.6 ARN bactériens

Contrairement aux eucaryotes, les ARN bactériens néo-synthétisés ne possèdent pas de coiffe m7-G en 5' mais un tri-phosphate (dans certains cas il peut s'agir de NAD+, de NADH ou de CoA), et ne possèdent pas d'introns (Fig. 9A et 9B). Chez les bactéries, la transcription et la traduction peuvent avoir lieu simultanément dans le cytoplasme de la cellule (transcription co-traductionelle) du fait de leur absence de noyau (Stent, 1964).

De plus, chez les eucaryotes, la plupart des molécules d'ARNm sont monocistroniques (c'est-à-dire qu'elles codent pour un seul polypeptide). Chez les bactéries et les archées, cependant, certains ARNm sont polycistroniques (Fig. 9C), c'est-à-dire que les gènes codant pour plusieurs protéines sont sous le contrôle d'un même promoteur, et leurs produits protéiques sont généralement impliqués dans la réalisation d'une même fonction physiologique. Les groupes de gènes qui donnent naissance aux ARNm polycistroniques sont des unités de transcription uniques appelées opérons (Kozak, 1983).

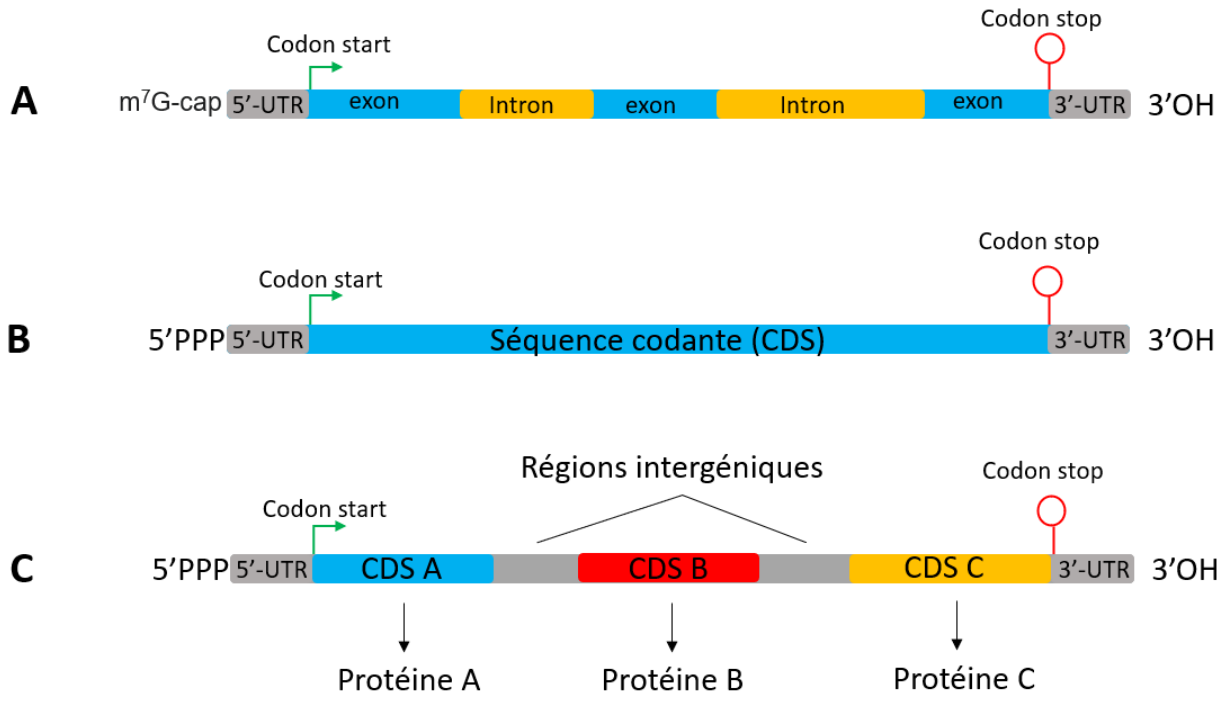


Figure 9 : ***Comparaison entre les ARN néo-synthétisés*** *(**A**) chez les eucaryotes, (**B**) chez les bactéries et (**C**) les ARN polycistroniques chez les bactéries.*

1.1.2.2 La traduction

La synthèse des protéines est une étape essentielle chez tous les êtres vivants et se fait par la traduction des régions codantes (CDS) des ARNm en peptides. La traduction requiert quatre composants principaux :

- Les ARNm contenant la séquence codant la protéine ainsi que des régions de reconnaissance pour l'initiation et la terminaison de la traduction
- Les ARN de transferts (ARNt) qui une fois chargé de leur acide aminé (processus réalisé par les aminocyl-ARNt- synthétases) vont mettre en relation un codon et un acide aminé
- Les aminoacyl-ARNt synthétases permettant donc de charger l'acide aminé sur l'ARNt correspondant à son anticodon
- Les ribosomes, gros complexe ribonucléoprotéique, qui coordonne la reconnaissance de l'ARNm par chaque ARNt et qui catalyse la formation de la liaison peptidique entre la chaîne polypeptidique en croissance et l'acide aminé porté par l'ARNt sélectionné.

1.1.2.2.1 Le ribosome

Le ribosome bactérien est une particule ribonucléoprotéique cytoplasmique dont la fonction principale est de servir de site de traduction de l'ARNm et de synthèse des protéines. Le ribosome se compose de deux sous-unités appelées 30S (petite sous-unité) et 50S (grande sous-unité). Lorsqu'elles sont réunies, le ribosome a un coefficient de sédimentation de 70S (Fig. 10).

La petite sous-unité 30S est composée de l'ARNr 16S et de 21 protéines ribosomales. Elle contient le centre de décodage dans lequel les ARNt chargés décodent les codons de l'ARNm. La grande sous-unité du ribosome contient le centre peptidyl-transférase, responsable de la formation des liaisons peptidiques. Elle est composée des ARNr 23S et 5S, qui constituent l'ossature du complexe ribosomique ainsi que de 33 protéines ribosomique (Deutscher, 2009). Le ribosome possède trois sites de liaison pour les ARNt, les site A, P et E. Le site A lie les ARNt chargés de leur acide aminé (ARNt aminoacylés), le site P lie les ARNt liés à la chaîne peptidique en cours de synthèse (peptidyl-ARNt) et le site E lie les ARNt libres (déchargé et décroché de la chaîne peptidique) avant leur sortie du ribosome (E pour « exit »). Chaque site de liaison est situé à l'interface entre les deux sous-unités du ribosome. Les ANRt liés occupent donc l'intervalle entre le centre peptidyl-transférase de la grande sous-unité et le centre de décodage de la petite sous-unité.

Les ARNr sont les déterminants catalytiques du ribosome. La plupart des protéines ribosomiques sont situées à la périphérie du ribosome et non en son centre. Les domaines fonctionnels que sont les centres de décodage et de peptidyl-transférase sont composés entièrement ou principalement d'ARN.

Lors de la synthèse des protéines, le ribosome se déplace le long d'une molécule d'ARNm, lit le codon et ajoute l'acide aminé correct (provenant de l'ARNt-aminoacyl correspondant) à la protéine en croissance.

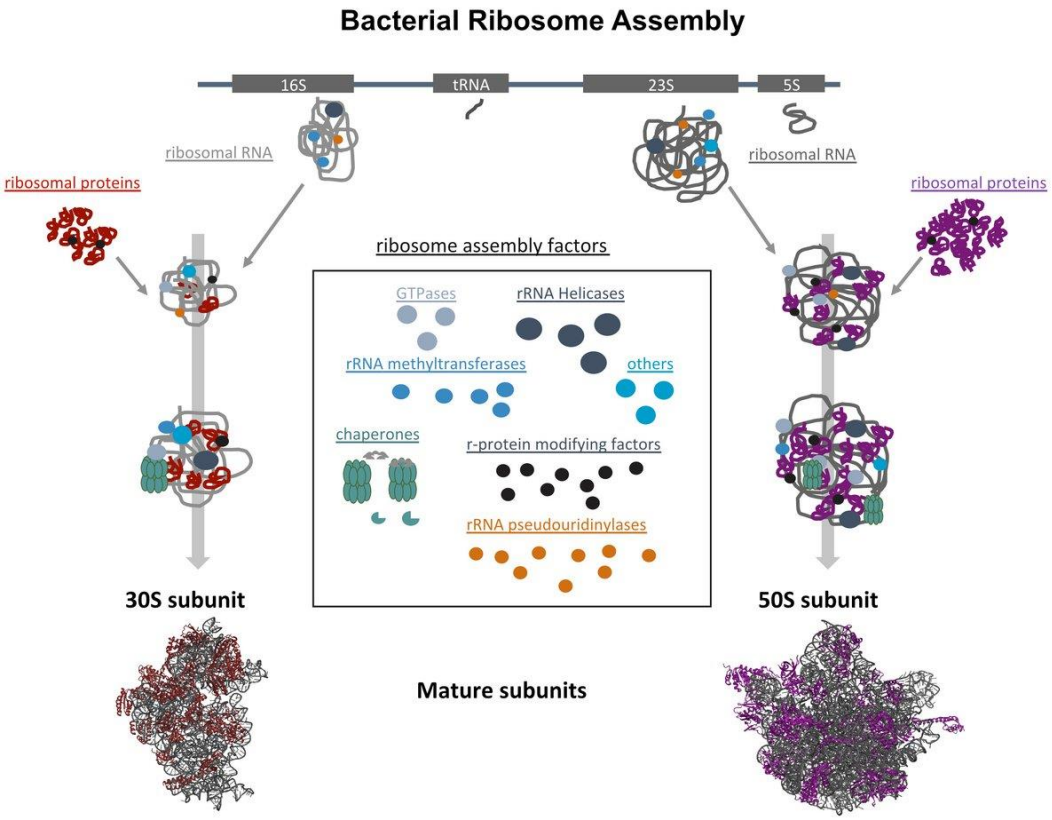


Figure 10 *:* ***Synthèse des sous unité ribosomique 30S et 50S****.Extraite du site de l'université de Konstanz (AG Deuerling team).*

1.1.2.2.2 Les ARN de transfert

Les ARN de transfert (ARNt) sont également des composantes principales de la machinerie de traduction. Il s'agit de courts ARN (70 à 100 nucléotides) non codants et repliés en plusieurs tige-boucles formant une structure secondaire caractéristique appelée feuille de trèfle et présentant une structure tridimensionnelle en forme de L. (Fig. 11). Chaque ARNt, de par son anticodon, est spécifique (complémentaire antisens) à l'un des 64 codons du code génétique codant pour les 20 acides aminés. Cette structure tertiaire en forme de L permet d'exposer l'anticodon pour la reconnaissance du codon tout en apportant l'acide aminé correspondant chargé sur son extrémité 3' CCA-OH (bras accepteur) par des aminoacyl-ARNt-synthétases. Cela permet d'apporter aux ribosomes l'acide aminé correspondant au codon de l'ARNm lors de la traduction (Shepherd & Ibba, 2015)**.** Les génomes possèdent donc plusieurs gènes codant pour des ARNt, par exemple 61 chez *S. aureus*, 86 chez *E. coli* et 86 chez *B. subtilis* (GtRNAdb).

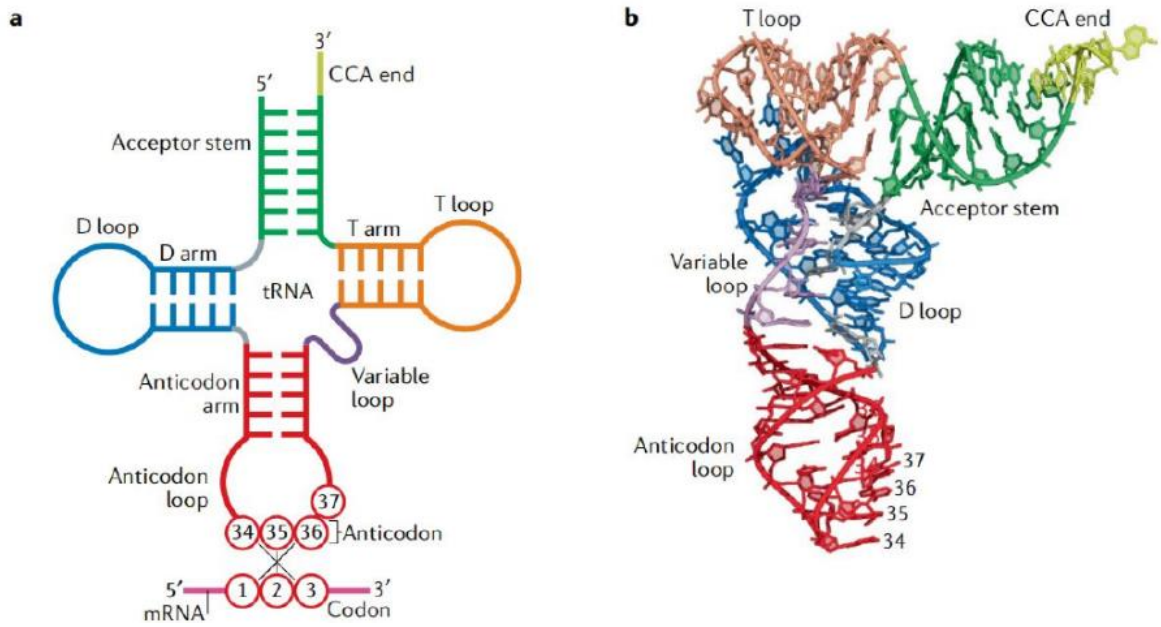


Figure 11 *:* ***Structure de base et fonction de l'ARNt.*** *(**a**) Structure en trèfle de l'ARNt avec appariement codon-anticodon : l'ARNt se compose de cinq parties : Le bras accepteur, le bras D, la tige de l'anticodon, la région variable et la boucle T. (**b**) Structure tertiaire de l'ARNt : les coordonnées sont tirées de l'entrée 1EHZ de la Protein Data Bank. Le code couleur est le même que dans la partie a. Extrait de* ***Ding et al. 2022 Journal of clinical medecine.***

1.1.2.2.3 L'initiation

Pour que l'initiation de la traduction soit correctement réalisée, il faut que le ribosome soit recruté sur l'ARNm, un ARNt chargé doit s'insérer au site P du ribosome et finalement le ribosome doit être positionné de façon très précise sur le codon d'initiation. Ce positionnement est critique car il établit le cadre de lecture utilisé. L'assemblage du ribosome se fait une sous-unité à la fois. Dans un premier temps, la petite sous-unité est chargée et chez les bactéries cela est réalisé *via* l'interaction de l'ARN 16S avec le site de fixation du ribosome ou « ribosome binding site » (RBS), appelé aussi la séquence de Shine-Dalgarno car initialement proposé par John Shine et Lynn Dalgarno. Ce site est situé six à 12 nucléotides en amont du codon d'initiation de la traduction et possède la séquence consensus 5'-GGAGG-3' complémentaire de la séquence CCUCCU localisée à l'extrémité 3' de l'ARNr 16S. Cela va permettre de guider le positionnement du codon de départ au site P du ribosome (Milón et al., 2012). Normalement les ARNt chargés entrent dans le ribosome au niveau du site A et n'atteignent le site P qu'une fois la formation de la liaison peptidique établie avec la chaîne peptidique en croissance et l'éviction de l'ARNt du site P qui est maintenant décroché de la chaîne peptidique. Cependant, lors de l'initiation, un ARNt chargé arrive directement au site P. Cet ARNt initiateur s'associe avec le codon d'initiation. Si le codon AUG reste généralement le codon d'initiation majoritaire (83% chez *E. coli* et 78% chez *B. subtilis*), les codon GUG (9% chez *B. subtilis*) et UUG (13% chez *B. subtilis*) sont également rencontrés (Rocha et al., 1999). Ces trois codons, quand ils sont situés au sein de la région codante vont être lu respectivement par des ARNt pour la méthionine, la valine et la leucine. L'ARNt initiateur ne porte aucun de ces trois acides aminés, mais une méthionine modifiée, une N-formylméthionine (fMet-ARNt$_i^{fMet}$). Ces étapes requièrent les facteurs d'initiation IF1, IF2 et IF3. IF1 empêche la fixation d'ARNt sur la partie du site A de la petite sous-unité. IF2, par ses interactions avec la petite sous-unité, IF1 et le fMet-ARNt$_i^{fMet}$, facilite l'association du fMet-ARNt$_i^{fMet}$ et empêche la fixation d'autres ARNt. Finalement, l'initiation de la traduction nécessitant une petite sous-unité libre, IF3 empêche qu'elle s'associe avec une grande sous-unité (Blattner et al., 1997)**,** (Starmer et al., 2006). L'appariement du codon d'initiation et du fMet-ARNt$_i^{fMet}$ conduit à un changement conformationnel de la petite sous-unité entraînant la libération d'IF3. La grande sous-unité peut alors se lier à la petite sous-unité entraînant la libération de IF1 et IF2. Le complexe ribosome-ARNm est alors prêt pour accueillir un ARNt chargé au site A et commencer la synthèse peptidique (Watson et al. 2009).

Outre le mécanisme canonique d'initiation de la traduction décrit ci-dessus, les bactéries peuvent initier la traduction par au moins deux autres mécanismes L'un d'eux est le balayage 70S, également appelé couplage translationnel ou readthrough (Fig. 12, centre).

Dans ce mécanisme, le ribosome ne se dissocie pas en sous-unités 30S et 50S après avoir terminé la traduction d'un cistron. Au contraire, il continue à être associé à l'ARNm et glisse en aval jusqu'à ce qu'il atteigne un RBS canonique où il peut commencer la traduction (Yamamoto et al., 2016), (Leiva et al., 2022). D'autre part, des ARNm sans région 5' UTR et donc sans Shine-Dalgarno ont été identifié dans de nombreux génomes bactériens et ils sont notamment plus nombreux dans les bactéries Gram positif (Tedin et al., 1999). Il a été proposé que la traduction des ARNm commencerait principalement par la liaison directe d'une particule 70S à un codon d'initiation AUG situé à l'extrémité 5′ d'un ARNm ou très proche de celle-ci (Fig. 13) (Udagawa et al., 2004), (O'Donnell & Janssen, 2002), (Beck et al., 2016).

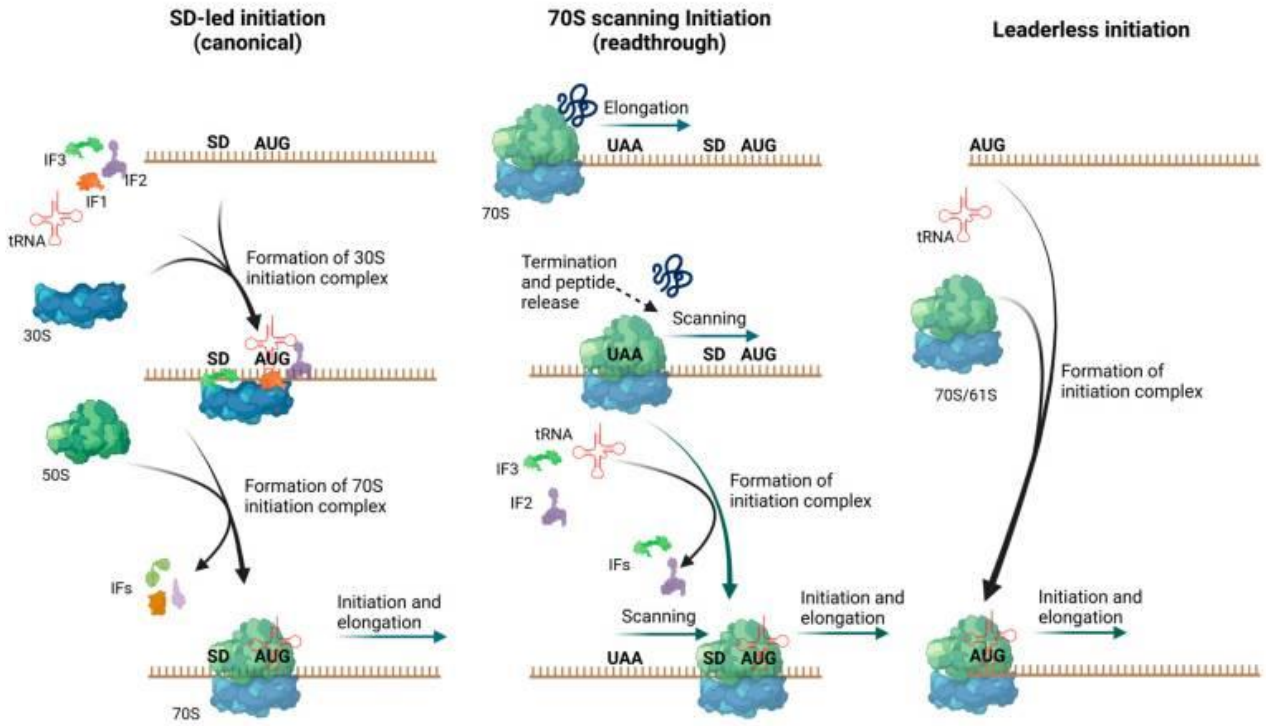


***Figure 12 :* Mécanismes d'initiation de la traduction chez les bactéries.**

Chez les bactéries, la traduction peut être initiée par trois mécanismes. La traduction dirigée par le Shine-Dalgarno (SD) (à gauche), le balayage 70S (au centre) et l'initiation sans leader (à droite). Comme on le voit, bien que la plupart des composants soient similaires entre ces trois mécanismes, les facteurs d'initiation de la traduction (IF) ont des rôles différents. Les flèches indiquent soit le mouvement de l'ARNm, soit la liaison/dissociation des composants de la machinerie de traduction. Le modèle de IF1 est basé sur la pdb 1AH9 (Sette et al., 1997) *et le modèle de IF2 sur la pdb 4b48* (Simonetti et al., 2013) *; le modèle de IF3 a été obtenu à partir d'AlphaFold (P0A707* (Jumper et al., 2021), (Varadi et al., 2022)*). Extrait de* (Leiva et al., 2022).

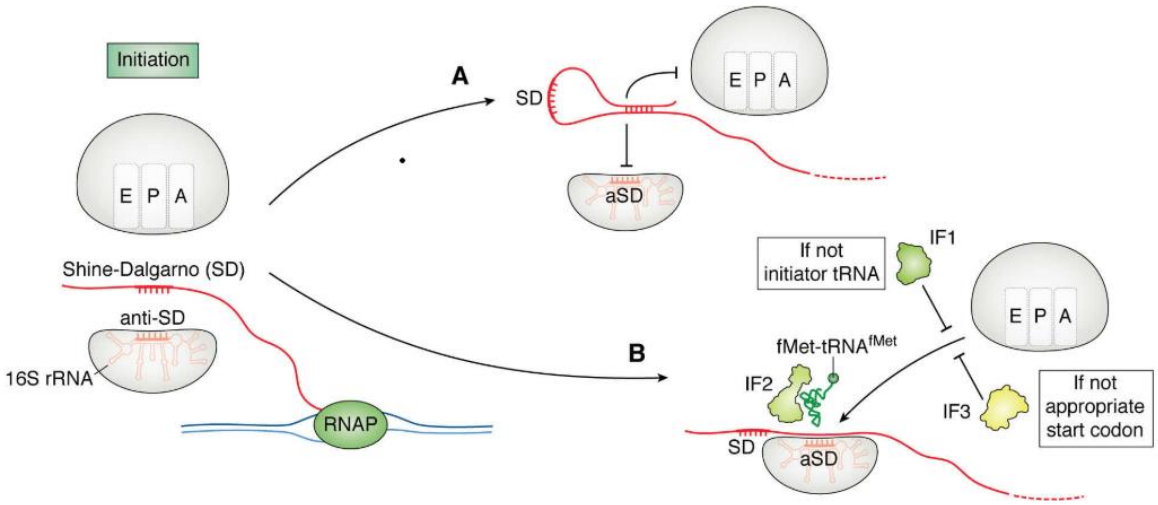


***Figure 13 :* Régulation de l'initiation de la traduction chez les bactéries.**

*Dans de nombreuses bactéries, l'initiation de la traduction se produit en aval d'une séquence SD à l'aide des facteurs d'initiation 1 à 3. (**A**), un mécanisme de régulation de l'initiation consiste à inhiber la liaison de la sous-unité ribosomique à l'ARNm en séquestrant la séquence SD par des événements de repliement de l'ARN. (**B**), lorsque le processus d'initiation de la traduction commence, IF1 inspecte l'ARNt qui est amené au ribosome par IF2. L'IF3 contribue à la fidélité de l'initiation en garantissant l'initiation au codon de départ approprié. Extrait de* (Tollerson & Ibba, 2020)

1.1.2.2.4 L'élongation

L'élongation de la chaîne polypeptidique implique l'addition d'acides aminés à l'extrémité carboxyle de la protéine en croissance (Fig. 14). Trois évènements sont requis. Tout d'abord un ARNt chargé est amené au site A du ribosome en concordance avec le codon qui y est exposé. Puis une liaison peptidique est établie entre l'ARNt chargé au site A et la chaîne polypeptidique attachée au peptidyl-ARNt au site P. Il y a alors transfert de la chaîne polypeptidique en croissance de l'ARNt situé au site P vers celui situé au site A. Ce dernier est ensuite transloqué au site P alors que l'ARNt qui s'y trouvait maintenant déchargé de sa chaîne polypeptidique est déplacé vers le site E. Après dissociation de l'ARNt du site E, on revient à l'étape initiale du cycle. L'ARNt entrant dans le site P provoque un changement de conformation permettant à un nouvel ARNt aminoacylé de se fixer au site A, grâce au facteur d'élongation-Tu (EF-Tu) qui « l'escorte » jusqu'au ribosome (Kothe et al., 2004). À chaque étape du cycle, les ARNt chargés diffusent de manière stochastique dans le site A du ribosome. Lorsqu'un ARNt porteur de l'anticodon complémentaire du codon de l'ARNm se fixe au site, il y a appariement.

La formation de la liaison peptidique, catalysée par l'ARNr 23S (présent dans la sous-unité 50S), lie le polypeptide en croissance à l'acide aminé dans le site A (Erlacher et al., 2006). La translocation, catalysée par le facteur d'élongation G (EF-G), permet le déplacement des ARNt et des codons correspondants du site A vers le site P et du site P vers le site E, continuant ainsi la traduction jusqu'à ce qu'un codon stop soit atteint sur l'ARNm (Mohr et al., 2002).

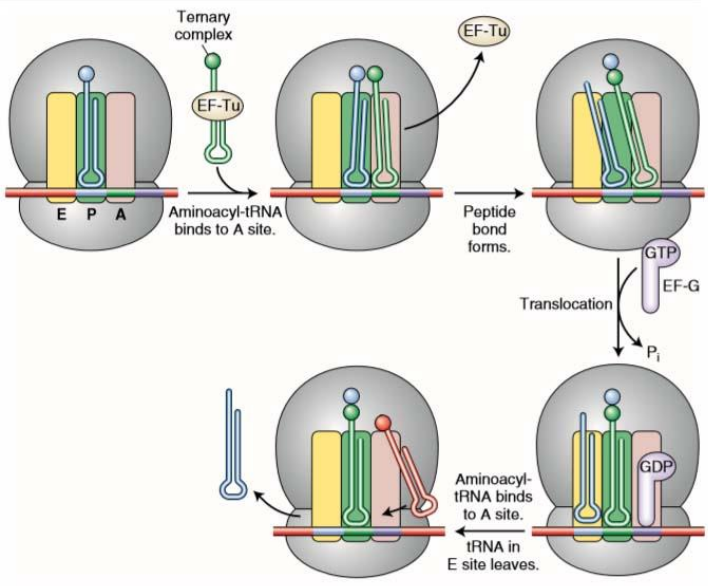


Figure 14 : Mécanisme d'élongation de la traduction chez les bactéries.

Un complexe ternaire composé d'un ARNt aminoacylé attaché à un facteur EF-Tu se fixe sur le site A. Lorsque son acide aminé s'est lié à la chaîne polypeptidique en croissance, un facteur EF-G se lie au site A tout en poussant les ARNt et leurs codons d'ARNm vers les sites E et P. Extrait de ***« Introduction to genetic analysis, Tenth edition 2012 »***

1.1.2.2.5 La terminaison

La terminaison se produit lorsqu'un des trois codons STOP (UAA, UAG et UGA) entre dans le site A (Fig. 15). Ces codons ne sont reconnus par aucun ARNt du code génétique standard, mais par des protéines appelées facteurs de libération (RF), RF1 qui reconnaît les codons stop UAA et UAG et RF2 qui reconnaît les codons stop UAA et UGA (Rodnina, 2018). Ces facteurs déclenchent l'hydrolyse de la liaison ester dans le peptidyl-ARNt et donc la libération de la protéine nouvellement synthétisée du ribosome. Un troisième facteur de libération, RF3, catalyse la libération de RF1 et RF2 à la fin du processus de terminaison (Rodnina, 2018).

Il est à noter que plusieurs ribosomes peuvent traduire simultanément un même ARN. Ce complexe ARNm associé à plusieurs ribosomes est alors appelé « polysome » (Munro et al., 1964)**.** Par ailleurs, les bactéries n'ayant pas de noyau, il arrive que les étapes de transcription et traduction soient couplées, c'est-à-dire que l'ARN est traduit alors qu'il est encore en cours de synthèse (Stent, 1964)**,** (Irastortza-Olaziregi & Amster-Choder, 2020)).

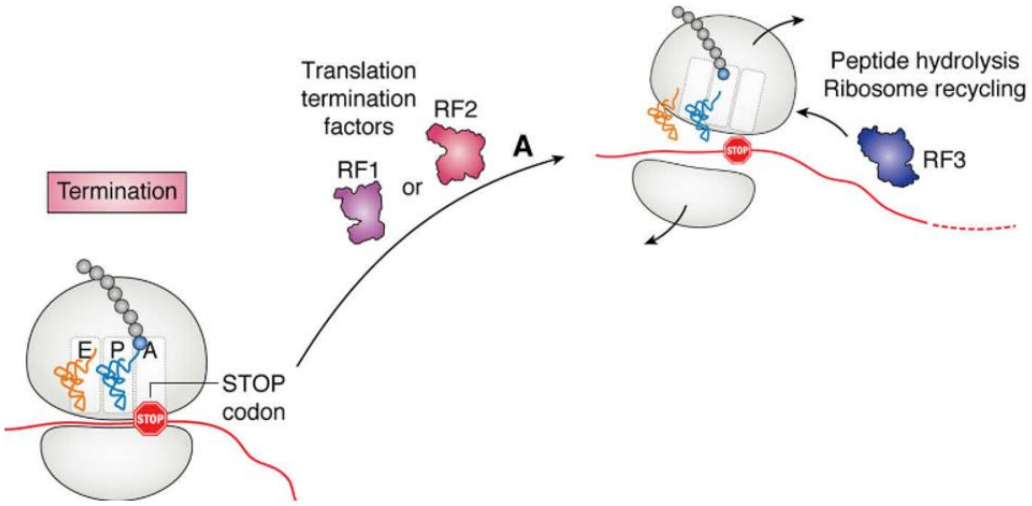


***Figure 15 :* Mécanisme de terminaison de la traduction chez les bactéries.**

Lorsqu'un ribosome atteint un codon stop en présence de RF1 et RF2, ces facteurs contribuent à l'hydrolyse des peptides et au recyclage des ribosomes. RF1 libère les ribosomes aux codons UAG et UAA, tandis que RF2 libère les ribosomes aux codons UGA et UAA. RF3 facilite le recyclage des ribosomes et la libération de RF1 ou RF2 du ribosome. Extrait de (Tollerson & Ibba, 2020)*.*

1.1.2.3 Régulateurs de l'expression génique

La régulation de l'expression des gènes fait référence aux divers mécanismes employés par la cellule pour moduler la production de produits génétiques, tels que les protéines ou les ARN. Cette régulation peut parfois être complexe et impliquer divers systèmes de régulation dont les acteurs peuvent s'influencer créant ainsi des réseaux de régulation. Les processus permettant la régulation de l'expression génique peuvent agir au niveau de la transcription, en régulant la production de l'ARNm par exemple, mais peuvent également agir sur la production des protéines par la régulation de la traduction ou encore sur le devenir des ARN par leurs modifications post-transcriptionnelles.

Dans les sections suivantes de ce chapitre, nous allons explorer les principaux acteurs de cette régulation chez les bactéries. Toutefois, étant donné leur grand nombre, nous ne prétendons pas être exhaustifs.

1.1.2.3.1 Facteur de transcription

Une classe très importante de régulateurs bactériens est celle des facteurs de transcription (TF) dont font notamment partie les facteurs sigma et qui activent et/ou répriment la transcription de gènes cibles, généralement en réponse à des signaux environnementaux ou cellulaires.

Les TF sont des protéines qui se lient à des séquences spécifiques de l'ADN à proximité de leurs gènes cibles, modulant ainsi l'initiation de la transcription. Les TF peuvent activer ou réprimer la transcription en fonction de l'endroit où ils se lient par rapport au site de départ de la transcription du gène cible (Browning & Busby, 2004). Chaque TF régule un ensemble de gènes, en réponse à des déclencheurs environnementaux et/ou intracellulaires spécifiques. Les facteurs de transcription bactériens sont caractérisés par la présence de deux domaines : un domaine de liaison à l'ADN (DBD) et un domaine de liaison aux ligands (LBD) (Helsen et al., 2012). Le LBD est un domaine qui est capable d'interagir avec des molécules exogènes telles que des ions métalliques ou des composés aromatiques, ainsi que des molécules endogènes telles que des sucres ou certains acides aminés. En interagissant avec ces différents ligands, les facteurs de transcription peuvent changer de conformation ce qui affecte leurs propriétés de liaison à l'ADN. Ces modifications conformationnelles peuvent conduire à leur dissociation de l'ADN ou, au contraire, renforcer l'affinité du facteur de transcription pour sa séquence cible. (Helsen et al., 2012)**.**

Les TF sont regroupés en deux groupes, les TF dits activateurs (Fig. 16 gauche) et les TF dits répresseur (Fig. 16 droite). Les activateurs se lient sur une région régulatrice en amont du promoteur et favorisent le recrutement de l'ARNP par la liaison avec un ou deux α-CTD (Fig. 16a) ou par interaction avec plusieurs sous unité de l'ARNP (Fig. 16b), favorisant donc la transcription du ou des gènes en aval.

Les répresseurs, ont plusieurs mécanismes permettant d'inhiber la transcription. Ils peuvent se fixent sur la région promotrice et empêcher la liaison de l'ARNP par encombrement stérique (Fig. 16c). Ils peuvent également inhiber la transition du complexe fermé au complexe ouvert. Le répresseur se lie alors soit en amont de l'ARNP, soit sur la face opposée de l'hélice d'ADN ce qui permet la liaison de l'ARNP, mais inhibe la fusion des brins d'ADN médiée par l'ARNP dans la région d'initiation (Fig. 16d). Finalement, les répresseurs peuvent inhiber la libération du promoteur. Le répresseur permet la fixation de l'ARNP et la formation de complexes ouverts. Cependant, le déplacement de l'ARNP est bloqué soit par un contact avec un répresseur lié en amont, soit par une collision avec un répresseur lié en aval (Fig. 16 e). Ce dernier cas peut être illustrer par l'opéron lactose (Fig. 17), ou en absence de lactose, la forme active du répresseur se lie à la région opérateur localisé en aval du promoteur et empêche ainsi l'ARNP de transcrire les gènes en aval.

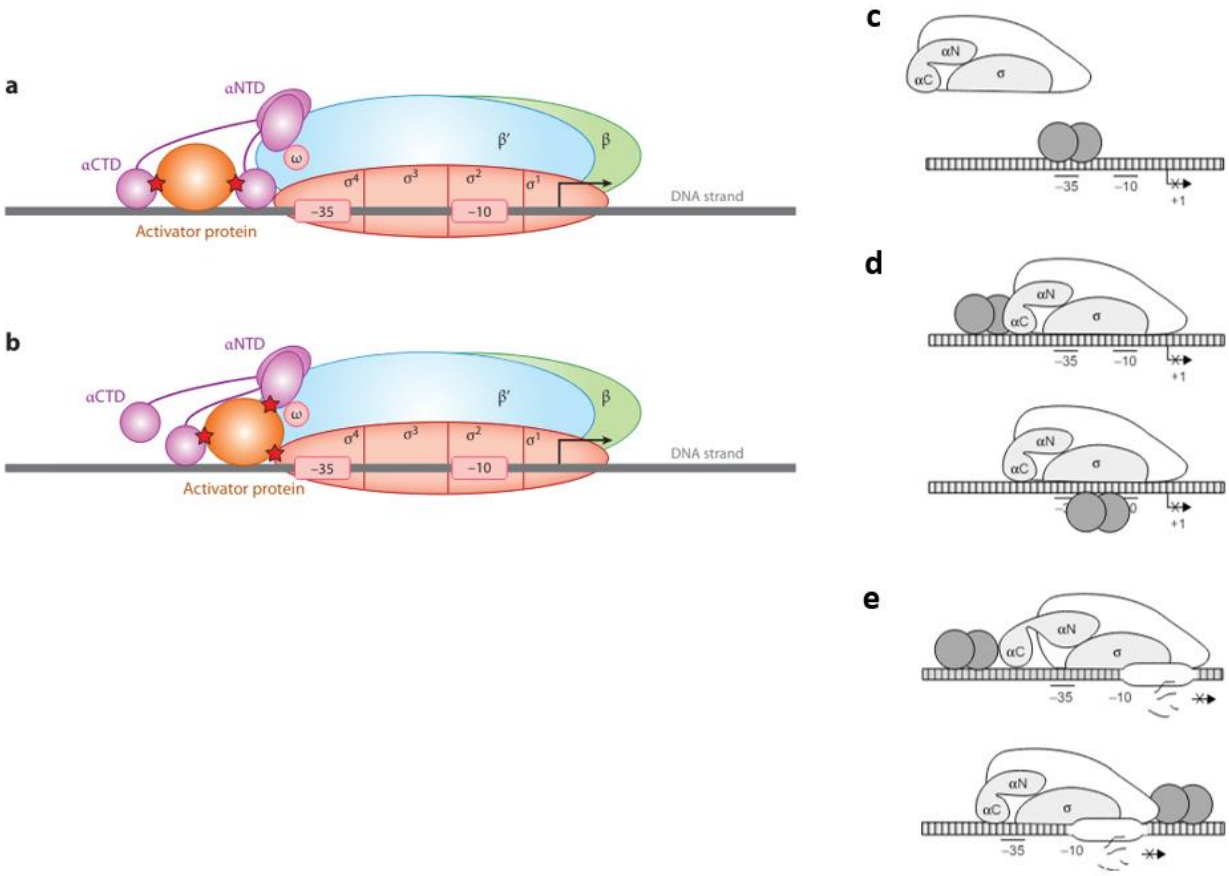


Figure 16 : ***Mécanisme d'activation et de répression de la transcription d'un gène par les facteurs de transcription.*** *Activation par recrutement (gauche). Extrait de* (D. J. Lee et al., 2012) *Représentation de différents mécanismes de répression (droite) Extrait de* (Rojo, 2001).

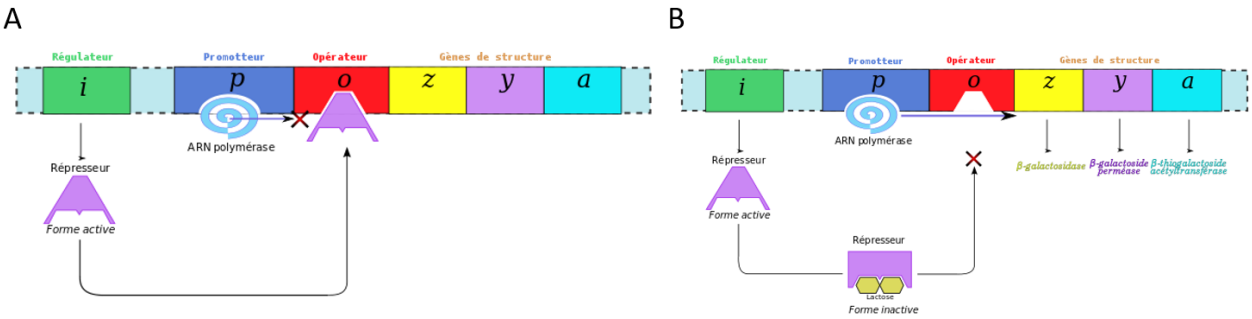


Figure 17 : ***Régulation négative de l'opéron lactose.*** *(**A**) En absence de lactose, le répresseur se lie sur la région opérateur et bloque l'ARNP. (**B**) En présence de lactose, le répresseur est sous forme inactive et l'ARNP peut transcrire les gènes aval.* ***Extrait de wikipedia « Par Nicolas Grandjean »***

Le génome d'*E. coli* code pour environ 270 TF, ce qui représente 6 % des gènes codants pour des protéines de cet organisme (Madan Babu & Teichmann, 2003)**.** Selon, les souches, les génomes de *S. aureus* code pour environ 112 TF soit 5% des gènes codants (Ibarra et al., 2013).

1.1.2.3.2 Les ARN régulateurs de l'expression génique

1.1.2.3.2.1 Riboswitch

Les riboswitches sont des domaines complexes d'ARN repliés qui servent de récepteurs à des métabolites spécifiques (Serganov & Nudler, 2013)**.** Ces domaines se trouvent dans les parties non codantes de divers ARN (UTR), où ils contrôlent l'expression des gènes en exploitant les changements structurels allostériques provoqués par la liaison des métabolites.

Les riboswitches sont composés de deux domaines : le domaine aptamère et la plateforme d'expression (Tucker & Breaker, 2005). Le domaine aptamère agit comme un récepteur qui lie spécifiquement un ligand (Fig. 18 partie verte). La plateforme d'expression agit directement sur l'expression des gènes grâce à sa capacité à basculer entre deux structures secondaires différentes en réponse à la liaison d'un ligand (Fig. 18 partie rouge).

La plupart des riboswitches, possèdent deux mécanismes principaux de contrôle. Le premier mécanisme régule la transcription de l'ARN. En fonction du ligand, le riboswitch peut être impliquer dans la formation soit d'une tige boucle terminatrice intrinsèque en amont d'une série de U (Fig. 18A) soit d'une tige boucle anti-terminatrice (Fig. 18B). Dans le premier cas, la tige boucle intrinsèque créer en amont d'une série de U va interrompre la transcription prématurément en induisant la dissociation de l'ARNP du duplex ARN-ADN. Dans le deuxième cas, la tige boucle anti-terminatrice va empêcher la formation d'une boucle intrinsèque et ainsi permettre la transcription du gène en aval (Mandal & Breaker, 2004)**.**

Le second mécanisme agit au niveau de l'initiation de la traduction. Tout comme les changements allostériques dans la structure des aptamères peuvent contrôler la formation de tiges terminatrices et anti-terminatrices intrinsèques, la liaison d'un ligand peut entraîner des changements structurels dans les ARNm afin de contrôler l'accès au RBS (Fig. 18C et 18D) (Mandal & Breaker, 2004).

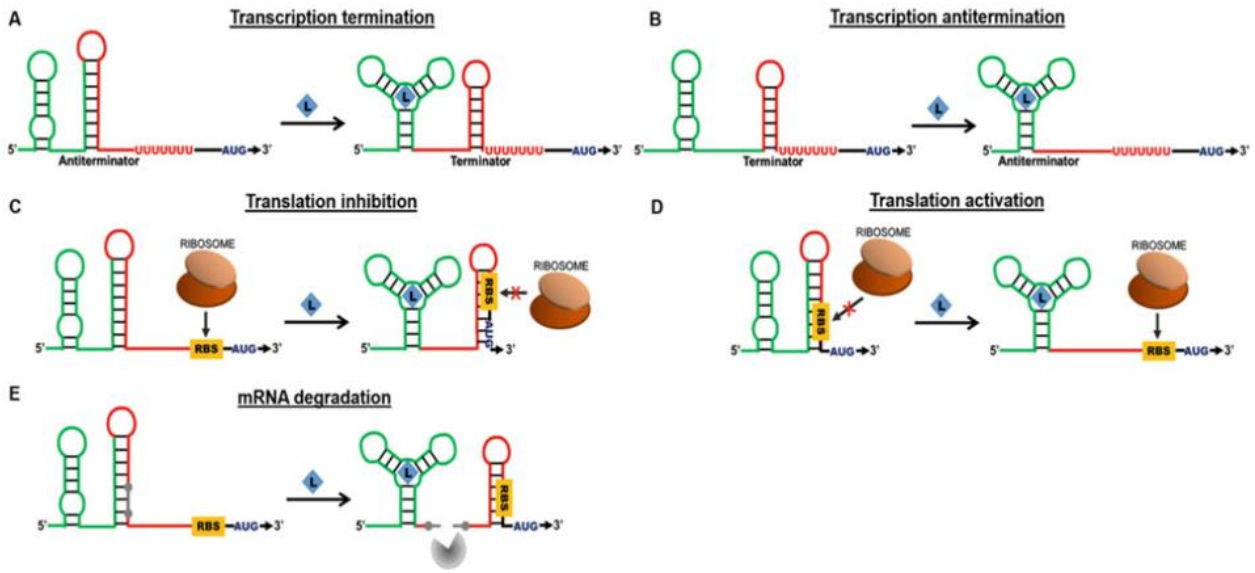


***Figure 18 :* Exemples de mécanismes de riboswitches.**

*Les riboswitches bactériens régulent à la hausse ou à la baisse l'expression des gènes au niveau de la transcription, de la traduction ou de la dégradation de l'ARN. L'aptamère est coloré en vert. La tige boucle terminatrice est colorée en rouge. (**a**) La liaison du ligand (L) à l'aptamère entraîne la formation d'une tige boucle terminatrice conduisant à l'arrêt de la transcription. (**b**) En l'absence de ligand, un terminateur intrinsèque arrête la transcription et la liaison du ligand à l'aptamère induit un changement de conformation qui forme un antiterminateur, permettant l'expression du gène en aval. (**c**) La liaison du ligand à l'aptamère entraîne la formation d'une structure alternative qui séquestre le site de liaison du ribosome (RBS) et empêche l'initiation de la traduction. (**d**) Le RBS est séquestré en l'absence de ligand en l'absence de ligand. Lorsque le ligand se lie à l'aptamère, l'ARN subit un changement de conformation, libérant le RBS et permettant la traduction. (**e**) Lorsque le ligand se lie à l'aptamère, le riboswitch adopte une conformation qui expose le site de clivage de la RNase (coloré en gris), ce qui entraîne la dégradation de l'ARN.* Extrait de (Harbaugh et al., 2019)

Au total, il existe plus de 20 classes de riboswitches en fonction de la nature de leur ligand. Certains riboswitches sont également capables de changer de conformation en fonction de la température. D'autres riboswitches, appelés les T-box, sont capable de reconnaître des ARNt par liaison avec sa boucle de l'anticodon (Green et al., 2010)**,** (Gutiérrez-Preciado et al., 2009)**.**

1.1.2.3.2.2 ARN anti-sens (cis et trans et actifs)

Les ARN anti-sens ou ARNas sont de petits ARN (~ entre 50 et 100 nucléotides) non codants qui s'apparient à des ARN cibles par complémentarité pour réguler leur expression au niveau de la transcription, de la traduction et de leur dégradation. Ces ARNas peuvent être codés en cis ou en trans.

Les ARNas cis-actifs sont codés sur le brin d'ADN opposé à leur ARN cible et ont donc une forte complémentarité avec leurs cibles. Chez *S. aureus*, des expériences de séquençage de petits ARN ont permis de montrer que pour plus de 75 % des ARN, il existe un ARNas codé en cis complémentaire. Ceci indiquerait la présence d'une transcription anti-sens (Lasa et al., 2011), qui permettrait de réguler une grande partie des ARN. Plus tard, cette transcription anti-sens, qualifiée de « transcription pervasive » (Wade & Grainger, 2014)**,** a été identifiée dans plusieurs bactéries (voire la plupart) et notamment chez *E. coli* (Thomason et al., 2015)**.** Pour le moment, toutes les études faites sur les ARNas cis-actif décrivent une régulation négative de l'expression des gènes par une inhibition de la transcription ou de la traduction ou en induisant la dégradation de leur ARN (Fig. 19A)

Les ARNas trans-actifs sont eux codés à partir d'un gène différent du gène codant pour leurs ARN cibles. Cela signifie que dans la plupart des cas les ARN trans-actifs ont une plus faible complémentarité avec leurs cibles que les ARNas cis-actifs. Ces ARNas codé en trans permettent également d'exercer un contrôle négatif sur la traduction (E. G. Wagner & Simons, 1994)**,** (Carrier et al., 2018) ainsi qu'une induction de la dégradation des ARN cibles (Bandyra et al., 2012). De plus, des études ont également montré que dans certains cas, la liaison d'un ARN trans-actif avec un ARN messager (ARNm) induisait un changement de conformation permettant la libération de la séquence RBS préalablement inaccessible (Papenfort & Vanderpool, 2015) (Fig. 19B)**.** Pour finir il a été montré que les ARNas trans-actifs pouvait former des duplex avec des ARNas cis-actifs permettant de réguler l'activité de ces derniers (Miyakoshi et al., 2015).

Aujourd'hui, il a été montré que les ARNas forment une classe très abondante de régulateurs de l'expression génique et jouent un rôle clé dans le métabolisme des bactéries, dans la réponse au stress ainsi que dans la virulence (Brantl, 2015)**.** Chez *S. aureus*, plusieurs centaines d'ARNas ont été identifiés et la caractérisation fonctionnelle d'un sous-ensemble a révélé qu'ils participent à tous les processus biologiques, y compris la virulence, l'adaptation métabolique et la résistance aux antibiotiques (Menard et al., 2022).

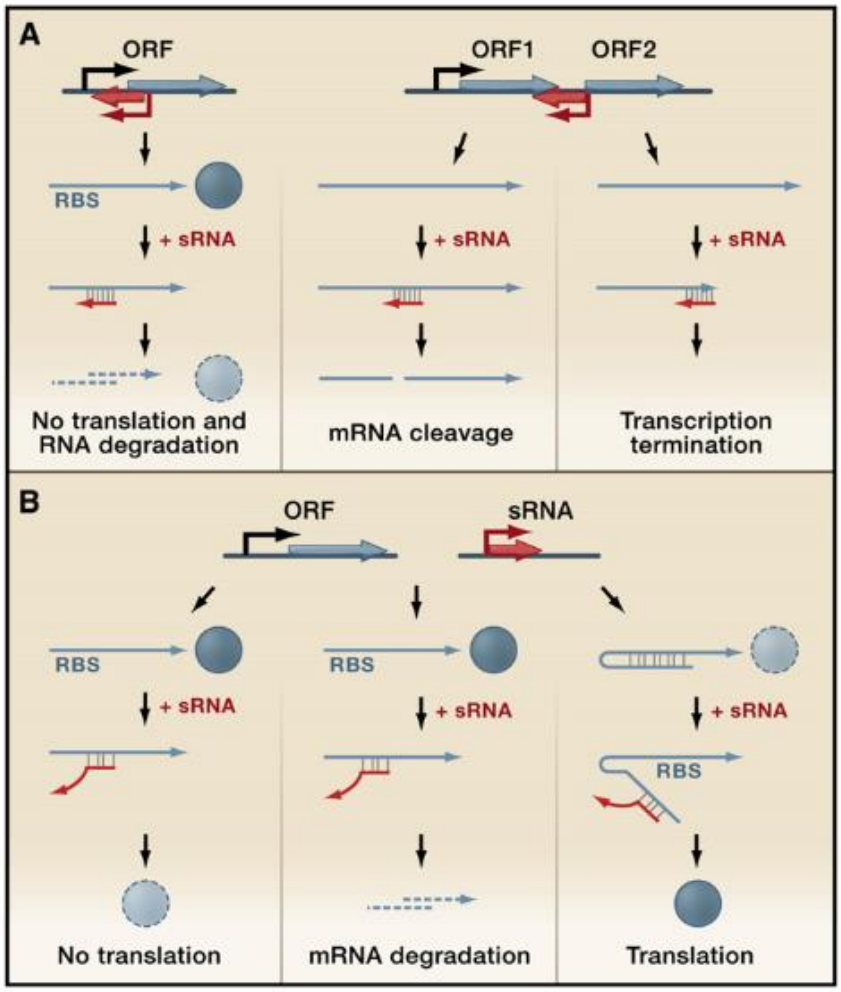


***Figure 19 :* Modes d'action des ARNas.**

*(**A**) Deux configurations possibles d'ARNas codés en cis (rouge) et de leurs ARN cibles (bleu), qui partagent une grande complémentarité. Panneau de gauche : Un ARNas codé à l'opposé du 5'-UTR de son ARNm cible. L'appariement des bases inhibe la liaison avec le ribosome et entraîne souvent la dégradation de l'ARNm cible. Panneau de droite : L'ARNas codé à l'opposé de la séquence séparant deux gènes dans un opéron. L'appariement des bases de l'ARNas peut permettre le recrutement de ribonucléases sur la région et provoquer le clivage de l'ARNm, ou l'ARNas peut provoquer l'arrêt de la transcription, ce qui entraîne une baisse de l'expression du gène. (**B**) Les gènes codant pour des ARNas codés en trans (rouge) ne sont pas situés sur les gènes codant pour leurs ARN cibles (bleu) et n'ont qu'une complémentarité limitée. L'ARNas trans peut agir négativement en s'appariant aux bases de l'UTR et en bloquant la liaison avec le ribosome (Panneau de gauche) et/ou en ciblant le duplex ARNas-ARNm pour qu'il soit dégradé par les ribonucléases (Panneau du milieu). L'ARNas codé en trans peut agir positivement (Panneau de droite) en empêchant la formation d'une structure inhibitrice, qui séquestre le ribosome. Extrait de* (L. S. Waters & Storz, 2009)

1.1.2.3.2.3 ARN CRISPR

Une classe unique d'ARN régulateurs récemment découverts est celle des ARN CRISPR (ARNcr), qui assurent la résistance aux bactériophages.

Les séquences CRISPR ont été trouvées dans 45 % des bactéries et 87 % des archées présent dans la base de données CRISPR (*https://crispr.i2bc.paris-saclay.fr/*) en 2017 (Ishino et al., 2018), ce qui souligne leur importance à grande échelle. Un locus CRISPR est une région spécifique du génome composée de trois éléments principaux. Premièrement, il contient des séquences répétées palindromiques courtes et régulièrement espacées, généralement de 24 à 48 nucléotides, qui se répètent de manière régulière et peuvent former des structures en épingle à cheveux lorsqu'elles sont transcrites en ARN (Fig. 20 triangle noirs). Deuxièmement, entre ces répétitions, se trouvent des séquences d'ADN non répétitives appelées espaceurs (spacers). Ces espaceurs sont des fragments d'ADN exogène, comme ceux provenant de phages ou de plasmides, intégrés dans le locus CRISPR après une infection et fonctionnant comme une mémoire immunitaire. Troisièmement, adjacents au locus CRISPR, se trouvent des gènes associés appelés gènes Cas (CRISPR-associated system), qui codent pour des protéines essentielles aux différentes étapes du mécanisme CRISPR-Cas, y compris l'acquisition des nouveaux espaceurs, le traitement de l'ARNcr et la destruction de l'ADN exogène (F. Jiang & Doudna, 2015).

Lorsqu'une bactérie est infectée par de l'ADN étranger (comme celui d'un phage), des fragments de cet ADN sont intégrés dans le locus CRISPR de son génome. Ces fragments sont insérés entre des séquences répétées palindromiques, il s'agit des espaceurs. Le locus CRISPR est transcrit en un long pré-ARNcr contenant à la fois les séquences répétées et les espaceurs. Ce pré-ARN est ensuite clivé par des endoribonucléases spécifiques, comme la protéine Cas, pour produire des ARNcr matures, chacun contenant une séquence d'espaceur unique dérivée de l'ADN exogène. Les ARNcr matures se complexent avec des protéines Cas. Ce complexe guide l'activité de la protéine Cas pour reconnaître et dégrader l'ADN exogène. L'ARNcr agit comme une sorte de guide grâce à sa séquence complémentaire au fragment d'ADN cible. Lorsque ce complexe rencontre l'ADN exogène correspondant, l'ARNcr s'hybride à l'ADN étranger par complémentarité de base. La protéine Cas, dotée d'une activité endonucléase, clive alors l'ADN exogène, neutralisant ainsi l'infection (F. Jiang & Doudna, 2015).

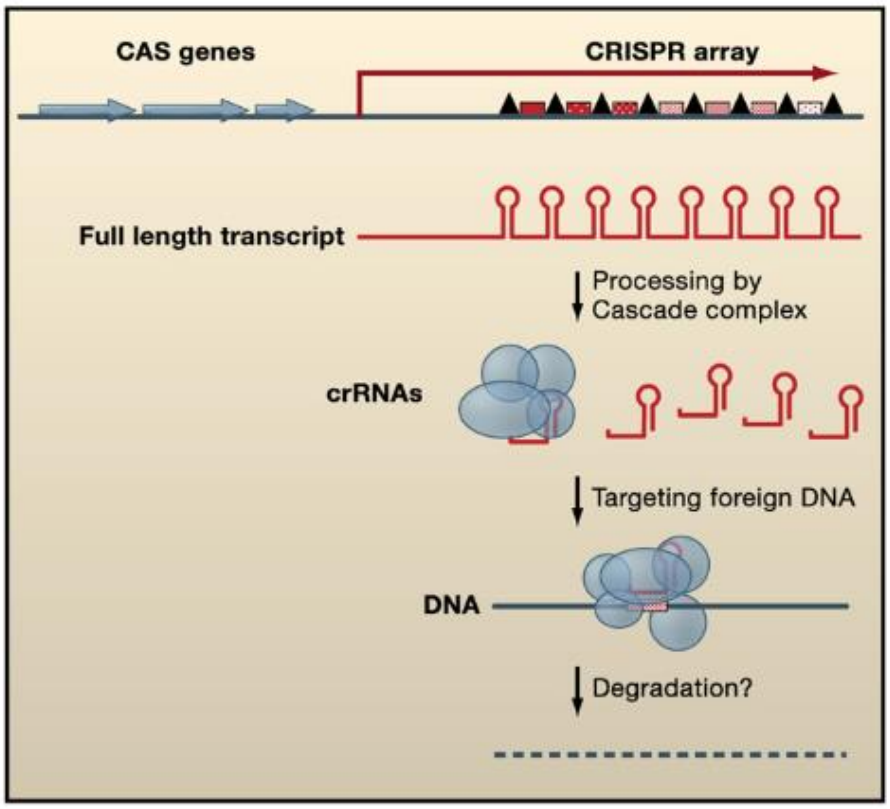


***Figure 20 :* Fonctionnement des ARN CRISPR.**

Les réseaux CRISPR sont composés de répétitions d'ADN (triangles noirs) séparées par des espaceurs uniques (boîtes mouchetées rouges). Les gènes CAS (bleus), qui codent des protéines qui interviennent dans le traitement de l'ARN CRISPR et/ou le silencing de l'ADN sont situés à proximité. Les réseaux CRISPR sont initialement transcrits sous la forme d'un long ARN, qui est ensuite traité par le complexe CAS (cercles et ovales bleus) en unités individuelles d'espaceurs répétés, appelées ARNcr. Extraite de (L. S. Waters & Storz, 2009)

1.1.2.3.3 Protéines de liaison à l'ARN

Les protéines de liaison à l'ARN (RBP) sont des protéines qui comme leur nom l'indique peuvent se lier aux ARN pour former des ribonucléoprotéines. Elles possèdent en général au moins un domaine de reconnaissance et de liaison à l'ARN.

Des RBP sont retrouvées dans tous les domaines du vivant. Ces RBP sont impliquées dans de nombreux système de régulation positif ou négatif de l'expression des gènes que ce soit au niveau de la transcription, de la dégradation des ARN ou encore de la traduction (Fig. 21).

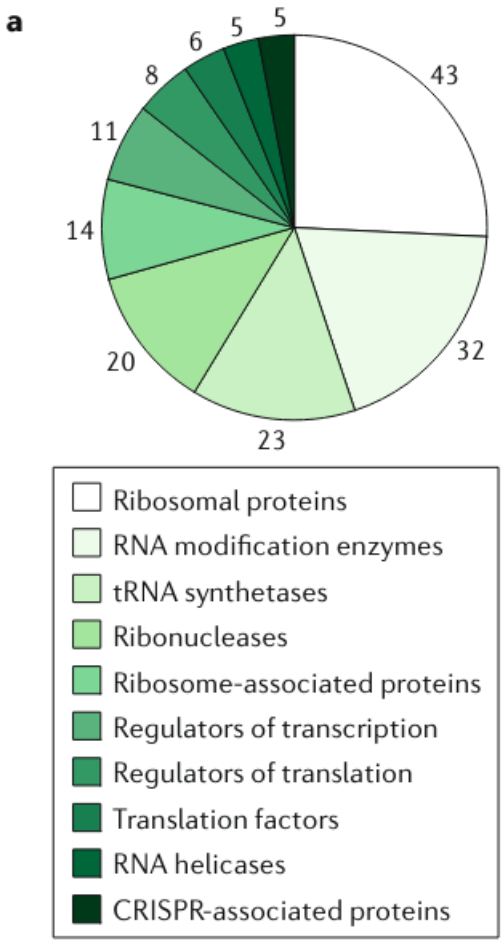


Figure 21 : Protéines de liaison à l'ARN chez E. coli. *Diagramme camembert de toutes les protéines d'E. coli classées comme « RNA binding » selon la Gene Ontology, réparties dans les catégories fonctionnelles indiquées. Extrait de* (Holmqvist & Vogel, 2018)

La régulation de la transcription par les RBP se fait principalement au niveau de l'étape de terminaison. La RBP, impliquée dans ce processus, la plus étudiée chez les bactéries est la protéine Rho qui, comme déjà évoqué, induit l'arrêt de la transcription. Par exemple, chez *E. coli*, il a été montré que Rho était liée à la terminaison d'environ 25 % des unités de transcription (Cardinale et al., 2008) (Holmqvist & Vogel, 2018). D'autres RBP permettent également de réguler cette étape, comme par exemple les protéines BglG et EutV qui induisent la formation de boucle anti terminatrice, ou comme la protéine NusA qui empêche la fixation du facteur Rho (Fig. 22B) et peut également induire la formation de tige boucle intrinsèque (Fig. 22A).

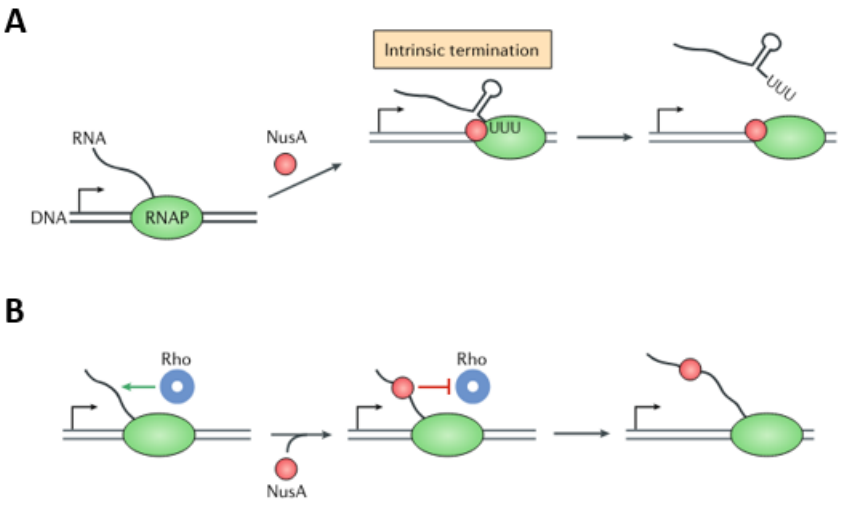


Figure 22 : ***Régulation de la terminaison intrinsèque*** *(**A**) et Rho-dépendante (**B**) par la RBP NusA. Extrait de* (Holmqvist & Vogel, 2018)

La plupart des RBP qui affectent la traduction fonctionnent au niveau de l'initiation de la traduction, c'est-à-dire sur l'association de la sous-unité ribosomique 30S avec l'ARNm au niveau du RBS. Les RBP comme Csr ou Rsm, peuvent entrer directement en compétition avec la sous-unité 30S pour la liaison à un ARNm ou induire un changement structurel de l'ARNm qui modifie l'accessibilité du RBS. Dans des cas ou le RBS serait masqué dans une structure secondaire, certaines RBP comme Hfq et S1 peuvent également se fixer sur la structure et la déstabiliser pour activer la traduction du gène en aval.

Par exemple, la protéine S1 favorise généralement la synthèse protéique en dépliant les ARNm structurés pour permettre l'arrimage de la sous-unité 30S et le positionnement optimal du codon de départ (Duval et al., 2013). Lorsque S1 est en excès, sa forme libre (non lié au ribosome) se lie sur son propre ARNm afin d'inhiber sa traduction et donc la production de protéines S1 supplémentaires.

Un autre mécanisme important par lequel les bactéries contrôlent la traduction des ARNm implique le recrutement de ARNas par les RBP, dont le principal exemple est le réseau de régulation formé par la protéine Hfq. En effet, le rôle principal de Hfq consiste à guider les ARNas trans-actifs vers les ARNm qu'ils régulent (Gorski et al., 2017). Cette action de guidage est très importante puisque par exemple, chez *E. coli*, seulement 100 ARNas peuvent reconnaître plus de 25 % de tous les

ARNm grâce à Hfq, ce qui représente des milliers d'interactions potentielles (Holmqvist & Vogel, 2018), (Vogel & Luisi, 2011), (Gorski et al., 2017), (Tree et al., 2014). Le modèle actuel propose que Hfq est capable de lier simultanément les ARNas et les ARNm. Un complexe ARNas-ARN se forme si les deux partenaires présentent une complémentarité suffisante. Le duplex d'ARN ainsi formé séquestre généralement le RBS, empêchant la traduction de l'ARNm.

De plus, la biogenèse et la fonction des ribosomes s'appuient sur plusieurs autres RBP. En effet, les protéines ribosomales, qui constitue le ribosome, font partie des RBP les plus conservées puisque certaines étaient déjà présentes dans le dernier ancêtre commun universel (LUCA) (Fox, 2010). Les RBP impliquées dans la dégradation des ARN sont nombreuses et variés. Elles peuvent être des ribonucléases et des protéines CAS qui dégradent les ARN. Les hélicases sont également des RBP et permettent de désenrouler les structures secondaires d'ARN pour permettre le traitement des ARN (dégradation, maturation, traduction, *etc*).

1.1.2.3.4 Processing des ARN

La synthèse et la dégradation de l'ensemble des ARN (ARNm, ARNt, ARNr et ARNnc) font appel à des voies de maturation, de modification et de contrôle-qualité. L'ensemble des voies sont en général regroupées sous le terme « processing de l'ARN ». Nous décrirons donc ces différentes voies avant d'aborder la description des machineries de dégradation des ARN.

Les ribonucléases (RNases) sont des enzymes clés de la régulation post-transcriptionnelle, impliquées dans la maturation et la dégradation des ARN. Elles agissent soit en interne sur la molécule d'ARN en tant qu'endoribonucléases, soit en tant qu'exoribonucléases en attaquant l'ARN à son extrémité 5' ou 3' pour initier la dégradation. Pour contrôler ces fonctions importantes, certaines RNases agissent au sein de complexes multiprotéiques.

Les bactéries à Gram positif et à Gram négatif possèdent un important répertoire de RNases dont le nombre et la composition varient d'une espèce à l'autre. Par exemple, l'organisme modèle à Gram négatif, *E. coli*, possède 15 RNases (Tejada-Arranz et al., 2020), alors que l'organisme modèle à Gram positif, *Bacillus subtilis*, possède 20 RNases (Cheng & Deutscher, 2003, 2005). La diversité du contenu en ribonucléases en fonction des espèces est visible sur la Figure 23 qui présente la présence/absence des homologues à l'ensemble des ribonucléases identifiées chez *E. coli* et *B. subtilis*. Dans ce manuscrit, nous avons fait le choix de décrire les ribonucléases majeures au niveau des processus de maturation et de dégradation dans lesquels elles interviennent, plutôt que décrire séparément chacune d'entre elles.

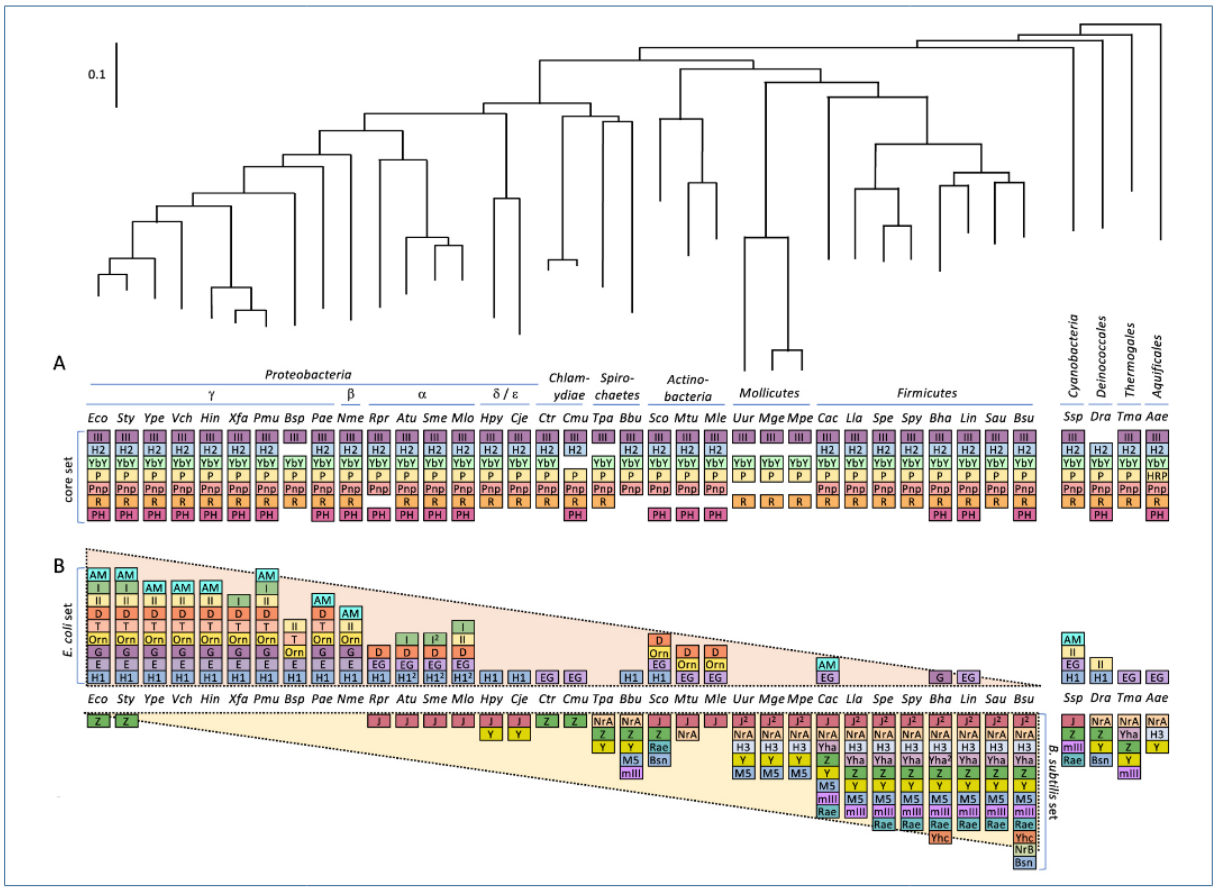


***Figure 23 :* Phylogénie des ribonucléases chez les bactéries.**

*(**A**) L'ensemble des ribonucléases « cœur » conservées dans la plupart des bactéries.*

*(**B**) Distribution de l'ensemble « E. coli » et de l'ensemble « B. subtilis » des ribonucléases dans l'arbre phylogénétique. Les abréviations des ribonucléases sont les suivantes : III = RNase III; P = RNase P; H2 = RNase HII; H1 = RNase HI; H3 = RNase HIII; E = RNase E; G = RNase G; EG = RNase E/G; YbY = YbeY; I = RNase I; Z = RNase Z; M5 = RNase M5; Y = RNase Y; mIII = Mini-III; Rae = Rae1; Maz = MazF; Rel = RelE; Yoe = YoeB; Hig = HigB; YfQ = YafQ; Yhv = YhaV; LS = RNase LS, Mqs = MqsR; YfO = YafO; Bsn = RNase Bsn; Bar = Barnase; AM = RNase AM; R = RNase R; Pnp = polynucleotide phosphorylase; PH = RNase PH; D = RNase D; Orn = oligoribonuclease; II = RNase II; T = RNase T; J = RNase J; BN = RNase BN; Yha = RNase YhaM. Les nombres en exposant font référence aux nombres de paralogues identifiés.Les abréviations des organismes sont les suivantes: Eco = Escherichia coli; Sty = Salmonella typhimurium; Ype = Yersinia pestis; Vch = Vibrio cholera; Hin = Haemophilus influenza; Xfa = Xylella fastidiosis; Pmu = Pasteurella multocida; Bsp = Buchnera sp.; Pae = Pseudomonas aeruginosa; Nme = Neisseria meningitidis; Rpr = Rickettsia prowazekii; Atu = Agrobacterium tumefaciens; Sme = Sinorhizobium meliloti; Mlo =Mesorhizobium loti; Hpy Helicobacter pylori; Cje =Campylobacter jejuni; Ctr =Chlamydia trachomatis; Cmu = Chlamydia muridarum; Tpa = Treponema pallidum; Bbu = Borrelia burgdorferi; Sco = Streptomyces coelicolor; Mtu = Mycobacterium tuberculosis; Mle = Mycobacterium leprae; Uur = Ureaplasma urealyticum; Mge = Mycoplasma genitalium; Mpe = M. pneumoniae; Cac = Clostridium acetobutylicum; Lla = Lactococcus lactis; Spe = Streptococcus pneumoniae; Spy = Streptococcus pyogenes; Bha = Bacillus halodurans; Lin = Listeria innocua; Sau = Staphylococcus aureus; Bsu = Bacillus subtilis; Ssp = Synechocystis sp.; Dra = Deinococcus radiodurans; Tma = Thermotoga maritima; Aao = Aquifex aeolicus.*

Extrait de (Condon et al., 2021)

1.1.2.3.4.1 Maturation

Chez les eucaryotes les ARNm ont besoin d'être maturés (épissage, ajout d'une coiffe m^7-G et d'une queue poly-A). Chez les bactéries, de tels mécanismes sont presque exclusivement absents, mais certains ARN non traduits ont tout même besoin de subir des maturations pour devenir fonctionnels, notamment les ARN ribosomiques et les ARN de transfert.

1.1.2.3.4.1.1 Processing des ARN de transferts

Les ARNt des bactéries sont synthétisés sous forme de molécules précurseurs plus longues (pré-ARNt), dont certaines contiennent plusieurs séquences individuelles d'ARNt. Certains ARNt peuvent aussi être dans les transcrits des pré-ARNr. Ces précurseurs d'ARNt subissent une série de transformations qui transforment un transcrit primaire en une molécule mature et fonctionnelle. La maturation de l'ARNt consiste à éliminer les résidus spécifiques des extrémités 5′ et 3′ du précurseur. Au niveau de l'extrémité 5′, cette élimination est réalisée par un seul clivage endonucléolytique par la quasi-universelle RNase P, qui génère l'extrémité 5′ mature de l'ARNt (Fig. 24a et 24b). La RNase P bactérienne est constituée par une chaîne d'ARN, appelé M1 RNA, et par une protéine, la protéine C5. Elle se distingue des autres RNases par le fait qu'il s'agit d'un ribozyme, c'est à que à dire que c'est la molécule d'ARN qui est responsable de l'activité catalytique de l'enzyme (Guerrier-Takada et al., 1983). Le rôle principal de la protéine C5 est de renforcer l'affinité de liaison au substrat et le pouvoir catalytique de l'enzyme ARN M1, probablement en augmentant l'affinité de l'ion métallique dans le site actif.

La maturation de l'extrémité 3′ semble également pouvoir être faite de manière endonucléolytique par la RNase Z en une seule étape (Fig. 24a et 24b). Cependant, chez *E. coli* le motif CCA présent au niveau des extrémités 3' des ARNt inhibe le clivage endonucléolytique de la RNase Z (Perwez & Kushner, 2006; Redko et al., 2007).

La RNase Z est une endonucléase qui appartient à la super famille des metallo-β-lactamase zinc dépendante. Tout comme la RNase P, elle permet de reconnaître le « bras T » et le « bras accepteur » des ARNt et de retirer le 3'-tailer en clivant juste après le nucléotide discriminant (Fig. 24a) (Garber & Gage, 1979), (Castaño et al., 1985), (Mayer et al., 2000). La RNase Z préfère cliver après un tri-nucléotide CCA. Bien que l'efficacité du clivage soit indépendante du nucléotide discriminant, elle dépend fortement du nucléotide situé immédiatement en aval (la cytosine étant la moins préférée). Les séquences 3′-trailer commençant par CC, CCA ou CCA incluant un ou deux nucléotides supplémentaires sont toutes faiblement clivées.

Par conséquent, le motif CCA présent dans tous les ARNt matures est un anti-déterminant général pour l'activité de la RNase Z (Mohan et al., 1999). Cela est logique d'un point de vue physiologique, car cela empêche le recyclage sans fin de l'extrémité 3′ des ARNt matures.

La maturation de l'extrémité 3' peut également se faire de manière exonucléolytique en impliquant un certain nombre d'enzyme différentes (Z. Li & Deutscher, 1996). Chez *E. coli*, ce processus est généralement initié par un clivage endonucléolytique catalysée par la RNase E suivi d'une dégradation 3'-5' par les RNases PH et T. Chez *B. subtilis,* il n'y pas d'homologues de la RNase E, et le processus est directement initié par la dégradation 3'-5' par la RNase PH.

La RNase PH est une exoribonucléase nucléotidyltransférase et contrairement aux enzymes hydrolytiques, il s'agit d'une enzyme phosphorolytique, ce qui signifie qu'elle utilise le phosphate inorganique comme réactif pour cliver les liaisons nucléotide-nucléotide, libérant ainsi des nucléotides diphosphates (NDP). Cette exonucléase joue un rôle majeur dans la dégradation des derniers nucléotides en aval du nucléotide discriminant. L'importance de la RNase PH est accrue chez *B. subtilis* où l'absence d'orthologues de la RNase T empêchent le recouvrement fonctionnel comme chez *E. coli*.

La RNase T est une exonucléase non processive de la famille DnaQ, présente chez les gamma-protéobactéries et qui agit sur l'extrémité 3' des acides nucléiques simple brin. Elle clive spécifiquement les substrats ayant une cytosine à leur extrémité 3' (Hsiao et al., 2014)**,** (Zuo & Deutscher, 2002). La RNase T coupe le résidu AMP après les séquences CCA en 3' des ARNt jusqu'à la séquence CC-3'. Les ARNt trop raccourcis sont réadditionner d'un A terminal 3' par des nucléotidyltransférases (Weiner 2004, Deutscher 1978). Chez E. coli, certaines séquences en 3' des précurseurs d'ARNt, comme ARNtAsp et ARNtVal, contiennent des cytosines en tandem et ne peuvent pas être clivées par la RNase T, nécessitant l'intervention de la RNase PH (Z. Li & Deutscher, 1996).

Dans une moindre mesure, la maturation exonucléolytique des extrémités 3' peut être réalisée par d'autres exonucléases telles que la RNase D et II chez *E. Coli* et RNase R, PNPase et YahM pour *B. Subtilis*. Ces RNases ont en général, peu d'affinité pour les ARNt « classique » mais sont tout de même capable de dégrader leur extrémité 3'. La RNase D par exemple, possèdent une grande affinité pour les ARNt qui n'ont pas ou pas entièrement le motif CCA (Zuo et al., 2005). Les RNase II et R appartiennent à la super famille des RNases R (RNR) et sont des 3'-5' exonucléases hydrolytiques non spécifiques et hautement processives. Les deux enzymes sont impliquées dans la dégradation des ARNm bactériens et possèdent toutes deux une forte affinité pour les séquences polyA (Cheng & Deutscher, 2002).

Contrairement à la RNase II, la RNase R, a la particularité de pouvoir dégrader les ARN structurés, à condition qu'il y ait une extension simple brin en 3' de plus de sept nucléotides (Vincent & Deutscher, 2006)**.** Chez *E. coli*, cela peut s'expliquer par le fait que la RNase R possède, en plus de son activité de dégradation, une activité hélicase lui permettant de dé-enrouler les structures secondaires (Tejada-Arranz et al., 2021). Chez d'autres bactéries comme *H. pylori par* exemple, cette capacité est acquise par interaction avec des enzymes hélicases (Tejada-Arranz et al., 2021)**.**

Comme la RNase PH, la PNPase n'est pas une hydrolase mais une ribonucléase phosphorolytique utilisant le phosphate inorganique pour dégrader les ARN, ce qui lui permet également de réaliser la fonction inverse, à savoir, synthétiser de l'ARN à partir de nucléoside diphosphate, cette activité de synthèse reste cependant très peu active. L'activité de la PNPase n'est pas restreinte aux ARNt puisqu'elle est aussi l'exoribonucléase 3'→5 principale chez *B. subtilis* et des expériences ont montré que sa délétion engendrait une sensibilité accrue au froid, à la tétracycline et des problèmes à effectuer la compétence (Luttinger et al., 1996). La PNPase nécessitent un site de liaison non structuré d'au moins 7-10 nt à l'extrémité 3′ du substrat (Godefroy, 1970)**,** (Symmons et al., 2000)**,** (Fazal et al., 2015). De plus, sa structure en forme de canal, avec son site catalytique situé au fond de ce canal, l'empêche de traiter les ARN double brin ainsi que les structures secondaires (Bechhofer & Deutscher, 2019). La PNPase interviendrait donc principalement après un clivage interne au niveau des extrémités 3' préalablement créées par les endoribonucléases.

Quant à YhaM, il s'agit d'une exoribonucléase 3'→5' identifiée chez les bactéries à Gram positif (Oussenko et al., 2002). Des expériences ont montré qu'elle était capable de dégrader les extrémité 3' des ARNt mais sa délétion n'engendre que de légers effets sur la maturation des ARNt (Wen et al., 2005).

Tous les ARNt ont la séquence CCA à leur extrémité 3′. Cette séquence est le site d'attachement des acides aminés et est donc nécessaire à la fonction de l'ARNt pendant la synthèse des protéines. L'extrémité CCA est codée dans l'ADN de certains gènes d'ARNt, mais pas peut être absente dans certains. Elle est alors ajoutée lors d'une étape de traitement de l'ARN par une enzyme, l'ARNt nucléotidyl-transférase ou « CCAse », qui reconnaît et ajoute la CCA à l'extrémité 3′ de tous les ARNt dépourvus de cette séquence.

Les ARNt se caractérisent aussi par la présence d'un grand nombre de nucléotides non-canoniques, ou nucléotides modifiés. Ces modifications de nucléotides sont diverses et incorporées post-transcriptionnellement par des enzymes spécialisées : méthylation, isomérisation, thiolation, réduction etc.

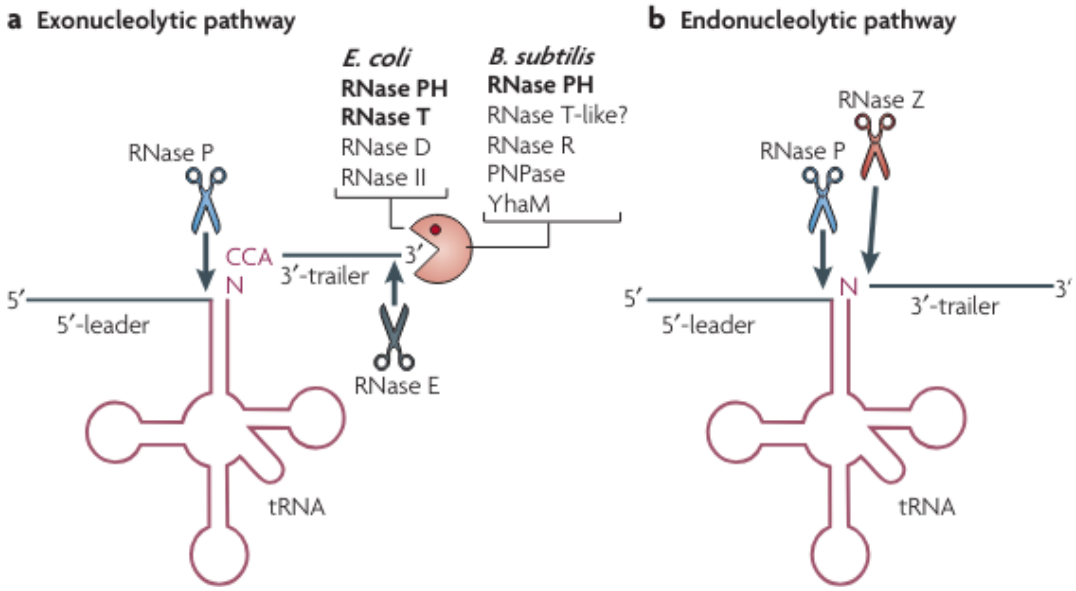


Figure 24 : Voies de maturation de l'ARN de transfert chez les procaryotes.

(a) *Voie exonucléolytique de maturation de l'extrémité 3' de l'ARN de transfert (ARNt) chez E. coli et B. subtilis. Les séquences 5'-leader et 3'-trailer sont indiquées. Les principales exonucléases (symbole Pacman) identifiées jusqu'à présent sont en caractères gras. La maturation de l'extrémité 5' par la RNase P est indiquée par les ciseaux bleus. (**b**) La voie endonucléolytique de la maturation de l'extrémité 3' de l'ARNt. Les clivages endonucléolytiques par la RNase P et la RNase Z sont indiqués par les ciseaux bleus et rouges, respectivement. La RNase Z clive en aval du nucléotide discriminant (N) de l'ARNt précurseur. Extrait de* (Redko et al., 2007)

1.1.2.3.4.1.2 Processing des ARN ribosomiques

Parallèlement à l'assemblage des ribosomes, les 3 ARNr (16S, 23S et 5S), qui sont initialement synthétisés sous forme d'un unique précurseur qui doit ensuite être clivée pour obtenir les trois ARNr pré-matures (Fig. 25) puis seront ensuite une nouvelle fois clivés pour obtenir les ARNr que l'on retrouve dans les unités ribosomales fonctionnelles.

Bien que l'étape initiale de clivage catalysée par la RNase III soit conservée entre *B. subtilis* et *E. coli*, les étapes suivantes de la maturation des ARNr individuels de *B. subtilis* se sont avérées être catalysées par des enzymes absentes d'*E. coli* dans chaque cas étudié jusqu'à présent (Condon et al., 2001)**,** (Britton et al., 2007).

La RNase III initie, ce processus de maturation car elle permet de cliver les ARN double brin, permettant ainsi à d'autres endoribonucléases, qui pour la plupart, ne clivent que l'ARN simple brin, de pouvoir cliver les ARNr (Robertson et al., 1968)**,** (Bram et al., 1980).

L'extrémité 5′ de l'ARNr 16S est maturée par la combinaison de l'activité endoribonucléasique de la RNase E, de la RNase G (extrémité 5') et de YbeY (extrémité 3') chez *E. coli* tandis que chez *B. subtilis*, cette réaction est catalysée par YqfG et par l'activité exoribonucléasique 5′ à 3′ de la RNase J1 (activité développée dans une section ultérieure).

La RNase E est une enzyme clé dans le métabolisme de l'ARN. Elle est impliquée dans la maturation des ARNr (ARN ribosomiques), des ARNt (ARN de transfert) et de manière générale dans la dégradation des ARN. La RNase G est un homologue de la RNase E qui clive le précurseur de l'ARNr 16S à une position spécifique pour générer l'extrémité 5' mature. Et bien que ces deux homologues puissent avoir des fonctions redondantes dans certaines conditions les fonctions de la RNase G ne sont pas aussi étendues que celles de la RNase E.

YqfG (*B. subtilis*) et YbeY (*E. coli*) sont des orthologues (72% de similarité) et sont des endoribonucléases qui influence la maturation des trois ARNr, avec un effet particulièrement fort sur la maturation des extrémités 3′ de l'ARNr 16S (Davies et al., 2010). En outre, YbeY est la seule endoribonucléase décrite qui a la capacité remarquable de cliver spécifiquement l'ARNr dans les ribosomes 70S défectueux, mais pas dans les sous-unités 30S et 50S dissociées dérivées de ces ribosomes 70S défectueux. Cette spécificité unique du substrat d'YbeY la distingue des RNases actuellement connues (Jacob et al., 2013).

L'ARNr 5S d'*E. coli* est maturé par une combinaison de RNase E et de RNase T, tandis que *B. subtilis* dispose d'une endoribonucléase double brin appelée RNase M5 dédiée à cette tâche (Condon et al., 2001). En effet, il a été montré qu'à part l'ARNr 5S, la RNase M5 n'avait que peu, voir pas de substrat chez *B. subtlis* (Condon et al., 2002)**.**

Enfin, chez *B. subtilis*, l'ARNr 23S est maturé par la mini-RNase III, une endoribonucléase capable de cliver les ARN double brins et qui serait plus efficace pour cliver l'ARNr pré-23S dans les sous-unités 50S assemblées (Redko et al., 2008). Cette RNase est absente chez *E. coli* et plus largement, absente chez les protéobactéries.

Certains de ces clivages peuvent ensuite être suivis d'une dégradation exonucléolytique soit 3'-5' pour *E. coli* par les RNases PH, R, T et II, soit 5'-3' chez *B. subtilis* par la RNase J1.

Ces ARNr doivent ensuite subir une série de modifications nucléolytiques pour générer les molécules matures que l'on trouve dans les particules fonctionnelles. Ces modifications correspondent principalement à des méthylations de bases et de sucres et des conversions d'uridines en pseudo-uridines (Deutscher, 2009)**,** (Giuliano & Engl, 2021).

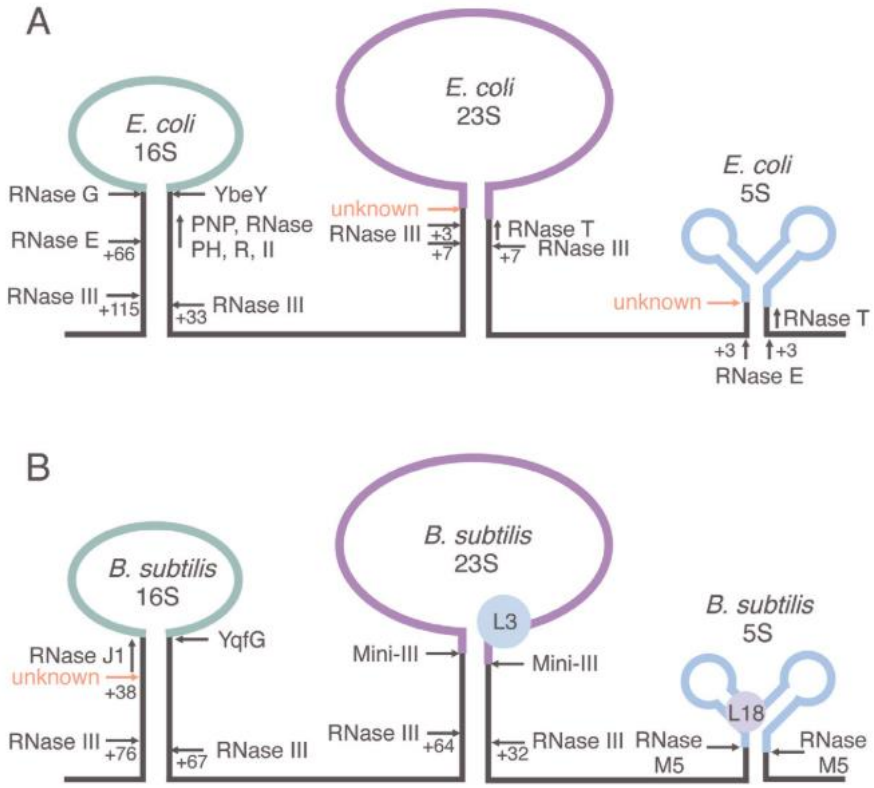


***Figure 25 :* Maturation de l'ARN ribosomal chez Escherichia coli (A) et Bacillus subtilis (B).** *Les ARNr 16S, 23S et 5S matures sont représentés par des lignes épaisses vertes, violettes et bleues, respectivement. Les sites de clivage des principales enzymes de transformation sont indiqués par des flèches et leurs distances par rapport aux extrémités 5' et 3' matures sont indiquées. Les enzymes inconnues sont indiquées en rouge. Les protéines ribosomales qui servent de cofacteurs dans les réactions de transformation de B. subtilis sont représentées par des sphères colorées. Extrait de* (Baumgardt et al., 2018)

1.1.2.3.5 Mécanismes de dégradation des ARN

L'arrêt de la transcription ne suffit pas à réguler l'expression des gènes. En effet, même si le gène n'est plus transcrit, si les ARNm produit avant ne sont pas dégradés et donc toujours présents, ils peuvent être traduits. La dégradation efficace de l'ARN a donc été reconnue comme absolument essentielle pour la régulation de l'expression génique. Cependant, bien que les machineries de dégradation de l'ARN doivent être très efficaces, elles doivent néanmoins être maintenues sous contrôle afin d'éviter la dégradation de tous les ARN présents dans la cellule.

Chez les bactéries, la demi-vie des ARN est courte et peut varier de quelques secondes à plus d'une heure, mais elle est généralement beaucoup plus courte que le temps de doublement de l'organisme (Laalami et al., 2014).

Cette instabilité est cruciale pour adapter le modèle d'expression des gènes. En effet, le renouvellement rapide de l'ARNm confère un avantage certain en permettant aux cellules d'adapter rapidement la synthèse des protéines à des changements environnementaux brusques, car les temps de réponse seraient beaucoup plus lents s'ils dépendaient uniquement de la modulation de la transcription. De plus, les coûts énergétiques de la traduction et l'avantage du recyclage des ribonucléotides exigent un mécanisme de destruction rapide des transcrits qui ne sont plus utiles. Toutefois, pour que les ARN aient le temps d'être traduits, l'événement qui déclenche la dégradation des ARN est supposé comporter un délai afin de s'assurer que l'ARN soit "vieux" (mesuré en secondes) avant de le dégrader. Ceci peut se faire, soit en séparant le lieu de synthèse de l'ARN des ribonucléases, soit en modifiant les ARN de manière épi-transcriptomique, par exemple en dé-enroulant les structures secondaires ou en déphosphorylant les extrémités 5' pour obtenir des ARN 5' monophosphorylés qui sont de meilleurs substrats pour les ribonucléases.

Les enzymes de dégradation des ARN possèdent une activité exoribonucléase qui permet de dégrader les ARN à partir des extrémités 5' ou 3' nucléotide par nucléotide, ou/et une activité endoribonucléase qui clive les ARN de manière interne et produit donc deux fragments d'ARN. Ces deux activités sont en général réalisées par des enzymes différentes, sauf dans le cas d'une famille particulière, les ribonucléases β-CASP dont certains membres possèdent un unique site catalytique pouvant porter les deux types d'activité hydrolytique, endo- et exoribonucléasique.

Les mécanismes impliqués dans ces processus restent cependant largement inconnus (Laalami et al., 2014), notamment car les enzymes impliquées dans la dégradation des ARN chez les bactéries sont variées et différentes. Cependant, il existe deux voies de dégradation majeures qui semblent conservées : la voie directe et la voie extrémité 5' dépendante (Hui et al., 2014).

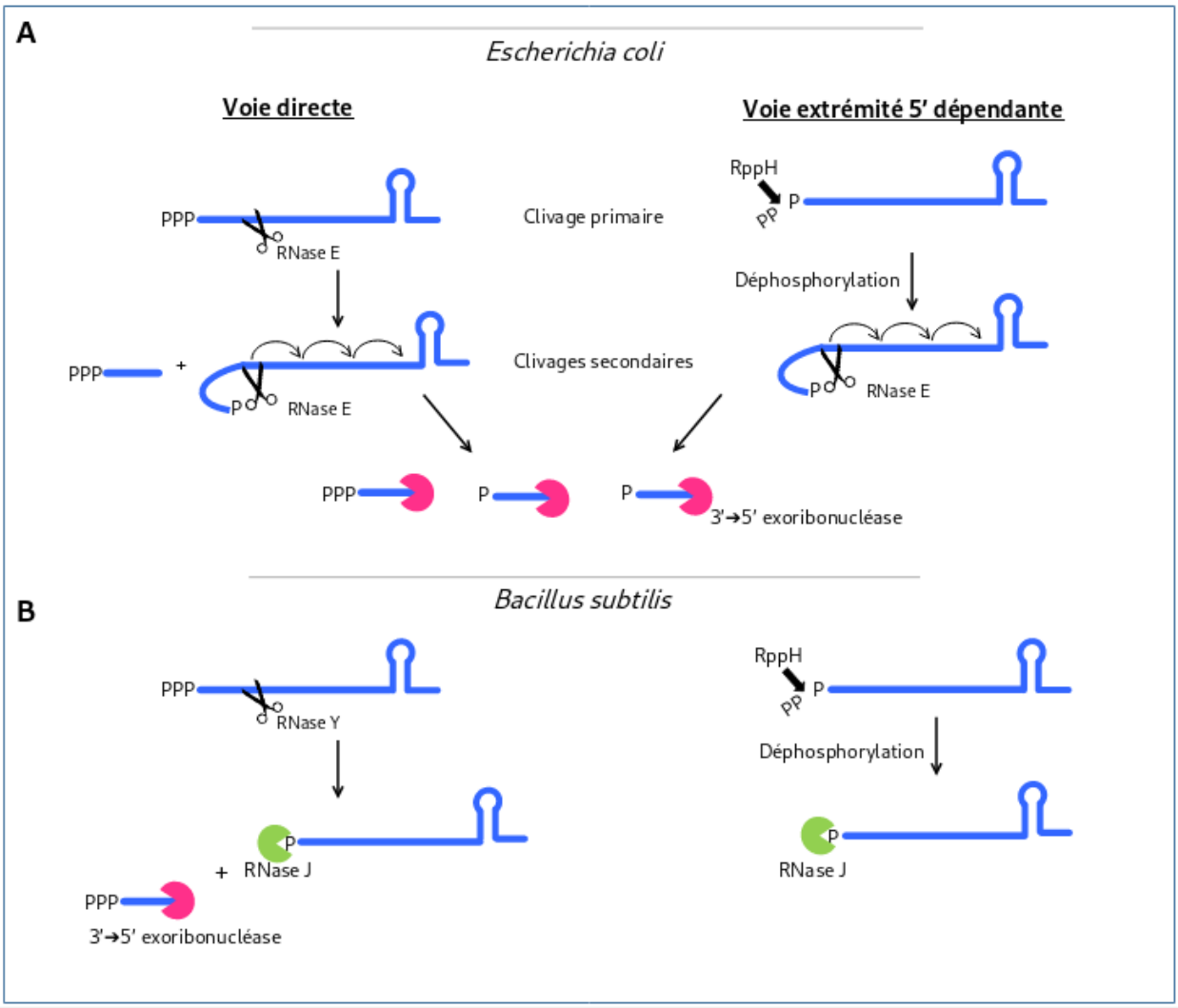
1.1.2.3.5.1.1 Voies de dégradation

- Voie directe (direct-entry pathway) :

Dans la voie directe (Fig. 26 gauche), la dégradation est initiée par une endoribonucléase qui clive l'ARNm de manière interne. Ce clivage génère un fragment 5’ en amont du clivage avec une extrémité 3’ instable dû à l’absence de structures secondaires en tige boucle généralement rencontrées aux extrémités des ARN bactériens intacts et qui protègent ces derniers de la dégradation. Ce fragment peut donc être dégradé par des exoribonucléases de l’extrémité 3’ vers l’extrémité 5’ (notée 3’→5’ par la suite ou 3’-exoRNase). Le fragment en aval de cette réaction initiale de clivage interne, le fragment 3’ de l'ARNm, possède maintenant une extrémité 5’P favorisant sa dégradation par différents types de ribonucléases. Pour les espèces bactériennes dont le génome code pour la RNase J, dont l’activité 5’→3’ exoribonucléasique a tout d’abord été mise en évidence chez *B. subtilis* alors que l’on pensait les bactéries dépourvues d’une telle activité, la présence d'un seul phosphate à cette extrémité expose ces intermédiaires à une dégradation 5'-exonucléolytique rapide. Chez les espèces dépourvues de la RNase J (comme *E. coli* et la majorité des γ-protéobactéries), ces intermédiaires de dégradation sont rapidement détruits par la RNase E, dont le pouvoir ribonucléolytique est considérablement accru lorsque l'extrémité 5' de l’ARN est monophosphorylée.

- Voie extrémité 5’ dépendante (5’-end dependent pathway) :

La dégradation par la voie extrémité 5’ dépendante (Fig. 26 droite) est déclenchée par un premier événement non nucléolytique : la conversion de l'extrémité 5' d'un triphosphate en un monophosphate. Réalisée par l'ARN pyrophosphohydrolase RppH, cette modification augmente considérablement la sensibilité de l'ARNm à la dégradation par la RNase E (augmentation de son activité ribonucléolytique) ou la RNase J par son activité 5’ exoribonucléasique.



***Figures 26 :* Voies de dégradation de l'ARN** *chez (A) E. coli et chez (B) B. subtilis. Adapté de* (Laalami et al., 2014)

1.1.2.3.5.1.2 Les endoribonucléases

La mise en évidence des endoribonucléases dans les organismes modèles a montré que cette activité était portée par la RNase E chez *E. coli*, la RNase Y et la RNase J chez *B. subtilis*. Ces protéines ne présentent entre-elles aucune similarité de séquences bien que l'on puisse noter que la structure 3D de la moitié catalytique N-terminale présente des similitudes avec celle de la RNase J (Laalami et al., 2014). Cette activité aurait donc été « inventé » au moins trois fois de façon indépendante au cours de l'évolution, reflétant un cas remarquable d'évolution convergente, montrant que les fonctions de ces enzymes sont d'une importance générale pour le métabolisme

de l'ARNm chez les bactéries. En effet, toutes les espèces bactériennes dont les génomes ont été séquencés possèdent au moins une enzyme apparentée aux RNases E/G, J ou Y, quelques soit leur phylum d'appartenance (Fig. 27). De plus, on peut trouver toutes les combinaisons possibles de ces enzymes dans un même organisme. Certaines espèces ne comptent qu'un seul membre, comme la plupart des β- et γ-protéobactéries, qui possèdent presque exclusivement une enzyme de type RNase E/G. D'autres au sein des Bacillota (ex Firmicutes) peuvent posséder les trois protéines (majorité des Costridia par exemple) ou ne posséder que les RNases Y et J comme *B. subtilis* et *S. aureus*.

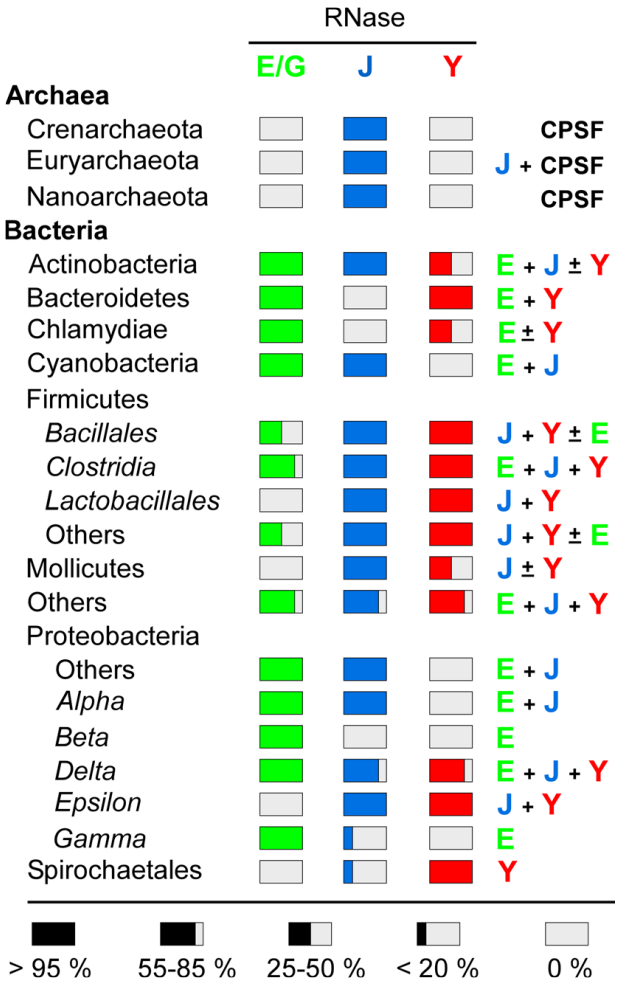


Figure 27 : Présence des RNases E, J et Y chez les procaryotes.

Pourcentage d'espèces contenant des RNases E/G (vert), J (bleu) et/ou Y (rouge) par rapport au nombre total d'organismes au sein d'un phylum. Les combinaisons de RNases présentes dans la majorité d'un phylum sont indiquées à droite. Toutes les archées contiennent une activité de type RNase J, mais elles peuvent être divisées en deux grandes subdivisions qui correspondent aux orthologues du facteur de spécificité de clivage et de polyadénylation eucaryote (CPSF73) et de la RNase J bactérienne (Dominski et al., 2013). *Adapté de* (Laalami et al., 2014).

La RNase G d'*E. coli* est une protéine homologue non essentielle de la RNase E qui présente un degré élevé de conservation avec le domaine catalytique de la RNase E (Z. Li et al., 1999)**,** (Wachi et al., 1999). Ces deux protéines sont à l'origine de la famille des endoribonucléases RNase E/G (Kaberdin et al., 1998)**,** (Carpousis, 2002)**,** (Condon & Putzer, 2002). Chez *E. coli*, la RNase E et la RNase G se distinguent par la grande extension C-terminale non catalytique de la RNase E. Une étude phylogénétique de ces deux protéines au sein des γ-protéobactéries a montré qu'elles forment deux groupes orthologues distincts qui ont été hérités verticalement et qui se seraient différenciés avant l'émergence de ce phylum (Aït-Bara et al., 2015, p. 20)**.**

1.1.2.3.5.1.2.1 Ribonucléase E : le modèle *E. coli*

Chez *E. coli*, le système de dégradation de l'ARN repose principalement sur l'activité de l'endoribonucléase E (RNase E)**.** Bien que la plupart des orthologues à la RNase E soient retrouvées dans les protéobactéries, beaucoup d'enzymes appartenant à la même famille sont retrouvées en dehors de ce phylum (Carpousis, 2007). La RNase E est une enzyme clé car elle est impliquée dans la dégradation de la plupart des ARNm (Bernstein et al., 2004)**,** (Clarke et al., 2014). Par exemple, 78% des ARNm de *Salmonella typhimurium* sont clivés par la RNase E (Chao et al., 2017). Elle est également impliquée dans la maturation de certains ARN ribosomiques et ARN de transferts (Z. Li & Deutscher, 2002)**.** De nombreuses études ont montré que si la RNase E ne dépend pas d'une séquence nucléotidique particulière pour le clivage, elle nécessite une région monocaténaire de préférence riche en résidus AU (Arraiano et al., 2010)**,** (Huang et al., 1998)**,** (McDowall et al., 1994)**.** De plus, comme nous l'avons dit précédemment, son pouvoir ribonucléolytique est fortement augmenté au voisinage d'une extrémité 5'P (par rapport à une extrémité 5'PPP) (Garrey et al., 2009)**,** (X. Jiang & Belasco, 2004, p. 200)**,** (Kime et al., 2008, p. 20)**,** d'où l'intérêt de la voie extrémité 5' dépendante. La RNase E est ancrée à la membrane cytoplasmique interne d'*E. coli* (Taghbalout & Rothfield, 2007)**,** (Khemici et al., 2008). Les bactéries dans lesquelles l'ancrage de la RNase E a été supprimé, sont viables, mais présentent un phénotype de croissance lente. La localisation membranaire est donc bénéfique, bien que les processus qui en dépendent n'aient pas encore été identifiés. Des travaux plus récents montrent également que la RNase E diffuse librement sur toute la surface de la membrane cytoplasmique interne (Strahl et al., 2015)**.**

La RNase E est un tétramère et les quatre sous-unités sont disposées en dimères de dimères. Chaque sous-unité de RNase E d'*E. coli* se compose d'un domaine amino-terminal bien conservé qui abrite le site catalytique et d'un domaine carboxy-terminal peu conservé qui comprend une hélice de liaison à la membrane, deux domaines de liaison à l'ARN riches en arginine et une région qui sert d'échafaudage pour l'assemblage d'un complexe ribonucléolytique appelé le dégradosome de l'ARN (Fig. 28)

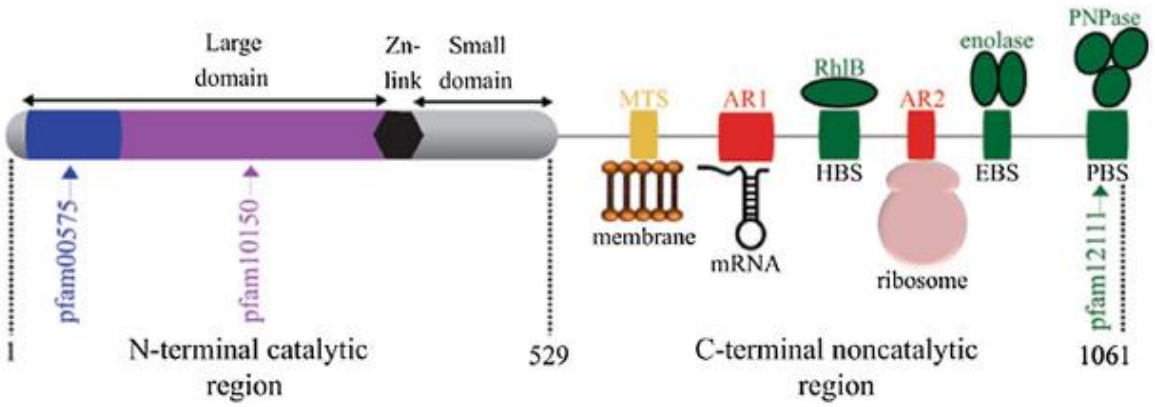


Figure 28 : Organisation en domaines de la RNase E d'*E. coli*.

Structure primaire de la RNase E d'E. coli montrant la région catalytique (résidus 1-529) et la région non catalytique (résidus 530-1 061). La région catalytique est composée d'un grand domaine, d'un lien en zinc (lien Zn) et d'un petit domaine. Le grand domaine contient un motif de liaison à l'ARN S1 (bleu, Pfam00575) et un site catalytique de liaison aux métaux (violet, Pfam10150). L'holoenzyme tétramérique de la RNase E est un dimère de dimères. Le grand domaine et le lien Zn ont un rôle structurel dans la formation du dimère ; le petit domaine a un rôle structurel dans les interactions dimères et dimères-dimères. La région non catalytique (résidus 530-1 061) contient des microdomaines responsables de l'interaction entre la RNase E et la membrane plasmique interne (jaune MTS, séquence de ciblage de la membrane, résidus 565-582), l'ARN (rouge AR1, riche en arginine 1, résidus 604-644 ; AR2, arginine-rich 2, résidus 796-814) et les protéines pour former le dégradosome canonique de l'ARN (vert HBS, site de liaison de l'hélicase, résidus 719-731 ; EBS, site de liaison de l'énolase, résidus 834-850 ; PBS (Pfam12111), site de liaison de la PNPase, résidus 1 021-1 061). Extrait de (Aït-Bara et al., 2015)

Le dégradosome est composé de la RNase E, de RhlB (hélicase à ARN), de l'énolase et de la PNPase (3' exoribonuclease) (Carpousis, 2007)**,** (Carpousis et al., 2009)**,** (Bandyra et al., 2012). RhlB déroule les structures d'ARN qui peuvent bloquer l'action de la PNPase, comme le montre l'effet des REP (Repeated Extragenic Palindromes) sur la PNPase (Newbury et al., 1987).

Ce rôle a été confirmé par la démonstration que le dégradosome a besoin de l'hydrolyse de l'ATP pour dégrader l'ARN contenant des REP (Py et al., 1996). Le rôle de l'énolase, une enzyme impliquée dans la glycolyse, reste énigmatique.

Étant donné, son implication dans le métabolisme glycolytique, elle pourrait détecter l'état énergétique de la cellule ou le stress lié au phosphosucre (Bonnin & Bouloc, 2015). Dans d'autres bactéries, les homologues de la RNase E forment des complexes de type dégradosome, bien que leur composition ne soit pas conservée. Cependant, ces complexes ont en commun l'interaction de la RNase E avec des exoribonucléases (PNPase ou RNase R), des ARN hélicases (ARN hélicases à boîte DEAD ou Rho) et des enzymes du métabolisme central du carbone (énolase ou aconitase) (Jäger et al., 2001)**,** (K. Lee & Cohen, 2003)**,** (Purusharth et al., 2005)**,** (Hardwick et al., 2011).

1.1.2.3.5.1.2.2 Ribonucléase Y : le modèle de *B. subtilis*

La RNase Y est une endoribonucléase appartenant à la famille des phosphohydrolases métallo-dépendantes qui a été initialement découverte chez *B. subtilis*. Chez les Bacillota, il n'y a généralement pas d'homologue de la RNase E et la RNase Y jouerait donc un rôle important dans la dégradation des ARN, voire même majeur chez *B. subtilis* où elle impacte la stabilité de 25 à 30% des ARN (Laalami et al., 2013) **,** (Durand et al., 2012), et où sa délétion engendre de sévères effets pléiotropes (Figaro et al., 2013)**.**

Chez *B. subtilis*, la RNase Y (520 aa) est composée d'un domaine transmembranaire N-terminal (des acides aminés 1-25), suivi d'une grande région désordonnée (aa ~30-210), d'un domaine KH de liaison à l'ARN (aa ~ 211–270), et d'un domaine HD de chélation des métaux (aa 336-46) qui contient le motif conservé His/Asp nécessaire à l'activité de la RNase (Fig. 29).

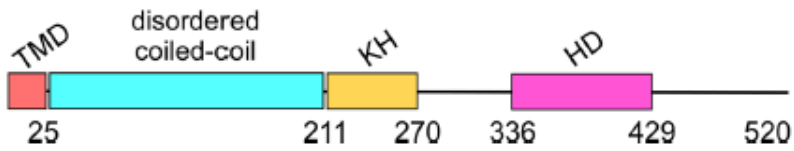


***Figure 29 :* Domaines composant la RNase J1 de B. subtilis (555 aa).** *Le domaine β-CASP est inséré dans le domaine β-lactamase auquel le domaine C-terminal est attaché par un linker. Extrait de* (Laalami et al., 2014)

Tout comme la RNase E, la RNase Y de *B. subtilis* est ancrée à la membrane cytoplasmique interne (Hunt et al., 2006)**,** (Shahbabian et al., 2009)*.* Cette localisation joue un rôle clé dans la dégradation des ARN et parfois même un rôle plus important que son activité enzymatique elle-même.

En effet, chez *B. subtilis*, il a été montré que la délétion de son ancre membranaire engendre des effets négatifs sur la croissance, qui sont plus importants que ceux engendrés par la mutation de son site actif (Khemici et al., 2015)**,** renforçant l'hypothèse qu'elle serait la base d'un complexe de dégradation multi-enzymes incluant d'autre ribonucléases (Lehnik-Habrink et al., 2010).

En revanche, même si ces deux ribonucléases partagent des similarités fonctionnelles, elles semblent différer quant aux caractéristiques de leurs substrats. Pour le moment, les caractéristiques des substrats préférentiels de RNase Y ne sont pas complètement bien compris, et aucune étude *in vivo* n'a pu montrer que la RNase Y était sensible à l'état de phosphorylation des extrémités 5' des ARN.

Chez *S. aureus*, la RNase Y semble jouer un rôle plus secondaire (Broglia et al., 2020)**,** étant donné qu'elle impacte la stabilité de seulement 6 à 15% des ARN et que sa délétion engendre un phénotype moins sévère que chez *B. subtilis* (Khemici et al., 2015)**.** Parmi les cibles de la RNase Y de *S. aureus*, des sites ont été trouvés dans la séquence codante de la RNase J2 et dans la 5′ UTR du cistron codant pour RpoY et la RNase J1. Ces clivages suggèrent un réseau de régulation dans lequel le niveau cellulaire de la RNase Y serait coordonné avec le niveau des RNases J1 et J2. Un site de clivage de la RNase Y a également été cartographié dans la 5′ UTR de son propre transcrit, ce qui suggère une autorégulation de son niveau par clivage de son propre transcrit, qui serait suivi d'une dégradation via une activité exonucléasique 5**′** (Khemici et al., 2015)**.**

L'absence de RNase E et la faible activité de RNase Y laissent penser qu'une autre ribonucléase serait l'actrice principale de la dégradation des ARN chez *S. aureus*.

1.1.2.3.5.1.2.3 Ribonucléase J : le modèle de *S. aureus* ?

Notre travail de thèse consistant en l'étude de l'activité de la ribonucléase J *chez S. aureus*, nous détaillerons davantage cette protéine.

1.1.2.3.5.1.2.3.1 Double activité de dégradation

La RNase J a été initialement découverte chez *B. subtilis* (Figaro et al., 2013) et, bien qu'elle n'ait pas d'homologue chez *E. coli,* elle a ensuite été identifiée chez plus de la moitié des espèces bactériennes dont le génome a été séquencé ainsi que dans la majorité des espèces d'euryarchées, à l'exception des espèces appartenant aux groupes des archeoglobales et des thermoplasmatales (Phung et al., 2020). Elle fait partie de la famille des β-CASP et plus précisément de la sous-famille des metallo-β-lactamase zinc dépendant (Dorléans et al., 2011)**.** Certains membres de cette famille, dont la RNase J, possèdent une activité duale, endo- et exoribonucléasique.

L’activité 5'→3’ exoribonucléasique de la RNase J a tout d’abord été mise en évidence chez *B. subtilis* (Even, 2005) alors que l’on pensait les bactéries dépourvues d’une telle activité. Cette activité est impliquée, chez les Bacillota, à la fois dans la maturation de certains ARN, par exemple l'ARNr 16S et dans la dégradation des ARNm (Mathy et al., 2007)**,** (Linder et al., 2014)**,** (Bugrysheva & Scott, 2010).

Lors d’études sur l’activité endoribonucléasique, celle-ci a généralement été démontré *in vitro* (Even, 2005)**,** (Hausmann et al., 2017)**,** (Pei et al., 2015). L'activité endoribonucléasique n'a été démontrée *in vivo* que sur quelque ARN dont l’ARNm de thrZ qui code pour une threonyl-tRNA synthétase (Laalami et al., 2014) et l’ARN leader trp (Deikus & Bechhofer, 2011), mais elle permettrait à la RNase J de contourner les protections en 5’, les structures secondaires en 3’ et les éventuelles protéines de liaison à l'ARN, en générant de nouvelles extrémités 5' qui pourraient ensuite être dégradées de manière 5'→3’ exoribonucléolytique Cette double activité, endoribonucléase et 5’→3’ exoribonucléase, lui permettrait donc de participer activement aux deux voies de dégradation (directe et extrémité 5’ dépendante).

De précédents travaux ont indiqué que l'activité exoribonucléasique de la RNase J de *B. subtilis*, *M. smegmatis*, *T. thermophilus* et *S. coelicolor* était presque totalement inhibée par un ARN 5’PPP (Mathy et al., 2007)**,** (Li de la Sierra-Gallay et al., 2008)**,** (Pei et al., 2015)**,** (Taverniti et al., 2011). Cette préférence pour les substrats monophosphorylés peut varier d’un facteur 3 à un facteur 10 selon les orthologues RNase J. Chez *S. aureus,* des expériences *in vitro* ont montré que la RNase J semblait moins sensible (préférence d’un facteur 3) que ces homologues aux états de phosphorylations de ses substrats (Hausmann et al., 2017). Par ailleurs, dans ces études, son activité endoribonucléolytique a été testée sur des substrats d’ARN 5’PPP et il semblerait que l’état de phosphorylation n’empêche pas cette activité. En revanche, il a donc été proposé que le choix d’une activité par rapport pourrait tout de même être modulé par l'état de phosphorylation 5′ (Li de la Sierra-Gallay et al., 2008)**,** (Pei et al., 2015, p. 201)**,** (Taverniti et al., 2011). Ainsi, les ARN 5’PPP serait préférentiellement dégradés par l’activité endoribonucléasique de la RNase J tandis que les ARN 5’P serait préférentiellement dégradés par son activité 5’→3’ exoribonucléasique. Cependant, il est difficile de statuer du fait que les activités endo- et exoribonucléasique partagent le même site catalytique (Hausmann et al., 2017). Si l'activité 5’→3’ exoribonucléasique de la RNase J est régulée par l'état de phosphorylation des extrémités 5' des ARN, les ribopyrophosphohydrolases (RPP) bactériennes devraient jouer un rôle majeur dans la régulation de la dégradation des ARN. En effet, les RPP modifient les extrémités 5' des ARN 5'PPP pour générer des extrémités 5'P (Srouji et al., 2017)**,** (Frindert et al., 2019)**,** (Deana et al., 2008).

Chez *S. aureus* aucune de ces RPP n'a encore été étudiée expérimentalement et donc leur rôle dans la régulation de la dégradation des ARN *in vivo* chez cette bactérie est encore mal compris.

La double activité de dégradation de la RNase J et le fait que la RNase Y aurait un rôle moindre chez *S. aureus* laissent penser que la RNase J serait l'élément central de la dégradation chez ce pathogène, comme c'est le cas dans d'autres organisme où la RNase J a déjà été étudiée telles que Mycobactérium tuberculosis (Bao et al., 2023) qui possèdent la RNase J en plus de la RNase E, *Deinococcus radiodurans* ***(****Zhao et al., 2015)*, *Streptomyces pyogenes* (Bugrysheva & Scott, 2010)***,*** *B. subtilis* (Even, 2005, p. 200) *et Helicobacter pylori* (Redko et al., 2016) qui possèdent la RNase J et la RNase Y. Par exemple, la délétion de la RNase J chez la bactérie *H. pylori* augmente, d'un facteur 2, la stabilité de 80% de ses ARNm (Redko et al., 2016) et, même si elle semble moins importante chez *B. subtilis*, elle impacterait tout de même la stabilité de 20 à 30% de ses ARNm (Durand et al., 2012). Il a également été montré que la plupart des *Archées* (Clouet-d'Orval et al., 2010) ainsi que certains Chloroplastes (Kleffmann et al., 2004) possèdent également une RNase J dont le rôle semble important.

Contrairement à la RNase E chez *E. coli* ou la RNase Y chez *B. subtilis*, la RNase J est cytosolique. Cependant, il a été montré que la RNase J pouvait occasionnellement interagir avec la RNase Y (Roux et al., 2011), et donc être retrouvée au niveau de la membrane cytoplasmique interne.

1.1.2.3.5.1.2.3.2 J-complexe

La plupart des organismes porte une seule copie du gène codant pour la RNase J, mais il est communément accepté que la majorité des génomes de Bacillota (ex Firmicutes) codent pour deux paralogues de la RNase J, Nous avons entrepris une première étude de 3771 génomes bactériens entièrement séquencés et bien annotés, dont 842 génomes de Bacillota correspondant à 426 espèces distinctes. Parmi les génomes de ces 426 espèces, 60 % codent pour deux protéines RNase J, 32 % pour une seule, et aucun gène RNase J n'a été détecté dans 2,4 % d'entre eux. En outre, parmi les 1409 génomes en dehors des Bacillota dans lesquels nous avons identifié la RNase J, la majorité ne possède bien qu'une copie de la protéine, à l'exception des Tenericutes dont la majorité des génomes renferment deux paralogues (Fig. 30).

La RNase J1 (auparavant YkqC) et la RNase J2 (auparavant YmfA) ont été découvertes chez *B. subtilis* et sont respectivement codées par les gènes *rnjA* et *rnjB*. Chez *S. aureus*, ces deux protéines sont codées par les gènes *rnj1* et *rnj2*.

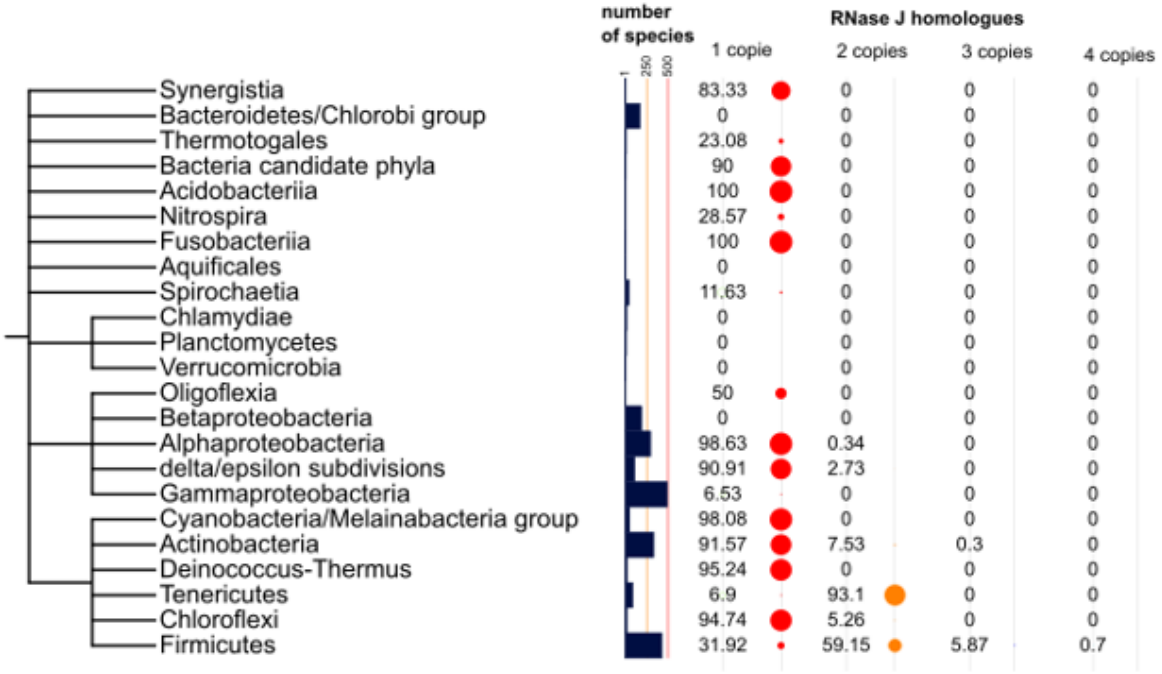


Figure 30 : Distribution phylogénétique des homologues de la RNase J.

Les RNases J ont été identifiées dans des génomes bactériens complets (une souche par espèce). Les résultats sont résumés en utilisant la taxonomie NCBI. Seuls les rangs taxonomiques contenant au moins cinq génomes sont indiqués. Le nombre d'espèces par rang taxonomique est indiqué par un histogramme. La fréquence du nombre de copies de RNase J par génome est indiquée par des valeurs et des cercles proportionnels à ces valeurs (communication personnelle).

Des interactions protéine-protéine entre les protéines paralogues J1 et J2 ont été démontrées par des expériences de double-hybride, de même qu'elles sont co-purifiées chez *S. aureus* et *B. subtilis* (Giraud et al., 2015), (Even, 2005), (Roux et al., 2011), (Lehnik-Habrink et al., 2010). De plus, la quantification des deux monomères montre qu'ils se trouvent environ en même quantité chez *B. subtilis* (Commichau et al., 2009). Ces éléments suggèrent que les RNases J1 et J2 forment un "complexe RNase J" *in vivo* qui constituerait l'unité fonctionnelle de l'activité RNase J (Linder et al., 2014), (X. Chen et al., 2015), (Mathy et al., 2010).

Les RNases J1 et J2 sont essentielles à la survie de certaines bactéries Gram-positives, comme par exemple chez *Streptococcus pyogenes* (Bugrysheva & Scott, 2010) et *Helicobacter Pylori* (Redko et al., 2013)**.** Chez *Streptococcus mutans* elles ne sont pas essentielles mais leurs délétions (même individuelles) provoquent de sévères effets pléiotropes, notamment des problèmes de croissance, de morphologie, de tolérance à l'acide, de compétence naturelle et de formation de biofilm. Ces effets sont exacerbés chez le simple mutant RNase J2 (X. Chen et al., 2015)**.**

Chez *S. aureus et B. subtilis*, il a d'abord été proposé que la RNase J1 était essentielle (Chaudhuri et al., 2009)**.** Cependant, (Linder et al., 2014) ont montré que les mutants de délétion de la RNase J1 sont viables mais provoquent de sévères effets pléiotropes. Chez *S. aureus*, par exemple, les mutants RNase J1 se divisent lentement, avec des septa de division mal placés, et ne se développent pas à 30°C ou 42°C (Linder et al., 2014)**.** La délétion du gène de la RNase J2 provoque également de sévères effets, en revanche, la mutation de son site actif n'engendre pas ou peu d'effet, suggérant un rôle enzymatique moindre.

D'ailleurs, il a été montré que l'activité 5'→3' exoribonucléase de la RNase J2 est presque inexistante chez *S. aureus* dû à sa faible affinité avec les substrats qui pourrait être expliquée par un motif β-lactamase différent de celui de la RNase J1 (HGHEDH → HGHDEN) (Mathy et al., 2010). Toutefois, l'activité endoribonucléolytique de la RNase J2 n'en serait pas impactée, ce qui fait d'elle une possible endoribonucléase fonctionnelle (Newman et al., 2011)**.** Par ailleurs (Gao et al., 2017) ont montré que la RNase J2 de la bactérie *Enterococcus faecalis* jouait un rôle très important dans la virulence.

Bien que les RNases J1 et J2 possèdent la même activité endoribonucléasique et les mêmes préférences pour les substrats, le complexe J1/J2 montre une préférence pour cliver à des sites différents de ceux des monomères, sur l'ARNm thrS (Mathy et al., 2010). D'autres cas similaires suggèrent que l'association de J1 et J2 modifie les propriétés endoribonucléolytiques de l'enzyme. En effet, il a été montré qu'une RNase J monomérique de *D. radiodurans* avait une activité exoribonucléasique plus faible mais une activité endoribonucléasique plus élevée qu'une RNase J multimérique (Y. Zhao et al., 2015). De plus, d'après la structure, la dissociation de la RNase J à l'état monomérique semble essentielle pour son activité endoribonucléasique, car l'accès au substrat ARN est bloqué par l'autre sous-unité du dimère (Dorléans et al., 2011). La structure quaternaire, en particulier chez les homologues bactériens de la RNase J, est fortement influencée par le segment polypeptidique C-terminal dont l'élimination conduit à une enzyme monomérique dont l'activité exoribonucléasique est altérée par rapport à l'enzyme complète (Raj et al., 2020).

Ainsi, la multimérisation de la RNase J permet de moduler ses activités endo et exoribonucléasiques, assurant une dégradation efficace des ARN.

1.1.2.3.5.1.2.3.3 Structure de la RNase J et du J-complexe

La structure tridimensionnelle de la RNase J est organisée en trois domaines (Fig. 31) (Raj et al., 2020)**,** (Li de la Sierra-Gallay et al., 2008)**,** (Newman et al., 2011). Un domaine β-lactamase, un domaine β-CASP et un domaine C-terminal.

Les résidus du site actif sont principalement retrouvés sur le domaine β-lactamase tandis que le domaine β-CASP aurait un rôle dans la reconnaissance du substrat. Quant au domaine C-terminal, il n'est pas systématiquement présent et sa fonction reste encore inconnue. Toutefois, il semblerait jouer un rôle dans la « dimérisation » puisque des expériences ont montré que sa délétion conduit à l'absence de multimères de RNases J, réduisant fortement l'activité exoribonucléasique (Li de la Sierra-Gallay et al., 2008)**.** La résolution de la structure du complexe RNase J liée à un uracile monophosphate (UMP), a permis de mieux comprendre le fonctionnement de l'activité exoribonucléase et notamment le fait que les ARN 5'P soient de meilleurs substrats, par rapport aux ARN 5'PPP (Li de la Sierra-Gallay et al., 2008).

En effet, il a été montré que le site actif de la RNase J se trouve localisé profondément dans la fente située entre les domaines β-CASP et β-lactamase, où deux ions métalliques (Mn2+ ou Zn2+) garantissent qu'une molécule d'eau peut effectuer l'attaque nucléophile sur le phosphate situé entre le premier et le deuxième nucléotide (Raj et al., 2020)**,** (Li de la Sierra-Gallay et al., 2008)**,** (Newman et al., 2011). Cependant, seul un ARN simple brin peut pénétrer dans cette fente étroite, de sorte que les structures secondaires ou la formation d'un duplex avec une autre molécule d'ARN peut protéger un substrat de la dégradation (Raj et al., 2020). En outre, il existe une poche de liaison pour le monophosphate 5' du nucléotide immédiatement en amont du site actif, mais cette poche semble être trop petite pour accueillir un triphosphate (Li de la Sierra-Gallay et al., 2008)**,** (Dorléans et al., 2011), expliquant la préférence de la RNase J pour les ARN 5'P.

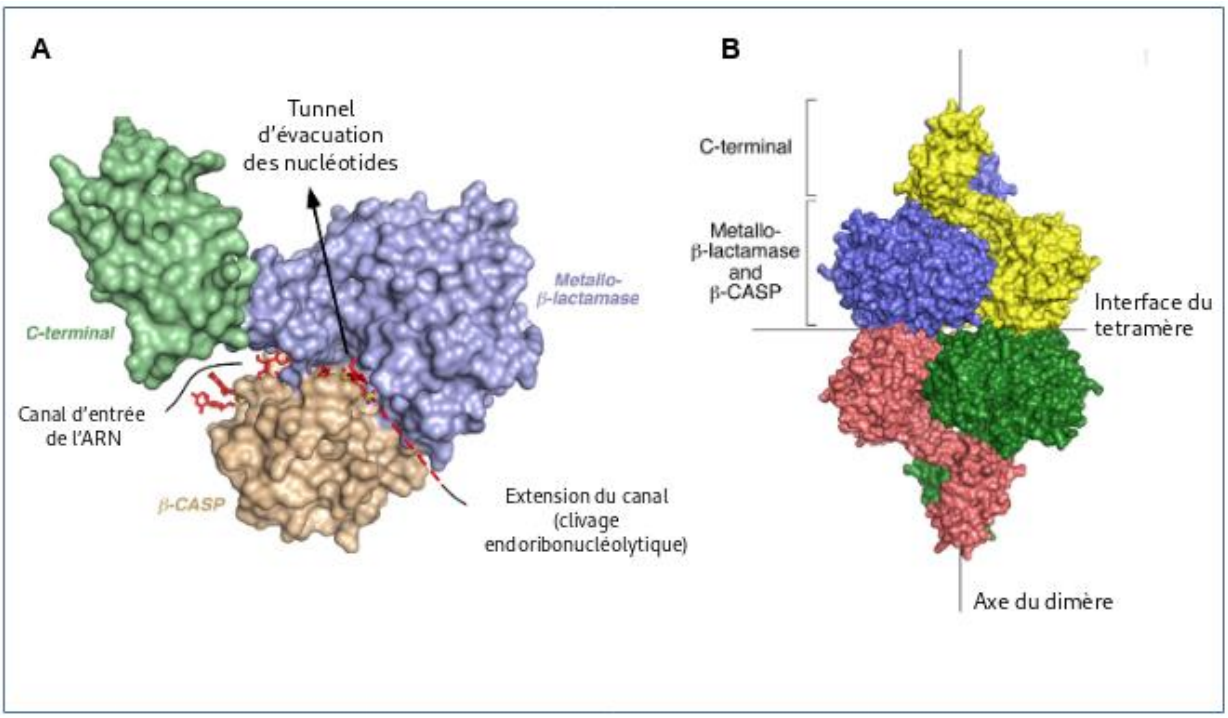


Figure 31 : Structure de la RNase J.

(A) *Monomère de RNase J portant un ligand oligoribonucléotidique prêt à subir une dégradation 5'-exonucléolytique. La RNase J existe normalement en tant que dimère ou tétramère et le canal d'entrée de l'ARN, son extension inoccupée et le tunnel d'évacuation des nucléotides putatifs ont été exposés dans cette image. Dans un dimère, les canaux et le tunnel seraient pris en sandwich entre les domaines métallo-β-lactamase (bleu) et β-CASP (beige) des deux sous-unités et cachés à la vue. Le domaine C-terminal (vert) ne semble pas entrer en contact avec l'oligonucléotide, qui est 2'-O-méthylé et représenté sous forme de bâtonnet avec tous ses atomes colorés en rouge, à l'exception du phosphore qui est représenté en jaune. Pour être dégradé par exonucléolyse, l'ARN doit se faufiler dans le canal d'entrée jusqu'à ce que son extrémité 5' atteigne le site de liaison de l'extrémité 5'. Pour être clivé par voie endonucléolytique, Dorléans et et Newman proposent que l'ARN occupe à la fois le canal d'entrée et son extension (ligne rouge en pointillés).* (Richards & Belasco, 2011)***. (B)*** *Tétramère de la RNase J. Les quatre sous-unités du tétramère (chacune de couleur distincte) s'assemblent comme un dimère de dimères. Les trois domaines de chaque sous-unité entrent en contact intermoléculaire à l'interface étendue du dimère, tandis que seuls les domaines métallo-β-lactamase et β-CASP se rencontrent à l'interface plus petite du tétramère. La figure représente l'enzyme de T. thermophilus ; la RNase J1 de B. subtilis forme un tétramère similaire.* (Richards & Belasco, 2011)

1.1.2.3.5.1.3 Le dégradosome chez les bactéries

Comme nous l'avons vu précédemment avec le dégradosome de *E. coli* centré sur la RNase E, l'assemblage multi-enzymes du dégradosome de l'ARN apporte des contributions essentielles au métabolisme de l'ARN : co-localisation de plusieurs ribonucléases et d'enzyme de decapping des ARN.

Le premier dégradosome des bactéries à Gram-positif qui a été proposé est celui de *B. subtilis,* suite à l'identification d'interaction par des approches double-hybride entre les composants. Il se compose de la RNase Y, de la RNase J1 et de son paralogue RNase J2, de la PNPase, de la DEAD-box hélicase CshA, et des deux enzymes glycolytiques l'énolase et la phosphofructokinase. (Commichau et al., 2009)**,** (Lehnik-Habrink et al., 2010), (Roux et al., 2011). Les interactions détectées entre les différents partenaires sont synthétisées sur la Figure 32A.

Le dégradosome de *S. aureus* a été identifié un peu plus tard. Les gènes orthologues aux gènes codant les composants du dégradosome de *B. subtilis* ayant été identifié dans le génome de *S. aureus*, cela laissait suggérer que les deux complexes de dégradation devaient être fortement conservés entre les deux espèces. Cette hypothèse a été confirmée également grâce à des approches de double-hybride (Roux et al., 2011). Cependant, des différences dans les interactions entre les partenaires ont été observées (Fig. 32B). La RNase J1 et la RNase J2 interagiraient plus fortement dans *S. aureus* et une interaction, absente dans *B. subtilis*, entre la RNase J1 et l'hélicase CshA a été identifiée. Par contre, aucune interaction significative n'a été détectée dans *S. aureus* entre la RNase J1 et la RNase Y ou la phosphofructokinase. L'hélicase CshA interagit comme dans *B. subtilis* avec la phosphofructokinase, l'énolase et la RNase Y mais pas avec la PNPase.

La RnpA, la sous-unité protéique de la RNase P, ayant été montré comme jouant un rôle dans la dégradation des ARN chez *S. aureus* (Olson et al., 2011), celle-ci a été ajoutée dans les approches double-hybride. Les résultats ont révélé que CshA interagit avec RnpA chez *S. aureus*, ce qui suggère que l'hélicase peut augmenter l'activité de la RNase RnpA cellulaire. Pour déterminer si cette interaction est conservée parmi les bactéries Gram-positives, les études de double-hybride ont été ensuite réalisée entre les deux protéines de *B. subtilis* et ont effectivement révélé la conservation de l'interaction entre CshA et RnpA (Roux et al., 2011).

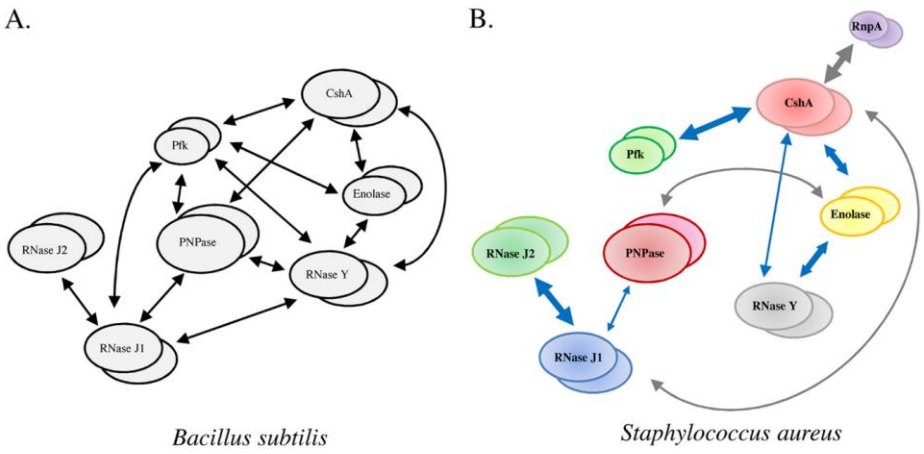


Figure 32 : *Complexe de type dégradosome à ARNm chez (A) B. subtilis et (B) S. aureus*.
Les flèches bleues représentent les interactions protéiques conservées en S. aureus et B. subtilis
La largeur des flèches indique la force de l'interaction déterminer par des résultats quantitatifs de l'activité β-galactosidase. Extrait de (Roux et al., 2011)

Un second complexe, le Y-complexe, a été mis en évidence chez *B. subtilis* et *S. aureus* décrivant les interactions des protéines YlbF, YmcA et YaaT avec la RNase Y. Il a été proposé que ces partenaires agiraient comme des facteurs de spécificité pour le clivage par la RNase Y de certains opérons et riboswitches (DeLoughery et al., 2018).

D'autres dégradosomes ont été décrits. Celui de *Helicobacter pylori*, bactérie pathogène Gram-négative, représente un dégradosome minimal à deux partenaires, la RNase J et la seule hélicase ARN à Dead-box de *H. pylori*, RhpA. Cette simplicité pourrait avoir comme origine la taille réduite du génome de *H. pylori* (1,6 Mb), sa niche de colonisation restreinte à l'estomac humain et à la capacité de la RNase J à agir à la fois comme endo et exoribonucléase (Redko et al., 2013). Finalement celui de *Mycobacterium tuberculosis* comprend quatre composants principaux, la RNase J, la RNase E, la PNPase et l'hélicase à ARN RhlE. D'autres protéines ont été trouvées associées à ce dégradosome (Fig. 33A). La RNase J a montré de fortes interactions bidirectionnelles avec la PNPase et RhlE, mais pas avec la RNase E. Les réseaux d'interaction protéine-protéine basés sur la copurification avec des appâts marqués à la GFP sont cohérents avec la présence soit d'un dégradosome mycobactérien unique comprenant les quatre composants mais excluant peut-être l'interaction directe entre la RNase E et la RNase J (Fig. 33B, diagramme inférieurs), soit la constitution de deux complexes différents impliquant l'une ou l'autre des deux nucléases (Fig. 33B diagrammes supérieurs). Des études menées sur d'autres micro-organismes suggèrent que les dégradosomes sont des assemblages dynamiques (Górna et al., 2012) et qu'ils

peuvent prendre des formes multiples. L'absence d'interaction entre la RNase J et la RNase E et la présence en excès de la PNPase par rapport à la RNase E et à la RNase J seraient en faveur de la formation de deux complexes différents, Le modèle d'un seul complexe pourrait refléter des situations physiologiques où la PNPase devient limitante (Płociński et al., 2019).

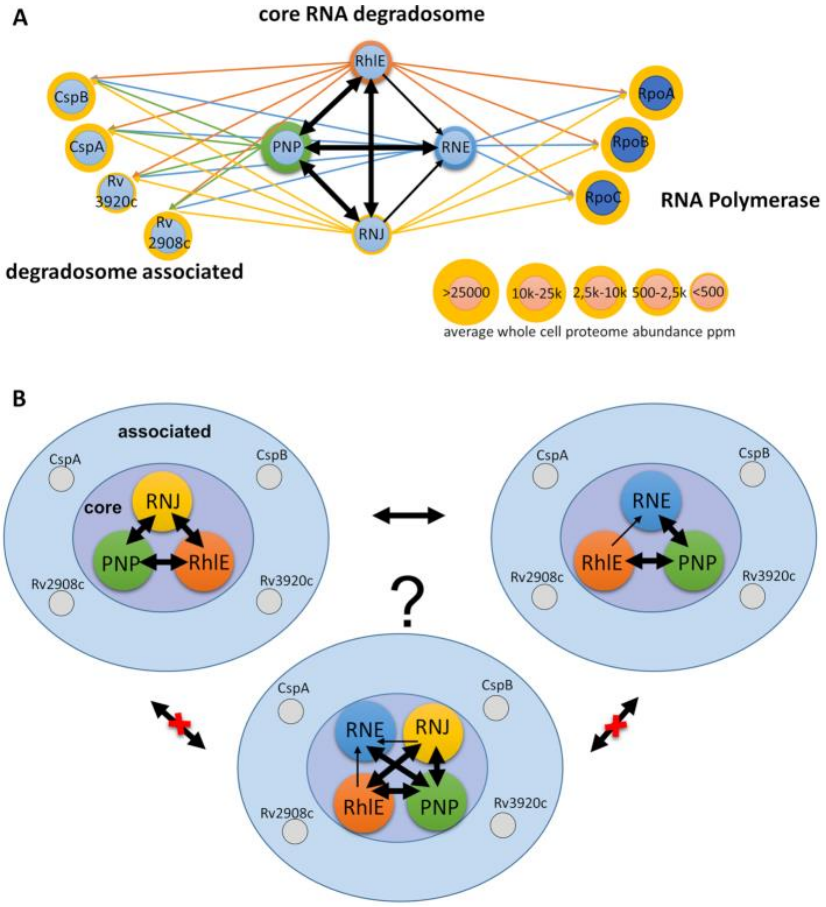


Figure 33 : ***Caractérisation des complexes de dégradation de l'ARN chez M. tuberculosis.***

*(**A**) Illustration de l'organisation possible du dégradosome de l'ARN chez les espèces mycobactériennes. Sur la base des résultats expérimentaux, les éléments putatifs centraux du dégradosome de l'ARN ont été prédits comme étant la PNPase, la RNase E, la RNase J et RhlE. Les chaperons d'ARN CspA, CspB ainsi que les protéines Rv3920c (Jag-like) et Rv2908c contenant un domaine KH ont été prédits comme des composants accessoires du dégradosome d'ARN. L'association de la machinerie de dégradation de l'ARN avec l'ARN polymérase est également suggérée. Les flèches épaisses indiquent l'interaction protéine-protéine bidirectionnelle observée dans l'expérience pull-down, tandis que les flèches fines représentent des interactions unidirectionnelles. (**B**) Modèles alternatifs de l'organisation des complexes de dégradation de l'ARN mycobactérien, où la RNase J peut former un complexe séparé, excluant l'interaction directe avec la RNase E, ou les quatre composants principaux du dégradosome forment un seul complexe. Des états intermédiaires et diverses combinaisons avec des protéines accessoires sont également possibles.* Extrait de (Płociński et al., 2019)

A la suite de ces quelques exemples, nous voyons donc que même si les acteurs sont différents, les dégradosomes bactériens sont composés de ribonucléases, d'hélicases à ARN et aussi dans certains cas d'enzymes impliquées dans le métabolisme énergétique.

Les hélicases à ARN permettent de déplier les structures secondaires de l'ARN, générant ainsi de l'ARN simple brin, qui peut être ciblé par des RNases spécifiques. De plus, il a été montré que CshA interagissait avec de nombreuses protéines liées à la dégradation des ARN et il a même été proposé qu'elle serait une des protéines centrales de la machinerie de dégradation des ARN chez *B. subtilis* (Lehnik-Habrink et al., 2010)**.** Effectivement, il a été montré que la délétion de CshA impactait la stabilité de plus de 200 ARNm de *B. subtilis* et environ 100 ARNm chez *S aureus* (Lehnik-Habrink et al., 2010)**,** (Giraud et al., 2015)**.**

Il est possible que les interactions entre ces composants soient transitoires étant donné qu'aucun complexe complet n'a jamais été purifié chez les bactéries à gram positive (Cascante-Estepa et al., 2016)**.** L'hypothèse serait même que ces grands complexes ne s'assemblent que de manière ponctuelle, lorsque plusieurs activités enzymatiques différentes sont nécessaires pour la dégradation d'un ARN en particulier (Giraud et al., 2015)**,** (Roux et al., 2011)**,** (Lehnik-Habrink et al., 2010)**,** (Raj et al., 2020).

1.1.2.3.5.2 Déphosphorylation des extrémités 5'

Comme nous l'avons vu dans les chapitres précédents, l'état de phosphorylation des extrémités 5' est impliqué dans la modulation de l'activité de certaines ribonucléases et donc dans la régulation de la dégradation des ARN (Laalami et al., 2014). L'extrémité 5' de l'ARN est tri-phosphorylé (5'PPP), puisque le premier nucléotide incorporé par l'ARNP est un nucléotide triphosphate (NTP). Dans de rares cas, il peut toutefois s'agir de NAD+, de NADH ou de CoA. Ces modifications ont en commun au moins une liaison pyrophosphate. Les triphosphates présents sur les extrémités 5' des ARN peuvent être modifiés par les ribopyrophosphohydrolases (RPP), qui clivent le pyrophosphate, c'est-à-dire les 2 phosphates β et γ pour laisser un 5'-monophosphate (5'P), laissant ainsi une marque épi-transcriptomique sur l'ARN. Chez les bactéries, dans l'état des connaissances actuelles, seules les hydrolases NUDIX (NUcleoside Diphosphate for X-moeity) ont été décrites comme ayant une activité ribopyrophoshydrolase.

La super-famille NUDIX, est présente dans les trois domaines du vivant et comprend plus de 80 000 protéines qui se répartissent dans quatre grandes classes fonctionnelles : les pyrophosphohydrolases, les adénine glycosylases spécifiques du mésappariement de l'adénine/guanine (adénine glycosylases spécifiques A/G), les isomérases isopentényl diphosphate

(IDI), et les protéines ayant des activités non enzymatiques telles que l'interaction protéique et la régulation transcriptionnelle (Srouji et al., 2017). Elles sont classées dans cinq familles Pfam (PF00293, PF14443, PF09296, PF14815, PF13869). La plupart des hydrolases NUDIX caractérisées possèdent un motif NUDIX conservé de 23 acides aminés (GX5EX7REUXEEXGU, où U est un résidu aliphatique volumineux et X n'importe quel acide aminé). Ce motif forme une structure boucle-hélice-boucle essentielle à la catalyse qui lie un ou plusieurs ions métalliques, qui à leur tour aident à orienter le groupement diphosphate présent dans tous les substrats des hydrolases NUDIX (Koonin, 1993)**,** (Bessman et al., 1996). Les substrats des hydrolases NUDIX sont variés et comprennent les nucléosides triphosphates, les sucres nucléotidiques, les dinucléosides polyphosphates, les dinucléotides coenzymes, les substrats non nucléotidiques, y compris les pyrophosphates d'inositol, et les extrémités 5' des ARN (McLennan, 2013)**,** (Srouji et al., 2017)**.**

En ce qui concerne les pyrophosphohydrolases d'ARN bactérien, elles ont été particulièrement étudiées chez *E. coli* et *B. subtilis*. Pour deux d'entre elles, RppH et NudC, Il a été montré qu'elles génèrent de l'ARN avec une extrémité 5' monophosphate (5'P). NudC a été caractérisée chez *E. coli* comme une enzyme de decapping de l'ARNm qui hydrolyse spécifiquement la liaison pyrophosphate de la coiffe NAD+ protégeant l'extrémité 5' d'un sous-ensemble d'ARNm (Cahová et al., 2015). La recherche de similarités a permis d'identifier des homologues de NudC chez de nombreuses bactéries à gram négatif, mais dans quasiment aucune espèce de Bacillota.

Il a été démontré que RppH, chez *E. coli* (Deana et al., 2008) et *B. subtilis* (Richards et al., 2011), éliminent les γ- et β-phosphates des extrémités 5'de l'ARN, mais ne présentent aucune similarité de séquences, hormis le motif NUDIX. Les deux protéines nécessitent la présence d'au moins deux (préférentiellement plusieurs) nucléotides simple brin à l'extrémité 5' de leurs substrats. Par rapport à EcRppH qui tolère une plus grande variation de la séquence 5', *Bs*RppH a des exigences plus strictes, avec un G en deuxième position, une préférence pour une purine en troisième position et une légère préférence pour A par rapport à G en position (Hsieh et al., 2013)**,** (Piton et al., 2013)**.**

Les propriétés du mutant de délétion *rppH* suggèrent que *Bs*RppH pourrait ne pas être la seule RPP chez *B. subtilis* (Richards et al., 2011). Cette observation a conduit Frindert et al. (2019) à étudier, *in vitro*, la capacité des cinq autres hydrolases NUDIX mal caractérisées de *B. subtilis*, YvcI, NudF, MutT, YjhB et YmaB, à éliminer les groupes phosphates de l'ARN 5'PPP (Frindert et al., 2019). A l'exception de *Bs*YjhB qui est complètement inactive, les quatre autres NUDIX libèrent les phosphates γ et β des extrémités 5' de l'ARN, *Bs*YvcI étant la plus active après *Bs*RppH. Il s'est avéré que YvcI et YmaB produisaient des ARN 5'P sans ARN diphosphate intermédiaire, tandis

que MutT et NudF produisaient un mélange d'ARN monophosphate et diphosphate. Cependant, des ARN 5'PP n'ayant été identifiés que chez *E. coli* (Luciano et al., 2018), il reste à déterminer leur existence chez *B. subtilis*. L'activité *in vivo* de l'ARN pyrophosphohydrolase de ces quatre protéines Nudix reste également à déterminer.

Une recherche par homologie de *Bs*RppH chez *S. aureus (38% d'identité de séquence)* a permis d'identifier *Sa*RppH (Bonnin & Bouloc, 2015). Toutefois, la comparaison entre les transcriptomes d'une souche sauvage et d'une souche ΔrppH montre très peu de différence avec seulement quatre transcrits stabilisés dans le mutant ΔrppH. Aucun d'entre eux n'avait de guanosine en deuxième position aucune caractéristique commune particulière n'a été identifiée. *Sa*RppH jouerait donc un rôle mineur dans les conditions testées, qui pourrait être dû à la présence d'une seconde enzyme de type NUDIX qui pourrait compenser l'absence de *Sa*RppH.

Ce rapide aperçu des pyrophosphohydrolases NUDIX met en évidence le fait que l'identification de l'activité d'une nouvelle pyrophosphohydrolase non décrite n'est pas simple car i) la même activité peut être réalisée par des hydrolases NUDIX sans similarité de séquence, ii) les pyrophosphohydrolases NUDIX présentent une ambiguïté de substrat (activité vers des substrats apparentés avec des différences significatives d'efficacité (McLennan, 2013) et iii) le même substrat peut être traité par plus d'une enzyme de cette classe.

1.1.2.3.5.3 Synthèse sur la dégradation des ARN chez *S. aureus*

La dégradation des ARN chez *S. aureus* implique diverses ribonucléases (Fig. 34) permettant de dégrader les ARN à partir des extrémités 3', 5' et/ou via des clivages internes.

Plusieurs études suggèrent que la RNase J joue un rôle central dans le processus de dégradation des ARN chez *S. aureus*. En effet, Linder et al. 2014 ont montré que le double mutant de délétion rnjA et rnjB présente une faible croissance à 42°C. En outre, le complexe RNase J est impliqué dans la maturation du précurseur de l'ARNr 16S, qui comme évoqué est une étape clé dans la formation des ribosomes. Toutefois, des nombreuses études (Hausmann et al., 2017; Laalami et al., 2014; Linder et al., 2014; Mathy et al., 2010) ont montré que la RNase J1 présente une activité enzymatique nettement supérieur à celle de la RNase J2. Par ailleurs, il a été proposé que cette dernière aurait un rôle dans la stabilisation du complexe RNase J. En effet, des expériences de délétion de la RNase J2 ont montré qu'une surexpression de la RNase J1 (même enzymatiquement inactive) sur un plasmide permet de compenser la perte de la RNase J2, et donc la perte de stabilité du complexe RNase J par la formation d'un homodimère J1/J1 (Linder et al. 2014).

Ribonucléase	Gène	Fonction	Essentialité
RNase III	rnc	Endoribonucléase double brin*	Non
Mini-III	mmC	Endoribonucléase double brin*	Non[b]
RNase Y	rny/cvfA	Endoribonucléase simple brin*	Non
RNase J1	rnjA	Forte activité 5'➔3' exoribonucléase* *Endoribonucléase simple brin ?*	Non**
RNase J2	rnjB	Faible activité 5'➔3' exoribonucléase* *Endoribonucléase simple brin ?*	Non**
RNase P	rnpA	Endoribonucléase + excision d'extranucléotides en 5' des précurseurs d'ARNt avec le ribozyme rnpB*	Oui
RNase Z	rnz	Endoribonucléase + excision d'extranucléotides en 3' des précurseurs d'ARNt °	Oui
RNase M5	rnmV	Endoribonucléase double brin + maturation de l'ARNr 5S °	Non
PNPase	pnpA	3'➔5' exoribonucléase*	Non
RNase R	rnr	3'➔5' exoribonucléase °	Oui
YhaM	yhaM	3'➔5' exoribonucléase °	Non
RNase HI	ypqD/rnhA	Ribonucléase de la famille des protéines HI dont la fonction est inconnue°	Non
RNase HII	rnhB	Endoribonucléase (dégradation des duplex ARN/ADN)°	Non
RNase HIII	nrnA	Oligoribonucléase, 3', 5'-nucléotidase bisphosphate°	Non
Rae1	SA0491	Endoribonucléase simple brin*	Non

***Figure 34 :* Ribonucléases chez S. aureus**.

** Fonction démontrée expérimentalement ; ° Annotation basée sur des études réalisées chez B. subtilis ou E. coli ; Oui = démontré expérimentalement* (Chaudhuri et al., 2009) ***;*** *Non[b] = basé sur des études réalisées chez B. subtilis ; Non** = Essentielle à 42°C mais pas à des températures plus basses. Extrait de* (Bonnin & Bouloc, 2015)

Cependant, la contribution du complexe RNase J dans la stabilité des ARN chez *S. aureus*, ainsi que celle de ses paralogues J1 et J2, reste encore à déterminer. Bien que les deux RNases J soit capable de cliver les ARN de manière endoribonucléolytique *in vitro*, aucune étude globale *in vivo* n'a encore été réalisée pour identifier ses sites de clivage exacts.

Contrairement à la RNase Y et à la RNase E, les RNases J ne sont pas ancrées à la membrane, ce qui pourrait leur permettre de dégrader les ARN pendant leur transcription, voire pendant leur traduction (Fig. 35) (Tejada-Arranz et al., 2020). Ceci suggère qu'un couplage entre la transcription, la traduction et la dégradation des ARN serait possible. Bien que la RNase J1 semble très active, ses substrats spécifiques restent à identifier. À ce jour, le seul facteur connu influençant légèrement son activité est l'état de phosphorylation des extrémités 5' des ARN (Hausmann et al., 2017).

Par ailleurs, bien que la RNase Y ne soit pas essentiel chez *S. aureus* (Marincola et al. 2012), des études ont montré que la délétion de cette dernière entraine une variation de l'expression d'entre 500 et 600 gènes (Bonnin et Bouloc 2015, Marincola 2012, Numata 2014). Et bien que seulement 99 sites de clivages endonucléolytique n'est été identifié chez S. aureus (Khemici et al. 2015), ces résultats indiquent que la RNase Y est un acteur important dans la régulation de l'expression génique chez *S. aureus*.

De plus, bien que l'endoribonucléase double brin RNase III participe au processing des ARNr, son rôle dans ce processus semble être secondaire chez *S. aureus* (Chevalier et al., 2008), comme chez *B. subtilis* (Herskovitz & Bechhofer, 2000). En revanche, les travaux de Chevalier et al. (2008) ont démontré que, chez S. aureus, la RNase III joue un rôle crucial dans la répression de l'expression de plusieurs ARNm codant pour des facteurs de virulence, soulignant ainsi son implication dans la régulation de la virulence.

Par la suite, il a été montré à de nombreuses reprises que les RNases III bactériennes sont impliquées dans la dégradation des duplex ARN/ARNas (Durand et al., 2012; Lasa et al., 2012; Le Rhun et al., 2016; Lybecker et al., 2014; Šetinová et al., 2018).

Dans l'étude menée par Lioliou et al., (2012) sur *S. aureus*, des expériences de co-immunoprécipitation impliquant la RNase III ont permis d'identifier des ARNas correspondant à 44 % des gènes annotés. De plus, les résultats de Mediati et al., (2022) ont d'ailleurs renforcé cette hypothèse en identifiant 133 interactions entre la RNase III et des duplex ARN/ARNm chez *S. aureus*. Tout ceci confirme que la RNase III est également une ribonucléase importance chez *S. aureus*.

S. aureus code pour trois orthologues d'exonucléases 3′→5′, la PNPase, la RNase R et YhaM (Bonnin & Bouloc, 2015). Le rôle global de la PNPase dans la dégradation de l'ARN de S. aureus a été déterminé à l'échelle du génome par Anderson & Dunman (2009). Leurs analyses ont révélé qu'après 5 minutes d'arrêt de la transcription, 1 287 transcrits (51 %), sont entièrement dégradés dans la souche sauvage.

En revanche, dans la souche où la PNPase a été délétée, seulement 444 transcrits, (18 %), sont entièrement dégradés. Ces résultats suggèrent que la PNPase joue un rôle important dans le processus de dégradation des ARN chez *S. aureus*.

Finalement, bien que la déphosphorylation des extrémités 5' des ARN soit une étape clé pour l'initiation de la dégradation par voie dépendante des extrémités 5' chez de nombreuses espèces bactériennes, chez *S. aureus*, la délétion de la SaRppH identifiée comme réalisant cette activité n'affecte que très peu le transcriptome de la bactérie (4 transcrits stabilisés) (Bonnin & Bouloc, 2015). Ces résultats laissent penser qu'une voire, plusieurs autres enzymes de type NUDIX pourrai(en)t être présente(s) chez *S. aureus*, permettant une compensation fonctionnelle en cas de perte d'une NUDIX hydrolase.

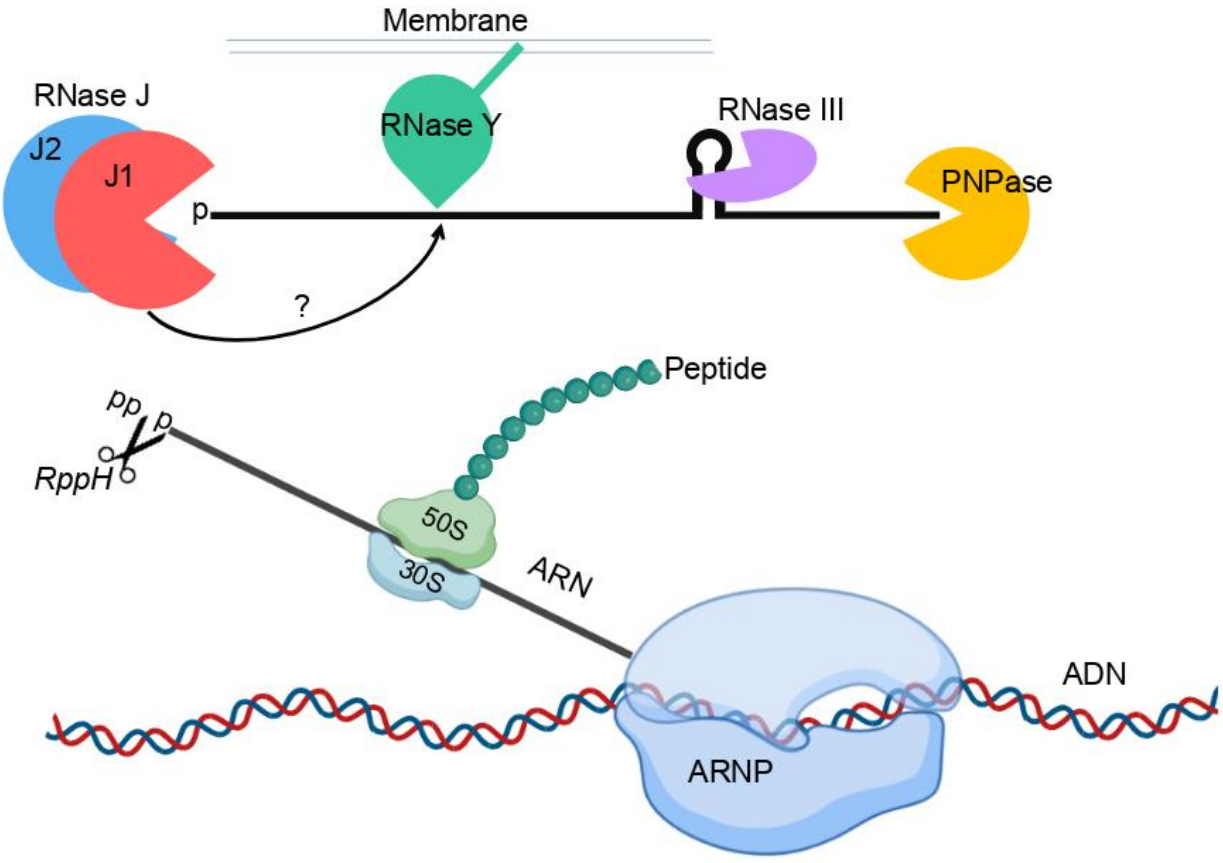


Figure 35 *:* ***Principaux acteurs de la dégradation des ARN chez S. aureus****. Le mécanisme de transcription co-traductionnelle est également représenté (Réalisée avec Biorender)*

1.2 Analyses transcriptomiques

Le terme "transcriptome" apparaît à la fin des années 90 (Velculescu et al., 1997) et fait référence à l'ensemble complet des molécules d'ARN présentes dans une cellule, un tissu ou un organisme à un moment donné. Ces molécules d'ARN, ou transcrits, sont produites par la transcription de l'ADN et peuvent représenter une grande variété de types d'ARN, notamment les ARNm mais aussi les ARNnc, les ARNr, les ARNt, *etc.* Contrairement à l'analyse du génome qui permet une représentation statique des organismes, l'analyse du transcriptome permet d'accéder à l'ensemble des transcrits produits dans une population de cellules en fonction de différents paramètres et/ou conditions environnementales. Les technologies permettant d'étudier le transcriptome ont considérablement évolué au fil du temps et permettent aujourd'hui de séquencer des transcriptomes bactériens entier en quelques heures.

1.2.1 Puces à ADN

Les premières puces à ADN font leur apparition dans les années 1990 (Stephen Fodor, Affimetrix). Les puces à ADN fonctionnent sur le principe de l'hybridation complémentaire, où deux séquences d'acides nucléiques complémentaires s'associent sous l'influence de la température et de la concentration saline (Southern et al., 1999). Ces dispositifs utilisent un support rigide, généralement en verre ou nylon, sur lequel sont fixées des milliers de courtes séquences d'ADN appelées "sondes", conçues pour être spécifiques à une séquence cible. Pour l'analyse, l'ARN total est extrait des échantillons et converti en ADN complémentaire (ADNc) marqué avec des fluorochromes. Ces ADNc marqués, ou "cibles", sont ensuite hybridés avec les sondes sur la puce. Après l'hybridation, les puces sont lavées pour éliminer les cibles non liées, puis scannées pour détecter les signaux fluorescents émis par les cibles hybrides. Après acquisition des images d'hybridation, la quantification des signaux d'hybridation reflète le niveau d'expression de chacun des gènes représentés sur la puce (Fig. 36). Cette méthode permet l'analyse quantitative beaucoup plus fine que les méthodes traditionnelles comme le Northern Blot et l'analyse d'un transcriptome complet en quelques semaines (si les puces sont disponibles) alors qu'il fallait jusque-là parfois des années pour identifier en aveugle un gène dans un processus donné sur la base de la modification de son expression dans une situation donnée.

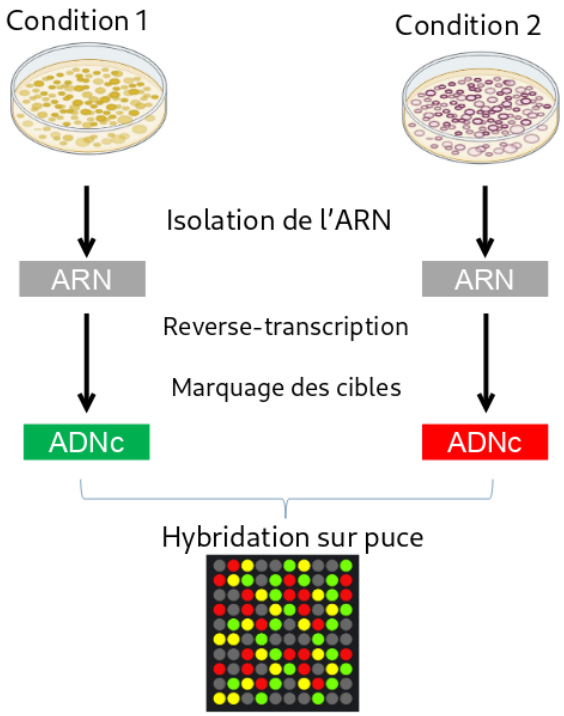


Figure 36 : Analyse différentielle d'expression de gène par puces à ADN.

Les ARN extraits de cultures subissent une transcription inverse en ADNc simple brin, en présence d'un fluorophore, vert pour une condition et rouge pour l'autre. Les ADNc sont hybridés sur une puce à ADN contenant des séquences complémentaires à tous les gènes du génome, situées à un endroit prédéterminé. La quantité de chaque fluorochrome est déterminée à l'aide d'un lecteur et d'un traitement informatique. Si le signal est vert, cela signifie que l'ARN est enrichi dans la condition 1, s'il est rouge, il est enrichi dans la condition 2 et s'il est jaune, l'ARN est en même quantité dans les deux conditions.

Les puces à ADN ont été une avancée majeure dans l'analyse du transcriptome, mais elles présentent certains inconvénients, parmi lesquels, la nécessité de devoir concevoir et fabriquer des sondes spécifiques pour chaque gène ou région d'intérêt. Cela peut être coûteux et fastidieux, et il est parfois difficile de concevoir des sondes spécifiques pour tous les gènes d'intérêt, en particulier dans le cas d'organismes peu étudiés ou présentant des variants génétiques. De plus, les puces à ADN peuvent avoir une sensibilité limitée pour détecter les ARN de faible abondance. Les ARN faiblement exprimés peuvent être sous-représentés ou ne pas être détectés, ce qui peut conduire à des résultats incomplets ou biaisés. Enfin, les puces à ADN peuvent présenter des problèmes de spécificité car une sonde peut s'hybrider avec des ARN non ciblés, conduisant à la détection de faux positifs ou à une surestimation de l'expression de certains gènes.

1.2.2 Séquençage de l'ARN

De nombreuses méthodes permettant le séquençage des ARN ont vu le jour et permettent le séquençage du transcriptome entier ou de certains ARN en particulier et le séquençage de séquences courtes ou longues. La plupart de ces méthodes effectue en réalité le séquençage d'ADN, les ARN à séquencer sont donc, en général, retro-transcrits en ADN complémentaire (ADNc).

1.2.2.1 Les débuts et le séquençage Sanger (1ère génération)

Le séquençage de première génération fait référence aux premières techniques développées pour déterminer la séquence des nucléotides dans l'ADN. La méthode la plus emblématique de cette génération est le séquençage de Sanger, développé par Frederick Sanger en 1977 (Sanger et al., 1977) (Fig. 37). La molécule d'ADN à séquencer est amplifiée par PCR avec un mélange de nucléotides (dNTP) et de nucléotides modifiés et marqués (ddNTP) dont la ligation qui arrêtent la synthèse de l'ADN lorsqu'ils sont incorporés dans la chaîne. La réaction de séquençage est effectuée dans quatre tubes séparés, chacun contenant la matrice d'ADN, l'amorce, l'ADN polymérase, les dNTPs, et une petite quantité d'un des quatre ddNTPs marqués. La synthèse de la nouvelle chaîne d'ADN commence à partir de l'amorce et continue jusqu'à ce qu'un ddNTP soit incorporé, ce qui arrête l'extension de la chaîne. Comme les ddNTPs sont présents en faibles concentrations par rapport aux dNTPs, la réaction produit un ensemble de fragments d'ADN de différentes longueurs, chacun se terminant par un ddNTP marqué. Ces molécules sont ensuite séparées par électrophorèse capillaire et grâce au marquage, il est possible d'identifier la base (A, G, T ou C) par exposition aux UV et en lisant le signal de couleur.

Le séquençage Sanger a été une révolution dans le monde de la génomique et de la biologie moléculaire, et, bien qu'il soit encore utilisé dans des cas spécifiques, son coût élevé, son faible débit et la lenteur de son processus fait que la technologie a dû évoluer vers des techniques haut débit permettant notamment l'étude du transcriptome entier.

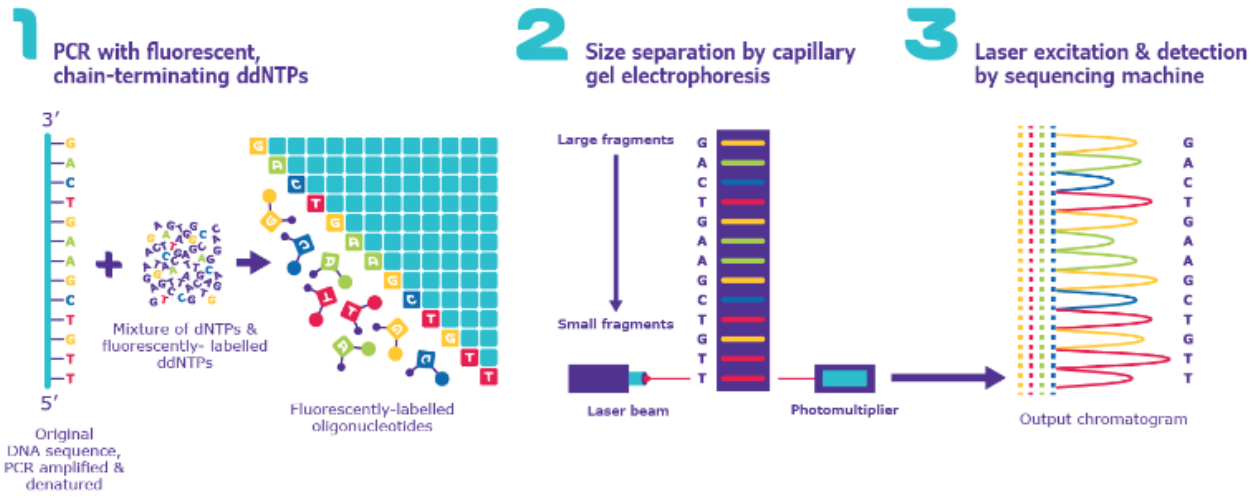


Figure 37 : Etapes du séquençage Sanger.

(1) PCR avec des ddNTPs fluorescents : Mélange de dNTPs et de ddNTPs marqués pour produire des oligonucléotides fluorescents de taille différents. (2) **Séparation par électrophorèse** : Les fragments d'ADN de différentes tailles sont séparés. (3) **Détection des signaux fluorescents** pour produire un chromatogramme de sortie. *Extrait du site web de "sigma aldrich"*

1.2.2.2 2ème génération et apparition du séquençage haut débit

Les technologies de séquençage de deuxième génération, également connues sous le nom de séquençage de nouvelle génération (NGS), ont fait évoluer le domaine de la génomique en permettant un séquençage à haut débit plus rapide, moins coûteux et plus efficace que les méthodes de première génération. Le RNA-Seq est une technologie NGS développée au début des années 2000 qui permet de séquencer l'ensemble des ARN présents dans un échantillon biologique au travers d'une numérisation rapide de millions de courtes séquences d'ADN. Ces méthodes ont profondément transformé la manière d'étudier l'expression génique et la régulation des gènes. Plusieurs méthodes de séquençage haut débit ont rapidement émergé, offrant des capacités de séquençage toujours plus grandes et des coûts de plus en plus bas.

1.2.2.2.1 « 454 Sequencing » (2005)

La technologie de séquençage 454, développée par Roche et commercialisée en 2005, est basée sur le séquençage par synthèse et l'utilisation du pyroséquençage. Il s'agit de l'une des premières technologies NGS commercialisées. Dans cette technologie, les fragments d'ADN sont liés à un adaptateur permettant leur ligation sur des billes.

Les billes liées à un fragment d'ADN sont récupérées et mises sur une plaque à puits (1 bille par puit). Les fragments d'ADN sont ensuite séquencés par synthèse. Pour chaque cycle de séquençage, un seul type de nucléotide (A, T, C ou G) est introduit dans la réaction. Lorsqu'un nucléotide complémentaire est incorporé à l'extrémité 3' d'un brin d'ADN, une lumière est émise. Les images sont analysées pour déterminer l'ordre d'incorporation des nucléotides et la séquence d'ADN est alors reconstruite en alignant les signaux lumineux pour chaque cycle de séquençage.

Bien que cette technologie ne soit plus largement utilisée, elle a été l'une des premières à offrir des lectures plus longues (jusqu'à 400 paires de base) que celles disponibles avec d'autres technologies de séquençage de deuxième génération.

1.2.2.2.2 Illumina (2007)

La technologie Illumina, anciennement connue sous le nom de « Solexa Sequencing », est l'une des plateformes de séquençage de deuxième génération les plus répandues. Elle repose sur l'amplification en grappes et le séquençage par synthèse. Les fragments d'ADN à séquencer sont immobilisés sur une surface solide et amplifiés en clusters. Ensuite, des nucléotides marqués sont ajoutés un par un, et la fluorescence est détectée pour identifier chaque base incorporée. Cette méthode permet de générer des lectures courtes (~150 bp), mais très fidèles, à un coût relativement faible.

De manière plus précise, l'ADN cible est tout d'abord fragmenté, et des adaptateurs sont ajoutés aux extrémités des fragments pour permettre leur attachement à une puce (flow cell). Une fois les fragments fixés, ils sont amplifiés localement via une méthode appelée "bridge amplification", formant des amas d'ADN ou « grappes ». Lors du séquençage, des nucléotides marqués par fluorescence sont ajoutés un par un, et chaque incorporation est détectée par un signal fluorescent. Les terminateurs réversibles sur les nucléotides garantissent qu'un seul nucléotide est ajouté à la fois. Les signaux fluorescents sont capturés et analysés pour déterminer la séquence d'ADN.

Le déroulement d'un cycle est le suivant : une base modifiée vient se fixer par complémentarité en face de la matrice, c'est-à-dire la molécule à séquencer. Les bases ne s'étant pas fixées sont lavées. Ensuite, un laser excite les fluorochromes pour détecter le signal fluorescent émis par les nucléotides incorporés, permettant ainsi de lire la séquence. Le terminateur est ensuite retiré pour permettre l'ajout du prochain nucléotide au cycle suivant (Fig. 38).

Le nombre de cycles effectué déterminera la taille des fragments séquencés (entre 50 et 600 cycles). Les données sont fournies sous forme d'un fichier texte de type FASTQ, avec autant de lignes qu'il y avait de molécules attachées à la cellule d'écoulement. A chaque molécule est associé un identifiant correspondant à la position de la molécule sur la cellule, la séquence lue, et la qualité de séquençage, reflet du degré de confiance quant à l'exactitude de la base lue à chaque position.

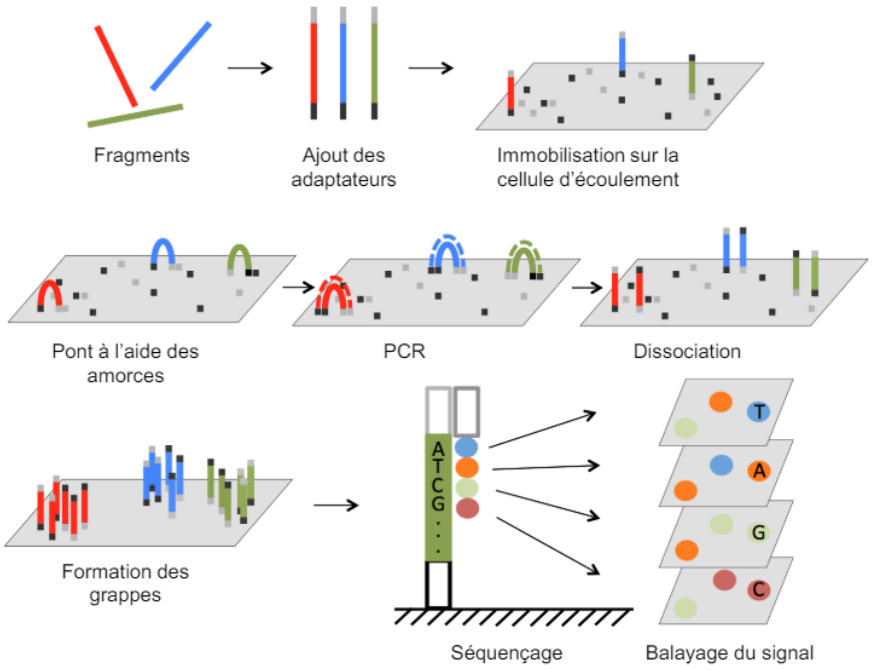


Figure 38 : **Étapes du séquençage avec la technologie Illumina.**

Préparation des fragments d'ADN par l'ajout des adaptateurs aux fragments et fixation sur la cellule d'écoulement (flow cell). Formation de ponts à l'aide des amorces et amplification par PCR. Création de grappes d'ADN clonales. Séquençage des fragments par synthèse. Détection et lecture des bases par fluorescence. *Extrait de* (Lu et al., 2016)

1.2.2.2.3 SOLid (Sequencing by Oligonucleotide Ligation and Detection) (2008)

SOLiD est une méthode enzymatique de séquençage. Dans un premier temps, les ADN à séquencer sont liés en 5' avec un adaptateur. Puis ils sont amplifiés et liés à des billes par PCR à émulsion. Ces billes sont ensuite déposées sur une surface en verre et une amorce de longueur N (violet) est hybridée à l'adaptateur. Les billes sont ensuite exposées à une bibliothèque de sondes de 8 nucléotides qui ont différents colorants fluorescents à l'extrémité 5'. Seule une sonde dont les 2 premiers nucléotides sont complémentaires s'hybridera à la séquence cible, adjacente à l'amorce. Une ligase d'ADN est ensuite utilisée pour lier la sonde de 8 nucléotides à l'amorce.

Le colorant fluorescent est ensuite clivé du fragment à l'aide d'ions argent, ce qui permet de mesurer la fluorescence (quatre colorants fluorescents différents sont utilisés) et de générer également un groupe phosphate qui peut subir une autre ligation. Une fois la première étape de séquençage terminée, le produit d'extension est dissocié, puis une deuxième étape de séquençage est réalisée avec une amorce de longueur N−1 puis une troisième avec une amorce de longueur N-2 et ainsi de suite (Fig. 39).

Cette méthode offre une grande précision, mais produit des lectures relativement courtes (entre 50 et 100 bp). De plus, son coût plus important que les autres plateformes la rend moins couramment utilisée depuis l'avènement de la technologie Illumina.

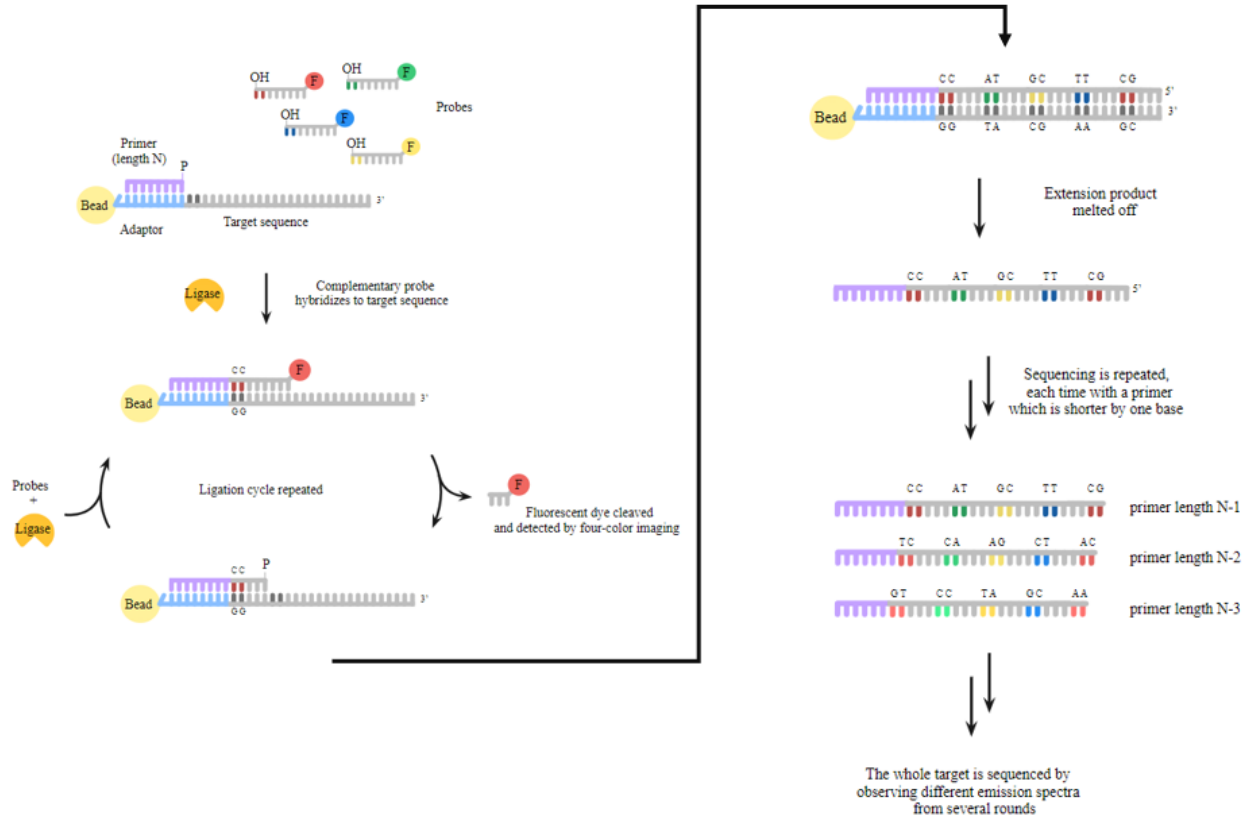


Figure 39 : Étapes du séquençage avec la technologie SOLiD.

*Hybridation de la séquence cible avec un amorce fixé sur une bille. Les sondes complémentaires marquées par fluorescence s'hybrident à la séquence cible. Les fluorophores sont clivés et détectés par imagerie à quatre couleurs. Le cycle de ligation et de détection est répété plusieurs fois. Chaque cycle de séquençage est répété avec une amorce raccourcie d'une base pour lire différentes positions. La séquence complète est déterminée en observant les différents spectres d'émission sur plusieurs cycles. Extrait du site web de « **atdbio** »*

1.2.2.2.4 Ion torrent (2010)

Cette technologie de séquençage repose sur la détection des ions d'hydrogène produits lors de l'incorporation de nucléotides par une ADN polymérase. Lorsque des nucléotides sont incorporés dans un brin d'ADN, ils libèrent des ions d'hydrogène qui sont détectés par un capteur à semi-conducteurs. Chaque type de nucléotide (A, T, C ou G) libère un nombre spécifique d'ions d'hydrogène lors de son incorporation. Cette méthode permet de générer des lectures relativement courtes mais à un coût compétitif.

1.2.2.3 Séquençage de 3ème génération

Il existe aujourd'hui des technologies de 3ème génération (TGS) telle que les technologies Pacific Biosciences (Pacbio 2010) et Oxford Nanopore Technologie (ONT 2014). Une des différences majeures des TGS par rapport au NGS réside dans l'utilisation des principes de séquençage de molécules uniques et de séquençage en temps réel pour permettre le séquençage de molécules de nucléotides (ADN ou ARN) sans qu'il soit nécessaire d'amplifier le modèle par PCR.

La technologie Pacbio reposent sur l'observation directe de la synthèse d'ADN en temps réel à partir de molécules individuelles (Fig. 40a). Cette technique repose sur le principe de l'activité de l'ADN polymérase, qui fondamentalement, se lie à l'ADN matrice, où chacune des quatre bases nucléotidiques (dNTP) est marquée de façon unique avec des fluorophores. Pendant la phase d'appariement des bases, l'incorporation de différents nucléotides émet des longueurs d'onde de lumière différentes, discernables par leurs pics et leurs longueurs d'onde respectifs (Fig. 40b). En outre, l'ADN polymérase joue un rôle essentiel dans l'obtention de lectures exceptionnellement longues (>50 kb), une caractéristique attribuée à son activité enzymatique (Athanasopoulou et al., 2021).

La technologie ONT repose sur la détection des variations du courant électrique provoquées par la désorganisation des protéines nanopores lorsque des brins d'ADN ou d'ARN les traversent (Fig. 40b). Contrairement aux méthodes NGS qui détectent des signaux secondaires, la technologie des nanopores identifie directement les changements dans le courant électrique qui sont produits en temps réel pendant le séquençage. Les modifications de courant sont distinctes pour chaque nucléotide, générant une signature unique pour chaque base. Le processus de traduction du signal électrique en bases nucléotidiques est appelé "basecalling".

Cependant, les systèmes ONT n'identifient pas les bases individuelles puisque le courant observable est déterminé par de courtes séquences de nucléotides (Wick et al., 2019). Ces séquences d'environ cinq bases sont appelées k-mers et 1024 signaux différents peuvent donc être détectés. L'ONT permet la production de lectures de séquençage ultra-longues (>4 Mb).

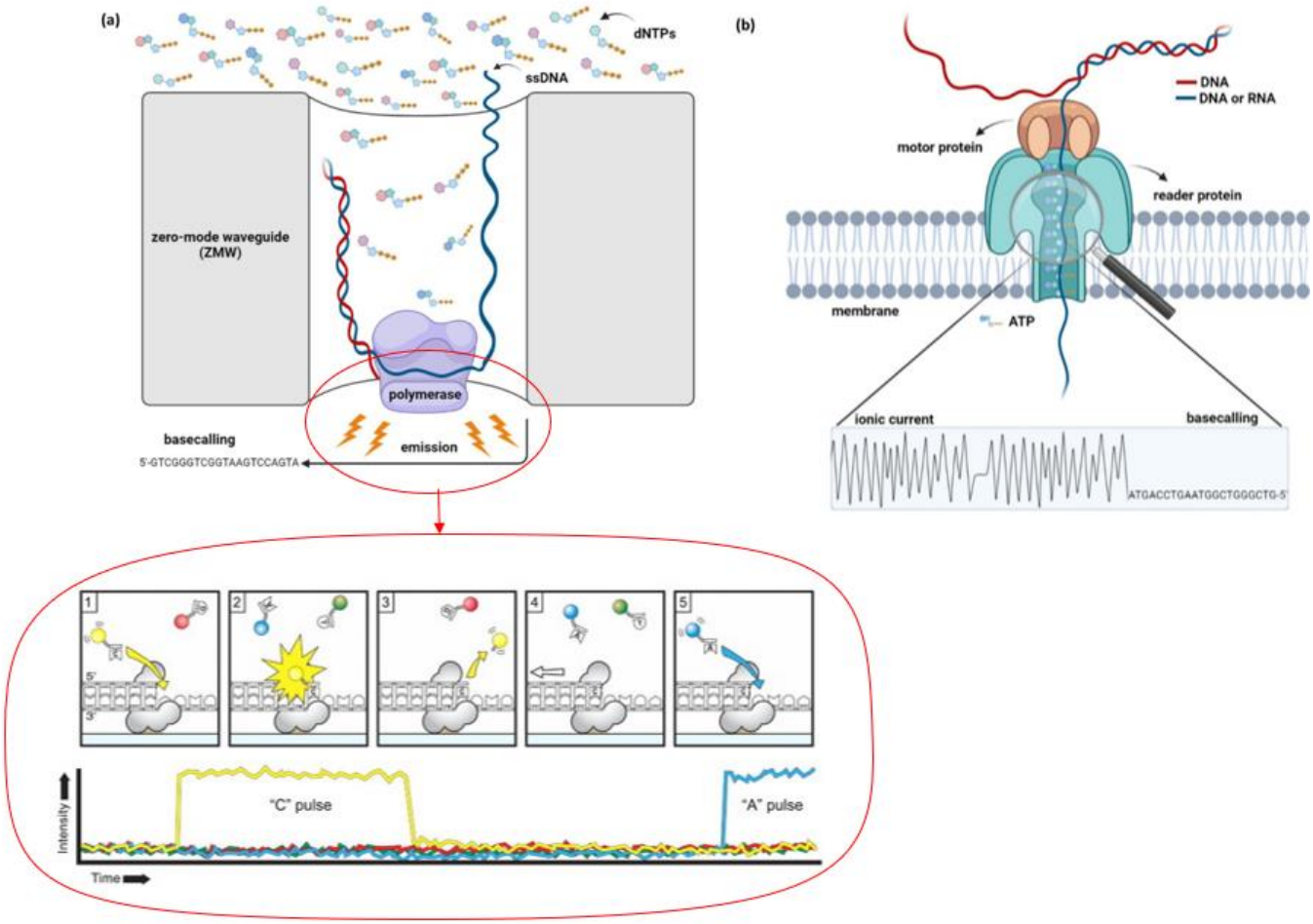


***Figure 40 :* Technologies de séquençage de troisième génération.**

*(**a**) Séquençage PacBio. La méthode est basée sur le séquençage de l'ADN par synthèse. L'ADN simple brin est lié à l'ADN polymérase, qui catalyse l'incorporation de dNTP marqués dans un brin d'ADN complémentaire nouvellement synthétisé. La réaction produit un signal différent en fonction du dNTP incorporé (**b**). Le signal est ensuite traduit en séquence de nucléotides, une procédure connue sous le nom de "basecalling". (**c**) Séquençage par nanopores. La méthode repose sur le guidage d'une molécule d'ADN ou d'ARN simple brin vers un nanopore, une protéine de lecture qui détecte les altérations du courant électrique qui se produit lors du passage de l'ADN/ARN. Extrait de* (Athanasopoulou et al., 2021) *et du site web « CD genomics »*

1.2.2.4 Préparation des librairies RNA-Seq

La préparation d'une librairie RNA-Seq est une étape essentielle avant le séquençage, car elle permet de convertir les molécules d'ARN en une forme compatible avec la technologie de séquençage utilisée (généralement en ADN). Cette préparation se compose de plusieurs étapes illustrées dans la Figure 41 et permettant, entre autres, l'amplification des molécules, l'élimination des contaminants ou encore le multiplexage d'échantillons.

La première étape consiste à extraire l'ARN total à partir de l'échantillon bactérien à l'aide de méthode de lyse chimique ou mécanique des cellules, suivie d'une purification de l'ARN total. L'étape suivante correspond à l'élimination de l'ADN contaminant à l'aide de nucléase afin de garantir que seul des molécules d'ARN sont présentes dans l'échantillon. Si l'objectif est de séquencer spécifiquement l'ARNm, une étape d'enrichissement de l'ARNm peut être réalisée pour éliminer l'ARN ribosomal (ARNr), qui est surreprésenté (~95% des ARN totaux) et d'autres ARN non-messagers. Cela peut être fait à l'aide kit de ribodéplétion ou des méthodes de sélection spécifique de l'ARNm. L'ARN total ou l'ARNm enrichi est ensuite fragmenté en morceaux plus courts par hydrolyse chimique ou par une fragmentation enzymatique contrôlée. Ces fragments d'ARN sont ensuite convertis en ADN complémentaire (ADNc) à l'aide d'une enzyme appelée transcriptase inverse ou rétro-transcriptase (RT). Cette étape permet également l'ajout d'un oligonucléotide au niveau des extrémités des brins d'ADNc qui servira pour la ligation des adaptateurs d'index (code-barres) lors des étapes suivantes. Ces oligonucléotides contiennent des amorces pour la PCR ultérieure ainsi que des séquences d'index uniques qui permettent de multiplexer plusieurs échantillons dans une seule étape de séquençage. Les brins d'ADNc avec les adaptateurs d'index sont ensuite amplifiés par PCR pour obtenir une quantité suffisante de matériel pour le séquençage. Pour éliminer les amorces d'amplification, les adaptateurs libres, les produits de PCR non désirés et d'autres contaminants, les ADNc peuvent être purifiés par électrophorèse. Les fragments migreront à travers le gel en fonction de leur taille et les bandes contenant les fragments d'ADNc de la taille souhaitée sont excisées du gel. Une fois que la librairie d'ADNc prête, les fragments contenus seront séquencés par des méthodes NGS et les résultats seront analysés à l'aide d'approches *in silico*.

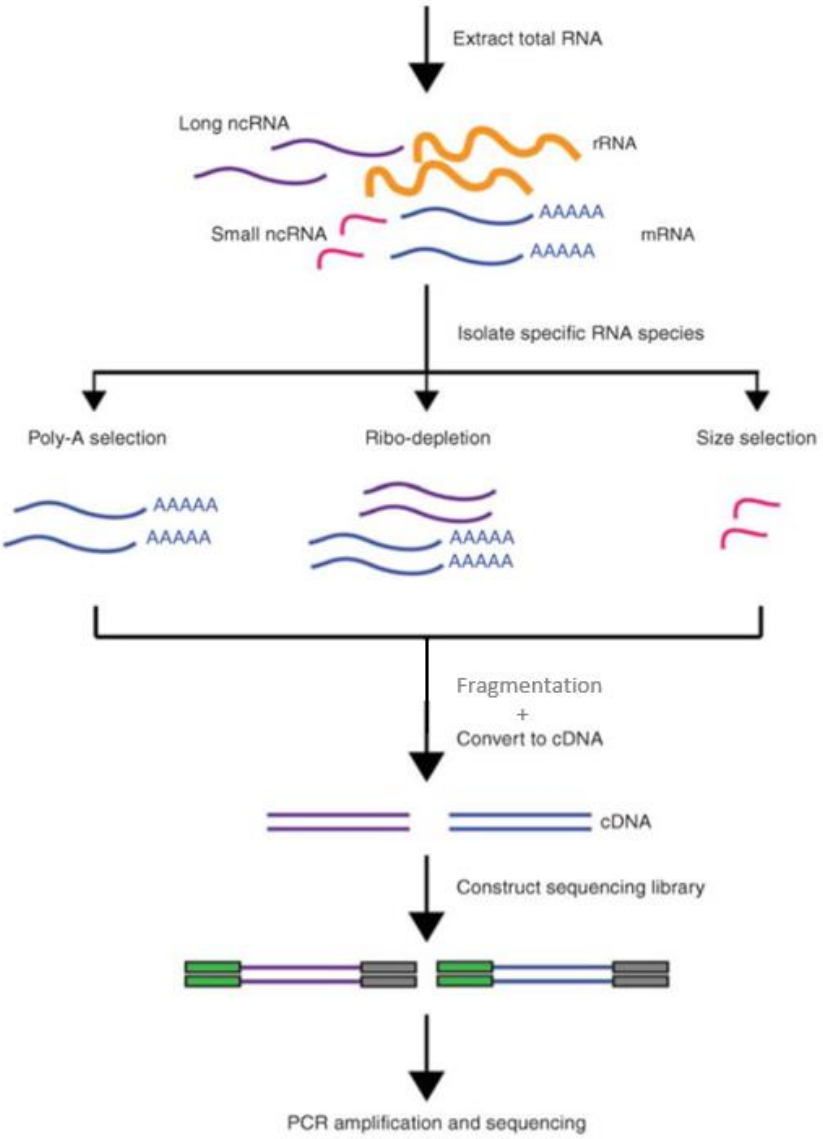


***Figure 41 :* Préparation d'une librairie RNA-Seq.**

Premièrement, l'ARN est extrait du matériel biologique choisi (cellules, tissus, etc.). Deuxièmement, des sous-ensembles de molécules d'ARN sont isolés à l'aide d'un protocole spécifique, tel que le protocole de sélection polyA pour enrichir les transcrits polyadénylés (ARNm eucaryotes) ou un protocole de déplétion ribosomique pour éliminer les ARN ribosomiques et/ou une purification sur gel pour sélectionner les ARN en fonction de leur taille. Ensuite, les ARN sont fragmentés puis convertis en ADN complémentaire (ADNc) par transcription inverse et les adaptateurs de séquençage sont ligaturés aux extrémités des fragments d'ADNc. Suite à l'amplification par PCR, la librairie RNA-Seq est prête pour le séquençage. Adapté de (Kukurba & Montgomery, 2015)

Le séquençage peut se faire en « Single-end » ou en « Paired-end ». Dans le séquençage single-end, chaque fragment d'ADN est séquencé à partir d'une seule extrémité (5’ ou 3’), générant une seule séquence pour chaque fragment (Fig. 42 haut). Le séquençage paired-end génère des lectures des deux extrémités d'un fragment d'ADN, produisant deux séquences pour chaque fragment. (Fig. 42 bas). Si le fragment à séquencer est au moins deux fois plus long que la taille d’une lecture, il y aura entre les deux un « gap » dont la longueur est appelée « inner mate distance ».

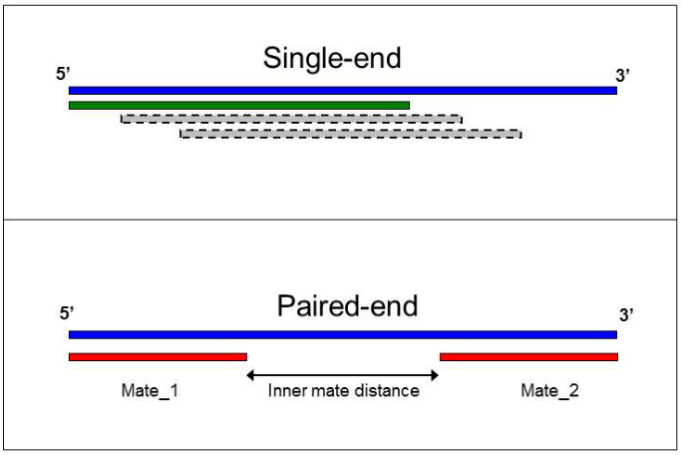


Figure 42 : Schéma d’un séquençage single-end (en haut) et paired-end (en bas)*. Extrait du manuscrit de thèse de Fidel BOTERO-CASTRO (2014).*

Ce protocole peut varier et comporter d’autres étapes plus spécifiques en fonction du type de molécule à séquencer (ARN totaux, ARNm, extrémités des ARN, *etc.*). Il convient de noter que chaque étape présente un risque d’introduction de biais, comme des biais de ligation, des biais d’amplification PCR, *etc.* Les protocoles sont donc en général conçus avec le minimum d’étape possible afin de diminuer les biais potentiels et d’obtenir des résultats au plus proche de la réalité biologique.

1.2.2.5 Applications et variantes du RNA-Seq

Le RNA-Seq possède des applications diverses et variées, allant de l'étude de l'expression génique à la détection des variants, en passant par la découverte de nouveaux ARN non codants. En fonction des objectifs de recherche, des modifications peuvent être apportées au protocole de préparation des librairies précédemment décrit. Cependant, ces modifications ne changent, en générale, pas la logique expérimentale et ce qui va le plus orienter les analyses subséquentes ce sont les ARN qui seront séquencés. En effet, dans certaines applications décrites juste après, la préparation de librairies ne se fait pas à partir des ARN totaux mais plutôt sur un sous-ensemble d'ARN préalablement sélectionnés et purifiés.

1.2.2.5.1 Catalogue de gènes exprimés et comparaison entre conditions

Le RNA-Seq permet de mesurer de manière précise et globale les niveaux d'expression des gènes dans différents états physiologiques cellulaires, de stress ou différentes conditions environnementales. Cela permet de déterminer quels sont les gènes exprimés et à quel niveau, ce qui est crucial pour étudier la régulation génique, les voies métaboliques et les réponses aux stimuli externes. Il permet également de comparer les profils d'expression génique entre différentes conditions afin de mieux comprendre les mécanismes des réponses adaptatives bactériennes (Lasa et al., 2011), (Liu et al., 2014), (Redko et al., 2016), (Pobre & Arraiano, 2015), (Leskinen et al., 2016).

Initialement, ces comparaisons se faisaient avec les méthodes de puces à ADN en comparant les intensités de fluorescence (variable quantitative continue). Depuis l'avènement du RNA-Seq, ce sont les comptages (variable quantitative discrète), reflétant le niveau d'expression des gènes qui sont comparés de manière statistique, souvent en prenant en compte des réplicats. Les librairies EdgeR (Robinson et al., 2010) et DESeq2 (Love et al., 2014), (Anders & Huber, 2010) permettent de faire ces comparaisons directement dans l'environnement de programmation R (dépôt Bioconductor). Ces libraires implémentent directement des méthodes de normalisation et de filtrage en plus de méthodes permettant l'analyse d'expression différentielle. Les deux utilisent des approches différentes. La méthode utilisée par EdgeR se base sur un test analogue au test exact de Fisher alors que celle de DESeq2 se base sur l'utilisation d'un modèle linéaire généralisé (GLM) suivi d'un test de Wald. Dans les deux cas, ces méthodes modélisent la distribution des comptages à l'aide d'un modèle basé sur la loi binomiale négative et utilisent un paramètre de dispersion pour modéliser la variance biologique et technique des échantillons.

1.2.2.5.2 Annotation de génome

Le RNA-Seq peut également être utilisé pour détecter de nouveaux gènes, transcrits et variants génétiques. En séquençant l'ARN messager, il est possible d'identifier des transcrits non annotés et des variants de séquence qui pourraient jouer un rôle important dans la physiologie de la cellule. Le RNA-Seq peut également être utilisé pour détecter et caractériser les ARN non codants (ARNnc), qui ne codent pas pour des protéines mais remplissent diverses fonctions régulatrices dans la cellule.

1.2.2.5.3 Détermination de la demi-vie des ARN

Le calcul des demi-vies des ARN à l'aide du RNA-Seq est une approche utilisée pour mesurer la stabilité relative des ARNm dans une cellule (Z. Chen et al., 2013)**,** (X. Chen et al., 2015)**,** (Kristoffersen et al., 2012)**.** La demi-vie d'un ARN fait référence à la durée nécessaire pour que la concentration d'un ARN spécifique dans une cellule diminue de moitié. En d'autres termes, c'est le temps nécessaire pour que la moitié des molécules d'un ARN spécifique présentes dans une cellule soient dégradées ou perdues. Cette demi-vie est importante pour étudier la régulation de l'expression génique et les processus cellulaires qui contrôlent la dégradation des ARNm. Dans un premier temps, la transcription est bloquée à l'aide d'un inhibiteur de la transcription, tel que la rifampicine. Cela permet d'arrêter la synthèse de nouveaux ARNm et donc de mesurer la dégradation des ARNm déjà synthétisés. Un échantillonnage à différents temps est ensuite effectué, c'est-à-dire que des échantillons cellulaires sont prélevés à des intervalles de temps définis après le blocage de la transcription, permettant de quantifier le niveau de chaque ARN à chacun des temps. Des librairies RNA-Seq sont préparées à partir de ces échantillons, puis, sont séquencées, permettant de quantifier les niveaux d'expression de tous les ARNm dans chaque échantillon et donc de mesurer la dynamique de dégradation des ARNm au fil du temps. Une fois la table de comptages générée, la demi-vie de chaque ARNm est calculée à l'aide d'un modèle linéaire (Fig. 43). Les comptages subissent une transformation logarithmique en base 2, et un modèle linéaire est ajusté pour chaque ARN (comptage en fonction du temps). La pente α de la courbe résultant du modèle donne la demi-vie (demi-vie = -1/α).

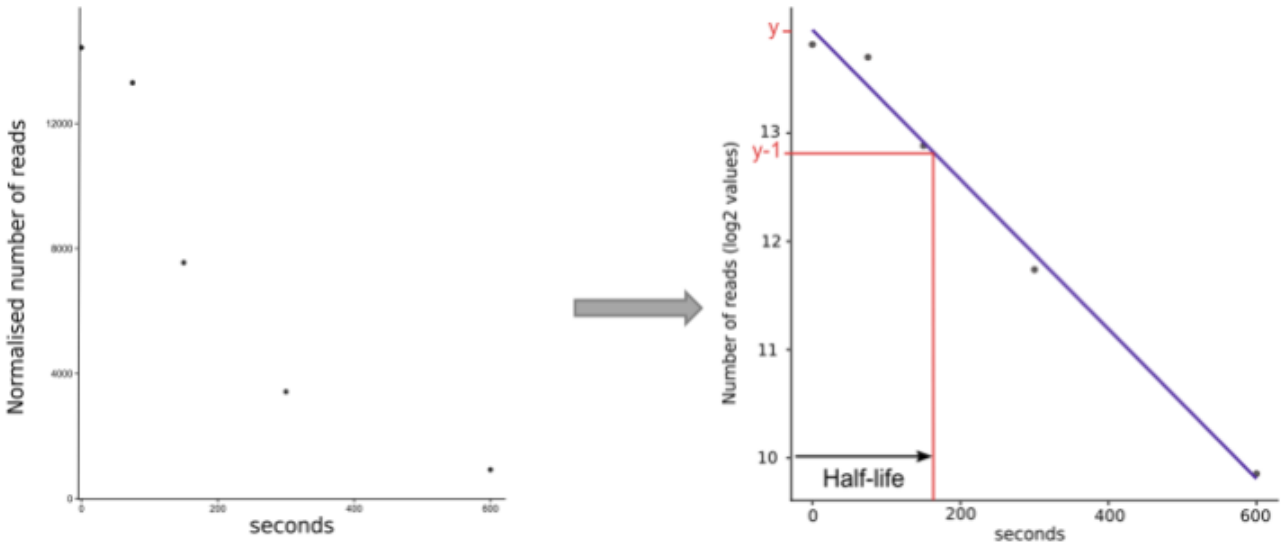


Figure 43 : ***Calcul de la demi-vie d'un ARN.*** *Sur le graphique de gauche, les nombres de lectures alignées sur une CDS (axe y) aux différents temps (axe x). Dans le graphique de droite, les valeurs logarithmiques en base 2 des nombres de lectures alignées sur la CDS sont représentées (axe y). La régression linéaire ajustée sur les comptages est représentée par la ligne bleue. La régression est utilisée pour calculer le temps (en secondes) nécessaire pour que le niveau d'ARN soit divisé par deux (ce qui signifie, en logarithme de base 2, le niveau d'origine moins 1 ; y-1). Extrait du rapport de stage de* ***Natalia Baumann.***

1.2.2.5.4 Etude de l'interactome ARN

Les RNA-Seq peut également permettre d'étudier les interactions impliquant l'ARN, que ce soit avec des protéines, d'autres ARN, des ribonucléoprotéines, de l'ADN ou des petites molécules.

Ces approches sont donc très utiles pour mieux comprendre les mécanismes régulant l'expression des gènes, au niveau de la transcription avec par exemple les interactions ARN- protéine Rho, au niveau de la traduction avec les interactions ARN-ribosome ou encore au niveau de la dégradation avec les interactions ARN-ribonucléases ou ARN-ARNas.

Le principe de ces approches est de fixer les interactions entre les ARN et la molécule d'intérêt puis des méthodes de purification telles que la chromatographie d'affinité ou le pull-down peuvent être utilisées pour isoler les complexes d'intérêt ainsi fixés. Les ARN engagés dans les complexes sont ensuite extraits, isolés puis séquencés afin d'obtenir une liste des régions des ARN interagissant avec la ou les molécules d'intérêt.

Par exemple, grâce à ces approches, les ARN directement ciblés par la RNase III ont pu être identifiés chez *S. aureus* (Lioliou et al., 2012) et chez *E. coli* (Lybecker et al., 2014). A partir de ce principe, plusieurs protocoles ont vu le jour et notamment les protocoles CLASH (Helwak & Tollervey, 2014) pour l'étude des interactions ARN-ARN, CRAC (Philippe et al., 2013) pour les interactions ARN-protéine et Ribo-seq (Ingolia et al., 2009) pour les interactions ARN-ribosome.

1.2.2.5.4.1 ARN-ARN: CLASH (Cross-Linking and ImmunoPrecipitation followed by High-Throughput Sequencing)

Les interactions ARN-ARN jouent un rôle essentiel dans de nombreux processus cellulaires, comme la régulation par les ARNas par exemple, mais leur étude est difficile et laborieuse. La méthode appelée « cross-linking ligation and sequencing of hybrids » (CLASH), permet l'identification à haut débit des sites d'interaction ARN-ARN (Fig. 44). Le protocole CLASH commence par la fixation des interactions ARN-protéine (protéine marquée avec une étiquette) par cross-linking, généralement avec des UV ou au formaldéhyde, suivie de la lyse des cellules pour libérer les complexes ARN-protéine. Ces complexes sont ensuite isolés par immunoprécipitation à l'aide d'anticorps spécifiques à l'étiquette de la protéine. Les ARN dans les complexes sont ligaturés entre eux pour former des hybrides ARN-ARN, qui sont ensuite purifiés. Après l'ajout d'un linker, la préparation d'une librairie d'ADNc et le séquençage à haut débit, les duplex ligaturés donnent naissance à des ADNc chimériques, qui permettent l'identification des sites d'interaction ARN-ARN (Helwak & Tollervey, 2014).

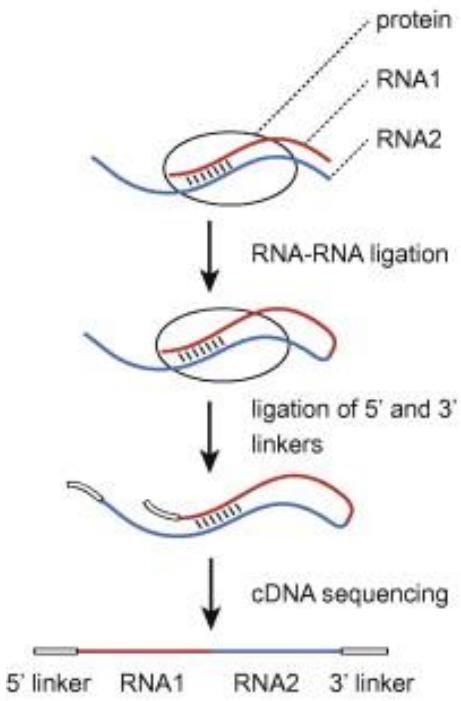


Figure 44 : CLASH cartographie les interactions ARN-ARN. *Dans cette méthode, les complexes ARN-protéines sont réticulés aux UV et purifiés par affinité. Les hybrides ARN-ARN sont ligaturés, isolés et retranscrits en ADNc. Le séquençage de l'ADNc fournit des lectures chimériques à haute résolution des interactions ARN-ARN. Extrait de* (Travis et al., 2014)

1.2.2.5.4.2 ARN-protéine: CRAC (Cross-Linking and Analysis of cDNA)

La méthode CRAC permet l'identification des sites d'interaction entre une protéine marquée et ses ARN cibles. Il est similaire à la méthode CLASH dans son principe de fixation des interactions aux UV. Les ribonucléoprotéines fixées sont ensuite purifiées grâce au marquage de la protéine d'intérêt. Les complexes sont ensuite traités avec un cocktail de ribonucléases non spécifiques (RNase IT = Rnase A + Rnase T1). Cela a pour effet de fragmenter, en petits oligonucléotides, les régions d'ARN libres dans le but de ne garder que les régions plus ou moins protégées par la protéine. Les extrémités sont ensuite modifiées par divers traitements (Fig. 45) puis liées à des adaptateurs. Une étape de purification sur gel permet ensuite de retirer les adapteurs non liés ainsi que les fragments d'ARN non protégés. Enfin, les protéines sont dissociées des ARN puis dégradées par une protéinase. Les ARN sont extraits, reverse transcrits, amplifiés par PCR et finalement séquencés.

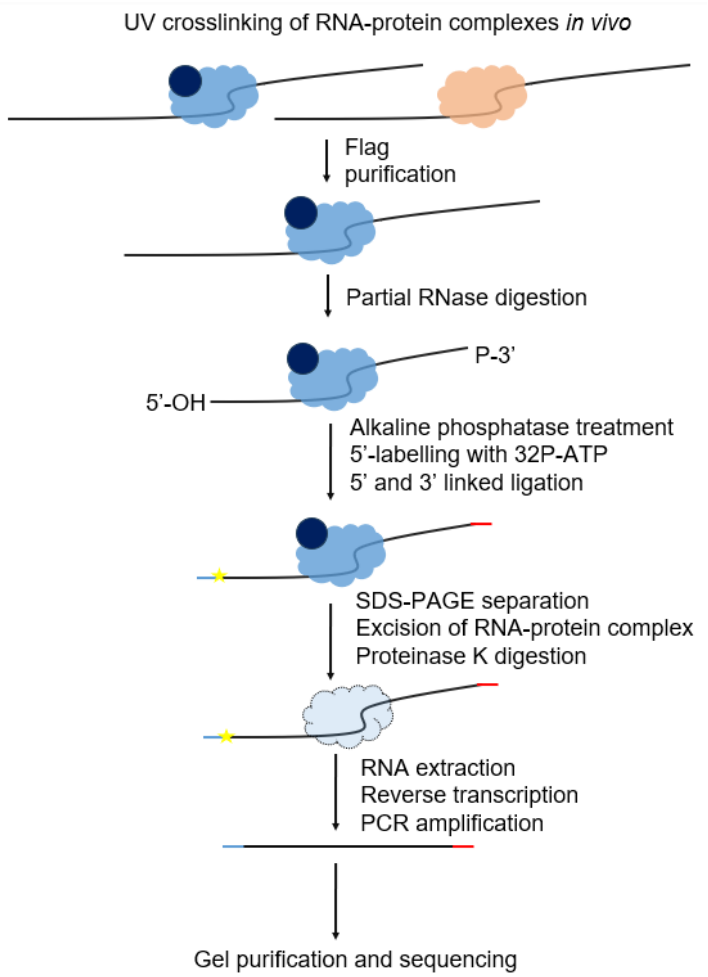


Figure 45 : Identification des sites de liaison à l'ARN à l'aide de la méthode CRAC. *Des ribonucléoprotéines partiellement digérées par les ribonucléases T1 et A, sont incubées avec des billes de nickel pour immobiliser des protéines (nuage bleu) étiquetées (ovales bleus marine) et les ARN fixés de manière covalente (lignes noire). Les extrémités 3' des ARN ainsi purifiées sont déphosphorylées et liées à un adaptateur adénylé (ligne rouge). Les extrémités 5' ont été marqués radioactivement (étoile jaune) avec la polynucléotide kinase, puis liées à un adaptateur (ligne bleue). Après libération par traitement à l'imidazole, les complexes radioactifs sont résolus sur des gels SDS-PAGE et transférés sur nitrocellulose. Les bandes correspondant à la masse moléculaire prédite de la protéine cible sont excisées et digérées à la protéinase K. Les ARN récupérés sont retro-transcrits et amplifiés par PCR. Les produits de la PCR sont finalement purifiés sur gel puis séquencés. Adaptée de* (Sy et al., 2018)**.**

1.2.2.5.4.3 ARN-ribosome : Ribo-Seq

Ribo-seq, ou « ribosome profiling » est un cas particulier puisqu'il permet de séquencer les fragments d'ARN protégés par les ribonucléoprotéines que sont les ribosomes. Ceci permet de cartographier l'emplacement des ribosomes sur les ARNm à l'échelle du transcriptome. Dans cette approche, les cellules sont traitées par un agent chimique pour stabiliser les complexes ARNm-ribosome. Ensuite, les ribosomes sont isolés et les fragments d'ARN protégés par les ribosomes sont séquencés (Fig. 46a). Cela permet de cartographier l'emplacement de ce qui est appelé les empreintes des ribosomes, les régions qui étaient en cours de traduction au moment où les interactions ARN-ribosome ont été fixées. Grâce à cette méthode il est donc possible d'identifier les régions codantes du transcriptome ainsi qu'un motif périodique le long de celles-ci (Fig. 46b). En effet, lorsqu'on capture les interactions ribosome-ARN, il y a plus de chances de tomber sur des positions correspondant à des multiples de 3 nucléotides, car les ribosomes couvrent et protègent des fragments d'ARNm correspondant à des codons entiers (3 nucléotides). Cette régularité crée un profil périodique qui reflète la lecture triplet par triplet des codons par les ribosomes.

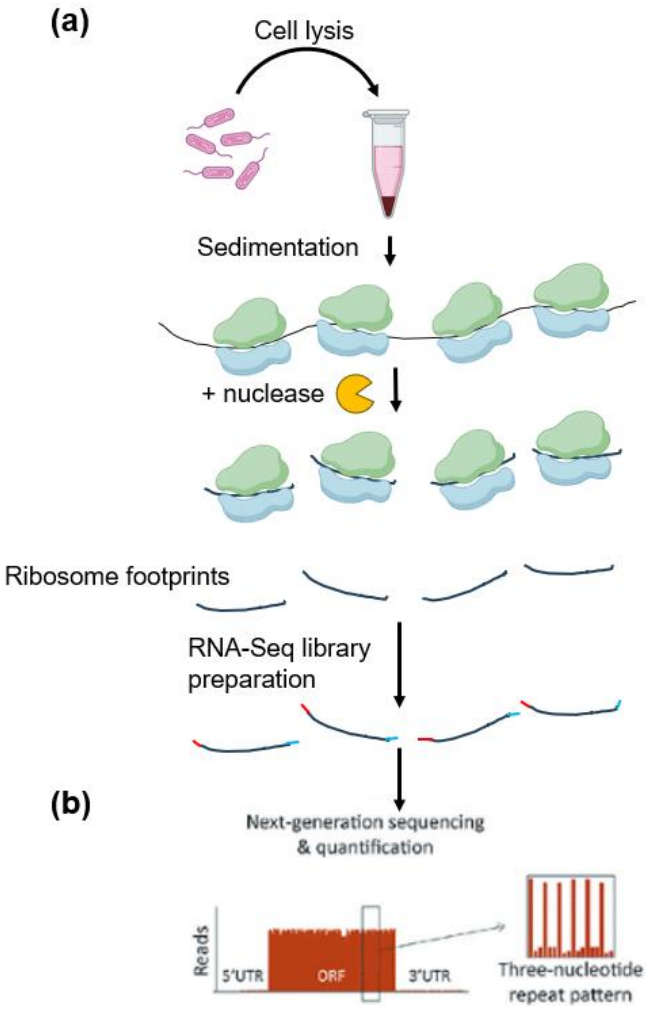


Figure 46 : ***Détection des régions protégées par les ribosomes avec la méthode Ribo-Seq.***

*Dans la méthode Ribo-seq, les ARNm liés aux ribosomes sont purifiés par sédimentation après lyse cellulaire. Un traitement à la nucléase est ensuite utilisé pour dégrader les sections d'ARN non masquées, laissant des fragments de 20 à 30 nucléotides de longs protégés par les ribosomes et connus sous le nom d'empreintes de ribosomes. Les empreintes sont ensuite soumises à la préparation de librairies et au séquençage. (**b**) Les densités de lecture peuvent être utilisées pour distinguer les CDS individuels et pour quantifier leurs taux de traduction, avec un motif d'empreinte périodique distinctif de trois nucléotides qui indique la translocation des ribosomes et qui est souvent visible sur toute la longueur de la CDS. Adaptée de* (Dermit et al., 2017)

1.2.2.5.5 Étude des extrémités des ARN

Le séquençage des extrémités des ARN vise principalement à identifier et à caractériser les extrémités des ARN dans un échantillon biologique donné. Les principaux objectifs de cette méthode sont l'identification de sites d'initiation et de terminaison de la transcription ou encore l'étude de l'ensemble des ARN en cours de dégradation.

Dans cette section, nous nous concentrerons exclusivement sur les méthodes spécifiques au séquençage des ARN bactériens, conformément aux méthodes utilisées dans ces travaux de thèse. Cependant, il est important de noter qu'il existe des techniques dédiées pour les ARN eucaryotes, basées sur la capture de la coiffe en 5'.

Le principe commun des différents protocoles permettant ce type de séquençage est le « RACE » (Rapid Amplification of cDNA Ends) (Fig. 47) dont la première étape consiste à ligaturer un oligonucléotide de séquence connue (en orange) aux extrémités 5'P des ARN, en excluant les ARN 5'PPP qui ne sont pas des substrats pour la ligase. Ensuite, une transcription inverse est réalisée pour générer de l'ADNc (en bleu clair), où pour les ARN ligaturés, l'ADNc se termine par le complément inverse de la séquence oligonucléotide (en jaune). Enfin, une PCR est effectuée pour amplifier l'ADNc et ajouter des séquences adaptatrices nécessaires pour le séquençage NGS (en rouge et marron). Seul l'ADNc des ARN ligaturés est amplifié, et le séquençage commence à partir de la séquence adaptatrice.

De plus, ces protocoles n'incluent pas d'étape de fragmentation des ARN afin de conserver l'information de l'extrémité 5' *in vivo* des ARN.

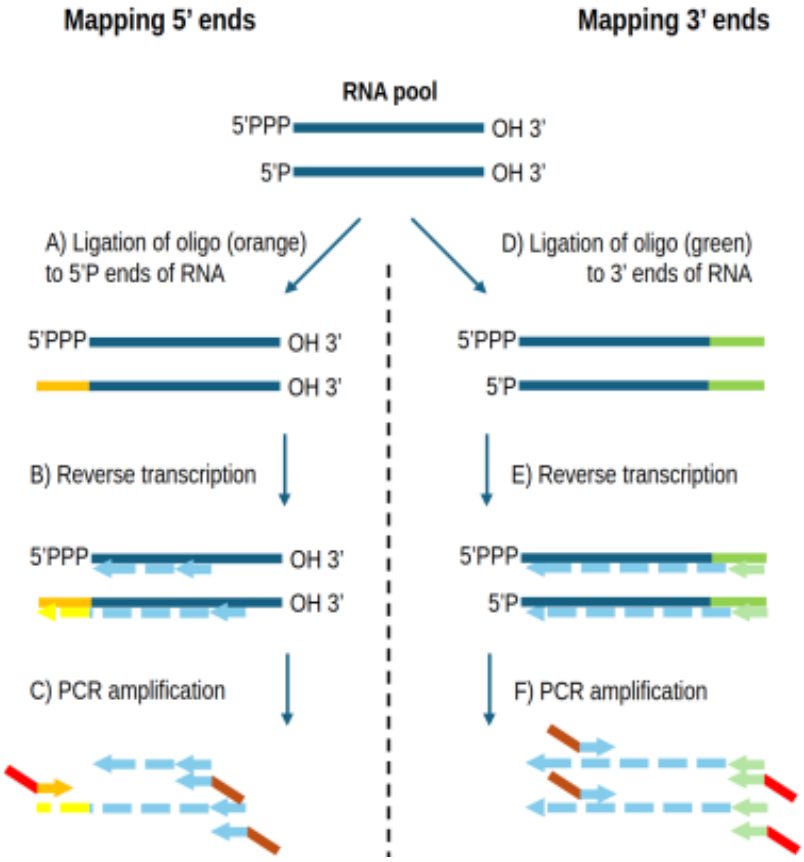


Figure 47 : ***Étapes clés des protocoles de préparation des librairies RNA-Seq pour déterminer les extrémités exactes des ARN.*** *Les détails varient en fonction du protocole utilisé. (**A**) Ligation d'un oligonucléotide de séquence connue (orange) aux extrémités 5'P des ARN. Les ARN 5'PPP ne sont pas acceptés comme substrat pour la ligase. (**B**) Transcription inverse pour générer de l'ADNc (bleu clair). Pour les ARN ligaturés, l'ADNc se terminera par le complément inverse de la séquence oligonucléotide (jaune). (**C**) PCR pour amplifier l'ADNc et ajouter des séquences adaptatrices pour le séquençage NGS (rouge et marron). Seul l'ADNc des ARN ligaturés sera amplifié. Le séquençage est effectué à partir de la séquence adaptatrice rouge.*

*Séquençage des extrémités 3' (panneaux D, E et F) : (**D**) Ligation d'un oligonucléotide de séquence connue (vert) aux extrémités 3'OH de l'ARN. € Rétro-transcription pour générer de l'ADNc (bleu clair) qui commencera par le complément inverse de la séquence de l'oligonucléotide (vert clair). Seuls les ARN ligaturés feront l'objet d'une transcription inverse. (**F**) PCR pour amplifier l'ADNc et ajouter des séquences adaptatrices pour le séquençage NGS (rouge et marron). Seul l'ADNc des ARN ligaturés sera amplifié. Le séquençage est effectué à partir de la séquence adaptatrice rouge.*

1.2.2.5.5.1 Exact Mapping Of Transcriptome Ends (EMOTE)

Chez les procaryotes, comme les ARN n'ont pas de coiffe en 5', la stratégie repose sur l'état de phosphorylation de l'extrémité 5' de l'ARN. Le protocole EMOTE (Redder, 2015) a été développé avec l'objectif de séquencer les extrémités 5'P des ARN. Dans un premier temps, les ARN totaux sont récupérés et purifiés. Les ARN ribosomiques sont ensuite éliminés. La deuxième étape consiste à lier spécifiquement un oligonucléotide (RpX) aux extrémités 5'P des ARN (Fig. 48), la ligase utilisée (T4 RNA ligase) n'acceptant pas les ARN 5'PPP.

Ces oligonucléotides peuvent contenir différents éléments tels que des Identifiants Uniques de Molécules (Unique Molecular Identifier – UMI) servant à identifier les duplicats PCR, ou des code-barre qui sont des index de multiplexage par exemple (Fig. 48).

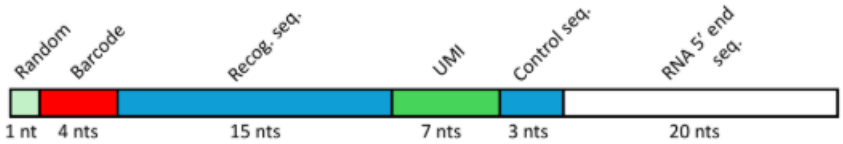


Figure 48 : ***Lecture EMOTE.*** *Exemple de structure d'une lecture de séquençage EMOTE de 50 nucléotides avec ses caractéristiques attendues : 1 nucléotide aléatoire, puis les 4 nucléotides du code-barre (parmi une liste prédéfinie) identifiant les échantillons multiplexés, puis la séquence de reconnaissance qui sert de site de liaison pour la RT, puis un UMI (7 nucléotides aléatoires), puis 3 nucléotides de la séquence de contrôle, et enfin l'extrémité 5' in vivo de l'ARN (20 nucléotides) qui peut être mappée sur la séquence de référence.*

Les ARN sont ensuite rétro-transcrits en ADNc qui sont par la suite amplifiés par PCR. La PCR est initiée à l'aide d'amorces complémentaires à une partie de l'oligonucléotide afin de n'amplifier que les ADNc issus d'ARN lié à l'oligonucléotide, donc 5'P. Les produits PCR sont finalement séquencés (Fig. 49).

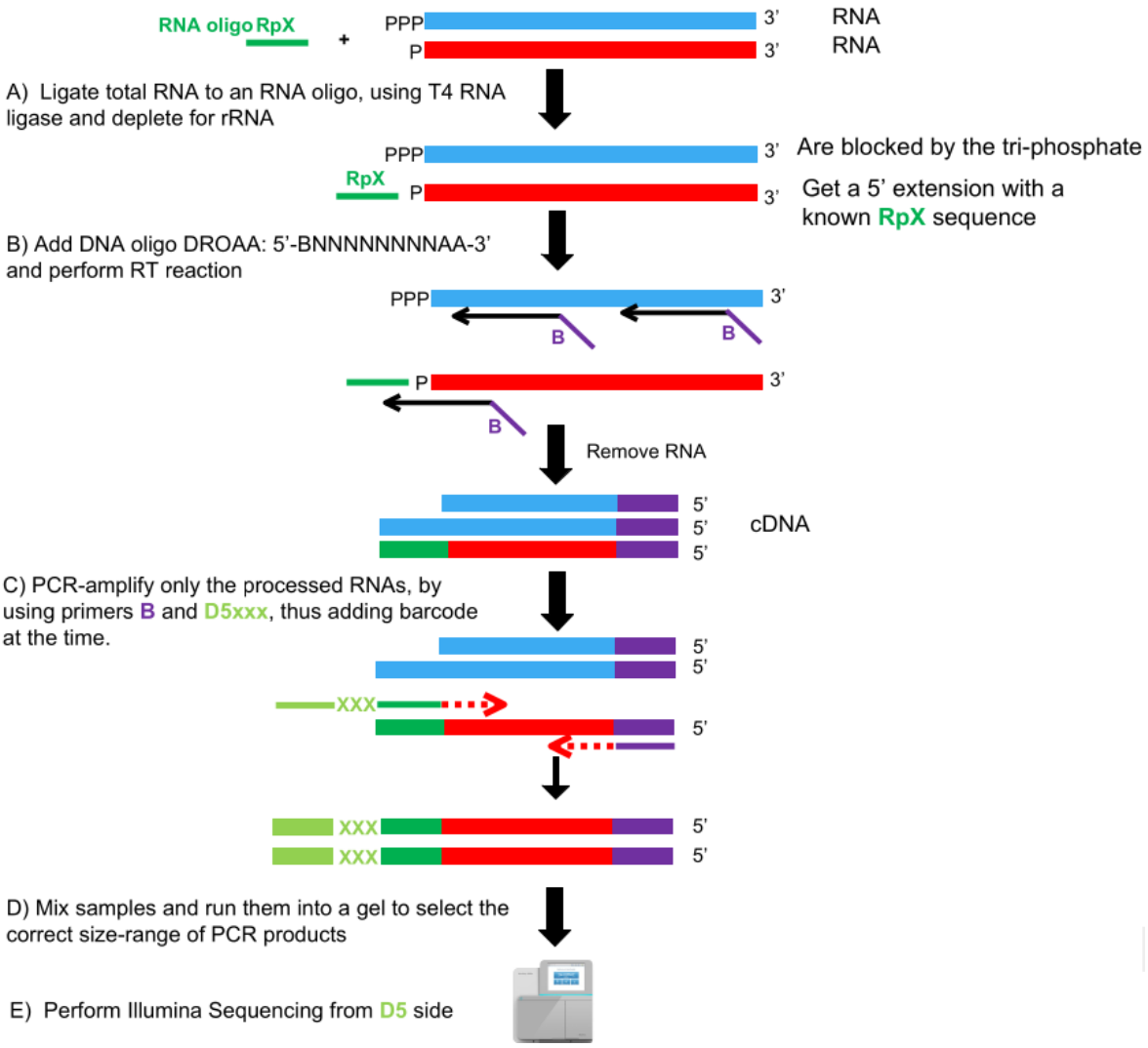


Figure 49 : ***Protocole EMOTE.*** *Les ARN totaux sont purifiés, et les extrémités 5'P (trait rouge) sont lié avec un oligonucléotide RpX (trait vert). Les ARN sont rétro-transcrit en ADNc puis les ARN restant sont dégradés. L'oligonucléotide contient un tag servant de reconnaissance pour les amorces PCR, ceci permet de n'amplifier que les ADNc dont l'ARN initial était lié à l'oligonucléotide (les 5'PPP ne seront donc pas amplifiés). Les produits PCR sont ensuite filtrer par taille puis séquencés.*

Dans cette section, nous avons choisi de présenter le protocole EMOTE car il a été développé par l'équipe avec laquelle nous collaborons, et une grande partie des données qui seront présentées dans ce manuscrit proviennent de l'utilisation de ce protocole.

1.2.2.5.5.2 TSS-EMOTE

Le séquençage des extrémités de l'ARN, et plus particulièrement, de l'extrémité 5' des ARN a été initialement motivé pour remédier à l'incapacité des protocoles RNA-Seq standard à déterminer avec précision l'extrémité 5' des molécules d'ARN *in vivo* et à leur faible performance dans l'identification et la quantification des sites d'initiation de la transcription (TSS). Etant donné que chez les bactéries, seuls les ARN dont les extrémités 5'PPP peuvent correspondre à des extrémités issues d'un début de transcription, Prados et al., (2016) ont développé la méthode PQ-EMOTE. Cette dernière est composée d'un protocole expérimental (Fig. 50 panneau gauche A à G) dérivé du protocole EMOTE et une méthode bio-informatique (Fig. 50 panneau droite H à N*)* permettant l'identification des TSS par une série de traitements et d'analyse *in silico*. Les différences entre ces protocoles résident, dans un premier temps, en une étape de dégradation des ARN 5'P par la RNase XRN1 afin de ne garder que les ARN 5'PPP pouvant correspondre à des TSS. Les ARN restants sont ensuite séparés en 2 échantillons, un échantillon (+RppH) qui sera traité avec la RppH de *E. coli* afin de déphosphoryler les 5'PPP en 5'P et ainsi les transformer en substrat pour la T4 RNA ligase et un échantillon non traité à la RppH (-RppH) et dans lequel aucune ligation n'est attendue ; cet échantillon sert de contrôle.

L'étape de traitement bio-informatique se fait à partir des fichiers FASTQ contenant les lectures issues du séquençage (Fig. 50H). Le logiciel EMOTE-conv (Yasrebi & Redder, 2015) est utilisé afin d'analyser les lectures, de les aligner sur le génome et d'effectuer la quantification (le principe de ces étapes sera décrit dans une section ultérieure) (Fig. 50I). Une table de comptage décrivant les positions génomiques et le brin sur lequel les lectures se sont alignées ainsi que le *pool* dont elles sont issues, et ainsi obtenue. Finalement, une position génomique donnée sera prédite comme un TSS si le nombre de lectures du pool "+RppH" alignées à cette position est statistiquement plus élevé que le nombre de lectures du pool "-RppH" en appliquant un test bêta-binomial tel que décrit dans (Prados et al., 2016) (Fig. 50J). Les p-valeurs obtenues de tous les réplicats biologiques sont ensuite combinées en une seule p-valeur (Fig. 50K). Seules les positions avec une p-valeur ajustée inférieure à 0,01 et au moins 5 lectures dans le pool +RppH sont considérées comme des TSS (Fig. 50L et 50M).

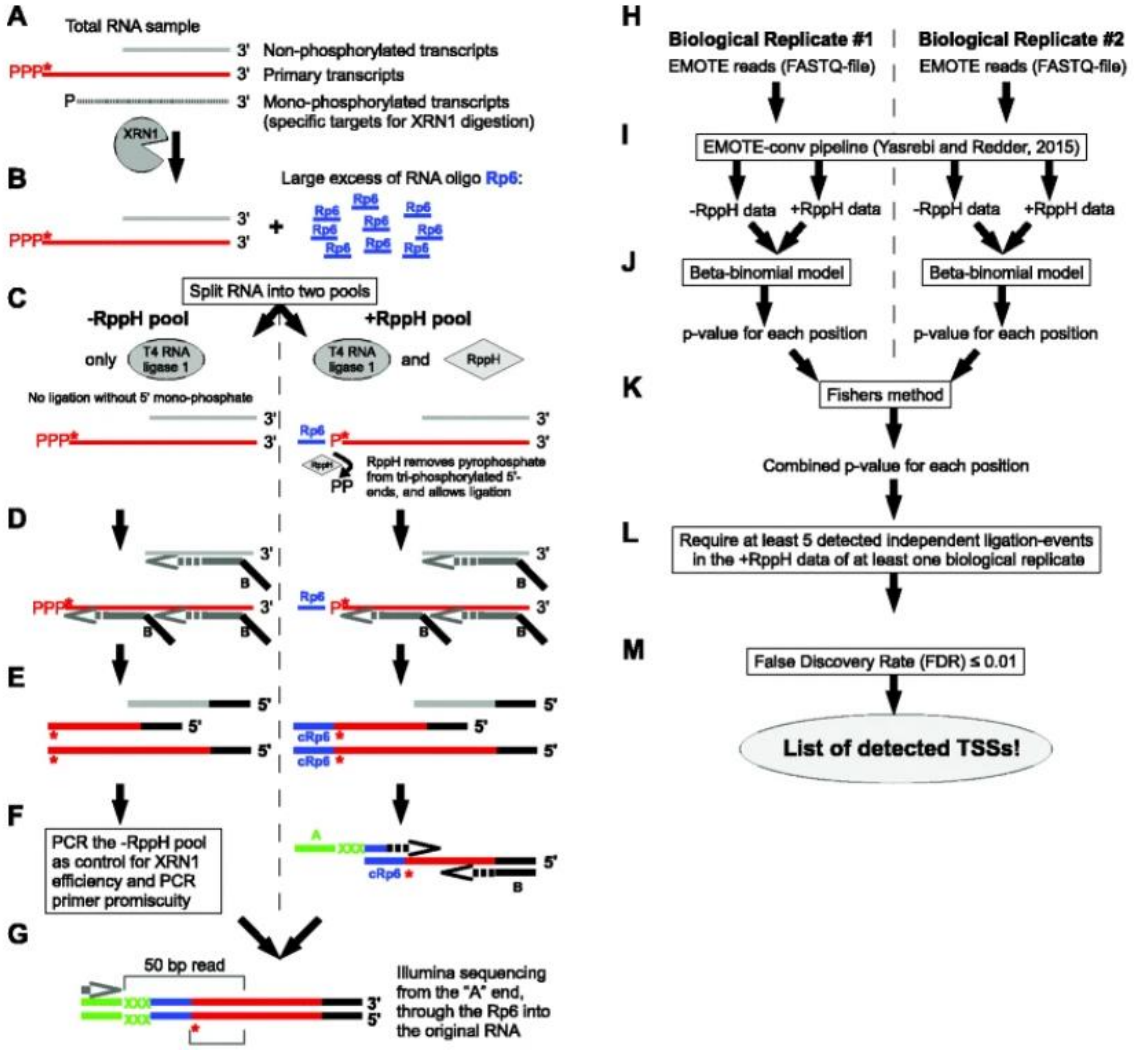


Figure 50 : Méthode TSS-EMOTE.

(A) *Digestion des ARN 5'P par la ribonucléase XRN1. Un astérisque marque continuellement l'extrémité 5' in vivo de l'ARN 5'PPP (ligne rouge fine). (**B**) Ajout de l'oligonucléotide Rp6 (bleu). (**C**) Division en deux pools dans lesquels la T4 RNA ligase 1 est ajoutée. Le pool +RppH est en plus traité avec RppH. (**D**) Les ARN sont rétro-transcrits et un adaptateur Illumina est ajouté. (**E**) Les ADNc issus d'ARN liés avec Rp6 contiennent une séquence complémentaire à Rp6 en 5' et un adaptateur Illumina en 3'. (**F**) Les ADNc sont amplifiés par PCR à l'aide d'amorces dirigées, une contre la partie complémentaire de Rp6, et une contre l'adaptateur Illumina. (**G**) Les produits PCR sont séquencés. (**H**) Les fichiers FASTQ servent de points de départ de l'analyse in silico. (**I**) Les fichiers FASTQ sont traités avec le logiciel EMOTE-conv* (Yasrebi & Redder, 2015). *(**J**) Un test beta-binomiale est effectué sur les positions génomiques sur lesquelles des lectures se sont alignées. (**K**) Les p-valeurs obtenues sont combinées par réplicats avec la méthode de Fisher. (**L**) Les positions sont filtrées pour ne garder que celles avec plus de 5 lectures alignées dans l'échantillon +RppH. (**M**) Les p-valeurs sont ajustées (False discovery rate) et les positions avec une p-valeur ajustée inférieure à 0,01 sont caractérisées comme des TSS. Extrait de* (Prados et al., 2016)***.***

1.2.2.6 Analyse de données RNA-Seq

Bien qu'il y ait une grande diversité de protocoles RNA-Seq et de multiples applications, les données résultantes sont toujours un fichier contenant des lectures dont le traitement englobe plusieurs étapes communes : le prétraitement des données, l'alignement des lectures sur un génome de référence, la quantification et le traitement des duplicats PCR, des étapes de normalisation et de filtrage (Fig. 51). Ce n'est qu'une fois que l'ensemble de ces étapes ont été réalisées, que des analyses statistiques peuvent être envisagées et qui dépendront de la nature de la question biologique posée. Nous allons, dans ce chapitre, décrire plus en détails ces étapes communes de traitement des données brutes.

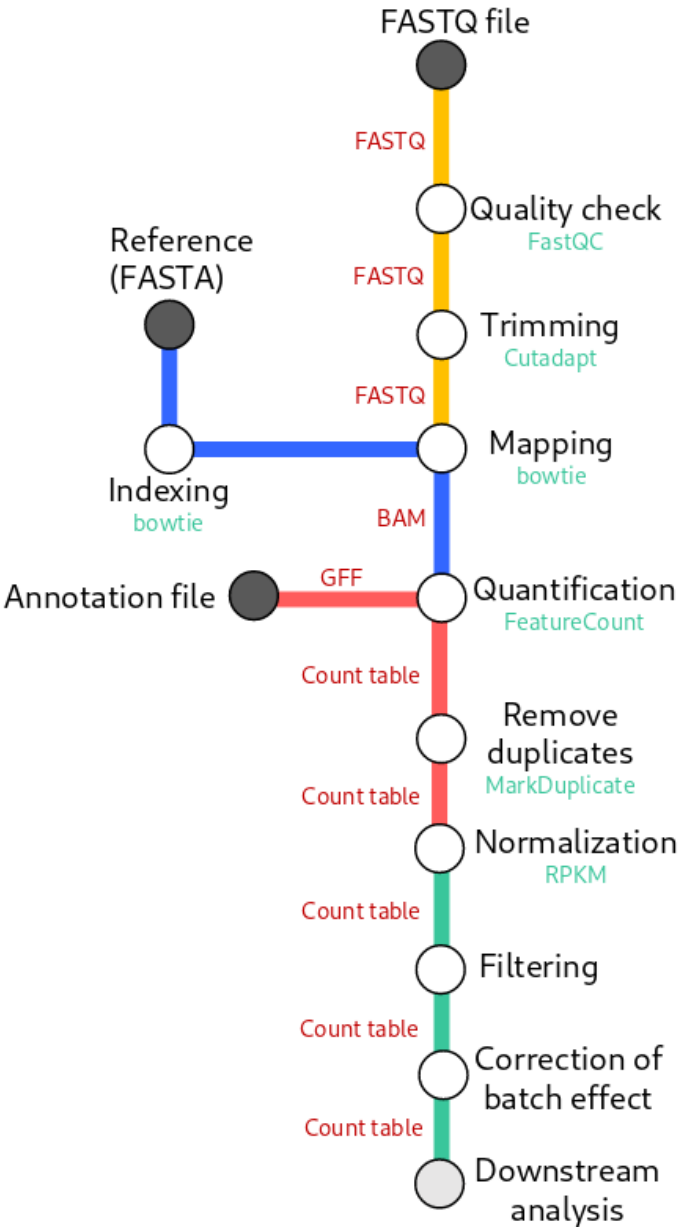


Figure 51 : Étapes courantes d'une analyse de données RNA-Seq.

Les points gris correspondent à des fichiers d'entrée et les points blancs à des étapes d'analyse. Les étapes sont regroupées par couleurs, jaune pour les étapes de pré-traitement, bleu pour l'alignement de lectures, rouge pour la quantification et vert pour le post-traitement. Le texte rouge correspondant au type de données et le texte en vert correspond à un exemple d'outils utilisé pour l'étape en question.

1.2.2.6.1 Pré-traitement

La première étape consiste à évaluer la qualité des données brutes issue du séquençage. Cela peut inclure l'examen de la qualité des bases, la détection d'adaptateurs ou de séquences de mauvaise qualité, ainsi que l'élimination des bases de mauvaise qualité à partir des extrémités des lectures.

Les données brutes de séquençage sont souvent fournies sous la forme de fichiers au format FASTQ, où chaque séquence est accompagnée d'une qualité de base associée à chaque nucléotide. Des outils tels que FastQC (*Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*, s. d.) sont utilisés pour évaluer la qualité en examinant les scores de qualité pour chaque position dans les lectures, ainsi que la contamination des lectures par la présence des adaptateurs d'index utilisés lors de la préparation des librairies. Les lectures de mauvaise qualité peuvent ensuite être filtrées à l'aide de critères tels que la qualité des lectures, ou la présence d'adaptateurs. L'étape de nettoyage des adaptateurs fait partie du contrôle de la qualité des données RNA-Seq. Cette étape vise à retirer les séquences d'adaptateurs qui peuvent être présentes à l'extrémité des lectures. Des outils tels que Cutadapt (Martin, 2011) et Trimmomatic (Bolger et al., 2014) sont couramment utilisés pour cette tâche. Ils permettent de détecter et de retirer les séquences d'adaptateurs ainsi que d'autres séquences que l'utilisateur peut juger comme indésirables des lectures.

1.2.2.6.2 Alignement des lectures

Après le séquençage, on ne connaît pas l'origine d'une lecture, c'est-à-dire, de quel ARN elle est issue mais seulement sa séquence. Pour le déterminer, il est nécessaire d'aligner ces lectures sur une séquence de référence en cherchant la position où sa séquence ou une sous-séquence correspondante se trouve. Cette position révèle l'origine de la lecture dans le génome de référence.

Avant l'alignement, le génome de référence (ou tout autre ensemble de séquences de référence) peut être indexé pour permettre une recherche rapide des séquences d'ADN. Cela implique de pré-calculer des structures de données pour stocker des informations sur le génome. Cela permet d'accélérer le processus d'alignement en trouvant toutes les correspondances exactes d'une lecture à l'aide d'une seule recherche dans la structure de données, sans avoir à parcourir l'ensemble du génome de référence du début à la fin pour chaque lecture.

Il existe plusieurs approches pour indexer un génome de référence dont les principales sont basées sur les arbres de suffixes (Weiner, 1973), sur la transformée de Burrows-Wheeler (BWT) (Burrows et al., 1994) ou sur les tables de hachage. Certaines méthodes combinent également différentes approches toujours dans l'optique d'augmenter la vitesse d'alignement et de diminuer la taille de l'index. Une fois que le génome de référence est indexé, les lectures pré-traitées sont alignées sur une séquence de référence à l'aide de programmes d'alignement de lectures (appelés « aligneurs » par la suite). L'objectif est de trouver la meilleure correspondance entre chaque lecture et une région de la référence, en tenant compte des variations génétiques telles que les substitutions, les insertions et les délétions.

Dans le cas idéal, chaque lecture correspond exactement à une seule région de la séquence de référence, permettant un alignement sans ambiguïté. Cependant, il peut arriver qu'en raison de diverses variations génétiques comme les polymorphismes nucléotidiques simples (SNP), les insertions et les délétions (Indels), souvent résultant de l'étape d'épissage chez les eucaryotes, peuvent compliquer l'alignement. De plus, les modifications post-transcriptionnelles ajoutent un niveau supplémentaire de complexité puisqu'elles peuvent apporter des modifications de la séquence des ARN qui ne sont pas visible sur la séquence de référence. Les lectures peuvent également provenir de régions répétées dans le génome, rendant leur attribution à une position unique difficile. En outre, la longueur des lectures peut varier considérablement, avec des lectures courtes posant des défis pour l'alignement en raison de leur moindre spécificité, tandis que les lectures longues, bien qu'offrant plus de contexte, peuvent inclure plus d'erreurs. Pour faire face à ces défis, des nombreux outils ont été développés au cours des dernières années (Fig. 52).

Les programmes d'alignement sont conçus pour trouver un équilibre entre la précision (alignements corrects) et la performance (rapidité et efficacité). La précision, ou exactitude des alignements, fait référence à la capacité d'un programme à aligner correctement les lectures RNA-Seq sur le génome de référence. La performance se réfère à la rapidité avec laquelle un programme peut traiter et aligner les données RNA-Seq. Certains programmes sont optimisés pour une grande précision, mais au détriment de la vitesse, tandis que d'autres privilégient la rapidité, notamment pour traiter de grands ensembles de données (bowtie ou bwa-mem). La variété des outils d'alignement RNA-Seq disponibles est donc en grande partie due à la nécessité de répondre à des besoins différents en termes de précision et de performance. Les chercheurs choisissent les outils en fonction de la taille et de la complexité de leurs jeux de données, des ressources informatiques disponibles et des exigences spécifiques de leur analyse, ce qui explique pourquoi il existe tant de programmes pour aligner les lectures RNA-Seq.

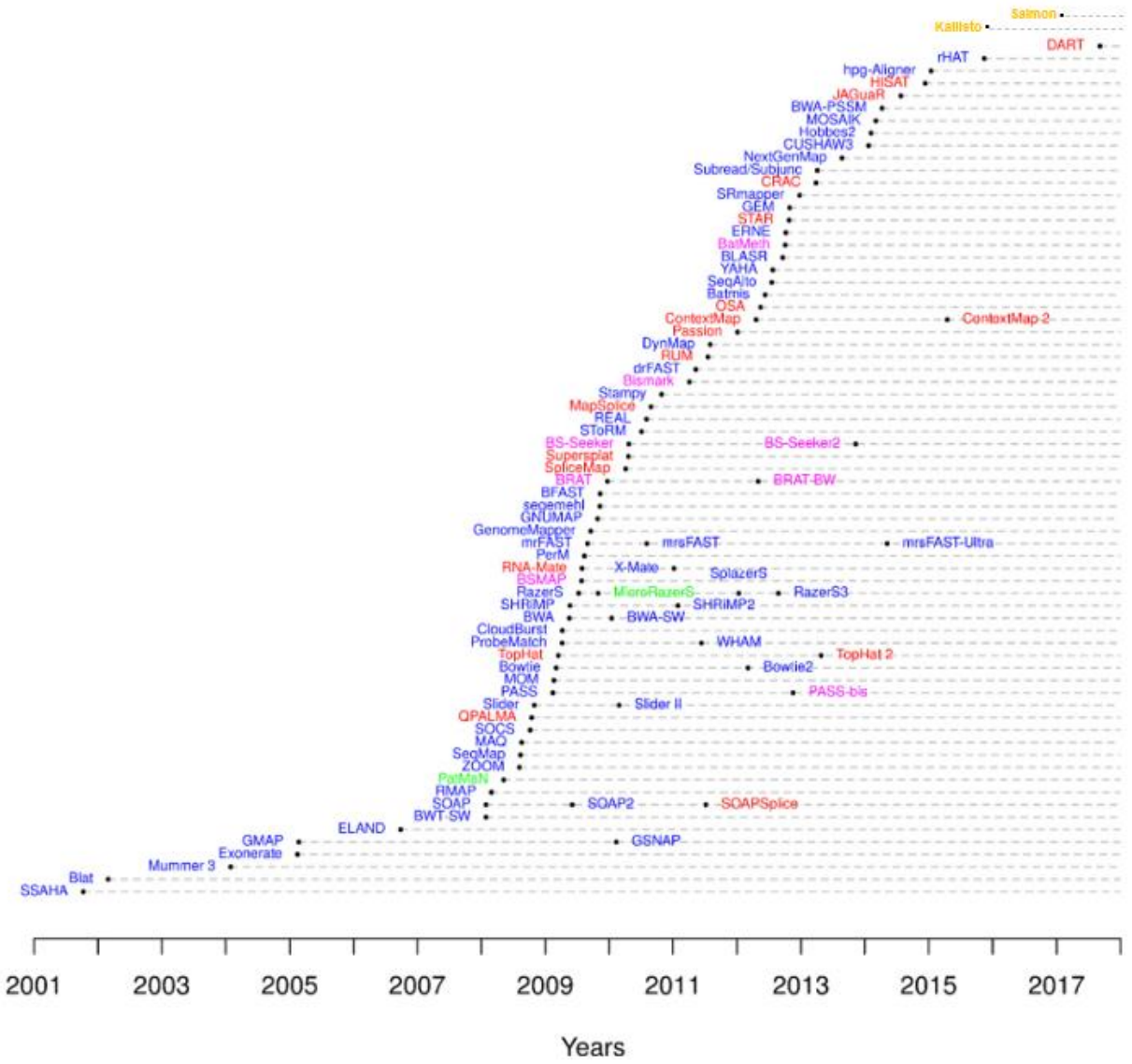


Figure 52 : ***Chronologie des aligneurs (2001-2017).*** *Les aligneurs d'ADN/ARN sont représentés en bleu, les aligneurs tenant compte de l'épissage en rouge, les aligneurs de miARN en vert, les aligneurs de bisulfite en violet et les quasi-aligneur en jaune. Les lignes grises en pointillé relient les aligneurs apparentés (extensions ou nouvelles versions). La chronologie n'inclut que les aligneurs ayant fait l'objet de publications évaluées par des pairs, et la date correspond à la première date de publication. (Adapté du site web de l'EBI)*

De plus, chez les eucaryotes, lorsque les lectures proviennent de transcrits matures, elles correspondent souvent à des séquences d'ARNm où les introns ont été excisés. Cela signifie que des lectures peuvent contenir des jonctions exon-exon, c'est-à-dire des séquences qui, dans le génome, ne sont pas contiguës mais sont directement adjacentes dans l'ARNm mature. Cela pose un défi car l'alignement naïf pourrait échouer à placer correctement ces lectures. Pour surmonter ces défis, des outils d'alignement RNA-Seq spécialisés, comme STAR (Dobin et al., 2013), TopHat2 (Kim et al., 2013), et HISAT2 (Kim et al., 2019), sont utilisés. Ces outils sont conçus pour détecter et aligner les jonctions exon-exon.

L'alignement génomique des petites séquences d'ARN (smRNA) telles que les microARN, peut également poser des défis considérables en raison de leur courte longueur (~21 nucléotides) ainsi que de la grande taille et de la complexité des génomes des plantes et des animaux (Ziemann et al., 2016).

De la même manière, différentes techniques de séquençage peuvent présenter des défis uniques pour l'alignement des lectures, comme le séquençage au bisulfite (Bi-Seq) par exemple qui est utilisé pour étudier les sites de méthylation de l'ADN (Grunau et al., 2001). Les cytosines non méthylées sont converties en thymidines, tandis que les cytosines méthylées restent inchangées. Après la PCR, les séquences obtenues sont alignées sur un génome de référence en tenant compte de cette conversion, permettant d'identifier les positions méthylées par comparaison des écarts T/C. De plus, comme les modifications des brins d'ADN ne sont pas symétriques, chaque brin nécessite un alignement distinct, ce qui ajoute une complexité supplémentaire à l'analyse (Campagna et al., 2013).

Jusqu'à récemment, l’assignement des lectures impliquait la production d'alignements exacts des lectures sur une séquence de référence. Cependant, des outils comme Kallisto (Bray et al., 2016) et Salmon (Patro et al., 2017) ont introduit une nouvelle approche. Ces méthodes, qualifié de « quasi-aligneurs » ont changé la façon dont les lectures sont assignées à des caractéristiques génomiques. En effet, ils permettent d'assigner des lectures directement à des transcrits sans produire d'alignements détaillés. Contrairement aux méthodes traditionnelles qui alignent les lectures sur un génome complet, les quasi-aligneurs utilisent une séquence de référence correspondant à un transcriptome, c'est-à-dire l'ensemble des transcrits, ou ARN messagers, produits par un organisme. Cette méthode présente plusieurs avantages. Premièrement, elle est beaucoup plus rapide car elle évite les calculs intensifs nécessaires pour un alignement exact.

Deuxièmement, pour de nombreuses applications, la précision des alignements n'est pas nécessaire et seule l'assignation correcte des lectures est important pour la quantification des transcrits.

Ces quelques exemples non exhaustifs révèlent la complexité des problèmes qu'il peut y avoir lors de l'alignement précis de millions de lectures.

Les résultats de l'alignement (hors quasi-aligneur) sont généralement enregistrés dans un format standardisé, tel que le format SAM (Sequence Alignment/Map) (Fig. 53) ou son équivalent compressé BAM (Binary Alignment/Map). Ces fichiers contiennent des informations sur l'emplacement de chaque lecture alignée sur le génome, ainsi que des informations sur la qualité de l'alignement.

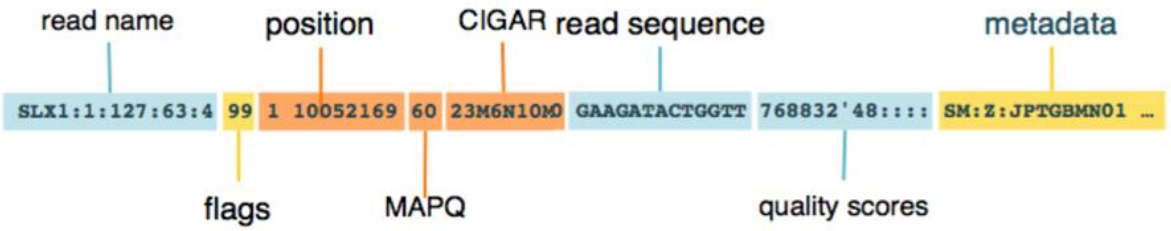


Figure 53 : Aperçu d'un fichier d'alignement au format SAM. *Un seul alignement est représenté. « Read name » correspondant à l'identifiant (déjà présent dans le fichier FASTQ) de la lecture. Le « Flag » est un indicateur binaire qui décrit l'état de l'alignement. « position » correspond aux positions dans la séquence de référence de début et de fin de l'alignement. « MAPQ » représente la qualité de l'alignement et « CIGAR » un indicateur décrivant l'alignement en indiquant les appariements, les insertions, les mésappariements, etc. Il y a ensuite la séquence de la lecture, sa qualité et des données additionnelle comme par exemple son emplacement sur la cellule d'écoulement. (Adapté du site web de* ***« GATK »)***

Le nombre total de lectures alignées sur le génome correspond à la taille de la librairie RNA-Seq. Il ne faut pas confondre ceci avec ce qu'on appelle la profondeur de séquençage qui correspond au nombre moyen de lectures par position alignée. Il est important de s'assurer une profondeur de séquençage élevée afin de maximiser la détection des transcrits qui sont présents en faible abondance. De plus, une profondeur de séquençage élevée permet de réduire le bruit de fond qui correspond aux lectures non spécifiques ou aléatoires qui ne représentent pas de véritables transcrits biologiques. Ce bruit de fond existe en raison d'imperfections techniques, telles que des erreurs d'amplification ou des séquences contaminantes.

1.2.2.6.3 Quantification et duplicats PCR

1.2.2.6.3.1 Intervalles génomiques

Dans une analyse RNA-Seq, l'étape de quantification vise à évaluer la quantité de lectures alignées sur des intervalles génomiques choisis par l'analyste. Traditionnellement, ces intervalles sont associés aux gènes, permettant ainsi d'estimer l'expression génique (Fig. 54 bleu). Toutefois, ces intervalles peuvent représenter d'autres caractéristiques génomiques telles que des opérons, des régions UTR, *etc.* La quantification peut également se faire selon un intervalle glissant de taille constante (Fig. 54 rouge) pour par exemple calculer une couverture génomique. Dans certains cas, la quantification peut même se faire à l'échelle d'une seule position génomique (Fig. 54 vert), où le nombre de lecture est compté par position exacte. L'objectif de la quantification consiste donc à mesurer le nombre de lectures alignées sur des intervalles génomiques choisis, correspondant à des caractéristiques génomiques ou pas, de taille constante ou variable, afin d'avoir une description relative de la quantité de lectures alignées.

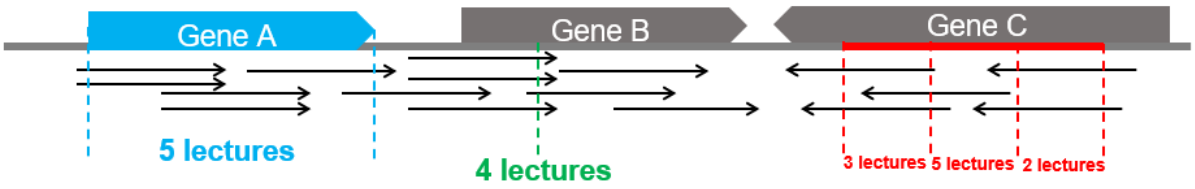


Figure 54 : Exemple d'intervalles génomiques quantifiés. *En bleu, une quantification sur un intervalle correspondant à un gène. En vert, une quantification par position exacte. En rouge, une quantification par intervalles régulier.*

Les intervalles à quantifier sont donc représentés par une position génomique de début, de fin et un brin.

1.2.2.6.3.2 Outils de quantification

Le principe de la quantification est de faire l'intersection entre deux jeux d'intervalles. Les intervalles correspondants aux alignements des lectures et ceux correspondant aux intervalles à quantifier (gènes ou autres). featuresCounts (Liao et al., 2014) et htseq-count (Anders et al., 2015)) sont des programmes, qui, à partir de fichiers BAM et de fichiers d'annotation dans lesquels sont décrits les intervalles correspondant à des caractéristiques génomiques, permettent de compter le nombre de lectures correspondant à des caractéristiques génomiques, dans les environnements de programmation respectifs R et python. Bamcoverage, est un autre programme appartenant à la suite deepTools (Ramírez et al., 2014) utilisé pour générer des profils de couverture, sans

nécessiter de fichier d'annotation. La couverture est calculée comme le nombre de lectures par « bin », où les bins sont de courts intervalles d'une taille définie par l'utilisateur. A noter que dans la plupart des cas, l'utilisateur peut choisir la manière dont les lectures chevauchantes entre deux intervalles ou à cheval sur le début ou la fin doivent être comptabilisées.

Les données de quantification sont généralement renvoyées sous forme de tableaux où chaque ligne représente un nombre de lectures alignées sur un intervalle génomique (une annotation). Il faut également garder en tête que pour des génomes de taille conséquente (milliards de paires de base), des formats synthétiques et/ou des fichiers compressés peuvent être privilégiés pour un stockage moins lourd.

1.2.2.6.3.3 Biais d'amplification PCR

Dans les protocoles de type RNA-Seq, l'étape d'amplification par PCR est généralement nécessaire, à la fois pour amplifier l'ADNc à un niveau suffisant pour le séquençage et pour enrichir les molécules correctement liées aux adaptateurs. Cependant, cette étape engendre des biais d'amplification, c'est-à-dire que certains fragments d'ADNc peuvent être amplifiés plus efficacement que d'autres en raison de leur composition en nucléotides. Par exemple, les régions riches en guanine-cytosine (GC) peuvent être amplifiées de manière moins efficace que les régions riches en adénine-thymine (AT) pendant la PCR, en raison des appariements plus ou moins stable entre les bases (Aird et al., 2011). Cela fausse la profondeur de séquençage des lectures et introduit des erreurs dans la quantification.

1.2.2.6.3.3.1 Méthodes de correction

1.2.2.6.3.3.1.1 Marquage et/ou suppression des lectures identiques

Une méthode pour éliminer ces biais d'amplification consiste à ne garder qu'une seule lecture parmi toutes celles qui sont identiques. Les outils comme MarkDuplicates (Picard Toolkit." 2019. Broad Institute) ou rmdup (SAMtools) (H. Li et al., 2009) sont souvent utilisés pour cette tâche (Ebbert et al., 2016). rmdup est une commande disponible dans SAMtools, un ensemble de programmes en ligne de commande, et permet de supprimer les duplicats PCR directement du fichier BAM. Son algorithme repose sur l'utilisation des coordonnées des alignements des séquences pour identifier les duplicats. Il examine les paires de séquences qui sont alignées aux mêmes positions sur le génome et conserve uniquement une copie de chaque paire, à condition qu'elles aient la même séquence, supprimant les duplicats subséquents. MarkDuplicates est inclus dans la suite logicielle Picard, développée par le Broad Institute. Son fonctionnement est similaire à rmdup, cependant,

une fois les duplicats identifiés, MarkDuplicates les marque dans le fichier BAM en ajoutant un tag d'identification à chaque séquence dupliquée et laisse à l'utilisateur le choix de les conserver ou de les supprimer en fonction de ses besoins. Il convient de noter que l'hypothèse selon laquelle les lectures identiques sont le résultat d'un biais de la PCR n'est que partiellement vrai car cela dépend grandement de la profondeur du séquençage.

1.2.2.6.3.3.1.2 Les identifiants moléculaires uniques (UMI : Unique Molecular Identifiers)

Pour suppléer au problème de la présence de duplicats dus à l'étape de PCR et donc identifier des molécules uniques dans des mélanges complexes, il a été proposé, au niveau expérimental, d'incorporer, dans les transcrits avant l'amplification PCR, des codes-barres constitués de séquences aléatoires de nucléotides afin de rendre chaque molécule unique (H. Hug & Schuler, 2003). Ces codes-barres sont désormais appelés identifiants moléculaires uniques (UMI) (Fig. 55). Grâce à l'incorporation d'UMI les ADNc individuels sont alors étiquetés de manière unique et peuvent être identifiés après le séquençage. Cela permet d'amplifier les librairies à un niveau suffisant pour détecter les transcrits de faible abondance, tout en conservant les informations d'expression moins biaisées, car les séquences identiques résultant de la préparation de la librairie peuvent être regroupées par séquence et par UMI pour être comptées comme une seule lecture (déduplication).

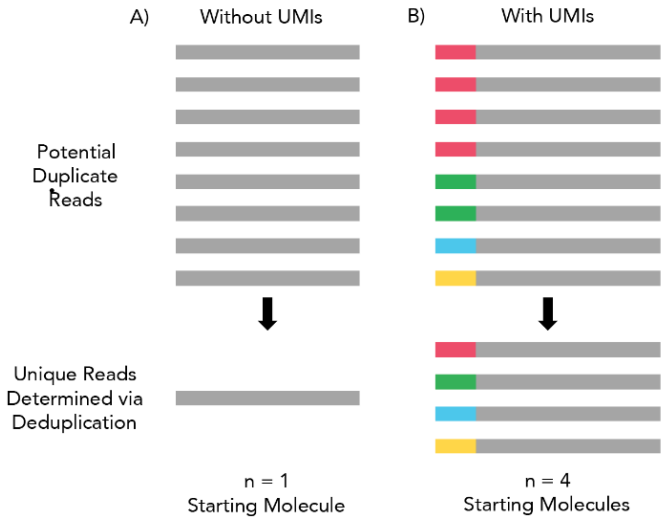


Figure 55 : Marquage des transcrits à l'aide d'identifiant unique de molécule UMI.

*Comparaison de l'élimination des duplicats de PCR sans et avec les UMI. (**A**) Sans UMI, les lectures avec la même séquence sont faussement réduites à une molécule de départ. (**B**) Les lectures de même séquence peuvent être regroupées par des UMI pour déduire correctement le nombre original de molécules de départ dans un échantillon.*

Les méthodes rmdup et MarkDuplicates proposent aujourd'hui des fonctionnalités permettant de tenir compte des UMI. Il existe également des outils conçus spécifiquement pour le traitement d'UMI tels que UMI-tools (Smith et al., 2017).

Cette étape de déduplication permet de réduire les erreurs de mesure pouvant résulter des différentes étapes expérimentales et donc d'améliorer la fiabilité de la quantification de l'expression génique, conduisant à une meilleure interprétation des données biologiques. Cependant, l'utilisation d'UMI n'est pas toujours primordiale, par exemple, pour des analyses qualitatives comme la détection de variant (variant calling) qui permet l'identification des variations génétiques (SNP, indels, *etc.*) pour lesquelles la quantification par la déduplication n'apporte que très peu du gain de précision (Ebbert et al., 2016).

1.2.2.6.3.3.1.2.1 Saturation des UMI

Dans les applications portant sur des librairies complexes, telles que le séquençage d'ARN totaux, seule une variété limitée d'UMI est nécessaire (~1024 séquences différentes pour des UMI de taille 5 nucléotides) (Saunders et al., 2020), car la variation des molécules à séquencer est énorme. Cependant, lorsqu'il s'agit de librairies moins complexes, par exemple pour le séquençage de sous-ensembles distincts du génome, tels que les microARN (Saunders et al., 2020)**,** (Fu et al., 2018) ou encore les extrémités 5' et 3' des ARN, il est nécessaire d'augmenter drastiquement le nombre de combinaisons en augmentant la complexité des UMI, par exemple en utilisant des UMI de taille plus importante, Saunders et al. proposent d'utiliser des UMI d'une taille supérieure à 8 pour les données de séquençage de microARN.

En effet, dans une préparation de librairie RNA-Seq classique, une étape de fragmentation aléatoire est effectuée. Cela signifie que même si plusieurs copies d'un même transcrit existent, les fragments résultants seront majoritairement uniques. Par conséquent, un nombre relativement faible de séquences d'UMI suffit pour les marquer de manière unique.

En revanche, dans des protocoles sans fragmentation, comme le séquençage des extrémités 5' de l'ARN (5'-end RNA-Seq), plusieurs copies du même transcrit conservent la même séquence. Cela nécessite donc une plus grande diversité de séquence d'UMI pour garantir que chaque copie puisse être marquée de manière unique. En effet, si les UMI utilisés sont trop courts et donc le nombre de séquences différentes d'UMI trop faible, un même UMI peut, de façon aléatoire, être ajouté à deux ou plusieurs molécules différentes d'un même transcrit (différentes molécules d'ARN issues de la transcription du même gène par exemple). Ces différentes molécules ayant des séquences identiques car correspondant à la transcription de la même région génomique, il devient donc

impossible de les différencier des duplicats PCR. On parle alors de la saturation des séquences d'UMI. Cela peut engendrer plusieurs problèmes dans l'analyse des données RNA-Seq. Lorsque la saturation se produit, cela entraîne une déduplication excessive des lectures, c'est-à-dire que des lectures qui ne sont pas des duplicats PCR seront supprimés, ce qui conduit à une sous-estimation significative des comptages, pouvant atteindre un facteur de 20 pour les transcrits les plus abondants (Saunders et al., 2020). Cette déduplication excessive biaise les résultats, rendant les estimations de l'abondance des transcrits peu fiables et faussant les conclusions biologiques. Il est donc crucial de prendre en compte ces limitations lors de l'interprétation des données RNA-Seq et de mettre en œuvre des stratégies appropriées pour atténuer les effets de la saturation des UMI. (Saunders et al., 2020) propose un modèle permettant de prévoir le nombre de combinaisons nécessaires pour une quantification exacte, ce qui aide à anticiper et à éviter la saturation des UMI (Formule 1). Toutefois, il n'existe pas, à notre connaissance, de solution *a posteriori* pour corriger cette saturation une fois qu'elle s'est produite. Pour éviter ces problèmes, il est essentiel d'adapter la complexité des UMI en fonction de la complexité de la librairie de séquençage. Pour les librairies moins complexes, augmenter la taille des UMI permet de minimiser les risques de saturation et d'assurer une identification unique des molécules.

$$\mathrm{P} = 1 - \Pr(\mathrm{UMInotbeingsampled}) - \Pr(\mathrm{UMIbeingsampledexactlyonce})$$

$$= 1 - \left(\frac{C-1}{C}\right)^{n} - n * \frac{1}{C} * \left(\frac{C-1}{C}\right)^{n-1}.$$

Formule 1 : Probabilité pour un UMI d'être utilisé deux fois ou plus. *Ou C représente au nombre de combinaisons d'UMI possible, n le nombre de lectures alignées à la même position et P la probabilité de sélectionner un UMI plus d'une fois. Ainsi, le nombre total d'UMI échantillonnées plus d'une fois compte tenu de ces paramètres est simplement égal à P*C. Extrait de* (Saunders et al., 2020)

1.2.2.6.4 Post-traitement (normalisation et filtrage)

1.2.2.6.4.1 Normalisation

Afin de tenir compte des différences de profondeur de séquençage entre les échantillons, ainsi que des facteurs techniques qui pourraient affecter l'expression génique mesurée, une étape de normalisation des comptages est nécessaire. L'objectif principal de la normalisation est de pouvoir comparer les niveaux d'expression entre les différents gènes, échantillons et jeux de données. Il existe trois niveaux principaux de normalisation de données RNA-Seq (Abbas-Aghababazadeh et al., 2018).

La normalisation au sein d'un même échantillon (within sample) est nécessaire pour comparer l'expression des gènes au sein d'un même échantillon (Y. Zhao et al., 2021). Elle permet d'ajuster les données en fonction de deux variables principales : la longueur des transcrits et la profondeur de séquençage. Les gènes plus longs ont souvent davantage de lectures alignées que les gènes plus courts pour un même niveau d'expression (Mortazavi et al., 2008). Par conséquent, leur niveau d'expression ne peut être comparé qu'après normalisation. En outre, le nombre de lectures de séquençage par échantillon peut varier. Cela peut également être corrigé par ce type de normalisation. Les normalisations au sein de l'échantillon les plus utilisées sont les normalisations RPKM/FPKM et TPM.

Les RPKM (lectures par kilobase de transcrit par million de lectures alignées) pour les données à extrémités uniques (single-end) et les FPKM (fragments par kilobase de transcrit par million de fragments alignés) pour les données à extrémités appariées (paired-end), permettent de corriger les différences de taille de librairie et de longueur des gènes (Mortazavi et al., 2008).

$$RPKM = 10^9 * \frac{Nombre\ de\ lectures\ alignées}{Nombres\ total\ de\ lectures\ alignées * Longueur\ du\ transcrit}$$

Formule 2 : ***Calcul de la valeur RPKM pour un transcrit***

Cependant, la présence de quelques gènes très exprimés peut influencer de manière disproportionnée la normalisation des autres gènes, rendant les comparaisons biaisées entre échantillons (G. P. Wagner et al., 2012). La normalisation RPKM/FPKM permet donc de mieux comparer l'expression des gènes au sein d'un seul échantillon. La normalisation TPM (Transcript Per Million) (B. Li & Dewey, 2011) est une alternative qui permet de mettre à l'échelle les unités RPKM pour que le total soit égale à 1 million.

La valeur TPM d'un transcrit correspond au nombre de lectures qu'on aurait eu si la taille de la librairie était de 1 million. Cela permet de comparer les valeurs d'expression entre échantillon contrairement au RPKM (S. Zhao et al., 2020). Les TPM sont calculés en divisant le nombre de lectures alignées pour un transcrit par sa longueur. Cette valeur est ensuite divisée par la somme des rapports de ce même calcul pour chaque transcrit de la librairie (RPKM). Elle est ensuite multipliée par un million pour faciliter les analyses ultérieures (G. P. Wagner et al., 2012), (Y. Zhao et al., 2021).

Les TPM permettent donc de normaliser les données RNA-seq en fonction de la profondeur de séquençage et de la longueur des transcrits.

$$TPM = \frac{Nombre\ de\ lectures\ alignées\ /\ Longueur\ du\ transcrit}{\sum(Nombre\ de\ lectures\ alignées\ /\ Longueur\ du\ transcrit)} * 10^6 = \frac{RPKM}{\sum RPKM} * 10^6$$

Formule 3 : Calcul de la valeur RPKM pour un transcrit

Bien que les TPM corrigent certaines des limitations des RPKM en maintenant la somme des TPM constante (1 million) dans chaque échantillon, ils peuvent encore être inadéquats si les échantillons présentent des variations d'expression importantes (conditions, souches différentes par exemple). Les TPM, tout comme les RPKM, ne prennent pas en compte les variations dans les distributions des niveaux d'expression entre les échantillons. (S. Zhao et al., 2020) a démontré que ces deux indices ne devraient jamais être utilisés pour des comparaisons quantitatives entre échantillons lorsque les teneurs en ARN total et leurs distributions sont très différentes. Par conséquent, des méthodes de normalisation RNA-Seq entre échantillons sont nécessaires pour obtenir des comparaisons appropriées des niveaux d'expression des gènes.

La normalisation entre échantillons (between samples) : Les échantillons d'un ensemble de données peuvent être normalisés simultanément comme un ensemble complet pour tenir compte des différentes variations techniques telles que la profondeur de séquençage (Bolstad et al., 2003). Le RNA-Seq est une mesure relative, et non absolue, de l'abondance des transcrits. Cela signifie que la population de transcrits dans son ensemble affecte les niveaux relatifs des transcrits (Robinson & Oshlack, 2010). Cela crée des biais pour les analyses d'expression génique, qui sont minimisés par des méthodes de normalisation entre les échantillons. Les méthodes de normalisation entre échantillons les plus utilisées sont les méthodes quantile, TMM et ARN de contrôle.

La méthode du quantile vise à rendre la distribution des niveaux d'expression des gènes identique pour chaque échantillon d'un ensemble de données (Bolstad et al., 2003). Elle suppose que les différences globales de distribution entre les échantillons sont toutes dues à des variations techniques. Les différences restantes sont probablement des effets biologiques réels. La procédure de normalisation quantile (Fig. 56) consiste d'abord à classer les gènes de chaque échantillon par ordre d'abondance puis à calculer la valeur moyenne des gènes occupant le même rang. Les valeurs de tous les gènes sont remplacées par la valeur moyenne de leur rang correspondant. L'étape suivante consiste à réorganiser les gènes de chaque échantillon dans leur ordre initial.

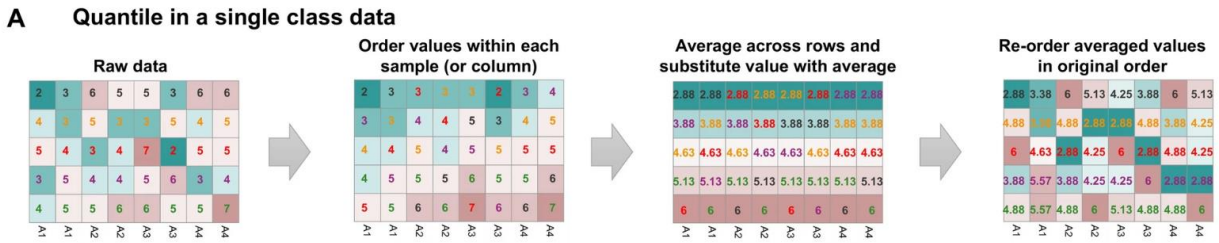


Figure 56 : Exemple d'une procédure de normalisation quantile. *Les gènes sont d'abord classés par ordre d'abondance. Le niveau d'abondance de chaque gène est remplacé par la valeur moyenne des abondances des gène de même rang. Les gènes sont finalement remis dans l'ordre initial.*

La méthode TMM (Trimmed Mean of M-values) repose sur l'hypothèse que la majorité des gènes ne sont pas différentiellement exprimés (Robinson & Oshlack, 2010)**.** Pour chaque échantillon, le facteur TMM est calculé en prenant un échantillon de référence et en comparant les autres échantillons (appelé « test ») à celui-ci. Pour chaque échantillon test, après avoir exclu les gènes les plus exprimés et ceux présentant les ratios logarithmiques les plus élevés (log(test/ref)), le facteur TMM est déterminé comme la moyenne pondérée (par l'inverse de la variance) des ratios logarithmiques entre cet échantillon test et l'échantillon de référence. En raison de l'hypothèse de faible expression différentielle, le facteur TMM devrait être proche de 1. Si ce n'est pas le cas, sa valeur indique le facteur de normalisation à appliquer à la taille de la libraire. Cette normalisation est très utilisée notamment pour les analyses différentielles d'expression de gènes où la taille des gènes n'a pas besoin d'être prise en compte.

Une autre approche consiste à utiliser un ensemble d'ARN de contrôle afin de normaliser les comptages. La première étape pour appliquer cette méthode est donc de sélectionner un ensemble d'ARN de contrôle approprié. Il s'agit en généralement d'ARN stables et exprimés de manière constitutive dans tous les échantillons comme, par exemple, les ARN correspondant aux gènes dits de ménage (housekeeping genes) qui sont couramment utilisés dans les études transcriptomiques. Pour chaque échantillon, un facteur de normalisation est calculé à partir des comptages des ARN de contrôles. Par exemple, le facteur de normalisation peut être calculé en divisant les comptages des ARN de contrôle par la médiane de ces comptages dans tous les échantillons, fournissant ainsi un facteur de normalisation relatif à la quantité d'ARN dans chaque échantillon. La sélection des ARN de contrôle appropriés, c'est-à-dire stables et exprimés de manière constante dans toutes les conditions expérimentales est donc une étape clé. Cependant, même dans ce cas, il est important de considérer les limites potentielles de cette méthode, telles que la stabilité et la variabilité naturelle des ARN de contrôle dans différentes conditions expérimentales. C'est pourquoi il est possible d'utiliser un ARN exogène, appelé spike-in, comme ARN de contrôle. Cet ARN exogène n'est pas

influencé par la variabilité biologique, ce qui en fait une référence externe idéale pour la normalisation. Néanmoins, il est essentiel de garantir une répartition égale du spike-in entre les échantillons, ce qui peut poser des défis techniques importants. A noter que ces méthodes ne prennent pas en compte la longueur des transcrits. Si l'objectif est de comparer l'abondance entre échantillons mais également entre les transcrits, il faudra combiner plusieurs niveaux de normalisation.

La normalisation entre jeux de données (across librairies) : Les chercheurs intègrent souvent des données RNA-Seq provenant de plusieurs études indépendantes. Ces ensembles de données sont généralement séquencés à des moments différents, avec des méthodes variables dans plusieurs installations, et parfois par différents manipulateurs (Leek et al., 2010). Il en résulte des effets de lot (ou effet batch) qui induisent des différences entre les échantillons qui ne sont pas liées à la conception de l'expérience. Dans le contexte des données de séquençage, deux séquençages à des moments différents peuvent déjà présenter un effet de lot (Sprang et al., 2022)**.** Bien que ces effets puissent être minimisés par des pratiques expérimentales plus rigoureuses et un plan d'expérience adapté (Grass, 2009), des effets de lot peuvent toujours survenir et il peut être difficile de les corriger. Les effets de lot sont connus pour interférer avec les analyses statistiques en aval, par exemple en introduisant des gènes différentiellement exprimés entre les groupes, qui ne sont détectés qu'entre les lots et qui n'ont aucune signification biologique.

De nombreuses méthodes ont été proposées pour traiter les effets de lot dans les études RNA-seq. Par exemple, Limma (removeBatchEffect) et ComBat (Johnson et al., 2007) restent les méthodes d'ajustement des effets de lot les plus populaires lorsque les effets proviennent de sources connues. Pour l'hétérogénéité provenant de sources inconnues, SVASeq (Leek & Storey, 2007) et RUVSeq (Risso et al., 2014) sont couramment utilisées. Des méthodes telles que Limma (Smyth, 2004) et ComBat éliminent les effets de lot lorsque les sources de variation, telles que la date de séquençage, sont connues (Ritchie et al., 2015). La normalisation au sein du jeu de données doit d'abord être appliquée avant la correction des effets de lots afin que les valeurs d'expression des gènes soient à la même échelle entre les échantillons. Limma et ComBat utilisent tous deux les statistiques empiriques de Bayes pour estimer les distributions de probabilité préalables à partir des données. Par exemple, pour chaque gène, Combat estime les effets de lot en utilisant un modèle linéaire. Les effets de lot sont soustraits des valeurs d'expression génique et les données sont ensuite ajustées pour conserver les variations biologiques.

Après correction des effets de lot connus, l'analyse des variables de substitution (SVA) peut être utilisée pour identifier et estimer les sources inconnues de variation (Leek et al., 2012). La méthode SVASeq de Bioconductor (Leek et al., 2012) peut détecter et corriger les effets de plusieurs sources de variations indésirables, y compris les lots. Et RUVSeq (Remove Unwanted Variables) utilise des contrôles négatifs (gènes ou échantillons connus pour ne pas être affectés par la condition biologique) pour modéliser et soustraire les effets non biologiques.

Cependant, ces méthodes doivent être utilisés avec précaution car ils peuvent détecter et supprimer par erreur des signaux biologiques, comme l'ont indiqué leurs auteurs.

1.2.2.6.4.2 Filtrage

Le filtrage est une étape qui vise à retirer les intervalles quantifiés dont les abondances sont trop faibles et pouvant correspondre à du bruit ou des artefacts, et/ou pour lesquelles de faibles erreurs de mesures engendrent des variations importantes du signal. Cette approche vise à améliorer la qualité des données utilisées pour l'analyse. Le seuil de filtrage est déterminé en fonction de critères spécifiques, tels que la distribution des valeurs dans les données ou des connaissances biologiques sur l'échantillon étudié. Par exemple, certaines études utilisent des seuils basés sur des quantiles de la distribution des valeurs, tandis que d'autres utilisent des seuils basés sur des valeurs arbitraires définies en fonction de la sensibilité de la technique de séquençage utilisée.

1.2.2.6.5 Bilan

Les étapes présentées constituent les procédures classiques et couramment utilisées pour le traitement des données RNA-Seq allant du pré-traitement des fichiers FASTQ, de l'alignement et de la quantification des lectures jusqu'au post-traitement des données. Cependant, en fonction de l'objectif de l'étude et des analyses postérieures prévues, ces étapes peuvent varier. Par exemple, le choix des programmes et des outils logiciels peut différer en fonction des besoins spécifiques de l'analyse, des préférences de l'équipe de recherche, ou des particularités des données. Après avoir effectué ces traitements, une variété d'analyses peut être réalisée pour étudier différents aspects comme évoqué précédemment.

1.3 Projet de thèse

Ce travail de thèse vise à mieux comprendre le rôle de la RNase J dans les mécanismes de dégradation des ARN chez *S. aureus*. Ce travail est réalisé en collaboration avec une équipe d'expérimentalistes du laboratoire dirigée par Peter Redder dont la thématique porte sur l'étude du flux des ARN et de la régulation génique chez *S. aureus*. Lors de mon travail de thèse, j'ai développé des méthodes statistiques et bio-informatiques adaptées au traitement et à l'analyse de données de séquençage des extrémités 5' des ARN, données obtenues par l'équipe de P. Redder ayant développé des protocoles expérimentaux permettant d'étudier *in vivo* et à l'échelle du génome l'activité du complexe RNase J. En plus de ces données, d'autres type de données hétérogènes, telles que des données de demi-vie ou encore des données CRAC, ont été prises en compte dans nos analyses.

La partie Résultats est organisée en trois chapitres principaux, chacune abordant des aspects spécifiques de l'étude. Dans le premier chapitre, nous présentons une analyse sur la stabilité des ARN chez *S. aureus,* l'impact de la RNase J sur cette stabilité ainsi que les spécificités de substrat de l'enzyme. De plus, une étude sur son activité de dégradation co-traductionnelle sera également présentée ainsi qu'une analyse exploratoire sur son activité endoribonucléolytique *in vivo*.

Le second chapitre se focalise sur l'identification de la ou des protéines NUDIX ribopyrophosphohydrolases (RppH) responsables de la conversion des ARN 5'PPP en 5'P en vue de leur dégradation par le complexe RNase J. Parmi les cinq RppH prédites *in silico* chez *S. aureus*, seule l'étude de la plus similaire à celle de *B. subtilis* a été publiée concluant à son faible impact sur la stabilité des ARN. Des mutants de délétion pour les cinq RppH candidates (délétions seules ou combinées) ont été obtenus par nos collaborateurs et leurs effets sur l'état de phosphorylation des ARN a été abordé par l'analyse bio-informatique des données issues d'un protocole expérimental spécifique (développé par l'équipe de P. Redder) permettant de déterminer l'état de phosphorylation des extrémités 5' des ARN.

Le troisième chapitre est dédié à la présentation du développement d'EMOTE-tk, une librairie R libre et gratuite, déposée sur le gitlab du CNRS. Cette librairie intègre les méthodes statistiques et bio-informatique que nous avons développé au cours de ce travail pour adapter les traitements aux protocoles expérimentaux originaux de nos collaborateurs. Cette partie permettra de détailler les concepts et les analyses effectué dans la 1ère et la 2ème partie. Le manuscrit se terminera par une partie « Discussions et Perspectives ».

Résultats

Chapitre 1

La RNase J1 est la principale architecte du paysage transcriptomique de *S. aureus*

Chapitre 2

États de phosphorylation des extrémités 5’ des ARN et identification des NUDIX RppH chez *S. aureus*

Chapitre 3

EMOTE toolkit : une librairie R conçue pour l’analyse de données de séquençage des extrémités des ARN

2 Résultats

2.1 Chapitre 1 : La RNase J1 est la principale architecte du paysage transcriptomique de *S. aureus*.

2.1.1 Dégradation des ARN par la RNase J chez *S. aureus*

Chez *Staphylococcus aureus*, il est supposé que la RNase J joue un rôle central dans la dégradation des ARN. Cette hypothèse repose sur plusieurs observations clés. Tout d'abord, la RNase J possède une double activité de dégradation, ce qui lui permettrait non seulement de contourner les protections des extrémités 5' des ARN grâce à un clivage endoribonucléolytique, mais également de dégrader efficacement les ARN via une activité exoribonucléasique.
De plus, les études ont montré que les autres ribonucléases présentes dans cette bactérie, telles que la RNase Y, la PNPase ou encore la RNase III, jouent des rôles plus spécifiques ou de moindre importance dans la dégradation des ARN, ce qui renforce l'idée que la RNase J a une fonction prépondérante. Enfin, les effets pléiotropes sévères observés lors de la mutation du site actif de la RNase J1, qui affectent de nombreux aspects de la physiologie bactérienne tels que la vitesse de croissance ou la résistance au froid, soutiennent fortement cette hypothèse.
Ces observations suggèrent que la RNase J1 n'est pas seulement une actrice parmi d'autres, mais un régulateur majeur de la stabilité et du taux de renouvellement (*turnover*) des ARN chez *S. aureus*.

Cependant, aucune étude *in vivo* n'a encore été réalisée pour mesurer l'impact de la RNase J sur la stabilité des ARN. Dans cette première partie de nos résultats, nous présentons une étude de la stabilité des ARN dans plusieurs souches sauvages et mutantes pour la RNase J. L'objectif est de mesurer l'impact de la RNase J sur la stabilité des ARN et d'identifier les facteurs pouvant réguler son activité.
De plus, afin de compléter l'ensemble d'ARN qui serait dégradé par la RNase J1 et pour mieux comprendre les spécificités de substrat nous proposons également une étude sur les interactions ARN-RNase J1 par l'analyse de données CRAC.
Ensuite, nous proposerons une analyse montrant qu'une partie de son activité 5'→3' exoribonucléasique se fait de manière co-traductionnelle. Enfin, nous présenterons les premières étapes d'une étude visant à déterminer l'activité endoribonucléasique *in vivo* de la RNase J1 à travers l'identification de ses potentiels sites de clivage.

2.1.2 Stabilité des ARN et impact de la RNase J

La plupart des ARNm bactériens ont des demi-vies remarquablement courtes, généralement inférieures à 2,5 minutes comme illustré dans la Figure 57 extraite de (Anderson et al., 2010). Les bactéries sont souvent soumises à des conditions environnementales changeantes et doivent donc être capables d'ajuster rapidement la quantité d'ARN. C'est pourquoi, la régulation de la quantité d'ARN présent dans la cellule est un processus, en général très dynamique et les ARN sont synthétisés en réponse à des signaux environnementaux ou cellulaires et sont rapidement dégradés une fois que ces signaux cessent. Avoir une demi-vie courte pour les ARN permet une régulation rapide et précise de l'expression génique, en éliminant les ARN codants pour des protéines non nécessaires ou potentiellement toxiques et ainsi éviter toute perturbation de la physiologie cellulaire. Chez *S. aureus*, il a été montré que 90 % des ARN avaient une demi-vie inférieure à 5 min (Roberts et al., 2006), (Olson et al., 2011).

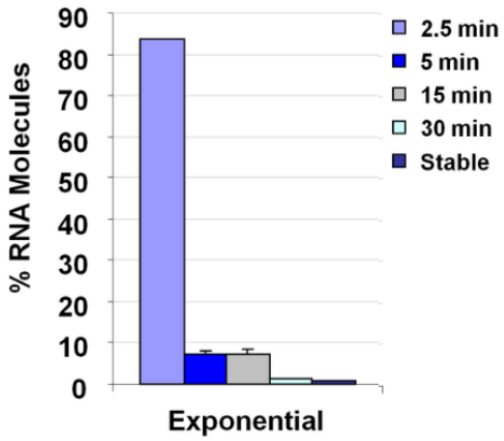


Figure 57 : Distribution des demi-vies des ARNm détectables pendant la phase exponentielle de croissance. *Pourcentage d'ARNm détectables (axe y) ayant une demi-vie ≤ 2,5, 5, 15, 30, ou >30 min pendant la phase exponentielle de croissance. Extrait de* (Anderson et al., 2010)

2.1.2.1 Plus de 80 % des ARN ont une demi-vie impactée par la délétion de la RNase J

Afin d'étudier le rôle ou l'importance de la RNase J dans la stabilité des ARN, des expériences permettant de mesurer la demi-vie des ARN ont été réalisées dans le but de comparer les demi-vies des ARN dans une souche sauvage avec celles dans des souches mutantes pour la RNase J. Dans le cadre, d'un projet réalisé par **Peter Redder** (LMGM), **Julien Prados** et **Natalia Baumann** (Université de Genève), des mutants du site actif de la RNase J1 ainsi que des mutants de délétion des gènes codants pour la RNase J1 et la RNase J2 ont été générés afin de mesurer la stabilité des ARN, et de la comparer à la souche sauvage. Pour mesurer la stabilité des ARN, des expériences de transcriptome ont été réalisées à différents temps après inhibition de la transcription par ajout de rifampicine à 0 s, 75 s, 150 s, 300 s, et 600 s notés t_0, t_{75}, t_{150}, t_{300} et t_{600}. A chaque temps après blocage de la transcription, une partie a été prélevée pour construire une librairie RNA-Seq permettant de mesurer l'abondance de chaque ARN. Chaque préparation de librairie a été conduite en duplicat (Fig. 58A), et a été réalisée sur les trois souches différentes de *S. aureus*. Ces souches correspondent à une souche sauvage SA564 (WT), une souche $J1^{AGA}$ dans laquelle deux histidines de coordination métallique dans le site actif de la RNase J1 ont été mutées en alanines, et une souche ΔJ1ΔJ2 dans laquelle les cadres ouverts de lecture de la RNase J1 et de la RNase J2 présentent tous deux de grandes délétions, garantissant ainsi l'absence complète des 2 protéines.

Après séquençage RNA-Seq, les fichiers FASTQ obtenus ont été pré-traités, alignés et quantifiés par gène afin d'obtenir une table de comptages du nombre de lectures alignées sur chacun des gènes annotés de *S. aureus*.
Les comptages aux différents temps sont issus d'échantillons différents. Il est donc nécessaire d'effectuer une normalisation afin de pouvoir tenir compte, notamment, de différences de profondeur de séquençage, et ainsi de pouvoir quantifier et comparer l'abondance d'ARNm pour chaque gène par rapport à son abondance au début de l'expérience à t_0.

Étant donné qu'après le blocage de la transcription, on s'attend à ce que le nombre d'ARN diminue au cours du temps, de par leur dégradation, les méthodes de normalisation se basant sur l'hypothèse que la grande majorité des gènes ne sont pas impactés et que les variations d'expression (sur-expression et sous-expression) peuvent se compenser (quantité globale d'ARN et distributions similaires de l'ensemble des abondances des transcrits) ne sont pas applicables.

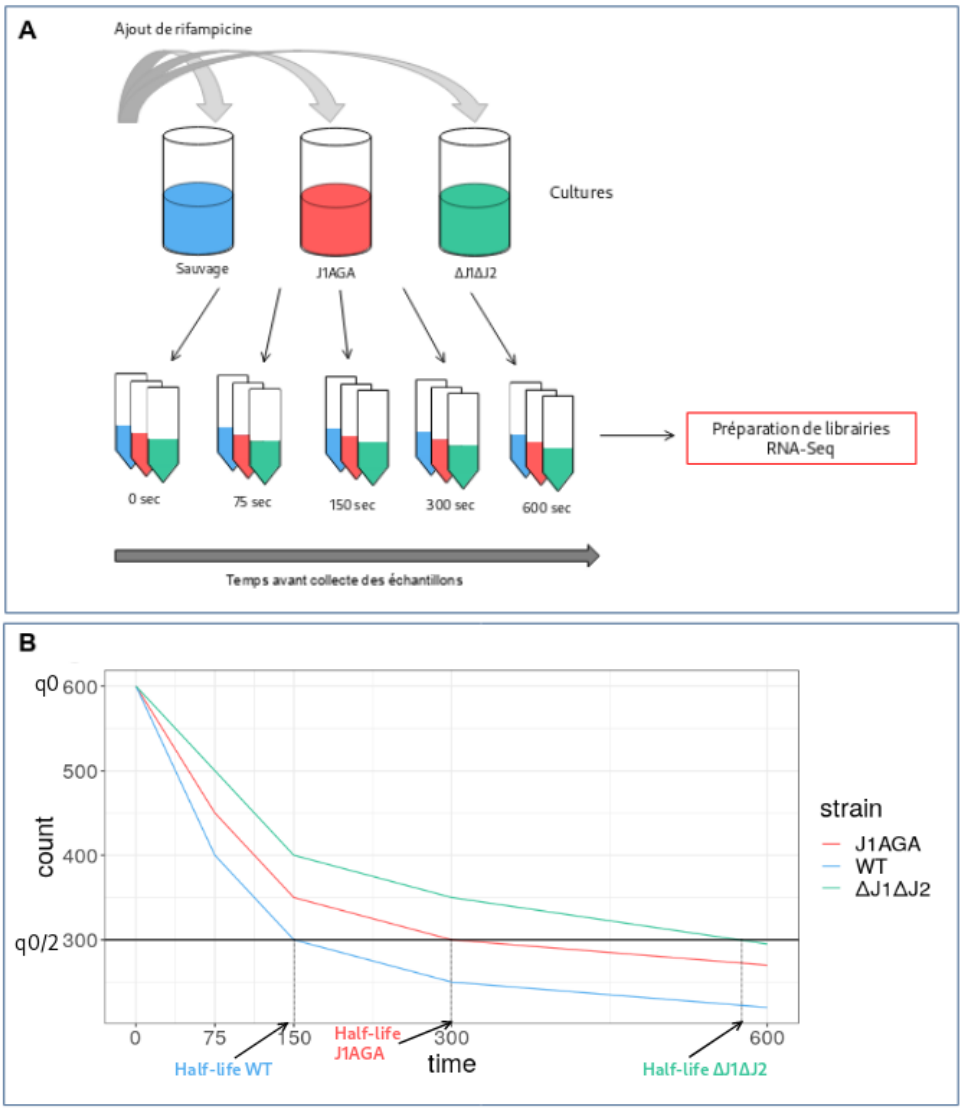


Figure 58 : ***Schémas des procédures expérimentales et analytiques pour les échantillons temporels de RNA-Seq et le calcul des demi-vies des ARN*** *(**A**) Procédure expérimentale pour la production d'un des réplicats des échantillons temporels de RNA-Seq. (**B**) Procédure mathématique pour le calcul de la demi-vie d'un ARN dont la quantité q_0 de départ à t_0 est de 600 dans les 3 souches.*

Pour normaliser les abondances par rapport à leur quantité de départ, il est nécessaire de disposer d'une référence telle que l'existence d'ARN très stables dont le niveau ne devrait pas varier (diminuer) en l'absence de transcription. L'ARN sélectionné est celui correspondant au gène *hu* qui code pour une protéine de liaison à l'ADN impliquée dans la structuration de la chromatine, et donc dans la condensation du chromosome (source : UniProtKB (Swiss-Prot)). Ce gène est connu pour être un gène fortement exprimé dont l'ARN est très stable (Giraud et al., 2015). Son nombre de lectures ne devrait donc pas varier entre les temps t_0 et $t_{600,}$ et il a été utilisé comme référence pour la normalisation des abondances aux différents temps et entre répétitions.

Les demi-vies des ARN dont la quantité initiale ne diminue pas de moitié avant le temps du dernier prélèvement (à 600 s) peuvent tout de même être calculées par extrapolation. Cependant, plus les valeurs de demi-vie sont grandes par rapport à t_{600}, plus la marge l'erreur sur la valeur extrapolée est importante. C'est pourquoi un seuil a été fixé arbitrairement à 1800 secondes sur la distribution des données de demi-vie.

Pour un ARN ayant une demi-vie de 1800 secondes, son abondance sur une échelle logarithmique serait à t_0 de $log_2(x_0)$ et à t_{1800} $log_2(x_0/2)$, on en déduit son abondance relative par rapport à t_0 pour $t_{600} = 2^{600\ log2(1/2)\ /\ 1800} = 0{,}79$. Cela impliquerait une diminution d'expression de seulement 20 % entre t_0 et t_{600}. Cette différence est relativement faible et difficile à détecter avec précision dans une expérience de RNA-Seq, en raison des erreurs de mesure et de normalisation des expressions. Afin d'illustrer ce point, la Figure 59 illustre l'écart de mesure entre les deux réplicats biologiques sauvage, lequel est estimé à environ 28 % sur l'ensemble des gènes. Comme cette valeur dépasse les 20 %, la précision sur l'estimation des demi-vies au-delà de 1800 secondes serait insuffisante avec ce plan d'expérience.

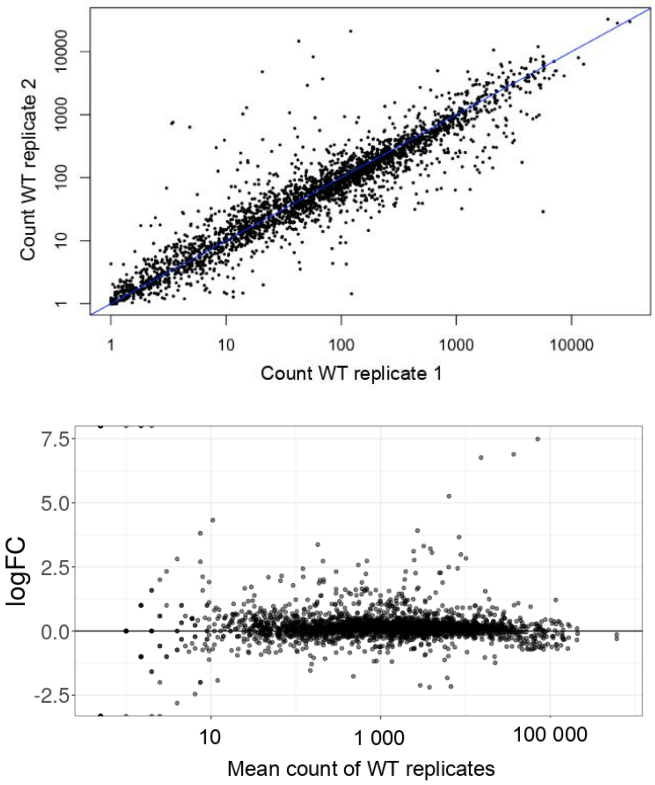


Figure 59 :* Comparaison des demi-vies des gènes entre réplicats sauvages.** (A***) *Demi-vie des gènes issus du premier réplicat sauvage en fonction de celles issus du deuxième réplicat.* (***B***) *Logarithmes des ratios des demi-vie des gènes entre les deux réplicats sauvages en fonction de la moyenne des comptages (log_{10}).*

Toutes ces étapes ont été réalisées par **Julien Prados** et **Natalia Baumann** (Université de Genève) qui nous ont fourni les données sous forme d'un tableau (Table 1) représentant les demi-vies des ARN correspondant aux gènes de *S. aureus* calculées (en secondes) comme décrit en introduction (Fig. 58B). A partir de ce tableau de demi-vie, nous avons pu réaliser un ensemble de traitements et d'analyses que nous allons maintenant décrire.

locus_tag	halflife.DJ1DJ2_rep1	halflife.DJ1DJ2_rep2	halflife.J1AGA_rep1	halflife.J1AGA_rep2	halflife.WT_rep1	halflife.WT_rep2
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
RT87_00005	115.	118.	106.	101.	138.	116.
RT87_00010	146.	147.	102.	103.	142.	120.
RT87_00015	171.	167.	155.	147.	258.	186.
RT87_00020	139.	145.	132.	129.	173.	148.
RT87_00025	138.	150.	139.	134.	179.	144.
RT87_00030	148.	149.	145.	142.	159.	122.
RT87_00035	232.	222.	217.	211.	233.	168.
RT87_00040	546.	560.	1409.	1174.	236.	271.
RT87_00045	316.	342.	289.	290.	152.	145.
RT87_00050	426.	461.	519.	440.	235.	190.

Table 1 : Tableau des demi-vies des gènes de S. aureus*. Premières lignes du tableau contenant les demi-vies des gènes (en secondes) de S. aureus en fonction des souches et réplicats.*

Dans un premier temps, nous avons réalisé une analyse en composante principale (ACP) afin de représenter la dispersion des échantillons en fonction des valeurs de demi-vie des gènes (Fig. 59). Les données sont représentées sur les deux premières composantes principales. La séparation des échantillons issus des souches mutantes, à droite, et sauvage, à gauche, sur la première composante principale, suggère qu'ils ont des profils bien distincts. La proximité sur les deux premières composantes principales des échantillons issus de la souche $J1^{AGA}$ et issus de la souche ΔJ1ΔJ2 indiquent qu'elles ont des profils de demi-vie assez semblables, d'autant plus que les deux premières composantes principales permettent d'expliquer 94 % de la variance.

Dans une cellule, l'expression génique est régulée de manière dynamique en réponse à divers signaux internes et externes. Par conséquent, à un moment donné et dans une cellule donnée, seuls certains gènes sont actifs et sont transcrits en ARN, tandis que d'autres sont inactifs.
Cela implique qu'il n'est pas possible de mesurer la demi-vie des ARN pour les gènes inactifs dans les conditions de l'expérience. De plus, certains gènes sont également exprimés à des niveaux trop bas pour être détectés de manière fiable. Ainsi, dans une analyse de demi-vie des ARN, seuls les ARN exprimés à des niveaux détectables sont pris en compte, reflétant ainsi les processus biologiques actifs dans la cellule à un moment donné. Les gènes peu exprimés peuvent avoir des niveaux d'expression proches du bruit de fond, ce qui rend leur estimation de demi-vie moins fiable.

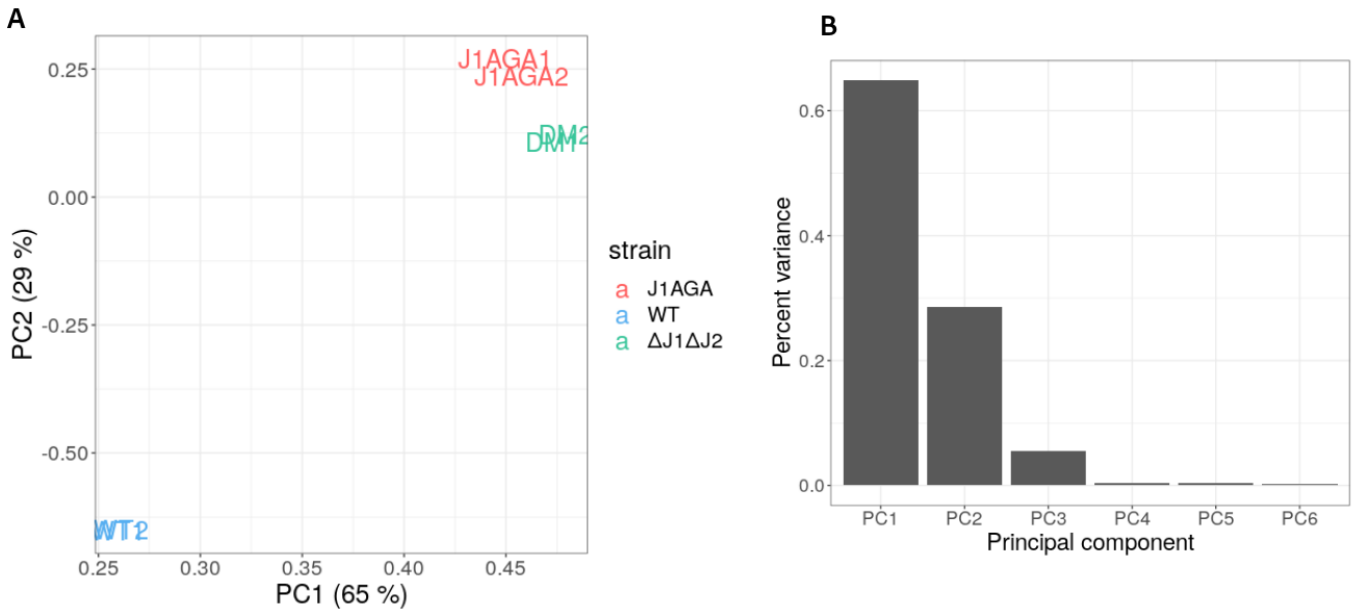


Figure 59 : ***Analyse des demi-vies des gènes par ACP*** *(**A**) Analyse en composantes principales (ACP) des demi-vies des gènes dans les différents échantillons. (**B**) Variances portées par les composantes.*

Dans un premier temps, nous avons donc effectué une étape de filtrage de ces gènes. Pour cela, nous avons récupéré les comptages normalisés qui ont servi au calcul des demi-vies. La normalisation effectuée entre les échantillons par rapport à l'ARN du gène *hu* ne tient pas compte de la longueur des gènes mais seulement du nombre de lectures alignées sur celui-ci. En effet, plus un gène est long, plus son ARN sera long, et plus l'étape de fragmentation générera de fragments à séquencer. Afin de pouvoir comparer l'expression des gènes au sein d'un même échantillon, nous avons donc appliqué une normalisation RPKM.

Une fois les données normalisées, nous pouvons filtrer les gènes avec une valeur de RPKM trop faible. Le filtre appliqué est arbitraire et correspond à filtrer les gènes pour lesquels la somme des RPKM dans les échantillons dans lesquels les transcrits sont les plus abondants (t_0 et t_{75}) est inférieure à 20 (Fig. 60).

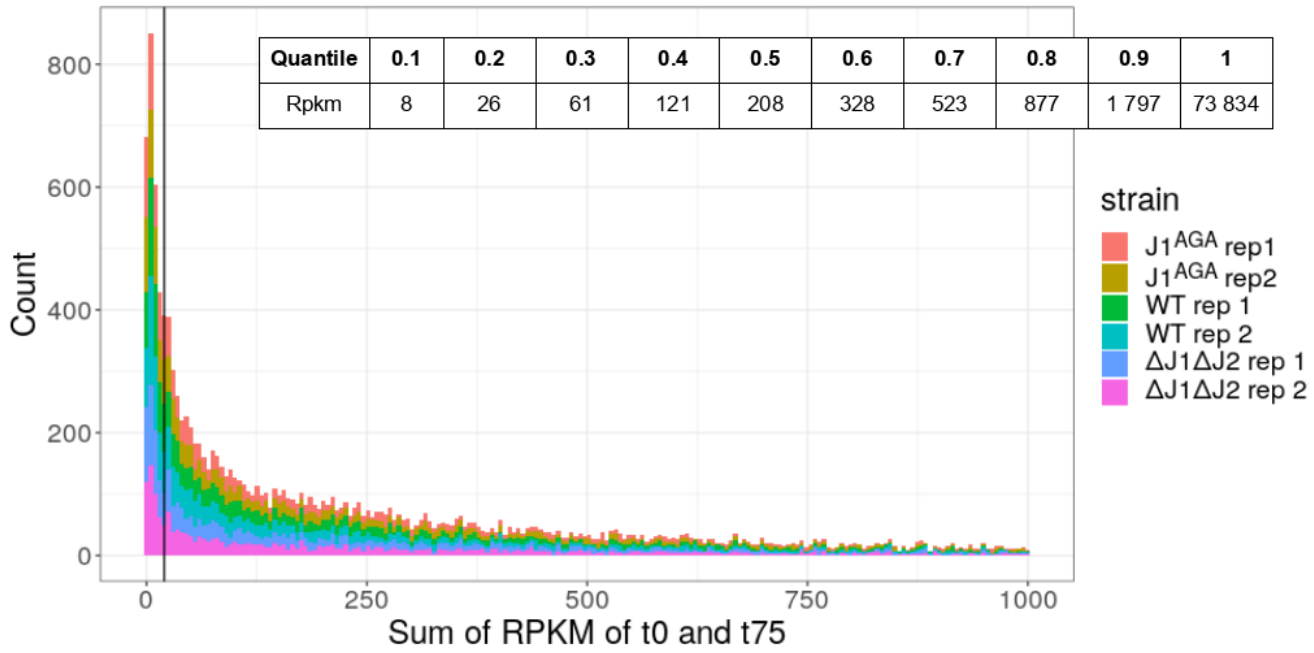


Figure 60 : ***Expression des gènes par échantillons temporels, souche et réplicats****. Histogramme des sommes de RPKM des gènes issus des échantillons t_0 et t_{75} coloriées par souches et réplicats. La ligne verticale représente le seuil de 20 RPKM utilisé pour le filtre.*

Au total, sur les 2703 gènes annotés dans la souche étudiée, 2073 gènes (77 %) sont considérés comme suffisamment exprimés.

Afin d'écarter les valeurs qui ne concordent pas entre les réplicats biologiques, nous avons appliqué un filtre supplémentaire. Le seuil a été fixé arbitrairement afin de supprimer les gènes pour lesquels il est observé un changement supérieur à un facteur de 1,5 entre les demi-vies des deux réplicats (Fig. 61). Sur les 2073 gènes exprimés, 2002 sont considérés comme suffisamment concordants. Cela correspond à 97 % des gènes, ce qui montre que les répétitions de l'expérience sont globalement concordantes et que les résultats devraient être reproductibles.

Conformément à ce qui a été observé précédemment dans cette souche et d'autres souches de *S. aureus* (Anderson et al., 2010), (Khemici et al., 2015), nous pouvons observer que la majorité des ARN dans la souche sauvage a une demi-vie d'environ 2 à 3 minutes, et que plus de 75 % des ARN a une demi-vie inférieure à quatre minutes (Table 2). Les demi-vies des ARN dans les souches $J1^{AGA}$ et ΔJ1ΔJ2 sont nettement différentes de celles de la souche sauvage comme illustré ci-dessous en Figure 62 et Figure 63 où l'on observe de manière générale une augmentation des demi-vies dans la souche $J1^{AGA}$ par rapport à la souche sauvage.

En effet, chez le mutant $J1^{AGA}$, plus de 75 % des ARN a désormais une demi-vie supérieure à 4 minutes (Table 2).

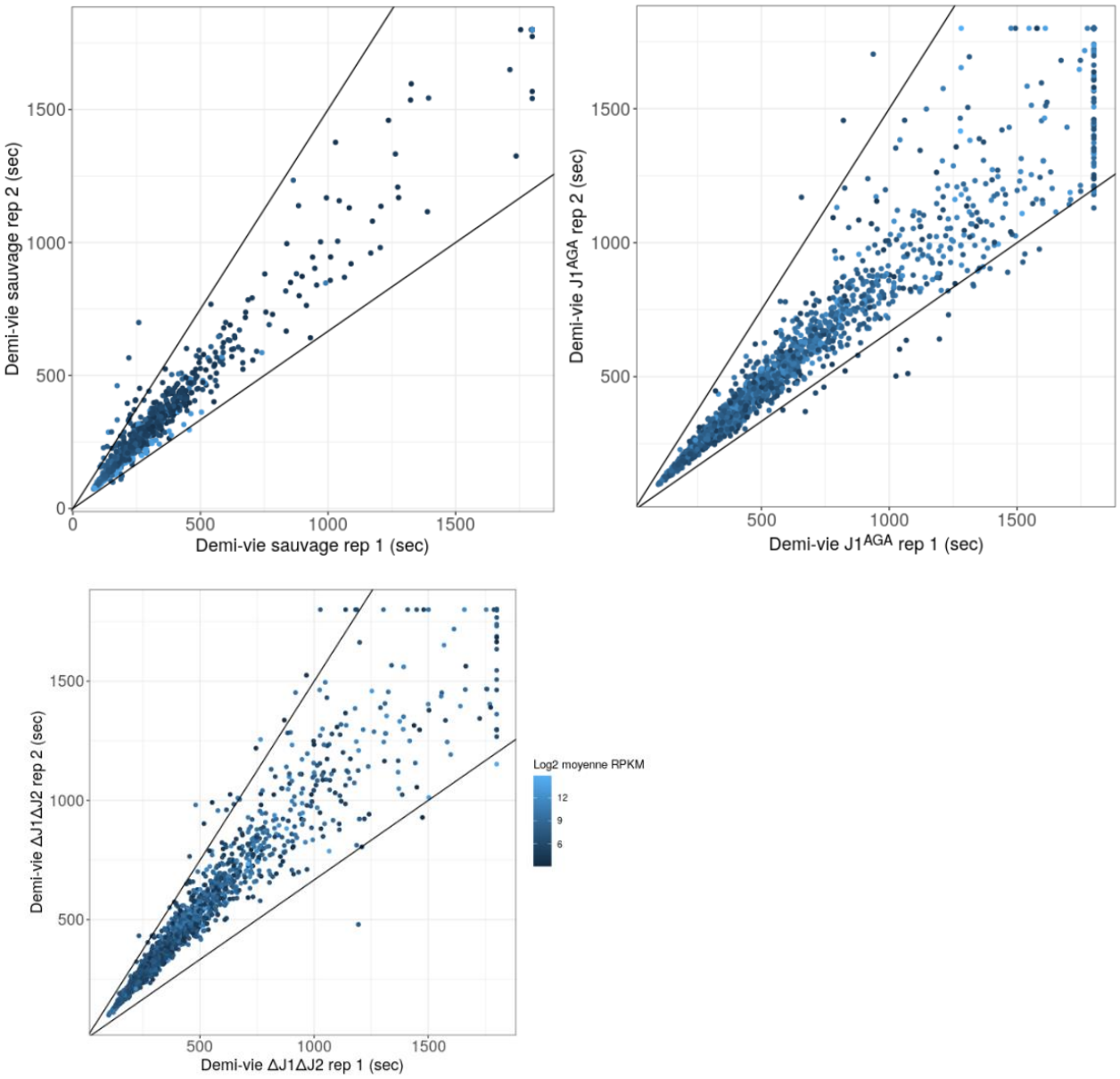


Figure 61 *:* ***Demi-vie des gènes par réplicat biologique****. Demi-vie des réplicats 1 (axe x) en fonction des réplicats 2 (axe y) par souches. Les lignes noires représentent les seuils utilisés pour le filtrage de la concordance des réplicats. Les points en dehors des cônes sont filtrés.*

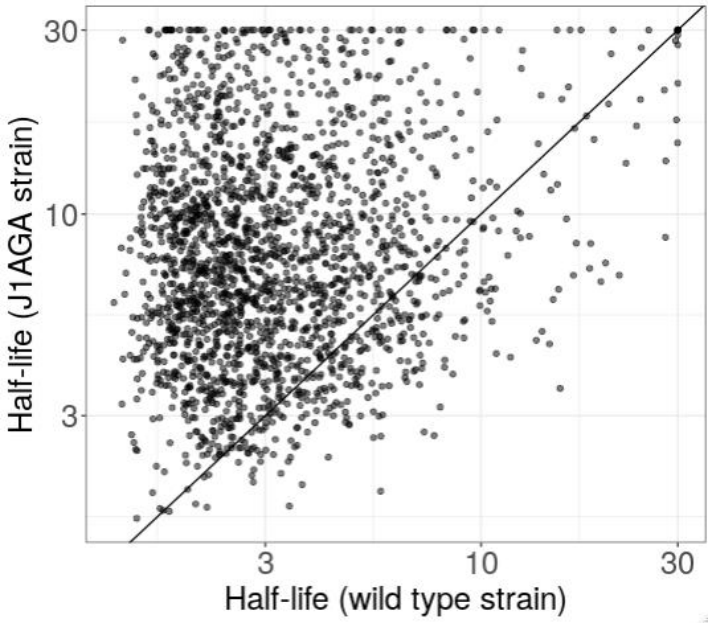


Figure 62 : ***Comparaison des demi-vies sauvage et mutante****. Demi-vies par gène (moyenne des répétitions) dans la souche J1^{AGA} (axe y) par rapport à celles dans la souche sauvage (axe x).*

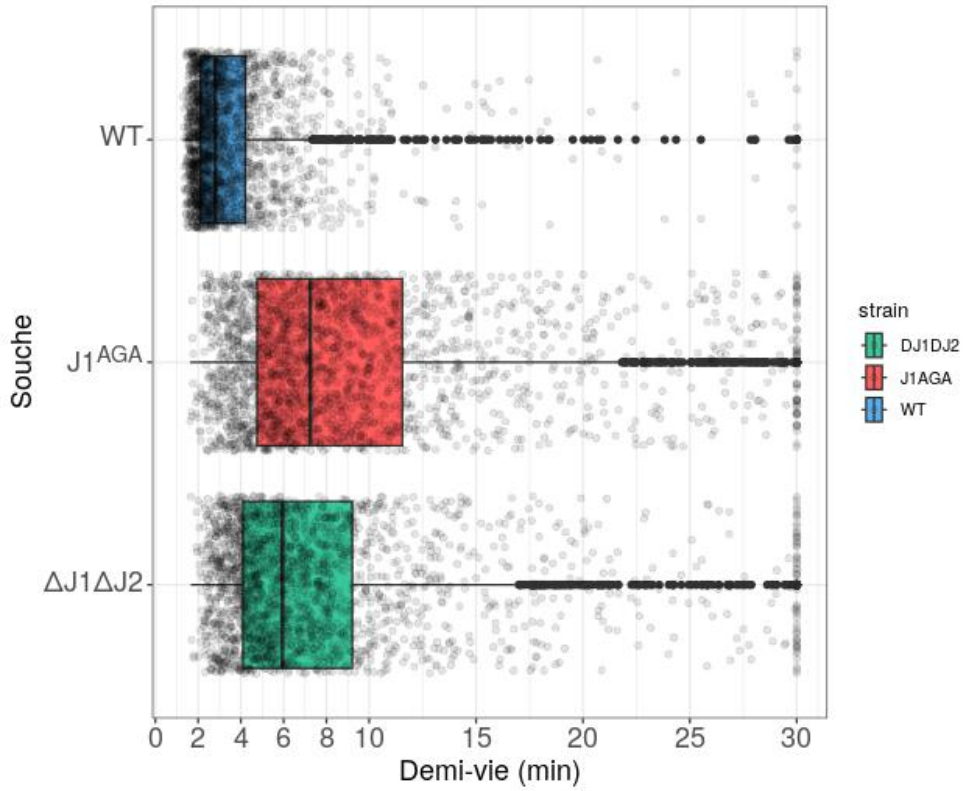


Figure 63 : ***Distribution des demi-vies des gènes par souches****. Les demi-vies des gènes correspondent à la moyenne des répétitions dans les souches sauvage (bleu), J1^{AGA} (rouge) et ΔJ1ΔJ2 (vert).*

Percentile → Souche ↓	0 %	25 %	50 %	75 %	100 %
WT	1,28	2,12	2,84	4,3	30,0
$J1^{AGA}$	1,62	4,81	7,42	12,0	30,0
DJ1DJ2	1,65	4,13	6,03	9,56	30,0

Table 2 : Demi-vie des gènes en minutes par souche et par percentile.

Les mutations dans le site actif de la RNase J1 impactent fortement la stabilité de la majorité des ARN de *S. aureus.* Il est intéressant de noter que les demi-vies dans la souche ΔJ1ΔJ2 sont légèrement, mais de manière statistiquement significative (p-valeur du test de Wilcoxon-Mann-Whitney = $4{,}3.10^{-14}$), moins élevées que dans la souche $J1^{AGA}$. Cela indique que la délétion de la RNase J2, et plus largement du J-complexe, n'entraîne pas davantage de stabilité des ARN par rapport à $J1^{AGA}$. Une hypothèse pour expliquer cette différence serait que, dans la souche $J1^{AGA}$, le complexe J inactif serait encore capable de se lier aux ARN, et pourrait entraver l'accès à d'autres ribonucléases, empêchant ainsi la dégradation de ces ARN.

2.1.2.2 La demi-vie des gènes est corrélée avec leur niveau d'expression

Linder et al. (2014) ont montré que les phénotypes des souches mutantes $J1^{AGA}$ et ΔJ1ΔJ2 ont un phénotype gravement impacté : croissance lente et sensibilité à la température. Afin d'étudier s'il existe un lien entre le niveau d'expression et la stabilité des ARN, nous avons cherché à déterminer si les gènes différentiellement exprimés entre les souches sauvage et mutantes sont les mêmes que ceux dont la demi-vie des ARN est impactée par l'inactivation du site actif de la RNase J1.

Pour ce faire, nous avons effectué une analyse d'expression différentielle de gènes à l'aide de la librairie EdgeR (Robinson et al., 2010) à partir des comptages normalisés avec la méthode TMM à t_0, correspondant donc à des conditions de croissance sans blocage de la transcription entre les souches étudiées.

Les résultats sont présentés de manière globale pour l'ensemble des gènes en Figure 64 avec pour chaque gène, sa variation d'expression par rapport à sa variation de demi-vie. En accord avec les résultats présentés dans Linder et al. (2014) et le phénotype de croissance sévèrement impacté de $J1^{AGA}$, nous observons que le transcriptome est modifié à l'échelle globale avec de nombreux gènes surexprimés et sous-exprimés (Fig. 64) accompagnés, la plupart du temps, d'une augmentation de leur demi-vie.

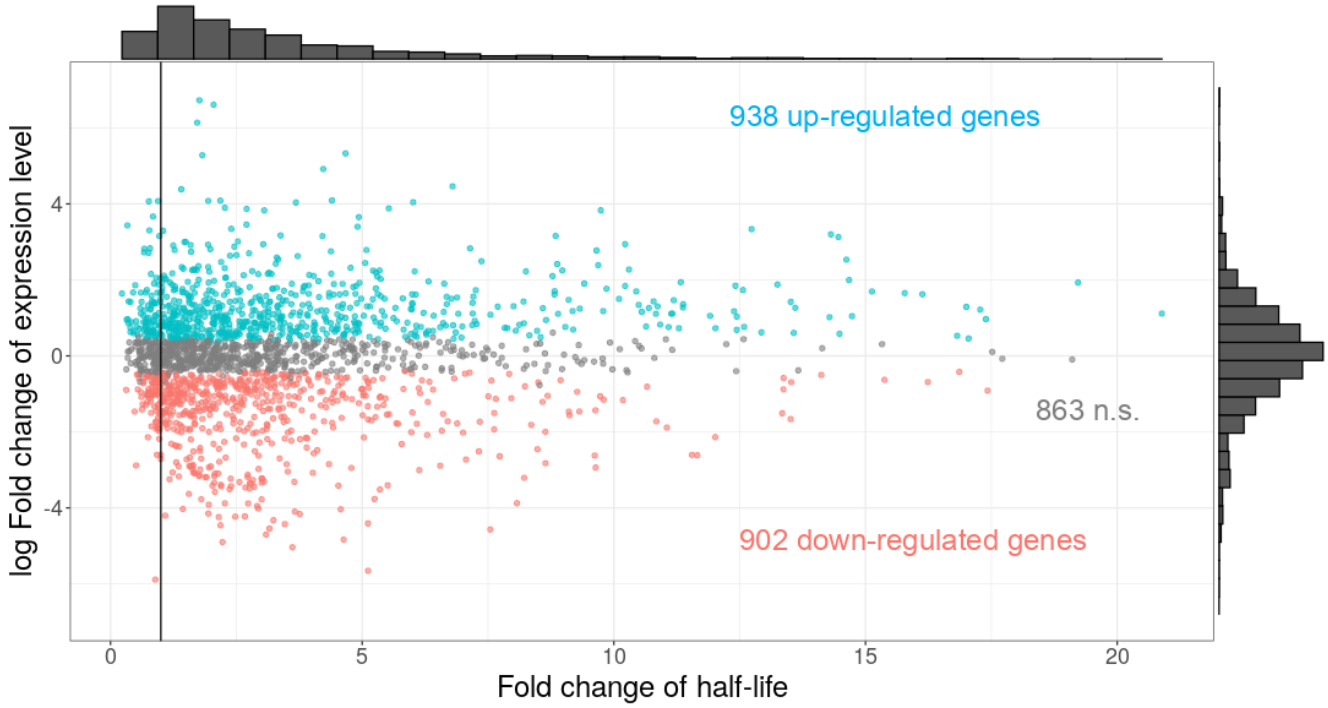


Figure 64 : ***Relation entre variation d'expression et variation de demi-vie des gènes entre souches.*** *Variation d'expression (axe y) des gènes en fonction de la variation de demi-vie (axe x) des gènes dans une souche sauvage par rapport à une souche $J1^{AGA}$ à t_0. Les histogrammes sur les côtés correspondent à la distribution des variations de demi-vie (en haut) et d'expression (à droite) de la souche mutante par rapport à la souche sauvage. La ligne verticale représente une variation de demi-vie d'un facteur 1.*

Une légère, mais significative, corrélation est observée (coefficient de corrélation de Pearson r = 0.25 ; p-valeur du test de Student associé < 2.10^{-16}) suggérant que les gènes dont l'expression augmente dans le mutant $J1^{AGA}$ par rapport à la souche sauvage, ont également tendance à avoir des ARN dont la demi-vie augmente.

De plus, nous avons également observé une de corrélation négative (r = -0,55 ; p-valeur<2.10^{-16}) entre la demi-vie des gènes en $\log_2$ dans la souche sauvage et le logarithme de leur niveau d'expression (Fig. 65 WT). Cette observation indique que les gènes fortement exprimés ont tendance à avoir des ARN ayant une demi-vie courte. En effet, les ARN ayant une demi-vie supérieure à 10 minutes correspondent presque tous à des gènes peu exprimés (niveau d'expression inférieur à 50 RPKM).

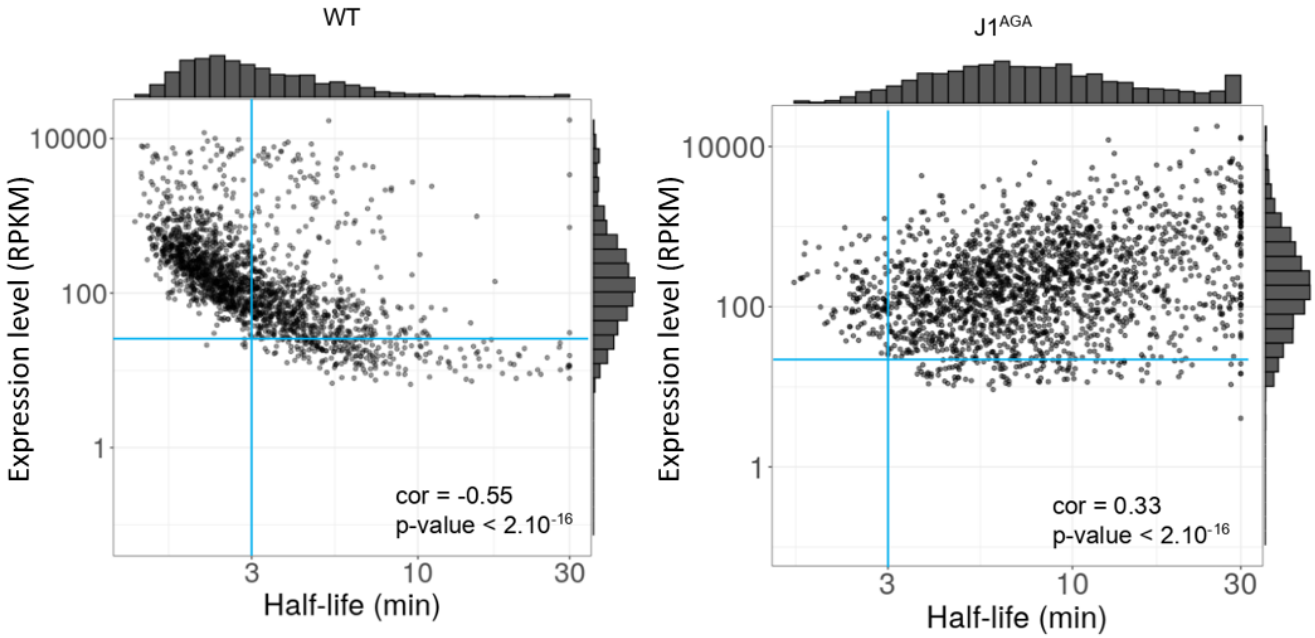


Figure 65 : Relation entre niveau d'expression et demi-vie des gènes par souches.

Niveau d'expression des gènes (axe y échelle logarithmique) en fonction de la demi-vie de leur ARN (axe x en logarithme) dans une souche dans la souche sauvage (gauche) et dans la souche J1^{AGA} (droite). Les lignes horizontales et verticales bleues représentent respectivement, 25 RPKM et 3 minutes.

Cette corrélation négative n'est plus observée dans la souche J1^{AGA} comme illustré en Figure 65 (J1^{AGA}), au contraire, on observe une faible corrélation positive (r = 0,33). Ainsi, nous pouvons observer que, globalement, les gènes fortement exprimés (expression supérieure à 100 RPKM) sont dégradés rapidement (demi-vies inférieure 3 minutes) dans la souche sauvage, et que cette dégradation semble être le fait de la RNase J1, car les demi-vies de ces gènes ont tendance à augmenter dans la souche mutante J1^{AGA}. Quant aux gènes faiblement exprimés (expression inférieure à 25 RPKM), leur demi-vie est assez élevée (supérieure à 3 minutes) dans la souche sauvage, et ne semble pas varier fortement dans le mutant J1^{AGA}.

2.1.2.3 Près de la moitié des gènes ont la dégradation de leurs transcrits qui est dépendante de la RNase J

Nous avons établi dans la partie précédente que certains gènes ont des ARN dont les demi-vies sont plus ou moins impactés par la mutation de la RNase J1. L'objectif est maintenant de déterminer ceux dont la dégradation est dépendante ou non de la RNase J1. Afin de faciliter le discours, nous dirons à partir de maintenant, qu'un gène est « impacté » si la dégradation de ses transcrits (estimée par les lectures alignées sur le gène) est dépendante de la RNase J1, et, au contraire, qu'il est « non impacté » si cette dégradation n'est pas dépendante de la RNase J1.

Pour ce faire, nous avons établi un ensemble de critères permettant de définir les deux groupes.

Nous avons considéré qu'un gène est impacté lorsque sa demi-vie (moyenne des réplicats) était au moins deux fois supérieure dans les mutants $J1^{AGA}$ et ΔJ1ΔJ2 par rapport à sa demi-vie dans la souche sauvage. De plus, afin de tenir compte de la variabilité naturelle et des erreurs de mesure non systématique, les expériences ont été menées en duplicats, ce afin de réaliser des tests statistiques. Nous avons utilisé des modèles linéaires pour filtrer les augmentations statistiquement significatives des changements de demi-vie.

Selon ces critères, nous ne serons pas en mesure d'identifier les gènes impactés si leur demi-vie dans la souche sauvage est supérieure ou égale à 900 s (15 min), car dans les souches mutantes, leur demi-vie ne pourra pas dépasser le seuil de variation (x2) étant donné que la demi-vie maximale estimée a été fixée à 1800 s (30 min).

Les 2002 gènes pour lesquels les mesures étaient exploitables (suffisamment exprimés et répétitions suffisamment concordantes) sont répartis dans la table de contingence en Table 3, par rapport à leur demi-vie dans la souche sauvage et dans la souche mutante $J1^{AGA}$. Cette table nous permet d'observer que pour 5 gènes, leur demi-vie est supérieure ou égale à 1800 s dans les deux souches, et que nous ne pourrons donc pas décider de l'impact de la RNase J1 sur leur stabilité. Il en va de même lorsque la demi-vie d'un gène est supérieure ou égale à 1800 s dans une des deux souches : 5 gènes ont une demi-vie entre 900 s et 1800 s dans la souche sauvage, et supérieure ou égale à 1800 s dans la souche mutante. Leur demi-vie ne pourra donc pas dépasser le seuil fixer pour détecter une variation suffisante (demi-vie au minimum 2 fois plus grande). A l'inverse, si un gène a une demi-vie inférieure à 900 s dans la souche sauvage (1ère colonne), il sera possible de déterminer si sa stabilité n'est pas impactée ou augmente significativement. Et dans la 2ème colonne pour laquelle la demi-vie dans la souche sauvage se situe entre 900 et 1800 s, il sera possible, dans certains cas, de déterminer qu'il y a une diminution significative, ou bien qu'il n'y a pas d'augmentation significative, et donc d'impact de la RNase J1.

WT → $J1^{AGA}$ ↓	≤ 900 s	900 s < X < 1800 s	≥ 1800 s	Total
≤ 900 s	1664	14	0	1678
900 s < X < 1800 s	257	9	3	269
≥ 1800 s	45	5	5	55
Total	1966	28	8	2002

Table 3 : Répartition des gènes en fonction de leur demi-vie*. Nombre de gènes ayant une demi-vie inférieure à 900 s, entre 900 et 1800 s et supérieure à 1800 s dans les souches sauvages et $J1^{AGA}$*

Les cases du tableau avec un fond vert correspondent aux gènes pour lesquels il sera possible d'observer une variation suffisante de la demi-vie dans la souche $J1^{AGA}$ par rapport à la souche sauvage. Cela correspond à 1 989 gènes. À l'inverse, les cases avec un fond rouge indiquent les gènes pour lesquels il ne sera pas possible de détecter une variation suffisante, ce qui correspond à 13 gènes.

Pour la suite des analyses, nous avons conservé uniquement les 1989 gènes pour lesquels il était possible de détecter une variation significative de leur demi-vie.

Les résultats des tests statistiques sont présentés sous forme de *volcano plot* dans la Figure 66, illustrant la variation de demi-vie des gènes entre les deux souches et la significativité statistique de cette différence.

Sur les 1 989 gènes pour lesquels des données exploitables ont pu être obtenues :

- 937 gènes ont une demi-vie augmente de manière statistiquement significative, c'est-à-dire qui a au moins doublé (p-valeur ajustée < 0,01) à la fois dans les mutants $J1^{AGA}$ et le mutant ΔJ1ΔJ2.
- 422 gènes ont une demi-vie qui ne varie pas de manière statistiquement significative, c'est-à-dire que la variation de leur demi-vie est inférieure à 1,5 en valeur absolue dans les mutants $J1^{AGA}$ et ΔJ1ΔJ2.
- 135 gènes ont une demi-vie qui varie peu ou de manière peu significative et pour lesquels il est difficile de décider d'une variation réelle, c'est-à-dire que leur demi-vie augmente ou diminue entre 1,5 et 2 fois dans les mutants $J1^{AGA}$ et ΔJ1ΔJ2.
- 17 gènes ont une demi-vie 2 fois moins élevée dans les mutants $J1^{AGA}$ et ΔJ1ΔJ2 que dans la souche sauvage.

Ces résultats montrent que la dégradation d'au moins 47,1 % (937/1989) des gènes transcrits nécessite une RNase J1 fonctionnelle.

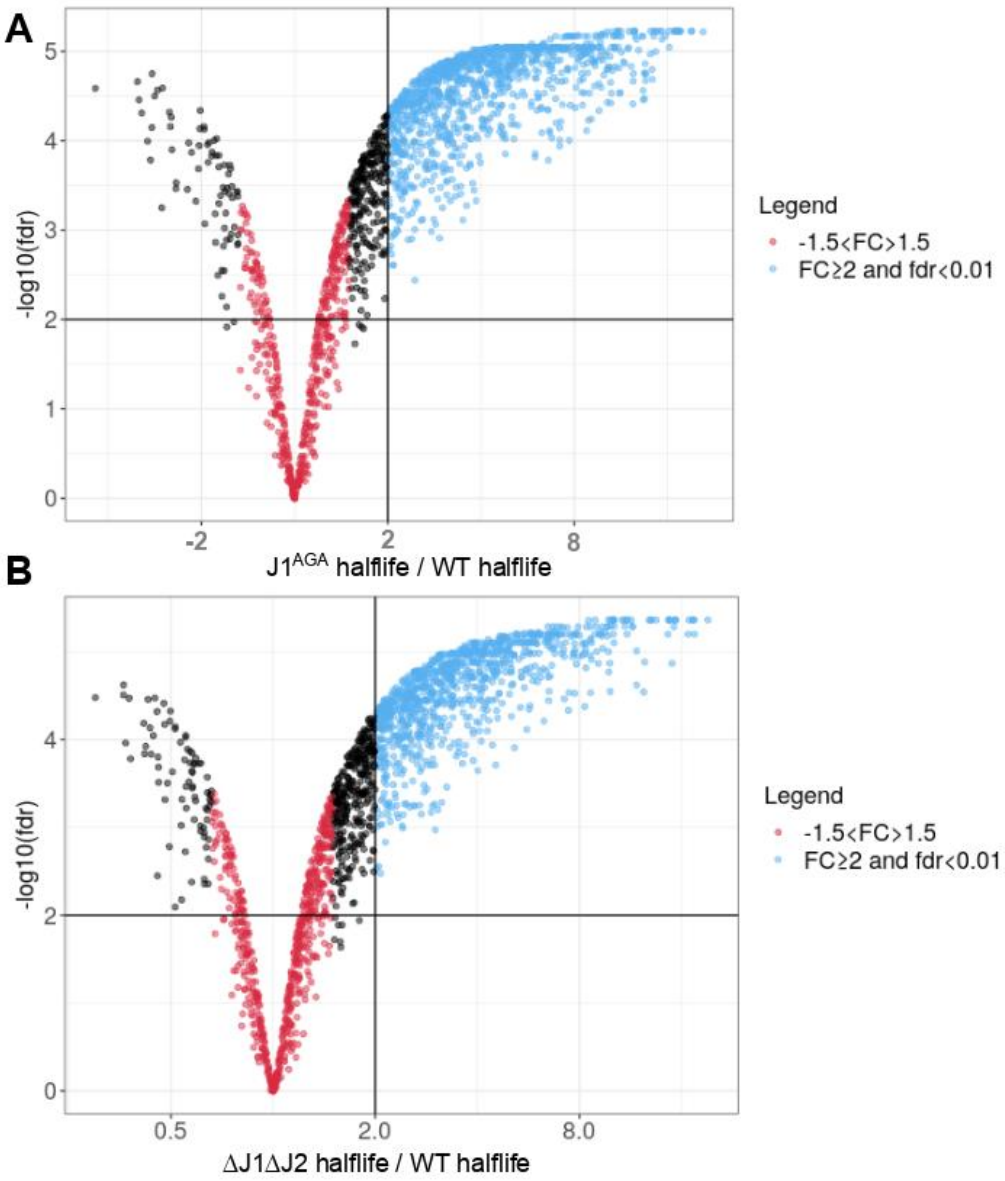


Figure 66 : ***Variation statistique de demi-vie entre la souche sauvage et les souches mutantes.*** *Volcano plots issus de la comparaison entre les demi-vies des gènes dans les souches sauvage et mutantes pour la RNase J. (**A**) Demi-vie des gènes transcrits issus de la souche sauvage versus ceux issus de la souche J1^{AGA} (**B**) Demi-vies des gènes transcrits issus de la souche sauvage versus ceux issus de la souche ΔJ1ΔJ2. Les points bleus correspondent aux gènes caractérisés comme ayant un ARN dont la stabilité est augmentée par la mutation. Les points rouges correspondent aux gènes caractérisés comme ayant un ARN dont la stabilité n'est pas impactée par la mutation. Les points noirs correspondent aux gènes non caractérisés.*

Comme l'illustrent les volcano plots (Fig. 66), les seuils utilisés (augmentation d'un facteur 2 et p-valeur ajustée inférieure à 0,01) ont été fixés arbitrairement, et qu'en fonction de ceux-ci, quelques dizaines de gènes pourraient être ajoutés ou retirés des deux groupes. Néanmoins, ces critères coïncident entre eux puisqu'aucun gène avec une variation de demi-vie supérieure ou égale à un facteur 2 n'a de p-valeur ajustée supérieure à 0,01. Il est également apparent qu'une grande partie des gènes est impactée.

Ces résultats viennent confirmer les précédentes observations montrant que la mutation du site actif de la RNase J1 suffisait à modifier la demi-vie des gènes, et que la délétion complète des deux RNase J1 et J2 ne permettait pas d'augmenter davantage leur stabilité. Cela suggère un rôle moindre de la RNase J2, comme cela a déjà été observé par (Linder et al., 2014). En revanche, il y a davantage de gènes dont la demi-vie augmente (106 gènes) dans la souche $J1^{AGA}$ par rapport à la souche double mutante ΔJ1ΔJ2 (Fig. 67), comme nous l'avions observé lors des analyses globales.

Nous pouvons donc conclure que la RNase J1, par son site actif, est responsable de la dégradation des ARN correspondants à près de la moitié des gènes de *S. aureus*.

Il est important de noter que les estimations de demi-vie ne permettent pas d'identifier tous les ARN dégradés par la RNase J, mais seulement ceux qui sont principalement, voire uniquement, dégradés par cette dernière. En effet, un ARN peut être la cible de multiples ribonucléases. Par conséquent, la demi-vie d'un gène donné peut ne pas être significativement affectée par la mutation d'une ribonucléase particulière si d'autres sont capables de le dégrader efficacement. Ainsi, notre analyse se limite à la détection des cibles dont la dégradation dépend principalement de la RNase étudiée, et peut en manquer de nombreuses, si elles sont dégradées de manière complémentaire par plusieurs ribonucléases. C'est pourquoi nous parlons de gènes dont la dégradation est dépendante (impacté) ou non dépendante (non impacté) de la RNase J.

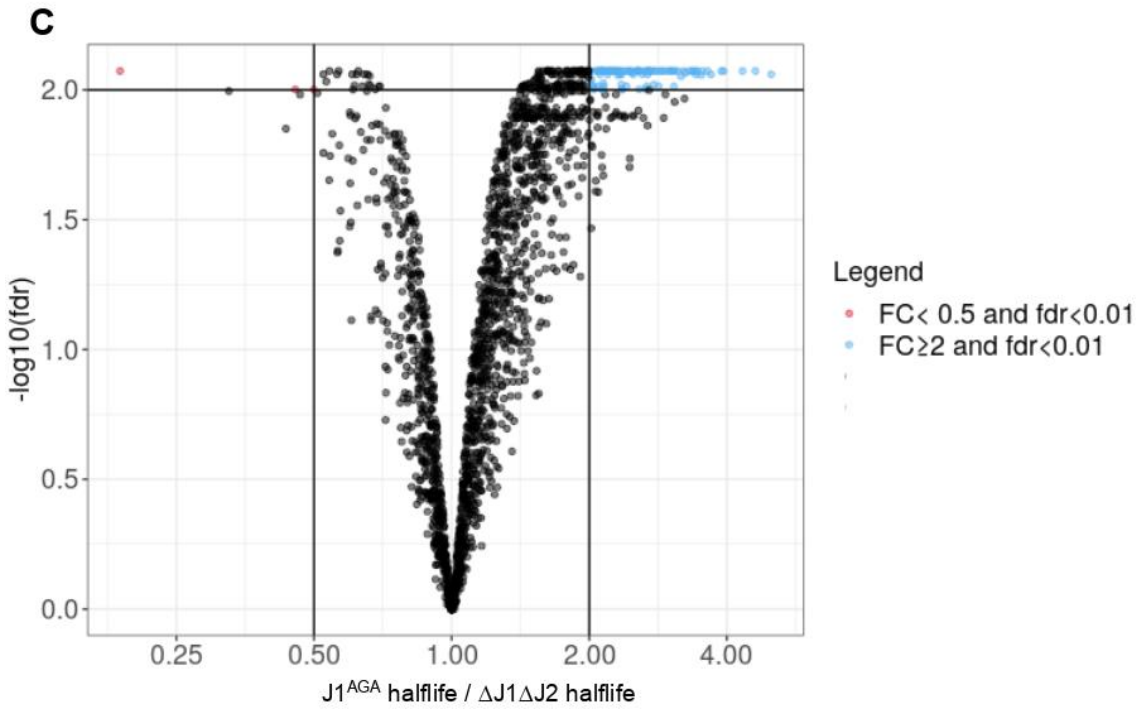


Figure 67 : Variation statistique de demi-vie entre les souches mutantes. *Volcano plot issu de la comparaison entre les demi-vies des gènes dans les souches mutantes pour la RNase J. Demi-vie des ARN issus de la souche J1^{AGA} versus ceux issus de la souche ΔJ1ΔJ2. Les points bleus correspondent aux gènes caractérisés comme ayant un ARN dont la stabilité est plus importante dans la souche J1^{AGA} par rapport à la souche ΔJ1ΔJ2. Les points rouges correspondent aux gènes caractérisés comme ayant un ARN dont la stabilité est moins importante dans la souche J1^{AGA} par rapport à la souche ΔJ1ΔJ2. Les points noirs correspondent aux gènes caractérisés comme ayant un ARN dont la stabilité ne varie pas significativement d'une souche à l'autre.*

2.1.2.4 Les gènes non impactés par la délétion de la RNase J1 sont dégradés par d'autres ribonucléases

Comme évoqué précédemment, il pourrait s'avérer que les ARN non impactés par l'absence de la RNase J pourraient tout de même être dégradés par celle-ci, mais non détectés dans l'analyse précédente si ces ARN étaient aussi dégradés par d'autres ribonucléases. Une autre hypothèse est que ces ARN soient très stables (> 5 min) et donc peu dégradés de manière générale.

Afin de tester cette dernière hypothèse, nous avons comparé les demi-vies, dans la souche sauvage, des deux groupes d'ARN précédemment identifiés à l'aide d'un test de Wilcoxon-Mann-Whitney. De plus, afin de déterminer si les demi-vies des ARN non impactés ont un niveau similaire

aux demi-vies des ARN impactés lorsque la RNase J1 est absente, nous avons également comparé les demi-vies des deux groupes dans la souche J1^{AGA}.

Comme illustré dans la Figure 68A, dans la souche sauvage, les demi-vies des gènes non impactés ont tendance à être plus élevées que celles des gènes impactés. En d'autres termes, cela signifie que les gènes non impactés sont globalement plus stables que ceux impactés.

En revanche, et comme on pouvait s'y attendre, dans la souche mutante J1^{AGA}, les demi-vies des gènes impactés sont nettement plus longues que celles des gènes non impactés (Fig. 68B). Ces gènes ont été classés comme impactés car leur demi-vie a significativement doublé par rapport à celle observée chez la souche sauvage. Il est intéressant de noter que l'inactivation de la RNase J1 stabilise ces gènes non seulement par rapport à la souche sauvage, mais de manière encore plus marquée en comparaison avec les gènes non impactés. En d'autres termes, la stabilité des gènes impactés, et donc leur demi-vie, est significativement plus élevée que celle des gènes non impactés dans le mutant J1^{AGA}.

Ces observations soutiennent, d'une part, l'hypothèse selon laquelle la dégradation des gènes impactés est principalement, sinon exclusivement, réalisée par la RNase J1 et, d'autre part, que les gènes non impactés sont néanmoins dégradés par d'autres ribonucléases, puisque leur demi-vie reste inférieure à celle des ARN qui ne sont plus dégradés (gènes impactés dans le mutant J1^{AGA}).

La plupart des ARN non impactés a une demi-vie relativement courte, et certains d'entre eux (25 %) ont même une demi-vie inférieure à 3 min (Table 4 ci-après). Ces résultats confirment que la RNase J n'est pas la seule ribonucléase active chez *S. aureus,* et bien qu'elle impacte la demi-vie de beaucoup d'ARN, d'autres ribonucléases sont capables de dégrader rapidement les ARN.

Impact/Percentile	0 %	25 %	50 %	75 %	100 %
Impacté (937)	1,28	1,86	2,27	2,92	12,6
Non impacté (489)	1,46	3.05	4.26	6.10	30.0

Table 4 : Distribution des demi-vies (en minutes) des gènes par percentile en fonction des groupes impacté et non impacté.

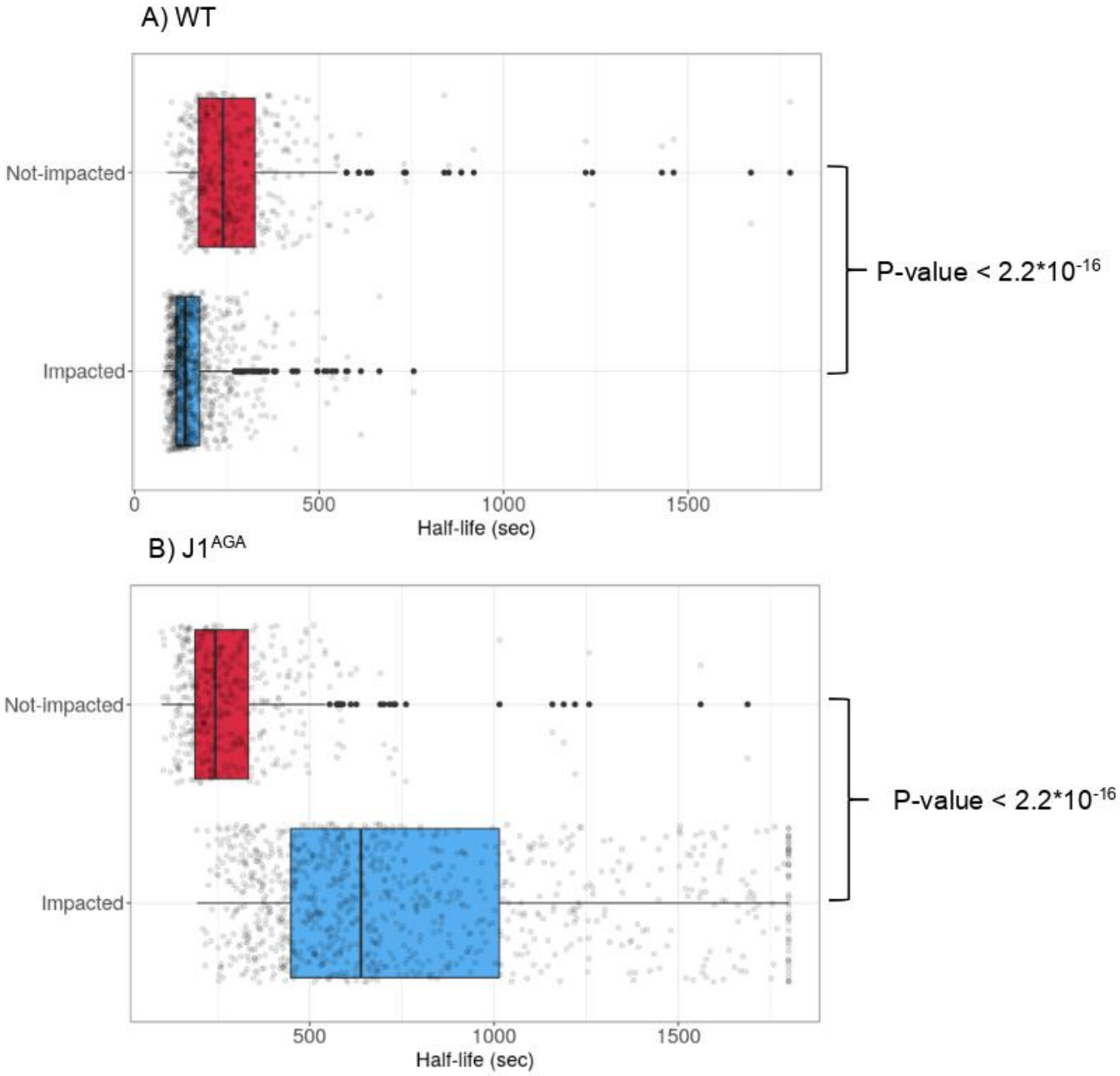


Figure 68 : Demi-vies des gènes en fonction de l'impact de la RNase J sur la stabilité de leur ARN. *Distribution des demi-vies (en secondes) des gènes par catégorie impactés et non impactés dans (**A**) la souche sauvage et dans (**B**) la souche mutante J1AGA.*

Dans une étude similaire réalisée sur l'endoribonucléase RNase Y, Khemici et al. (2015) ont observé 230 gènes dont la demi-vie et/ou l'expression augmentaient fortement par la délétion de cette enzyme. Parmi les 489 gènes non impactés par la RNase J1, il est intéressant de remarquer que 31 d'entre eux sont impactés par la délétion de la RNase Y Khemici et al. (2015).

La demi-vie des gènes semble donc pouvoir être en mesure d'identifier des liens entre une RNase et des ARN mais cette relation pourrait être indirecte. Dans la partie qui suit, nous allons chercher à déterminer les cibles directes de la RNase J1.

2.1.3 ARN ciblés directement par la RNase J1

Comme évoqué précédemment, les ARN qualifiés de non impactés pourraient tout même être dégradés par la RNase J1. En effet, la RNase J1 pourrait agir de manière redondante avec d'autres ribonucléases, de sorte que son inactivation serait compensée par l'action d'autres enzymes de dégradation. Ceci expliquerait que la demi-vie de certains ARN ciblés par la RNase J1 reste courte même en cas de délétion de cette dernière. De plus, il est possible que la RNase J1 ne dégrade l'ARN de certains gènes que partiellement. Ainsi, pour le calcul des demi-vies, il y aurait néanmoins suffisamment de lectures alignées sur le gène ciblé pour le considérer comme non impacté avec les seuils utilisés. Pour finir, des mécanismes de régulation compensatoires pourraient être activés en réponse à la délétion de la ribonucléase, comme l'augmentation de l'expression d'autres ribonucléases par exemple, maintenant ainsi la demi-vie de l'ARN au même niveau dans la souche $J1^{AGA}$.

2.1.3.1 ARN avec lesquels la RNase J1 interagit

La RNase J1 interagit physiquement avec les ARN qu'elle dégrade. Il est possible de fixer ces interactions par des méthodes expérimentales afin d'identifier les ARN qui seraient dégradés directement par la RNase J1, et non pas seulement impactés indirectement par la délétion de celle-ci. Le protocole CRAC (détaillé en introduction) est une approche qui pourrait donc permettre de répondre à ces incertitudes. En effet, il consiste à fixer les interactions protéines-ARN, puis à dégrader les parties des ARN non protégées par des protéines liées. Les ARN sont ensuite purifiés et séquencés, et les lectures résultantes correspondent donc aux parties des ARN qui étaient liées à une protéine. Afin de ne garder que les ARN liés à une protéine en particulier, la protéine d'intérêt est étiquetée, ce qui permet de purifier uniquement les complexes protéine-ARN dans lesquels cette protéine d'intérêt est impliquée.

Nous avons pu disposer de données générées par ce protocole pour la RNase J1 dans la souche JKD6009 de *S. aureus* qui est assez proche de la souche SA564. Ces données, encore non publiées, nous ont été fournies par Jay Tree (UNSW – Equipe Tree) et son équipe, et ont été produites dans un projet réalisé portant également sur d'autres organismes et s'intéressant aux interactions ARN-protéine.

À partir de ces données, nous devrions donc être en mesure de déterminer, pour certains gènes, si leurs ARN sont directement ciblés par la RNase J1. Malheureusement, la réciproque n'est pas vraie, c'est-à-dire que l'absence d'observation d'interaction RNase J-ARN n'implique pas que la RNase J n'interagit pas avec ces ARN. En effet, il se peut que dans les conditions de l'expérience, certains ARN ne sont pas ou pas suffisamment exprimés (selon la profondeur de séquençage) pour être détectés.

Dans ce qui suit, nous allons tout d'abord détailler les données fournies, puis le traitement effectué pour déterminer les cibles directes de la RNase J1. Ensuite, nous mettrons en relation ces résultats avec les résultats obtenus sur la classification des gènes considérés impactés ou non impactés par la RNase J. Pour cela, nous effectuerons l'identification des gènes orthologues dans les deux souches considérées : SA564 et JKD6009.

Les données fournies consistent en des librairies de séquençage Illumina avec trois répétitions pour l'obtention des régions en interaction avec la RNase J1, et trois répétitions pour des échantillons de contrôle. Dans les échantillons de contrôle, aucune protéine n'a été étiquetée. Ceci permet de s'assurer que seules les régions des ARN liés à une protéine étiquetée dans la condition d'intérêt peuvent être séquencées. Bien que la technologie utilisée permette le séquençage de 150 nucléotides, la plupart des lectures fait entre 20 et 75 nucléotides de long.

Les données brutes correspondent donc à 6 fichiers au format FASTQ contenant les lectures issues des différents échantillons contrôles et « RNase J1 étiquetée ».

Pour le pré-traitement et la quantification des fichiers FASTQ, nous avons appliqué la méthode décrite par (Sy et al., 2018) comme suit :

- Les adaptateurs ont été supprimés et la qualité des lectures a été filtrée à l'aide programme Flexbar (Dodt et al., 2012)**,** (Roehr et al., 2017)**.**
- Les lectures ont été alignées sur le génome de référence de *S. aureus* JKD6009 (NZ_LR027876.1) avec bowtie2 en autorisant 1 mésappariement et en ne gardant que les alignements uniques. L'annotation RefSeq de ce génome indique qu'il comporte 2778 gènes codant pour des protéines.

– Les lectures ont été quantifiées par gène à l'aide du programme pyreadcounters, inclus dans la suite de programmes pyCRAC (Webb et al., 2014). Afin d'éliminer les duplicats PCR, nous avons utilisé l'option « --block », qui permet de ne compter qu'une seule fois les lectures s'alignant exactement au même endroit.

Le nombre de lectures contenues dans chacune des librairies est donnée dans la Table 5. Le peu de lectures obtenues dans les échantillons de contrôle semble déjà indiquer que celles obtenues dans les échantillons pour la RNase J1 seraient spécifiques des interactions RNase J1-ARN. En effet, les tailles de librairies montrent que les échantillons de contrôle ont très peu de lectures alignées. Ce sont des résultats positifs, car ils indiquent que peu de lectures sont séquencées lorsque aucune protéine n'est étiquetée, et donc que les lectures présentes dans les échantillons « RNase J étiquetée » correspondent bien, pour la plupart, à des interactions RNase J-ARN comme indiqué dans la Table 5 ci-dessous.

Échantillon	Taille de la librairie	Nombre de lectures alignées	Nombre de lectures dédupliquées
RNase J taguée réplicat 1	7 436 129	5 491 593	45 846
RNase J taguée réplicat 2	10 877 019	10 094 826	218 711
RNase J taguée réplicat 3	331 671 534	27 254 737	98 927
Contrôle réplicat 1	4 214 696	2 452	118
Contrôle réplicat 2	11 896	3 658	99
Contrôle réplicat 3	11 850	4 432	215

Table 5 : Nombre de lectures totales, nombre de lectures alignées et nombre de lectures dédupliquées dans chacun des 6 échantillons.

Afin d'éliminer le bruit de fond pouvant correspondre à des ARN non protégés mais non dégradés dans la mise en œuvre du protocole expérimental, nous avons conservé uniquement les gènes dont au moins 5 lectures étaient alignées après élimination des duplicats de PCR. Ensuite, nous avons normalisé les comptages avec la méthode TPM afin de tenir compte, tout d'abord de la longueur des gènes, et ensuite de la taille de la librairie, pour être en mesure de comparer les comptages entre gènes au sein d'un même échantillon, et également entre les différents échantillons pour un

même gène ou des gènes différents. Une fois cette normalisation effectuée, nous avons résumé les répétitions en prenant la médiane des comptages pour chaque gène.
Après les différentes étapes de normalisation et de filtrage précédentes, nous avons considéré que les ARN d'un gène étaient ciblés par la RNase J1 si sa valeur était supérieure ou égale à 20 TPM.

Avec ces paramètres, sur les 2778 gènes codants, nous pouvons identifier 1915 gènes ciblés par la RNase J1, soit 68,9 %. Par conséquent, 863 gènes sont classés comme non ciblés. Ces résultats semblent indiquer que la RNase J serait capable de dégrader les ARN de davantage de gènes que ce que nous avions pu identifier dans la section précédente en considérant la variation de leur demi-vie chez le mutant $J1^{AGA}$, c'est-à-dire qu'elle ciblerait presque 70 % des gènes, et non pas seulement les 47 % de gènes impactés identifiés précédemment.
Afin de faire le lien avec les gènes impactés précédemment identifiés, nous avons réalisé l'identification des orthologues entre les deux souches comme suit, à l'aide du programme Panaroo (Tonkin-Hill et al., 2020) qui permet de construire un graphe de pangénome bactérien en utilisant une méthode de clustering pour regrouper les gènes similaires entre les différents génomes.
Nous avons ainsi pu identifier 2357 gènes orthologues entre les 2778 gènes de la souche JKD6009 et les 2703 gènes de la souche SA564.
Parmi les 1 989 gènes dont les durées de demi-vies étaient exploitables pour déterminer une variation significative, 1 866 ont un orthologue identifié chez JKD6009.Nous disposons donc de 1 866 gènes pour lesquels une demi-vie est exploitable et un orthologue est identifié entre les deux souches. Ces informations sont résumées dans la Table 6 ci-après.

	JKD6009	Orthologues SA564	Orthologues SA564 avec demi-vie
Gènes totaux	2778	2357	1866
Cibles RNase J1	1915	1736	1590
Non cibles RNase J1	863	621	276

Table 6 : Nombre de gènes identifiés comme ciblé ou non ciblé par la RNase J1 en fonction des gènes orthologues dont nous avons pu calculer une demi-vie chez SA564.

Les 1915 gènes cibles identifiés sur JKD6009 ont ainsi 1736 orthologues, dont 1590 ont une demi-vie exploitable. Les 863 gènes non cibles identifiés chez JKD6009 ont 621 gènes orthologues, et seulement 276 de ces gènes ont une demi-vie exploitable.

La Table 7 de contingence ci-dessous résume la répartition de ces 1866 gènes entre les catégories impacté/non déterminé/non impacté, et les catégories cibles/non cibles.

	Cible	Non cible	Total
Impacté	833 (44,6 %)	80 (4,3 %)	913 (48,9 %)
Non déterminé	480 (25,7 %)	89 (4,6 %)	564 (30,5 %)
Non impacté	277 (14,8 %)	107 (5,7 %)	384 (20,6 %)
Total	1590 (85,2 %)	276 (14,8 %)	1866 (100 %)

Table 7 : Table de contingence représentant les intersections entre les groupes de gènes impactés et non impactés avec les groupes de gène ciblés et non ciblés.

Dans ces résultats, nous pouvons remarquer que presque tous les gènes impactés sont également ciblés par la RNase J1 (833 sur 913, soit 91,2 %). Ceci vient soutenir l'hypothèse que les gènes dont la stabilité augmente significativement dans la souche $J1^{AGA}$ sont des cibles directes de la dégradation par la RNase J1.
En revanche, comme il avait été suggéré plus tôt, on remarque que les gènes non impactés peuvent tout de même être ciblés par la RNase J1 (277 sur 384, soit 72,1 %). Ceci vient soutenir l'hypothèse qu'en l'absence de la RNase J1, la dégradation des ARN correspondant à ces gènes serait prise en charge par une autre ribonucléase qui reste à déterminer.
Au total, parmi les 1 866 gènes considérés, 1 590 gènes (soit 85,2%) sont ciblés par la RNase J1.

Le fait que nous n'observons pas de comptages avec le protocole CRAC sur certains ARN dont la demi-vie est néanmoins impactée par la délétion de la RNase J pourrait s'expliquer par le fait que la RNase J aurait un effet indirect sur la demi-vie des ARN, et que ces ARN ne seraient donc pas dégradés par cette dernière. Par ailleurs, les conditions de culture ainsi que les souches étudiées sont différentes, et ceci pourrait expliquer des différences sur les observations effectuées sur la demi-vie des ARN d'une part, et sur leurs interactions avec la RNase J1 d'autre part. Enfin, la profondeur de séquençage est relativement faible après élimination des duplicats de PCR, et il est possible d'avoir des faux négatifs, c'est-à-dire des gènes classé non cibles alors qu'ils sont en réalité cibles.

La normalisation des niveaux d'expression pour calculer les demi-vies a été effectuée avec le gène *hu*, car celui-ci est très stable. Ceci suppose qu'il n'est pas ciblé par la RNase J, ou tout du moins, pas dégradé par cette dernière. Nous avons donc voulu vérifier, dans les données générées avec le protocole CRAC, que nous n'observions pas de lecture sur celui-ci. Or, nous observons une médiane de 20 comptages sur ce gène, la plupart située à environ 150 nucléotides du TSS, indiquant que la RNase J interagit avec les ARN de *hu*. Ceci pourrait suggérer que les transcrits du gène *hu* sont tout de même dégradés mais faiblement, car sa demi-vie est très stable (> 1800 s). Une autre hypothèse serait que la RNase J puisse interagir avec des ARN sans pour autant les dégrader. Ces interactions pourraient correspondre à la RNase J qui entre en interaction avec un ARN mais qui ne réalise pas la dégradation, par exemple via un domaine d'interaction situé en dehors de son site actif.

Pour conclure sur cette partie, les données obtenues avec le protocole CRAC sur la RNase J1 permettent de compléter les données de demi-vie en indiquant que la RNase J1 pourrait interagir avec les ARN de près de 85 % des gènes de *S. aureus*, renforçant encore l'hypothèse qu'elle serait l'actrice principale de la dégradation des ARN chez cette bactérie.

2.1.3.2 La RNase J1 impacte et cible moins souvent le premier gène d'un opéron que les suivants

Dans les parties précédentes, nous avons pu déterminer si le niveau d'expression des gènes était modifié ou non par la mutation du site actif de la RNase J1 *via* l'étude de la stabilité des ARN messagers en considérant les lectures RNA-Seq alignées sur les gènes codant pour des protéines. Nous avons également étudié si la RNase J1 interagissait avec ces ARN.

Dans cette partie, nous allons nous intéresser à l'impact de la RNase J1 sur l'expression des gènes et à sa modulation, notamment *via* la stabilité des ARN, au sein des unités de transcription. En effet, certains ARN sont polycistroniques, c'est-à-dire qu'ils arborent plusieurs cadres ouverts de lecture correspondant à plusieurs gènes co-transcrits sur le même ARN. Les gènes co-transcrits sont souvent impliqués dans un même processus biologique. Le fait d'être portés par un même transcrit pourrait présenter certains avantages : concernant la régulation de leur expression, concernant la stœchiométrie de leur produit ainsi que leur localisation, ou encore concernant leur transmission verticale ou horizontale au cours de l'évolution. Malgré ces avantages, et notamment celui de la co-régulation, il a été observé pour de nombreux systèmes que les gènes d'une même unité de transcription peuvent néanmoins avoir des niveaux d'expression différents (en termes de nombre de copies d'ARNm).

Historiquement, un ARNm polycistronique sous le signal d'un même promoteur pour leur transcription est appelé un opéron. Initialement, les gènes d'un même opéron étaient décrits comme participants à une même voie métabolique. Cette définition s'est révélée parfois inappropriée. Certains opérons pouvant inclure des gènes impliqués dans des processus biologiques distincts. De même, leur régulation sous un même *opérateur* a été remise en question, et de fait, la définition d'un opéron a évolué. Par exemple, il a également été proposé d'appeler opéron, une unité de transcription arborant un seul gène, autrement dit, un opéron monocistronique.

Dans ce qui suit, nous nous intéressons aux gènes co-transcrits et à leur stabilité sous l'effet de la RNase J1. Nous avons donc opté pour la définition d'opéron suivante : ensemble de un ou plusieurs gènes pouvant être co-transcrits sur un même ARN. Cette définition relativement générale nous permet de simplifier leur analyse, et de nous placer dans un cadre assez général portant sur des gènes co-transcrits. Bien sûr, la réalité biologique est plus complexe, de par l'existence de différents départs de transcription produisant des ARN portant les mêmes gènes. De même, des départs alternatifs de transcription peuvent être localisés sur ou après le premier gène d'un opéron, produisant des transcrits portant un sous-ensemble des gènes de l'opéron. Par ailleurs, bien que co-transcrits sur le même ARN, les gènes de l'opéron peuvent aussi présenter des différences d'expression ou de stabilité causées par une dégradation partielle des ARN résultant soit d'une dégradation exoribonucléolytique arrêtée par une structure secondaire ou un autre facteur, soit d'une coupure de l'ARN par l'action d'une endoribonucléase.

La RNase J étant principalement caractérisée par son activité 5'→3' exoribonucléasique, nous avons entrepris d'examiner si le premier gène des opérons pouvait montrer des variations de stabilité chez le mutant RNase J1, c'est-à-dire de demi-vie, différentes de manière statistiquement significative par rapport aux autres gènes situés en aval.

Afin d'identifier le premier gène des opérons, nous avons tout d'abord annotés les opérons à partir d'une prédiction. Ensuite, nous avons comparé les demi-vies et leurs variations avec le test de Wilcoxon-Mann-Whitney effectué entre les gènes en première position et ceux en position ultérieure dans les opérons. A partir de cette classification, premier gène ou pas ainsi que gène impacté ou pas et gène ciblé ou pas, nous avons effectué un test d'indépendance du χ^2 afin de déterminer s'il existe un lien entre le fait qu'un gène est caractérisé comme impacté ou non impacté, et ciblé ou non ciblé, et le fait qu'il soit en première position ou pas dans l'opéron.

Pour la prédiction des opérons, nous avons utilisé OperonMapper (Taboada et al., 2018)**.** Nous avons ainsi identifié 1 505 opérons dans le génome de *S. aureus* (souche SA564). La majorité de ces opérons (66 %) ne contient qu'un seul gène, et 520 opérons (1 718 gènes) ont été identifiés comme contenant plus d'un gène. (Fig. 69).

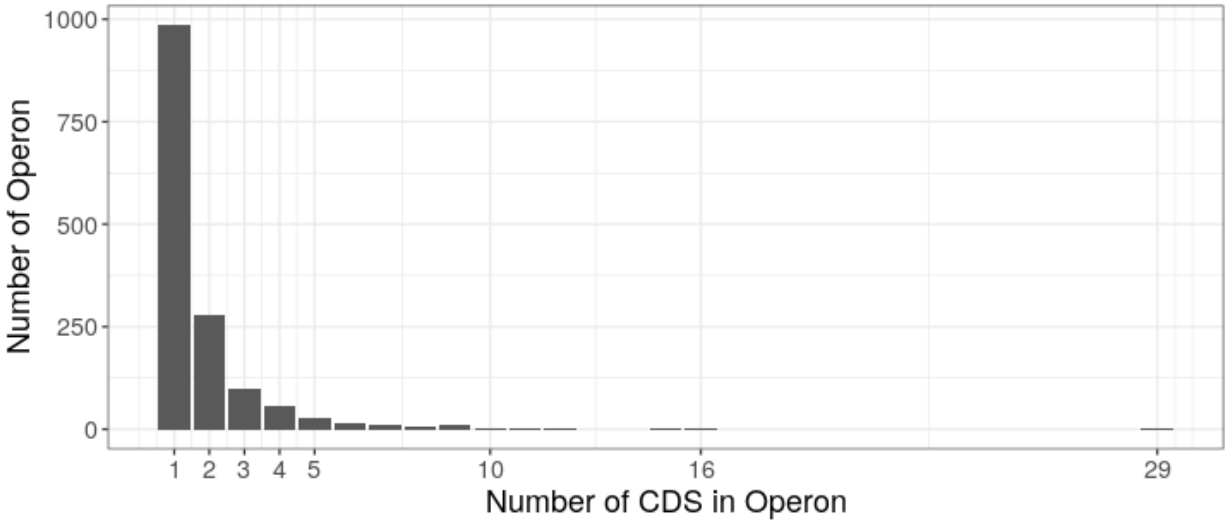


Figure 69 *:* ***Taille des opérons.*** *Histogramme représentant la distribution du nombre de gènes dans les opérons*

Parmi ces 1 505 opérons :

- 224 ne contiennent que des gènes non impactés,
- 394 ne contiennent que des gènes impactés,
- 881 ne contiennent que des gènes ciblés,
- 266 ne contiennent que des gènes non ciblés,
- 344 ne contiennent que des gènes ciblés et impactés,
- 46 ne contiennent que des gènes non ciblés et non impactés.

Parmi les 1 359 gènes ayant pu être classés impacté ou non impacté, 700 sont en première position de l'opéron et 659 dans les suivantes.
Concernant les 422 gènes non impactés, 274 (64,9 %) d'entre eux sont en première position (en comptant ceux ne contenant qu'un seul gène), tandis que parmi les 937 gènes impactés, 426 (45,5 %) sont en première position. Ceci indique que les gènes caractérisés comme étant non impactés sont majoritairement en première position contrairement aux gènes impactés. De plus, sur les 659 gènes n'étant pas en première position, 511 sont impactés (77,5 %). Ceci suggère que les gènes qui ne sont pas en première position sont davantage impactés par la RNase J que ne le sont les gènes en première position.

Nous pouvons donc remarquer que les gènes non impactés sont plus souvent en première position que ne le sont les ARN impactés. Ceci est illustré en Figure 70 montrant que 64,9 % des gènes non impactés sont premiers tandis que seulement 45,5 % des gènes impactés le sont.
De plus, nous observons que 60,7 % des gènes non ciblés et non impactés par la RNase J sont premiers tandis que seulement 45,7 % des gènes ciblés et impactés le sont. Ces résultats montrent une légère tendance de la RNase J qui impacte davantage et qui cible préférentiellement les gènes ne se trouvant pas en première position dans un opéron.

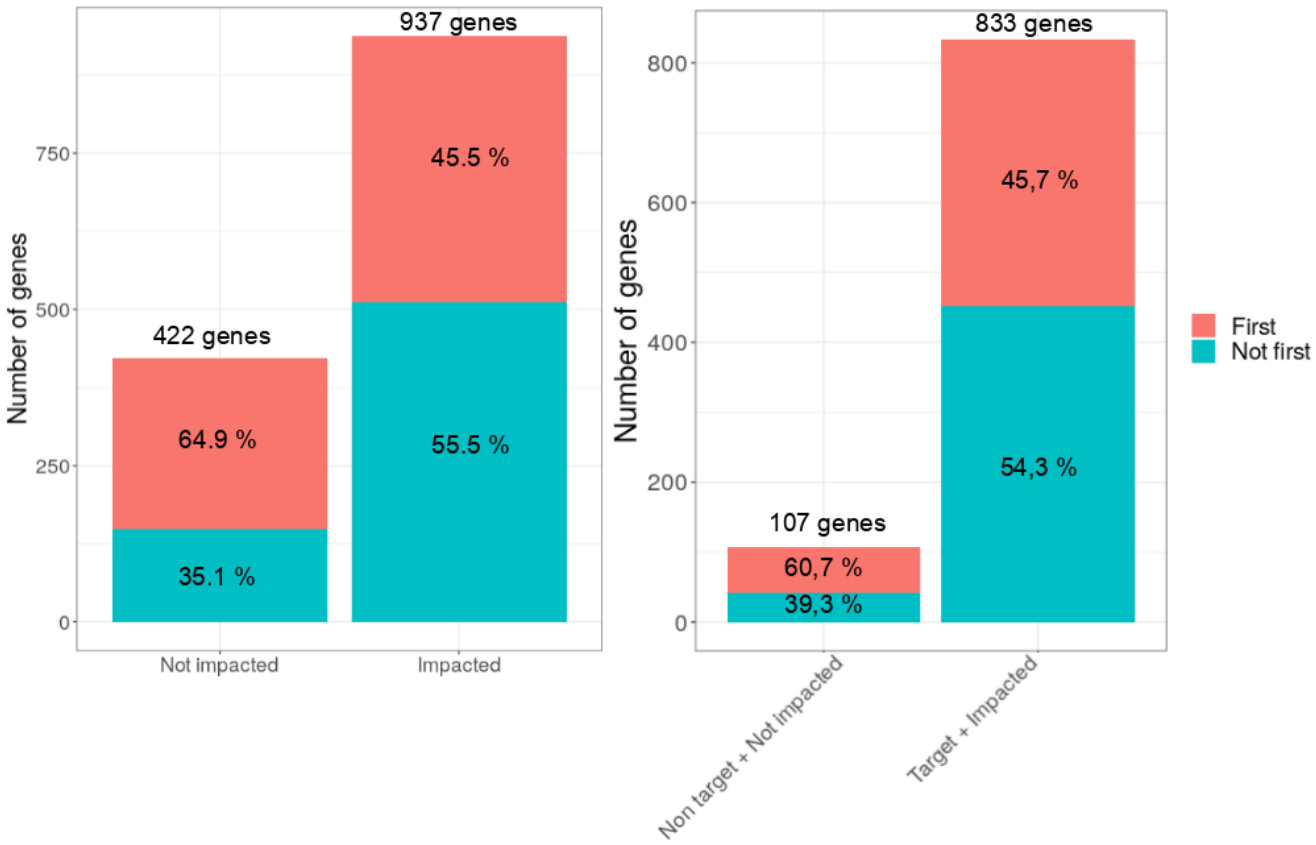


Figure 70 : Répartition des gènes selon leur position dans les opérons et leur impact.

Nombre de gènes en première position d'un opéron ou dans les suivantes en fonction des groupes non impactés et impactés (gauche) et des groupes non impactés non ciblés et impactés ciblés (à droite).

Ces observations sont synthétisées dans les tables de contingence présentées en Table 8. Les tests de X^2 effectués obtiennent respectivement une p-valeur de $1{,}9.10^{-14}$ et une p-valeur de $4{,}7.10^{-3}$. Il n'y a donc pas indépendance entre ces deux facteurs et cela indique que la répartition observée est significativement différente de celle attendue sous l'hypothèse d'indépendance des variables. Ces résultats confirment les observations et indiquent que la RNase J impacte davantage et cible préférentiellement les gènes qui ne sont pas en première position dans l'opéron.

	Premier	Non premier	Total
Impacté	426	511	937
Non impacté	274	148	422
Total	700	669	1359

	Premier	Non premier	Total
Ciblé + Impacté	381	452	833
Non ciblé + Non impacté	65	42	107
Total	517	423	940

Table 8 : ***Tables de contingence représentant les intersections*** *entre :*

- *les groupes de gènes impactés et non impactés avec les groupes de gène premiers et non premiers (haut).*
- *les groupes de gènes impactés ciblés et non impactés non ciblés avec les groupes de gène premiers et non premiers (haut).*

De manière plus générale, la Figure 71 illustre les comparaisons réalisées à l'aide d'un test de Wilcoxon-Mann-Whitney. Les résultats montrent que les gènes en première position d'un opéron ont une demi-vie significativement plus longue (p-valeur = $2{,}8.10^{-10}$) et une variation de demi-vie suite à la mutation du site actif de la RNase J1 significativement plus courte (p-valeur < $2{,}2.10^{-16}$). Ceci confirme donc que la RNase J1 impacte moins les gènes en première position d'un opéron qui sont également plus stables que les gènes n'étant suivants dans l'opéron.

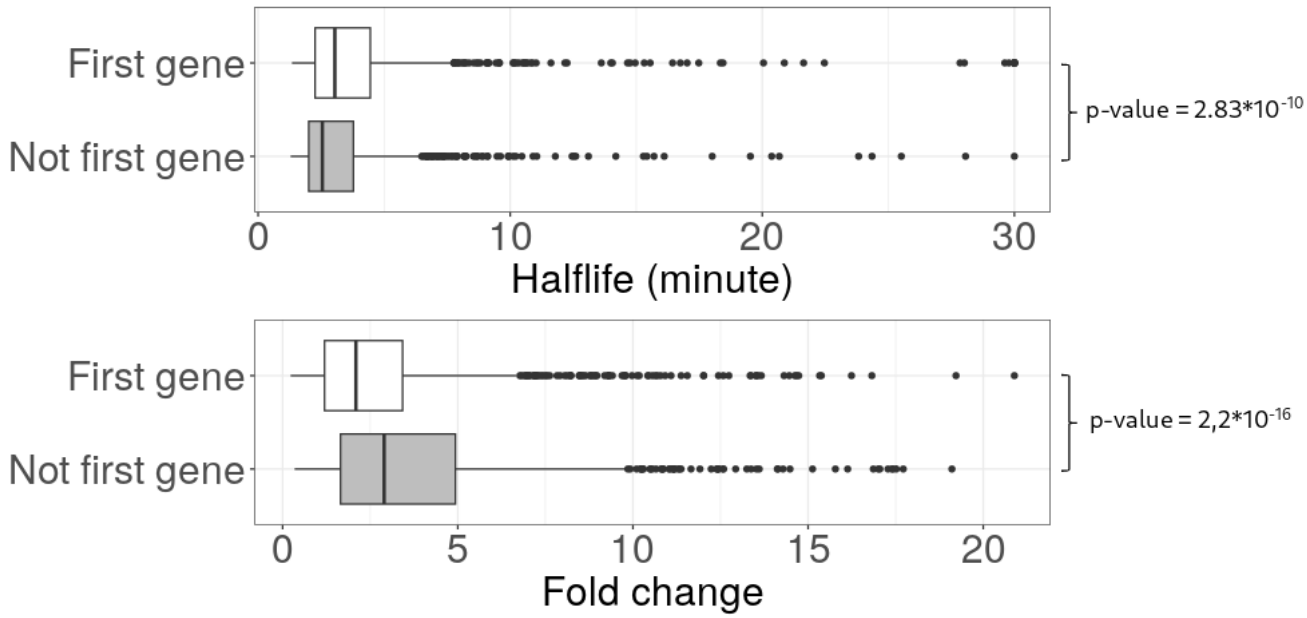


Figure 71 : ***Demi-vie et variation de demi-vie des gènes en fonction de leur place dans l'opéron.*** *Distribution des demi-vies en log_2 (en haut) et des variations de demi-vie (en bas), des* premiers et non premiers gènes d'op*éron. Les p-valeurs ont été obtenues avec un test de Wilcoxon-Mann-Whitney.*

Il convient de noter qu'en raison de la présence de TSS alternatifs, par exemple sur ou après le premier gène de l'opéron, certains gènes considérés dans cette analyse comme non premiers peuvent tout de même être situés en première position d'un transcrit nouvellement synthétisé.

Étant donné que la RNase J a une activité exoribonucléasique 5'→3', on pourrait s'attendre à ce qu'elle impacte principalement le premier gène des opérons puisqu'il est plus proche de l'extrémité 5'. Nous pouvons observer que ce n'est pas le cas pour 274 des 700 gènes en première position considérés. Une hypothèse pour expliquer ce phénomène pourrait être que des facteurs protègent les extrémités 5' des ARN, empêchant ainsi sa dégradation par la RNase J, tandis que la présence de sites de clivage endoribonucléolytique sur les transcrits permettrait néanmoins la dégradation des gènes qui ne sont pas en première position.

Dans la prochaine section, nous examinerons de plus près les extrémités 5' des transcrits, tandis qu'une section ultérieure présentera une étude préliminaire sur les sites de clivage endoribonucléolytique de la RNase J.

2.1.3.3 Aucune caractéristique commune n'a été identifiée entre les extrémités 5' des ARN impactés ou celles des ARN non impactés

Dans la partie précédente, nous avons vu que, bien que la RNase J module l'expression génique de près de la moitié des gènes de *S. aureus*, son activité 5' exoribonucléasique ne semble pas cibler directement ou préférentiellement l'extrémité 5' des opérons. Nous avons pu notamment observer que la stabilité des ARN du premier gène d'un opéron pouvait dans certains cas être impactée par l'absence de la RNase J, et dans d'autres non, et que cette dernière ciblait autant, si ce n'est davantage, les autres gènes de l'opéron. Dans cette partie, nous allons nous intéresser aux caractéristiques des extrémités 5' des ARN pouvant, d'une part, favoriser leur dégradation par l'activité 5' exoribonucléasique de la RNase J, ou bien, d'autre part, les en protéger.

La première hypothèse que nous pouvons formuler est que l'extrémité 5' tri-phosphate des ARN nouvellement synthétisés les protégerait de leur dégradation et qu'une déphosphorylation serait nécessaire pour que la RNase J puisse opérer. Dans une première partie, nous allons étudier s'il existe un lien entre l'état de phosphorylation de l'extrémité 5', c'est-à-dire la proportion d'extrémités 5'P par rapport à celle de 5'PPP, et la demi-vie des ARN.

Une autre possibilité concernant la protection des ARN et pouvant expliquer leur stabilité serait la présence de certains signaux ou propriétés au niveau de la séquence en ribonucléotides de leur extrémité 5'. Dans cette optique, nous avons recherché des motifs conservés, d'une part, sur les séquences des extrémités impactées et/ou ciblées pouvant favoriser leur dégradation, et d'autre part, sur les extrémités non impactées et/ou non ciblées pouvant les protéger de l'action de la RNase J. De plus, la formation de structures stables pourrait aussi empêcher l'action de la RNase J. Nous avons donc recherché des indices de formation de telles structures *via* la recherche d'un enrichissement en ribonucléotides G ou C pouvant former des structures stables, ainsi que *via* la prédiction de structures secondaires au niveau des extrémités 5'.

Dans un premier temps, afin d'étudier les extrémités 5' des ARN, il est nécessaire de les répertorier dans la souche étudiée et de les associer aux gènes transcrits pour mettre en relation leurs caractéristiques avec leur stabilité et/ou sa variation chez le mutant RNase J.

L'annotation du génome de la souche étudiée, SA564, ne contient pas les TSS. En revanche, Prados et al. (2016) ont réalisé une étude pour leur identification sur une souche très proche, MW2. Au cours de cette étude, il a notamment été mis au point un protocole expérimental, TSS-EMOTE, et une méthode statistique d'identification des TSS est également proposée, comme présenté dans la partie introduction. Pour rappel, cette méthode est basée sur le séquençage spécifique des extrémités des ARN.

Un échantillon contenant les ARN 5'PPP est alors comparé statistiquement avec un même échantillon biologique (un échantillon « contrôle ») dans lequel les ARN 5'PPP ne devraient pas pouvoir être séquencées.
Dans un premier temps dans cette partie, nous allons donc réaliser l'annotation des TSS dans la souche SA564. Pour cela, nous avons effectué un transfert des annotations des TSS de la souche MW2 vers la souche SA564 en identifiant les régions orthologues de leur génome.
Les deux souches considérées étant assez proches (98,9 % d'identité et 94,5 % de couverture), nous pouvons nous attendre à un pourcentage d'identités de séquence très élevé entre les TSS orthologues. Pour identifier les TSS orthologues, nous avons donc aligné les 100 premiers nucléotides des TSS identifiés dans la souche MW2 sur le génome de la souche SA564.

Pour rappel, pour les différentes copies ou transcrits d'un gène nécessaire à son expression, les départs de transcription peuvent varier de quelques nucléotides autour d'un site majoritaire de départ de transcription. On appelle « clusters de TSS », les groupes de TSS très proches les uns des autres, généralement à quelques nucléotides de distance (Fig. 72). Ces clusters résultent de l'ARNP démarrant la transcription à des points légèrement différents à l'intérieur d'une petite région. Ceci est souvent dû à des variations dans l'efficacité de l'initiation de la transcription et des fluctuations dans la reconnaissance des sites de démarrage par l'ARNP.
Un gène n'est pas forcément associé à un seul cluster de TSS, car plusieurs de ces clusters peuvent être le site d'initiation de la transcription. On parlera alors de TSS alternatifs (Fig. 72). Ces TSS alternatifs correspondent aux différents sites de démarrage de la transcription pour un même gène, permettant la production de différentes isoformes d'ARN. Ces sites peuvent être séparés, chez les bactéries, par plusieurs dizaines de nucléotides et sont régulés par des mécanismes comme les facteurs de transcription, les modifications épigénétiques, et les conditions cellulaires ou environnementales. Les TSS alternatifs permettent une diversité dans les produits transcrits et donc une régulation fine de l'expression génique.

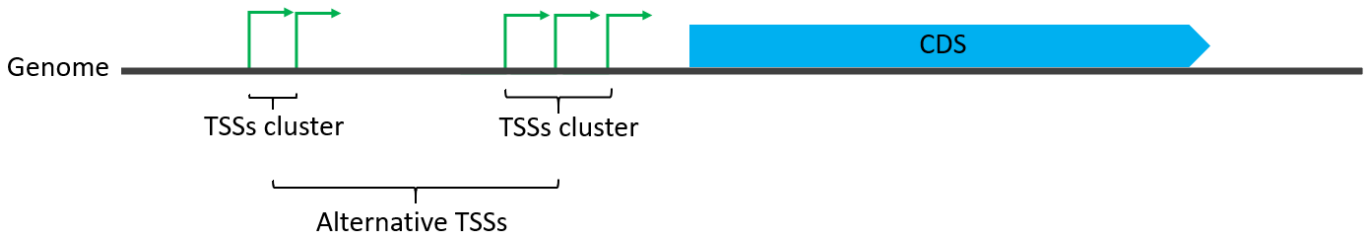


Figure 72 : Départs alternatifs de transcription. *Représentation schématique d'un gène possédant deux clusters de TSS alternatifs contenant respectivement deux et trois TSS.*

Au total, nous avons identifié 3 575 TSS dans la souche SA564 orthologues avec ceux de la souche MW2, ce qui correspond à 2 395 clusters de TSS, c'est-à-dire des ensembles de TSS distants les uns des autres de moins de 10 nucléotides.

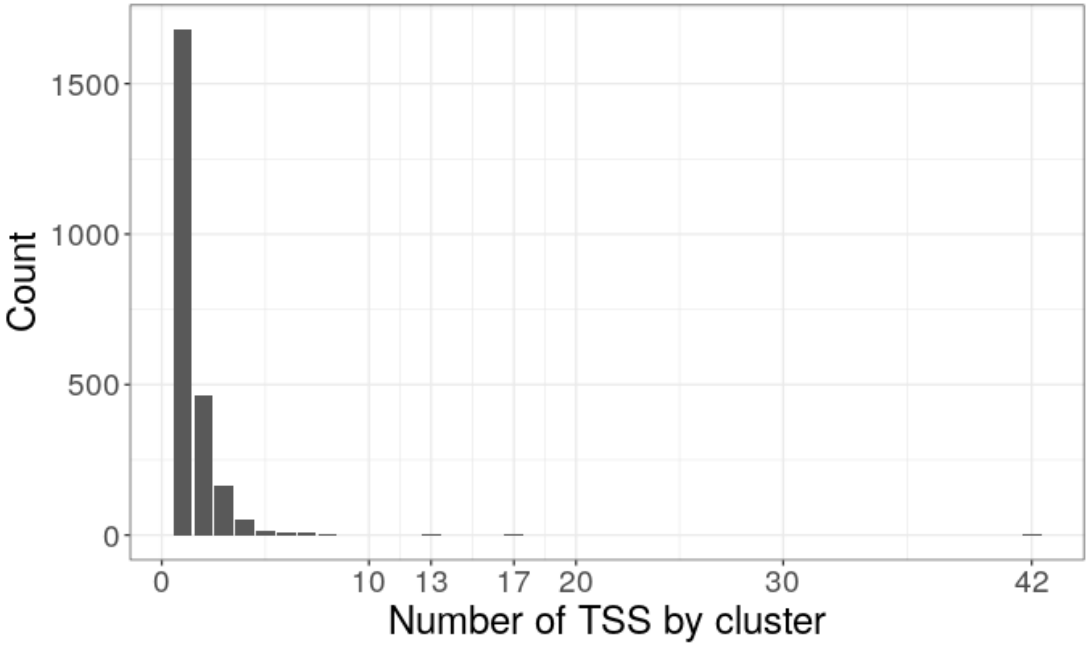


Figure 73 : ***Taille des clusters de TSS.*** *Histogramme du nombre de TSS par clusters dans la souche SA564* de *S. aureus.*

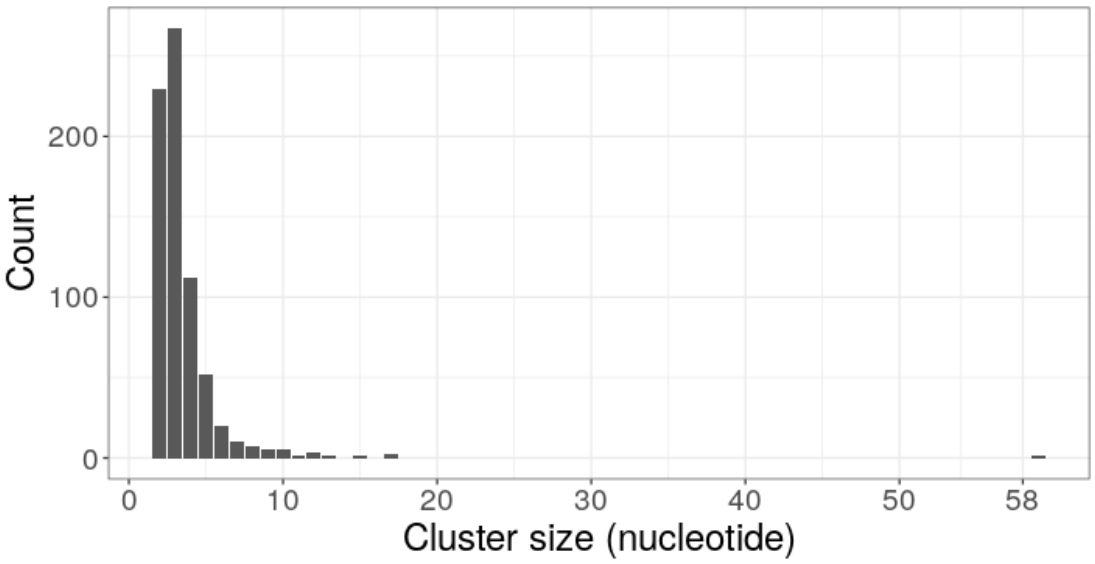


Figure 74 : ***Largeur des clusters de TSS.*** *Histogramme des tailles des clusters de TSS (en nucléotides), dans la souche SA564* de *S. aureus, pour ceux contenant plus d'un TSS.*

Étant donné que les 5'-UTR sont relativement courtes chez *S. aureus* (*Fig. 75*), nous avons associé aux premiers gènes des opérons, tous les TSS situés à moins de 500 nucléotides en amont du début du gène. Pour les autres gènes d'un opéron, nous avons associé les TSS du premier gène de l'opéron ainsi que les TSS identifiés entre le début du premier gène de l'opéron et le début de l'autre gène considéré. Au total, dans la souche SA564, nous avons pu associer au moins un TSS à 1951 gènes sur les 2703 du génome, et ainsi associer au moins un TSS à 1073 opérons sur les 1505 prédits précédemment.

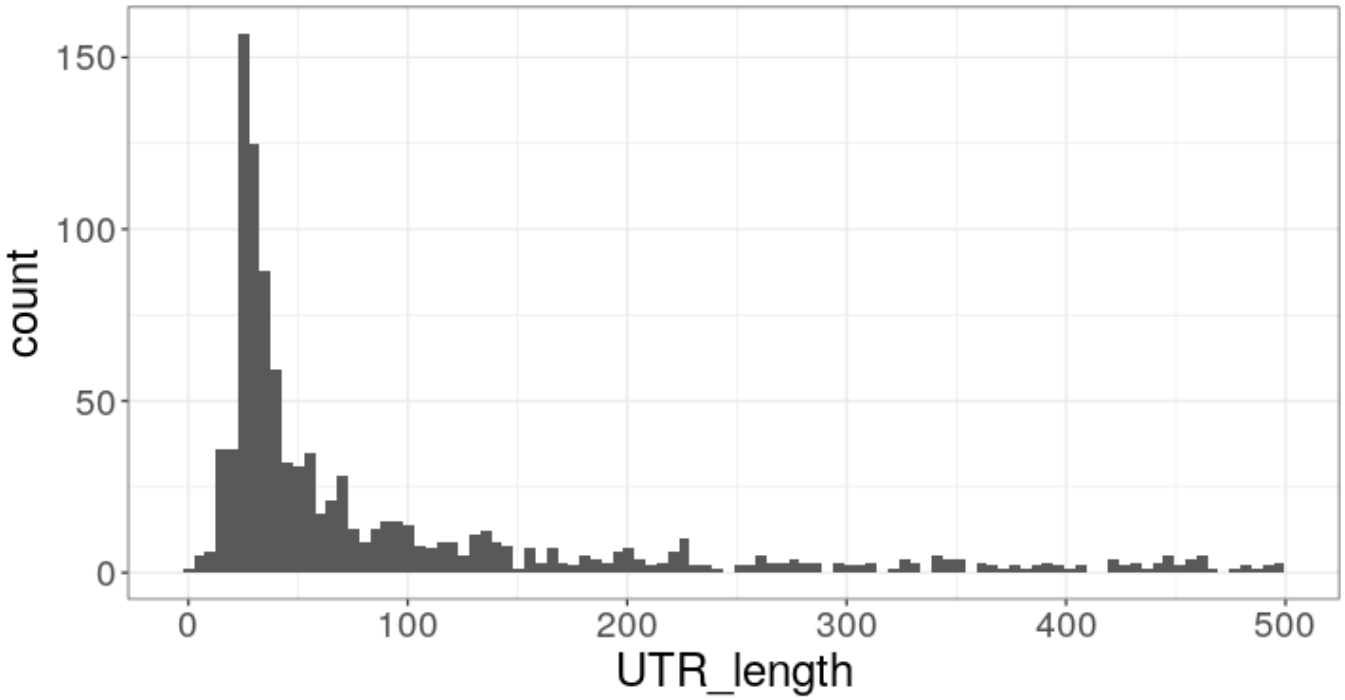


Figure 75 *:* ***Taille des régions 5' non traduites (5'-UTR).*** *Histogramme des tailles de 5'-UTR (en nucléotides)* dans la souche SA564 de *S. aureus.*

671 (34 %) gènes ne sont associés qu'à un seul TSS, et la majorité des 1 951 gènes (55 %) ne sont associés qu'à un seul cluster de TSS (Fig. 76).

La proportion de gènes associés à un TSS est de 1 951 sur 2 703, soit 72 %. Ce résultat n'est pas surprenant, car tous les gènes ne sont pas exprimés. Par exemple, pour le calcul des demi-vies, nous avions identifié 2 073 gènes exprimés sur les 2 703. Par conséquent, il n'y a que 322 gènes pour lesquels une demi-vie a été déterminée, mais dont aucun TSS n'a pu être identifié.

Une fois les extrémités 5’ des ARN annotées, nous avons pu étudier et comparer leurs caractéristiques entre les groupes impacté et non impacté, afin de tenter d’identifier les facteurs qui modulent l’activité de la RNase J.

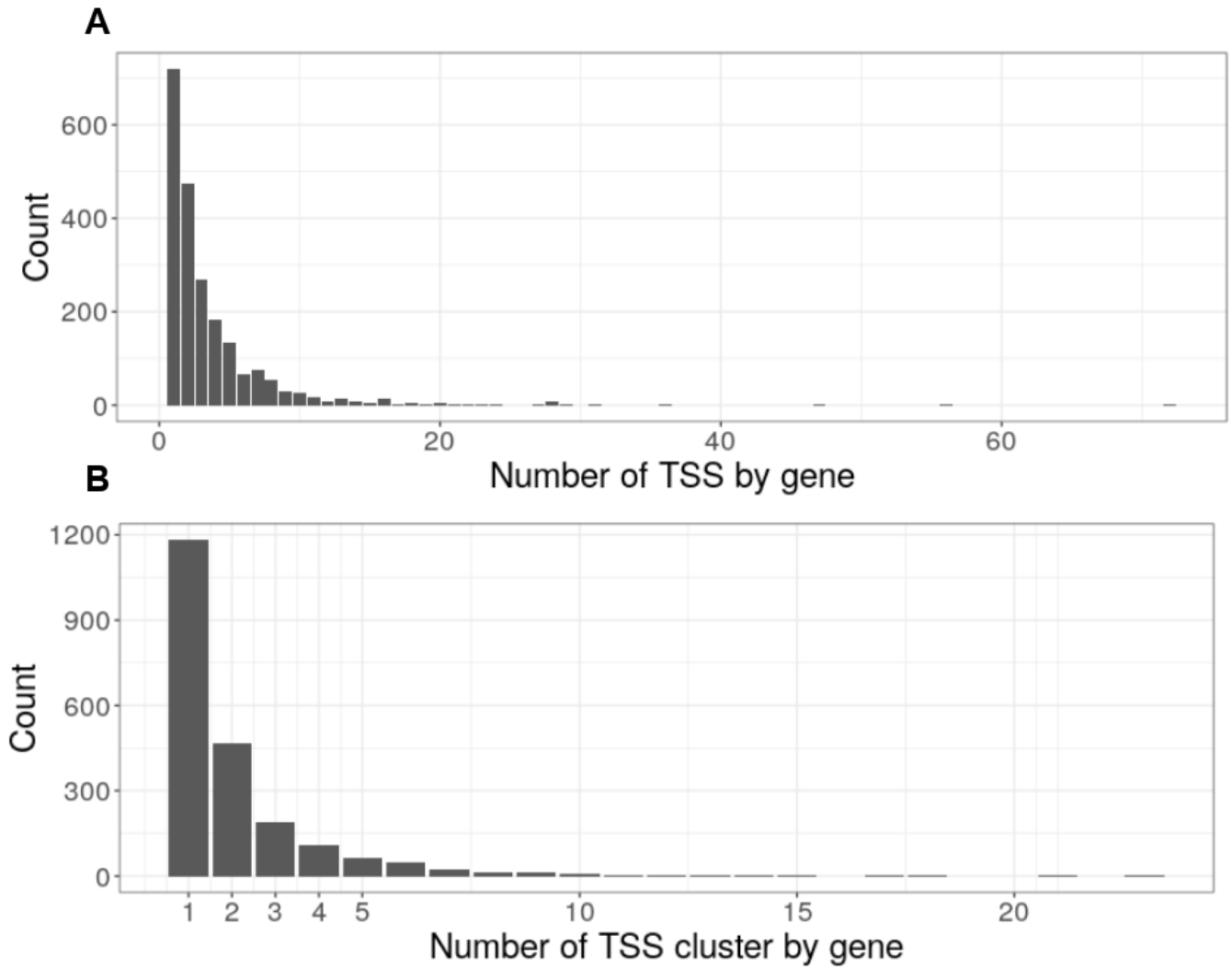


Figure 76 : ***Nombre de TSS associé à chaque gène.*** *Histogrammes des distribution (**A**) du nombre de TSS associés par gène et (**B**) du nombre de cluster de TSS associés par gène.*

2.1.3.3.1 Les états de phosphorylation des extrémités 5’ des ARN des gènes ciblés sont similaires aux non ciblés

Une des hypothèses principales serait que la différence entre les gènes impactés et non impactés soit due à l’état de phosphorylation des extrémités 5’ de leur ARN car l’activité de dégradation 5’→3’ exoribonucléase de la RNase J préfère légèrement les extrémités 5’P aux extrémités 5’PPP.
Pour tester cette hypothèse, nous avons comparé les proportions de forme 5’P des ARN ciblés *vs* non ciblés à l’aide d’un test de Wilcoxon-Mann-Whitney. Pour examiner l'état de phosphorylation de chaque transcrit, l’équipe de Peter Redder a développé le protocole PQ-EMOTE (Phosphorylation Quantification Exact mapping Of Transcriptome Ends). Le principe de ce protocole est de diviser un échantillon en deux pour préparer deux librairies de 5’-end RNA-Seq en parallèle, une contenant les ARN 5’P ($pool^{mono}$), et l’autre contenant les ARN 5’P et 5’PPP ($pool^{total}$). Pour chaque extrémité, il est alors possible de calculer (avec une méthode décrite en partie 2) sa proportion de forme 5’P en faisant le ratio de son abondance dans l’échantillon $pool^{mono}$ sur son abondance dans l’échantillon $pool^{total}$.

Pour la mise en œuvre de cette analyse, nous disposons de trois paires d’échantillons (issus de la librairie PQ-EMOTE 58 dont le détail sera présenté dans la partie 2 des résultats), soit 6 fichiers FASTQ à traiter. Les traitements effectués correspondent à une chaîne de traitements que nous décrirons en détails dans la partie 2. Brièvement, cela consiste à vérifier l’intégrité des lectures et à retirer les éventuels adaptateurs encore présents, puis à aligner les lectures sur le génome de référence, et à quantifier la première position alignée de chacune des lectures. Les comptages sont ensuite dédupliqués, filtrés et normalisés afin de pouvoir calculer une proportion de forme 5’P pour chacune des extrémités 5’.

Les TSS pouvant varier de quelques nucléotides, cela entraîne la production de plusieurs espèces d'ARNm pour un même gène, présentant des extrémités 5' différentes.
Si les états de phosphorylation influent effectivement sur la stabilité des ARN, la demi-vie des ARN correspondant à un gène ne dépend donc pas uniquement de l’état de phosphorylation d’une seule extrémité, mais peut être considérée comme résultant de la moyenne pondérée de la proportion d’extrémités monophosphorylées (5'P) de toutes les extrémités 5' observées pour ce gène.

Sur les 9 millions de lectures alignées en moyenne, nous avons pu calculer une proportion sur leur état de phosphorylation ($pool^{mono}$ / $pool^{total}$) pour 3 570 positions sur les 2 783 537 du génome.

Parmi ces 3 570 positions, 1 376 correspondent à une position caractérisée comme étant un TSS. Nous avons pu calculer une proportion pour 1 554 gènes, dont 611 parmi les 866 gènes ciblés et premiers de leur opéron et 61 parmi les 157 gènes non ciblés et premiers de leur opéron.
Les résultats montrent que la variation de demi-vie (donc l'impact de la délétion de la RNase J) des ARN ne semble pas être corrélée avec l'état de phosphorylation de leur extrémité 5' comme illustré en (Fig. 77). De même, nous n'observons pas de différence statistiquement significative (p-valeur = 0,62) de proportions entre les groupes cible et non cible.

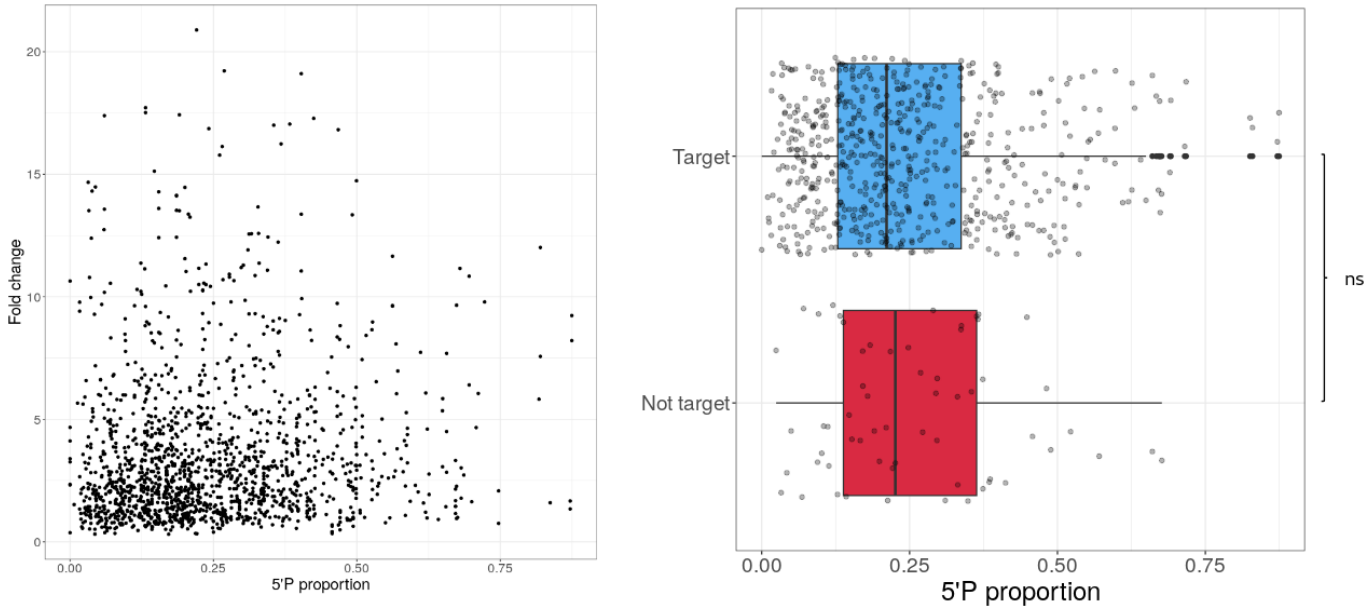


Figure 77 : Proportion d'extrémité 5'P des gènes en fonction de leur variation de stabilité.
*(**A**) Proportion d'extrémités 5'P des TSS représentants de leur cluster en fonction de la variation de demi-vie entre les souches sauvage et J1^{AGA} pour les gènes premiers de leur opéron, (**B**) Distribution des proportions d'extrémités 5'P par groupe cible ou non cible.*

Les résultats obtenus avec les expériences menées *in vitro* présentés par Hausmann et al. (2017) montrent que l'état de phosphorylation des ARN de *S. aureus* n'impacte que très peu l'activité de la RNase J contrairement à certains de ses homologues bactériens. Ces résultats ne sont donc pas très surprenants et permettent de soutenir l'hypothèse que la RNase J1 est capable de dégrader les ARN 5'P et les ARN 5'PPP.
Nous avons également essayé de comparer les proportions des gènes non ciblés et non impactés *vs* ciblés et impactés, mais nous n'avons pu calculer une proportion d'extrémités 5'P que pour respectivement 24 (parmi les 65) et 286 (parmi les 381), ce qui est trop peu pour effectuer une comparaison représentative.

2.1.3.3.2 Les séquences primaires, les pourcentages en GC et les énergies libres des extrémités ne sont pas différentes entre les ARN ciblés impactés et les non ciblés non impactés

Il a été montré par des analyses transcriptomiques chez *S. aureus* que la RNase Y préfère cliver juste après une guanine (Khemici et al., 2015), et bien que ce signal soit faible pour, à lui seul, guider un clivage, cela suggère que les activités des ribonucléases peuvent être, en partie, régulées par des motifs. A partir de cette observation, nous avons cherché à déterminer si l'activité de la RNase J pouvait elle-même être modulée par des motifs.

Par ailleurs, il a été montré par des analyses transcriptomiques chez *E. coli* que, sur beaucoup de substrats de la RNase E, une structure en tige boucle en amont du site de clivage est présente, suggérant que les motifs structurels des ARN pourraient être parfois nécessaires pour la reconnaissance par la RNase E (Campo et al., 2015).
Nous avons donc cherché à déterminer si l'activité de la RNase J pouvait également être modulée par la présence de structures secondaires. De plus, les régions riches en GC sont plus susceptibles de former des structures secondaires stables. Nous avons donc également considéré le pourcentage en GC des extrémités 5' des ARN.

Pour tester ces hypothèses nous avons commencé par comparer les séquences des extrémités 5' des ARN ciblés et impactés avec celles des ARN non ciblés et non impactés. Nous avons réalisé des weblogos qui sont des représentations graphiques pouvant permettre de mettre en lumière des motifs conservés au sein d'un ensemble de séquences alignées. Ces weblogos se basent sur la théorie de l'information et utilisent la notion d'entropie et d'information afin d'identifier des nucléotides conservés dans un ensemble de séquences. Les séquences d'intérêt, ici les séquences des extrémités 5', doivent d'abord être alignées. Nous avons aligné uniquement les TSS associés aux premiers gènes des opérons, et pour chaque cluster de TSS, nous n'avons gardé que celui identifié comme étant le représentant de son cluster par Prados et al. (2016). Cela correspond au TSS avec le plus de lectures alignées. Les séquences sélectionnées sont ensuite alignées par rapport au 1er nucléotide du TSS.
Pour chaque position de l'alignement, il est possible de calculer l'information relative ou *bits* d'information qui indique à quel point cette position est conservée ou variable. L'hypothèse étant que si un motif est présent, il devrait se trouver non loin de l'extrémité 5'. Nous avons construit les weblogos à partir des 30 premières positions des TSS représentants.

Les programmes, utilisent généralement la même probabilité pour chacun des nucléotides (soit 0,25). Cependant, chez *S. aureus* le pourcentage en GC est autour de 33 % (16 % C et 17 % G). Nous avons donc utilisé le programme Weblogo3 (Crooks et al., 2004) qui permet de prendre en compte ce pourcentage en GC dans le calcul de l'information.

Dans un second temps, nous avons comparé les capacités de ces régions à former des structures secondaires stables. Pour cela, nous avons utilisé le programme RNAfold issu de la suite de programme Vienna (Hofacker et al., 1994)**,** (Lorenz et al., 2011) pour prédire la présence de structure secondaires sur ces séquences. RNAfold se base sur des modèles thermodynamiques pour prédire la structure secondaire la plus probable d'une séquence d'ARN donnée, c'est-à-dire celle qui minimise l'énergie libre (Minimum Free Energy, MFE). L'énergie libre est une mesure de la stabilité thermodynamique de la structure, une énergie plus basse indique une structure plus stable, et c'est cette mesure que nous avons comparée entre les ARN des deux groupes étudiés. Ici, nous sommes principalement intéressés par la prédiction de tiges boucles dans la région des 5'-UTR. Cependant, plus les séquences données en entrée des programmes comme RNAfold sont longues, plus ils sont susceptibles de prédire des structures grandes, c'est pourquoi nous avons utilisé le programme RNAfold sur les mêmes séquences que celles utilisées pour la création des weblogos, c'est-à-dire les premiers 30 nucléotides en aval des TSS représentants. Cette analyse a également été réalisée sur des séquences de 40 et 50 nucléotides de long. Les résultats étant similaires, seuls ceux obtenus sur des séquences de 30 nucléotides de long sont présentés.
Nous avons ensuite comparé les énergies libres des structures secondaires prédites associées aux extrémités issues des deux groupes à l'aide d'un test de Wilcoxon-Mann-Whitney.

Au total, nous avons pu construire les weblogos à partir de 56 extrémités issues d'ARN non ciblés et non impactés, et 531 issues d'ARN ciblés et impactés.
Les résultats illustrés en Figure 78A et 78B ne semblent pas montrer de motif relatif à l'un des groupes. Une guanine est légèrement conservée en première position dans les deux groupes, au regard de sa fréquence théorique, en 1ère position. Cependant, il est connu que le 1er nucléotide transcrit est souvent un A ou un G comme observé ici.

La capacité des ARN à former des structures secondaires stables est estimée à partir de l'énergie libre ou le pourcentage en GC. Les distributions des valeurs obtenues sont illustrées en Figure 79. Il ne semble pas y avoir de différences entre les deux groupes. En effet, les distributions des énergies libres prédites par RNAfold et celle des pourcentages en GC ne sont statistiquement pas différentes (p-valeurs respectives = 0,23 et 0,41).

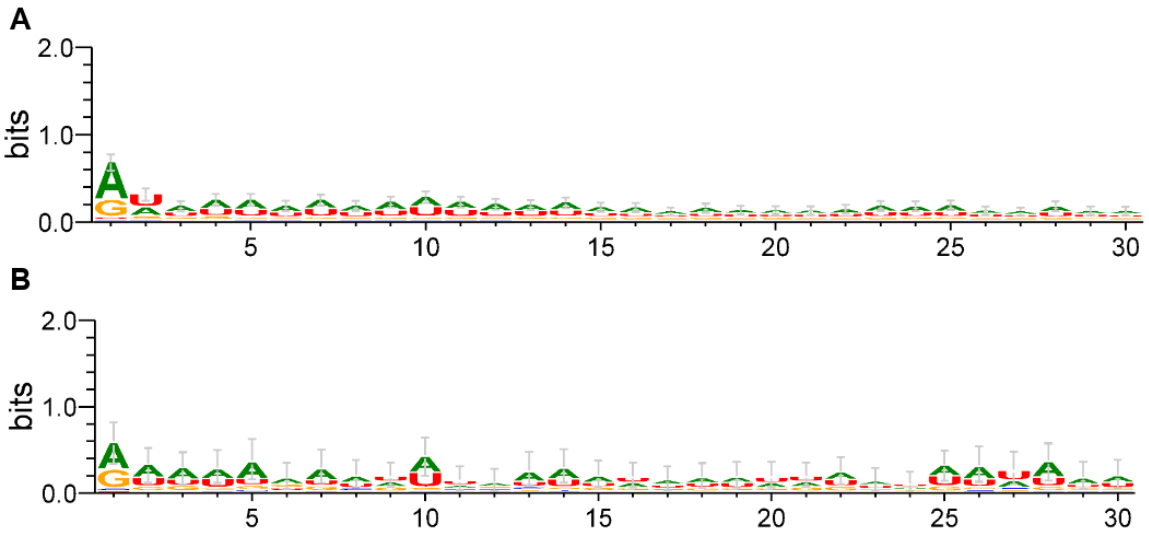


***Figure 78 :* Conservation des séquences des extrémités 5' des ARN.** *Weblogos générés en tenant compte du pourcentage en GC de S. aureus (**A**) des extrémités 5' des ARN non impactés non ciblés et (**B**) des extrémités impactées ciblés.*

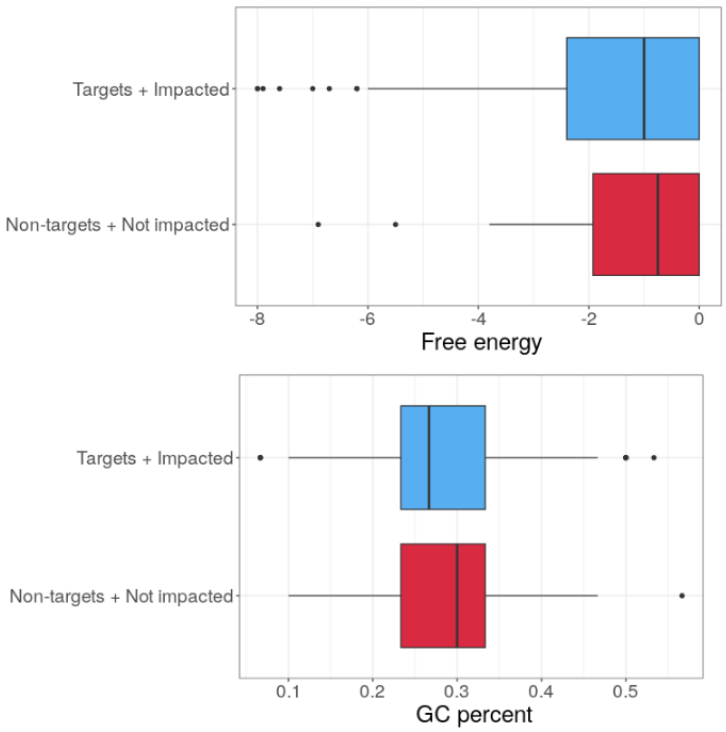


***Figure 79 :* Énergies libres et pourcentage en GC des extrémités 5' des ARN.** *Distribution des (**A**) pourcentages en GC et (**B**) énergies libres, associées aux extrémités des ARN impacté (bleu) et non impactés (rouge)*

L'ensemble de ces analyses n'a pas permis de montrer que la composition en ribonucléotides, le pourcentage en GC, ou encore la présence de structures secondaires au niveau des extrémités 5' pouvait avoir un impact sur le ciblage de la RNase J1.

2.1.4 La RNase J est la principale actrice de la dégradation co-traductionnelle chez *S. aureus*

Chez la levure, l'exonucléase 5'→3' Xrn1 peut dégrader les ARNm en cours de traduction en suivant le dernier ribosome. Cette dégradation co-traductionnelle se manifeste par les intermédiaires de dégradation, c'est-à-dire les extrémités 5' monophosphorylées des ARNm en cours de dégradation, qui, lorsqu'alignées sur le génome et quantifiées, révèlent une période de 3 nucléotides. Cette périodicité résulte de l'effet des ribosomes qui se déplacent sur les ARNm par pas de 3 nucléotides (un codon). La périodicité commence 14 nucléotides en amont du codon de départ, ce qui correspond aux nucléotides protégés par le ribosome lorsque le codon de départ se trouve dans le site P. Une abondance d'extrémités 5' est retrouvée à cette position, étant donné que l'initiation de la traduction est plus lente que l'élongation de la traduction. Un autre enrichissement peut être observé 17 nucléotides en amont des codons *stop*, dû au recrutement lent des facteurs de libération du ribosome lorsqu'un codon stop se trouve dans le site A. La périodicité est en effet absente après cette position (Pelechano et al., 2015)**.**

Une périodicité a ensuite été observée pour plusieurs bactéries, dont *S. aureus* et *B. subtilis*, pour lesquelles il a été supposé que l'effet soit causé par l'activité exoribonucléase de la RNase J1 et/ou J2 (Huch et al., 2021). Toutefois, cette attribution n'était pas certaine, car pour de nombreuses bactéries des genres Eubacterium, Gerdnerella et Caulobacter, possédant des homologues de la RNase J, il n'est pas observé de périodicité sur les intermédiaires de dégradation des ARNm (Huch et al., 2021)**.**

Notre objectif est donc d'étudier si le phénomène de dégradation co-traductionnelle observé chez *S. aureus* est effectivement dû à la RNase J.

Pour ce faire, nous avons utilisé des données générées à partir du protocole EMOTE (Exact Mapping Of Transcriptome Ends) (Redder 2018) qui a été réalisé sur une souche sauvage et sur la souche mutante $J1^{AGA}$.

Les données ont été analysées à l'aide de fonctions que nous avons développées et qui seront décrites en partie 3, et qui s'inspirent du travail de (Nersisyan et al., 2020).

Les données correspondent à des fichiers FASTQ dans lesquels sont présentes les lectures correspondant aux extrémités 5' des ARN 5'P, incluant donc les intermédiaires de dégradation.

Brièvement, le principe est de considérer la position du 1er nucléotide aligné de chaque lecture par rapport à une caractéristique génomique (ici, le codon de départ ou d'arrêt de traduction).

Seules les extrémités s'alignant dans une fenêtre de 100 nucléotides autour de la caractéristique sont considérées. Afin de tenir compte du niveau d'expression des gènes, les comptages sont normalisés par rapport au nombre total d'extrémités alignées dans la fenêtre. Pour finir, les comptages sont agrégés par position relative à la caractéristique. De cette manière, un profil de méta-comptages est obtenu autour de la caractéristique étudiée, qui nous permettra d'analyser la distribution des intermédiaires de dégradation le long des séquences codantes.

Nous étudierons la présence d'une périodicité sur les profils de méta-comptages générés à partir des souches sauvage et $J1^{AGA}$. En effet, un profil de méta-comptages peut être vu comme un signal, sur lequel nous effectuons une analyse de transformée de Fourier rapide (FFT) pour étudier la périodicité.

Pour analyser des signaux, il y a deux types principaux d'analyse utilisant la transformée de fourrier, la CFT (Continuous Fourier Transform) pour les signaux continue et la DFT (Discrete Fourier Transform) pour les signaux discrets. Nos profils de méta-comptages étant des valeurs discrètes (points spécifiques dans le temps) nous avons eu recours à une analyse DFT. La FFT est une méthode optimisée pour le calcul de la DFT dont l'utilisation dans R est rendue possible et simple au travers d'une fonction de la librairie de base *stats*. La valeur qui nous intéresse étant la période. Et enfin, afin d'étudier s'il y a un biais dans la distribution des extrémités 5' le long des régions codantes, nous analysons comment sont réparties les extrémités 5' alignées parmi les trois cadres de lecture possibles.

Les lectures contenues dans les quatre fichiers FASTQ correspondant aux deux réplicats sauvage et aux deux réplicats $J1^{AGA}$ ont été alignées sur le génome de référence de la souche SA564 de *S. aureus* (NZ_CP010890) avec 1 seul mésappariement autorisé et en ne gardant que les alignements uniques. En moyenne, 1,3 millions de lectures ont été alignées. Étant donné que seule l'extrémité 5' des ARN nous intéresse ici, les alignements contenus dans les fichiers BAM ont ensuite été traités pour ne garder que la position de la première position alignée de chacune des lectures. Au total, 605 759 positions génomiques (sur les 2 783 537) ont au moins une extrémité 5' alignée.

Les profils de méta-comptages que nous avons générés à partir des deux réplicats sauvages sont présentés en Figure 80. Ils correspondent à l'agrégation de 2 532 codons d'initiation et codons d'arrêt.

Ces profils révèlent une accumulation d'extrémités 5’ s'alignant 14 et 11 nucléotides en amont du codon de départ, et 17 nucléotides en amont du codon d'arrêt. Ces accumulations, déjà observées dans d'autres organismes, s'expliquent par le comportement du ribosome avant la traduction : celui-ci marque une longue pause lorsque le codon de départ est dans son site P. Pendant cette pause, le ribosome protège, par encombrement stérique, une région de l'ARN des exoribonucléases 5’→3’, qui ne peuvent pas dégrader les ARN au-delà de 11 nucléotides en amont du codon de départ. De manière similaire, en fin de traduction, avant de se dissocier de l'ARN, le ribosome effectue une longue pause lorsque le codon d'arrêt est dans le site A, ce qui bloque les exoribonucléases 5’→3’.

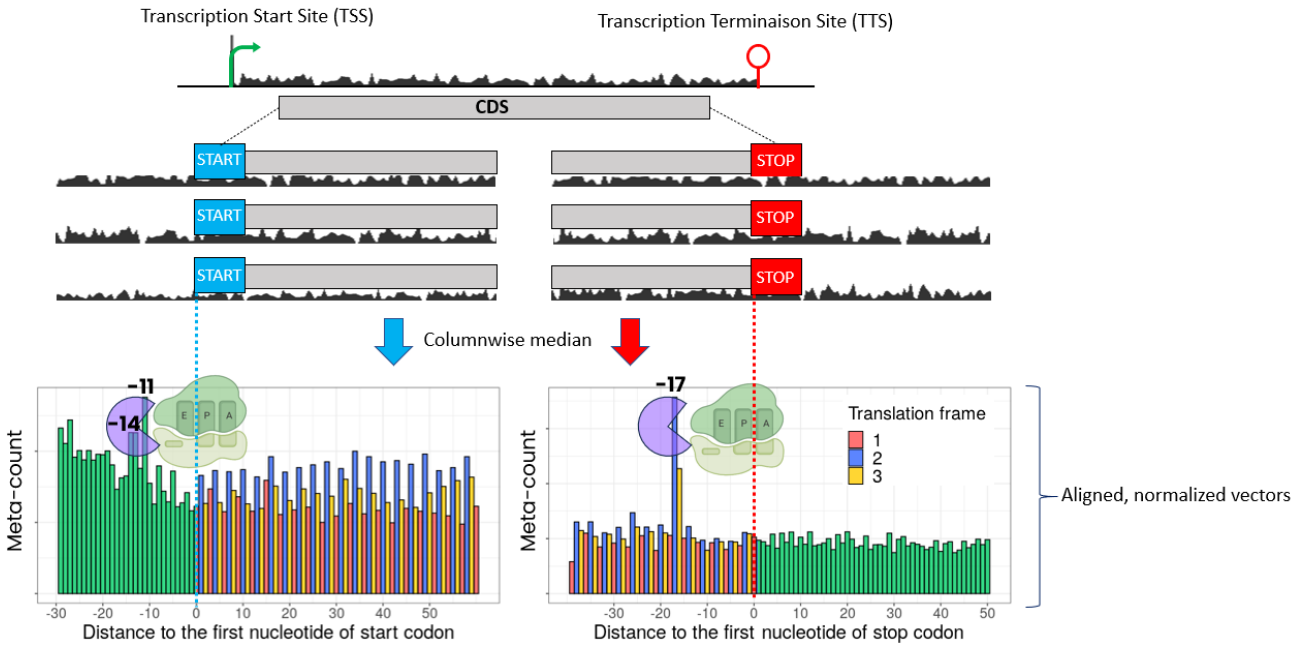


Figure 80 : Profils de méta-comptages autour des régions start et stop des gènes.

Profils des méta-comptages des intermédiaires de dégradation chez une souche sauvage de S. aureus SA564 autour des codons de départ (gauche) et autour des codons d’arrêt (droite). Les différents cadres de lectures sont coloriés d’une couleur différente. Le ribosome et le pacman symbolisant une exoribonucléase 5’→3’ ont été ajoutés manuellement.

En outre, il semblerait que des accumulations moins importantes se produisent entre les deux longues pauses, le long de la séquence codante, tous les trois nucléotides, avec une préférence pour le cadre de lecture numéro 2, colorié en bleu sur la Figure 80.

La Figure 81 présente les résultats de l’analyse FFT, l'axe des abscisses ("Period") représente les différentes périodes détectées dans le profil de méta-comptages, tandis que l'axe des ordonnées ("Signal") montre l'amplitude du signal FFT correspondant à chaque période. Un signal élevé à une

certaine période indique une forte régularité dans la distribution des méta-comptages à cette échelle temporelle, suggérant une périodicité.
Ces résultats montrent un fort signal pour une période de trois nucléotides, confirmant ainsi les observations faites sur les profils de méta-comptages.

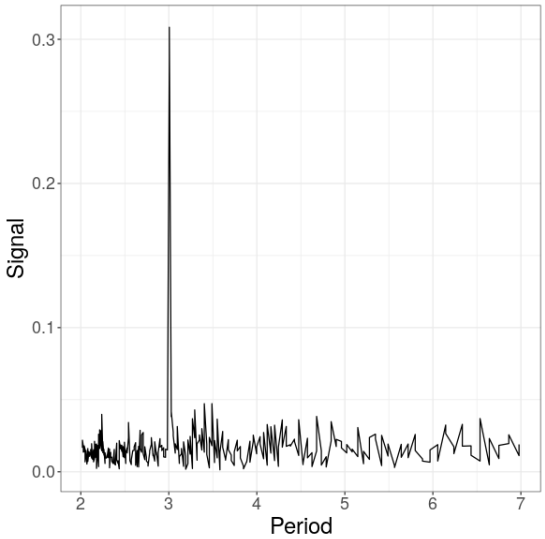


Figure 81 : Profil FTT dans la souche sauvage. *Analyse FFT des profils de méta-comptages effectuées sur les régions codantes de la souche sauvage.*

Ces résultats confirment les observations précédentes faites chez *S. aureus*, à savoir que, dans la souche sauvage, il existe un mécanisme de dégradation co-traductionnelle.

Les résultats obtenus sur l'analyse de préférence de cadres de lecture illustrés en Figure 82 montrent que les extrémités 5' correspondent davantage (45 %) au cadre de lecture numéro 2, et confirment l'observation faite sur les profils de méta-comptages.

Ces résultats confirment les observations précédentes faites chez *S. aureus*, à savoir que, dans la souche sauvage, il existe un mécanisme de dégradation co-traductionnelle. Après avoir confirmé les résultats attendus chez la souche sauvage, nous allons maintenant analyser les résultats issus de la souche mutante et déterminer si ces observations sont également présentes ou bien absentes.

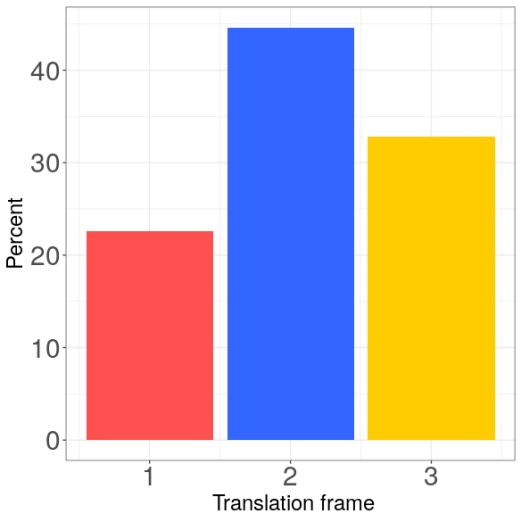


Figure 82 : Distribution des intermédiaires de dégradation par phase de lecture dans la souche sauvage. *Pourcentage d'extrémité 5' alignées sur chacune des trois phases de lecture des régions codante dans la souche sauvage.*

La Figure 83 présente les profils de méta-comptages générés à partir des deux réplicats issus de la souche J1^{AGA}. Les accumulations observées dans la souche sauvage, 11 nucléotides en amont des codons d'initiation et 17 nucléotides en amont des codons d'arrêt, sont nettement moins importantes dans la souche J1^{AGA}. Cela montre que l'exoribonucléase 5'→3' présente chez *S. aureus*, qui est bloquée par les ribosomes lors de leurs pauses, est bien la RNase J1.
Nous pouvons également observer une accumulation d'extrémités 5'P trois nucléotides en aval des codons de départ. Cette accumulation est absente dans la souche sauvage.

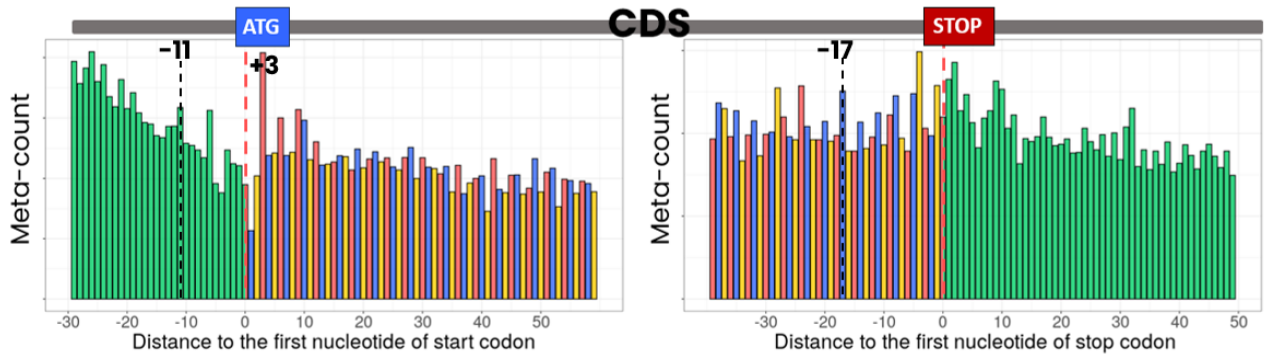


Figure 83 : Profils des méta-comptages des intermédiaires de dégradation chez une souche J1^{AGA} autour des codons de départ (gauche) et autour des codons d'arrêt (droite).

De plus, la Figure 84, illustrant l'analyse FFT réalisée sur le profil de méta-comptages de la souche $J1^{AGA}$, montre qu'il existe toujours une période de trois nucléotides. Bien que le signal de cette périodicité soit beaucoup moins important que dans la souche sauvage, elle suggère qu'un autre acteur de la dégradation des ARN permet une dégradation co-traductionnelle.

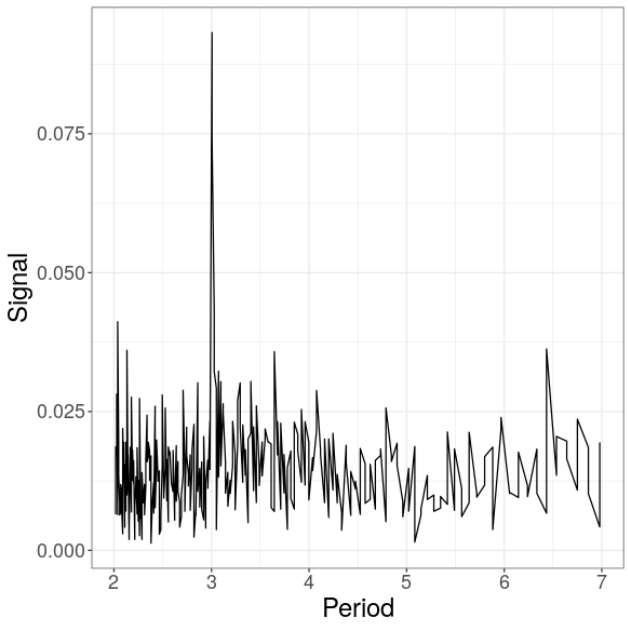


Figure 84 : Profil FTT dans la souche $J1^{AGA}$. *Analyse FFT des profils de méta-comptages effectuées sur les régions codantes de la souche $J1^{AGA}$.*

Enfin, la répartition des comptages sur les cadres de lecture, présentée en Figure 85, montre que la préférence pour le cadre numéro 2, observée chez la souche sauvage, n'est pas présente chez la souche $J1^{AGA}$. Ces résultats renforcent l'hypothèse que l'activité de la RNase J1 est responsable du mécanisme de dégradation co-traductionnelle dans le cadre de lecture numéro 2.

Il convient de noter que ces analyses ont également été réalisées sur des données issues de la souche ΔJ1 ΔJ2. Cependant, les résultats des trois analyses (méta-comptages, FFT et cadre de lectures) étant similaires nous ne présenterons pas les figures. Cela permet tout de même de confirmer le fait que la légère périodicité observée dans le mutant $J1^{AGA}$ mais aussi chez la souche double mutante ΔJ1 ΔJ2, n'est pas due à la RNase J2.

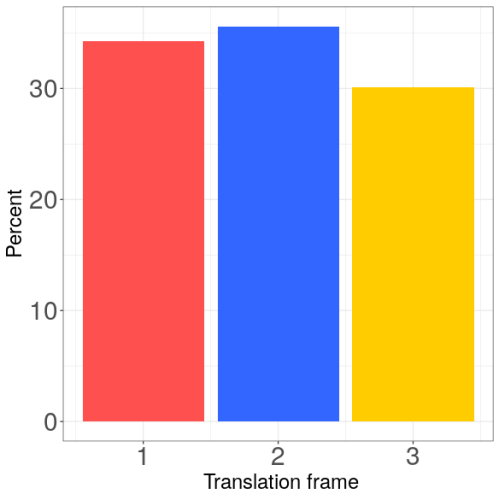


Figure 85 : ***Distribution des intermédiaires de dégradation par phase de lecture dans la souche J1^{AGA}.*** *Pourcentage d'extrémité 5' alignées sur chacune des trois phases de lecture des régions codante dans la souche J1^{AGA}.*

Il est intéressant de noter que, chez *B. subtilis*, les pauses des ribosomes au début et à la fin de la traduction se produisent respectivement 11 et 14 nucléotides en amont des codons start et stop. Cependant, chez *S. aureus*, les pauses se produisent à 14 et 17 nucléotides en amont des codons start et stop, ce qui correspond au modèle observé chez la levure. Cette dernière, qui ne possède pas de RNase J, fait partie des eucaryotes et est donc beaucoup plus éloignée phylogénétiquement de *S. aureus* que ne l'est *B. subtilis*. En revanche, il existe également une forte abondance d'extrémités 5' à 11 nucléotides en amont du codon start, comme chez *B. subtilis*. Cependant, cette abondance ne peut pas correspondre au début de la périodicité. En effet, lors de la pause du ribosome au début de la transcription, le codon start est dans le site P du ribosome, et lors de la pause finale, le codon stop est dans le site A. Ces deux sites étant espacés de 3 nucléotides, les positions des pauses initiale et finale doivent avoir un écart de trois nucléotides (17-14 = 3) et non de six (17-11 = 6).
Pour le moment, nous ne pouvons pas expliquer pourquoi ces pauses diffèrent entre les deux bacillota, ni pourquoi il y a deux pauses chez *S. aureus* au niveau de la région start en -11 et -14. Cela démontre néanmoins que le modèle de *B. subtilis* n'est pas universel parmi les bacillota et que *S. aureus* nécessite son propre modèle de dégradation co-traductionnelle.

Fait remarquable, un schéma de périodicité alternatif sous-jacent a été révélé dans les résultats de la souche J1^{AGA}, où une période de 3 nucléotides est observée (Fig. 83 et Fig. 84).

Cependant, cette périodicité semble commencer dans la séquence codante (à une distance de 3 nucléotides en aval premier nucléotide du codon start), ce qui n'est pas compatible avec une 5'→3' exoribonucléase. Plusieurs ribonucléases faisant partie de systèmes toxine-antitoxine peuvent cliver dans le site A (Mansour et al., 2022), mais à notre connaissance, elles laissent toutes un groupement hydroxyle (OH) à l'extrémité 5' de leur produit, et de tels événements de clivage ne seraient pas détectés par le protocole EMOTE utilisé ici (Redder, 2018). Au lieu de cela, nous pouvons émettre l'hypothèse que l'endoribonucléase Rae1 serait responsable, car il a été proposé que Rae1 entre dans le site A du ribosome et clive l'ARNm entre les codons du site P et du site A, tout en laissant une extrémité 5'P. Une manière de tester cette hypothèse serait l'analyse de données générées à partir du protocole EMOTE issues d'une souche mutante à la fois pour la RNase J et pour Rae1.

Dans la souche de type sauvage, les extrémités 5' des ARN en cours de dégradation sont distribuées de manière biaisée dans le cadre de lecture numéro 2 (Fig. 82). Ce biais n'est pas observé dans la souche mutante J1^{AGA} (Fig. 85) pour laquelle il ne semble pas y avoir de biais pour un cadre particulier. Cela signifie que nous pourrions être en mesure de déterminer l'origine des extrémités 5' en fonction du cadre de traduction sur lequel elles s'alignent.

2.1.4.1 Analyse de la périodicité par sous ensemble de gènes

2.1.4.1.1 Ciblés impactés *vs* non ciblés non impactés

Les analyses de périodicités des intermédiaires de dégradation s'appuient sur le fait d'agréger les données de toutes les régions codantes pour des protéines. Néanmoins, si la profondeur de séquençage était suffisante, il semblerait tout à fait envisageable de faire ces analyses sur des sous-ensembles de régions, par exemple, les régions correspondant aux ARN ciblés et impactés par la RNase J1 identifiés précédemment ainsi que celles identifiées comment non ciblées et non impactées. Sous l'hypothèse que la RNase J est la principale ribonucléase réalisant la dégradation co-traductionnelle, les profils obtenus entre les gènes ciblés et impactés et les gènes non cibles et non impactés devraient être différents.
Pour ce faire, nous avons utilisé un ensemble de 59 fichiers FASTQ contenant des lectures issues d'une préparation de librairies de séquençage EMOTE, dont 4 proviennent de données publiques (Khemici et al., 2015) et 55 ont été générés par l'équipe collaboratrice (Redder). Les fichiers FASTQ ont été traités de la même manière que précédemment.

Aucun de ces échantillons n'est issu d'une souche mutante pour la RNase J, ce qui signifie que nous ne pourrons pas observer la disparition des éventuelles périodicités observées, car les deux seuls réplicats issus de la souche $J1^{AGA}$ ne permettrons pas de détecter un signal sur le petit ensemble de gènes que représente le groupe des gènes non impactés et non ciblés. Comme la périodicité, ainsi que les pauses ribosomales de début et de fin, sont principalement dues à l'activité de la RNase J1, nous pouvons nous attendre à une réduction du signal pour la période de 3 nucléotides dans l'ensemble des gènes non ciblés et non impactés. En revanche, pour le sous-ensemble des gènes ciblés et impactés, nous pouvons nous attendre à observer un profil encore plus marqué.

Nous avons donc généré les profils de méta-comptages, effectué les analyses FFT et les analyses de préférence de cadre de lecture sur les deux sous-ensembles : 107 gènes non ciblés et non impactés, et 833 gènes ciblés et impactés.

En moyenne, 6,3 millions de lectures ont été alignées ce qui correspondant à 1 594 920 positions (sur les 2 783 537 du génome).

Le profil des gènes ciblés et impactés est présenté en Figure 86A et montre une accumulation 14 et 11 nucléotides en amont des codons d'initiation. De plus, la périodicité dans le cadre lecture numéro 2 est très apparente sur le profil, et est par ailleurs confirmée par l'analyse FFT et la distribution des extrémités selon les cadres de lectures, comme illustré en Figure 86B et 86C. Ces observations confirment nos hypothèses, c'est-à-dire que le phénomène de dégradation co-traductionnelle se distingue davantage que par rapport aux résultats obtenus précédemment à partir de tous les gènes (F2 = 50%). Néanmoins, ceci pourrait aussi être dû au fait que l'analyse est effectuée à partir de davantage de données RNA-Seq.

En outre, les profils de méta-comptages générés à partir des gènes non ciblés et non impactés, présenté en Figure 86D révèlent également les accumulations 14 et 11 nucléotides en amont des codons d'initiation et nous décelons également une périodicité en phase 2. Bien que ces observations soient moins marquées que pour l'autre groupe, comme le montre l'analyse FFT en Figure 86E et la distribution des cadres de lecture présentée en Figure 86F, il est surprenant d'observer ces résultats.

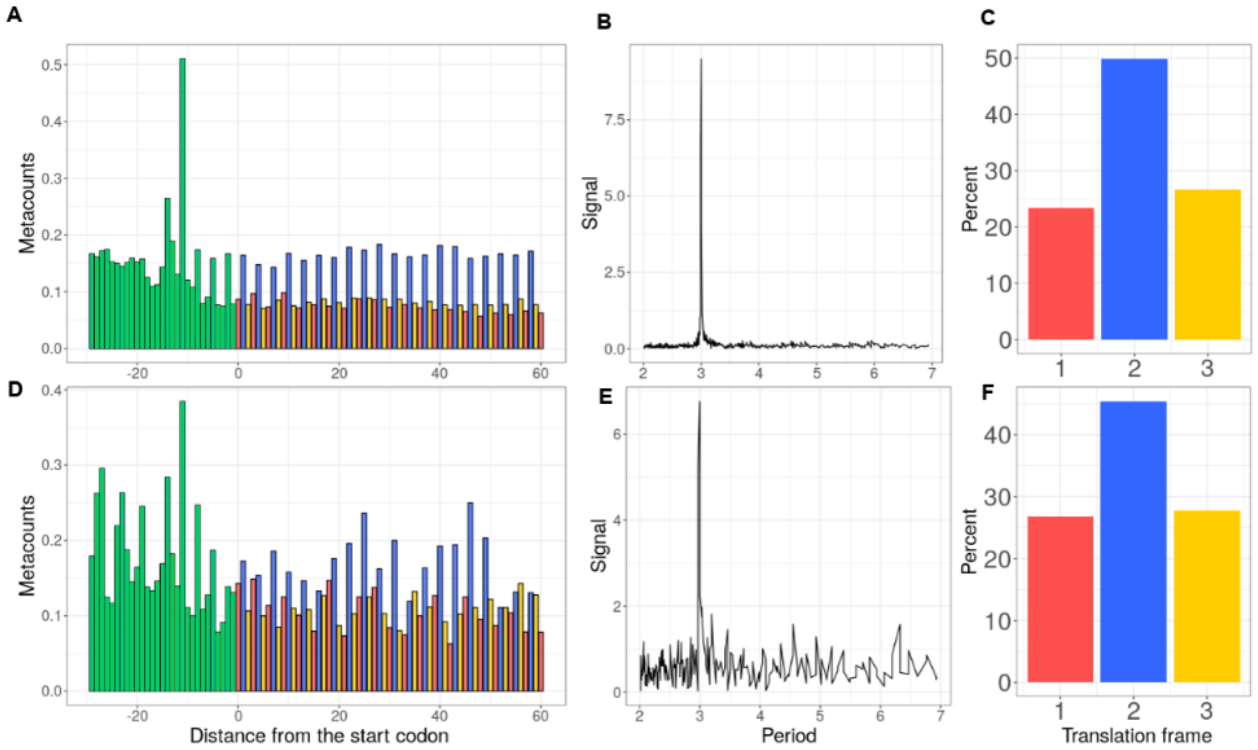


Figure 86 : ***Analyse de la périodicité des groupes de gènes impactés ciblés et non impactés non ciblés.*** *Profils de méta-comptages par rapport aux codons de départ (panneau de gauche). Les analyses FFT correspondantes sont présentées dans le panneau central et les distributions des extrémités 5' par rapport aux cadres de lectures sont présentées dans le panneau de droite. Les trois graphiques du dessus correspondent au groupe impacté ciblé et ceux du dessous au groupe non impacté non ciblé.*

Les résultats générés à partir du groupe des gènes non ciblés et non impactés ne correspondent pas à nos hypothèses de départ. En effet, il semble qu'une périodicité de trois nucléotides dans le cadre de lecture 2 émerge de manière globale, malgré le faible nombre de gènes agrégés. Cela suggère que, parmi ces gènes, une bonne partie serait dégradée de manière co-traductionnelle par la RNase J1.

L'hypothèse principale pouvant expliquer ces résultats est que, de la même manière que les gènes non impactés peuvent néanmoins être ciblés, il est possible que les résultats issus de l'analyse des données issues du protocole CRAC comportent des faux négatifs. Autrement dit, certains gènes ciblés par la RNase J1 pourraient ne pas avoir été identifiés comme tels en raison de comptages CRAC insuffisant. Cela peut être dû au seuil de filtrage de cinq comptages ou à une profondeur de séquençage insuffisante.

Pour ces différentes raisons, il est important d'utiliser différentes approches telles que l'étude de la demi-vie, des interactions ARN-protéine et des intermédiaires de dégradation. Ainsi, si parmi les 107 gènes non ciblés et non impactés, nous observons encore une périodicité, cela indique que la RNase J cible encore plus de gènes qu'identifier précédemment. C'est pourquoi, dans la prochaine section, nous allons tenter d'étudier la périodicité par gène afin de déterminer s'il est possible d'identifier des cibles de la RNase J avec cette méthode.

2.1.4.1.2 Périodicité par gène

Nous avons vu précédemment que si la quantité de données par rapport au signal était suffisante, les analyses de périodicité étaient possibles. Nous allons, dans cette partie, montrer que ce type d'analyses est aussi possible pour une seule région, mais, bien sûr, sous réserve que la profondeur de séquençage soit suffisante pour la région d'intérêt. Cette application n'a pas été utilisée, à notre connaissance, de manière aussi précise, c'est-à-dire sans profils de méta-comptages, et il s'agirait d'une nouvelle approche permettant, dans notre cas, d'identifier les gènes dégradés par la RNase J chez *S. aureus*.

Nous avons utilisé le même jeu de données que dans l'analyse précédente avec les 59 échantillons afin de maximiser la profondeur de séquençage disponible pour chaque gène. Néanmoins, le nombre de lectures s'alignant sur une région ou un gène dans les conditions de l'expérience dépendra en tout premier lieu de son niveau d'expression.

Nous avons construit les profils de comptages et effectué les analyses FFT de chacun des 2703 gènes annotés dans la souche SA564, et nous avons choisi de présenter les résultats pour 3 d'entre eux, un gène fortement exprimé et dont l'ARN est impacté (*pheT*), un gène fortement exprimé et dont l'ARN est non impacté et est très stable (*hu*), et un gène fortement exprimé dont l'ARN est caractérisé comme non impacté (*rplm*).

Les résultats sur le profil correspondant au gène *pheT* codant pour la sous-unité β de la Phenylalanine-ARNt ligase montre qu'il est effectivement possible d'observer une périodicité (apparente également de manière visuelle) sur un gène individuel lorsque ses transcrits sont dégradés par la RNase J (Fig. 87).

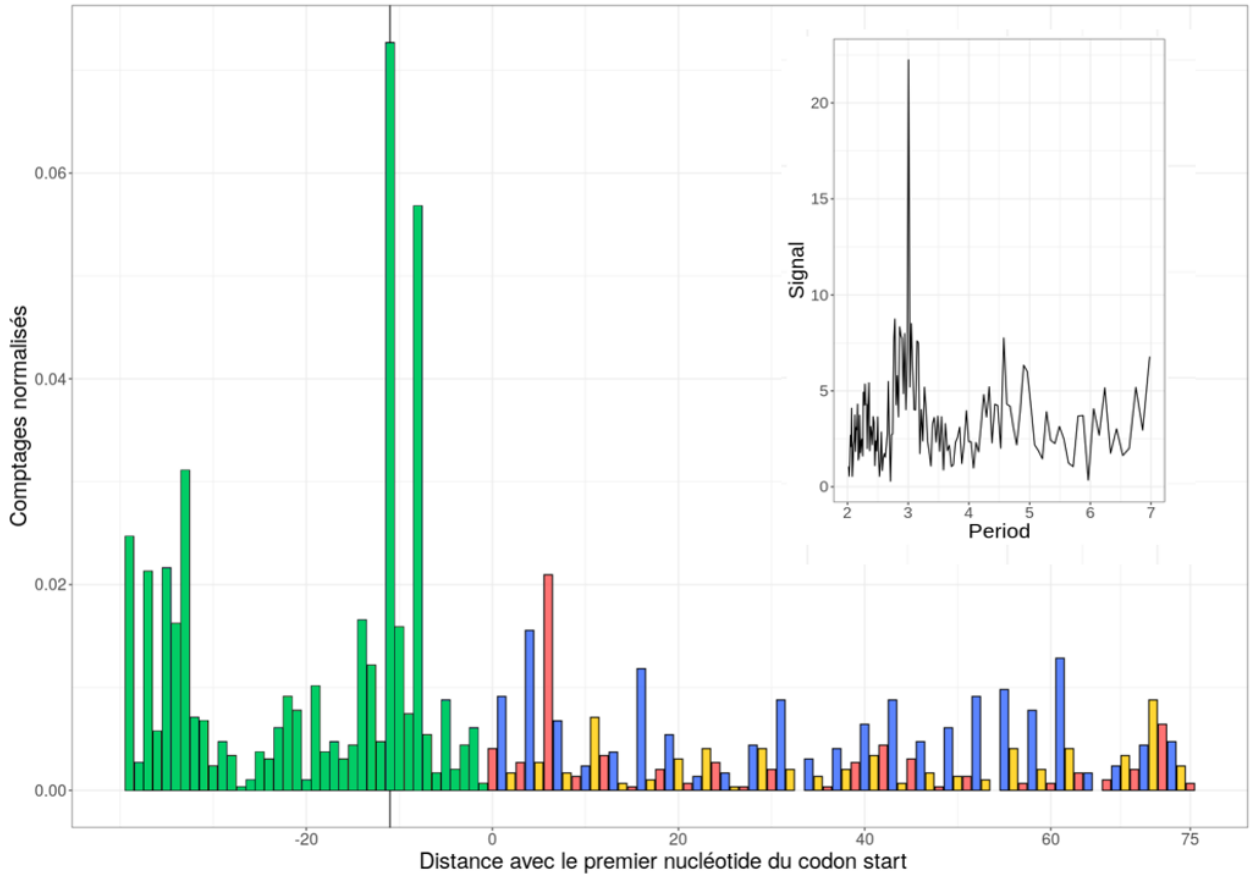


Figure 87 : ***Distribution des intermédiaires de dégradation et analyse de la périodicité du gène pheT.*** *Comptages des extrémités 5'P par rapport au codon d'initiation du gène RT87_ 05580 (pheT). Le panneau en haut à droite correspond au résultat de la transformée de Fourier rapide (FFT) sur les comptages de la région.*

Le profil fait sur le gène *hu* ne présente pas de périodicité (Fig. 88). En effet, l'immense majorité (94 %) des extrémités 5' s'alignent sur le TSS et très peu d'intermédiaires de dégradation sont observés. Ceci indique que les ARNm de *hu* ne seraient pas dégradés par la RNase J, d'où sa forte stabilité. Ces résultats sont en contradiction avec les données générées par le protocole CRAC qui montrent que la RNase J semble interagir avec l'ARN de *hu* (entre 100 et 150 nucléotides après le codon d'initiation). A moins que la RNase J ne se lie à l'ARN sans le dégrader, nous n'avons pas d'hypothèse quant à cette divergence de résultats.

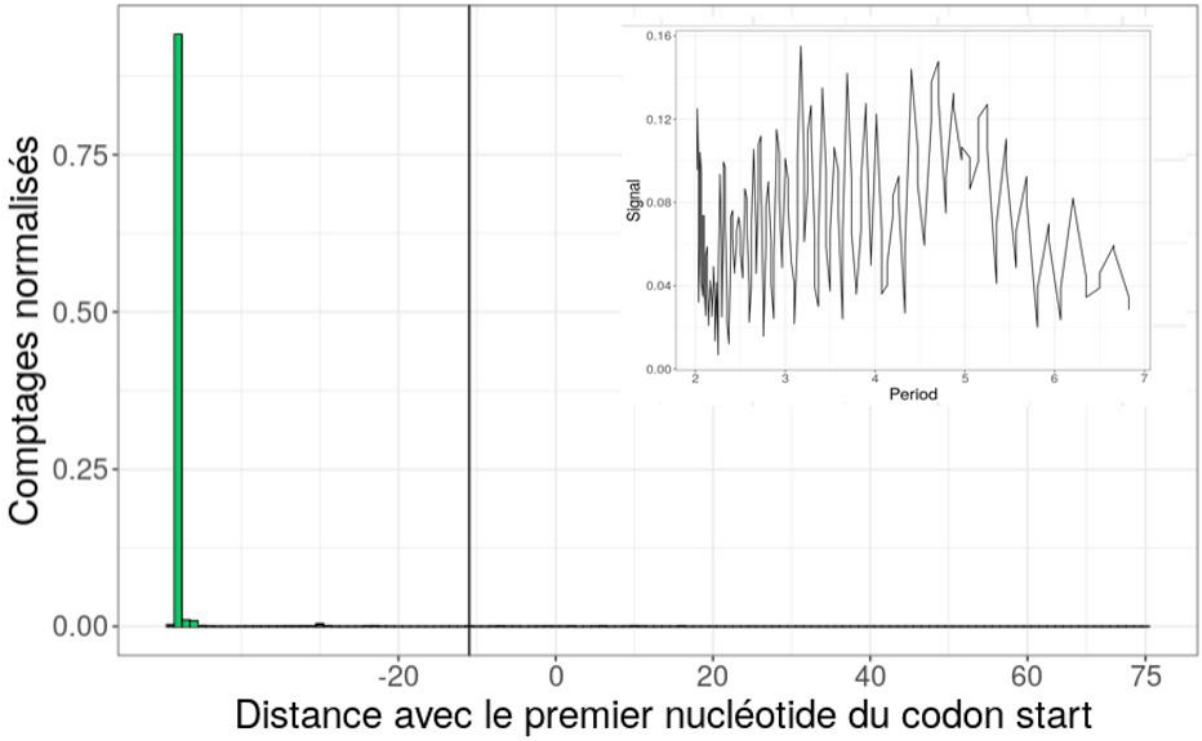


Figure 88 : ***Distribution des intermédiaires de dégradation et analyse de la périodicité du gène hu****. Comptages des extrémités 5'P par rapport au codon start du gène RT87_ 07375 (hu). Le panneau en haut à droite correspond au résultat de la transformée de Fourier rapide (FFT) sur les comptages de la région.*

Pour finir, le gène *rplm* code pour une protéine ribosomale de la sous-unité 50S du ribosome. Cette protéine est l'une des premières protéines assemblées dans la sous-unité ribosomale 50S. Son profil de comptages indique qu'il est également possible d'observer une périodicité sur des gènes catégorisés comme non impactés (Fig. 89), et donc de prédire si les ARN d'un gène peuvent être dégradés par la RNase J.

Une absence de périodicité sur une région correspondant à un gène codant n'implique pas forcément que le gène n'est pas dégradé par la RNase J. En effet, il n'est pas toujours possible de distinguer l'absence de signal par manque de profondeur et l'absence de signal du fait que la RNase J ne dégrade pas cet ARN. De même, la périodicité est observable sur un gène de par le fait qu'il est en cours de dégradation co-traductionnelle, ce qui nécessite suffisamment de copies de l'ARN qui soient en cours de traduction.

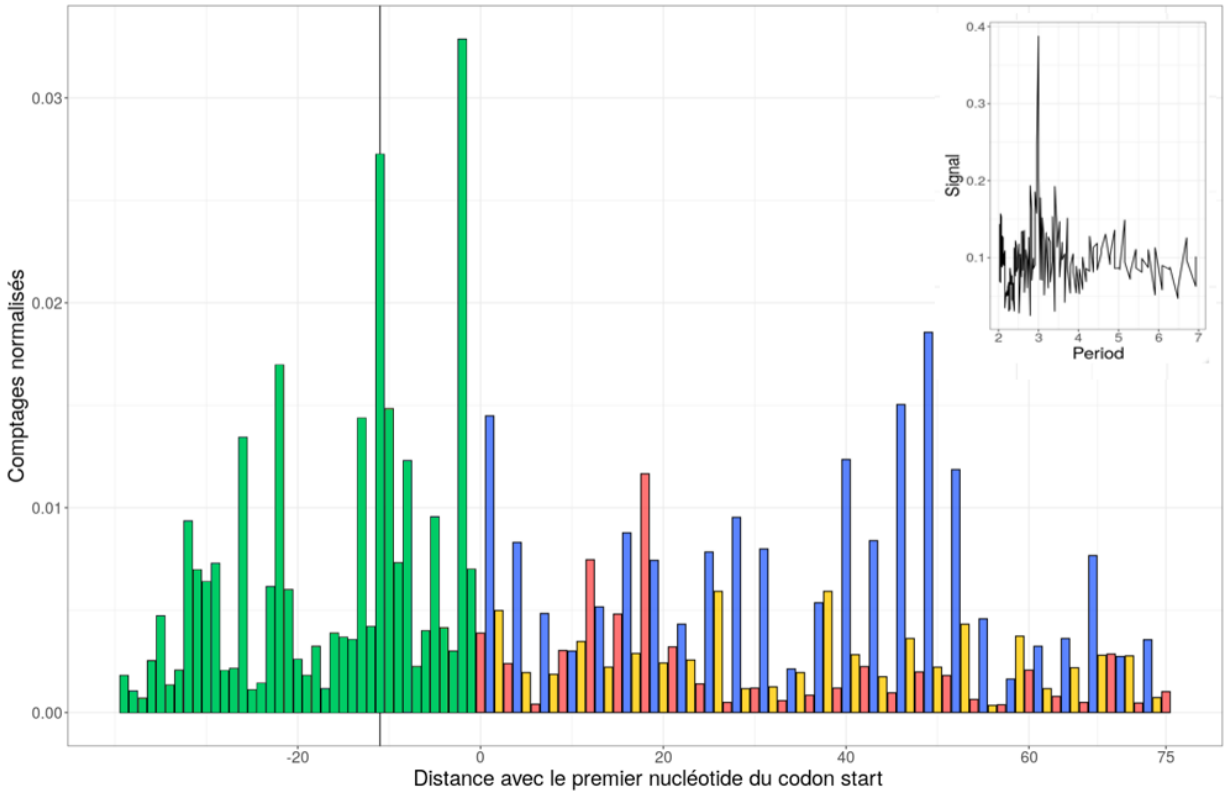


Figure 89 : ***Distribution des intermédiaires de dégradation et analyse de la périodicité du gène rplm.*** *Comptages des extrémités 5'P par rapport au codon start du gène RT87_08430 (rplm). Le panneau en haut à droite correspond au résultat de la transformée de Fourier rapide (FFT) sur les comptages de la région.*

Par ailleurs, même si le signal était moins fort, une périodicité reste tout de même observable dans les mutants RNase J. Ceci suggère l'existence d'autres facteurs participant à l'apparition du signal, et il n'est donc pas certain que la périodicité observée sur les gènes pris individuellement corresponde à une dégradation par la RNase J ou par d'autres acteurs. En revanche, seule la périodicité de la RNase J semble avoir un biais pour le cadre numéro 2 de traduction. Cette information pourrait donc permettre de faire la différence entre la périodicité due à la RNase J et la périodicité générée par les autres.

Pour identifier les gènes dégradés par la RNase J, nous avons défini un ensemble de critères. Dans un premier temps, afin de filtrer les gènes avec une profondeur de séquençage par position insuffisante, nous avons exclu ceux présentant en moyenne moins d'une extrémité alignée par position comme illustré en Figure 90.

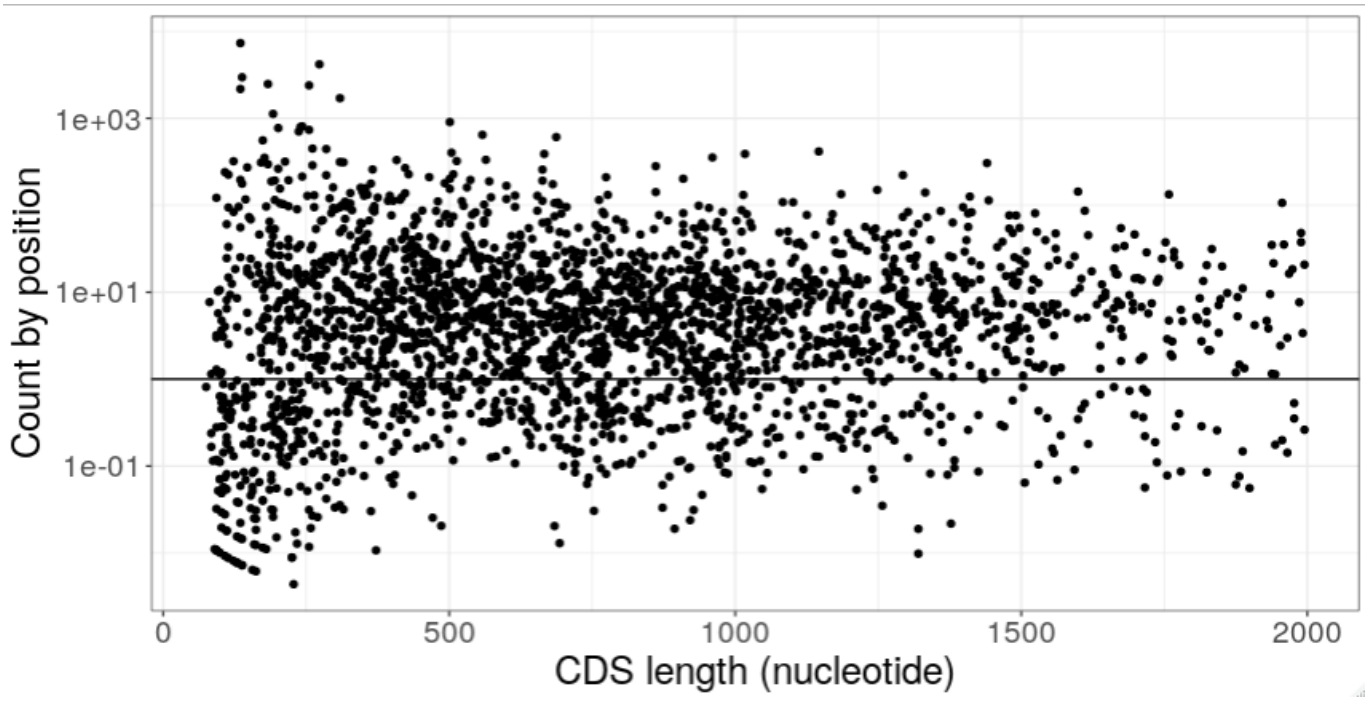


Figure 90 : ***Profondeur de séquençage par position.*** *Moyenne du nombre d'extrémités alignées par position en fonction de la taille des régions codantes des gènes (en nucléotide)*.

Sur les 2703 gènes annotés dans la souche SA564, 2028 présentent, en moyenne, au moins une extrémité 5' alignée par position sur la longueur de la séquence codante.

Par la suite, afin d'identifier les gènes pour lesquels nous observons une période de 3 nucléotides, nous avons besoin d'une méthode qui permette de détecter un pic d'une période de 3 dans les résultats des analyses FFT.
Dans le cadre d'étude de signaux il existe une multitude de méthodes pour faire la détection de pics, comme par exemple des méthodes basées les maximas locaux, sur le calcul de dérivés première et seconde, ou encore des méthodes de détection de valeurs extrêmes (outliers). Pour cette dernière approche, le z-score est une mesure statistique qui décrit l'éloignement d'une valeur par rapport à la moyenne de l'ensemble de valeurs, mesurée en termes d'écarts-types. La formule pour calculer le z-score d'une valeur x est : $z = (x - \mu) / \sigma$ où : μ est la moyenne de l'échantillon et σ est l'écart-type de l'échantillon. Les z-scores peuvent être utilisés pour identifier des valeurs extrêmes dans un ensemble de signaux en quantifiant à quel point chaque signal s'écarte de la moyenne. Si le z-score d'une valeur est très élevé (typiquement au-dessus de 3), cela peut indiquer que cette valeur correspondre à un pic.

Les z-scores basés sur la moyenne et l'écart-type sont sensibles aux valeurs extrêmes parce que ces mesures elles-mêmes peuvent être influencées par des valeurs aberrantes. Une seule valeur très élevée ou très basse peut affecter considérablement la moyenne et l'écart-type, ce qui rend le z-score moins fiable pour la détection de pics. Pour rendre le z-score plus robuste, nous avons utilisé la médiane et l'écart absolu médian (Median Absolute Deviation, MAD) à la place de la moyenne et de l'écart-type.

Nous avons choisi d'utiliser cette approche et pour cela, nous avons effectué un calcul de z-score, basé sur les médianes, sur chaque signal par profil FFT. Cette méthode est simple à mettre en œuvre et semble efficace puisque la Figure 91, représentant les z-scores en fonction des signaux, montre qu'il est possible de séparer les valeurs extrêmes des autres valeurs du profil avec un seuil fixé à 5.

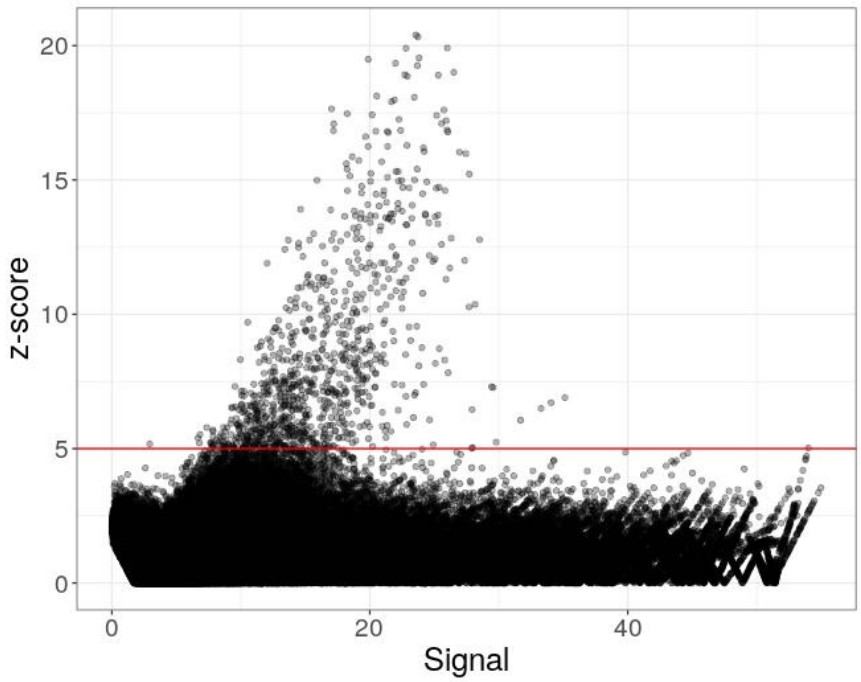


Figure 91 : z-scores par profil en fonction des signaux. *La ligne horizontale rouge représente le seuil de 5 qui a été choisi.*

Avant de calculer le z-score, nous avons procédé à un filtrage des signaux. Les signaux correspondant à des périodes supérieures à 10, jugés peu pertinents, ainsi que ceux égaux à 0, ont été exclus. Cette démarche vise à éviter une médiane artificiellement basse qui pourrait entraîner une surestimation des z-scores.

Parmi les 2 028 gènes, 671 présentent un profil FFT avec un pic à 3 nucléotides plus élevé que les autres et dont le signal a un z-score supérieur ou égal à 5.
L'observation d'une période de 3 nucléotides n'est pas suffisante pour attribuer la dégradation à l'action de la RNase J, nous avons identifié les gènes sur lesquels un biais de distribution des extrémités 5' alignées est observé pour le cadre de lecture numéro 2.

Pour cela nous avons calculé la répartition des intermédiaires de dégradation selon les cadres de lecture pour les 2 028 gènes. Nous n'avons gardé que les gènes où au moins la moité des intermédiaires de dégradation sont situé sur le cadre de lecture numéro 2. Ceci correspond à ne garder que les gènes ayant un pourcentage d'intermédiaire de dégradation en phase 2 supérieur ou égal à 50. Ce seuil est illustré en Figure 92 par un diagramme ternaire.
Nous avons identifié 911 gènes avec un pourcentage d'intermédiaires de dégradation en phase 2 supérieur ou égale à 50.

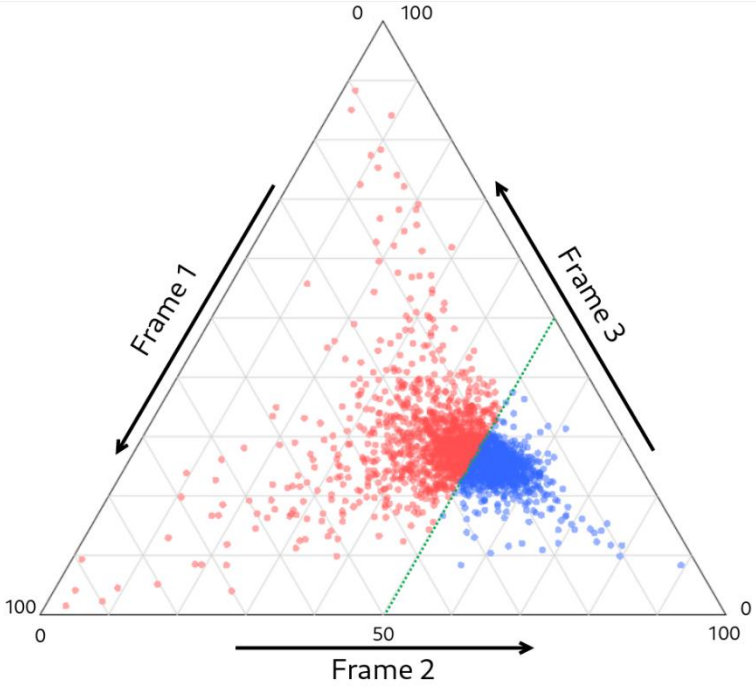


Figure 92 : Diagramme ternaire des pourcentages de lectures alignées dans les trois cadres de lecture possibles par gène.
Chaque point représente un gène, et les trois côtés du triangle représentent les axes portant les pourcentages dans les trois cadres de lecture possibles. La ligne verte correspond à un pourcentage de 50 % dans le cadre numéro 2 et les points bleus représentent les gènes considérés comme ayant un biais en phase 2.

La Figure 92 montre que le seuil choisi de 50 % est arbitraire et qu'une variation de 1 ou 2 % dans ce seuil modifierait l'ensemble de gènes sélectionné. Dans leur article, Nersisyan et al. (2020) effectuent un test de Student par gène à partir des comptages spécifiques de phase. En d'autres termes, pour déterminer s'il y a une différence entre la phase 2 et 1, il s'agit d'une comparaison de moyennes par gène entre les comptages présents sur la phase 2 et ceux sur la phase 1.

Cette approche permettrait d'avoir un seuil moins arbitraire que celui que nous avons utilisé, c'est pourquoi nous projetons d'essayer cette méthode.

Parmi les 2 028 gènes pour lesquels présentant une profondeur de séquençage dépassant le seuil minimum fixé pour détecter un signal, 462 présentent un pic de périodicité de 3 nucléotides ainsi qu'un biais de distribution dans le cadre de lecture numéro 2. Par conséquent, nous considérons que ces 462 gènes sont dégradés par la RNase J1. Bien que ce chiffre soit bien inférieur aux 1 590 gènes ciblés par la RNase J, il est probable que de nombreux faux négatifs soient présents en raison du seuil élevé pour la profondeur de séquençage, notamment pour les gènes faiblement exprimés et/ou courts, pour lesquels il est plus difficile de détecter un signal de périodicité. Pour vérifier ceci, nous pourrions étudier individuellement chacun des profils des gènes ciblés afin de déterminer s'il s'agit vraiment d'un manque de profondeur (peu d'intermédiaires de dégradation ou trop de bruit).
Néanmoins, cette approche offre un nouveau type d'information, nous permettant désormais de déterminer les gènes impactés, ciblés et/ou dégradés par la RNase J1.

Afin de synthétiser les résultats obtenus par l'analyse de la stabilité des ARN *via* leur demi-vie, leur ciblage par la RNase J1 à partir de données générées par le protocole CRAC, ainsi que leur profil de dégradation co-traductionnelle, nous avons comparé les ensembles de gènes formés par les catégories impactés/non impactés, ciblés/non ciblés, et dégradés. Les intersections entre ces ensembles sont calculées sur la population des 1 866 gènes orthologues entre les souches SA565 et JKD6009, pour lesquels nous avons pu calculer une demi-vie. Ces intersections sont présentées en Figure 93 sous la forme d'un lollipop plot, permettant de visualiser les différentes approches et les résultats obtenus.

Tout d'abord, nous avons identifié 913 gènes dont la demi-vie est impactée par la mutation de la RNase J1 (première ligne de la Figure 93), ainsi que 384 gènes dont la stabilité n'était pas impactée (deuxième ligne de la Figure 93), et 564 gènes pour lesquels nous n'avons pas pu déterminer l'un ou l'autre.
Par la suite, l'analyse des données générées par le protocole CRAC nous a permis d'identifier 1590 gènes dont les ARN interagissent directement avec la RNase J1 (troisième ligne de la Figure 93), et 276 pour lesquels aucune interaction n'a pu être mise en évidence (quatrième ligne de la Figure 93).

Enfin, grâce aux données de séquençage des extrémités 5’ des ARN (EMOTE) et au fait qu’une dégradation par la RNase J1 peut se traduire par une périodicité de trois nucléotides en phase 2, nous avons pu identifier 432 gènes dont les ARN sont dégradés par la RNase J1 de manière co-traductionnelle (cinquième ligne).

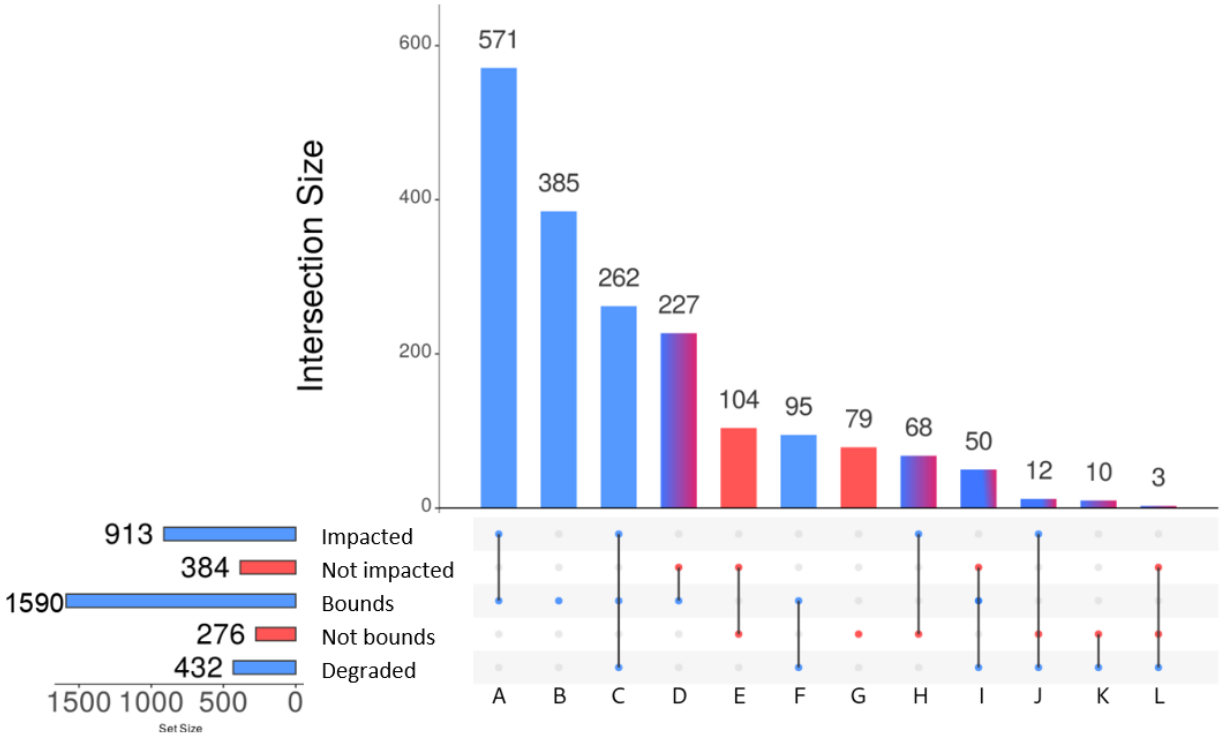


Figure 93 : Nombre de gènes par intersection entre les différents ensembles impactés, non impactés, ciblés, non ciblés, et dégradés. *Ce graphique indique les tailles des intersections entre ces ensembles, permettant d'identifier le nombre de gènes communs ou exclusifs à chaque combinaison. Les barres horizontales à gauche indiquent la taille totale de chaque catégorie de gènes : impactés (913), non impactés (384), ciblés (276), non ciblés (432), et dégradés (1590). Les barres verticales montrent la taille des intersections entre les groupes, avec des lignes reliant les points en bas indiquant les ensembles inclus dans chaque intersection.*

En croisant les résultats de ces approches, nous avons pu identifier 833 gènes ciblés et impactés par la RNase J1 (colonnes A et C) ainsi que 107 gènes non ciblés et non impactés (colonne E et L).

En intégrant les informations issues de la dernière approche, nous avons identifié un total de 262 gènes (colonne C) caractérisés comme impactés, ciblés et dégradés par la RNase J.

Nous avons observé que les gènes non impactés pouvaient néanmoins être ciblés par la RNase J1, et le fait d'en avoir trouvé 50 (colonne I) qui sont également dégradés le confirme.
Cette figure met également en lumière les limites de ces approches. En effet, nous avons observé que des gènes ne sont classés ni comme impactés, ni comme non impactés, et de même, peu de gènes dégradés peuvent être identifiés. C'est pourquoi la colonne B (deuxième plus grand effectif) correspond aux gènes ciblés pour lesquels nous n'avons pas pu déterminer l'impact sur la stabilité, ni la dégradation co-traductionnelle, sans pour autant pouvoir exclure ces phénomènes.
Encore une fois, cela démontre l'intérêt d'utiliser différentes approches complémentaires, chacune offrant une perspective différente. Cette complémentarité des approches permet d'obtenir une vision plus complète et plus nuancée des mécanismes de dégradation des ARN. Par exemple :

- Les ARN peuvent interagir avec la RNase J sans pour autant être dégradés.
- Les ARN peuvent ne pas être impactés par la délétion de la RNase J, mais être tout de même dégradés par cette dernière.
- Les ARN peuvent être dégradés par la RNase J sans subir une dégradation co-traductionnelle marquée ou sur toute leur longueur.

Néanmoins, ces résultats montrent que les trois approches sont concordantes, car il y a peu d'effectifs (93) dans les intersections entre catégories incohérentes, c'est-à-dire des gènes non ciblés mais qui seraient tout de même impactés et/ou dégradés (colonnes H, J, K et L).

2.1.5 Activité endoribonucléasique de la RNase J1

2.1.5.1 125 potentiels sites de clivage endoribonucléolytique de la RNase

Il a été montré *in vitro* que la RNase J avait une activité de dégradation endoribonucléolytique. Cependant, cette activité n'a été étudiée *in vivo* que sur quelques ARN, et aucune étude globale n'a été réalisée dans le but d'identifier les sites de clivage endoribonucléolytiques de la RNase J.
Notre objectif est donc de proposer une étude préliminaire quant à l'identification de ces sites.
Pour ce faire, nous avons utilisé une méthode basée sur celle de Khemici et al. (2015) qui a permis d'identifier ceux de la RNase Y à partir de données de séquençage des extrémités 5' des ARN. En effet, comme expliqué et décrit par (Broglia et al., 2020), d'une coupure endoribonucléolytique résulte une extrémité 5'P en aval du clivage et une extrémité 3'OH en amont du clivage.

A partir des données générées avec le protocole EMOTE que nous avons utilisées pour étudier les profils de méta-comptages sauvage (2 réplicats) et $J1^{AGA}$ (2 réplicats), il serait donc possible d'observer les extrémités 5'P qui devraient être plus abondantes en termes de comptage que les intermédiaires de dégradation exoribonucléolytiques, et devraient disparaître dans la souche J1AGA.

Dans un premier temps, nous avons filtré les positions génomiques ayant moins de 50 CPM (count per million) afin d'éliminer les lectures correspondant à des intermédiaires de dégradation exoribonucléolytiques. Puis, pour identifier les positions dont le nombre de lectures alignées diminue dans la souche $J1^{AGA}$ par rapport à la souche sauvage, nous avons fait une analyse d'expression différentielle avec EdgeR à partir des comptages, normalisé avec la méthode TMMwsp, des positions restantes entre les échantillons issus des deux souches. La normalisation TMMwsp est plus adaptée que la méthode TMM pour des valeurs proches de 0, en raison d'une étape supplémentaire de « winsorization » après avoir calculé les log-ratios qui permet d'ajuster les valeurs extrêmes de ces derniers. Cela peut aider à réduire l'impact des valeurs extrêmes dans l'estimation du facteur de normalisation. Étant donné qu'on s'attend à observer des valeurs réduites à 0 dans le mutant J1AGA, cette normalisation est plus adaptée.

Les positions présentant une diminution statistiquement significative (p-valeur ajustée < 0,05) et importante (supérieure à un facteur 2) du nombre de lecture alignées dans la souche $J1^{AGA}$ par rapport à la souche sauvage sont considérées comme des sites potentiels de clivage endoribonucléolytique de la RNase J.

La RNase J ayant une double activité de dégradation, il ne sera pas possible de différencier les accumulations d'extrémités 5' qui sont dues à des sites de clivage endoribonucléolytique de la RNase J, de celles qui sont dues à des pauses dans son activité exoribonucléolytique. Cette analyse est donc un travail prospectif qui pourra servir de point de départ pour de futures études complémentaires.

Parmi les sites identifiés, certains pourraient correspondre à des sites de clivage endoribonucléolytique de la RNase Y. Nous avons donc également vérifié que ces sites sont toujours exprimés dans une souche dans laquelle la RNase Y a été délétée (Khemici et al., 2015). Comme présenté antérieurement, le profil d'expression génique dans la souche $J1^{AGA}$ avant blocage de la transcription (t_0) est sévèrement modifié.

Dans l'analyse d'expression différentielle entre la souche sauvage et la souche J1^{AGA}, les positions correspondant à des sites de clivage endoribonucléolytique devraient correspondre à des positions avec des comptages d'au moins 50 CPM dans la souche sauvage et avec aucun comptage dans la souche J1^{AGA}. Cependant, il se pourrait que la région dans laquelle se trouve le site considéré ne soit plus suffisamment exprimée dans le mutant J1^{AGA}. Nous avons donc vérifié que les régions concernées sont toujours exprimées en utilisant les données d'expression présentées au début de ce chapitre et utilisée pour estimer les demi-vies des gènes.

Nous avons identifié 237 extrémités 5' ayant une p-valeur ajustée inférieure à 0,05 et un niveau d'expression dans le mutant J1^{AGA} au moins deux fois inférieur à celui observé dans la souche sauvage (Fig. 94).

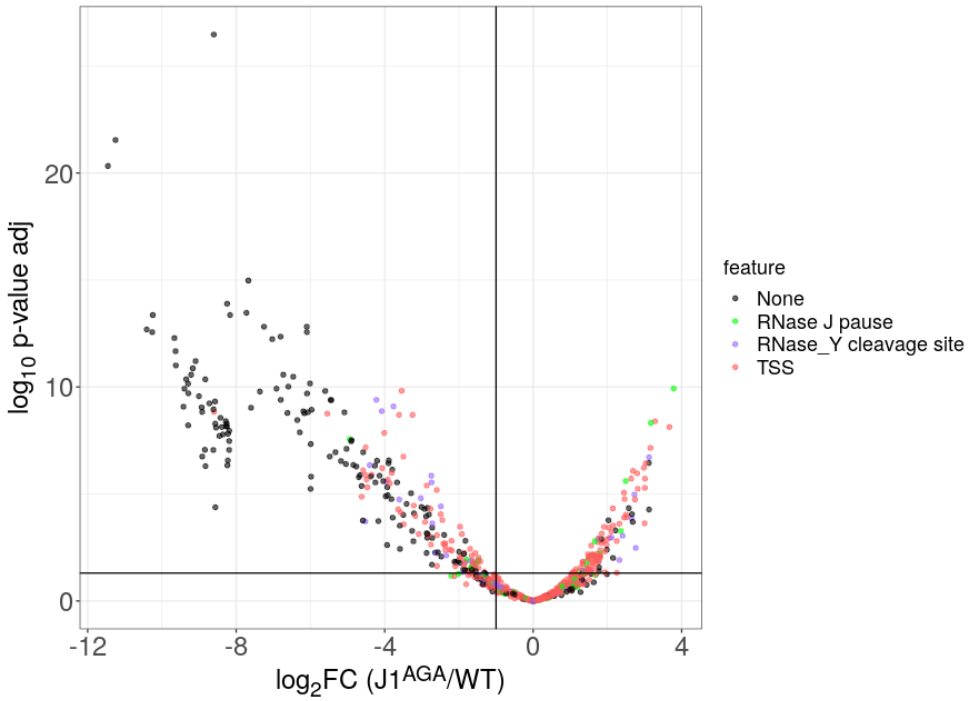


Figure 94 : Variation statistique du nombre de lectures alignées par position entre la souche sauvage et la souche J1^{AGA}. *Volcano plot montrant la p-valeur ajustée -log10 (axe y) par rapport au logFC (axe x). La couleur représente la caractéristique attribuée à ces positions génomiques (cleavage_site : site de clivages endoribonucléolytique de la RNase Y ; tous les autres : début de la caractéristique). La ligne horizontale indique le seuil de p-valeur de 0,01 et la ligne verticale indique le seuil de variation d'un facteur 2.*

Nous avons ensuite filtré les sites correspondant aux TSS ainsi qu'à ceux associés aux pauses de la RNase J, identifiées précédemment comme se situant à 14 et 11 nucléotides en amont des codons d'initiation et à 17 nucléotides en amont des codons d'arrêt de la traduction.

Parmi les 237 positions initiales, 72 ont été exclues car elles avaient déjà été identifiées comme des TSS, et 40 ont été retirées car considéré comme en lien avec les pauses des ribosomes. Nous identifions ainsi 125 sites potentiels pouvant correspondre à des sites de clivage endoribonucléolytique de la RNase J1 parmi lesquels 17 ont également été identifié comme des sites de clivage endoribonucléolytiques de la RNase Y.

Parmi ces 125 sites, 61 présentent une accumulation de comptages dans les conditions WT, passant à zéro dans les souches mutantes de la RNase J. Les 64 autres sites montrent une diminution significative des comptages dans les mutants.
Pour illustrer les résultats obtenus, nous présentons plus en détails deux sites potentiels dont les coordonnées sur le génome sont 1 711 092 et 1 711 167, situés dans le gène codant pour le facteur d'initiation de la transcription IF-3. Dans la Figure 95 illustrant les comptages des extrémités 5' alignées le long du génome, nous pouvons observer deux accumulations d'extrémités 5' à ces positions, tant dans les souches sauvages que mutante pour la RNase Y. Ces accumulations disparaissent dans la souche mutante J1^{AGA} (Fig. 95).



Figure 95 : Distribution des extrémités 5'P autour de la position génomique 1 711 092 par souche. *Profile EMOTE représentant le nombre d'extrémité 5' alignées (axe y) en fonction des positions génomiques (axe x). En bleu, dans une souche sauvage, en rouge dans la souche mutante J1^{AGA} et en rouge dans la souche mutante pour la RNase Y.*

Pour nous assurer que ces sites ne sont pas des sites de clivage de la RNase Y qui n'auraient pas été caractérisés comme tels en raison des différentes méthodes et seuils utilisés par Khemici et al. (2015), nous avons vérifié la présence des 125 sites dans la souche mutante RNase Y. Parmi les 125 sites, 100 sont encore observables dans une souche mutante pour la RNase Y, ce qui renforce l'hypothèse de sites spécifiques à la RNase J1.

(Mathy et al., 2010) ont montré que chez *B. subtilis*, un site de clivage endoribonucléolytique de l'ARN codant pour ThrS par la RNase J1 et la RNase J2 se situe environ (estimé à partir d'un gel d'électrophorèse) 175 ou 280 nucléotides en aval du TSS. Cependant, aucun des 125 sites identifiés ne se situe sur le transcrit codant pour ThrS chez *S. aureus*.

2.1.5.2 Les sites potentiels de clivage de la RNase J1 sont situés juste après une guanine et sont souvent proches d'une région structurée

Dans le but d'identifier des caractéristiques communes aux 125 sites potentiels pouvant expliquer la présence d'une accumulation d'extrémités 5', nous avons d'abord étudié la séquence en ribonucléotides et sa capacité à former des structures secondaires autour de ces sites. Ensuite, nous avons analysé les gènes sur lesquels ces sites ont été identifiés et avons mis en relation ces informations avec les groupes impactés, non impactés, ciblés et non ciblés.

Étant donné qu'il a été démontré, comme illustré en Figure 96, que la RNase Y préfère cliver juste après une guanine (Khemici et al., 2015), nous avons retiré les 15 sites parmi les 125 qui avaient également été identifiés comme des sites de clivage de la RNase Y. En effet, bien qu'il soit possible que les deux RNases partagent le même site de clivage, nous avons préféré les exclure de cette analyse pour éviter de perturber ou de brouiller un éventuel signal spécifique à la RNase J1.

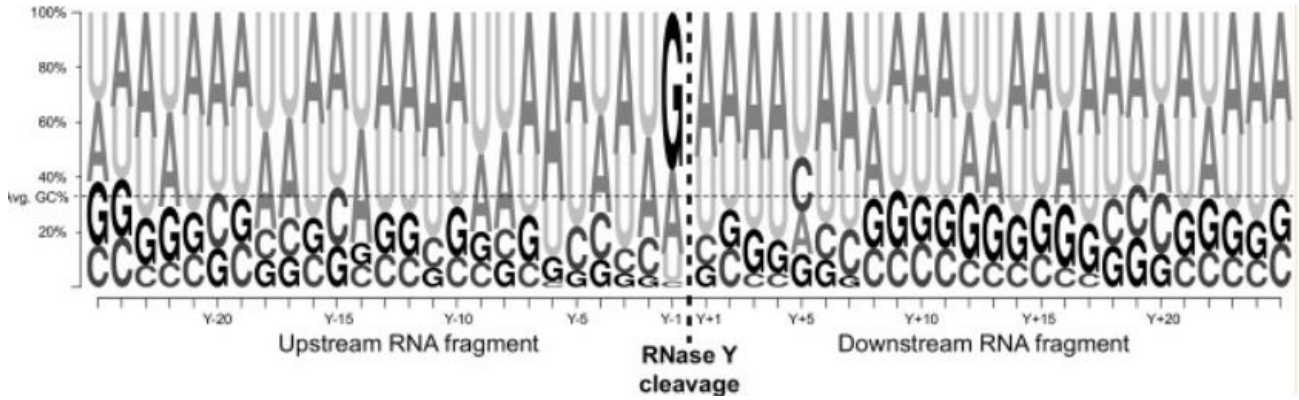


Figure 96 : Alignement des séquences autour des 99 sites de clivages de la RNase Y chez S. aureus. *Extraite de* (Khemici et al., 2015)

La distribution des ribonucléotides autour des 110 (125-15) potentiels sites de clivage de la RNase J1 est présentée en Figure 97. Les résultats montrent qu'une guanine est présente juste avant le site de clivage dans environ 45 % des cas et qu'une guanine est également présente 25 % des cas, 7 nucléotides en amont du site de clivage. Étant donné que la fréquence des guanines dans le génome de *S. aureus* est de 17 %, ces observations sont particulièrement intéressantes.
Bien que ce signal soit trop peu spécifique pour constituer à lui seul un motif de reconnaissance, sa fréquence accrue par rapport au génome suggère une certaine pertinence. Il est possible que ce signal soit combiné à d'autres facteurs, rappelant la distribution des sites de clivage de la RNase Y, ce qui pourrait expliquer pourquoi certains sites sont communs aux deux ribonucléases.

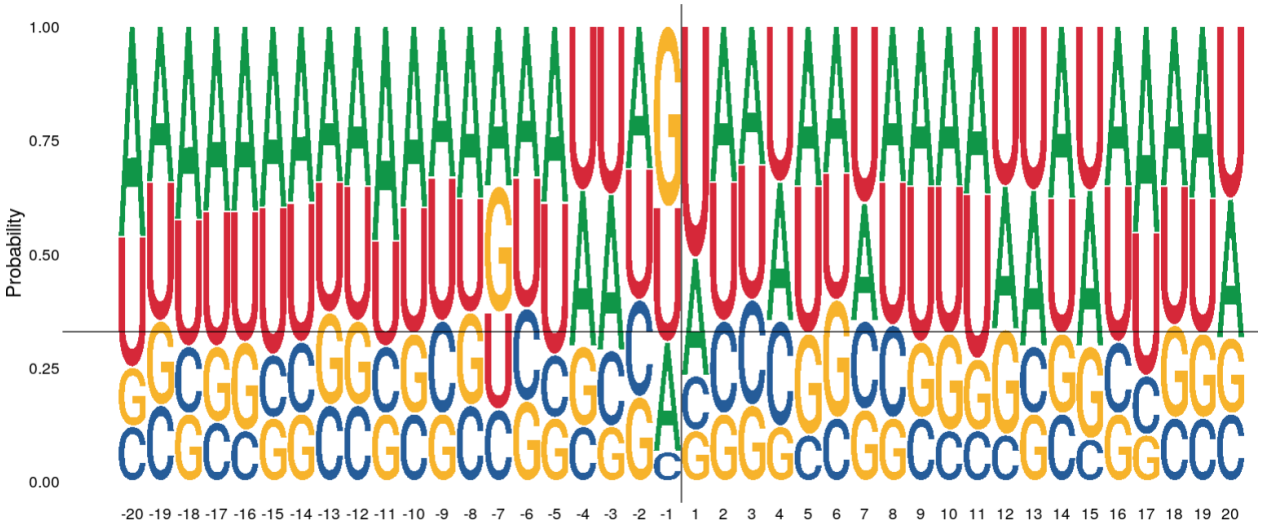


Figure 97 : Alignement des séquences autour des potentiels sites. *Distribution de la composition en ribonucléotides des régions autour des sites potentiels de clivage endoribonucléolytique de la RNase J. La ligne verticale indique le site de clivage, et la ligne horizontale, le pourcentage en GC de S. aureus.*

Les programmes de prédiction des structures de l'ARN prédisent systématiquement des structures, qui peuvent parfois être proches de la réalité. Cependant, il est bien établi que ces prédictions sont souvent inexactes. De plus, ces programmes sont sensibles aux régions considérées : l'ajout ou la suppression d'un nucléotide peut modifier complètement certaines prédictions.
Ainsi, nous cherchons une mesure capable d'indiquer la probabilité qu'un nucléotide, dans son contexte, soit impliqué dans une structure. Pour ce faire, nous avons adopté la méthode décrite par (Broglia et al., 2020) et que nous avons illustré en Figure 98, qui considère des régions de 25 nucléotides autour du nucléotide d'intérêt. Ce nombre est choisi pour offrir un contexte suffisant sans diluer les effets locaux.

Nous calculons ensuite, à l'aide du programme RNAfold, un score qui correspond à la moyenne des énergies libres prédites pour ces 25 régions. Une valeur élevée d'énergie libre à une position donnée indique que ce nucléotide a moins de chances d'être impliqué dans une structure. En revanche, une énergie libre basse suggère qu'un grand nombre de structures secondaires pourraient se former dans la région de ce nucléotide.

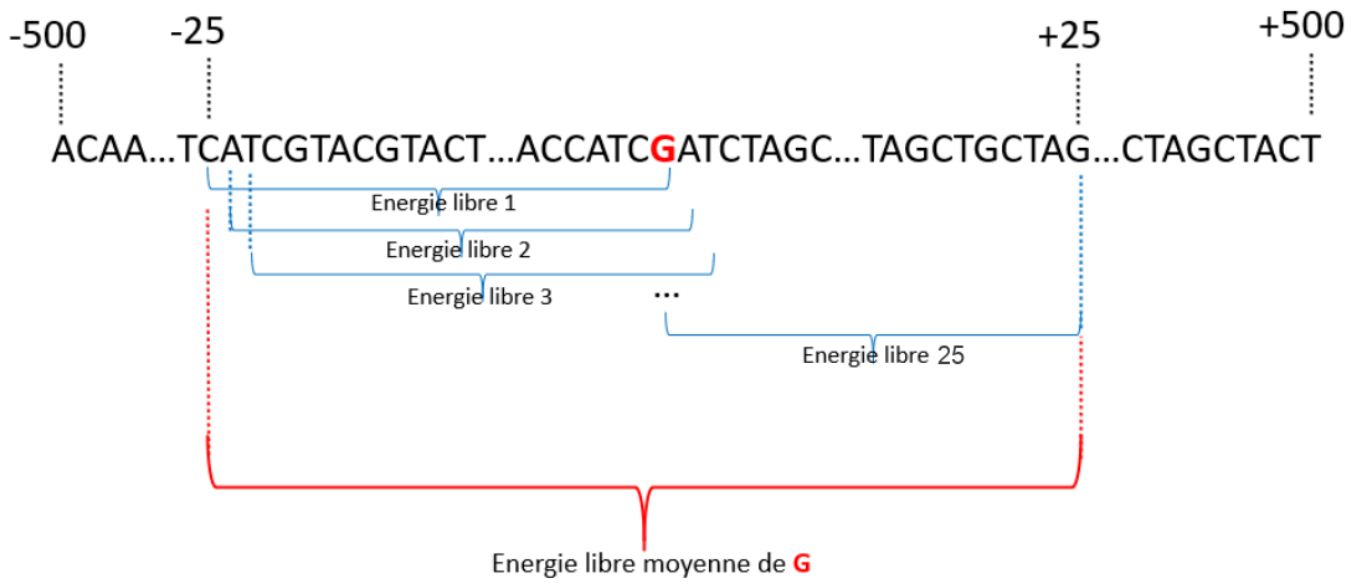


Figure 98 : Méthode d'estimation de l'énergie libre d'un nucléotide.

Le profil des scores (en KJ/mol) par position génomique autour des sites précédemment illustrés (position 1 711 092 et 1 711 167) est représenté en Figure 99. Nous constatons une forte diminution de l'énergie libre moyenne à proximité des deux sites (indiqués par les lignes verticales orange), présumés être des sites de clivage par la RNase J1.

Nous avons observé ce type de profils dans la majorité des 125 sites, suggérant que la proximité d'une structure secondaire pourrait expliquer leur présence. Cependant, la présence de régions structurées pourrait également induire des pauses de la RNase J1 pendant son activité exoribonucléasique.

Il serait intéressant d'étudier les gènes concernés par les potentiels sites de clivage, et de mettre en parallèle ces résultats avec les groupes de gène ciblés, non ciblés, impactés et non impactés.

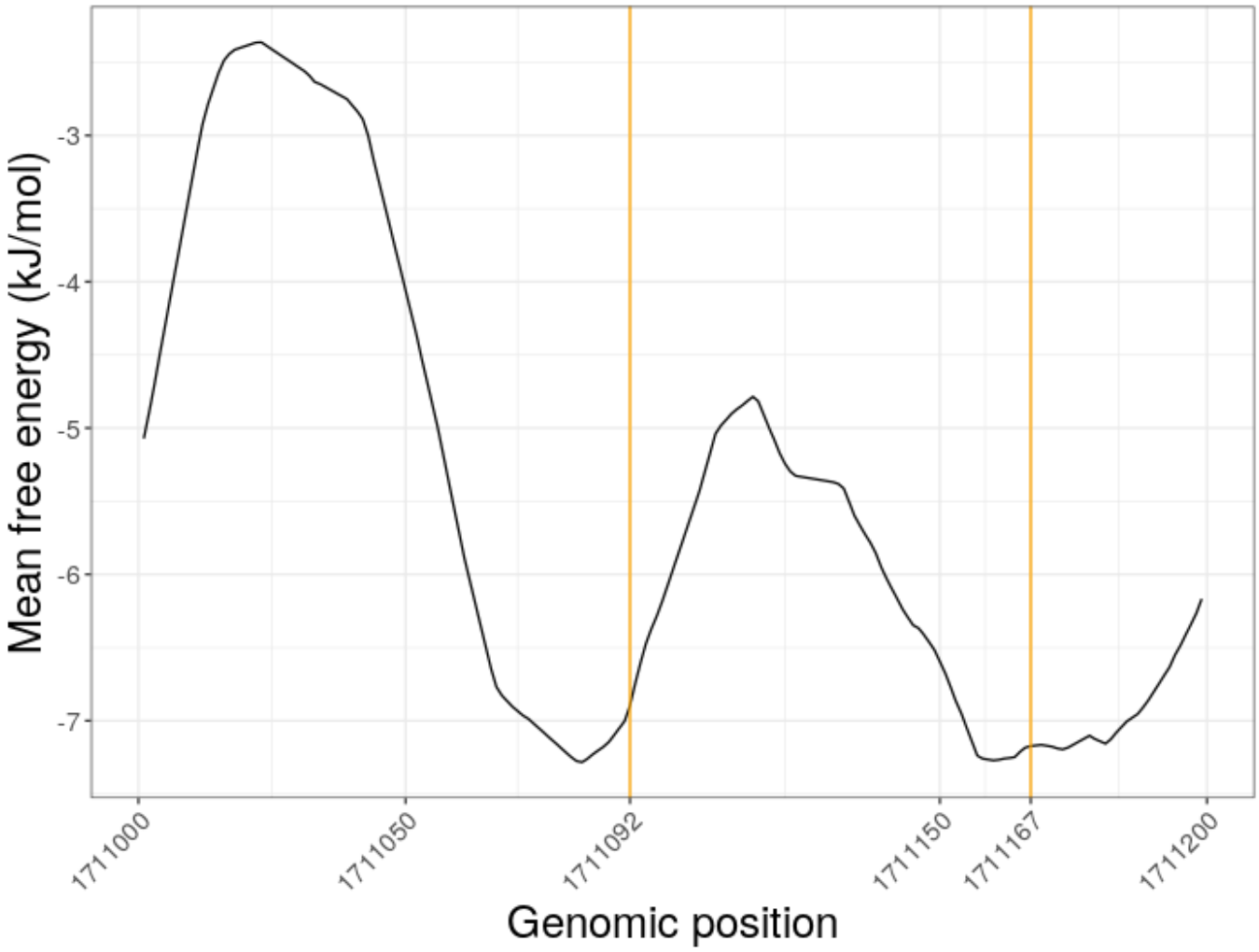


Figure 99 : ***Structuration de la région autour des sites 1 711 092 et 1 711 167.*** *Énergie libre en fonction des positions génomique autour des deux sites potentiel situés aux positions 1 711 092 et 1 711 167 (lignes verticales jaunes) qui représente l'énergie libre (axe x) par position génomique (axe y).*

Sur ces 125 sites, 76 se trouvent à l'intérieur d'un cadre ouvert de lecture. 36 sites sont situés dans la région codante ou dans la région 5'-UTR de gènes codants pour des protéines ribosomales. Ces résultats suggèrent que la RNase J pourrait jouer un rôle dans la maturation de certains transcrits impliqués dans l'assemblage des ribosomes. Toutefois, il est également possible que la sur-représentation de ces sites soit due à l'expression élevée des gènes codant pour les protéines ribosomales.

Nous avons ensuite mis en relation les gènes sur lesquels ont été identifié un site de clivage potentiel avec les groupes de gènes impactés, non impactés (Table 9 A), de gènes ciblés et non ciblés (Table 9 B) et les combinaisons de ces quatre groupes (Table 9 C).

(A)

Groupe	Nombre de sites
Impacté	86
Non impacté	10
Non déterminé	29

(B)

Groupe	Nombre de sites
Ciblé	120
Non Ciblé	3
Non déterminé	2

(C)

	Impacté	Non impacté	Non déterminé
Ciblé	83	9	28
Non ciblé	2	1	0
Non déterminé	1	0	1

Table 9 : ***Distributions des sites potentiels de clivage de la RNase J1*** *dans :*

- *les groupes de gènes impactés et non impactés*
- *les groupes de gènes ciblés et non ciblés*
- *les groupes de gènes ciblés, non ciblés, impactés et non impactés*

Nous observons que la majorité des sites potentiels de clivage (86 sur 125, soit 68,8 %) se situent sur des gènes impactés, et parmi ceux-ci, 120 sur 125 (soit 96 %) ont été identifiés comme des cibles directes de la RNase J1. Ces résultats sont prometteurs, car ils suggèrent que les gènes présentant des sites de clivage sont également caractérisés par une demi-vie impactée par la RNase J1 et une interaction directe avec cette dernière.

Par ailleurs, si l'ensemble des 125 sites potentiels sont effectivement des sites de clivage endoribonucléolytique de la RNase J1, c'est plus que les 99 sites de clivages de la RNase Y. Ceci renforce les conclusions quant au rôle majoritaire de la RNase J dans la dégradation des ARN chez *S. aureus*.

Cependant, parmi les 833 gènes ciblés et impactés, seulement 83 sites ont été identifiés, représentant environ 10 %. Cela pourrait indiquer que l'activité endoribonucléolytique de la RNase J1 est probablement minoritaire.

Il est également envisageable que notre approche et le seuil de filtrage que nous avons utilisé n'aient pas permis de détecter l'ensemble des sites de clivage endoribonucléolytique de la RNase J1 présents dans les conditions étudiées.

Malgré ces premiers résultats encourageants, il est possible que les accumulations que nous attribuons, avec cette approche, à l'activité endoribonucléasique de la RNase J1, pourraient également correspondre à des pauses dans son activité exoribonucléasique 5'→3'. Par ailleurs, étant donné que les deux activités de la RNase J partagent le même site actif, il est complexe d'inactiver spécifiquement l'activité 5'→3' exoribonucléasique. Néanmoins, comme discuté dans l'introduction, la dimérisation de la RNase J pourrait jouer un rôle crucial dans la modulation de ses deux activités. Cette observation pourrait ouvrir la voie à de futurs projets de recherche.

Pour pouvoir distinguer les pauses et le clivage par la RNase J, il serait nécessaire de quantifier également les extrémités 3'OH résultant d'un clivage endoribonucléolytique. Pour cela, il est envisageable d'analyser les données de séquençage des extrémités 3' des ARN, comme le montrent (Broglia et al., 2020)**,** afin de vérifier la présence d'une extrémité 3'OH juste en amont du site de clivage. En effet, les pauses ne génèrent pas ce type d'extrémités.

2.1.6 Discussion

Dans ce chapitre, nous avons présenté une analyse globale visant à étudier la dégradation des ARN chez *S. aureus* et le rôle des RNase J, principalement la RNase J1, dans les mécanismes associés. Nous avons également exposé des analyses variées sur les substrats de la RNase J1 dans le but d'identifier des caractéristiques communes à ces ARN.

L'étude de l'impact de la dégradation par les RNases J est complexe car elles possèdent deux activités de dégradation *via* le même site catalytique : une activité exonucléasique en 5' des ARN et une activité endoribonucléasique. De plus, la présence de deux paralogues, la RNase J1 et la RNase J2, complexifie le modèle, car la multimérisation en un complexe RNase J permet de moduler l'une et/ou l'autre des activités. De plus, bien que ces deux paralogues partagent les mêmes fonctions enzymatiques et interagissent l'une avec l'autre, il a été montré qu'elles jouent des rôles différents dans la dégradation des ARN et la régulation de l'expression génique.

Nous avons montré que les ARN de 47 % des gènes de *S. aureus* exprimés dans les conditions d'étude ont une stabilité qui augmente lorsque le site actif de la RNase J1 est muté ou lorsque que les deux RNases J sont délétées. Cela signifie que près de la moitié des gènes a des ARN dont la

dégradation dépend de la RNase J1. En comparaison, la délétion de la RNase E chez *E. coli* augmente la stabilité des ARN de 60 % des gènes (Stead et al., 2011). Chez *B. subtilis*, la délétion de la RNase Y entraîne l'augmentation de la stabilité des ARN de 25 à 30 % des gènes (Laalami et al., 2014), et la délétion de la RNase J entraîne l'augmentation du niveau d'expression de 30 % des gènes (Durand et al., 2012)**.**

Par la suite, nous avons montré que les gènes, pour lesquels la mutation du site actif de la RNase J1 ou la délétion des deux gènes codant pour les RNases J n'augmentent pas la stabilité des ARN, pouvaient être quand même dégradés par la RNase J1. En effet, l'étude proposée dans ce chapitre visant à identifier l'interactome RNase J1-ARN (données générées avec le protocole CRAC), nous a permis de montrer que cette dernière interagit avec les ARN de 85 % des gènes exprimés dans les conditions d'étude. Par comparaison, des expériences similaires chez *E. coli* ont montré que la RNase E interagit avec les ARNm de 75 % des gènes (S. A. Waters et al., 2017). Ceci montre que la RNase J1 est en mesure de se lier *in vivo* avec presque l'ensemble des ARN et donc potentiellement les dégrader.

L'ensemble des résultats présentés dans ce chapitre supportent le fait que la RNase J1 est l'actrice majeure de la dégradation des ARN chez *S. aureus* puisqu'elle serait en mesure de dégrader les ARN de 85 % des gènes et serait nécessaire pour la dégradation des ARN de 47 % des gènes.

Nous avons également montré que la RNase J1 est la principale contributrice au phénomène de dégradation co-traductionnelle, indiquant par la même occasion que son activité de dégradation exoribonucléasique 5'→3 ' est au moins aussi rapide que l'activité de traduction du ribosome. Grâce à ces résultats nous avons développé une nouvelle approche permettant l'identification des gènes dont les ARN sont dégradés de manière co-traductionnelle par la RNase J1. Nous avons identifié que c'est le cas pour 21 % des gènes exprimés dans les conditions d'étude, ce qui est peu comparé au 85 % de gènes ciblés. Cette approche nécessite cependant une profondeur de séquençage considérable et une optimisation des seuils ou de la méthode de détection de la période de 3 nucléotides en phase 2 de traduction pour améliorer sa sensibilité et permettre la diminution du nombre de faux négatifs.

Néanmoins, ceci permet de montrer qu'en plus d'être l'actrice majeure de la dégradation des ARN, la RNase J1 fait également partie intégrante de la dynamique d'un système où transcription, traduction et dégradation peuvent avoir lieu simultanément, permettant une grande rapidité dans la régulation du profil d'expression de gène.

En revanche, nous avons découvert que les pauses des ribosomes sur les codons de début et de fin de traduction ne se traduisent pas de la même manière chez *S. aureus* que chez les autres

organismes, y compris *B. subtilis*. Par exemple, nous avons mis en évidence un blocage de la RNase J1 à 14 nucléotides en amont du codon d'initiation de la traduction. Ce blocage est suivi (3 nucléotides après) d'un autre blocage, à 11 nucléotides en amont du codon d'initiation. La présence de ces deux blocages successifs n'a pas encore été observée chez d'autres bactéries, y compris chez celles possédant également des RNases J. Actuellement, nous n'avons aucune hypothèse pouvant expliquer ce deuxième blocage.

Par ailleurs, un autre blocage de la RNase J1 a également été observé 17 nucléotides en amont du codon d'arrêt. Un blocage similaire a été identifié chez *Saccharomyces cerevisiae* qui n'a pourtant pas de RNase J mais une autre exonucléase 5'→3' XRN1, et chez *B. subtilis,* ce blocage se situe 14 nucléotides en amont du codon d'arrêt.

Il pourrait paraître surprenant que pour des organismes phylogénétiquement proches tels que *S. aureus* et *B. subtilis* ayant des séquences conservées des RNase J, on observe une différence à l'échelle moléculaire avec un écart de 3 nucléotides pour ces blocages. Il est probable que cet écart soit dû à un encombrement stérique légèrement différent des RNases J.

Bien qu'un modèle de régulation du métabolisme de l'ARN à partir d'un organisme comme *B. subtilis* puisse servir de référence et soit conservé, nos résultats montrent que l'accumulation de particularités et/ou de subtilités peuvent motiver l'élaboration d'un modèle spécifique à l'organisme d'intérêt.

Par ailleurs, Hausmann et al. (2017) a montré que chez *S. aureus*, contrairement à d'autres organismes tels que *M. smegmatis*, *T. thermophilus* et *S. coelicolor*, la RNase J est peu sensible aux états de phosphorylation des extrémités 5' des ARN. Nous avons confirmé ces résultats en n'observant aucune différence statistiquement significative de proportion d'extrémité 5'P entre les gènes impactés et les gènes non impactés. Ces résultats soulignent encore une fois la particularité de la RNase J1 de *S. aureus* par rapport à celle des autres organismes.

En étudiant les caractéristiques des gènes impactés et non impactés par la RNase J1, nous avons observé que la RNase J1 ne semble pas particulièrement sensible à la composition en ribonucléotides des extrémités 5' des ARN, ou à des structures présentes sur celles-ci. Cela suggère que la RNase J1 serait capable de dégrader tous les ARN, à l'exception de quelques cas particuliers qu'il serait intéressant d'étudier.

Nous avons également montré que la RNase J1 impacte et cible davantage les gènes qui ne sont pas en premières position sur un ARN polycistronique. Cela pourrait signifier que la dégradation des ARN du premier gène n'est pas dépendante de la RNase J1, ou que la RNase J1 dégrade avec

une légère préférence les gènes qui suivent suite à l'action d'autres facteurs qui restent à identifier. Néanmoins, la double activité de la RNase J1 signifie qu'elle pourrait initier la dégradation de certains ARN par un clivage endoribonucléolytique, puis les dégrader de manière exoribonucléolytique.

Néanmoins, notre étude préliminaire sur l'identification des sites de clivage endoribonucléolytique de la RNase J1 n'a révélé que 125 sites potentiels, soit un nombre de cibles relativement restreint. Il serait néanmoins pertinent de poursuivre cette recherche. La majorité de ces 125 sites ont été localisés à proximité de structures secondaires, ce qui pourrait indiquer que la RNase J1 est capable de cliver les ARN structurés pour en permettre l'accès et la dégradation. Cela expliquerait l'impact étendu de la RNase J1 sur divers ARN et sa capacité d'interaction avec un grand nombre d'ARN. De plus, parmi ces 125 sites, 17 se chevauchent avec les sites de clivage endoribonucléolytique de la RNase Y, suggérant une redondance fonctionnelle. Cette redondance pourrait permettre à la RNase J1 de compenser partiellement la perte de la RNase Y, ce qui expliquerait que la délétion de cette dernière n'engendre que peu d'effet phénotypique (réduction de la pathogénicité).

La conservation évolutive des RNases J1 et J2 indique un intérêt pour les organismes à posséder les deux enzymes. Notre travail, qui s'est principalement porté sur l'étude de la RNase J1, a bien montré l'utilité de la RNase J1 pour les cellules. Toutefois, la question reste ouverte quant à la persistance de la RNase J2, qui semble enzymatiquement peu active, ce que nous avons pu également confirmer par le fait qu'elle ne permet pas de compenser la perte du site actif de la RNase J1 dans la souche mutante $J1^{AGA}$. De plus, sa délétion complète, ajoutée à celle de la RNase J1 dans la souche double mutante ΔJ1Δ2 n'apporte pas d'augmentation de stabilité supplémentaire. Pour étudier plus spécifiquement l'impact de la RNase J2 sur la stabilité des ARN, il serait nécessaire de réaliser des expériences sur une souche mutante pour le site actif de la RNase J2. Cette souche nous permettrait également d'étudier son activité endoribonucléolytique. En outre, une souche mutante où la RNase J2 serait complètement délétée pourrait également nous permettre d'évaluer son impact sur l'activité de la RNase J1.

Par ailleurs, certains des résultats présentés dans cette partie, notamment ceux concernant l'étude sur les demi-vies et ceux sur l'étude de la périodicité, seront soumis pour une publication scientifique dont la rédaction est en cours de finalisation.

Dans le chapitre suivant, nous nous intéresserons de nouveau à la dégradation des ARN chez *S. aureus*, mais cette fois-ci à travers l'étude d'un mécanisme particulier : la déphosphorylation des extrémités 5' des ARN. Bien que nous ayons démontré, à l'instar de Hausmann et al. (2017), que les états de phosphorylation des extrémités 5' des ARN ne jouent pas un rôle prépondérant dans

l'activité de la RNase J1, il a été montré qu'ils pouvaient légèrement moduler son activité. Les états de phosphorylation des extrémités 5' des ARN n'ont jamais été étudiés à l'échelle globale chez *S. aureus*. C'est pourquoi, dans le prochain chapitre, nous présentons une analyse globale de ces états.

2.2 Chapitre 2 : États de phosphorylation des extrémités 5’ des ARN et identification de la NUDIX chez *S. aureus*

2.2.1 Candidats NUDIX et quantification des états de phosphorylation *in vivo* des extrémités 5’ des ARN

La dégradation d’un ARN bactérien peut commencer par la conversion de son extrémité 5′ triphosphates (5’PPP) en 5’ monophosphate (5’P), créant ainsi un meilleur substrat pour la dégradation par certaines ribonucléases (RNase E chez *E. Coli* et RNase J chez *B. subtilis* et *S. aureus*).

À notre connaissance, chez les bactéries, les seules enzymes identifiées comme étant capables de déphosphoryler ces extrémités appartiennent à la famille des protéines NUDIX hydrolases. Comme évoqué en introduction, les NUDIX capables d’hydrolyser les phosphates β et γ des extrémités 5’ des ARN sont appelées « RNA PyroPhosphoHydrolase (RppH). Certaines de ces RppH ont été identifiées chez *E. coli* (Mackie, 1998)**,** (Celesnik et al., 2007)**,** (Deana et al., 2008), *H. pylori* (Bischler et al., 2017) ou encore *B. subtilis* (Richards et al., 2011). Chez *S. aureus*, (Bonnin & Bouloc, 2015) ont identifié une protéine RppH par homologie et synténie avec celle de *B. subtilis*. Cependant, leurs résultats montrent que la délétion de cette RppH ne stabilise que 4 transcrits, ce qui suggère la présence d’une ou plusieurs autres RppH. Notre objectif est d’identifier d’autres enzymes capables d’effectuer cette activité de déphosphorylation. Nous avons donc étudié l’impact des mutants de délétions des enzymes NUDIX (construits par l’équipe collaboratrice) sur les états de phosphorylation des extrémités 5’ des ARN et comparer ces résultats avec ceux issus d’une souche sauvage, par le traitement statistique des résultats d’une approche *in vivo* à l’échelle du génome.

Acquisition des données biologiques :

Pour identifier les protéines NUDIX codées par le génome de la souche N315 de *S. aureus*, nous avons recherché les produits des gènes de cette souche présentant une similarité avec l'un des cinq profils Pfam (PF00293, PF14443, PF09296, PF14815, PF13869) correspondant à la famille des NUDIX (Srouji et al., 2017). Chaque profil a été utilisé comme requête avec le programme *hmmsearch* de la suite logicielle HMMER3 (Eddy, 2011) contre l’ensemble des protéines codées par le génome. Le meilleur hit a été conservé. Nous avons identifié cinq pyrophosphohydrolases NUDIX potentielles (SA0425, SA1330, SA1501, SA1612 and SA2278). Une recherche réciproque par *BLASTP* utilisant soit chaque hydrolase NUDIX de *S. aureus* contre le protéome de *B. subtilis*

168, soit chacune des six hydrolases NUDIX de *B. subtilis* contre le protéome de *S. aureus*, n'a permis d'identifier que trois paires de séquences similaires : SA1330 avec *Bs*NudF (69 % de similarité) ; SA1501 avec *Bs*YmaB (57 % de similarité) et SA1612 avec *Bs*RppH (58 % de similarité). La protéine SA1612 correspond à la RppH identifiée et étudiée par Bonnin and Bouloc (2015).

Par la suite, l'équipe collaboratrice a construit des souches mutantes de *S. aureus* dans lesquelles les candidats ont été délétés de manière à pouvoir étudier l'impact des délétions sur les états de phosphorylation des extrémités 5' des ARN. Comme évoqué en introduction, les NUDIX peuvent avoir des fonctions redondantes entre elles, c'est-à-dire que d'autres NUDIX peuvent compenser la perte de fonction d'une NUDIX. En créant des simples mutants (un seul candidat délété), cette compensation peut masquer l'effet de la délétion, et une autre protéine NUDIX pourrait réaliser la déphosphorylation de l'extrémité 5' des ARN. Ceci rend difficile l'identification de la fonction spécifique de la protéine, d'où la nécessité de construire des mutants multiples.

A partir de ces cinq candidats, huit souches mutantes ont été construites, trois simples mutants (Δ0425 et Δ1612 testés phénotypiquement, et Δ2778 non testé phénotypiquement), deux doubles mutants (Δ1612-Δ1330 et Δ1612-Δ0425 non utilisé pour l'analyse des extrémités 5' des ARN), deux triples mutants (Δ1612-Δ1330-Δ0425 et Δ1612-Δ0425- Δ2778) et finalement un quadruple mutant (Δ1612-Δ1330-Δ0425- Δ2778). Pour le moment, aucune souche ne comporte de délétion pour le candidat SA1501 en raison de son essentialité qui pose de nombreux défis pour la construction d'une souche mutante. Des analyses sur les phénotypes de croissance des différents mutants ont été menées par Gabriel Tejero (Equipe Redder). Ces analyses montrent que les mutants NUDIX ne présentent aucun phénotype de croissance par rapport à une souche sauvage, contrairement au mutant J1^{AGA} qui, comme montré par Linder et al. 2014, a une croissance plus lente (Fig. 100 et 101).

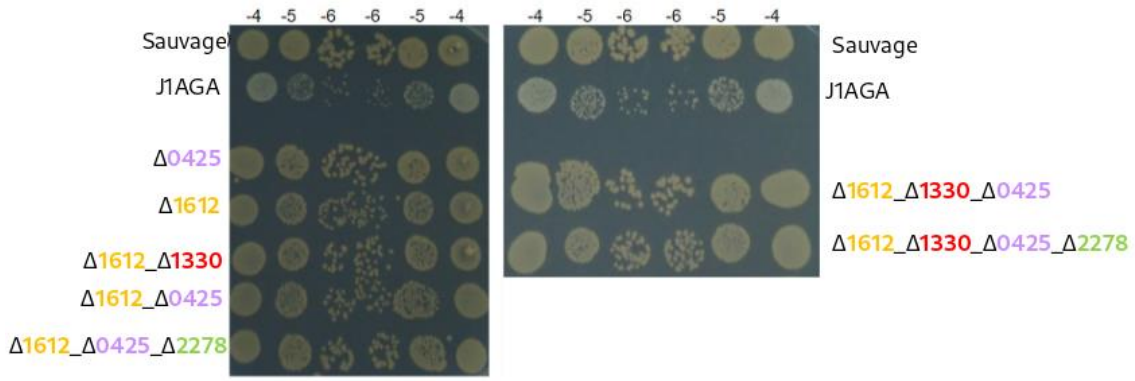


Figure 100 : ***Croissance des bactéries S. aureus après 48h en fonction des souches.***

Les colonnes représentent les niveaux de dilution. (Adapté du rapport de stage Gabriel Tejero)

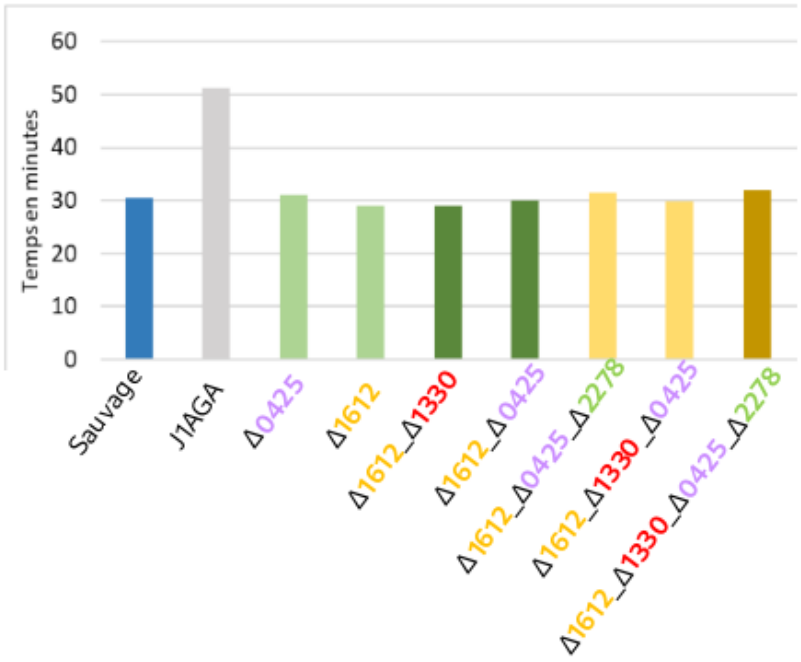


Figure 101 : Temps de génération moyen de chaque souche en minutes. *(Adapté du rapport de stage de Gabriel Tejero)*

Pour étudier l'impact des mutants de délétions des enzymes NUDIX sur les états de phosphorylation des extrémités 5' des ARN, et comparer ces résultats avec ceux issus d'une souche sauvage, une approche *in vivo* à l'échelle du génome, PQ-EMOTE (Phosphorylation Quantification EMOTE) a été développée par l'équipe de notre collaborateur, Peter Redder. Jusqu'à présent, le pourcentage d'ARN 5'P *in vivo* n'a été mesuré que pour quelques dizaines de transcrits. Ces résultats sont insuffisants pour examiner les effets régulateurs que la conversion en 5'P pourrait avoir sur la dégradation globale de l'ARN. PQ-EMOTE est une nouvelle méthode transcriptomique permettant de distinguer les extrémités 5'P des extrémités 5'PP et 5'PPP des ARN et de quantifier le pourcentage de molécules 5'P pour chaque espèce d'ARN. Ce protocole est composé de plusieurs étapes (Fig. 102).

Comme généralement dans des expériences de RNA-Seq, l'ARN total est purifié et contient un mélange d'ARN 5'PPP et 5'P (lignes bleues et rouges épaisses, respectivement de la Fig. 102). Chaque échantillon d'ARN est traité avec un kit de ribodéplétion pour retirer un maximum d'ARNr. Il convient de noter que la très grande abondance des transcrits d'ARNr dans les cellules ne permet pas à ce traitement d'éliminer l'ensemble des séquences d'ARNr dont une quantité résiduelle variable persiste dans les échantillons. Chaque échantillon est ensuite divisé en deux : **pool^{mono}** et **pool^{total}**, mais seul ce dernier est traité avec une RppH commerciale de *E. coli*, EcRppH, pour convertir tous les ARN 5'PPP en ARN 5'P. L'ARN T4 ligase 1 est utilisée pour lier un oligoribonucléotide synthétique (**vert**) à chaque molécule d'ARN 5'P. La transcription inverse est effectuée avec des amorces aléatoires (flèches noires) qui sont suivies par un adaptateur Illumina (**fine ligne cyan**). L'ADNc résultant n'est amplifié par PCR que si les molécules d'ARN d'origine ont été liées à l'oligonucléotide synthétique (c'est-à-dire que les ARN dont l'extrémité est 5'PPP dans **pool^{mono}** ne seront pas amplifiés). Les produits PCR sont ensuite purifiés par taille sur gel et les produits de petite taille (~250 pb) sont excisés afin de supprimer les produits PCR trop courts Ces produits peuvent notamment être dus à une hybridation des amorces aléatoires trop près de l'extrémité 5' des transcrits, entraînant des produits PCR courts avec trop peu d'information. Le reste des produits PCR est ensuite séquencé.

À partir des souches mutantes construites, quatre librairies PQ-EMOTE ont été générées (Table 10). Dans chacune des librairies, des échantillons issus d'une souche sauvage ont également été préparés. Ces librairies ont été réalisées au cours de plusieurs années, pendant lesquelles les paramètres du protocole ont évolué pour son amélioration, et les différences sont représentées dans la Table 10. La librairie PQ-EMOTE 62 est celle qui diffère le plus des autres puisque ses expériences ont été menées en quadruplicat, les lectures du séquençage sont plus longues (100 pb), et l'oligonucléotide RpX lié lors du protocole est différent, Rp10 à la place de Rp8 (les différences entre les deux seront expliquées plus loin). De plus, l'étape de purification sur gel a été effectuée avec un seuil de filtrage plus bas, ce qui signifie que certains ARN non séquencés dans les autres librairies en raison de leur faible taille seront retrouvés dans la librairie PQ-EMOTE 62.

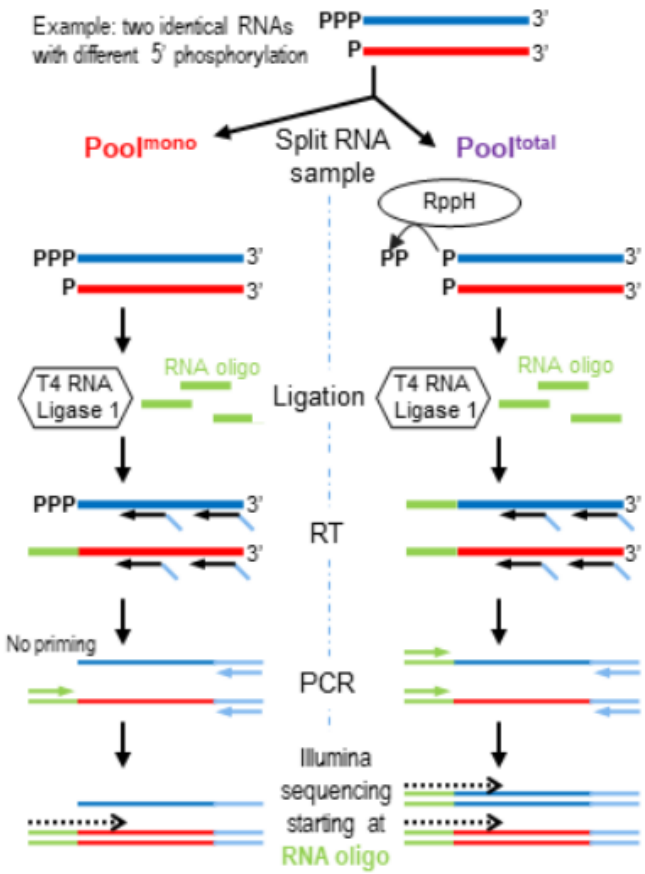


Figure 102 : Préparation de la librairie de séquençage PQ-EMOTE

Purification de l'ARN total, traitement par ribodéplétion pour éliminer l'ARNr, division en deux échantillons (pool^{mono} et pool^{total}), traitement du pool^{total} avec EcRppH pour convertir l'ARN 5'PPP en 5'P. Liaison de l'ARN 5'P avec un oligoribonucléotide synthétique, transcription inverse avec amorces aléatoires, ajout d'un adaptateur Illumina, amplification PCR des ARN liés. Purification et excision des produits PCR courts (~250 pb), puis séquençage.

Librairie	Souche	Nombre de réplicats	Illumina	Oligonucléotide RpX
PQ-EMOTE 49	Sauvage	2	50 pb	Rp8
PQ-EMOTE 49	Δ1612	2		
PQ-EMOTE 49	Δ1612_Δ0425_Δ2278	2		
PQ-EMOTE 49	J1AGA	2		
PQ-EMOTE 56	Sauvage	3	50 pb	Rp8
PQ-EMOTE 56	Δ0425	3		
PQ-EMOTE 56	Δ2278	3		
PQ-EMOTE 58	Sauvage	3	50 pb	Rp8
PQ-EMOTE 58	Δ1612_Δ1330	3		
PQ-EMOTE 58	Δ1612_Δ1330_Δ0425	3		
PQ-EMOTE 62	Sauvage	4	100 pb	Rp10
PQ-EMOTE 62	Δ1612_Δ1330_Δ0425_Δ2278	4		

Table 10 : Librairies PQ-EMOTE triées dans l'ordre de génération.

2.2.1.1 Traitement statistique des données de PQ-EMOTE

2.2.1.1.1 Développement d'une chaîne de traitement de données PQ-EMOTE

L'absence de chaîne de traitement bio-informatique pour analyser les données issues du protocole PQ-EMOTE, nous a conduit, dans un premier temps, à concevoir une telle chaîne qui permet de traiter les données, du pré-traitement des fichiers FASTQ jusqu'aux analyses différentielles de proportion des formes 5'P pour chaque extrémité 5' d'ARN entre les deux souches testées.

Les différentes étapes de ce pipeline sont représentées sur la Figure 103, et nous allons décrire chacune d'entre elles en mettant l'accent sur les méthodes développées pour gérer les problèmes rencontrés

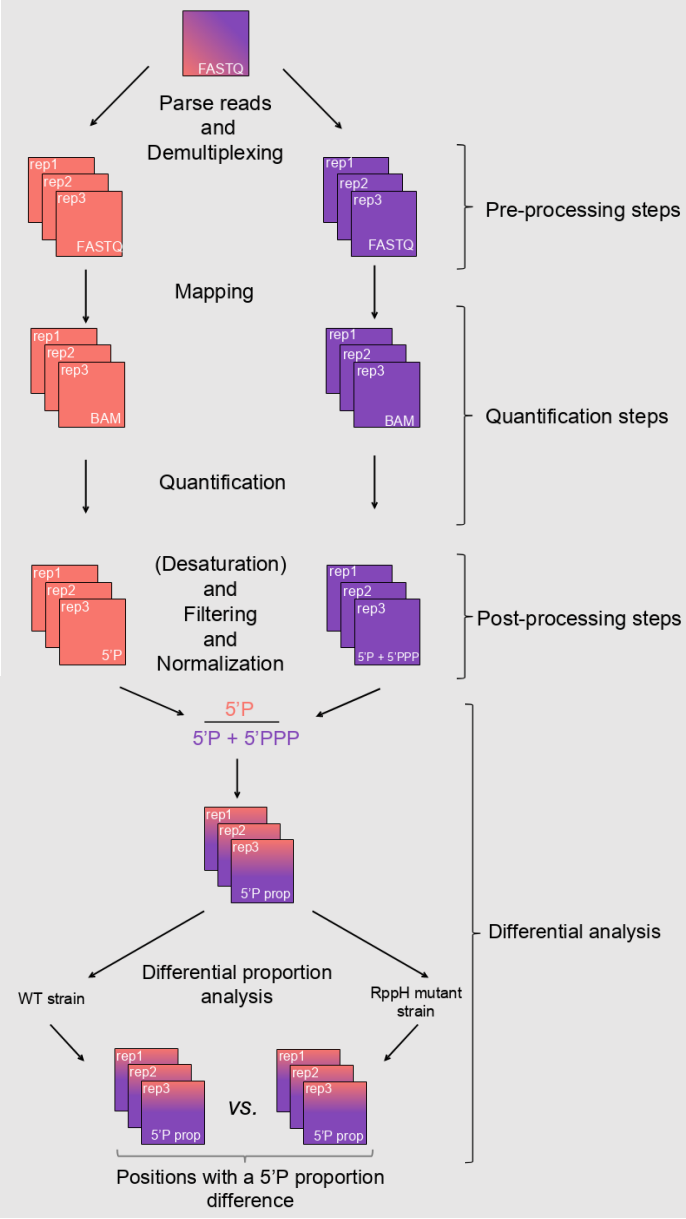


Figure 103 : Chaîne de traitement bio-informatique mise en place pour l'analyse de données PQ-EMOTE.

Les étapes sont regroupées en 4 thèmes : l'étape de pré-traitement, les étapes de quantification qui inclue l'alignement des lectures et la quantification, les étapes de post-traitement dont la désaturation, le filtrage et la normalisation et enfin l'étape d'analyse différentielle de proportion. Les carrés rouges indiquent les échantillons issus de pool^{mono} et les carrés violets ceux issus du pool^{total}.

2.2.1.1.2 Étape de Pré-traitement des données brutes

Les librairies PQ-EMOTE générées, correspondent à des fichiers FASTQ multiplexés (un fichier pour plusieurs conditions expérimentales, souches et réplicats par librairie) dont les lectures contiennent, en plus de la séquence de l'extrémité 5' des ARN, un nucléotide aléatoire pour la fixation de la ligase, un code-barre pour identifier la condition expérimentale et l'oligonucléotide RpX formé. L'oligonucléotide RpX est constitué des éléments suivants : une séquence de reconnaissance pour l'étape de PCR, une séquence UMI (Unique Molecular Identifier) pour distinguer les duplicats PCR de plusieurs événements de ligation avec des ARN ayant la même séquence, et une séquence de contrôle pour vérifier la validité des lectures et identifier le premier nucléotide de l'extrémité 5' du transcrit (Fig. 104).

Les oligonucléotides Rp8 et Rp10 se distinguent uniquement par la longueur de leur UMI. Rp8 possède un UMI composé de 7 nucléotides, tandis que Rp10 en possède un de 17 nucléotides. Cette adaptation vise à améliorer la précision de la quantification des transcrits alignés. Comme abordé précédemment et détaillé dans une section ultérieure, l'utilisation d'UMI courts peut induire des biais dans la quantification du fait d'un phénomène de saturation, c'est-à-dire que le nombre d'UMI de séquences différentes étant trop faible, la même séquence UMI est liée plusieurs fois à la même séquence d'ARN, la faisant apparaître comme un duplicat de PCR. Par ailleurs, l'augmentation de la longueur de l'UMI implique la nécessité de séquencer davantage de paires de bases (passant de 50 à 100), afin de préserver un nombre suffisant de nucléotides correspondant à l'extrémité 5' des ARN pour être en mesure de les aligner de manière unique sur le génome.

Il est à noter que les séquences des oligonucléotides Rp8 et Rp10 ne contiennent pas de U/T car l'étape de PCR utilise une amorce aléatoire incluant un dinucléotide AA à l'extrémité 3' et pouvant donc cibler les dinucléotides TT. Il est donc important que cette amorce ne puisse pas se fixer sur l'oligonucléotide, car cela conduirait uniquement à l'amplification de celui-ci, mais aussi à l'amplification d'oligonucléotides non liés qui sont présents en excès.

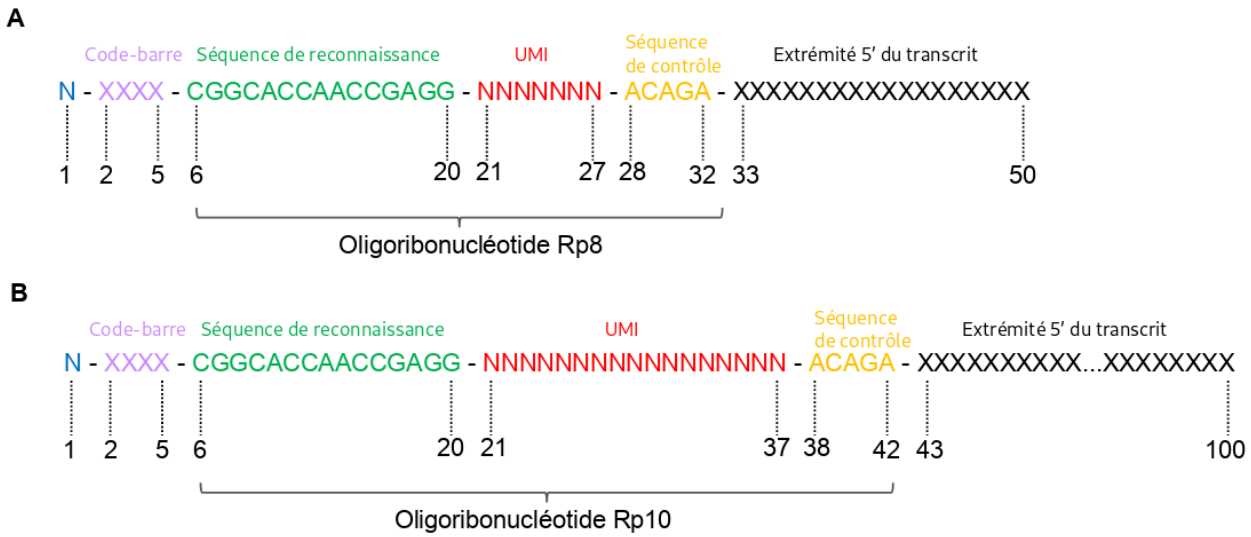


***Figure 104 :* Structures des lectures PQ-EMOTE dans les librairies.**

*(**A**) 49, 56 et 58, et (**B**) 62. Le premier nucléotide est un nucléotide aléatoire permettant l'hybridation de la ligase. Les 4 nucléotides suivants correspondent au code-barre qui permet d'identifier la condition expérimentale de la lecture. Ensuite, vient la partie correspondant à l'oligonucléotide de synthèse composé d'une séquence de reconnaissance permettant à la reverse transcriptase d'initier la transcription uniquement des ARN liés cet oligonucléotide. Une séquence aléatoire servant d'UMI est présente à la suite, ainsi qu'une séquence de contrôle permettant de vérifier la validité des lectures et de déterminer l'extrémité 5' des transcrits. A la fin des lectures se trouve la partie correspondant à la séquence de l'extrémité 5' du transcrit.*

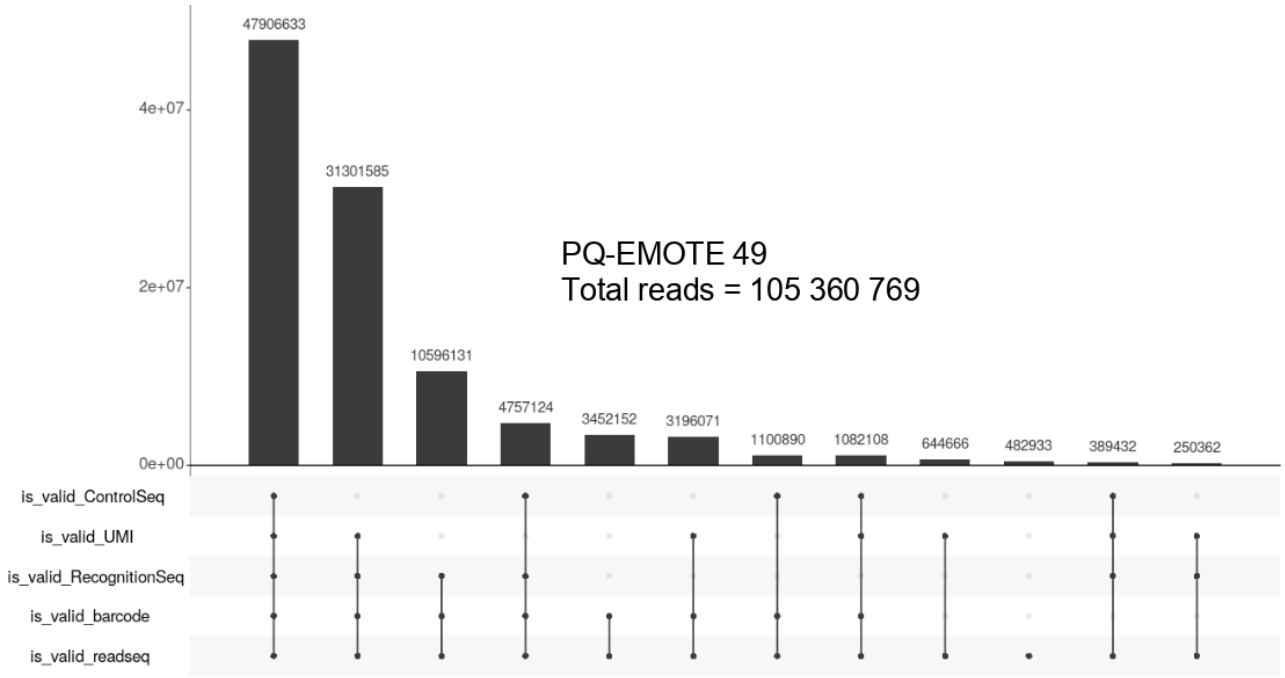
Cette première étape va donc consister à évaluer la qualité des données brutes issue du séquençage et vérifier que le protocole expérimental s'est bien déroulé. Nous vérifions : (i) l'absence d'au plus un nucléotide non déterminé (N) dans la séquence correspondant à l'extrémité 5' et (ii) que les éléments attendus, à savoir la séquence code-barre, la séquence de reconnaissance et la séquence de contrôle correspondent bien à celles présentes dans les lectures. Étant donné la longueur de la séquence de reconnaissance, nous autorisons une erreur de séquençage. Par contre, la séquence code-barre permettant de vérifier que la lecture appartient bien au jeu de données étudié doit correspondre strictement à celle connue, car la présence d'une erreur pourrait signifier une mauvaise attribution de la condition expérimentale lors du démultiplexage. Concernant les UMI, nous vérifions l'absence de T dans la séquence puisque ceux-ci sont générés aléatoirement à partir des bases A, C et G uniquement, contrainte liée à l'amplification PCR (*cf.* ci-dessus).

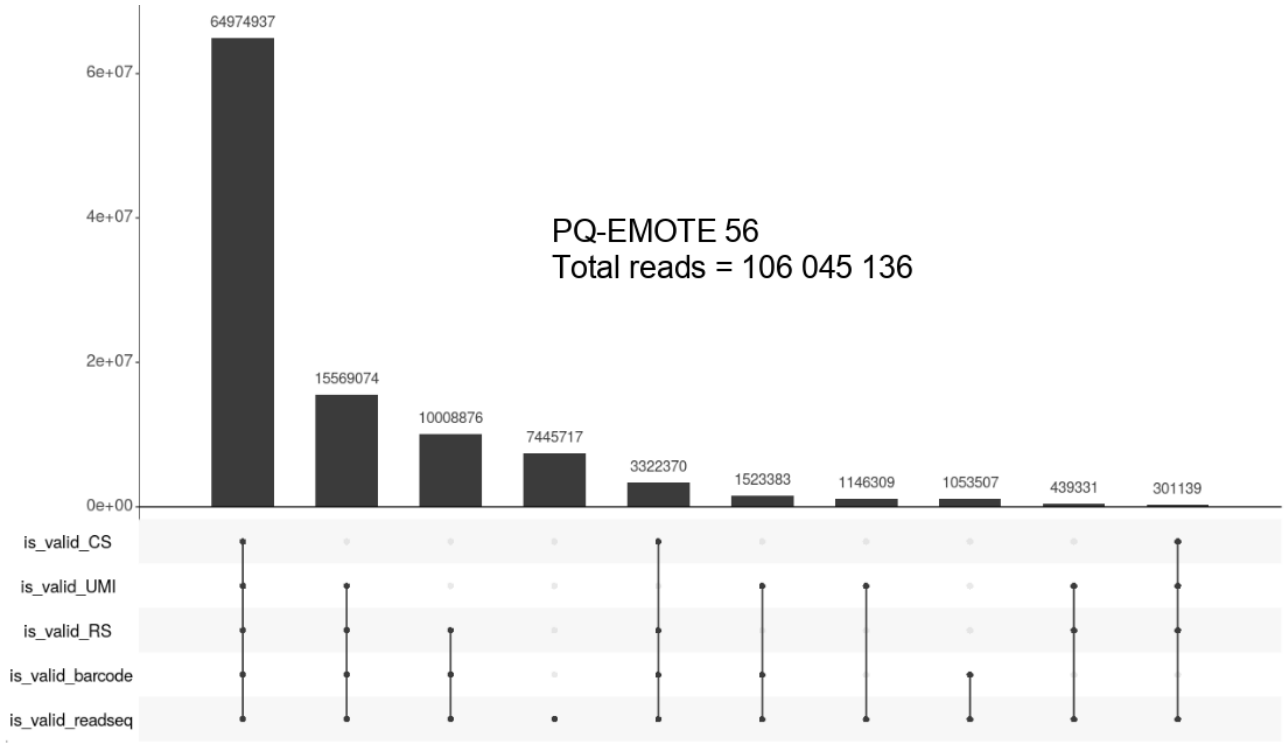
La séquence code-barre est ensuite utilisée pour attribuer les lectures à une condition expérimentale donnée et la partie des lectures correspondant à l'extrémité 5' des ARN est extraite et sauvegardée dans un fichier FASTQ. Ce fichier contient donc uniquement les lectures valides issus d'une même condition expérimentale prêtes à être alignées sur la séquence de référence.

Pour réaliser et automatiser ce travail, nous avons développé un ensemble de fonctions R génériques quelle que soit la constitution de l'oligonucléotide synthétique. Pour cela, l'utilisateur doit remplir la table EMOTE_read_features pour décrire la structure et le contenu des lectures de séquençage brutes (type de motif attendu et emplacement sur les lectures). Une fois les caractéristiques définies, le prétraitement du fichier FASTQ des lectures brutes est réalisé grâce à la sélection et l'extraction des caractéristiques, et la validation des lectures. En plus du ou des fichiers FASTQ contenant uniquement les lectures valides à aligner, un rapport est généré, contenant des statistiques pour chaque échantillon sur le contenu et la qualité du fichier FASTQ de lectures brutes (nombre de lectures correspondant à chaque caractéristique, et une synthèse du nombre et du pourcentage de lectures correspondant à toutes les caractéristiques).

Nous avons appliqué ces fonctions sur les quatre librairies PQ-EMOTE, et les rapports de démultiplexage sont présentés sous forme de Lollipop plot dans les Figures 105. Le pourcentage de lectures contenant tous les éléments attendus valides s'élève à 45 % pour la librairie PQ-EMOTE 49, à 61 % pour la librairie PQ-EMOTE 56, et à 48 % pour la librairie PQ-EMOTE 58. En revanche, la librairie PQ-EMOTE 62 ne possède que 22 % de lectures valides. La différence entre cette dernière librairie et les trois autres, réside dans l'oligonucléotide synthétique utilisé (Rp10 *versus* Rp8) qui contient des UMI plus longs (17 nucléotides versus 7 nucléotides pour les trois autres). Au final, les librairies construites avec l'oligonucléotide Rp8 possèdent toutes plus de 47 millions de lectures utilisables, alors que seulement 14 millions de lectures valides sont identifiées dans la librairie PQ-EMOTE 62 générée. Cette différence pourrait s'expliquer par la longueur plus grande des lectures issues de PQ-EMOTE 62 (UMI plus longues) qui sont donc plus susceptibles d'avoir subi au moins une erreur de séquençage.

Dans tous les cas, il semblerait que la plus grosse source d'invalidité des lectures est la séquence de contrôle (*control.seq*), peu importe la librairie. Ceci assez surprenant étant donné qu'il s'agit d'une séquence assez courte. Pour la librairie PQ-EMOTE 62, la seconde source d'invalidité est la séquence non conforme de l'UMI qui, bien que plus longue, ne doit pas contenir d'erreur (présence de T ou de N).


47906633
31301585
10596131
4757124
3452152
3196071
1100890
1082108
644666
482933
389432
250362
4e+07
2e+07
0e+00
PQ-EMOTE 49
Total reads = 105 360 769
is_valid_ControlSeq
is_valid_UMI
is_valid_RecognitionSeq
is_valid_barcode
is_valid_readseq


64974937
15569074
10008876
7445717
3322370
1523383
1146309
1053507
439331
301139
6e+07
4e+07
2e+07
0e+00
PQ-EMOTE 56
Total reads = 106 045 136
is_valid_CS
is_valid_UMI
is_valid_RS
is_valid_barcode
is_valid_readseq

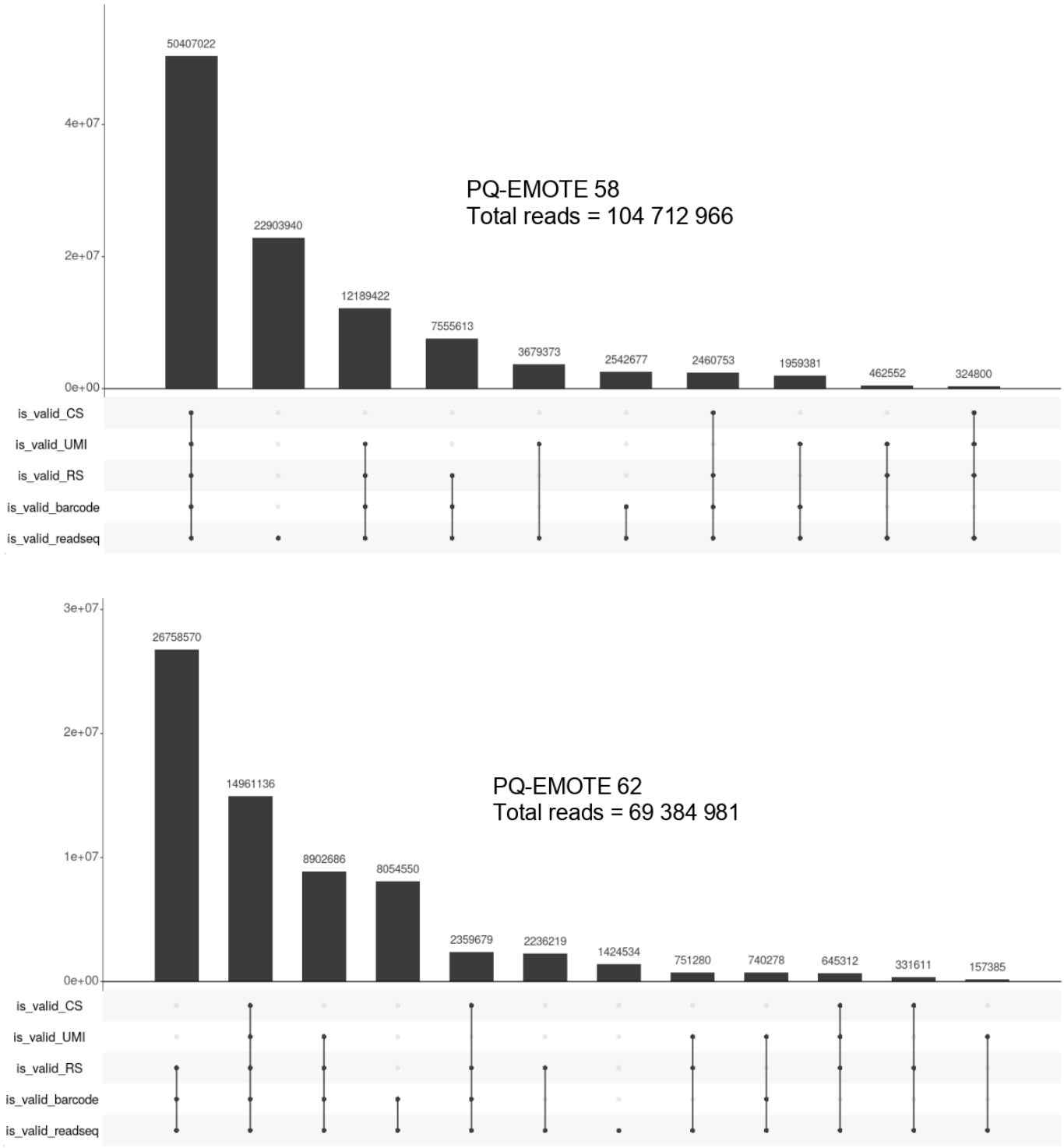


Figure 105 : ***Rapports de démultiplexage sous forme de Lollipop plot des librairies.***

*(**A**) PQ-EMOTE 49, (**B**) PQ-EMOTE 56, (**C**) PQ-EMOTE 58 et (**D**) PQ-EMOTE 62. Ces rapports représentent le nombre de lectures contenant les différentes associations d'éléments valides.*

2.2.1.1.3 Alignement des lectures et quantification des extrémités 5’

Les lectures sont ensuite alignées avec le programme Bowtie sur le génome de référence de la souche *S. aureus* SA564. Pour cette étude, aucun mésappariement n’est autorisé. Pour les librairies PQ-EMOTE 49, PQ-EMOTE 56 et PQ-EMOTE 58, les séquences à aligner font 18 nucléotides de long ce qui est généralement suffisant pour les aligner de manière unique. Dans la librairie PQ-EMOTE 62, les séquences à aligner font 57 nucléotides de long.

Il est tout de même possible que certaines lectures s’alignent à plusieurs endroits si elles correspondent à des régions génomiques répétés comme les ARNr ou les ARNt par exemple. Nous aurons besoin des lectures correspondant aux ARNr dans une étape de normalisation qui sera décrite dans une section ultérieure. Nous avons donc paramétré Bowtie pour conserver les lectures multi-alignées. Pour celles-ci, nous ne gardons que la position du premier alignement.

Concernant l’alignement des lectures, toutes les librairies présentent un pourcentage de lectures alignées supérieur à 74 % avec en moyenne 6,5 millions de lectures alignées (Fig. 106).

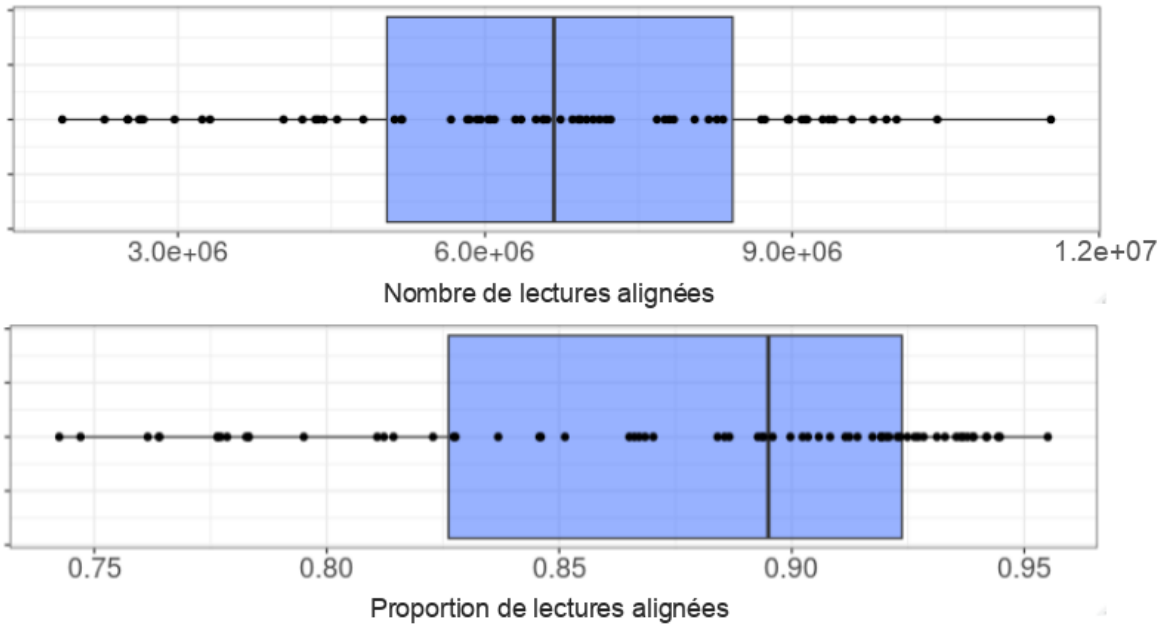


Figure 106 : Distribution du nombre et de la proportion de lectures alignées.

Pour finir, étant donné que nous nous intéressons aux états de phosphorylation des extrémités 5' des ARN, les fichiers BAM des alignements sont traités pour ne tenir compte que de la 1ère position alignée de chaque lecture. Nous avons donc développé une fonction R qui renvoie une table de comptage (Table 11) à partir d'un fichier BAM. Cette table contient en colonnes : l'identifiant de la séquence de référence (*rname*), la position (position) et le brin (*strand*) du 1er nucléotide aligné, le code-barre (*sample_id*) qui a servi à multiplexer, le nombre de lectures alignées (reads) et le nombre d'UMI unique alignés (count).

rname	position	strand	sample_id	reads	count
CP010890.1	312	+	GGTA	23	15
CP010890.1	313	+	AAGT	50	18
CP010890.1	313	+	ACAC	16	13
CP010890.1	313	+	TACA	50	26
CP010890.1	313	+	ATTG	66	32
CP010890.1	313	+	CAAG	478	155
CP010890.1	313	+	GGTA	873	199
CP010890.1	313	+	TCGG	354	144
CP010890.1	313	+	TGTT	64	26
CP010890.1	2131	+	GGTA	21	19
CP010890.1	3601	-	GGTA	19	16

Table 11 : Table de comptage. *Premières lignes de la table de comptage de la librairie EMOTE49.*

A noter que dans la librairie PQ-EMOTE 62, selon les échantillons, entre 66 et 78 % des lectures s'alignent sur l'ARNr 5S. Ceci entraîne une forte réduction de la profondeur de séquençage des autres positions du génome. Ce résultat s'explique par le fait que, dans cette librairie, le seuil de purification sur gel pour la taille des fragments avant l'étape de séquençage est plus bas que dans les autres librairies, ce qui a empêché l'élimination efficace d'une grande partie de cet ARN 5S qui fait 115 nucléotides de long.

Cette table de comptage nous a permis de vérifier si la délétion des gènes NUDIX a été effective. Pour cela, nous avons étudié si des lectures étaient alignées à des positions correspondant à la région codante de ces gènes (Table 12). Dans l'exemple donné ci-dessous (Fig. 107) du quadruple mutant Δ1612-Δ1330-Δ0425-Δ2778, nous n'observons aucune lecture alignée avec une grande partie centrale de la région codante pour les délétions Δ0425, Δ2778 et Δ1330. Seules quelques lectures s'alignent aux extrémités en amont et en aval de la région (en moyenne 7 lectures dans les échantillons $pool^{total}$ du quadruple mutant sur la région codante du gène SA0425, 2 sur celle codant pour SA2278, 4 sur celle codant pour SA1330 et 8 sur celle codant pour SA1612).

Gène	Position de début	Position de fin	Brin
SA0425	433 661	434 102	+
SA2278	2 531 844	2 531 422	-
SA1330	1 544 600	1 544 006	-
SA1612	1 846 773	1 845 687	-
SA1501	1 710 276	1 709 668	-

Table 12 : Coordonnées génomiques et brins des régions codantes pour les 5 candidats NUDIX identifiés.

Comme nous pouvons l'observer sur les données issues de la librairie PQ-EMOTE 62 en Figure 107, mais qui sont également observables pour l'ensemble des 4 librairies PQ-EMOTE, seule une partie des gènes codants pour les candidats RppH est délétée. Ceci s'explique par le fait que les délétions ont été effectuées par un processus de double recombinaison homologue « *in frame* » pour ne pas trop perturber l'expression des gènes en aval et en amont. Cependant ces délétions partielles suffisent à supprimer l'activité de la protéine produite par l'ARN transcrit puisque près de l'entièreté de la séquence codante n'est plus présente.

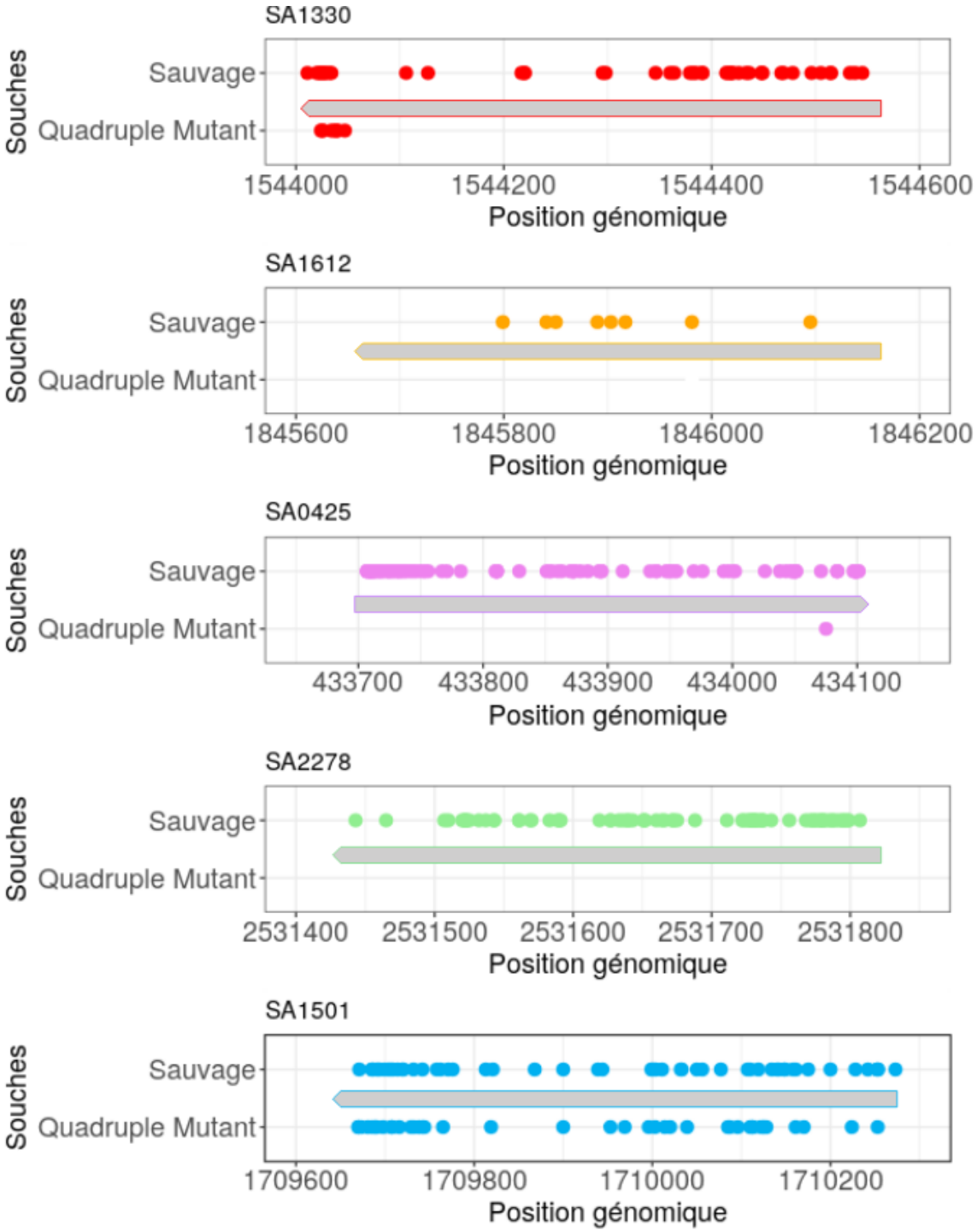


Figure 107 : Vérification de la délétion des NUDIX dans la souche quadruple mutant.

Présence des comptages sur les intervalles génomiques correspondant aux gènes des 5 candidats RppH parmi les 4 délétés dans la souche quadruple mutante. Pour chaque intervalle, les distributions sont regroupées par souches. Les flèches représentent les début, fin et brin des gènes.

2.2.1.1.4 Post-traitement des données de la table de comptage

Après l'étape de quantification des données RNA-Seq, nous obtenons une table de comptages qui pourrait être utilisée telle quelle pour certaines analyses. Cependant, comme mentionné en introduction, il est courant dans la plupart des études de RNA-Seq de procéder à plusieurs étapes de post-traitement des comptages. Ces étapes sont essentielles pour améliorer la fiabilité des résultats et incluent, entre autres, le filtrage des gènes peu exprimés dont de faible variation dans le nombre de comptages peut entraîner une grande variation de proportion, et la normalisation des comptages pour permettre des comparaisons entre échantillons. Pour les données de type PQ-EMOTE, ces étapes classiques sont légèrement différentes et nécessitent des méthodes spécifiques, que nous allons maintenant décrire.

2.2.1.1.4.1 Désaturation des UMI

2.2.1.1.4.1.1 Développement d'une méthode de désaturation des UMI

L'étape de PCR n'amplifie pas les transcrits de manière uniforme. Ceci entraîne une amplification différente des transcrits, comme illustré dans la Figure 108. La ligation d'un Identifiant Unique de Molécule (UMI) à l'extrémité 5' de chaque transcrit permet, après amplification, de distinguer les lectures issues de duplicats PCR de celles provenant de copies distinctes d'un même transcrit. Toutefois, lorsque le nombre de séquences différentes d'UMI est insuffisant, une même séquence d'UMI peut se lier à différents transcrits issus d'un même site de départ de transcription (et donc ayant la même séquence). Ceci rend impossible la distinction entre les duplicats PCR et les multiples copies d'un même transcrit après la PCR, comme également illustré dans la Figure 108.

Les données de séquençage des extrémités 5' sont particulièrement sensibles à la saturation des UMI. En effet, dans les librairies PQ-EMOTE 49, 56 et 58, les UMI consistent en des séquences aléatoires de seulement 7 nucléotides, composées uniquement de 3 des 4 bases possibles (A, C et G). Ainsi, il existe au total 2187 séquences d'UMI différentes (3^7). Selon le modèle théorique de (Saunders et al., 2020), illustré dans la Formule 4, une saturation de 5 % est estimée être atteinte lorsque 235 lectures sont alignées à la même position pour ces 2187 séquences d'UMI différentes. 5 % des UMI pourront donc, de façon aléatoire, être liés à plus d'une copie distincte d'un même transcrit.

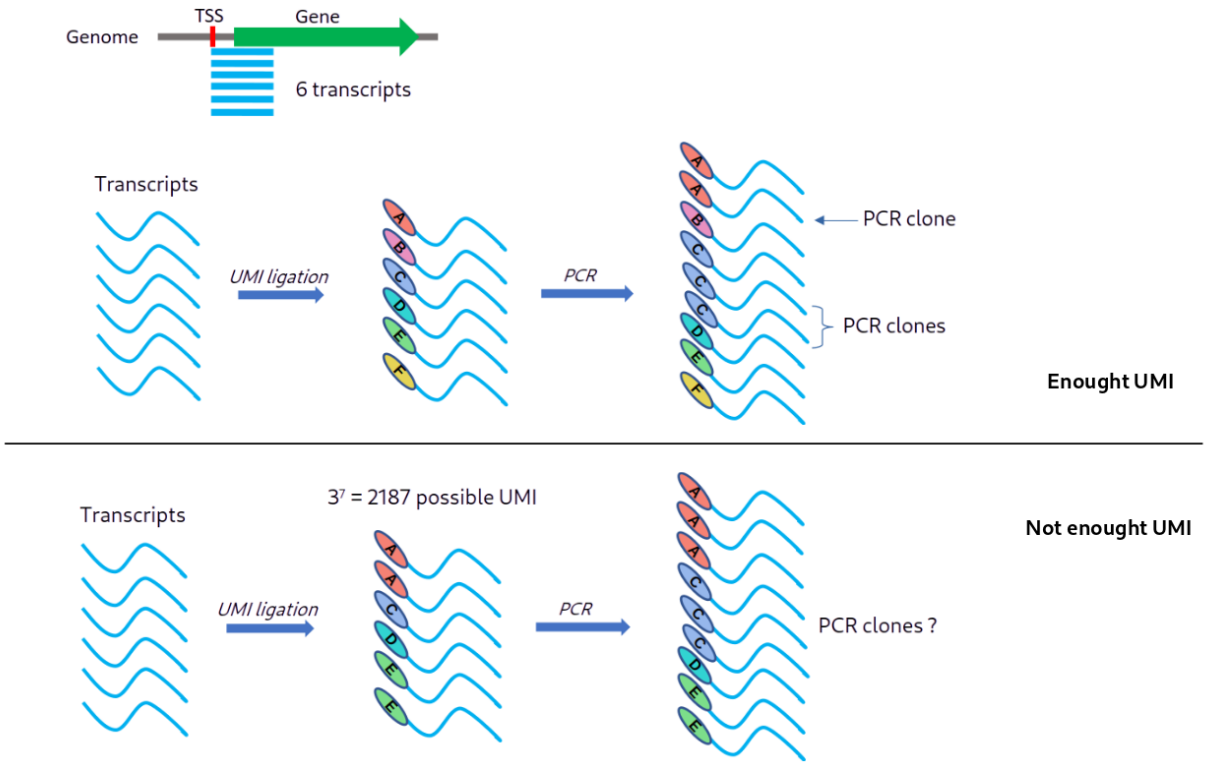


Figure 108 : Amplification différentielle des transcrits et de saturation des UMI.

(**A**) Avec suffisamment d'UMI pour marquer chaque molécule individuellement, (**B**) Avec un nombre insuffisant d'UMI pour marquer chaque molécule individuellement. Les UMI sont représentés par les ovales de couleurs en 5' des ARN (traits bleus).

$$\text{Nombre d'UMI réutilisés} = 1 - \left(\frac{C-1}{C}\right)^{n} - \frac{n}{C} * \left(\frac{C-1}{C}\right)^{n-1} * C$$

$$= 1 - \left(\frac{2187-1}{2187}\right)^{235} - \frac{235}{2187} * \left(\frac{2187-1}{2187}\right)^{235-1} * 2187 = 11{,}71$$

$$\text{Pourcentage d'UMI réutilisés} = \frac{11{,}71}{235} \simeq 5\ \%$$

Formule 4 : Calcul du pourcentage de saturation des UMI à l'aide du modèle théorique de Saunders et al. (2020). *C correspond au nombre total de séquences d'UMI et n au nombre de lectures alignées à une même position. Ici C = 2187 et n = 235 lectures alignées à la même position.*

Ce constat souligne que le comptage basé sur le nombre d'UMI uniques est biaisé et ne reflète pas fidèlement le nombre réel de transcrits, d'où la nécessité de développer une approche plus précise et fiable pour cette quantification.

Par ailleurs, le modèle théorique de Saunders et al. (2020) repose sur l'hypothèse d'une distribution uniforme des séquences d'UMI. Pour tester l'adéquation de ce modèle aux données PQ-EMOTE, nous avons d'abord étudié la distribution des séquences d'UMI présentes dans l'oligonucléotide Rp8 dans la librairie PQ-EMOTE 58. Pour calculer cette distribution, nous avons utilisé uniquement les UMI présents dans les lectures alignées à des positions où aucune duplication ni saturation n'a été observée. Pour cela, nous avons filtré la table de comptage pour ne garder que les positons où le nombre d'UMI différents est égal au nombre de lectures alignées. Nous avons ensuite extrait les séquences d'UMI correspondant à ces positions et calculé leur fréquence. Au total, nous avons pu calculer la fréquence des UMI à partir de 3 923 374 positions (6 020 551 lectures). Une même position mais pour des conditions expérimentales différentes dans la librairie PQ-EMOTE 58 peut contribuer plusieurs fois au calcul de cette distribution. La fréquence de chaque UMI est calculée en faisant le rapport du nombre de fois que cet UMI est observé divisé par le nombre total de lectures (6 020 551)

La Figure 109, représentant les fréquences d'utilisation des séquences d'UMI calculées sur la librairie PQ-EMOTE 58. Cette distribution n'est pas uniforme. Contrairement à une distribution uniforme où chaque UMI aurait une fréquence de 1/2187 (environ 0,000457), on observe de grandes variations. La séquence UMI la plus fréquente a une fréquence d'environ 0.004, tandis que la moins fréquente est près de 1000 fois moins fréquente ($4{,}4.10^{-6}$).

Cette distribution est également observée dans les librairies PQ-EMOTE 49 et 56 qui utilise également l'oligonucléotide Rp8. En effet, les tests de corrélation de Spearman effectués sur les fréquences des UMI entre les librairies PQ-EMOTE 49 et 56, entre PQ-EMOTE 49 et 58, et entre les librairies PQ-EMOTE 56 et 58 révèlent des coefficients de corrélation de 0,93, 0,99 et 0,94 respectivement.

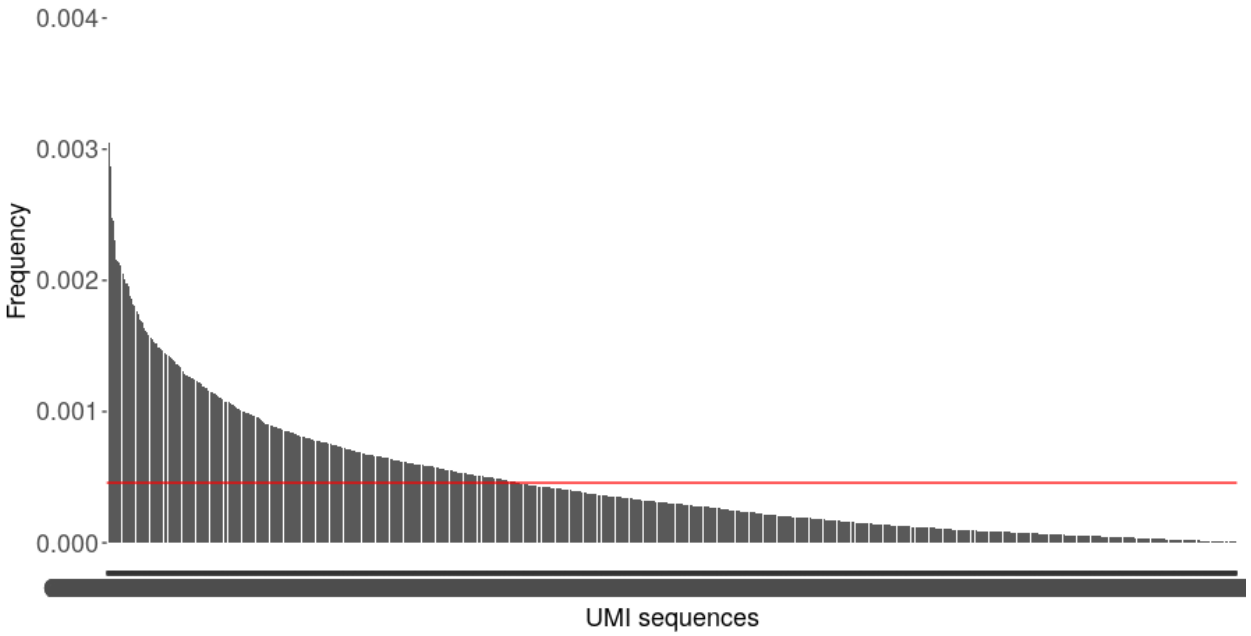


Figure 109 : Distribution des fréquences des séquences d'UMI. *Ce diagramme en bâton illustre la fréquence (axe y) de chacune des 2187 séquences UMI possibles (axe x), classées par fréquence décroissantes. La ligne rouge représente la fréquence pour une distribution uniforme attendue (1/2187).*

Afin de déterminer s'il existe un lien entre la fréquence des UMI et leur séquence en ribonucléotides, nous avons réalisé deux weblogos, un basé sur les 20 séquences d'UMI les plus fréquentes et un basé sur les 20 séquences les moins fréquentes. Les résultats présentés dans la Figure 110 révèlent que les UMI les plus fréquents présentent principalement deux guanines en position 6 et 7 de leur séquence. En revanche, les UMI les moins fréquents montrent une prédominance de cytosines sur toutes les positions de la séquence, notamment en position 1 à 5. De plus, ces 20 séquences ne contiennent aucune adénosine aux positions 2 à 5.

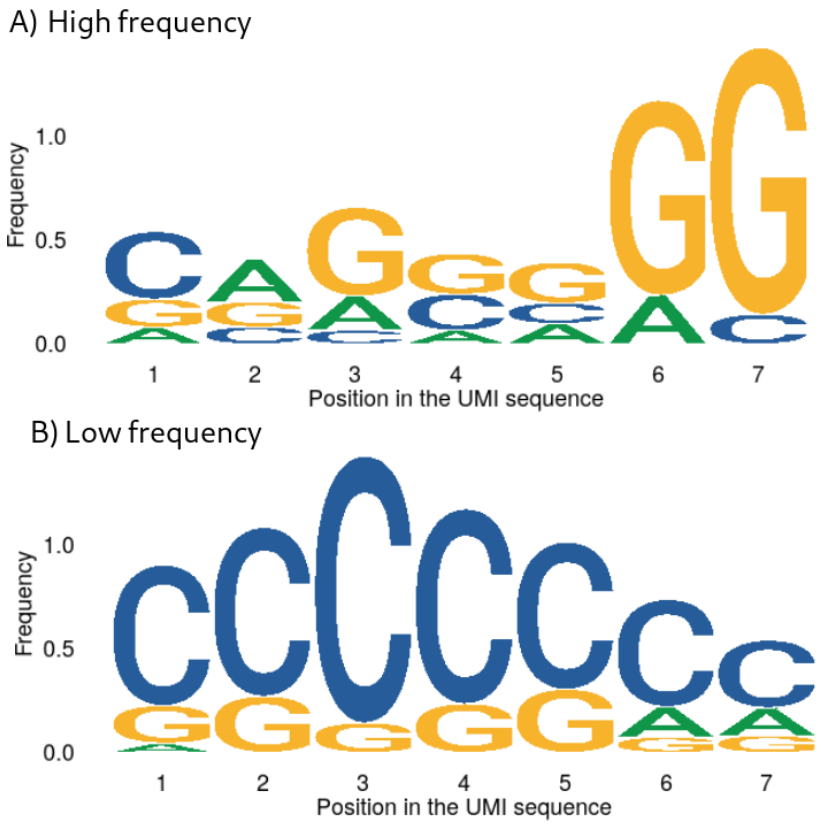
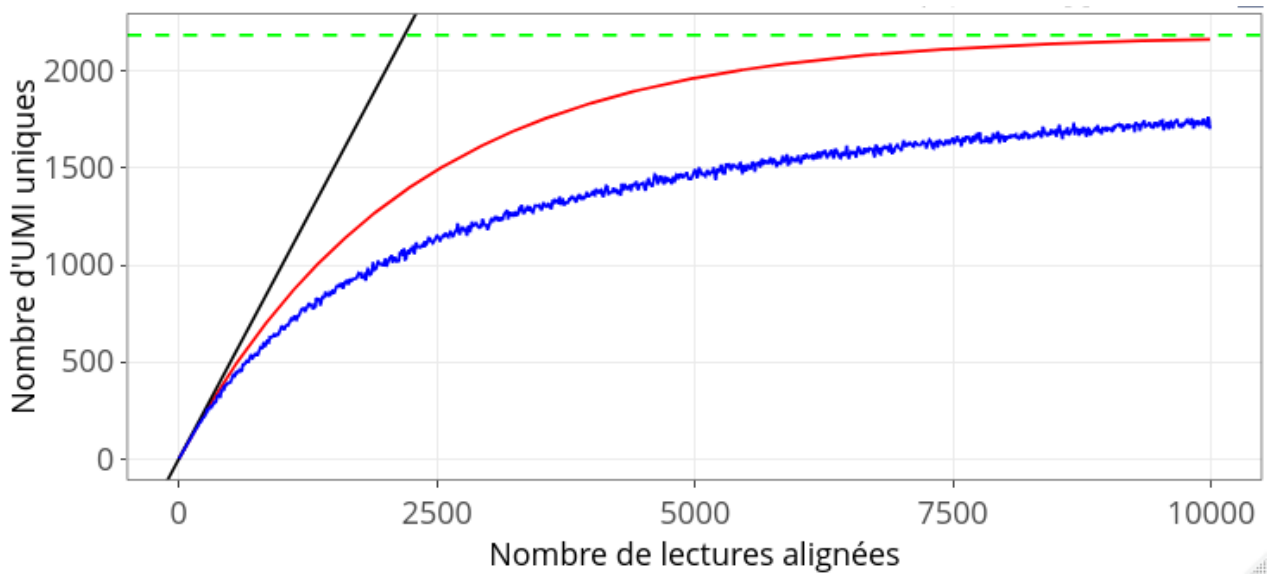


Figure 110 : Weblogos des 20 séquences d'UMI. *(**A**) les plus fréquentes et (**B**) les moins fréquentes*

Les hypothèses pouvant expliquer la distribution non uniforme des séquences d'UMI sont, d'une part, la présence possible d'un biais de ligation conduisant certaines séquences d'UMI à se lier plus efficacement, et d'autre part, la possibilité qu'il y ait un biais de fabrication des UMI, conduisant à des concentrations différentes des UMI dans les produits commandés. La fréquence des UMI n'étant pas uniforme, nous avons procédé à des simulations pour estimer la saturation des UMI à partir de leur fréquence estimée. Pour un nombre donné de lectures alignées *n* à une même position, nous effectuons un tirage aléatoire avec remise de *n* UMI en tenant compte de leur probabilité de tirage, c'est-à-dire de leur fréquence estimée précédemment. Ensuite, nous comptabilisons le nombre de séquences UMI distinctes obtenues. Pour chaque nombre de lectures alignées, nous répétons ce processus 1 000 fois, et rapportons la moyenne des nombres d'UMI différents. La simulation réalisée à partir des fréquences calculées pour des nombres de lectures alignées variant de 1 à 10 000 est présentée dans la Figure 111 (la Figure 111B étant un agrandissement de la Figure 111A).

(A)



(B)

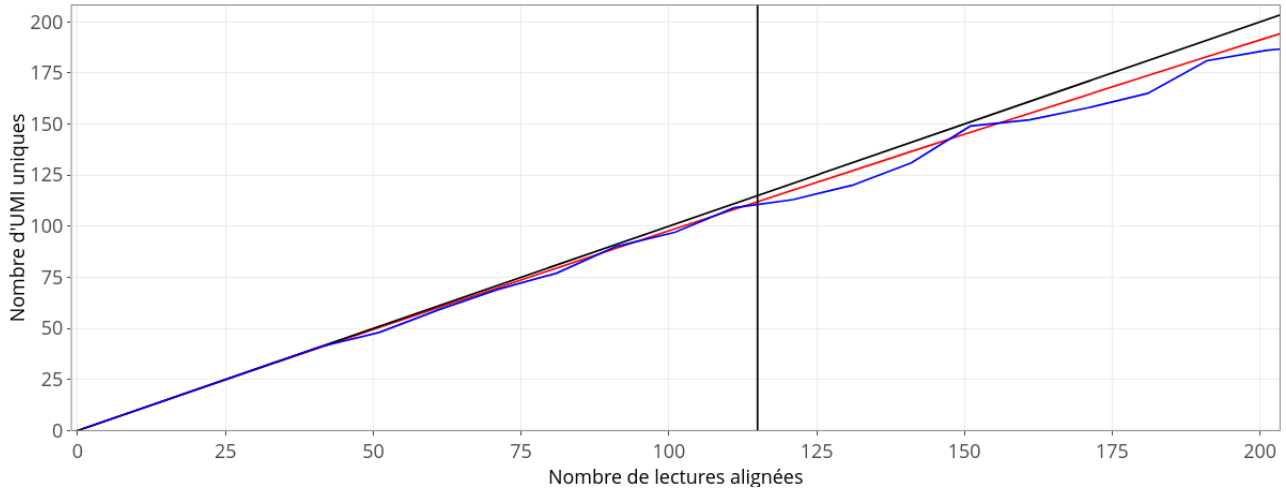


***Figure 111 :* Saturation des UMI en tenant compte de leur fréquence d'utilisation.**

La courbe bleue représente le nombre d'UMI uniques tirés aléatoirement (axe y) en fonction du nombre de lectures alignées à une position (axe x). La courbe rouge représente le modèle théorique pour une distribution uniforme des fréquences d'UMI. La droite verte en pointillé représente le nombre de séquences d'UMI possibles (2187) et la courbe noire représente le nombre unique estimé lorsqu'il est égal au nombre d'UMI unique alignés à une position la diagonale.

A partir de ces résultats, nous constatons une amplification du phénomène de saturation par rapport à un modèle uniforme. Cela se traduit par un nombre de lectures plus faible (~115) à partir duquel 5 % des séquences d'UMI sont réutilisées. Ces résultats soulignent la nécessité de désaturer les UMI dans les librairies 49, 56 et 58.

Nous avons donc développé une méthode pour identifier les duplicats PCR tout en tenant compte du phénomène de saturation et de la fréquence d'utilisation des UMI.

Cette méthode de désaturation utilise un modèle binomial. De la même manière que l'on peut calculer la probabilité d'obtenir X fois un 6 lors de Y lancers d'un dé à six faces, sachant que la probabilité de tomber sur un 6 est de 1/6, on peut déterminer la probabilité d'observer X occurrences d'un UMI à une position donnée sur Y lectures alignées, en sachant que la probabilité d'utilisation de X, p(X) correspond à sa fréquence calculée précédemment. Pour chaque UMI à chaque position, nous utilisons ce modèle comme illustré dans l'exemple de la Figure 112.

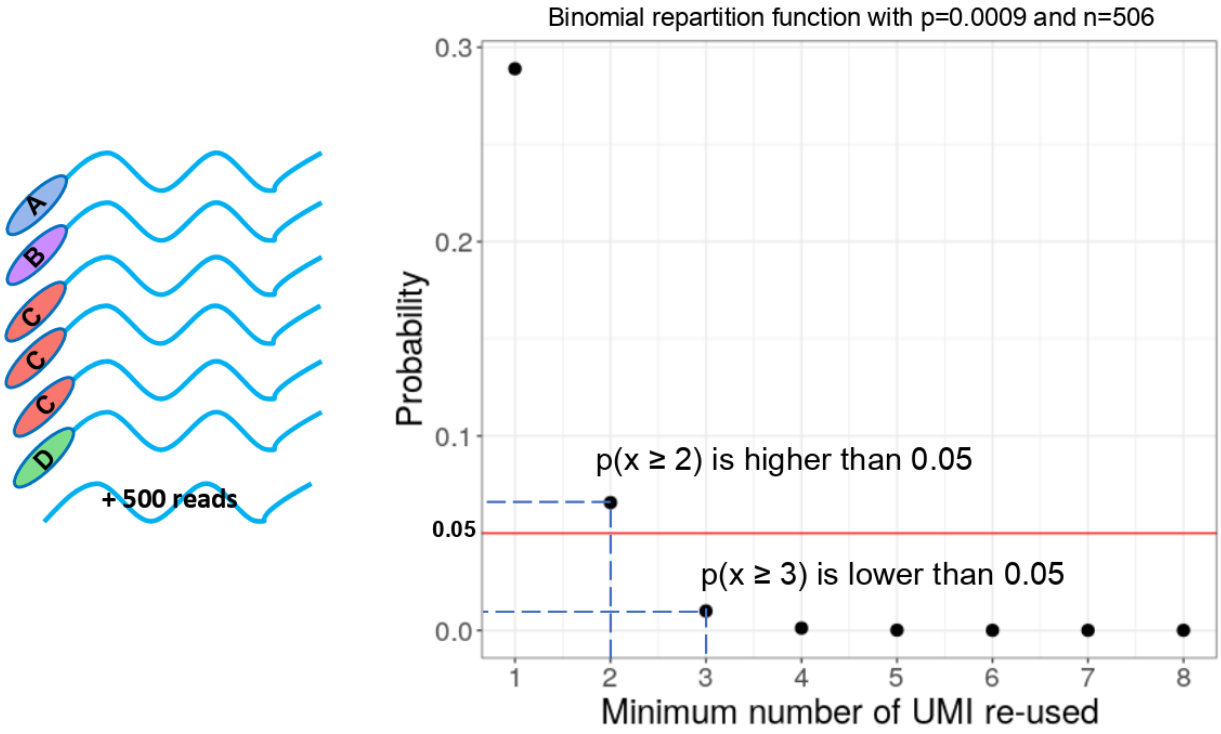


Figure 112 : Exemple de l'application d'un modèle binomiale utilisé la méthode de désaturation.

Dans cet exemple, nous avons 506 lectures alignées à une même positions, et seul 6 d'entre eux ont été représentés, les ovales de couleurs représentent les séquences d'UMI et nous pouvons voir qu'il y a 3 lectures liées avec l'UMI C rouge. Sachant que la probabilité de tomber sur l'UMI rouge est de 0,0009. Nous avons représenté la courbe de répartition avec les paramètres p = 0,0009 et n = 506. Nous observons qu'il est très improbable ($p(UMI_C \geq 3)$ est inférieure à 0,05) de réutiliser 3 fois l'UMI C rouge parmi les 506 lectures. En revanche la probabilité d'en obtenir 2 ou moins est supérieure à 0,05. Nous supprimons donc une lecture contenant un UMI C rouge. En d'autres termes, même avec le biais introduit pour la ligation (fréquences observées des UMI), il est très peu probable que cette lecture soit due à la saturation, et donc bien plus probable qu'elle soit due à un duplicat de PCR. Cette suppression diminue le total de lectures alignées à cette position (506 -> 505). Il faut donc réitérer les calculs avec les nouveaux paramètres (p = 0,0009 et n = 505), et déterminer s'il est probable d'obtenir 2 UMI C rouges sur les 505 lectures alignées. Cette suppression des lectures attribuées aux duplicats de PCR est répétée jusqu'à ce que le nombre de lectures alignées devienne stable, c'est-à-dire que plus aucune lecture n'est identifiée comme un duplicat de PCR.

Cette méthode permet donc de ne supprimer, à chaque itération, les UMI dont il est très peu probable que ce ne soit pas des duplicats PCR.

2.2.1.1.4.1.2 Performance de la méthode de désaturation

Pour évaluer les performances de la méthode de désaturation proposée plus haut, nous avons utilisé les données de la librairie PQ-EMOTE 62, caractérisée par l'utilisation de l'oligoribonucléotide Rp10 qui contient des UMI de 17 nucléotides de long, soit 10 nucléotides de plus que dans les autres librairies. Le nombre total de séquences d'UMI différentes possibles est donc de 129 140 163 (3^{17}). Avec plus de 129 millions de séquences d'UMI possibles, un nombre considérable de lectures alignées à une même position serait nécessaire pour observer un phénomène de saturation. Selon le modèle de Saunders et al. (2020) une saturation de 5% serait observée lorsqu'environ 13 millions lectures sont alignées à une même position, ce qui n'a jamais été le cas dans les données utilisées ici.

Les performances de la méthode de désaturation ont été évaluées en raccourcissant la taille des UMI de 17 nucléotides à 10, 7 et 5 nucléotides respectivement afin de simuler artificiellement la saturation, comme illustré dans la Figure 113.

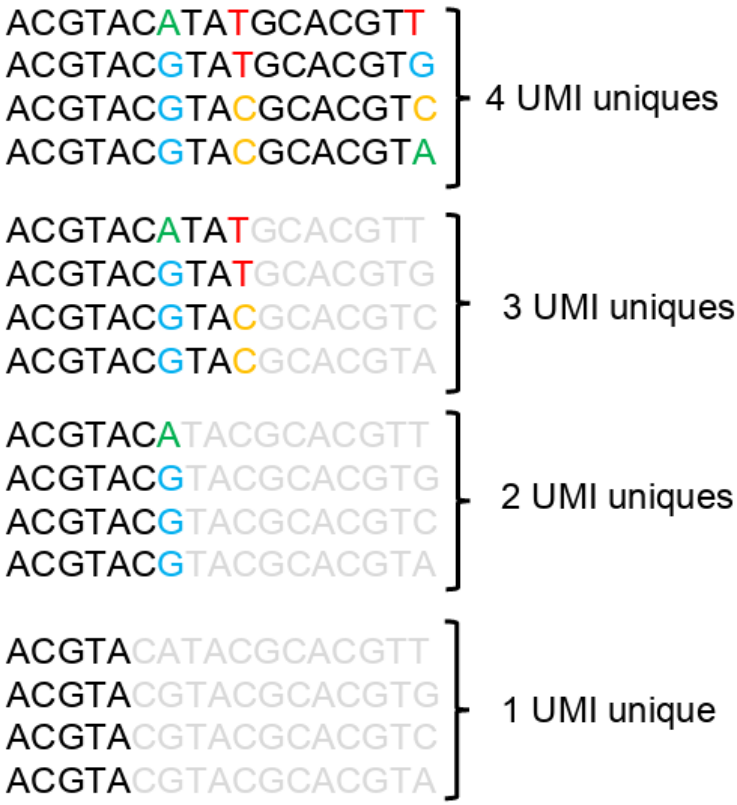


Figure 113 : Exemple d'ajout artificiel de saturation à partir d'UMI de 17 nucléotides de long*. Les nucléotides colorés représentent les différences entre les séquences d'UMI (une couleur par base). Bien que dans le protocole PQ-EMOTE, aucun T ne devrait être présent dans les UMI, nous en avons quand même représenté pour faciliter la compréhension de l'exemple.*

Les lectures contenant ces UMI raccourcis ont ensuite été désaturées à l'aide de notre méthode, et les résultats ont été comparés avec ceux obtenus pour les UMI de taille 17.

La Figure 114A présente les résultats obtenus en désaturant les UMI de taille 7 pour une position où le nombre de lectures alignées est de 947 (point noir). Le nombre d'UMI uniques de taille 7 est de 503 (point bleu foncé), tandis que celui des UMI uniques de taille 17 est de 707 (point violet). Nous avons également illustré que la différence entre le nombre de lectures et le nombre d'UMI de taille 17 correspond aux lectures dupliquées tandis que l'écart entre le nombre d'UMI de taille 17 et celui des UMI de taille 7 non désaturés représente le nombre d'UMI réutilisés.

Après désaturation, nous constatons que la valeur estimée des UMI de taille 7 désaturés est de 722, ce qui se rapproche davantage du nombre réel de transcrits comparé au nombre de lectures ou aux UMI de taille 7 non désaturés.

La Figure 114B illustre les résultats de la désaturation des UMI de taille 7 pour des positions où le nombre de lectures alignées se situe entre 800 et 1 000. Cette représentation offre une vue d'ensemble sur le comportement de la méthode de désaturation. Nous observons que les valeurs estimées par notre méthode de désaturation (points bleus) sont presque toutes très proches des valeurs réelles (points violets), ce qui suggère de bonnes performances pour cette méthode.

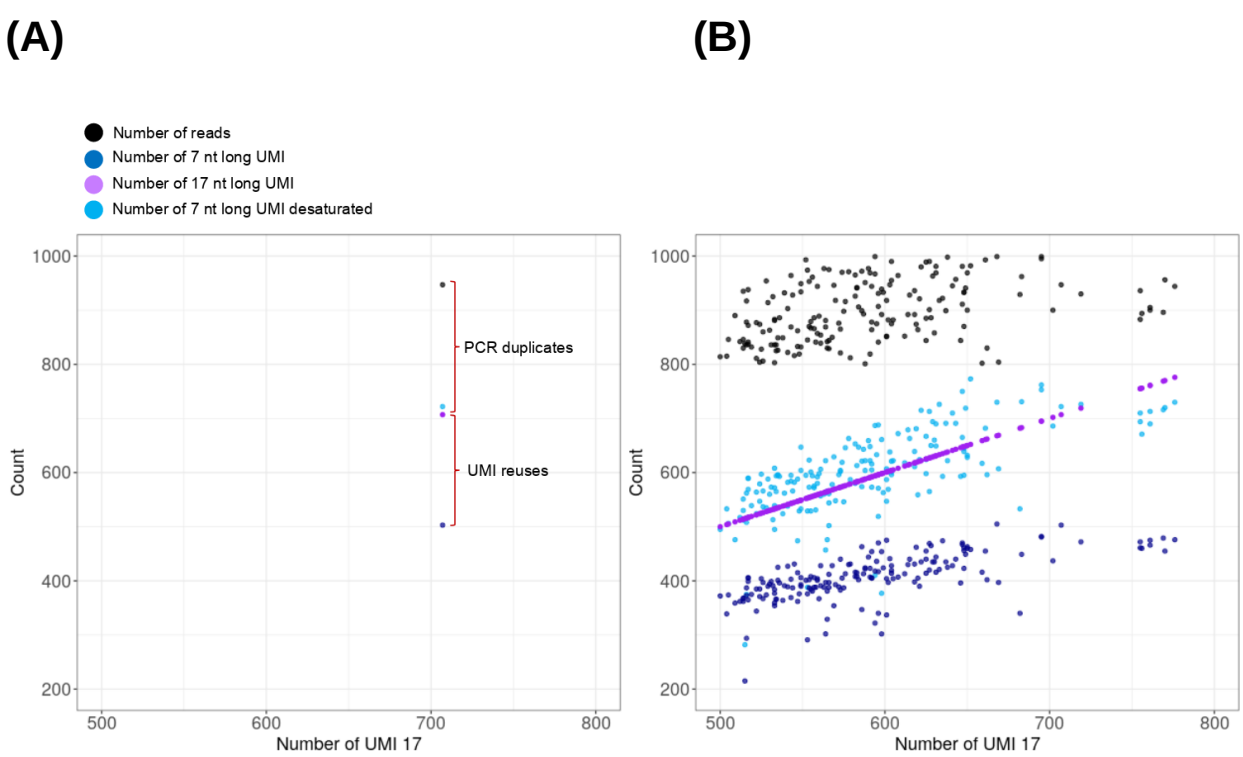


Figure 114 : ***Désaturation des UMI de taille 7.*** *Nombre de lectures alignées (noir), d'UMI unique de taille 7 (bleu foncé), d'UMI de taille 7 désaturé (bleu clair) et nombre d'UMI de taille 17 (violet) pour (**A**) une position et (**B**) pour un ensemble de positions dont le nombre de lectures est compris entre 800 et 1000.*

A partir de ces valeurs, nous avons calculé un pourcentage d'erreur (Formule 5) pour chacune des positions en fonction des comptages utilisés pour l'estimation.

$$\text{Pourcentage d'erreur} = \frac{|\ \textit{Valeur estimée} - \textit{valeur réélle}\ |}{\textit{valeur réelle}} * 100 = \frac{|\ \textit{Nombre d'UMI estimé} - \textit{Nombre d'UMI 17}\ |}{\textit{Nombre d'UMI 17}} * 100$$

Formule 5 : ***Calcul du pourcentage d'erreur.***

La valeur estimée correspond à une valeur de comptage pour une extrémité 5' et la valeur réelle correspond au nombre d'UMI unique de taille 17 pour cette même extrémité.

Ces pourcentages d'erreur sont présentés sous forme de courbes de régression loess par nombre de lectures alignées en Figure 115A, et par valeurs estimées en Figure 115B.

La Figure 115A présente les pourcentages d'erreur en fonction du nombre de lectures alignées à une même position, pour différentes tailles d'UMI ainsi que s'ils ont été désaturés ou non.

- UMI de taille 5 non désaturés (**vert foncé**) : Cette courbe montre une augmentation rapide de l'erreur, atteignant un plateau à un pourcentage d'erreur de 75%, indiquant que la saturation de cette UMI induit beaucoup des erreurs dans l'estimation de la vraie valeur de transcrit alignés. Ce n'est pas surprenant étant donné que le nombre de séquence d'UMI de taille possible est de seulement 3^5 soit 243.
- UMI de taille 7 non désaturés (**bleu foncé**) : On observe une montée progressive de l'erreur, qui dépasse les 50% à partir de 2 200 lectures alignées. Bien que le pourcentage d'erreur croie plus lentement qu'avec les UMI de taille 5, ces résultats soulignent le problème d'utilisation d'UMI trop courts.
- UMI de taille 10 non désaturés (**rouge foncé**) : La courbe rouge foncé démontre que l'erreur augmente, mais reste en dessous de 25, indiquant, comme on pouvait s'y attendre avec ses 59 049 séquences possibles, une meilleure résistance à la saturation comparée aux UMI plus courts.
- UMI de taille 5 désaturés (**vert clair**) : Cette courbe montre une réduction notable du pourcentage d'erreur par rapport aux UMI de taille 5 non désaturés, bien que l'erreur augmente avec le nombre de lectures, elle reste largement inférieure à celle des UMI non désaturés.
- UMI de taille 7 désaturés (**bleu clair**) : La courbe bleu clair indique que la désaturation des UMI de taille 7 réduit significativement le pourcentage d'erreur, qui reste stable et inférieur à celui des UMI non désaturés, montrant ainsi l'efficacité de la méthode de désaturation.
- UMI de taille 10 désaturés (**rouge clair**) : La courbe en rouge clair présente peu de différences par rapport à celle des UMI de taille 10 non désaturés. Cela s'explique par le fait que le nombre maximal de lectures alignées illustré dans la Figure 115A n'atteint pas la saturation de ces UMI.
- Nombre de lectures (**noire**) : La courbe noire montre un pourcentage d'erreur élevé mais plutôt constant autour de 50 %. Ce pourcentage d'erreur est supérieur à l'ensemble des autres courbes jusqu'à 50 lectures et reste au-dessus des UMI de taille 7 non désaturé jusqu'à 2 200 ce qui confirme l'importance de l'utilisation d'UMI pour la quantification des transcrits alignés.

Globalement, les résultats montrent que la désaturation des UMI réduit efficacement les pourcentages d'erreur. Ces observations suggèrent que la désaturation est une méthode performante pour améliorer la précision de la quantification des lectures alignées.

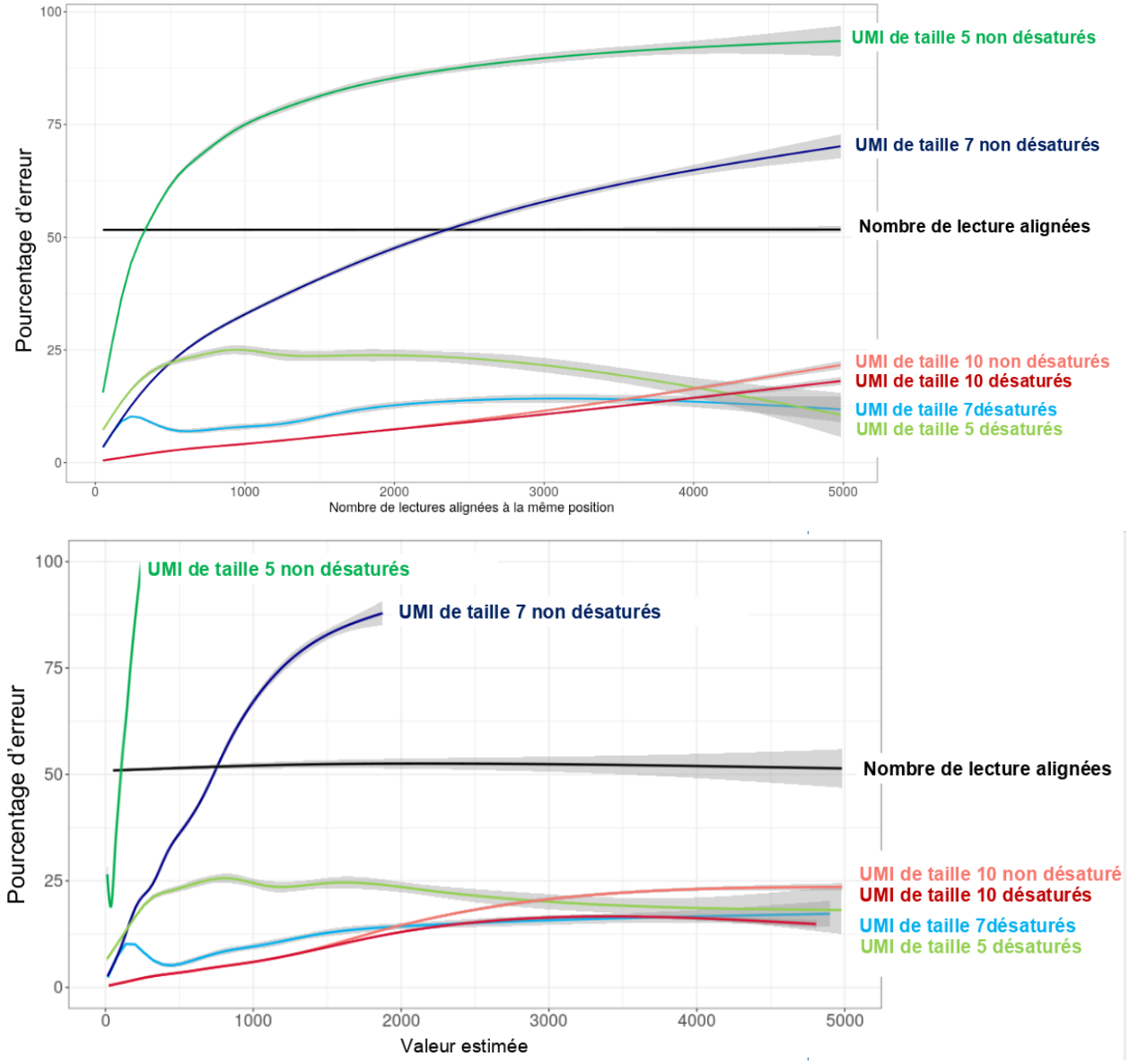


***Figure 115 :* Performance de la méthode de désaturation.**

*Pourcentage d'erreur fonction (**A**) du nombre de lectures alignées à une même position (**B**) de la valeur estimée pour différentes tailles d'UMI. Les courbes illustrent les performances des UMI de taille 5, 7, et 10, à la fois désaturés et non désaturés, par rapport à la valeur de référence des UMI de taille 17. **Vert foncé** : UMI de taille 5 non désaturés. **Bleu foncé** : UMI de taille 7 non désaturés, **Rouge foncé** : UMI de taille 10 non désaturés, **Vert clair** : UMI de taille 5 désaturés, **Bleu clair** : UMI de taille 7 désaturés, **Rouge clair** : UMI de taille 10 désaturés, **Ligne noire horizontale** : Nombre de lectures alignées.*

Ces résultats ont ensuite été synthétisés dans la Figure 116 sous forme de boîtes à moustaches, permettant d'étudier les distributions des taux d'erreur (pourcentage d'erreur / 100) en fonction de valeur utilisée pour estimer le nombre de transcrits.

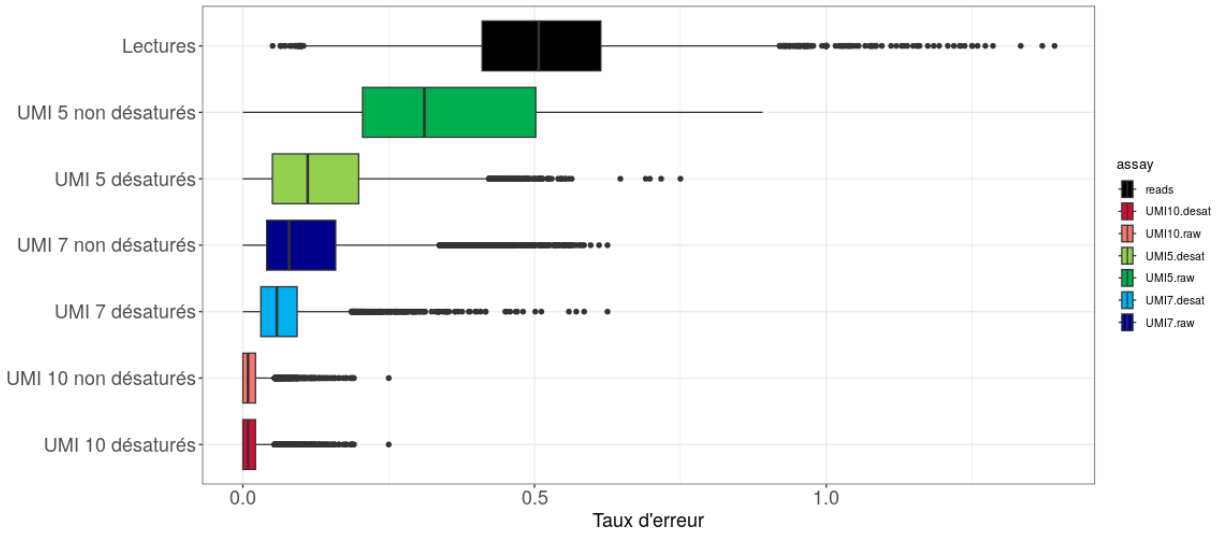


Figure 116 : Distribution des taux d'erreur en fonction des comptages utilisés pour estimer le nombre de transcrit alignés. ***Vert foncé*** *: UMI de taille 5 non désaturés.* ***Bleu foncé*** *: UMI de taille 7 non désaturés,* ***Rouge foncé*** *: UMI de taille 10 non désaturés,* ***Vert clair*** *: UMI de taille 5 désaturés,* ***Bleu clair*** *: UMI de taille 7 désaturés,* ***Rouge clair*** *: UMI de taille 10 désaturés,* ***Noir*** *: Nombre de lectures alignées.*

Étant donné que dans les données, 90 % des positions comptent moins de 972 lectures alignées et 75 % en comptent moins de 348, les courbes de régression présentées ne reflètent pas fidèlement l'impact de la désaturation sur ces données. Par conséquent, nous avons effectué un zoom dans la Figure 117 pour nous concentrer sur la région de 0 à 1000 lectures alignées. De plus, nous n'avons affiché que les courbes correspondant aux UMI de taille 7, car il s'agit du cas rencontré dans les librairies PQ-EMOTE 49, 56 et 58. Afin d'obtenir une vision plus complète du comportement de la méthode pour cette taille d'UMI, nous avons également affiché le pourcentage d'erreur correspondant aux UMI de taille 7 désaturés pour chaque position.

Comme anticipé, nous observons une grande dispersion des pourcentages d'erreur, qui, cependant ne dépassent que rarement les 25 %. L'agrandissement de la figure permet de voir qu'en moyenne que les UMI désaturés ont un pourcentage d'erreur plus faible que les UMI non désaturés à partir de 130 lectures alignées. Bien que le pourcentage d'erreur moyen des UMI désaturés soit relativement stable, il varie tout de même entre 0 et 25 %.

Ces observations confirment les conclusions tirées des figures précédentes : la désaturation des UMI réduit efficacement les pourcentages d'erreur, en particulier pour les positions avec un nombre modéré de lectures alignées. Malgré une dispersion notable, l'utilisation de la désaturation est clairement préférable, renforçant l'importance de cette méthode pour améliorer la précision des analyses de séquençage.

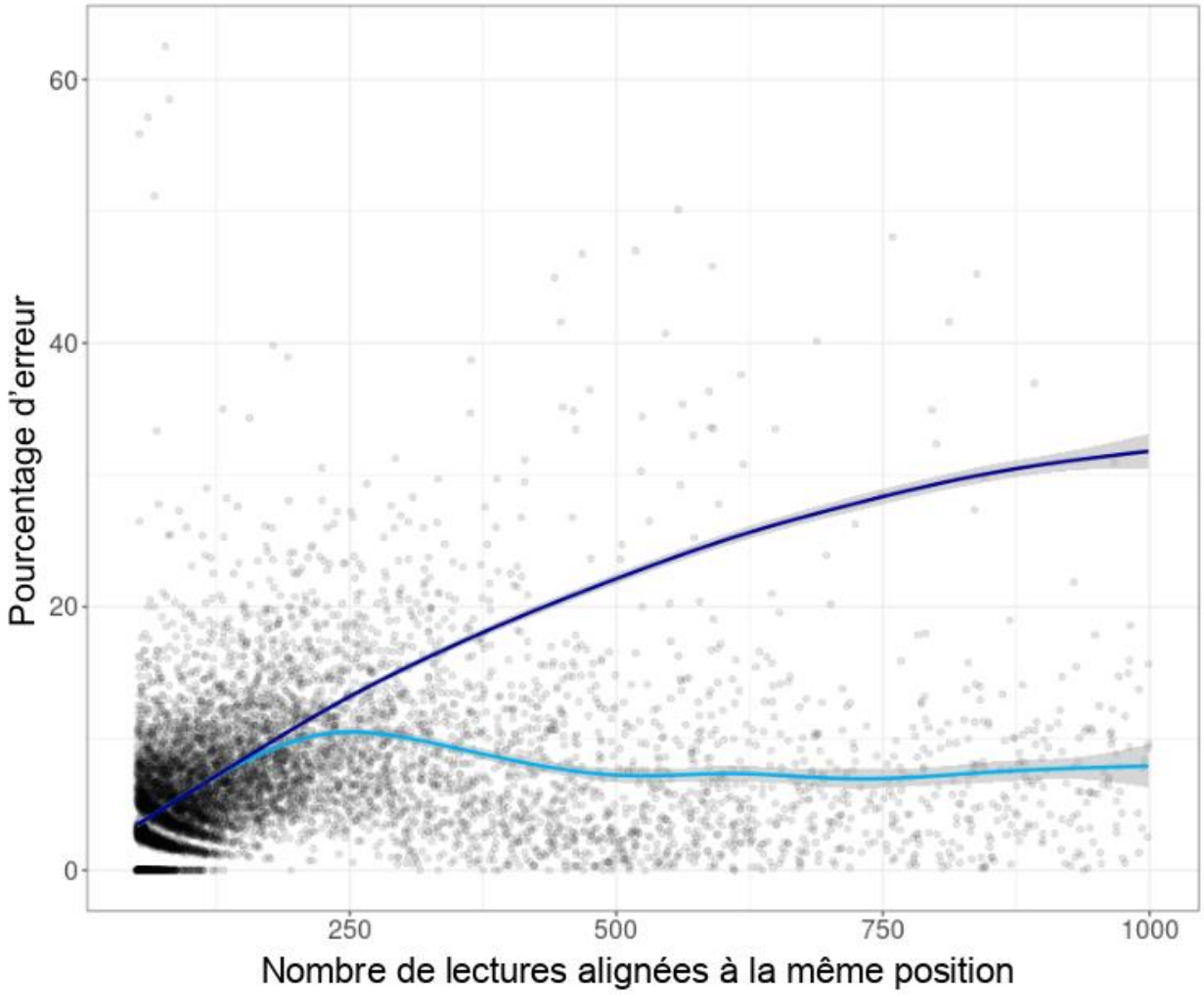


Figure 117 : Performance de la désaturation d'UMI de taille 7 pour les positions avec moins de 1 000 lectures alignées. *Pourcentage d'erreur en fonction du nombre de lectures alignées à une même position pour les UMI de taille 7.* ***Bleu foncé*** *: UMI de taille 7 non désaturés,* ***Bleu clair*** *: UMI de taille 7 désaturés.*

Notre méthode de désaturation permet notamment une meilleure estimation du nombre réel de transcrits que (i) l'utilisation du nombre de lectures, et aussi (ii) le nombre d'UMI uniques à une position (si la taille de l'UMI est insuffisante et qu'il y a saturation). Bien que des erreurs persistent, notre méthode contribue à les réduire.

Ainsi, cette approche va nous permettre d'analyser les données obtenues avec des UMI de 7 nucléotides en atténuant les biais liés à l'amplification PCR et à la saturation des UMI. Toutefois, il est recommandé d'utiliser dès le départ des UMI d'une taille appropriée pour une meilleure précision.

2.2.1.1.4.2 Filtrage des données

Dans le but d'éliminer les positions peu abondantes qui pourraient correspondre à du bruit de fond ou à des erreurs de mesure, nous avons filtré les données de comptages désaturés. Il est crucial de ne filtrer les positions considérées que par rapport à l'échantillon correspondant au $pool^{total}$, car l'objectif de l'analyse est de comparer les proportions d'ARN 5'P. Cette proportion est calculée en prenant, pour chaque position, le rapport des comptages de l'échantillon $pool^{mono}$ (ne contenant que les transcrits initialement 5'P) sur les comptages de l'échantillon $pool^{total}$ (contenant à la fois les transcrits initialement 5'P et 5'PPP). Il est donc possible que certains ARN aient une proportion d'extrémité 5'P égale à 0 dans le cas où ils n'auraient pas ou pas encore été déphosphorylés après transcription, ce qui signifie que le nombre de lectures dans l'échantillon $pool^{mono}$ est égal à 0. Cette information doit donc être conservée. Notons également que les positions du $pool^{mono}$ associées aux positions supprimées du $pool^{total}$ seront également retirées de l'analyse. Pour l'analyse de l'ensemble des données PQ-EMOTE, nous avons utilisé un seuil de 50 comptages afin qu'une variation d'un seul comptage dans l'échantillon $pool^{mono}$ ne fasse varier la proportion d'extrémité 5'P que de 2% ($\frac{1}{50}$ = 2 %).

2.2.1.1.4.3 Normalisation des données

Pour calculer la proportion d'extrémités 5'P, il est nécessaire de normaliser les deux échantillons $pool^{mono}$ et $pool^{total}$ afin de tenir compte du fait que, dans le protocole expérimental, la séparation de l'échantillon d'ARN total en deux pools ne permet pas de diviser la quantité de matériel de manière exacte.

Nous avons donc procédé à la normalisation des échantillons $pool^{mono}$ et $pool^{total}$. Les méthodes de normalisation conventionnelles, qui prennent en compte la longueur des gènes et les différences de profondeur de séquençage, ne sont pas adaptées dans ce contexte. En effet, la quantification

se fait par position exacte et non par gène, rendant toutes les longueurs de position égales à 1. De plus, les extrémités séquencées dans la condition pool^{total} correspondent aux ARN 5'P et 5'PPP alors qu'elles correspondent seulement aux ARN 5'P dans la condition pool^{mono} qui sont attendues en plus faibles quantités. Les distributions des lectures sur les deux pools ne devraient donc pas correspondre.

Pour normaliser les paires d'échantillons, nous avons donc choisi d'utiliser une méthode basée sur un ensemble de référence, c'est-à-dire un ou plusieurs ARN de contrôle. Ces ARN de contrôle doivent être présents en quantités identiques dans les deux échantillons, ce qui implique qu'ils doivent être exclusivement sous la forme 5'P afin que l'ajout de l'enzyme *Ec*RppH dans le pool^{total} n'influence pas la quantité de lectures obtenues par rapport au pool^{mono}.

Le facteur de normalisation (nf) est calculé en faisant la médiane des proportions de forme 5'P des ARN de contrôle (Formule 6). Par exemple, pour un ensemble de 2 ARN de contrôle ARN_1 et ARN_2 :

$$\text{nf} = \text{médiane}\left(\frac{comptages\ désaturés\ ARN_1\ pool^{mono}}{comptages\ désaturés\ ARN_1\ pool^{total}} ; \frac{comptages\ désaturés\ ARN_2 pool^{mono}}{comptages\ désaturés\ ARN_2\ pool^{total}}\right)$$

Formule 6 : Calcul du facteur de normalisation global.

Les comptages normalisés sont ensuite obtenus en divisant les comptages du pool^{mono} par le facteur de normalisation.

Nous avons opté pour l'utilisation de la médiane plutôt que la moyenne en raison de sa moindre sensibilité aux valeurs extrêmes. En effet, dans le cas de la moyenne, une valeur de proportion nettement plus élevée ou plus faible que les autres peut fortement influencer le facteur de normalisation et le rendant peu représentatif de la majorité des ARN de contrôle. La médiane, en revanche, offre une mesure centrale plus robuste, assurant une représentation plus fidèle des données dans leur ensemble.

Nous avons testé 3 ensembles d'ARN de contrôle :

- Les trois ARN ribosomiques (5S, 16S et 23S) qui sont toujours monophosphorylés en 5' et très abondants (en moyenne 300 000 comptages pour leurs extrémités 5' dans le pool^{total})
- Les ARNr plus les extrémités 5' correspondant aux 99 sites de clivages endoribonucléolytiques de la RNase Y identifiés par Khemici et al. (2015), eux aussi toujours 5'P. Il convient de noter que certains de ces sites, de par leur faible abondance, ont été éliminés lors de l'étape de filtrage des données, car leur valeur de comptage était inférieure au seuil fixé de 50 comptages. Le nombre d'ARN de contrôle utilisé pour le calcul du facteur de normalisation peut donc varier d'une condition expérimentale à l'autre
- L'ensemble des extrémités toujours 5'P, c'est-à-dire, par exemple, les ARNr, les sites de clivage endoribonucléolytiques, les intermédiaires de dégradation exoribonucléolytiques, les ARNt

ARNr

La Figure 118 représente la valeur de proportion de forme 5'P pour chacun des trois ARNr dans chacun des échantillons, calculé à partir des comptages normalisés. Nous pouvons observer que même après normalisation, ces valeurs, bien que centrées sur 1, sont assez variables (entre 0,6 et 1,6). Lorsque l'on normalise deux échantillons avec cette méthode, on procède à une normalisation linéaire. Cela signifie que, indépendamment du facteur de normalisation appliqué aux comptages, les différences relatives entre les valeurs resteront constantes. En d'autres termes, la normalisation linéaire préserve les écarts initiaux entre les ARN, ce qui signifie que les écarts observés entre les proportions des extrémités 5'P des 3 ARNr resteront toujours aussi importants.

Ces variations sont surprenantes car elles mettent en lumière une difficulté dans l'estimation précise des proportions de forme 5'P, suggérant une répartition non homogène des transcrits dans le protocole expérimental lors de la séparation en deux pools.

Comme cela a été décrit dans une section précédente, plus le nombre de lectures alignées à une même position est important, plus il sera difficile d'estimer correctement le nombre réel de transcrits, or le nombre de comptages correspondant aux ARNr est très élevé (>300 000 comptages en moyenne). Bien que la méthode de désaturation permette de réduire le pourcentage d'erreur, ce pourcentage augmente néanmoins avec le nombre lectures alignées à la même position. Ceci peut rendre l'estimation de la proportion de forme 5'P erronée, et amener à ces variations, notamment si le nombre de lectures de chaque ARNr est différent du fait que le phénomène de saturation n'est

pas linéaire. Cependant, les proportions issues de la librairie PQ-EMOTE 62, où aucune saturation n'est présente, sont toutes aussi variables.

Ces observations nous ont amené à augmenter la taille de notre ensemble d'ARN de contrôle afin de mieux estimer la dispersion dans l'estimation des proportions d'extrémités 5'P.

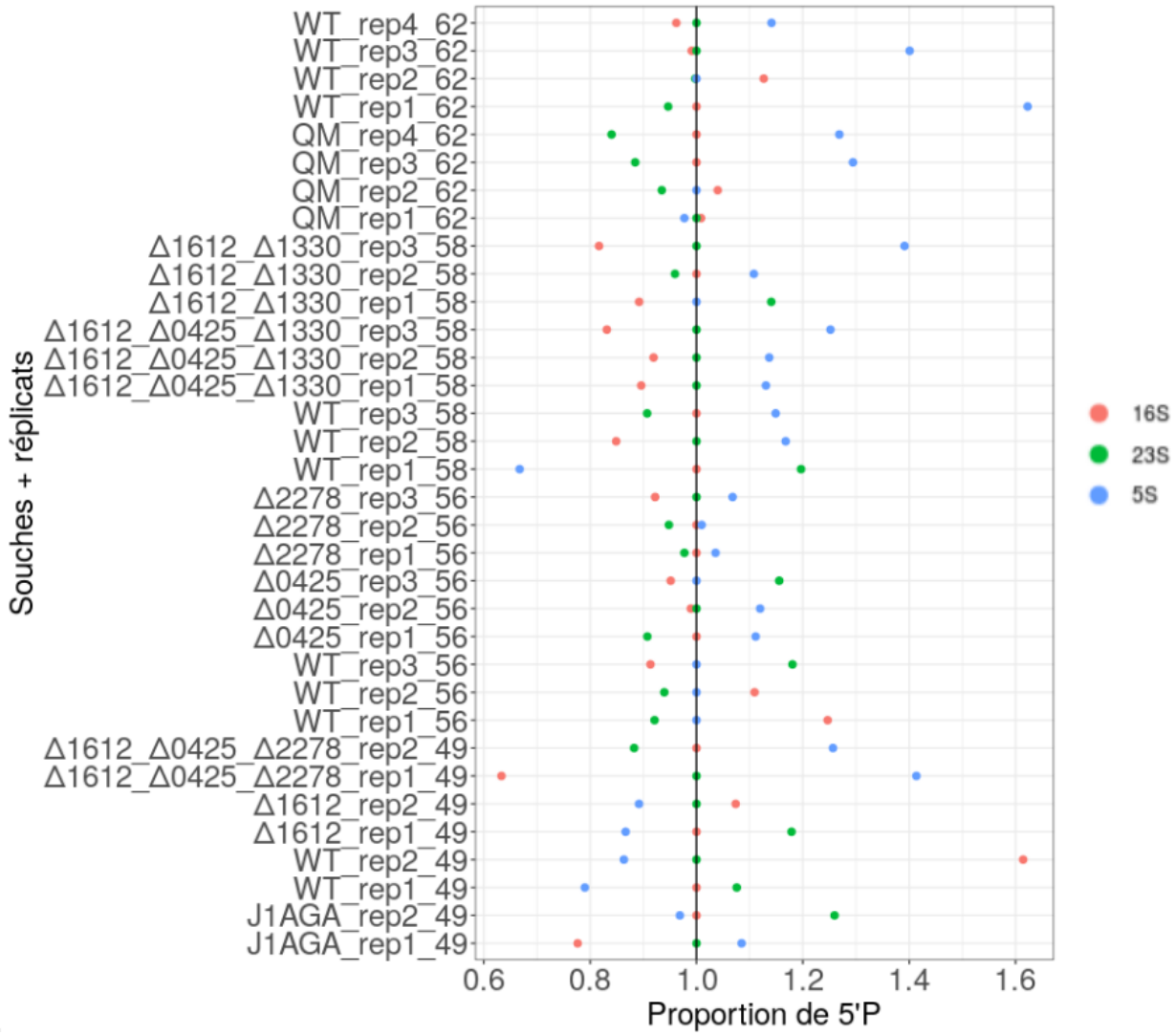


Figure 118 : ***Proportion de forme 5'P (axe x) des ARN ribosomiques après la normalisation des différentes souches et réplicats****. La ligne verticale correspond à une proportion d'extrémités 5'P égale à 1. Les couleurs correspondent aux différents ARNr, rouge pour le 16S, vert pour le 23S et bleu pour le 5S.*

ARNr + sites de clivages endoribonucléolytiques de la RNase Y

L'augmentation du nombre d'ARN de contrôle va permettre de mieux estimer la dispersion des proportions d'extrémités 5'P. En utilisant plusieurs ARN de contrôle, l'impact des fluctuations associées à un seul gène de contrôle est réduit. En utilisant plusieurs contrôles, l'effet de ces erreurs devrait être moindre sur valeur centrale utilisée comme facteur de normalisation, ce qui devrait conduire à une normalisation plus robuste.

En plus des 3 ARNr, nous avons pris en compte les extrémités 5' des 99 positions correspondant à des sites de clivage endoribonucléolytiques de la RNase Y identifiés par Khemici et al. (2015). Du fait du clivage endoribonucléolytique, nous sommes certains qu'ils sont toujours monophosphorylés en 5', et donc devraient être également en même nombre dans chacun des deux pools de l'échantillon.

La Figure 119 représente la valeur de proportion de forme 5'P, calculé à partir des comptages normalisés, pour chaque ARN de contrôle utilisés dans chacun des échantillons. Bien que les distributions de proportion d'extrémité 5'P des ARN de contrôle soit centrée sur 1, nous observons également une variation importante dans l'estimation de ces proportions (plus ou moins 0,5). En effet, la proportion d'extrémité 5'P censée être égale à 1 peut être estimée entre 0,5 et 1,5. Dans le premier cas, l'erreur sera d'un facteur 2 et dans le second d'un facteur 1,5.

On remarque également que la taille des ensembles d'ARN de contrôle utilisés pour normaliser les échantillons issus de la librairie PQ-EMOTE 49, est faible puisque seulement 5 sites ont pu être utilisés en moyenne. En revanche, dans les autres librairies, en moyenne 55 sites sont utilisés pour le calcul du facteur de normalisation.

Ces résultats montrent, à partir de davantage d'ARN de contrôle (plutôt que seulement les 3 ARNr), que la dispersion des valeurs de proportions observées pour des extrémités toujours 5'P est globalement comprise entre 0,5 et 1,5.

Par exemple, la condition expérimentale « QM_rep1_62 », correspondant au réplicat 1 de la souche Δ1612_Δ1330_Δ0425_Δ2278 issue de la librairie PQ-EMOTE 62, semblait présenter une faible variabilité lorsque seuls les trois ARNr étaient pris en compte. Cependant, l'ajout de nouveaux ARN de contrôle a révélé une variation comparable à celle observée dans les autres conditions expérimentales. Grâce à cela, nous pouvons conclure que le niveau de variation dans l'estimation des proportions d'extrémités 5'P n'est pas lié à une condition expérimentale ou à une librairie en particulier.

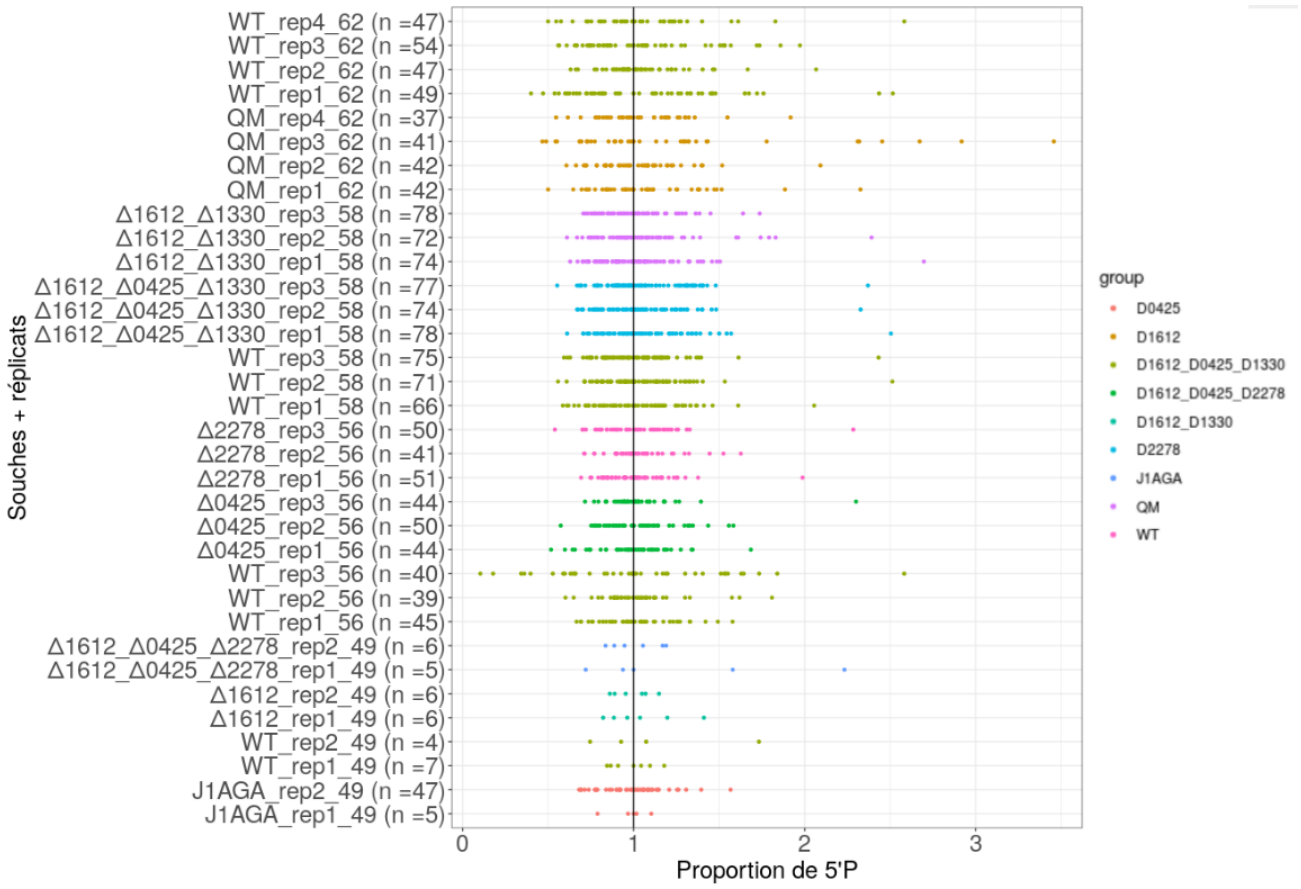


Figure 119 : Proportion de forme 5'P des ARN ribosomiques et les sites de clivages endoribonucléolytiques de la RNase Y après la normalisation des différentes souches et réplicats*. La ligne verticale correspond à une proportion d'extrémités 5'P égale à 1.*

La plupart des écarts de proportion par rapport à 1 est de plus ou moins 0,5. Cela confirme les observations quant à la variabilité des proportions d'extrémités 5' des ARN pourtant toujours 100 % 5'P. De la même manière que précédemment, le problème semble être dû à l'étape de séparation des échantillons en 2 pools, et non aux traitements des données. En effet, les sites de clivages endoribonucléolytiques de la RNase Y utilisés pour le calcul des facteurs de normalisation ont en moyenne 213 comptages dans le pool^{total}. Cela signifie que pour la librairie PQ-EMOTE 62, l'estimation du nombre de transcrits alignés à ces positions est très fiable, et pourtant le même niveau de variation est observé.

Toutefois, l'utilisation de réplicats biologiques devrait permettre de tenir compte de la variabilité, et donc de l'incertitude sur les mesures effectuées, et ainsi augmenter la confiance dans ces estimations puisqu'ils fournissent une mesure plus robuste et fiable des tendances réelles, renforçant ainsi la validité des conclusions que nous pourrons tirer de ces expériences.

Tous les 5’P

La méthode proposée par (Martini et al., 2021) est conçue pour normaliser deux échantillons du même type que PQ-EMOTE, le premier contenant des extrémités 5’P ($pool^{mono}$) et le second contenant des extrémités 5’P et 5’PPP ($pool^{total}$). Cette méthode repose sur le même principe d’ARN de contrôle, mais au lieu de définir *a priori* un ensemble d’ARN de contrôle, la méthode utilise toutes positions censées être des extrémités toujours 5’P, y compris les positions correspondant aux extrémités 5’ des intermédiaires de dégradation exoribonucléolytiques. Pour ce faire, les $\log_2$ des rapports $\frac{5'P+5'PPP}{5'P}$ de l’ensemble des positions d’un échantillon sont représentés sous forme d’une courbe de densité (supportée par un histogramme) (Fig. 120). Cette courbe est construite à partir de deux populations sous-jacentes mélangées : à gauche, un pic correspondant aux extrémités 100 % 5’P (résultants de clivages ou de dégradation) qui devrait donc être centré sur 0 ($\log_2(1) = 0$), et à droite un continuum de proportions 5’P pour les TSS. Le pic de gauche devrait donc être centré sur 0. La partie de droite représente les extrémités correspondant à des départs de transcription (TSS) dont le ratio est variable. Le facteur de normalisation est alors calculé de manière à ajuster la distribution pour que le pic de gauche soit centré sur 0.

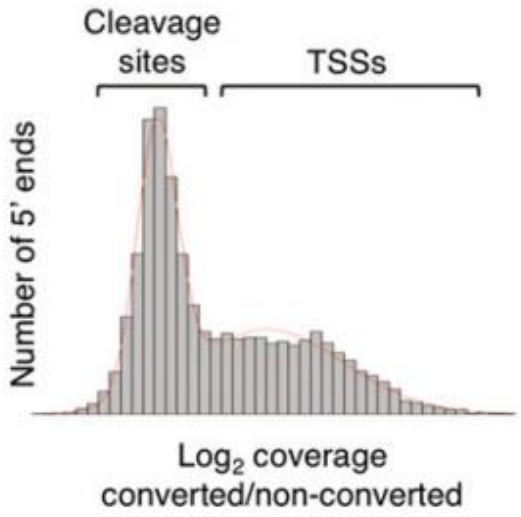


***Figure 120 :* Principe de la normalisation bimodale.** *Courbe de densité et histogramme des ratios des abondances entre deux échantillons 5’P+5’PPP (converted) et 5’P (non-converted). Extrait de* (Martini et al., 2021)

Étant donné la profondeur de séquençage, l'immense majorité des intermédiaires de dégradation sont supprimés lors de l'étape de filtrage à 50 comptages. Pour utiliser cette approche, nous avons donc abaissé le seuil de filtrage. Plusieurs valeurs ont été testées (1, 5, 10, 20 et 50) et le seuil retenu est celui à 5 comptages, car c'est celui qui permet de conserver un maximum de positions correspondant aux intermédiaires de dégradation tout en filtrant une grande partie du bruit dû aux positions faiblement abondantes dans le pool[total].

Cette approche est très intéressante car elle permet de prendre en compte toutes les extrémités pouvant servir de contrôle, ce qui conduit à une normalisation plus stable et robuste. Cependant, elle dépend fortement de la profondeur de séquençage et du seuil de filtrage utilisé. Dans le séquençage des extrémités 5' des ARN, les extrémités correspondent majoritairement aux TSS. Pour observer suffisamment d'intermédiaires de dégradation afin de calculer des ratios fiables et de correctement séparer les deux populations, une grande profondeur de séquençage est nécessaire.

Nous présentons les courbes de densité des $\log_2$ des rapports $\frac{5'P+5'PPP}{5'P}$, avant normalisation, pour chaque paire d'échantillon des 4 librairies dans la Figure 121. Nous observons pour l'ensemble des paires d'échantillons un pic plutôt centré sur 0 qui correspondrait aux intermédiaires de dégradation, sites de clivages endoribonucléolytique (points rouges) et autres ARN toujours 5'P. En revanche, la région correspondant aux TSS est difficilement discernable voire indiscernable. Encore une fois, nous sommes confrontés à une dispersion des proportions d'extrémités 5'P très importante autour de la valeur centrale.

Bien que, théoriquement, seul le pic de gauche soit requis pour appliquer cette normalisation, cette dispersion et le mélange des deux populations rendent difficile le calcul du facteur de normalisation. C'est pourquoi nous avons décidé de ne pas utiliser cette méthode de normalisation, et nous avons utilisé celles avec les ARNr et les sites de clivages endoribonucléolytiques.

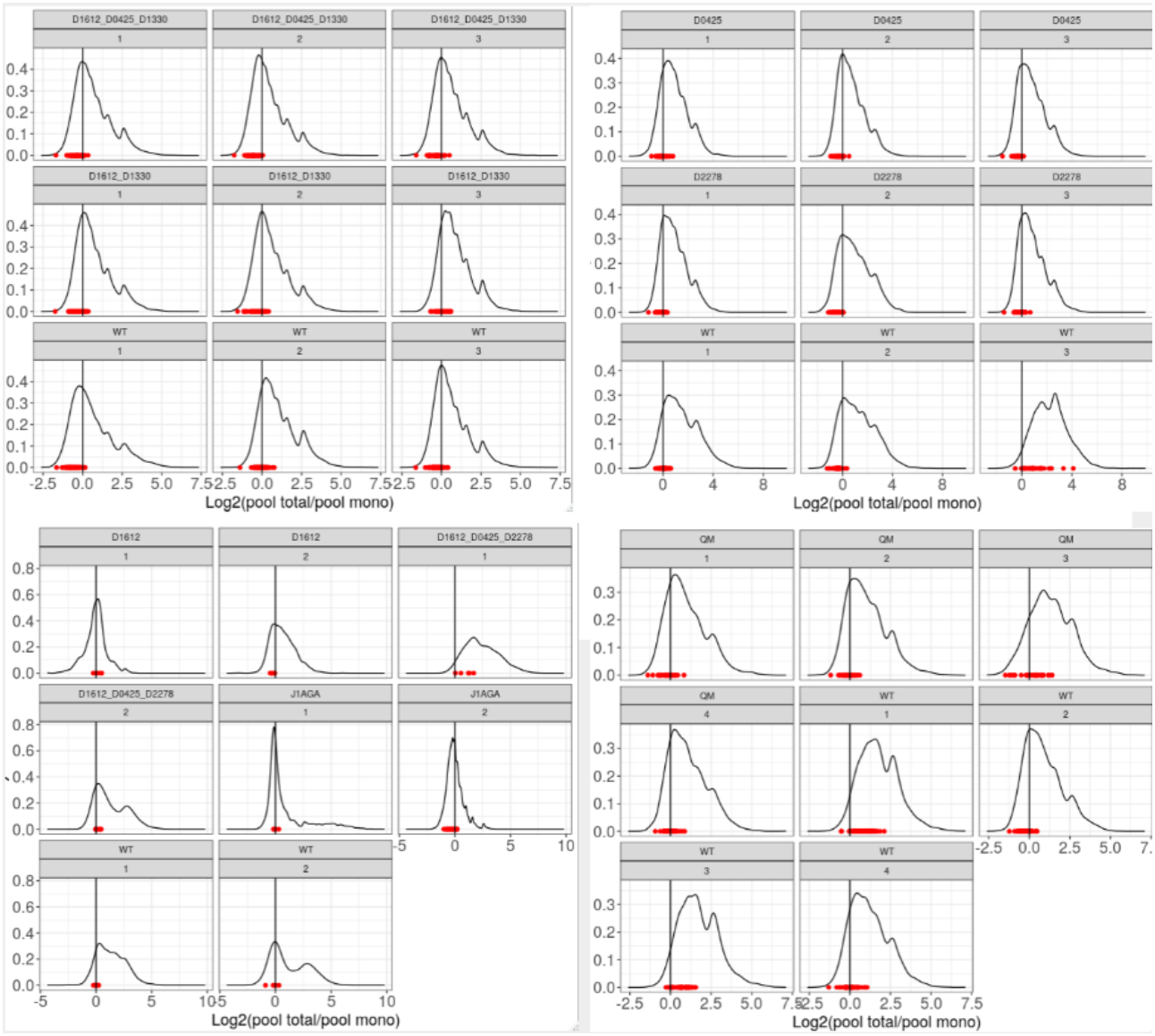


Figure 121 : Courbes de densité des logs en base 2 des ratios pool^{total} / pool^{mono} des extrémités 5' des ARN*. Les échantillons pool^{total} a été filtré à 5 comptages. Les points rouges représentent les 5' utilisés comme ARN de contrôle précédemment.*

Pour conclure, indépendamment de la méthode de normalisation employée, la dispersion des proportions d'extrémités 5'P reste très importante. La présence de plusieurs réplicats biologiques dans chacune des librairies PQ-EMOTE peut améliorer la fiabilité des valeurs calculées. Cependant, il est impératif de mener des investigations approfondies, notamment sur l'étape de séparation des échantillons, afin de comprendre les causes de ces variations.

De plus, l'utilisation d'une valeur ou d'un facteur de normalisation unique pour toutes les positions d'un échantillon pose des problèmes. En effet, bien que les distributions des proportions d'extrémités 5'P des ARN de contrôle soit centré sur 1, la variation n'est pas uniforme selon les positions. Ainsi, quel que soit le facteur de normalisation utilisé, il sera adéquat pour certaines positions et inadéquat pour d'autres. Une approche plus fine et position-spécifique pourrait être nécessaire pour obtenir des estimations plus précises et fiables.

Malgré cela, pour identifier la ou les RppH de *S. aureus*, une diminution globale de la proportion d'extrémités 5'P dans la souche mutant ΔrppH est attendue. Cette réduction devrait être détectable, au moins de manière globale, même en présence de telles variations.

Après avoir effectué, la désaturation sur les librairies PQ-EMOTE 49, 56 et 58, puis filtré et normalisé l'ensemble des 4 librairies en utilisant l'ensemble de contrôle contenant les ARNr et les sites de clivages de la RNase Y, nous obtenons, par paires d'échantillon, un certain nombre de positions sur le génome pour lesquelles il nous est possible de calculer une proportion d'extrémité 5'P. Ces chiffres sont présentés dans la Table 13.

Maintenant que les données sont nettoyées, c'est-à-dire désaturées, filtrées et normalisées, nous sommes en mesure d'entamer leur analyse approfondie. Nous procéderons notamment à une analyse en composantes principales (ACP) afin d'examiner comment se regroupent les échantillons en fonction de leurs proportions d'extrémités 5'P et d'identifier d'éventuels effets de lot. Cette étape cruciale nous permettra de vérifier la cohérence des échantillons et de déceler toute variation systématique non biologique pouvant influencer les résultats.

Souche	Réplicat	Librairie	Nombre de positions
Sauvage	1	PQ-EMOTE 49	373
Sauvage	2	PQ-EMOTE 49	252
Δ1612	1	PQ-EMOTE 49	155
Δ1612	2	PQ-EMOTE 49	184
Δ1612_Δ0425_Δ2278	1	PQ-EMOTE 49	329
Δ1612_Δ0425_Δ2278	2	PQ-EMOTE 49	280
J1AGA	1	PQ-EMOTE 49	3476
J1AGA	2	PQ-EMOTE 49	2067
Sauvage	1	PQ-EMOTE 56	1728
Sauvage	2	PQ-EMOTE 56	1571
Sauvage	3	PQ-EMOTE 56	1663
Δ0425	1	PQ-EMOTE 56	889
Δ0425	2	PQ-EMOTE 56	1125
Δ0425	3	PQ-EMOTE 56	958
Δ2278	1	PQ-EMOTE 56	1093
Δ2278	2	PQ-EMOTE 56	1337
Δ2278	3	PQ-EMOTE 56	1482
Sauvage	1	PQ-EMOTE 58	2558
Sauvage	2	PQ-EMOTE 58	2617
Sauvage	3	PQ-EMOTE 58	2977
Δ1612-Δ1330	1	PQ-EMOTE 58	2749
Δ1612-Δ1330	2	PQ-EMOTE 58	2490
Δ1612-Δ1330	3	PQ-EMOTE 58	2641
Δ1612-Δ0425-Δ1330	1	PQ-EMOTE 58	2871
Δ1612-Δ0425-Δ1330	2	PQ-EMOTE 58	2598
Δ1612-Δ0425-Δ1330	3	PQ-EMOTE 58	2744
Sauvage	1	PQ-EMOTE 62	1426
Sauvage	2	PQ-EMOTE 62	1723
Sauvage	3	PQ-EMOTE 62	1789
Sauvage	4	PQ-EMOTE 62	1530
Δ1612-Δ1330-Δ0425-Δ2778	1	PQ-EMOTE 62	1380
Δ1612-Δ1330-Δ0425-Δ2778	2	PQ-EMOTE 62	1542
Δ1612-Δ1330-Δ0425-Δ2778	3	PQ-EMOTE 62	1293
Δ1612-Δ1330-Δ0425-Δ2778	4	PQ-EMOTE 62	1235

Table 13 : ***Nombre de positions par souches, réplicats et librairies pour lesquelles nous avons pu calculer une proportion d'extrémité 5'P.***

2.2.1.1.4.4 Effets de lots

Étant donné que chacune des librairies intègre des échantillons issus de la souche sauvage, il pourrait être intéressant de les traiter ensemble. Ceci augmenterait le nombre de réplicats biologiques pour cette souche, et nous permettrait donc d'obtenir des valeurs plus fiables. Cependant, les librairies PQ-EMOTE ont été réalisées sur plusieurs années et par différentes personnes, ce qui laisse supposer l'existence possible d'effets de lots entre ces librairies. L'objectif est donc de déterminer s'il existe des effets de lots et, le cas échéant, de les corriger.

L'analyse en composantes principales (ACP) peut aider à détecter des effets de lots dans les données. L'ACP réduit les données à un petit nombre de composantes principales qui capturent la majeure partie de la variance présente dans les données.

S'il n'y a pas d'effets de lots, les échantillons devraient se regrouper par réplicats et par souches, au moins ceux issus de la souche sauvage. En revanche, si les effets de lots existent, ils peuvent constituer une source majeure de cette variance et être capturés par les premières composantes principales. Les échantillons pourraient ainsi se regrouper par librairies dans cet espace et non par souche. Ceci indiquerait une variation systématique entre les lots, ce qui est un signe d'effet de lots.

Cependant, il est crucial de s'assurer qu'il y a suffisamment de variables communes entre les lots avant d'appliquer une analyse en composantes principales (ACP). La raison en est que les variables communes permettent de comparer les lots de manière cohérente. Si les variables diffèrent entre les lots, l'ACP ne pourra pas correctement capturer les variations entre eux, ce qui pourrait biaiser les résultats ou les rendre non représentatifs de la véritable structure des données. C'est pourquoi, dans un premier temps, nous avons déterminé le nombre de positions pour lesquelles nous avons pu obtenir des valeurs (après filtrage) pour l'ensemble des échantillons des 4 librairies. Nous avons éliminé les positions correspondant aux ARNr, car les ribodéplétions n'étant pas homogènes, les effets détectés ne seraient pas imputables aux lots.

La Figure 122 illustre le nombre de positions communes parmi l'ensemble des échantillons des quatre librairies PQ-EMOTE. Il est observé que seules 50 positions présentent des valeurs disponibles dans l'ensemble des échantillons analysés.

Ce nombre est trop bas pour qu'il soit pertinent d'effectuer une ACP qui ne sera pas représentative de l'ensemble des positions de chaque échantillon. Par conséquent, il serait également peu approprié de rechercher et corriger des effets de lots. En effet, les méthodes d'identification et de correction des effets de lots, telles que Combat (Johnson et al., 2007) ajustent les valeurs en utilisant les informations partagées entre les lots.

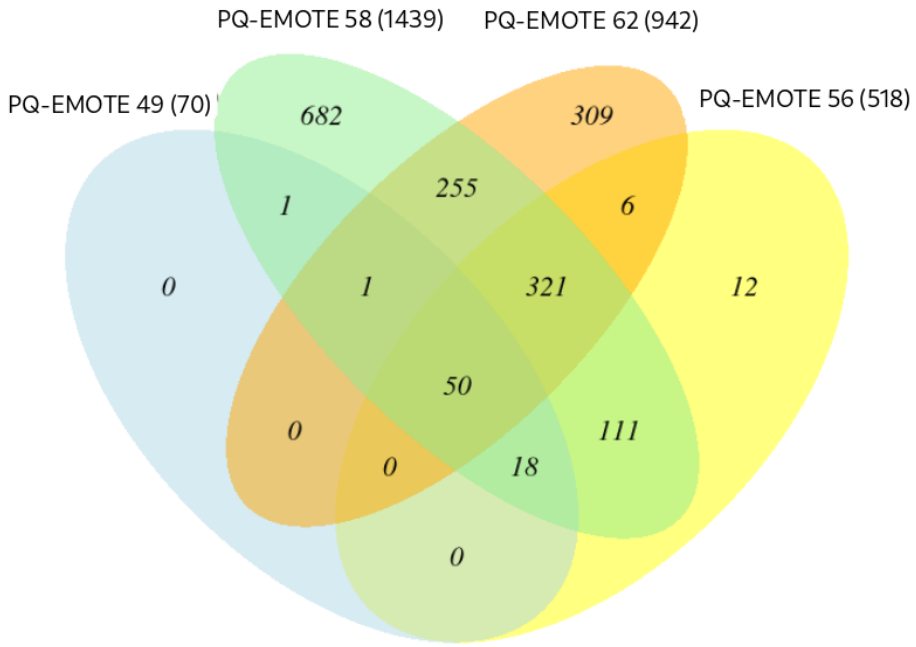


Figure 122 : Nombre de positions pour lesquelles des valeurs de proportions sont disponibles et ont pu être calculées dans l'ensemble des réplicats de chaque librairie PQ-EMOTE.

Chacune des librairies PQ-EMOTE présentent au moins 2 réplicats issus de la souche sauvage. En se restreignant à cette dernière, nous auront peut-être suffisamment de positions pour effectuer une ACP et identifier des éventuels effets de lots.

La Figure 123 montre que nous obtenons cette fois 120 positions avec des valeurs disponibles dans l'ensemble des échantillons. De nouveau, il ne serait pas pertinent de rechercher des effets de lots dans un ensemble si peu représentatif de positions. Bien qu'il soit difficile de déterminer le nombre exact de positions communes nécessaires pour détecter des effets de lots, on peut estimer qu'il faudrait au moins que la moitié des positions soient communes afin d'obtenir un équilibre informatif entre les données disponibles et celles manquantes. Par conséquent, même si l'on retirait la librairie PQ-EMOTE 49, qui est la moins « profonde », il ne resterait que 463 positions communes. Seule la librairie PQ-EMOTE 56 aurait alors plus de la moitié de ses positions en commun, ce qui reste insuffisant pour une analyse robuste des effets de lots.

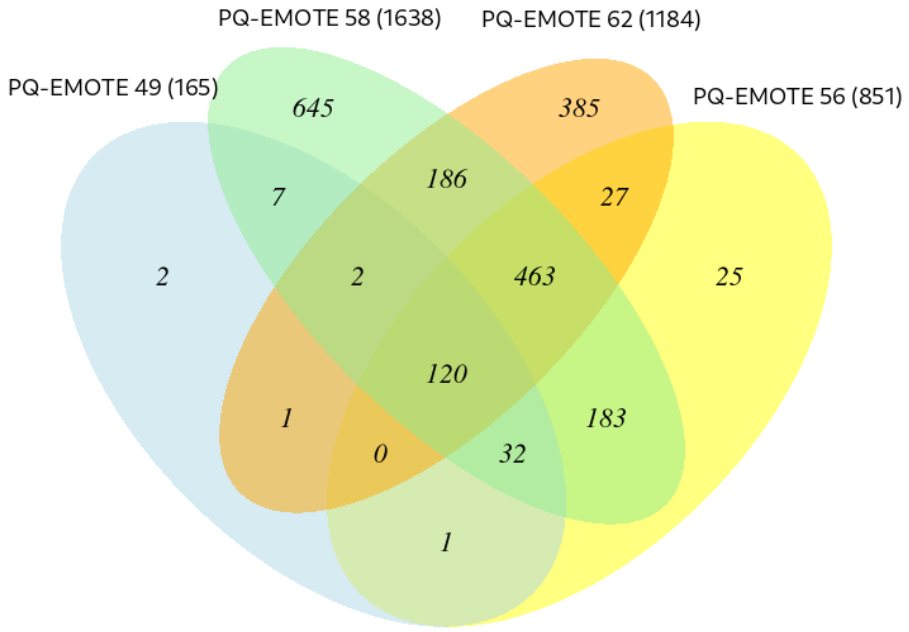


Figure 123 : Nombre de positions pour lesquelles des valeurs de proportions sont disponibles et ont pu être calculées dans l'ensemble des réplicats issus de la souche sauvage de chaque librairie PQ-EMOTE.

Bien que l'imputation de valeurs soit envisageable, le nombre élevé de valeurs manquantes rend cette approche inappropriée, car elle introduirait trop de valeurs artificielles, créant ainsi une fausse impression de précision qui ne refléterait pas l'incertitude inhérente aux valeurs manquantes. De plus, les valeurs imputées peuvent, par définition, être erronées, ce qui pourrait fausser les résultats et les conclusions.

Comme nous n'avons pas pu déterminer s'il y a des effets de lots ou non, nous avons décidé de ne pas traiter les 12 échantillons sauvages comme des réplicats biologiques. Par conséquent, nous traiterons les librairies individuellement, ce qui est possible grâce à la présence d'échantillons issus de la souche sauvage dans chacune des librairies.

Il est important de souligner que les données PQ-EMOTE peuvent ne pas être sensibles aux effets de lots, étant donné qu'elles correspondent à des proportions. En effet, ajuster les valeurs par lots reviendrait à appliquer un facteur d'ajustement aux comptages. Cependant, les échantillons appartenant aux mêmes lots auraient des comptages ajustés de manière proportionnelle, ce qui ne devrait pas altérer la valeur de la proportion d'extrémités 5'P calculée par la suite. Néanmoins, étant donné que nous n'avons pas pu vérifier cette hypothèse, nous préférons analyser chaque librairie individuellement.

2.2.1.1.5 Analyses des données PQ-EMOTE : Comparaison des proportions d'extrémités 5'P

Une fois les données post-traitées et prêtes à être analysées, nous procéderons au calcul des proportions de forme 5'P aux extrémités des ARN. Ces analyses visent principalement à évaluer deux aspects essentiels. Premièrement, nous déterminerons la proportion médiane d'extrémités 5'P dans la souche sauvage de *S. aureus*, ce qui nous permettra d'estimer la prévalence des ARN servant de substrat privilégié pour la RNase J. Deuxièmement, afin d'étudier l'effet sur les états de phosphorylation des extrémités 5' des ARN des mutants des différents candidats NUDIX testés, nous comparerons les distributions des proportions de la forme 5'P entre les souches sauvages et mutantes.

2.2.1.1.5.1 Analyse des résultats obtenus sur les souches sauvages

Pour évaluer la proportion d'extrémités 5'P dans les souches sauvages, nous avons calculé, pour chaque position, le rapport des comptages après désaturation, filtrage et normalisation (avec ensemble de contrôle composé des ARNr et des sites de clivage de la RNase Y) issus du pool^{mono} sur ceux issus du pool^{total}. Les TSS sont les seules positions pouvant posséder une forme tri-phosphorylée en 5' susceptible d'avoir été déphosphorylée par la RppH endogène de *S. aureus***.** Nous avons donc restreint notre analyse aux TSS identifiés par orthologie avec ceux identifié dans la souche MW2 de *S. aureus* par (Prados et al., 2016). Cela nous permet aussi de réduire le bruit de fond apporté par les proportions d'extrémités 5'P d'intermédiaires de dégradation, des sites de clivage endoribonucléolytiques, et des ARNr par exemple.

Afin d'évaluer la reproductibilité des données au sein et entre les librairies, et par la suite avoir une vue d'ensemble de la proportion d'extrémités 5'P chez la souche sauvage de *S. aureus*, nous avons examiné et comparé les distributions de ces proportions :

- par réplicat,
- par librairie,
- pour tous les échantillons regroupés.

La Figure 124 représente la distribution des proportions des extrémités 5'P par réplicat (panel A) et par librairie (panel B), en utilisant la médiane des proportions par position entre réplicats.

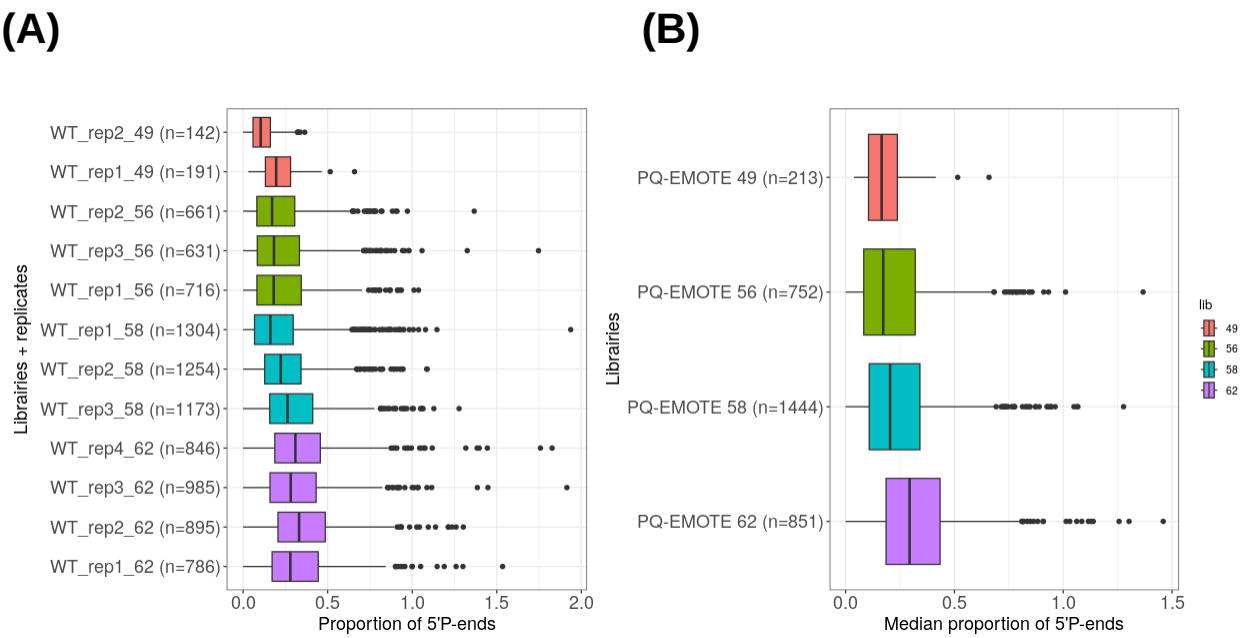


Figure 124 : Proportion d'extrémités 5'P dans la souche sauvage par librairie.

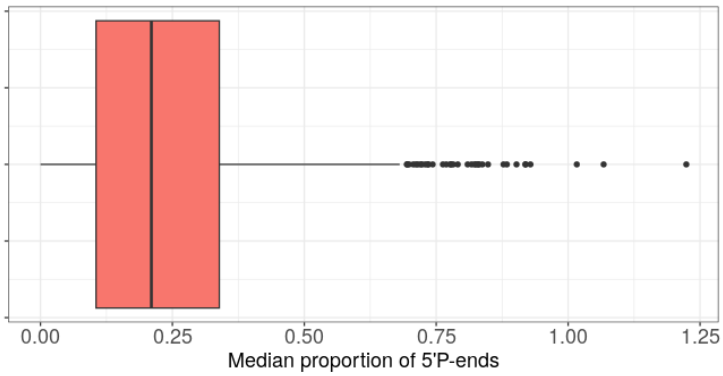
*Boîtes à moustaches représentant les distributions des proportions des extrémités 5'P dans les échantillons issus de la souche sauvage des différentes librairies. (**A**) les données regroupées par réplicat biologique. (**B**) les données sont regroupées par libraire (médiane des proportions à la même position dans les différents réplicats).*

Les réplicats sauvages provenant des mêmes librairies montrent une reproductibilité peu satisfaisante puisque les p-valeurs ajustée (BH) obtenues à partir des tests de Wilcoxon-Mann-Whitney montrent des différences statistiquement significatives entre les deux réplicats issus de la librairie PQ-EMOTE 49 et entre tous les réplicats de la librairie PQ-EMOTE 58. De plus, des différences statistiquement significatives sont également observées entre les réplicats 1 et 2, 1 et 4, 2 et 3, 3 et 4 de la librairies PQ-EMOTE 62.

Nous observons également des différences de proportion d'extrémité 5'P entre les librairies puisque les p-valeurs ajustée (BH) obtenues par les tests de Wilcoxon-Mann-Whitney sont toutes inférieures à 0,05. Ces résultats montrent que les proportions d'extrémités 5'P des TSS sont statistiquement inférieures dans la librairie PQ-EMOTE 49 par rapport à celles de la librairie PQ-EMOTE 56. De même, les proportions dans la librairie PQ-EMOTE 56 sont inférieures à celles de la librairie PQ-EMOTE 58, qui sont elles-mêmes inférieures à celles de la librairie PQ-EMOTE 62.

Ces différences sont tout de même assez faibles comparé à ce qui est attendu dans les mutants NUDIX puisque par exemple, la médiane des proportions d'extrémités 5'P des TSS issus de la librairie PQ-EMOTE 62 est de 0,29 tandis que celle de la librairie PQ-EMOTE 49 est de 0,20.
Comme mentionné précédemment, les librairies ont été générées sur plusieurs années avec un protocole qui a évolué. Ceci peut entraîner des variations, particulièrement pour des mesures dont, comme nous l'avons également vu, l'estimation est peu précise (plus ou moins 0,5).

Cette constatation souligne l'importance cruciale de générer des données à partir de souches sauvages pour chaque librairie. Cependant, cela confirme également qu'il ne sera pas possible d'utiliser tous les 12 réplicats sauvages ensemble pour les comparer à chaque mutant de délétion NUDIX. Néanmoins, afin d'avoir une vision globale de la proportion d'extrémité 5'P des 1 489 TSS pour lesquels nous avons pu calculer une proportion d'extrémités 5'P chez la souche sauvage de *S. aureus,* nous avons regroupé tous les échantillons sauvages en prenant par position la médiane des proportions observées. La distribution des proportions obtenues à partir de tous les échantillons est illustrée dans la Figure 125.



Percentile	0	25 %	50 %	75 %	100 %
Proportion	0,00	0,11	0,21	0,34	1,22

Figure 125 : Distribution des proportions d'extrémités 5'P par percentile dans les souches sauvages.

Dans cette figure et dans le tableau (avec les percentiles), nous pouvons observer pour le premier quartile que 25 % des TSS ont une proportion de forme 5'P inférieure à 0,11. Il sera *a priori* difficile de détecter pour ces TSS une baisse significative de ces proportions dans un mutant NUDIX, notamment d'après les difficultés rencontrées lors des étapes précédentes de désaturation et de

normalisation. Nous pouvons, en revanche, observer que la moitié des TSS a une proportion de forme 5'P d'au moins 0,21, et pour le dernier quartile d'au moins 0,34. Il semblerait que, pour ces TSS, il soit réalisable de détecter une baisse significative statistiquement, au moins pour certains TSS. Dans le cas où le mutant NUDIX considéré correspondrait à la RppH de *S. aureus*, nous devrions néanmoins observer une chute globale des proportions de forme 5'P dans la souche mutante pour la plupart des TSS.

2.2.1.1.5.2 Comparaison par librairie des proportions d'extrémités 5'P entre souche sauvage et mutante

Afin d'étudier l'effet des mutations des candidats RppH sur les états de phosphorylation des extrémités 5' des ARN, nous allons comparer les proportions d'extrémité 5'P par librairie à l'aide de tests de Wilcoxon-Mann-Whitney. Nous commencerons ensuite par une analyse globale des résultats avant d'examiner plus en détails les conséquences des différentes mutations sur l'état de phosphorylation des extrémités 5' par rapport à la souche sauvage, en analysant chaque TSS individuellement. Cette seconde analyse a pour objectif d'identifier les ARN substrats de la ou des protéines NUDIX.

2.2.1.1.5.2.1 Librairie PQ-EMOTE 49

La Figure 126 représente la distribution des proportions d'extrémité 5'P pour les TSS dans la librairie PQ-EMOTE 49. Nous observons que le nombre de TSS pour lesquels nous avons pu calculer une proportion d'extrémité 5'P est très faible comparé aux 3 575 TSS identifiés dans l'ensemble du génome. En effet, nous disposons de proportions pour seulement 191 et 142 TSS pour les réplicats de la souche sauvage. A partir de ces deux réplicats, nous avons pu calculer une proportion pour 213 TSS (6 %) en prenant pour un TSS la médiane des proportions observées dans chacun des réplicats ou bien la proportion dans le réplicat pour lequel nous disposions d'une proportion si la proportion n'était pas disponible dans l'autre. De même, nous disposons seulement de 89 TSS pour la souche Δ1612 (2,5 %), et seulement 210 TSS (5,9 %) pour le triple mutant. Ceci suggère que les résultats seront peu représentatifs des états de phosphorylation des TSS dans les souches étudiées dans cette librairie. De plus, nous observons que les réplicats issus de la souche $J1^{AGA}$ sont peu reproductibles, et il sera *a priori* difficile d'interpréter les valeurs de ces échantillons.

Toutefois, les résultats obtenus sur les deux réplicats issus de la souche Δ1612 apparaissent reproductibles de même que ceux issus de la souche triple mutant Δ1612_Δ0425_Δ2278. Nous avons donc testé si les distributions de proportion d'extrémités 5'P étaient différentes entre les souches sauvages et la souche Δ1612 ou la souche triple mutant avec un test de Wilcoxon-Mann-Whitney. La proportion d'extrémités 5'P est plus élevée (p-valeur ajustée = $2{,}57.10^{-13}$) dans le mutant Δ1612 (médiane = 0,32 ; moyenne = 0,33) par rapport au sauvage (médiane = 0,17 ; moyenne = 0,18). La proportion d'extrémités 5'P est légèrement plus faible (p-valeur ajustée = $8{,}76.10^{-8}$) dans le triple mutant (médiane = 0,11 ; moyenne = 0,14) par rapport au sauvage (médiane = 0,17 ; moyenne = 0,18).

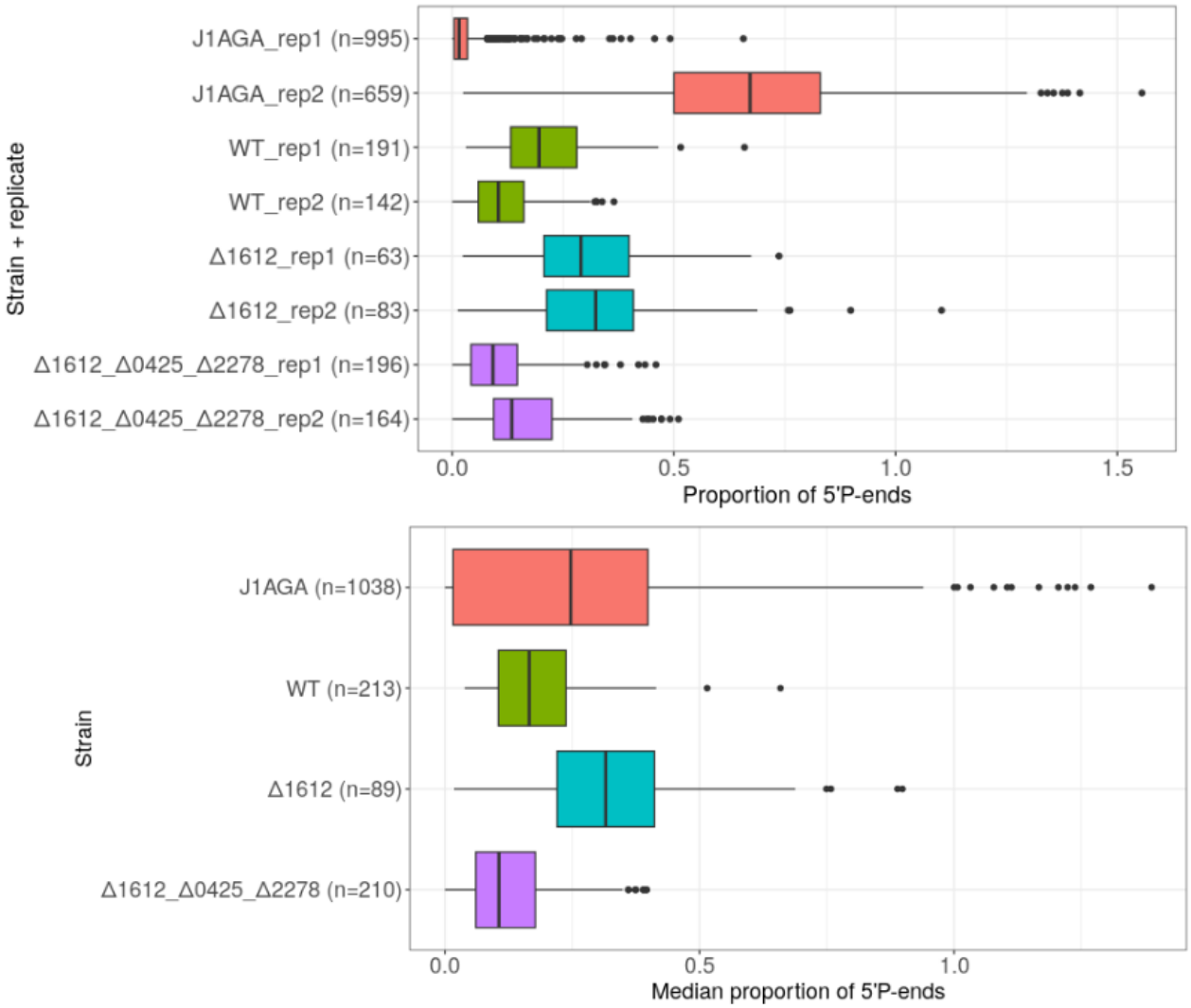


***Figure 126 :* Proportion d'extrémités 5'P dans la librairie PQ-EMOTE 49.**

*Boîtes à moustaches représentant les distributions des proportions des extrémités 5'P dans les échantillons issus de la librairies PQ-EMOTE 49. (**A**) les données regroupées par réplicat biologique. (**B**) les données sont regroupées par souche (médiane des proportions à la même position dans les différents réplicats).*

Ces résultats ainsi que ceux des autres librairies seront discutés et mis en relation à la fin de la section.

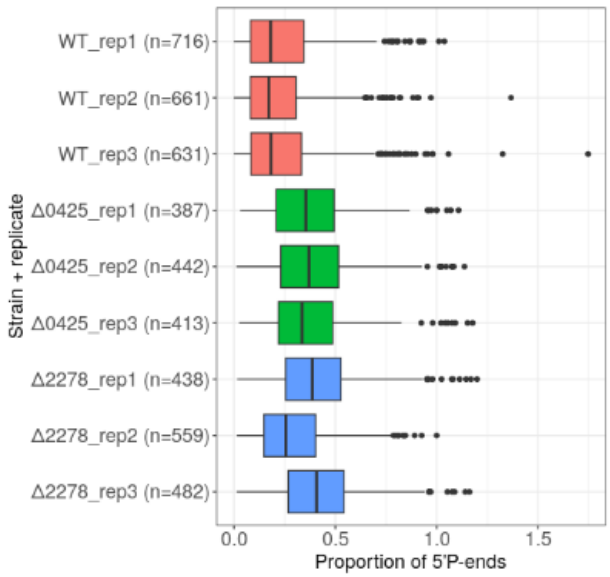
2.2.1.1.5.2.2 Librairie PQ-EMOTE 56

La Figure 127 représente la distribution des proportions des extrémités 5'P pour les TSS identifiés dans la librairie PQ-EMOTE 56. Comme dans la partie précédente, la reproductibilité des expériences a tout d'abord été étudiée au sein de la librairie pour les différents mutants en s'intéressant à la distribution des résultats au sein de chaque réplicat (Fig. 127A). Comparé à la librairie PQ-EMOTE 49, nous avons pu calculer une proportion d'extrémité 5'P pour davantage de TSS : 752 (21 %) dans la souche sauvage, 469 (13,1 %) dans la souche Δ0425 et 604 (16,8 %) dans la souche Δ2278. Cette représentativité est meilleure que celle de la librairies PQ-EMOTE 49, cependant, nous sommes encore loin de 3 575 TSS.

Nous observons également que les valeurs de médianes sont proches pour les trois réplicats du mutant Δ0425 et de la souche sauvage. Dans le cas du mutant Δ2278, le second réplicat présente une valeur de médiane un peu plus faible que les deux autres. La reproductibilité des données pour chacune des souches est donc bien meilleure que dans le cas de la librairie PQ-EMOTE 49 et nous allons donc pouvoir comparer les états de phosphorylation des extrémités 5' des TSS identifiés entre chaque mutant et la souche sauvage en utilisant la distribution des proportions sur l'ensemble des réplicats.

Nous observons cette fois, une augmentation statistiquement significative de la proportion des extrémités 5'P dans la souche mutante Δ0425 (médiane = 0,35 ; moyenne = 0,38 et p-valeur ajustée de $2{,}1.10^{-40}$) et la souche Δ2278 (médiane = 0,32 ; moyenne = 0,37 et p-valeur ajustée = $1{,}5.10^{-35}$) par rapport à la souche sauvage (médiane = 0,17 et moyenne = 0,23) (Fig. 127B). Ces résultats vont à l'encontre de l'hypothèse de départ qui supposait qu'en délétant une RppH, la proportion d'extrémités 5'P devrait diminuer.

(A)



(B)

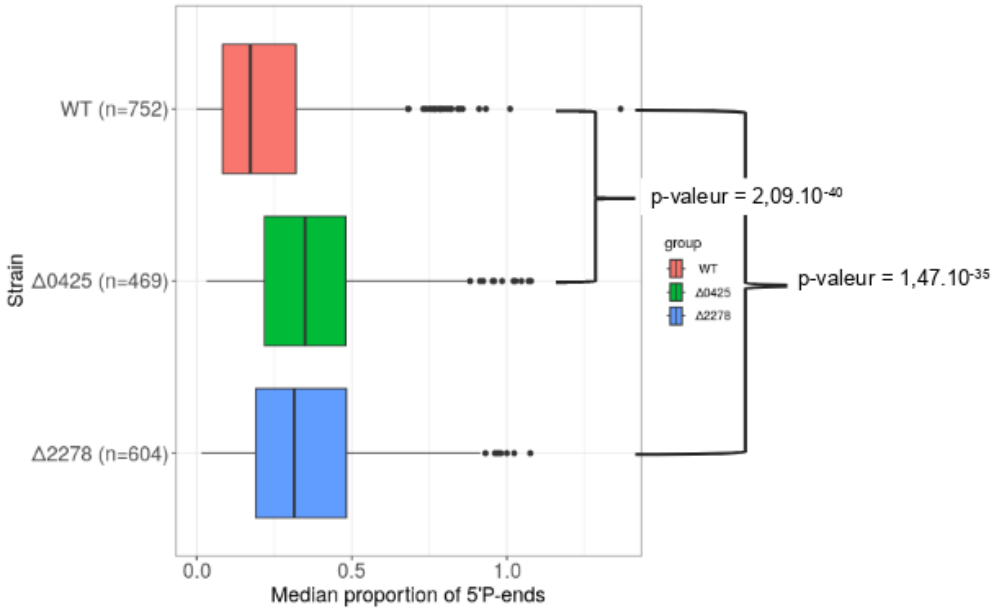


Figure 127 : ***Proportion d'extrémités 5'P dans la librairie PQ-EMOTE 56.***

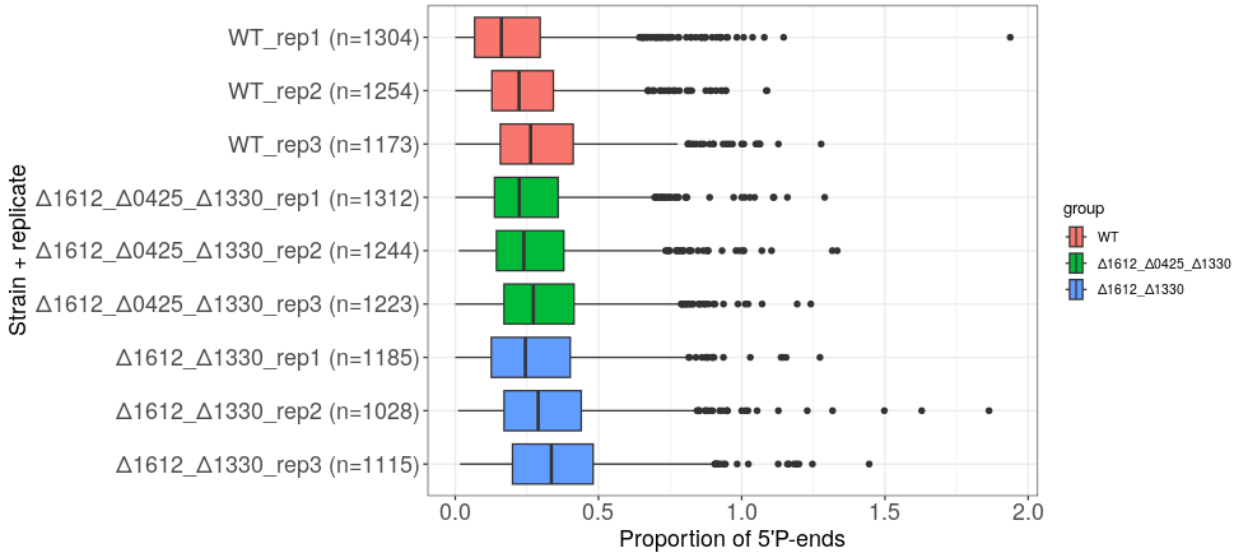
*Boîtes à moustaches représentant les distributions des proportions des extrémités 5'P dans les échantillons issus de la librairies PQ-EMOTE 56. (**A**) les données regroupées par réplicat biologique. (**B**) les données sont regroupées par souche (médiane des proportions à la même position dans les différents réplicats). Les accolades illustrent les différences* statistiquement *significatives.*

2.2.1.1.5.2.3 Librairie PQ-EMOTE 58

Nous avons de nouveau, dans un premier temps, vérifié la reproductibilité des données au sein de la librairie pour les différents mutants en s'intéressant à la distribution des proportions au sein de chaque réplicat (Fig. 128A). Nous observons une bien meilleure représentativité de l'ensemble des TSS puisque nous avons pu calculer une proportion d'extrémités 5'P pour 1444 (40,4 %) d'entre eux dans la souche sauvage, 1359 (38 %) dans la souche Δ1612-Δ1330 et 1225 (34,3 %) dans la souche Δ1612-Δ0425-Δ1330.

Nous observons, comme dans le cas de la librairie PQ-EMOTE 56, des valeurs de médianes proches pour les trois réplicats de la souche sauvage et des souches double et triple mutantes. La comparaison des états de phosphorylation des extrémités 5' des TSS entre chaque mutant et la souche sauvage a donc pu être réalisée en utilisant les distributions de proportions sur l'ensemble des réplicats. Nous observons une légère augmentation statistiquement significative (p-valeur ajustée $5{,}7.10^{-21}$) des proportions des extrémités 5'P dans le cas de la comparaison entre le double mutant Δ1612-Δ1330 (médiane = 0,28 ; moyenne = 0,33) et la souche sauvage (médiane = 0,20 ; moyenne = 0,26). Des résultats similaires sont également observés dans le cadre de la comparaison entre le triple mutant Δ1612-Δ0425-Δ1330 (médiane = 0,24 ; moyenne = 0,28) et la souche sauvage, c'est-à-dire qu'il semblerait qu'il y ait une légère augmentation statistiquement significative de la proportion d'extrémités 5'P (p-valeur ajustée $1{,}6.10^{-8}$).

(A)



(B)

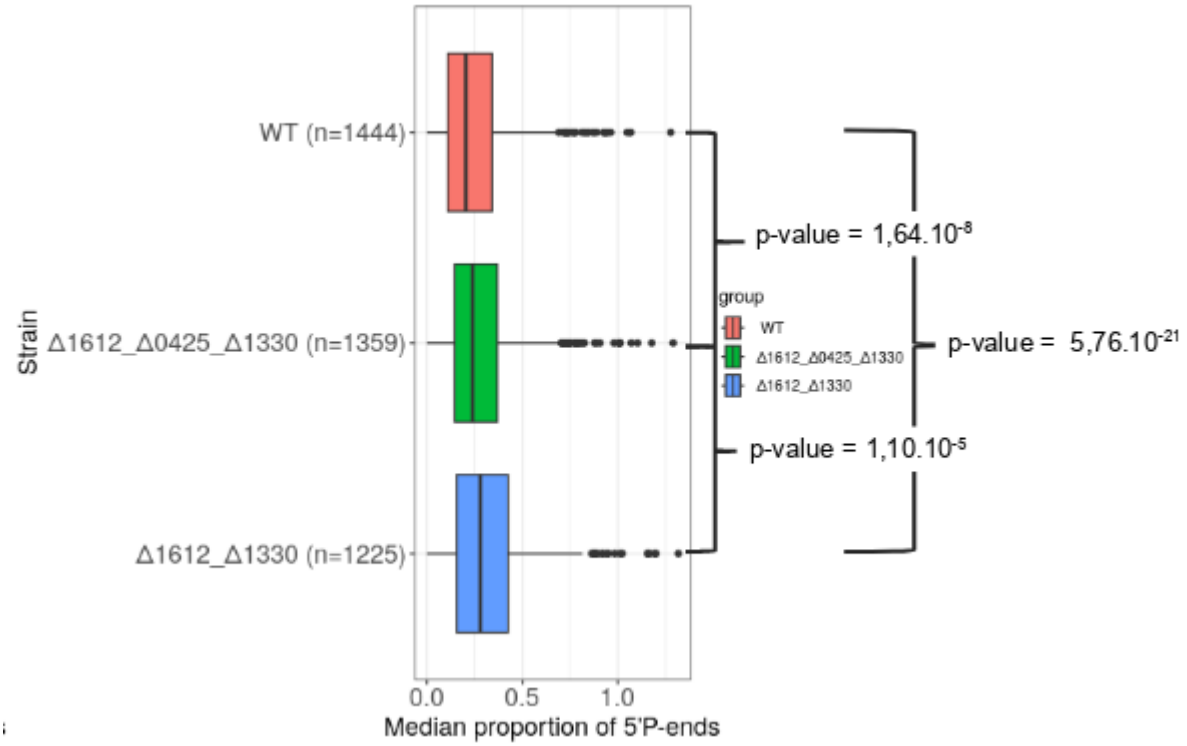


Figure 128 : ***Proportion d'extrémités 5'P dans la librairie PQ-EMOTE 58.***

*Boîtes à moustaches représentant les distributions des proportions des extrémités 5'P dans les échantillons issus de la librairies PQ-EMOTE 58. (**A**) les données regroupées par réplicat biologique. (**B**) les données sont regroupées par souche (médiane des proportions à la même position dans les différents réplicats). Les accolades illustrent les différences* statistiquement *significatives.*

2.2.1.1.5.2.4 Librairie PQ-EMOTE 62

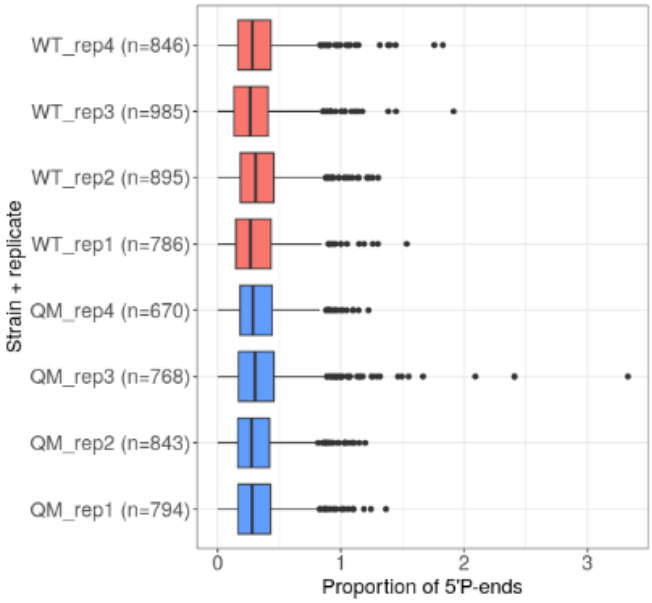
Dans le cas de la librairie PQ-EMOTE 62, nous disposons de quatre réplicats pour l'étude de chacune des souches, sauvage et quadruple mutant (Δ1612-Δ1330-Δ0425-Δ2778). Cependant, le manque de profondeur de séquençage dû à la sur-représentation de l'ARNr 5S a réduit le nombre de TSS pour lesquels nous pouvons calculer la proportion d'extrémités 5'P à seulement 851 (23,8 %) sites de départ de transcription (TSS) dans la souche sauvage et 742 (20,8 %) dans la souche quadruple mutant. Bien que ces chiffres soient supérieurs à ceux obtenus pour les librairies PQ-EMOTE 49 et 56, ils restent inférieurs à ceux de la librairie PQ-EMOTE 58. Cela est regrettable car PQ-EMOTE 62 est la librairie avec le plus grand nombre de protéines NUDIX inactivées, le plus grand nombre de réplicats et les UMI les plus longs, offrant ainsi *a priori* la quantification la plus précise et fiable parmi les quatre librairies.

Les données apparaissent reproductibles pour chacune des souches comme le montre la Figure 129A. Cependant, la comparaison des proportions des extrémités 5'P pour les TSS identifiés ne révèle pas de différence statistiquement significative (p-valeur ajustée = 0,76) entre la souche sauvage (médiane = 0,30 ; moyenne = 0,34) et la souche délétée des quatre gènes NUDIX (médiane = 0,29 ; moyenne = 0,35) sur les cinq identifiés dans le génome.

Ces résultats sur les tendances globales des proportions de formes mono-phosphorylées des ARN sont contraires aux attentes initiales. En effet, nous observons que les proportions globales de 5'P sont similaires, voire plus élevées pour les souches mutantes que pour la souche sauvage pour la librairie PQ-EMOTE 56. Cela contredit l'hypothèse selon laquelle la suppression d'un gène codant pour une RppH devrait normalement réduire la proportion d'extrémités 5'P des ARN ciblés par cette RppH. Il est donc possible que la ou les RppH de *S. aureus* ne soient pas incluses parmi les 4 mutants testés. Il est également envisageable que la combinaison de mutants pouvant inhiber l'activité de déphosphorylation des extrémités 5' ne fasse pas partie de celles construites.

Cependant, malgré ces observations, nous constatons néanmoins des effets sur les états de phosphorylation des extrémités 5' des ARN, comme par exemple une augmentation globale de la proportion d'extrémités 5'P dans les souches simples mutantes Δ0425 et Δ2278. Les enzymes NUDIX sont d'une grande importance pour les cellules et jouent un rôle crucial dans la régulation de l'expression des gènes. Ainsi, l'inactivation de ces enzymes n'est pas anodine puisque les NUDIX peuvent avoir plusieurs fonctions et intervenir dans différents processus biologiques. Il est donc plausible que d'autres mécanismes compensatoires, aux efficacités variables, interviennent pour pallier leur perte.

(A)



(B)

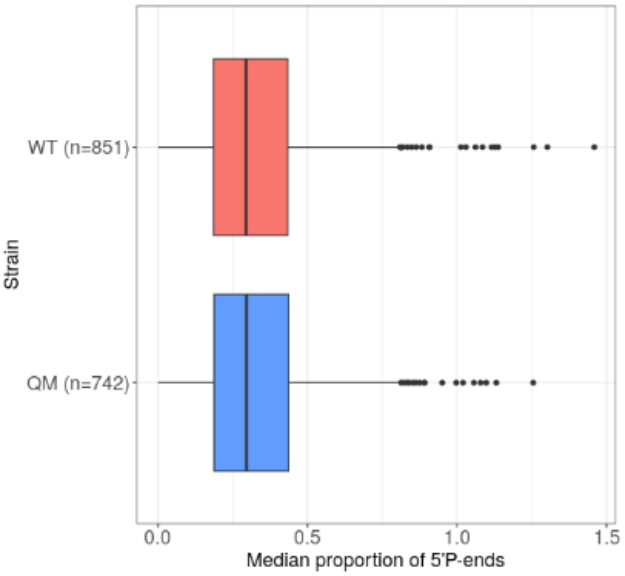


Figure 129 : ***Proportion d'extrémités 5'P dans la librairie PQ-EMOTE 62.***

*Boîtes à moustaches représentant les distributions des proportions des extrémités 5'P dans les échantillons issus de la librairies PQ-EMOTE 62. (**A**) les données regroupées par réplicat biologique. (**B**) les données sont regroupées par souche (médiane des proportions à la même position dans les différents réplicats).*

Bien que cette augmentation de la proportion d'extrémité 5'P soit très surprenante, toute variation à une position donnée (augmentation ou diminution) peut être interprétée comme une cible potentielle des NUDIX. En cas de diminution de la proportion d'extrémités 5'P, les transcrits ciblés par la NUDIX inactivée ne sont plus déphosphorylés, ce qui réduit la proportion d'extrémités 5'P observée. En revanche, en cas d'augmentation, l'inactivation de la NUDIX ne déphosphoryle plus le transcrit, mais cette fonction pourrait être compensée par un autre mécanisme ou par une autre NUDIX possédant une efficacité supérieure, augmentant ainsi la proportion d'extrémités 5'P des transcrits. Le principal défi réside dans la détection des transcrits cibles des NUDIX dont la perte de la déphosphorylation de leur extrémité 5' pourrait être compensée par d'autres mécanismes ayant une efficacité similaire à celle de la NUDIX perdue.

Toutefois, il est important de garder à l'esprit que les ensembles des TSS comparés représentent à chaque fois moins de la moitié des 3575 TSS identifiés chez *S. aureus* ou des 3 145 TSS associés aux gènes considérés comme exprimés (dans le chapitre précédent). En conséquence, il manque au minimum, la moitié de l'information totale, ce qui limite considérablement notre capacité à dresser un profil complet des états de phosphorylation des extrémités 5'P des ARN. Une augmentation de la profondeur de séquençage serait donc nécessaire pour obtenir une meilleure représentativité et permettre une analyse plus exhaustive et précise.

De plus, la grande variation dans l'estimation des proportions d'extrémités 5'P peut masquer les effets des mutations. Cette variabilité élevée complique la détection de différences significatives et pourrait potentiellement masquer les véritables impacts des mutations RppH.

Enfin, il se pourrait également que la régulation des ARN se fasse par la variation de l'abondance des extrémités 5'P plutôt que par leur proportion. En d'autres termes, lorsque l'activité de RppH est inhibée, les transcrits sont moins dégradés, ce qui pourrait entraîner une diminution de la transcription de ces gènes. Par conséquent, même si la quantité totale d'ARN change, la proportion des extrémités 5'P reste constante. Cela signifie que la mesure de la proportion des extrémités 5'P pourrait ne pas refléter fidèlement les changements dans l'activité de RppH. Cependant, il est attendu qu'en cas de mutation de l'ensemble des RppH, les extrémités 5'P devraient être totalement absentes.

2.2.1.1.5.3 Identification des substrats ARN potentiels de la RppH

Afin de déterminer les extrémités 5’ dont la proportion d’extrémités 5’P est impactée par les mutations des candidats RppH, nous avons procédé à des analyses de comparaison de proportion. Le but ici est d’identifier les ARN individuels pour lesquels la proportion d’extrémité 5’ P sera diminuée dans la souche mutée dans un ou plusieurs gènes NUDIX par rapport à cette proportion dans la souche sauvage. Cela nécessite de comparer les rapports 5'P / (5'P + 5'PPP) entre deux souches pour chaque TSS comme illustré dans un exemple en Figure 130.

Les expériences RNA-Seq ayant été réalisées avec des réplicats, cela nous permet d'estimer la variabilité biologique et technique contenue dans les données de séquençage, assurant ainsi la fiabilité des résultats et permettant des analyses statistiques appropriées tout en minimisant les faux positifs. Comparer les moyennes des proportions ne permet pas de pondérer les valeurs en fonction des abondances du $pool^{total}$ utilisée pour le calcul des proportions. Comparer les proportions des moyennes des comptages par réplicats ne permet pas de prendre en compte la dispersion entre réplicats (locale). Il est donc crucial d'utiliser une méthode qui prend en compte ces réplicats.

		Wild type strain						NUDIX mutant strain					
		Replicate 1		Replicate 2		Replicate 3		Replicate 1		Replicate 2		Replicate 3	
		$Pool^{mono}$	$Pool^{total}$	$Pool^{mono}$	$Pool^{total}$	$Pool^{mono}$	$Pool^{total}$	$Pool^{mono}$	$Pool^{total}$	$Pool^{mono}$	$Pool^{total}$	$Pool^{mono}$	$Pool^{total}$
TSS A	Count	50	100	50	100	70	120	3	50	2	70	0	20
	5’P prop	**0.5**		**0.5**		**0.58**		**0.06**		**0.03**		**0**	
TSS B	Count	1	100	2	200	1	20	0	50	35	75	100	100
	5’P prop	**0.01**		**0.01**		**0.05**		**0**		**0.47**		**1**	

Global dispersion
Local dispersion

Figure 130 : Exemple de comparaison de comparaison de proportion d’extrémité 5’P.

Les dispersion globale et local sont représenté respectivement par des rectangles bleu et vert.

Différentes méthodes ont été proposées pour réaliser les analyses différentielles de données RNA-Seq. Le choix de l’approche va dépendre de la nature des données comme par le nombre de réplicats et l’adéquation à la loi normale des échantillons. Le choix va également dépendre des attentes quant au fonctionnement de la méthode comme par exemple la prise en compte des dispersions locale et/ou globale.

Les méthodes basées sur la loi normale et le test exact de Fisher ne répondent pas à ce critère. Ces méthodes impliquent généralement l'agrégation des comptages de lectures des différents réplicats de chaque souche et cette approche fait l'hypothèse que les données proviennent d'une distribution identique et ignore donc la dispersion des réplicats biologiques. De plus, pour un faible nombre de répétitions des mesures, la normalité de l'échantillon n'est souvent pas vérifiée et une plus grande dispersion des données est souvent observée.

C'est pourquoi les modèles linéaires généralisés (GLM) sont souvent utilisés pour comparer des valeurs, y compris des proportions car ils permettent d'être plus robuste en prenant en compte la dispersion entre les réplicats (locale). Cependant, ils ne permettent pas d'accorder plus ou moins de confiance à une proportion en fonction des comptages du pool^{total} utilisé pour la calculer. Pourtant, il est crucial de pouvoir accorder plus de confiance à une proportion lorsque les abondances utilisées pour son calcul sont élevées, car elle devient moins sensible aux variations.

En général, les analyses différentielles à partir de données RNA-Seq sont effectuées avec des programmes comme DESeq2 ou EdgeR dans R. Cependant, ces méthodes sont conçues pour comparer des comptages et non des proportions. Elles supposent que les comptages suivent une distribution binomiale négative, ce qui n'est pas le cas pour les proportions que nous souhaitons comparer ici.

Le test de Wald (Formule 7) est couramment utilisé pour comparer des proportions tout en tenant compte de la variance entre réplicats. Cependant, il est considéré comme un test quelque peu simpliste car il ne prend pas en compte la dispersion globale des données ni les niveaux d'abondance. De plus, le test de Wald repose sur l'approximation normale de la distribution des estimations de proportion. Lorsque les proportions sont proches de 0 ou de 1, la distribution réelle peut être très différente de la normale, rendant l'approximation incorrecte.

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\mathrm{var}(\hat{\theta})}$$

***Formule 7** : Formule pour le calcul de la statistique de Wald où θ_0 correspond à la proportion théorique et $\hat{\theta}$ à la proportion estimée.*

DSS est un package R développé par (Wu et al., 2013) et (Feng et al., 2014), conçu pour détecter les loci différentiellement méthylés à partir de données de séquençage à résolution d'un seul nucléotide. Leur approche repose sur l'utilisation d'un test de Wald pour la comparaison statistique des proportions (nombre de sites méthylés divisé par le nombre total de sites dans une région). Ce test est enrichi par l'estimation d'un paramètre de dispersion à travers un modèle hiérarchique bayésien, permettant ainsi de d'estimer à la fois la variabilité biologique entre réplicats pour une région (dispersion locale) et la variabilité sur l'ensemble des positions ou régions au sein d'un même échantillon (dispersion globale).

Cette méthode est décrite par les auteurs comme très efficace dans les situations où le nombre de réplicats biologiques est limité, comme c'est souvent le cas dans les analyses de type RNA-Seq, comme l'ont souligné (Park & Wu, 2016). De plus, elle intègre les niveaux d'abondance utilisés pour calculer les proportions dans l'évaluation des p-valeurs, ce qui renforce sa robustesse statistique.

Enfin, il est important de noter que DSS ne permet pas de traiter les positions où une proportion supérieure à 1 est calculée. Cependant, dans les données, la variabilité technique peut parfois conduire à des proportions dépassant cette limite. Par conséquent, nous avons dû ajuster ces valeurs en les plafonnant à 1, en ramenant les comptages de l'échantillon $pool^{mono}$ au même niveau que ceux dans $pool^{total}$, avant d'utiliser DSS. Cet ajustement peut entraîner une réduction de la variabilité observée dans les données, potentiellement dissimulant des différences significatives. Il est donc important de reconnaître que cela peut entraîner une perte partielle d'information.

2.2.1.1.5.3.1 Librairie PQ-EMOTE 49

La Figure 131 illustre les comparaisons entre les proportions d'extrémités 5'P (PQ-EMOTE 49) issues de la souche sauvage et :

- celles issues de la souche Δ1612 (Fig. 131A),
- celles issues de la souche Δ1612_ Δ0425_ Δ2278 (Fig. 131B),
- celles issues de la souche $J1^{AGA}$ (Fig. 131C).

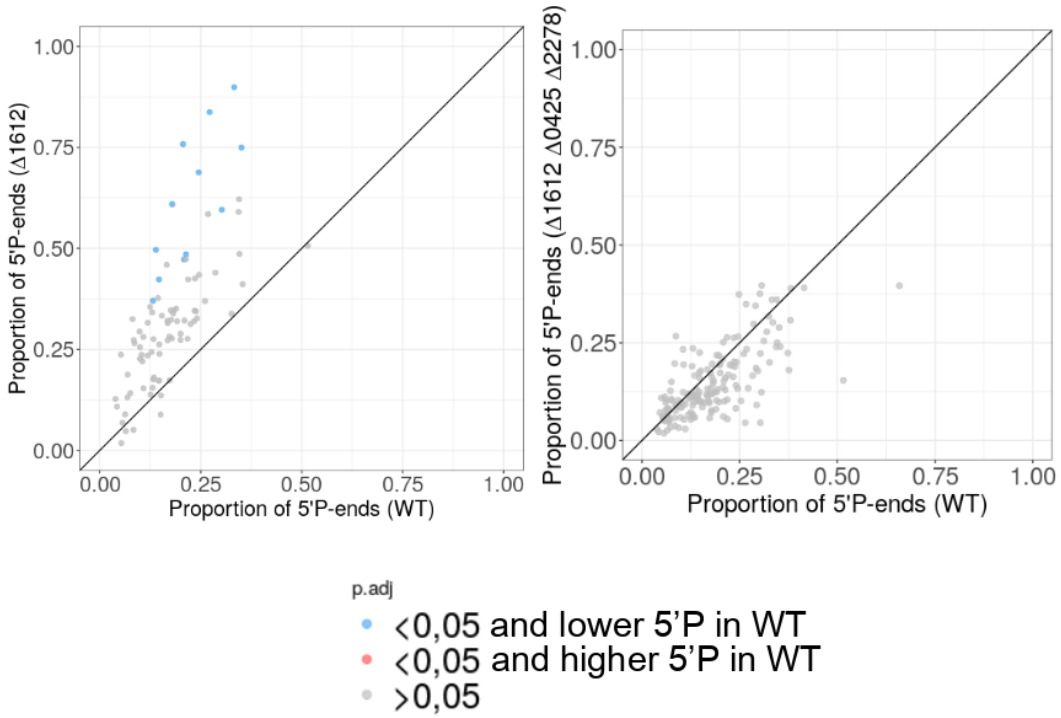
Les TSS dont la proportion d'extrémités 5'P augmente statistiquement, c'est-à-dire avec une p-valeur ajustée (BH) inférieure à 0,05, dans la souche mutante par rapport à la souche sauvage sont représentés en bleu et ceux dont elle diminue statistiquement (p-valeur ajustée inférieure à 0,05) sont représentés en rouge.

Les résultats sont obtenus sur un faible nombre de TSS que nous sommes en mesure de comparer : 86 TSS pour la comparaison avec ∆1612, 167 TSS avec Δ1612_ Δ0425_ Δ2278, et 209 TSS avec $J1^{AGA}$. De plus, nous ne disposons que de deux répétitions biologiques par souche dans lesquelles on a parfois des écarts importants entre les valeurs observées pour un même TSS dans les mêmes conditions expérimentales, ce qui limite le nombre d'extrémités 5' montrant une différence de proportion significative même si la différence est grande.

Le mutant ∆1612 présente 12 TSS où la proportion d'extrémité 5'P augmente de manière significative (points bleus) par rapport à la souche sauvage. Aucun TSS n'a de proportion d'extrémité 5'P qui diminue de manière statistiquement significative. Comme cela a déjà été observé dans l'analyse globale précédente sur l'ensemble des positions (boites à moustache) l'a indiquée, ces résultats ne sont pas conformes aux attentes, car le candidat SA1612 est le favori pour être une RppH. Cependant, compte tenu du nombre limité de TSS étudiés, il n'est pas surprenant de ne pas observer les résultats attendus.

Dans le cas du triple mutant NUDIX et du mutant $J1^{AGA}$, aucune différence statistiquement significative n'est observée. Pour ce dernier, on se serait attendu à ce que, la RNase J1 n'étant plus fonctionnelle, les extrémités 5'P soient moins dégradées que dans une souche sauvage, et donc que la proportion d'extrémités 5'P augmente dans la souche $J1^{AGA}$. Les résultats observés tendent effectivement à montrer une telle augmentation dans la souche $J1^{AGA}$. Toutefois, le manque de réplicats et la grande dispersion des valeurs rendent ces différences observées non statistiques.

(A) **(B)**



(C)

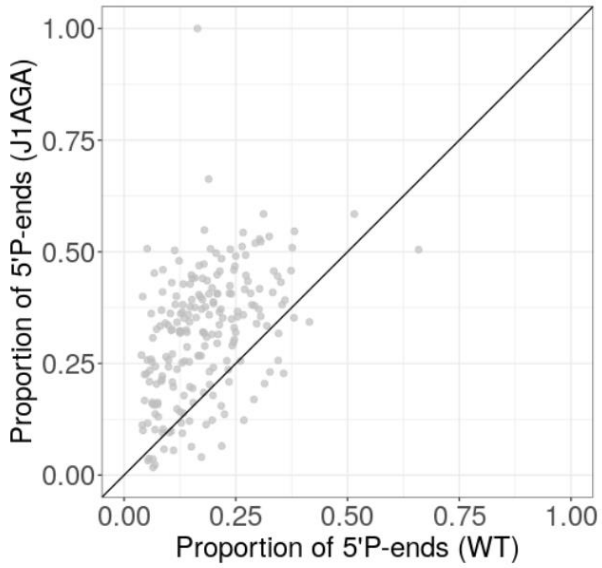


Figure 131 : ***Analyse différentielle de proportion d'extrémité 5'P.***

*Proportions d'extrémités 5'P des TSS dans une souche sauvage (axe x) et dans une souche mutante (axe y) (**A**) Δ1612, (**B**) Δ1612_Δ0425_Δ2278 et (**C**) $J1^{AGA}$. Chaque point représente une extrémité 5' correspondant à un TSS et les proportions affichées correspondent aux moyennes pondérées des réplicats retournées par DSS. Code couleur : point gris quand la différence n'est pas significative ; point bleu quand une augmentation statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage ; point rouge quand une diminution statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage*

Ces résultats ainsi que ceux des autres librairies seront discutés et mis en relation à la fin de la section.

2.2.1.1.5.3.2 Librairie PQ-EMOTE 56

Avec la librairie PQ-EMOTE 56, nous avons pu effectuer les comparaisons de proportions d'extrémités 5'P pour 460 TSS dans la souche Δ0425 et 588 TSS dans la souche Δ2278. Les résultats obtenus, illustrés en Figure 132 montre que 180 (mutant Δ0425, Fig. 132A) et 142 (mutant Δ2278, Fig. 132B) TSS, dont 106 en commun, ont une proportion d'extrémités 5'P qui augmente de manière statistiquement significative par rapport à la souche sauvage. Seulement 7 TSS dans le mutant Δ0425 et aucun dans le mutant Δ2278, montrent une diminution statistiquement significative de la proportion d'extrémités 5'P par rapport au sauvage. Ces résultats obtenus pour chaque TSS sont plus détaillés que les comparaisons globales effectuées précédemment (boîtes à moustaches sur les proportions de chaque échantillon et comparaisons de moyennes), et viennent soutenir les observations précédentes, c'est-à-dire une tendance globale à l'augmentation des proportions d'extrémités 5'P dans les souches mutantes par rapport à la souche sauvage.

Ces résultats suggèrent que dans les deux souches mutantes étudié dans cette librairie, les mécanismes de déphosphorylation sont amplifiés ce qui n'est pas ce qui était attendu dans le cas où une RppH serait inactivée. Cependant, les mutants étudiés étant des simples mutants, nous pouvons émettre l'hypothèse qu'une des quatre autres protéines NUDIX pourrait compenser l'absence de l'enzyme testée. L'augmentation de la proportion d'extrémités 5'P dans le cas de certains ARN, pourrait indiquer que le système permettant de réguler les états de phosphorylation des extrémités 5' des ARN est perturbé par la délétion de ces NUDIX. En effet, les NUDIX ayant des activités multiples et parfois redondantes, il se pourrait que des processus biologiques soit dérégulés, entraînant l'augmentation de proportion d'extrémité 5'P.

Ces résultats suggèrent que, dans les deux souches mutantes étudiées dans cette librairie, les mécanismes de déphosphorylation sont amplifiés, ce qui n'est pas attendu en cas d'inactivation de RppH. Cependant, étant donné que les mutants étudiés sont des mutants simples, il est plausible qu'une des quatre autres protéines NUDIX puisse compenser l'absence des enzymes testées. De plus, les NUDIX possédant des activités multiples, il se pourrait que la régulation des états de phosphorylation des extrémités 5' des ARN est perturbée par la délétion de ces NUDIX sans pour autant être des RppH.

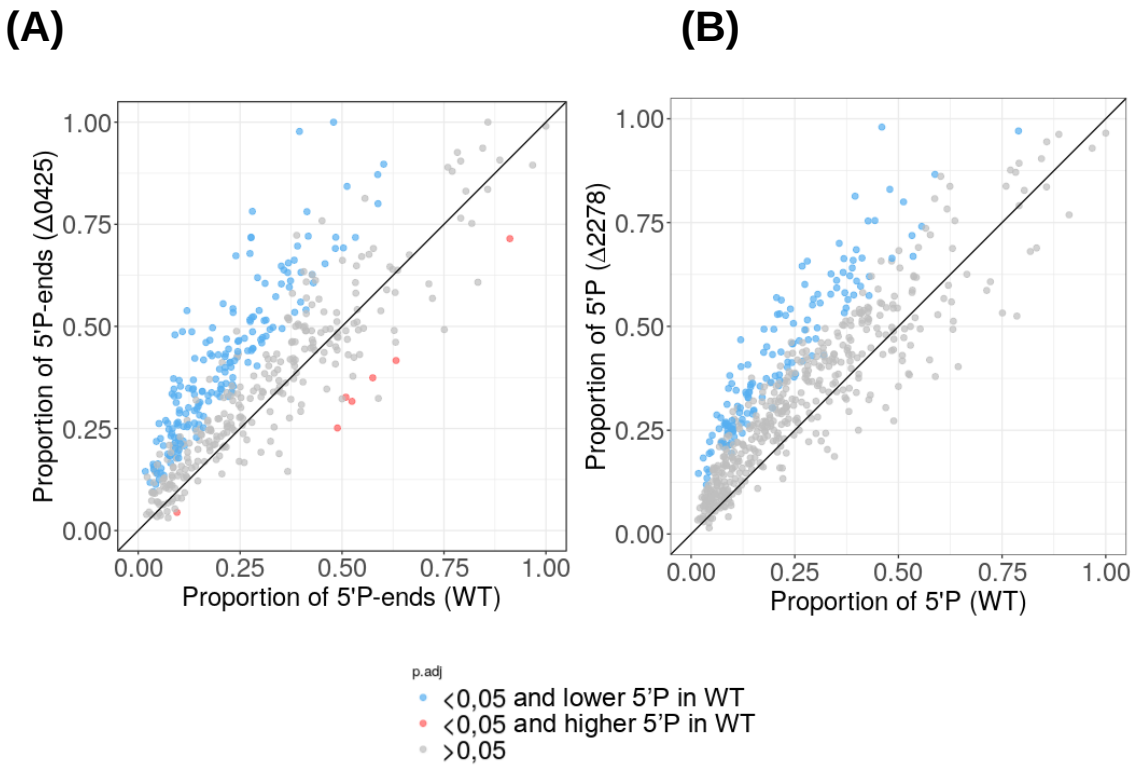


Figure 132 : ***Analyse différentielle de proportion d'extrémité 5'P.***

*Proportions d'extrémités 5'P des TSS dans une souche sauvage (axe x) et dans une souche mutante (axe y) (**A**) Δ0425 et (**B**) Δ2278. Chaque point représente une extrémité 5' correspondant à un TSS et les proportions affichées correspondent aux moyennes pondérées des réplicats retournées par DSS. Code couleur : point gris quand la différence n'est pas statistiquement significative ; point bleu quand une augmentation statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage ; point rouge quand une diminution statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage*

2.2.1.1.5.3.3 Librairie PQ-EMOTE 58

Avec la librairie PQ-EMOTE 58, nous avons pu effectuer les comparaisons de proportions pour 1 205 TSS dans le mutant Δ1612_Δ1330, et 1 310 TSS dans le mutant Δ1612_Δ0425_Δ1330. Ces comparaisons révèlent 1 TSS dont la proportion d'extrémité 5'P diminue significativement dans le double mutant Δ1612-Δ1330, et 1 TSS dans le triple mutant Δ1612-Δ1330-Δ0425 par rapport à la souche sauvage, ainsi que 9 TSS et 2 TSS dont la proportion augmente. Pour ces mutants, les résultats obtenus pour chaque TSS viennent également soutenir les observations antérieures faites lors de la comparaison globale des distributions de proportions dans les échantillons.

Néanmoins, nous pouvons remarquer dans les deux souches mutantes, un nombre important de TSS (points gris non significatifs) sont localisés au-dessus de la diagonale, indiquant que, même si l'augmentation n'est pas statistiquement significative, les proportions des extrémités 5'P sont plus élevées dans le mutant par rapport au sauvage pour un grand nombre d'ARN. Ceci pourrait expliquer la conclusion obtenue lors de la comparaison des distributions où une légère augmentation statistiquement significative des proportions des extrémités 5'P a été observée.

Le fait que l'on retrouve un profil de proportions de 5'P presque similaire à celui de la souche sauvage dans ces multiples mutants pourrait indiquer que les NUDIX délétées (SA0425, SA1330 et SA1612) n'ont pas l'activité RppH recherchée.

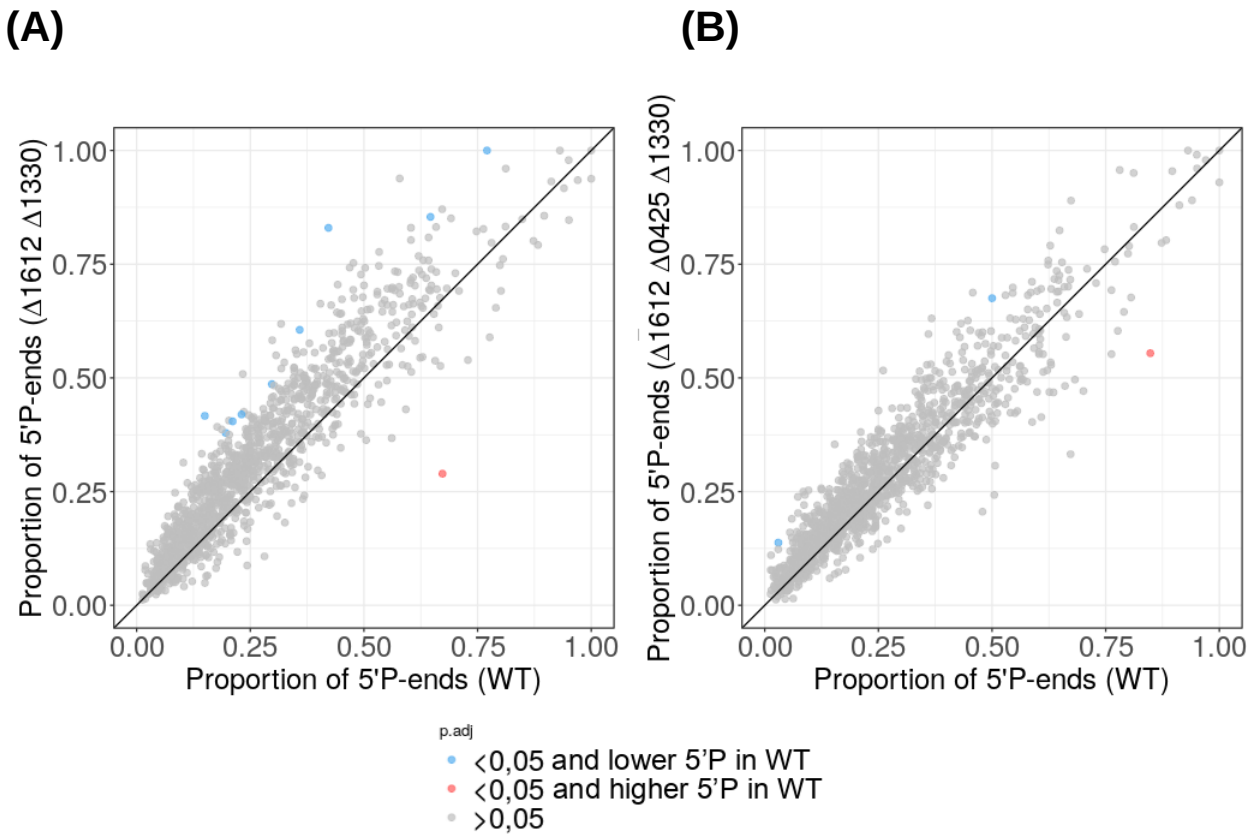


Figure 133 : Analyse différentielle de proportion d'extrémité 5'P.

*Proportions d'extrémités 5'P des TSS dans une souche sauvage (axe x) et dans une souche mutante (axe y) (**A**) Δ1612_Δ1330 et (**B**) Δ1612_Δ0425_Δ1330. Chaque point représente une extrémité 5' correspondant à un TSS et les proportions affichées correspondent aux moyennes pondérées des réplicats retournées par DSS. Code couleur : point gris quand la différence n'est pas statistiquement significative ; point bleu quand une augmentation statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage ; point rouge quand une diminution statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage*

2.2.1.1.5.3.4 Librairie PQ-EMOTE 62

Dans le cas de la librairie PQ-EMOTE 62 où quatre des cinq candidats RppH ont été délétés, les résultats des comparaisons effectuées sur 720 TSS révèlent un profil des états de phosphorylation des extrémités 5' des TSS dans la souche quadruple mutant très similaire à celui de la souche sauvage. Seulement 1 TSS a une diminution statistiquement significative de la proportion d'extrémité 5'P, et aucun ne présente une augmentation statistiquement significative. Cela rejoint les observations faites lors de la comparaison globale des distributions de proportions qui n'avait pas révélé de différence statistiquement significative entre la souche sauvage et la souche délétée des quatre gènes NUDIX.

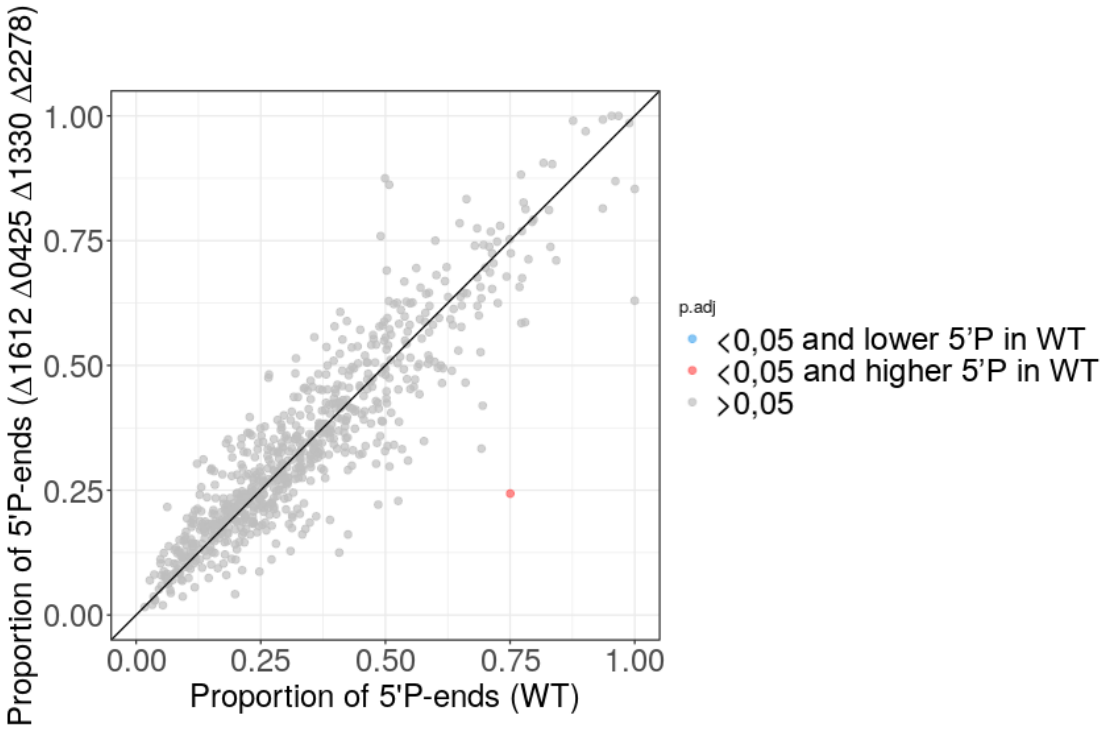


Figure 134 : Analyse différentielle de proportion d'extrémité 5'P.

Proportions de 5'P des TSS dans une souche sauvage (axe x) et dans une souche quadruple mutante Δ1612_Δ0425_Δ1330_Δ2278 (axe y). Chaque point représente une extrémité 5' correspondant à un TSS et les proportions affichées correspondent aux moyennes pondérées des réplicats retournées par DSS. Code couleur : point gris quand la différence n'est pas statistiquement significative ; point bleu quand une augmentation statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage ; point rouge quand une diminution statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage

Pour conclure sur les résultats de ces analyses, nous avons d'abord observé que les données issues de la librairie PQ-EMOTE 49 montrent que la délétion de la NUDIX SA1612 entraîne une légère augmentation de la proportion d'extrémité 5’P par rapport à la souche sauvage, tandis que la souche triple mutant ne révèle aucune différence statistiquement significative au niveau des TSS. Cependant, ces résultats sont peu informatifs en raison du nombre limité de TSS comparés. Cette librairie, l'une des premières générées par l'équipe collaboratrice, a permis d’améliorer le protocole, notamment par l'ajout d'un réplicat biologique supplémentaire, passant de 2 à 3 réplicats par souche. Cela est regrettable car cette librairie contient l'un des mutants les plus intéressants, le ∆1612, qui est la NUDIX présentant le plus de similarité avec la RppH de *B. subtilis*.

Les librairies PQ-EMOTE 56 et 58 sont de meilleure qualité car elles comprennent plus de réplicats biologiques et une profondeur de séquençage plus élevée, permettant ainsi la comparaison de la proportion d'un nombre beaucoup plus important de TSS (entre 2 et 10 fois plus que PQ-EMOTE 49). Les comparaisons entre les souches sauvages et mutantes issues de la librairie PQ-EMOTE 56 révèlent une augmentation de la proportion d'extrémité 5’P des TSS, tandis que celles issues de la librairie PQ-EMOTE 58 ne montrent presque aucun TSS différentiellement phosphorylé, malgré la délétion de certains des mutants ayant causés une augmentation chez PQ-EMOTE 56.

Bien que ces résultats soient à interpréter avec prudence en raison de la variation de plus ou moins 0,5 dans l'estimation des proportions des ARN toujours 5’P, il semble que les souches simples mutantes soient caractérisées par une augmentation de la proportion d'extrémités 5’P, tandis que les multiples mutants ne montrent aucune différence.

Pour améliorer la quantification et la puissance statistique dans la détection des TSS différentiellement phosphorylés, l'équipe collaboratrice a généré des données avec des UMI de taille 17, ne requérant ainsi plus l'étape de désaturation, qui, comme nous l'avons vu, peut engendrer des erreurs d'estimation sur les comptages d’extrémité 5’. Un réplicat biologique supplémentaire a également été généré, faisant de la librairie PQ-EMOTE 62 la plus fiable en termes de quantification.

Pourtant, la moyenne des écarts-types par TSS entre réplicats est de 0,164 pour PQ-EMOTE 62, contre 0,089 pour PQ-EMOTE 58, 0,090 pour PQ-EMOTE 56 et 0,308 pour PQ-EMOTE 49.

En effet, un changement de seuil lors de l'étape de filtrage sur gel a entraîné une perte de profondeur de séquençage, réduisant le nombre de TSS comparés. Néanmoins, les 720 TSS comparés ne révèlent aucune différence entre la souche mutante et la souche quadruple mutant pour 4 des 5 candidats, y compris les 3 qui, pour les simples mutants, avaient provoqué une augmentation de la proportion d'extrémité 5'P.

Finalement, il est difficile de tirer des conclusions définitives, principalement en raison de la variabilité dans l'estimation des proportions. Néanmoins, nous pouvons formuler certaines hypothèses sur les résultats observés :

– **Interactions croisées entre enzymes :** Il est possible qu'il y ait de nombreuses interactions croisées entre les enzymes (cross-talk), et que le dernier candidat (SA1501) possède à lui seul l'activité RppH responsable de la déphosphorylation des extrémités 5' observée dans le quadruple mutant (QM).

– **Amélioration du protocole :** Le protocole pourrait nécessiter des améliorations, notamment une augmentation de la profondeur de séquençage, une réduction du nombre de lectures invalides dues à la séquence de contrôle, et une diminution de la variation dans l'estimation des proportions résultant d'une séparation peu homogène des deux pools.

– **Limitations du pipeline d'analyses :** Le pipeline d'analyses lui-même pourrait présenter des limitations, notamment dans sa capacité à désaturer et/ou normaliser les données, ce qui pourrait empêcher la détection de différences de phosphorylation peu importantes.

Pour évaluer ces différentes hypothèses, plusieurs approches ont été mises ou sont en train d'être mises en place :

<u>Nouvelles données PQ-EMOTE</u>

Les données de meilleure qualité et les plus profondes (58 et 62) avec le plus de NUDIX inactivé suggèrent que la mutation des 4 NUDIX n'engendre pas, ou très peu, de variations dans les états phosphorylation des extrémités 5'P des ARN. Cependant, il serait utile de générer une librairie de haute qualité, incluant toutes les optimisations proposées, à savoir des UMI longs, 4 réplicats par souches, une profondeur de séquençage élevé et un seuil de filtrage sur gel permettant la suppression d'un maximum d'ARN 5S. L'idéal serait de généré ces données à partir des souches mutante Δ1612 et quintuples mutants pour les 5 candidats NUDIX. A partir de la première, nous pourrions vérifier si l'augmentation visible sur les quelques TSS dans la librairie PQ-EMOTE 49 peut se généraliser et en même temps nous pourrions voir si l'augmentation de la proportion d'extrémités

5’P dans les souches simples mutante est un effet librairie ou non. La souche quintuple mutante, quant à elle, permettrait de déterminer si en inactivant l’ensemble des NUDIX prédites, l’activité RppH de déphosphorylation perdure encore.

Les souches mutantes inductibles Δ1501 et Δ1612_Δ0425_Δ2278_Δ1330_Δ1501 ont été construites et des expériences PQ-EMOTE vont prochainement être réalisées.

Evaluation de la robustesse de l’estimation des proportions

Afin de tester la robustesse de l’estimation des proportions d’extrémités 5’P, nous avons prévu de comparer la proportion d’extrémités 5’P estimer par le protocole et la chaine de traitement PQ-EMOTE avec celles calculé par Nothern Blot, une approche expérimentale permettant la quantification de molécule d’ARN sur gel.

L’idée est assez similaire à celle de PQ-EMOTE puisque pour quantifier les états de phosphorylation des extrémités 5' des ARN par Northern Blot, un échantillon d'ARN est divisé en deux pools, l’un traitée avec l'enzyme *Ec*RppH pour déphosphoryler les extrémités 5'PPP, et l'autre laissé non traitée. Ces échantillons sont séparés par électrophorèse sur gel d'agarose ou de polyacrylamide, puis transférés sur une membrane de nylon ou de nitrocellulose. Une sonde radioactive spécifique à l'ARN d'intérêt, capable de uniquement aux extrémité 5'P est utilisée pour hybridation. Après lavage pour éliminer les sondes non spécifiques, la membrane est exposée à un film radiographique ou un détecteur d'image pour révéler les bandes radioactives des ARN liés à la sonde. Les images obtenues montrent les bandes correspondant aux ARN d’intérêt et l'intensité des bandes est mesurée. La proportion d'ARN avec une extrémité 5'P est déterminée en divisant l'intensité de la bande des ARN non traités par celle des ARN traités avec *Ec*RppH.

Cependant, cette approche n’est pas une approche globale et il sera possible de comparer uniquement quelque extrémités 5’. Les résultats pourront néanmoins nous fournir un aperçu de la robustesse de PQ-EMOTE.

Un Northern blot a déjà été effectué sur les ARN des gènes hu et hisS dont les proportions d’extrémités 5’P ont respectivement été estimées à 0,36 et 0,41 par PQ-EMOTE et les images sont en cours d’analyse pour l’estimation par Nothern Blot.

Test de la chaine de traitement sur des données non issues d’un protocole PQ-EMOTE

Pour évaluer l'efficacité et la robustesse de notre chaîne de traitement, nous allons, dans la prochaine section, l'appliquer à des données issues d'un autre protocole que PQ-EMOTE. En

particulier, nous traiterons des données publiées dans l'étude de (Bischler et al., 2017), qui examine la RppH de *Helicobacter pylori*. Cette démarche nous permettra de vérifier si notre chaine de traitement est capable de reproduire des résultats déjà validés et d'identifier les cibles de RppH dans un contexte différent. En utilisant des données externes et bien caractérisées, nous pourrions ainsi confirmer que notre chaîne de traitement est fiable et adaptable à d'autres ensembles de données et protocoles expérimentaux.

2.2.2 Traitement des données publiées sur la RppH de *H. pylori*

Afin de vérifier la fiabilité de notre chaîne de traitement, nous avons testé sa capacité à reproduire les résultats obtenus par Bischler et al. (2017) dans l'identification des cibles de la RppH de *H. pylori*. Les alignements locaux réalisés entre la RppH de *H. pylori*, la RppH de *E. coli* et les cinq NUDIX identifiés chez *S. aureus* révèlent, comme illustré dans la Table 14, une faible identité de séquence entre *Hp*RppH et *Ec*RppH, mais une haute couverture d'alignement. Cela n'est pas surprenant, car *H. pylor*i et *E. coli* sont toutes deux des protéobactéries.

En revanche, les NUDIX de *S. aureus* présentent des couvertures d'alignement très faibles avec *Hp*RppH, ce qui n'est pas inattendu étant donné que ces bactéries ne sont pas étroitement apparentées. Néanmoins, il est notable que la protéine SA2278 de *S. aureus* présente une couverture de 41 % et une identité de séquence de 35 % avec *Hp*RppH sur cette région.

NUDIX	**Couverture**	**Identité**
EcRppH	92 %	36 %
SA1612	25 %	41 %
SA0425	15 %	54 %
SA2278	41 %	35 %
SA1330	17 %	40 %
SA1501	Pas de similarité	

Table 14 : Alignements locaux entre HpRppH, EcRppH et les 5 candidats NUDIX de chez S. aureus. *Tableau représentant le pourcentage de couverture et d'identité des alignements retournés pas le programme BLASTP*

La RppH de *H. pylori* a été identifiée *in vitro*, puis ses cibles ont été analysées à l'aide d'un protocole similaire à PQ-EMOTE. Comme pour les données PQ-EMOTE, celles produites par Bischler et al. doivent être analysées par groupe d'échantillons. Outre les échantillons pool^{mono} (5'P) et pool^{total} (5'P + 5'PPP), ils proposent également un échantillon contenant uniquement les 5'PPP, qui ne nous est cependant pas utile dans notre chaîne de traitements. Leurs expériences n'ont pas été conduites en réplicats. En plus d'une souche sauvage et d'une souche mutante pour la RppH, ils ont également réalisé des expériences à partir d'une souche mutante pour la RppH mais complémentée.

Les lectures ne contiennent ni oligonucléotides (donc pas d'UMI), ni code-barres. En revanche, les séquences des transcrits peuvent contenir des séquences de poly A, résultant d'une étape de leur protocole expérimental. Nous avons retiré les séquences polyA afin de pouvoir aligner les lectures sur le génome de références. De la même manière que les auteurs, nous n'avons pas gardé les lectures qui, après suppression des polyA, font moins de 12 nucléotides de long.

Les lectures ont été alignées avec Bowtie en autorisant un mésappariement et les lectures multi-alignées ont été conservées.

Comme nous l'avons fait pour les données PQ-EMOTE, nous avons effectué un filtrage des positions pour lesquelles moins de 50 comptages sont observés dans l'échantillon pool^{total}.

Pour la normalisation, nous avons vérifié si la normalisation proposée par (Martini et al., 2021) et présentée plus haut, qui utilise la distribution bimodale, était réalisable. Nous avons également prêté une attention particulière aux extrémités 5' correspondant aux ARNr afin d'étudier leur proportion d'extrémités 5'P et leur variabilité.

La méthode d'identification des cibles RppH mise en œuvre par Bischler et al. (2017), illustrée dans la Figure 135, repose sur deux types de comparaisons :

- Une comparaison des abondances des transcrits dans le pool -TEX/+TAP (5'PPP + 5'P). Les comptages sont agrégés dans une fenêtre de 100 nucléotides autour des TSS et les valeurs sont comparées entre souches à l'aide du programme Gfold.
- Une comparaison des proportions de 5'P et 5'PPP. Les comptages sont agrégés dans une fenêtre de plus ou moins 5 nucléotides autour des TSS. Les valeurs, sous forme de table de contingence sont comparées à l'aide d'un test exact de Fisher.

Les ARN dont les TSS ont une augmentation statistiquement significative d'abondance dans le pool^{total} ainsi qu'une augmentation statistiquement significative de proportion d'extrémité 5'P dans la souche Δ*rppH* par rapport aux souche sauvage et C*rppH* sont considérés comme des cibles RppH.

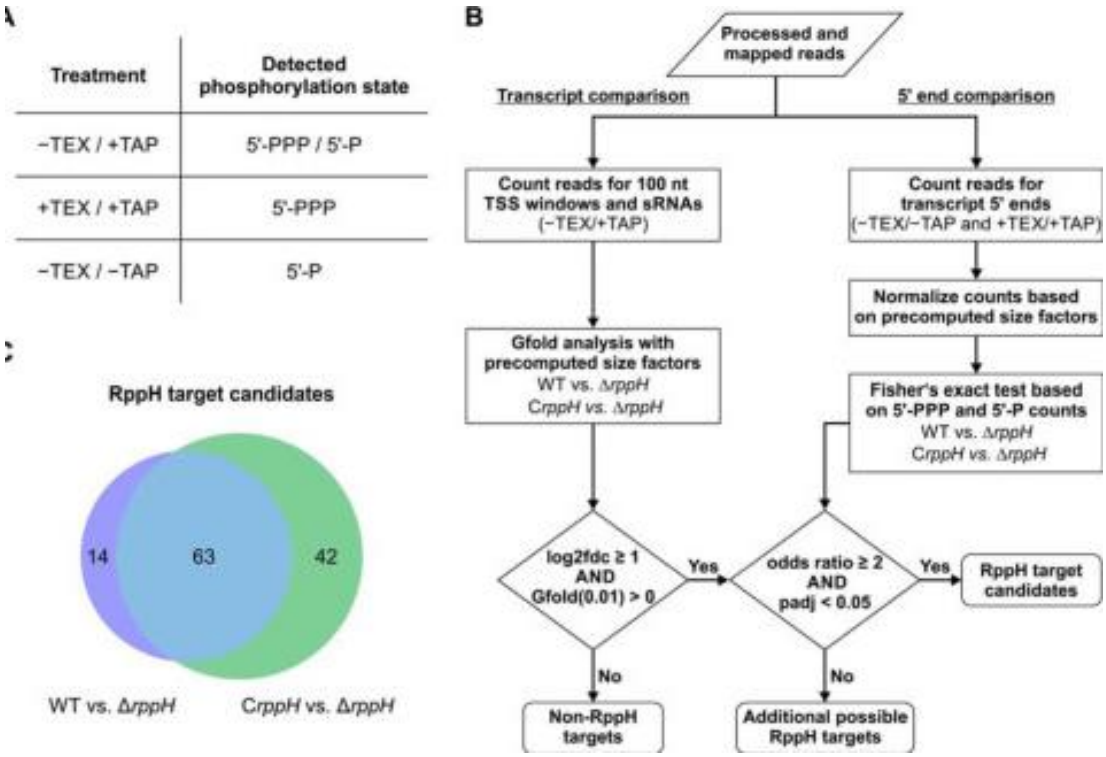


Figure 135 : Méthode utilisée par Bischler et al. (2017) pour l'identification des cibles RppH. *(**A**) Traitement effectué sur les échantillons d'ARN et états de phosphorylation détectés. (**B**) Procédure bio-informatique d'identification des cibles RppH. (**C**) Nombre de cibles RppH identifiées par comparaison. Extrait de* (Bischler et al., 2017)

La Table 15 présente le nombre total de lectures et le nombre de lectures alignées pour chaque échantillon. En moyenne, nous observons 4 millions de lectures alignées par échantillon, ce qui est inférieur de 2 millions par rapport aux données PQ-EMOTE. Toutefois, il est important de noter que la comparaison directe des profondeurs de séquençage entre ces deux ensembles de données n'est pas entièrement pertinente en raison des différences de taille des génomes. Le génome étudié ici est 1,5 fois plus petit en termes de nombre de paires de bases (1 667 892) que celui de la souche SA564 de *S. aureus* (2 783 537).

Échantillon	Nombre de lectures	Nombre de lectures alignées
WT (mono)	4 554 168	3 436 075
WT (total)	4 723 461	4 047 876
ΔrppH (mono)	5 186 438	3 741 936
ΔrppH (total)	5 705 741	4 870 704
ΔrppH complémenté (mono)	5 829 281	4 068 473
ΔrppH complémenté (total)	4 691 092	3 972 578

Table 15 : Nombre de lectures et nombre de lectures alignées en fonction de l'échantillon

– Filtrage

Sur les 1 667 892 positions du génome, 978 645 positions présentent au moins une extrémité 5' alignée et 7 832 après filtrage du pool^{total}.

– Normalisation

La Figure 136 représente les courbes de densité des logarithmes en base 2 des ratios pool^{total} / pool^{mono} des extrémités 5' des ARN, et permet de mettre en évidence les deux populations attendues (pic pour les extrémités toujours 5'P à gauche, et TSS à droite). Par ailleurs, nous observons que pour les trois souches, le pic à gauche, correspondant aux intermédiaires de dégradation et sites de clivages endoribonucléolytique, est presque déjà centré sur 0. De même, les trois points rouges représentants les ARNr sont très proches les uns des autres (bien plus que dans les données PQ-EMOTE), et presque centré sur 0 également. Le pic à gauche relativement resserré autour du mode proche de 0 indique que la dispersion des proportions obtenues par Bischler et al. semble être moins sujette aux variations techniques que celle obtenue avec le protocole PQ-EMOTE.

Ces résultats montrent que la méthode de normalisation basée sur le pic attendu en 0 est applicable pour cet échantillon. Le fait que cela n'ait pas été pas possible avec les données PQ-EMOTE ne semble pas dû à un manque de profondeur de séquençage étant donné que l'on dispose de moins de lectures alignées ici pour *H. pylori*.

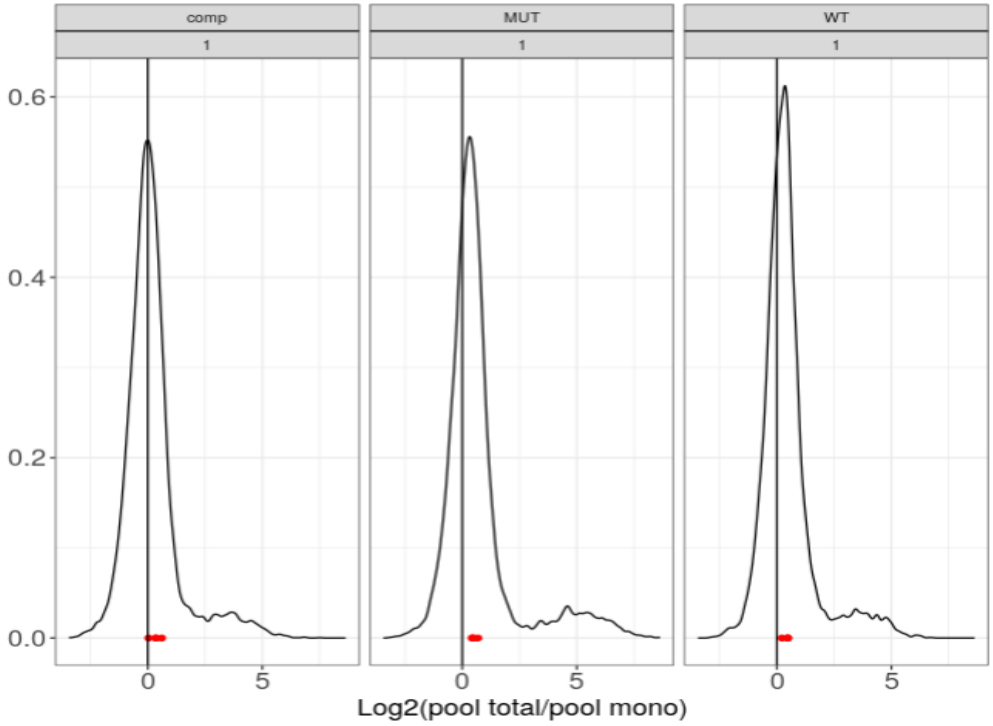


***Figure 136* : Courbes de densité des logarithmes en base 2 des ratios pool^{total} / pool^{mono} des extrémités 5' des ARN**. *Les points rouges représentent les 5' correspondant aux ARNr de H. pylori.*

- Proportion d'extrémités 5'P

La Figure 137 présente les distributions des proportions de 5'P dans les trois différentes souches. Bien que leur méthode inclue l'agrégation des comptages autour des TSS, nous ne disposons que des TSS correspondant aux 63 ARN ciblés par *Hp*RppH publiés. En conséquence, nous n'avons pas pu sélectionner uniquement les TSS comme nous l'avions fait avec les données PQ-EMOTE. Cependant, nous avons filtré les positions présentant une proportion supérieure à 2 afin de réduire le bruit que pourrait apporter les extrémités issues d'intermédiaires de dégradation ou d'autres ARN toujours 5'P. Ainsi, nous sommes passés de 7 832 positions à 7 416 pour lesquelles une proportion a pu être calculée. Aucune différence statistiquement significative n'a été observée entre les proportions des souches sauvage et mutante complémentée. En revanche, une diminution statistiquement significative de la proportion d'extrémités 5'P a été observée entre la souche ΔrppH et les deux autres souches. Ces résultats confirment les conclusions des auteurs, démontrant ainsi le rôle de RppH dans la déphosphorylation des extrémités 5' des ARN chez *H. pylori*.

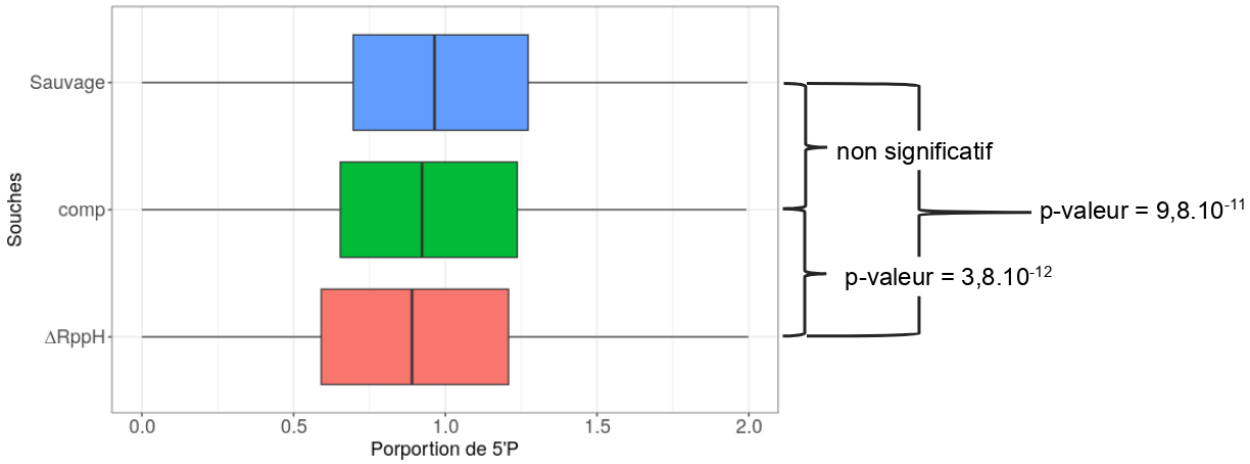


Figure 137 : Distribution des proportions de 5' des extrémités 5' dans les souches sauvage (bleu), ∆rppH complémentée (vert) et ∆rppH (rouge).

Il convient de noter que, dans les souches sauvage et mutante complémentée, la proportion médiane d'extrémités 5'P est respectivement de 0,97 et 0,94, ce qui est significativement plus élevé que dans les souches sauvages de SA564. Cela s'explique par le fait que, comme mentionné précédemment, nous n'avons pas pu sélectionner uniquement les TSS. En conséquence, la distribution est influencée par des intermédiaires de dégradation et d'autres positions toujours 5'P, ce qui tend à faire monter les distributions vers 1.

- Analyse différentielle de proportion

Nous avons comparé à l'aide de DSS les proportions d'extrémités 5'P entre les souches sauvage et mutante Δ*rppH* et entre les souches mutantes complémentée et mutante Δ*rppH*.

Les résultats, illustrés en Figure 138, révèlent un grand nombre d'extrémités 5' différentiellement phosphorylées. En effet, nous observons respectivement 341 et 332 extrémités 5' (dont 175 en commun) avec une diminution statistiquement significative (p-valeur ajustée < 0,05) de la proportion d'extrémités 5'P dans la souche la souche ΔrppH par rapport aux souches sauvage et ΔrppH complémentée (points rouges)**.**

Nous observons également 123 et 56 extrémités 5' avec une augmentation statistiquement significative (p-valeur ajustée < 0,05) de la proportion d'extrémités 5'P (points bleus)**.**

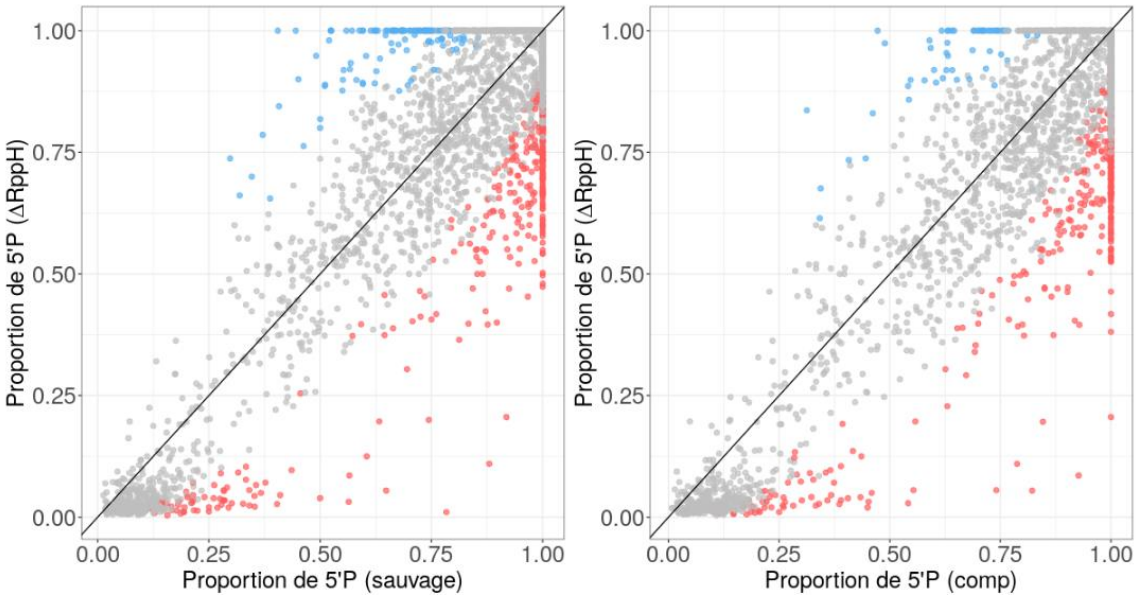


Figure 138 : *Analyse différentielle de proportion d'extrémités 5'P.*

Proportions d'extrémité 5'P dans la souche sauvage (panel gauche) et dans la souche mutante complémentée (panel droit) (axe x) et dans la souche quadruple mutante ΔrppH (axe y). Chaque point représente une extrémité 5'. Code couleur : point gris quand la différence n'est pas statistiquement significative ; point bleu quand une augmentation statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage ; point rouge quand une diminution statistiquement significative des proportions des extrémités 5'P est observée dans le mutant par rapport au sauvage

Les auteurs ont identifié 63 cibles de RppH. Cependant, leur méthode consiste à agréger les comptages des positions se trouvant dans une fenêtre de 5 nucléotides autour des TSS. Leurs résultats sont donc difficiles à comparer avec notre approche, qui considère chaque extrémité 5' individuellement. Nous avons toutefois regroupé les 341 et 332 positions se trouvant à moins de 5 nucléotides d'écart, obtenant ainsi 212 et 214 clusters de positions. Parmi ces clusters, nous avons respectivement identifié 3 et 0 clusters en commun avec ceux identifiés par Bischler et al. (2017).

Le fait de ne pas avoir détecté les mêmes cibles que Bischler et al. (2017) pourrait s'expliquer par le fait que les positions n'ont pas franchi le seuil de 50 lectures alignées dans l'échantillon pool^{total}, en raison de la non-agrégation des comptages, qui sont donc moins élevés.

Nous avons donc entrepris l'étude de cette possibilité et avons analysé le nombre de lectures alignées dans l'ensemble des pools^{total} pour les positions situées autour (± 5 nucléotides) des 63 TSS ciblés par RppH.

La Figure 139 illustre les distributions obtenues, révélant que parmi les 630 positions potentielles, 228, 242 et 272 d'entre elles présentent au moins une lecture alignée respectivement dans les souches sauvages, les mutants complémentés et ΔrppH. Parmi ces positions, seulement 7 (formant 6 clusters), 6 (formant 6 clusters) et 22 (formant 12 clusters) d'entre elles montrent au moins 50 lectures alignées. Cette observation permet d'expliquer, pourquoi toutes les cibles identifiées dans l'étude de Bischler et al. (2017) n'ont pas été détectées par notre approche.

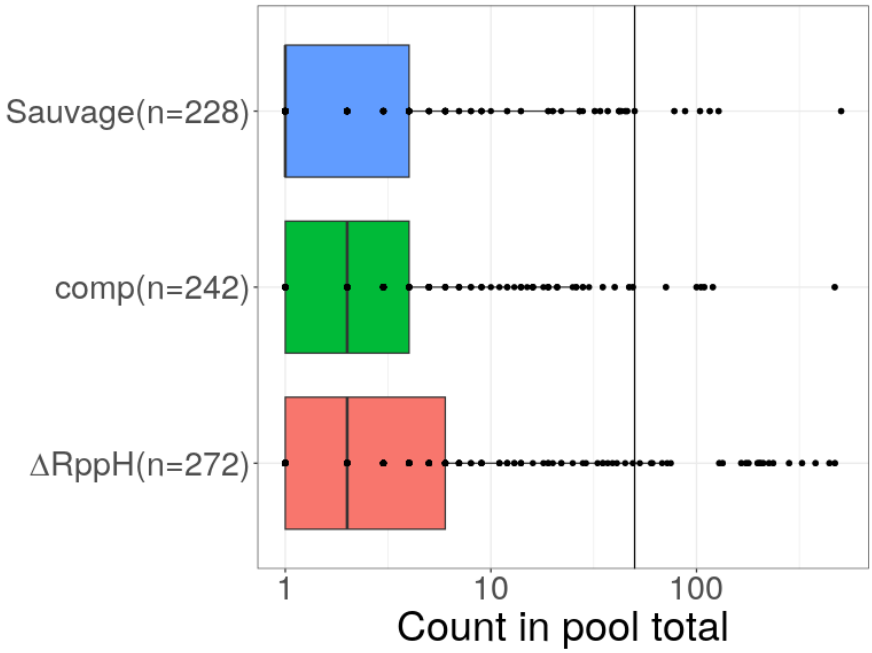


Figure 139 : Nombre de lectures alignées dans les échantillons pool^{total} par souches.

Souche sauvage (bleu), souche mutante ΔrppH complémenté (vert) et souche mutante ΔrppH (rouge). Seules les positions dans la région des 63 TSS sont affiché. La ligne verticale indique le seuil de 50 choisi pour l'étape de filtrage.

Ces résultats montrent que notre approche, du fait de sa précision (extrémité 5' individuelle) nécessite une plus grande profondeur de séquençage que l'approche de Bischler et al. (2017).

Pour conclure, notre approche permet bien de détecter une baisse globale de la proportion d'extrémité 5'P dans la souche Δ*rppH* par rapport aux souches sauvage et mutante complémentée.

Cependant, il est difficile de comparer les deux approches quant à l'identification des cibles de cette RppH. Chacune présente des avantages et des inconvénients distincts.

Notre méthode se distingue par sa capacité à identifier individuellement des extrémités dont la proportion en forme 5'P varie entre conditions, contrairement à l'approche de Bischler et al. (2017) qui se concentre sur l'identification des transcrits ciblés par RppH la rendant plus sensible par l'agrégation des comptages. De plus, leur utilisation d'un test exact de Fisher ne permet pas de tenir compte des réplicats biologiques.

L'étape supplémentaire de comparaison des abondances des pool^{total} est également une approche intéressante qu'il sera judicieux d'essayer dans les données PQ-EMOTE.

Par ailleurs, notre méthode semble prometteuse pour découvrir de nouvelles cibles de RppH chez *H. pylori*. Cependant, il serait préférable à l'avenir de pouvoir sélectionner spécifiquement les TSS, afin de potentiellement réduire le nombre de faux positifs. Cette approche pourrait ouvrir des perspectives intéressantes pour des études futures.

2.2.3 Spécificité de la ou des RppH de *S. aureus*

Bien que nous n'ayons pas réussi à identifier la ou les RppH de *S. aureus*, nous pouvons néanmoins étudier son activité de déphosphorylation *in vivo*, notamment ses spécificités de substrat, en analysant la proportion d'extrémité 5'P des extrémités 5' des ARN en fonction de leur composition en ribonucléotides.

Pour ce faire, nous avons comparé, à l'aide de tests de Wilcoxon-Mann-Whitney, les proportions de 5'P dans la souche sauvage en fonction du premier et du deuxième nucléotide, ainsi que du premier di-nucléotide. Nous avons utilisé les données issues de la librairie PQ-EMOTE 58 car il s'agit de la librairie offrant la plus haute profondeur de séquençage et la meilleure concordance de réplicats.

Nous avons étudié uniquement les extrémités 5' identifiées comme des TSS, afin d'éviter une distribution biaisée par les intermédiaires de dégradation ou les sites de clivage endoribonucléolytique qui sont toujours 5'P.

Il convient de noter qu'il a été observé que la RppH d'*E. coli* présentait des biais de séquence cible *in vitro*, dont le plus important se situait à la deuxième position nucléotidique où les 5'PPP-GR sont préférés aux 5'PPP-GY (où R et Y sont des purines et des pyrimidines, respectivement) pendant les 8 minutes examinées (Foley et al., 2015). Dans le protocole PQ-EMOTE, les pools^{total} sont traités avec *Ec*RppH pendant 4 heures, ce qui permet normalement à l'enzyme de convertir même les ARN non préférés en 5'P.

Néanmoins, si tous les ARN 5'PPP n'ont pas été convertis par *Ec*RppH dans le pool^{total}, le nombre d'ARN dans ce pool^{total} commençant par GR devrait être plus élevé que celui des ARN commençant par GY. Cela induit un ratio (pool^{mono} / pool^{total}) plus élevé pour les ARN commençant par GY que par GR.

De plus, le même article montre également une préférence de *Ec*RppH pour le di-nucléotide AG par rapport à AC et une préférence pour le di-nucléotide GA par rapport à GU.

Les résultats issus des 1 444 TSS pour lequel nous avons pu calculer une proportion d'extrémité 5'P dans au moins un réplicats issus de la souche sauvage, sont présentés dans la Figure 140 et montrent que les extrémités 5' dont le premier ribonucléotide est une guanosine possèdent une proportion de forme 5'P significativement plus élevée que les trois autres (p-valeur ajustée < $1{,}14.10^{-5}$). En outre, nous constatons que les extrémités dont le deuxième nucléotide est une adénosine présentent une proportion de forme 5'P statistiquement supérieure à celles dont le deuxième nucléotide est une uracile (p-valeur ajustée = 0,002) ou une cytidine (p-valeur ajustée = 0,01).

Par ailleurs, les comparaisons des proportions de 5'P par di-nucléotide, illustrées en Figure 141A, révèlent que les extrémités 5' dont le premier di-nucléotide correspond à GR présentent une proportion d'extrémités 5'P statistiquement plus élevée (p-valeur ajustée = $1{,}07.10^{-6}$) que celles dont le premier di-nucléotide est un GY.

De plus, nous avons observé une augmentation statistiquement significative pour GA comparée à GU (p-valeur ajustée = $2{,}00.10^{-3}$) et une augmentation qui n'est pas statistiquement significative de la proportion d'extrémité 5'P des AG par rapport aux AC (Fig. 141B).

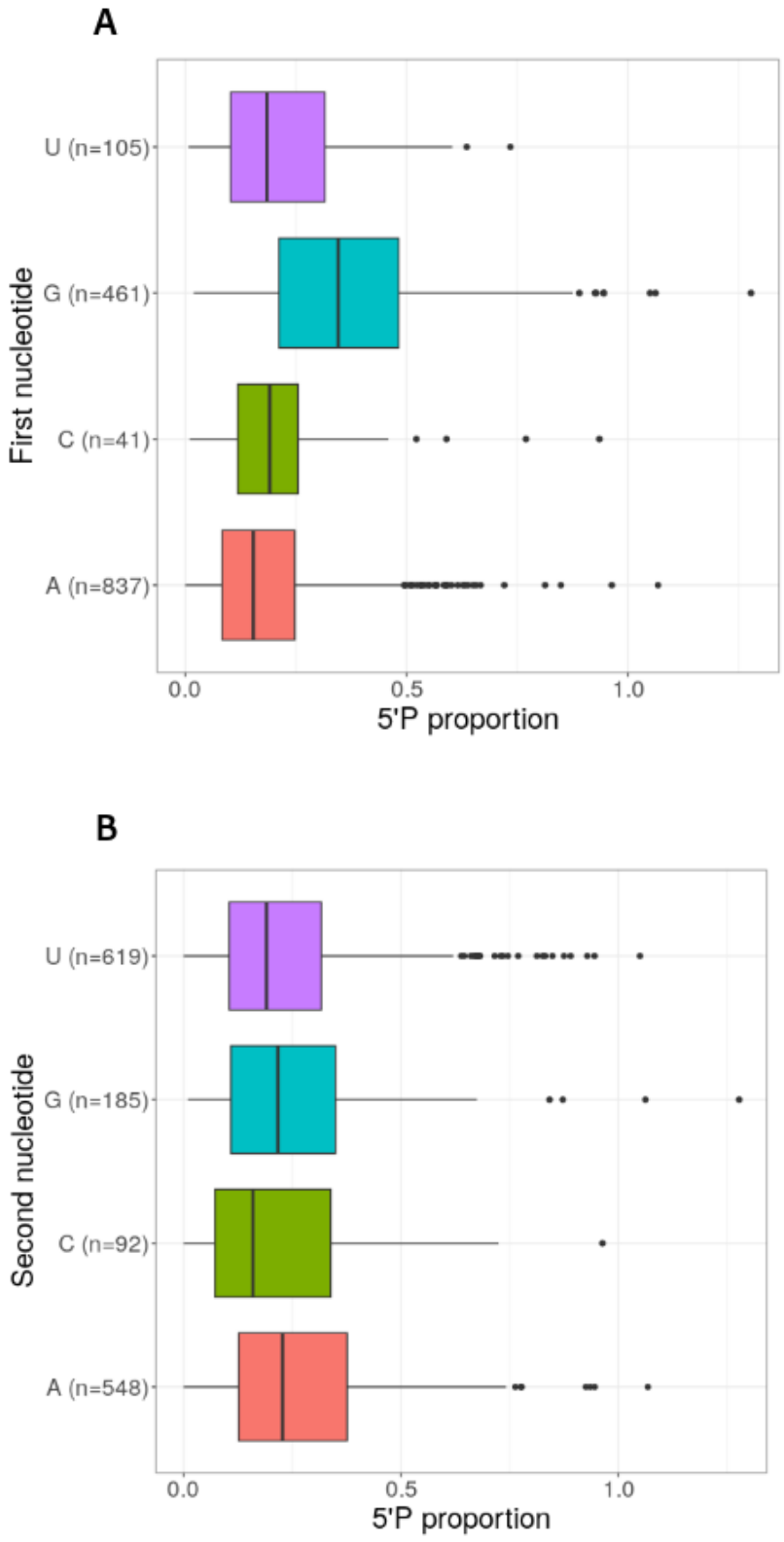
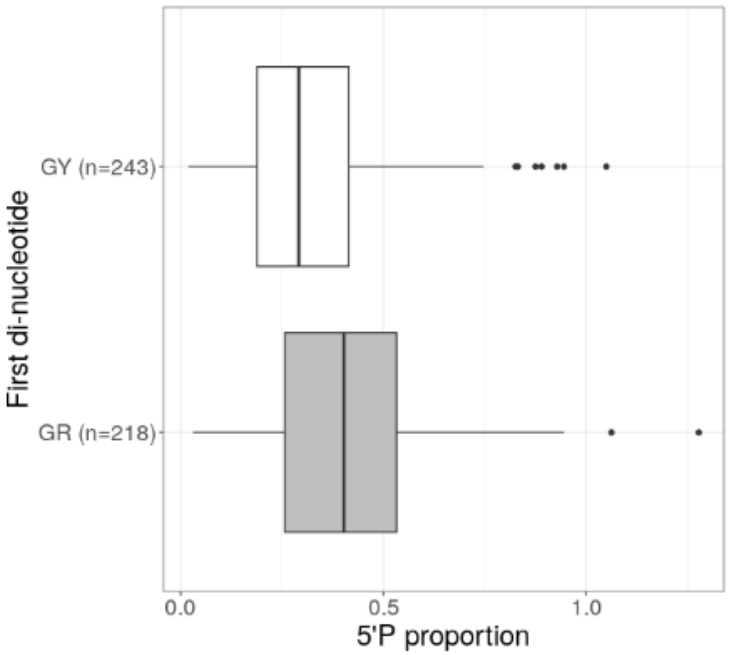


Figure 140 : ***Distribution des proportions d'extrémités 5'P en fonction (A) du premier ou (B) du deuxième nucléotide de l'extrémité 5'.***

(A)



(B)

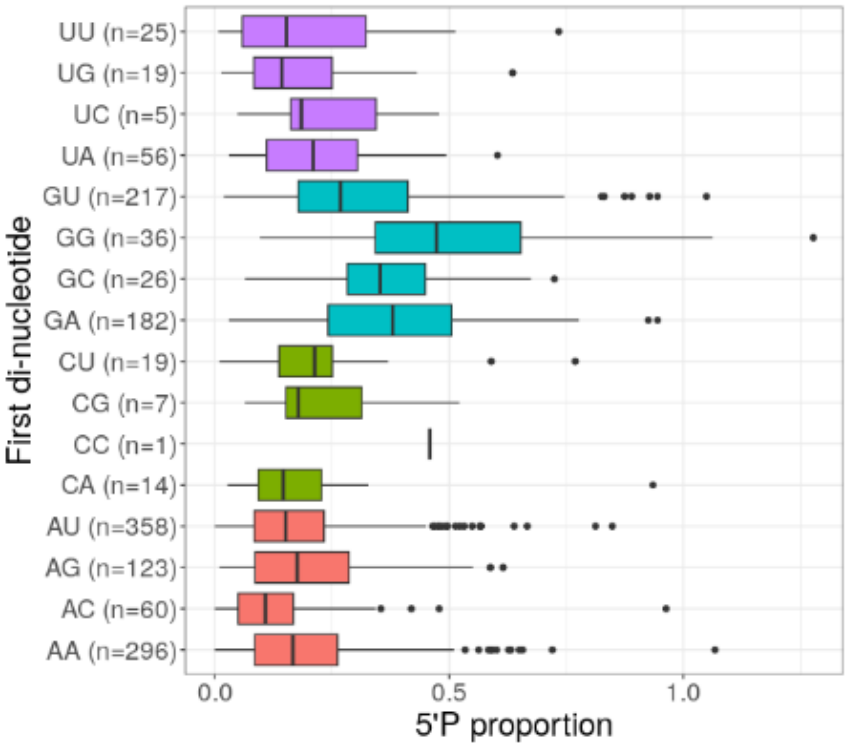


Figure 141 : ***Distribution des proportions de 5'P en fonction du premier di-nucléotide de l'extrémité 5'.***

Ces résultats sont à l'inverse de ce qui été attendu dans le cadre d'un biais de proportion d'extrémité 5'P provoqué par les préférences de *Ec*RppH. Ceci indique que l'utilisation de *Ec*RppH dans le pool^{total} ne biaise pas l'estimation des proportions d'extrémité 5'P.

En outre, nous observons d'autres différences statistiquement significatives entre les ensembles de di-nucléotides dont le premier nucléotide est le même :

– GU diminution par rapport à GG (p-valeur ajustée = 1.10^{-3})

– AA augmentation par rapport à AC (p-valeur ajustée = $1{,}20.10^{-2}$)

Cela suggère que parmi les ARN dont le premier ribonucléotide est une adénosine, ceux dont le deuxième est une cytidine sont moins déphosphorylés que les autres.

De même, parmi les ARN dont le premier ribonucléotide est une guanosine, ceux dont le deuxième est une autre guanosine sont plus déphosphorylés que les autres.

Pour conclure, il semblerait que les ARN dont le premier ribonucléotide est une guanosine soit préférentiellement déphosphorylés chez *S. aureus* ce qui contraste avec les préférences de *Ec*RppH (Foley et al., 2015) et *Bs*RppH (Hsieh et al., 2013) illustré en Figure 142 panel A et B.

Les travaux de Hsieh et al. (2013) et de Foley et al. (2015) ont révélé que la RppH de *B. subtilis* et celle de *E. coli* déphosphorylent plus efficacement les extrémités 5' dont le premier ribonucléotide est une adénosine, comparativement à celles débutant par une guanosine. Ce comportement est inverse de celui observé chez *S. aureus*.

En outre, ces études ont démontré une préférence marquée (presque exclusive) de *Bs*RppH pour les extrémités 5' dont le deuxième ribonucléotide est une guanosine. Chez *E. coli*, cette préférence est également présente mais moins prononcée que chez *B. subtilis*. Dans nos résultats concernant *S. aureus*, nous n'observons pas de préférence statistiquement significative pour une guanosine en deuxième position. Cependant, cette tendance se manifeste par une préférence pour le di-nucléotide GG par rapport aux autres di-nucléotides.

En revanche, comme chez *E. coli*, il semble que les ARN de *S. aureus* débutant par les di-nucléotides GR soient préférentiellement déphosphorylés par rapport aux GY. De même, les ARN débutant par AG sont déphosphorylés plus efficacement que ceux débutant par AC, et ceux débutant par GA sont préférentiellement déphosphorylés par rapport à ceux débutant par GU.

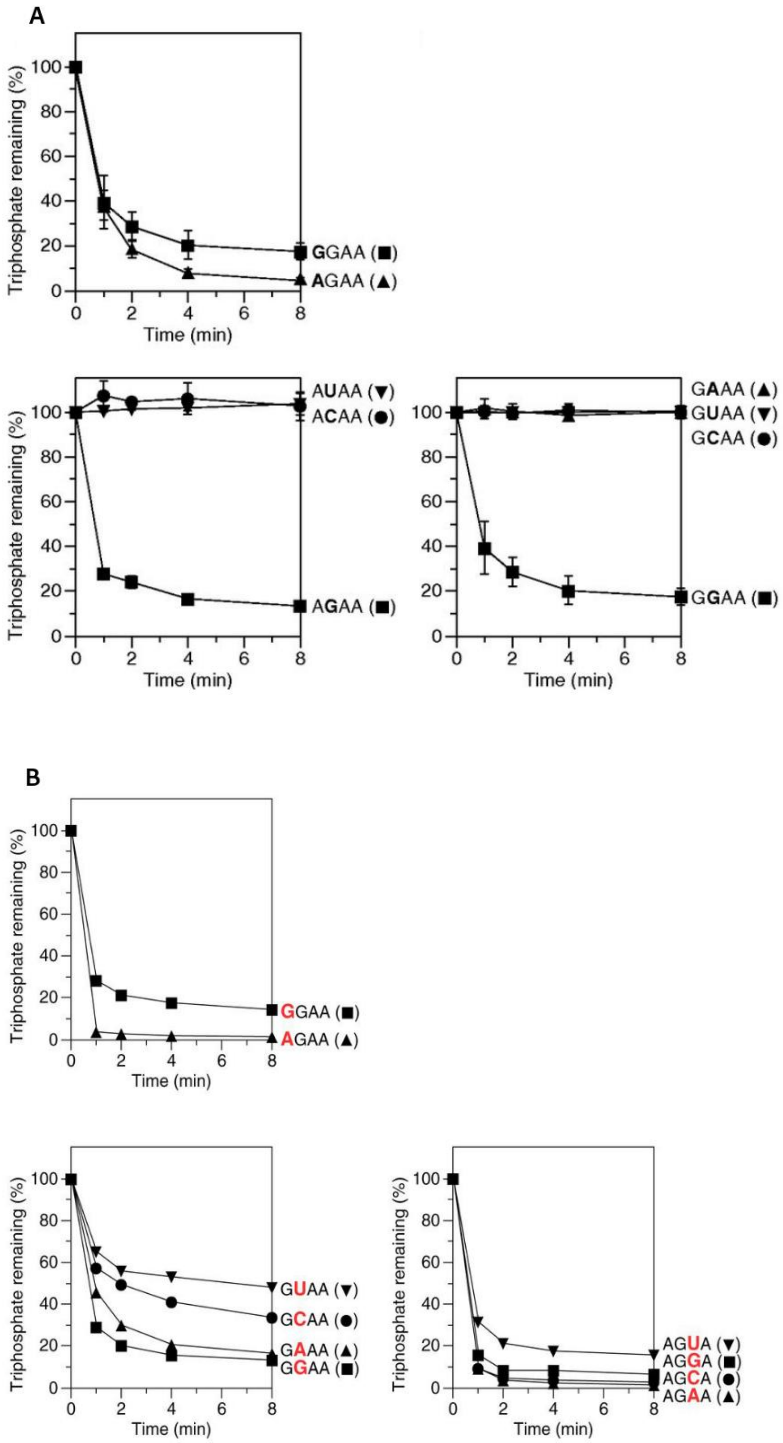


Figure 142 : ***Préférences des RppH de E. Coli et de B. subtilis.*** *Effet de la séquence des deux premiers ribonucléotides dans l'activité in vitro de (**A**) la RppH de E. coli et (**B**) de la RppH de B. subtilis. L'activité est mesurée en pourcentage de 5'PPP restant au cours du temps. Extrait de* (Hsieh et al., 2013) *et de* (Foley et al., 2015).

Les résultats de Foley et Hsieh proviennent d'expérimentations *in vitro*, où les conditions sont beaucoup plus contrôlées (quantité de RppH, présence d'autres enzymes, ARN testés, *etc.*) que dans les environnements *in vivo* à partir desquels nos résultats ont été obtenus. En effet, dans un modèle *in vivo*, il est possible que les états de phosphorylation observés ne soient pas uniquement dus à la ou aux RppH de *S. aureus*. Par exemple, si la RNase J dégrade davantage les 5'P, et que les ARN commençant par G sont plus souvent mono-phosphorylés, le nombre d'ARN mono-phosphorylés en G pourrait diminuer en raison de cette dégradation, ce qui entraînerait une proportion d'extrémités 5'P plus faible que celle observée.

Si les préférences observées sont effectivement dues aux NUDIX, il serait intéressant de vérifier si elles disparaissent dans les souches mutantes pour les NUDIX. Nous avons effectué les mêmes comparaisons des distributions des proportions d'extrémités 5'P dans chacune des 7 souches mutantes. L'ensemble des souches mutantes (sauf celles issues de PQ-EMOTE 49 car il n'y pas assez de TSS pour faire des comparaisons) présentent également une plus grande proportion d'extrémités 5'P statistique pour les ARN dont le premier ribonucléotide est une guanosine, pour les ARN dont le premier di-nucléotide est un GR par rapport aux GY et pour les ARN dont le premier di-nucléotide est GG par rapport aux autres di-nucléotides. Ces résultats suggèrent que le ou les facteurs responsables des préférences ribonucléotidiques en termes de déphosphorylation sont encore actifs dans les souches mutantes ∆0425, ∆2278, Δ1612_Δ1330, Δ1612_Δ0425_Δ1330 et Δ1612_Δ0425_Δ1330_∆2278.

2.2.4 Discussion

Dans ce chapitre, nous avons présenté des analyses approfondies visant à étudier les états de phosphorylation des extrémités 5' des ARN chez *S. aureus*. Ces analyses reposent sur des données issues d'un protocole innovant de séquençage des extrémités 5' des ARN et de quantification de leurs états de phosphorylation, dénommé PQ-EMOTE (Phosphorylation Quantification Exact Mapping Of Transcriptome Ends). Parallèlement à l'optimisation de ce protocole, nous avons développé une chaîne de traitement dédiée à l'analyse des données générées.

Nos travaux ont démontré que ce développement conjoint a permis d'améliorer significativement le protocole expérimental, notamment par l'allongement des UMI qui offre une quantification plus fiable des lectures alignées, ainsi que par l'augmentation du nombre de réplicats, ce qui renforce la puissance statistique des tests effectués *via* la méthode de DSS.

Par ailleurs, afin de valoriser les données issues de protocoles non optimisés, notamment avec des UMI trop courts, nous avons élaboré une méthode pour désaturer les UMI. La saturation des UMI est un problème peu documenté dans la littérature bien qu'il soit notable, comme nous l'avons démontré. À notre connaissance, la désaturation n'avait jamais été abordée auparavant et notre méthode permet de réduire de manière grandement les taux d'erreur dans l'estimation du vrai nombre de transcrits.

Malgré ces avancées dans la quantification des lectures, plusieurs améliorations restent envisageables, notamment en ce qui concerne la profondeur de séquençage. Comme nous l'avons montré, celle-ci est cruciale pour :

- améliorer la fiabilité de l'estimation des proportions d'extrémités 5'P
- augmenter le nombre de TSS étudiés
- détecter et corriger les effets de lots.

De plus, il serait également important d'investiguer plus en détails le fait que la séquence de contrôle présente les oligonucléotides RpX, bien que ne faisant que 5 nucléotides de long, soit la principale source d'erreur dans la validation des lectures.

Même avec une profondeur de séquençage adéquate, notre étape de normalisation a révélé une grande dispersion dans l'estimation des proportions d'extrémités toujours 5'P. Cette variation semble inhérente à la profondeur de séquençage, au nombre de réplicats, et à la taille des UMI. Notre hypothèse actuelle est que cette variation serait due à la séparation des échantillons en deux pools qui rendrait la répartition des ARN non homogène. En effet, lors de l'analyse des données de (Bischler et al., 2017)**,** où les échantillons étaient préparés indépendamment, une telle variation n'a pas été observée, ce qui soutient notre hypothèse.

Par ailleurs, nous avons justement comparé notre chaine de traitements avec celle proposé par Bischler et al. (2017) qui permet d'identifier les ARN ciblés par les RppH. Cette comparaison a révélé que notre chaîne de traitements est moins sensible, identifiant moins de cibles avec un seul réplicat. Cependant, notre méthode se distingue par une précision accrue, permettant l'identification individuelle des extrémités ciblées par RppH.

Par la suite, nous avons toutefois pu démontrer que dans la souche sauvage SA564 de *S. aureus*, le niveau médian de proportion d'extrémités 5'P est de 21 %, ce qui implique que 79 % des ARN se trouvent sous forme 5'PPP. Cela suggère que l'étape de transcription est plus rapide que les mécanismes de déphosphorylation et/ou que les ARN 5'P sont rapidement dégradés.

Étant donné que la RNase J1 a une légère préférence pour les ARN 5'P et que la proportion d'extrémités 5'P augmente légèrement, bien que de manière non statistique, dans la souche mutante J1^{AGA}. Cette observation suggère que la deuxième hypothèse est cohérente. De plus, nous avons identifié cinq protéines NUDIX chez *S. aureus*, qui pourraient potentiellement posséder une activité RppH et être responsables de la déphosphorylation des extrémités 5' des ARN. Afin d'identifier la ou les RppH de *S. aureus*, 7 souches mutantes (simples ou multiples) ont été construites et des données PQ-EMOTE ont été générées et analysées. Les trois souches simples mutantes pour les NUDIX SA0425, SA1612, et SA2278 présentent une augmentation de la proportion d'extrémités 5'P par rapport à la souche sauvage. Cependant, il est actuellement difficile d'être certain que cette augmentation ne soit pas due à un effet de librairie, car deux des trois souches proviennent de la même librairie et la troisième de la librairie la moins profonde et la moins fiable. Il serait donc crucial de régénérer des données pour une souche simple mutante en utilisant les optimisations récentes du protocole.

L'ensemble des souches multiples mutante ne présente presque aucune différence de proportion d'extrémités 5'P, même la souche quadruple mutante pour quatre des cinq NUDIX identifiées. Ces résultats sont surprenants même si des interactions croisées sont attendues.

La souche mutante pour le dernier NUDIX (SA1501), encore non testée, a été construite par l'équipe collaboratrice, et la souche quintuple mutante est en cours de construction. Une fois les données PQ-EMOTE générées, nous serons en mesure de tirer des conclusions sur le rôle de cette dernière NUDIX dans la régulation des états de phosphorylation des extrémités 5' des ARN.

Si aucun effet n'est observé il deviendra alors crucial de procéder à des expérimentations *in vitro* pour identifier la ou les NUDIX ayant une activité RppH, comme l'ont fait (Richards et al., 2011) pour identifier la RppH de *H. pylori*, (Bischler et al., 2017) pour celle de *B. subtilis* ou encore (Deana et al., 2008) pour celle de *E. coli*.

Même sans avoir identifié la ou les RppH responsables de la déphosphorylation des ARN de *S. aureus*, nous avons néanmoins pu montrer que les ARN dont le premier ribonucléotide est une guanosine sont préférentiellement 5'P. Ces résultats contrastent avec les observations faites pour *Ec*RppH et *Bs*RppH, qui déphosphorylent préférentiellement les ARN commençant par une adénosine plutôt que par une guanosine. Il est toutefois important de noter que les préférences relatives de *Bs*RppH et *Ec*RppH ont été démontrées *in vitro*, tandis que nos résultats proviennent de données *in vivo*. Cela signifie que d'autres facteurs peuvent intervenir dans la régulation des états de phosphorylation des extrémités 5' des ARN. Par exemple, il est possible que la préférence apparente pour une guanosine en première position soit due à une aversion de la RNase J pour les

ARN commençant par une guanosine, entraînant une augmentation relative de la proportion d'extrémités 5'P des ARN 5'-G en raison d'une dégradation réduite. Nos analyses dans le chapitre 1 n'ont pas identifié de telles préférences, mais si cet effet est faible, il pourrait être difficilement détectable. En revanche, à l'instar de *Bs*RppH et *Ec*RppH, les di-nucléotides GR sont préférentiellement 5'P chez *S. aureus*. Chez *B. subtilis*, *Bs*RppH montre une préférence presque exclusive pour les ARN présentant une guanosine en deuxième position. Cette préférence se manifeste également, bien que de manière moins exclusive, chez *S. aureus*, par une préférence pour le di-nucléotide GG par rapport aux autres possibilités. Cela suggère que *Bs*RppH fonctionne différemment de la RppH de *S. aureus* ou bien qu'il existe plusieurs RppH chez *S. aureus*, dont l'une serait plus permissive quant à la composition du deuxième ribonucléotide. Les sept souches mutantes pour les NUDIX étudiées dans ce chapitre présentent toutes une proportion d'extrémités 5'P plus élevée pour les ARN 5'-G, 5'-GR et 5'-GG, indiquant que les NUDIX SA0425, SA1612, SA2278 et SA1330 ne seraient pas responsables de ces observations.

Dans le prochain chapitre nous présentons le développement d'EMOTE-tk, une librairie R libre et gratuite, disponible sur le gitlab du CNRS. Cette librairie intègre les méthodes statistiques et bio-informatique que nous avons développé au cours de ce travail pour adapter les traitements aux protocoles expérimentaux originaux de nos collaborateurs. Cette partie permettra de détailler les concepts et les analyses effectué dans la 1ère et la 2ème partie en lien avec les données issues de séquençage des extrémités 5' des ARN. Nous y aborderons également les besoins, les choix de conception ainsi que la mise en œuvre.

2.3 Chapitre 3 : EMOTE toolkit : une librairie R pour l'analyse de données de séquençage des extrémités des ARN

Dans les parties précédentes, nous avons étudié les effets de la RNase J et des protéines candidates NUDIX RppH sur le transcriptome de *S. aureus in vivo*, notamment à partir de séquençage RNA-Seq des extrémités 5' des ARN. Contrairement aux analyses conventionnelles de séquençage à haut débit, qui impliquent le comptage des lectures alignées par plages génomiques lors de la quantification des ARN, l'étape de quantification d'une analyse de séquençage de leurs extrémités ne prend en compte que les lectures alignées à la même position (extrémité 5' ou 3'). La revue de (Broglia et al., 2023) donne un aperçu global et fournit une description détaillée des différents protocoles de préparation des librairies RNA-Seq permettant de caractériser et de quantifier les extrémités 5' ou 3', notamment afin d'étudier en détails l'action des exoribonucléases, des endoribonucléases ainsi que des RppH chez les bactéries.

Au cours de ce travail de thèse, nous avons travaillé à partir de multiples fichiers FASTQ de lectures, produits à partir de librairies 5'end RNA-Seq. La plupart de ces analyses bio-informatiques a été réalisée avec R et des librairies de BioConductor. Les raisons ayant orienté ce choix sont que (i) R est un environnement pour faire des statistiques, et nous en réalisons, (ii) la multitude de librairies R/BioConductor déjà disponibles pour des analyses bio-informatiques, notamment de RNA-Seq, et (iii) afin de rendre accessible les scripts et traitements réalisés à notre équipe collaboratrice, dirigée par Peter Redder, qui est à l'aise avec ce langage et cet environnement. Néanmoins, malgré l'abondance de librairies disponibles dans l'écosystème de R/BioConductor, nous n'en avons pas trouvé qui permette de réaliser certaines étapes des analyses que nous avions à effectuer, et qui sont communes aux traitements à réaliser sur des librairies RNA-Seq provenant de protocoles de préparation pour séquencer les extrémités 5' ou 3' des ARN. Au fur et à mesure de l'avancée de nos travaux, nous avons donc développé un ensemble de fonctions permettant de réaliser les traitements nécessaires. Cet ensemble de fonctions s'est enrichi selon les questions que nous nous sommes posées, et, petit à petit, il nous est apparu pertinent de les rassembler et de les organiser au sein d'une librairie spécialisée sur les traitements de librairies RNA-Seq d'extrémités des ARN afin de les rendre disponibles à la communauté.

Dans cette partie, nous allons dans un premier temps décrire la conception et le développement de cette librairie et des fonctionnalités qu'elle propose pour répondre aux besoins que nous avons identifiés.

Dans un deuxième temps, nous présenterons dans la partie réalisation comment ces fonctionnalités ont été implantées en illustrant leur utilisation et leur application pour produire les résultats présentés dans cette thèse.

Nous allons nous focaliser sur les parties concernant le traitement de librairies RNA-Seq générées pour les extrémités 5' des ARN. Certaines fonctions de la librairie sont spécifiques aux extrémités 3' (alignement sur les séquences de référence avec un autre programme et quantification en 3' plutôt qu'en 5'), et la librairie a été conçue pour en tenir compte, mais nous ne les aborderons pas ici, car elles ont été développées par un autre membre de l'équipe dans le cadre d'une autre étude, et plus particulièrement parce qu'elles n'ont pas été utilisées au cours de ce travail.

2.3.1 Analyse

Dans cette partie, nous allons commencer par recenser les besoins à travers des cas d'utilisation typiques que nous avons été amenés à réaliser au cours de cette thèse dans nos analyses.

Cas d'utilisation n°1 : Identification des TSS

Bien que nous n'ayons pas présenté ce type d'analyse dont nous avons utilisé les résultats dans le chapitre 1, nous avons été amenés à les réaliser. Les analyses bio-informatiques débutent à partir d'un fichier de données FASTQ généré, par exemple, avec le protocole TSS-EMOTE (Prados et al., 2016). Dans ce fichier, plusieurs conditions expérimentales sont multiplexées et les lectures de chaque condition expérimentale sont distinguées grâce à l'incorporation d'un code-barre lors de la préparation de la libraire RNA-Seq. Les deux conditions expérimentales ont été réalisées avec des réplicats. La première condition correspond à un échantillon dans lequel les extrémités des ARN 5'PPP sont séquencées. La deuxième condition correspond au contrôle dans lequel aucune extrémité 5' n'est attendue.

A partir du fichier FASTQ, les étapes de l'analyse sont les suivantes :

1. **Contrôle qualité** (FastQC) des lectures contenues dans le fichier FASTQ.
2. **Validation** du contenu des lectures (code-barre, séquence de contrôle, UMI, séquence de reconnaissance) et démultiplexage des lectures correspondant aux conditions expérimentales dans des fichiers FASTQ séparés. La sortie correspond à des fichiers FASTQ ; un par condition.

3. **Alignement** des lectures sur le génome de la souche étudiée. La sortie correspond à des fichiers BAM ; un par condition.

4. **Quantification** des lectures alignées à chaque position du génome correspondant aux extrémités 5' des ARN. La sortie correspond à une table de comptages par condition et par réplicat.

5. **Élimination des duplicats de PCR**. La sortie correspond à la table de comptages mise à jour.

6. **Identification** des positions génomiques correspondant aux TSS avec l'utilisation d'un test statistique basé sur une distribution béta-binomiale permettant d'identifier les différences significatives entre les comptages à une position génomique correspondant aux répétitions de la condition 5'PPP et les comptages correspondant aux répétitions de la condition contrôle

Cas d'utilisation n°2 : Etude de la dégradation co-traductionnelle de la RNase J1

Dans le chapitre 1, à partir d'un fichier de lectures FASTQ généré avec le protocole EMOTE, nous avons pu déterminer que la RNase J1 avait une activité de dégradation 5' exoribonucléasique qui s'effectue de manière co-traductionnelle, c'est-à-dire qu'elle dégrade les ARN en cours de traduction, en suivant le dernier ribosome. Cette activité se manifeste par une périodicité dans les abondances des extrémités 5' des ARN des intermédiaires de dégradation. Les lectures multiplexées dans le fichier FASTQ de départ correspondent à deux souches avec deux réplicats : la souche sauvage et la souche $J1^{AGA}$. Les lectures correspondent aux ARN ayant une extrémité 5'P.

A partir du fichier FASTQ, les étapes ci-dessous de l'analyse sont les suivantes (les étapes 1 à 5 sont les mêmes que dans le cas d'utilisation n°1) :

1. **Contrôle qualité** (FastQC) des lectures contenues dans le fichier FASTQ.

2. **Validation** du contenu des lectures (code-barre, séquence de contrôle, UMI, séquence de reconnaissance) et démultiplexage des lectures correspondant aux conditions expérimentales dans des fichiers FASTQ séparés. La sortie correspond à des fichiers FASTQ ; un par condition.

3. **Alignement** des lectures sur le génome de la souche étudiée. La sortie correspond à des fichiers BAM ; un par condition.

4. **Quantification** des lectures alignées à chaque position du génome correspondant aux extrémités 5' des ARN. La sortie correspond à une table de comptages par condition et par réplicat.

5. **Élimination des duplicats de PCR**. La sortie correspond à la table de comptages mise à jour.

A partir de ces tables de comptage, trois traitements sont effectués selon les régions génomiques considérées.

A) Pour les **profils de méta-comptages** autour du codon de départ de la traduction :

6. A partir d'une annotation des positions de départ de traduction de l'ensemble des gènes codants pour des protéines, récupération des comptages dans une fenêtre de 100 nucléotides en amont et en aval de ces positions. La sortie correspond aux comptages sur chacune des fenêtres.

7. **Normalisation** des comptages par fenêtre.

8. **Agrégation** des profils normalisés en profils de **méta-comptages** pour les deux conditions sauvage et $J1^{AGA}$.

B) Pour l'**étude de la périodicité** le long des séquences codantes au moyen d'une transformée de Fourier :

6. A partir d'une annotation des positions de départ et de fin de traduction de l'ensemble des gènes codants pour des protéines, les comptages sont récupérés pour chacun des gènes. Ceux-ci ayant des longueurs variables, seulement les comptages correspondant au début des gènes sont conservés de manière à avoir des comptages pour au moins 75 % des gènes (les comptages des gènes plus courts sont conservés et une valeur de 0 comptage est utilisée lorsque le gène est trop court).

7. **Normalisation** des comptages par fenêtre.

8. **Agrégation** des profils normalisés en profils de méta-comptages pour les deux conditions.

9. **Analyse de la périodicité** du signal par transformée de Fourier sur les deux profils de méta-comptage obtenus.

C) Pour l'étude de la **préférence de phase de traduction** :

6. A partir d'une annotation des positions de départ et de fin de traduction de l'ensemble des gènes codants pour des protéines, les comptages sont récupérés pour chacun des gènes.

7. **Normalisation** des comptages par fenêtre (correspondant chacune à un gène).

8. **Agrégation** des comptages par cadre de lecture et par gène.

9. **Agrégation** des comptages de tous les gènes.

Cas d'utilisation n°3 : Détermination des proportions des formes mono- et tri-phosphorylées des extrémités 5' des ARN et analyse différentielle de proportions

Dans les deux premiers chapitres, nous avons pu estimer les proportions des états de phosphorylation des extrémités 5' des ARN à partir des librairies de séquençage RNA-Seq obtenues avec le protocole PQ-EMOTE. Pour chaque souche et réplicat, nous disposons de deux conditions expérimentales fournissant des lectures correspondant aux extrémités 5'P ($pool^{mono}$) pour l'une, et aux extrémités 5'P et 5'PPP ($pool^{total}$) pour l'autre. Après calcul des proportions de formes 5'P de chaque extrémité 5', nous avons comparé ces proportions par extrémité 5' entre une souche sauvage et une souche délétée d'un ou plusieurs gènes (candidats NUDIX RppH).

A partir du fichier FASTQ, les étapes de ces analyses par couple souche sauvage-souche mutante sont les suivantes (les étapes 1 à 4 sont les mêmes que dans les cas d'utilisation n°1 et n°2) :

1. **Contrôle qualité** (FastQC) des lectures contenues dans le fichier FASTQ.

2. **Validation** du contenu des lectures (code-barre, séquence de contrôle, UMI, séquence de reconnaissance) et démultiplexage des lectures correspondant aux conditions expérimentales dans des fichiers FASTQ séparés. La sortie correspond à des fichiers FASTQ ; un par condition.

3. **Alignement** des lectures sur le génome de la souche étudiée. La sortie correspond à des fichiers BAM ; un par condition.

4. **Quantification** des lectures alignées à chaque position du génome correspondant aux extrémités 5' des ARN. La sortie correspond à une table de comptages par condition et par réplicat.

5. **Désaturation** des comptages de chaque extrémités 5' à partir des comptages de lectures alignées, des UMI et de leur fréquence d'apparition. La sortie correspond à la table de comptages désaturés par condition et réplicat.

6. **Filtrage** des extrémités avec des comptages trop faibles dans les échantillons $pool^{total}$.

7. **Normalisation** des comptages entre le $pool^{mono}$ et le $pool^{total}$ de chaque condition et réplicat à partir d'un ensemble de contrôle (rRNA et sites clivages de la RNase Y).

8. **Utilisation de la librairie DSS** pour identifier les extrémités dont les proportions sont différentes de manière statistiquement significative entre la souche sauvage et la souche mutante.

Cas d'utilisation n°4 : Analyse d'expression différentielle pour l'identification de sites de clivage endoribonucléolytique de la RNase J1

A la fin du chapitre 1, nous avons proposé une analyse d'expression différentielle entre la souche sauvage et le mutant $J1^{AGA}$ afin de proposer des sites potentiels de clivage endoribonucléolytique par la RNase J1. Les données utilisées correspondent à des échantillons pour lesquels nous disposons (avec des réplicats) des extrémités 5'P pour la souche sauvage et pour la souche mutante $J1^{AGA}$.

A partir du fichier FASTQ, les étapes de cette analyse sont les suivantes (les étapes 1 à 5 sont les mêmes que dans les cas d'utilisation n°1 et n°2) :

1. **Contrôle qualité** (FastQC) des lectures contenues dans le fichier FASTQ.

2. **Validation** du contenu des lectures (code-barre, séquence de contrôle, UMI, séquence de reconnaissance) et démultiplexage des lectures correspondant aux conditions expérimentales dans des fichiers FASTQ séparés. La sortie correspond à des fichiers FASTQ ; un par condition.

3. **Alignement** des lectures sur le génome de la souche étudiée. La sortie correspond à des fichiers BAM ; un par condition.

4. **Quantification** des lectures alignées à chaque position du génome correspondant aux extrémités 5' des ARN. La sortie correspond à une table de comptages par condition et par réplicat.

5. **Élimination des duplicats de PCR**. La sortie correspond à la table de comptages mise à jour.

6. **Filtrage** des valeurs trop faibles attribuées à des intermédiaires de dégradation plutôt qu'à des sites de clivage endoribonucléolytique.

7. **Utilisation de la librairie EdgeR** pour identifier les extrémités 5' dont les abondances diffèrent entre la souche sauvage et la souche mutante $J1^{AGA}$.

Analyse de l'existant et besoins

Les quatre cas d'utilisation présentés font ressortir des étapes communes, notamment dans le prétraitement des fichiers FASTQ de départ afin de pouvoir réaliser les alignements sur le génome de référence, mais dont les détails peuvent varier. Par exemple, pour la validation des lectures et leur répartition dans des fichiers FASTQ pour chaque condition multiplexée pour être prêtes à être alignées sur le génome de référence, la structure des lectures attendues peut changer : code-barres des échantillons multiplexés, longueur des UMI, ou encore présence de séquences poly A (comme dans les échantillons présentés de *H. pylori* dans le chapitre 2).

Des programmes comme *Trimmomatic* et *Cutadapt* permettent de tronquer les lectures afin de retirer divers éléments, y compris des régions poly A. Cependant, ces méthodes ne permettent pas de tester la présence et la validité des éléments attendus. *Cutadapt* permet également de démultiplexer un fichier FASTQ à partir d'une liste de code-barres à utiliser et donc de répartir les lectures dans différents fichiers en fonction de ces code-barres identifiés. Pour la prise en compte des UMI, les programmes tels que *rmdup* et *Markduplicate* ne sont pas adaptés. En effet, *rmdup* supprime directement les lectures correspondant à des duplicats PCR, ce qui n'est pas possible pour utiliser notre méthode de désaturation. *MarkDuplicate* permet d'étiqueter les duplicats et donc de les conserver mais son fonctionnement ne permet la prise en compte que des UMI-duplex (deux chaînes de même longueur séparées par quelques nucléotides), ce qui est trop spécifique pour nos besoins.

Nous nous sommes interrogés sur l'utilisation de ces programmes dans la librairie. Etant donné que nous souhaitions pouvoir générer des rapports concernant les erreurs identifiées dans les lectures (est-ce la séquence de reconnaissance ou la séquence de contrôle qui n'est pas valide, ou bien les deux ?), nous avons décidé de réaliser par nous-mêmes ces fonctionnalités qui ne semblaient pas très complexes à mettre en œuvre *a priori*. De plus, nous souhaitons que l'utilisateur puisse installer notre librairie pour réaliser l'ensemble des traitements à partir de R (et non à partir de la ligne de commande) et n'ait pas à installer des programmes en dehors de R, notamment par rapport au système d'exploitation utilisé. Les opérations nécessaires ici devraient être réalisables avec la librairie *ShortRead* afin d'accéder au contenu d'un fichier FASTQ,

Une fois les lectures validées et débarrassées des séquences ajoutées pendant la préparation de la librairie pour le séquençage, les lectures sont alignées sur le génome de référence. Plusieurs librairies BioConductor permettent d'utiliser des programmes d'alignement à partir de R (*Rbowtie*, *Rbowtie2*, *Rsubread*, *Rhisat2*, …), et cette étape devrait donc fonctionner sur tous les systèmes d'exploitation. A partir des fichiers contenant les alignements (fichiers BAM), seulement la position du premier nucléotide des alignements est considérée pour produire les tables de comptages des extrémités des ARN sur le génome de référence. Les fichiers BAM peuvent être analysés avec la librairie *Rsamtools*. Pour les comptages obtenus sous forme de table, nous avons choisi la méta-librairie *tidyverse* pour la lisibilité et la maintenance du code. *tidyverse* inclut *tidyr* et *dplyr* permettant de nombreuses opérations sur les tables (*tibbles*), *ggplot2* pour générer des graphiques pour la visualisation, et *magrittr* pour la lisibilité du code. Après cette étape de quantification, les comptages peuvent nécessiter des post-traitements plus spécifiques que nous avons à implanter :

- Élimination des duplicats de PCR ou désaturation, si les lectures contenaient des UMI.
- Filtrage des positions génomiques pour lesquelles les comptages sont trop faibles dans la condition $pool^{total}$ pour déterminer les proportions d'extrémités 5'P. Dans ce cas, il s'agit d'éliminer également ces positions génomiques des tables de comptages de la condition $pool^{mono}$ correspondante.
- Normalisation des comptages sur une fenêtre donnée avant d'agréger les profils pour des analyses par transformée de Fourier ou de préférence de phase de traduction. Les fenêtres à considérer peuvent être déterminées à partir de l'annotation du génome (les régions codantes pour des protéines).
- Normalisation des comptages entre les conditions $pool^{mono}$ et $pool^{total}$ appariées à partir d'un ensemble de contrôle pour estimer les proportions d'extrémités 5'P.

Ensuite, une fois les tables de comptage post-traitées, différents types d'analyse peuvent être effectués :

- Identification des TSS : pour cela, nous utiliserons la librairie *VGAM* fournissant la distribution béta-binomiale, et la librairie *survcomp* nous permettra de combiner les p-valeurs obtenues pour les différents réplicats.
- Analyse des profils des intermédiaires de dégradation : la librairie *stats* fournit l'analyse par transformée de Fourier (fonction *fft*).
- Analyse différentielle de proportions 5'P des TSS : la librairie *DSS* sera utilisée pour déterminer des différences statistiquement significatives de proportions en tenant compte des comptages et des réplicats.
- Identification de site de clivage endoribonucléolytique : la librairie *EdgeR* permet de faire ce type d'analyse différentielle.

Dans la section qui suit, nous présentons des choix de conception pour la réalisation des fonctionnalités nécessaires aux étapes des analyses présentées ci-dessus.

2.3.2 Conception

2.3.2.1 Pré-traitement d'un fichier FASTQ avant l'alignement des lectures sur un génome

La plupart des protocoles incorpore, lors de la préparation des librairies pour le séquençage, des informations qui seront contenues dans les lectures. Ces informations sur la structure attendue d'une lecture peuvent être décrites comme une liste d'éléments attendus. Ces éléments peuvent être des types suivants :

- une séquence donnée à une certaine position des lectures, par exemple une séquence de contrôle, en d'autres termes une chaîne de caractère à une position.
- un ensemble de séquences possible à une certaine position, par exemple le code-barre d'un échantillon, en d'autres termes une liste de chaînes de caractères autorisées à une certaine position.

- une séquence aléatoire formée de certains nucléotides sur une région de la lecture, par exemple un UMI ou la partie de la lecture correspondant à la séquence de l'extrémité ARN à aligner sur le génome de référence.
- une expression régulière à rechercher sur une région de la lecture, par exemple une séquence poly A.

Bien que tous ces éléments puissent être décrits par des expressions régulières, le traitement de chaînes constantes ou formés à partir d'un certain alphabet est réalisé beaucoup plus rapidement que la recherche d'une expression régulière. Cette recherche devant être réalisée dans la plupart des cas sur des millions de lectures, une petite économie de calcul et de temps sur chaque lecture pourra se révéler appréciable. De plus, pour la recherche de chaînes de caractères constantes, on pourra autoriser jusqu'à un certain nombre de différences entre la chaîne attendue et la chaîne observée. Ceci permettra notamment d'autoriser une différence, par exemple, sur une séquence de contrôle qui devrait très probablement correspondre à une erreur de lecture par le séquenceur, si tous les autres éléments de la lecture sont valides.

Une fois ces éléments validés dans une lecture, la région correspondant à la séquence de l'extrémité ARN sera écrite dans un fichier FASTQ contenant les lectures valides à aligner sur le génome de référence. En cas de démultiplexage, les code-barres présents sur la lecture permettront de générer un fichier FASTQ par code-barre.

Dans le cas où les lectures contiennent des UMI et que l'utilisateur souhaite réaliser une désaturation, il est nécessaire de conserver la séquence UMI associée à chaque lecture. En effet, cette information sert généralement à éliminer les duplicats de PCR, et bien qu'il soit possible d'éliminer les lectures ayant la même séquence ARN et le même UMI (et éventuellement le même code-barre en cas de multiplexage) dès cette étape, nous proposons de transférer cette information dans l'identifiant de la lecture (*readid*). En effet, nous nous en servirons pour la désaturation des séquences des UMI associées aux lectures alignées sur la référence pour calculer la distribution des fréquences d'utilisation des UMI. Cette distribution des fréquences d'UMI sert ensuite à déterminer si le nombre de fois qu'un UMI est associé à une extrémité ARN est plutôt issu de duplicat PCR ou de la saturation à cause d'un nombre trop faible d'UMI différents possibles.

Nous proposons à l'utilisateur de fournir la liste des éléments sous la forme d'une table dans laquelle chacune des lignes décrit un élément avec les paramètres suivants :

- ***name*** pour le nom de l'élément
- ***start*** pour la position de début
- ***width*** pour la longueur de l'élément
- ***pattern_type*** pour le type de l'élément : ce type pouvant être :
    - ***constant*** pour une chaîne de caractères constante ou une liste de chaînes de caractères constante
    - ***nucleotides*** pour une séquence aléatoire de nucléotides formés à partir d'une liste de nucléotides autorisés
    - ***regexp*** pour une expression régulière
- ***pattern*** pour la ou les chaînes de caractères spécifiant l'élément attendu. Par exemple, "ACG" pour l'alphabet autorisé des UMI, "ACAGA" pour la séquence de contrôle, ou encore "AAAAA.*" pour l'expression régulière correspondant au poly A
- ***max_mismatch*** pour le nombre d'erreurs autorisées. Ceci n'est pas possible pour l'expression régulière.
- ***readid_prepend*** pour indiquer si la région de l'élément (dans notre cas, l'UMI présent sur la lecture) doit être conservée, c'est-à-dire transférée en tant que préfixe de l'identifiant de la lecture.

Par exemple, pour les lectures de la librairie PQ-EMOTE 49 utilisée dans le chapitre précédent, la table avec les éléments est donnée en Table 16. Le dernier élément nommé *readseq* correspond à la séquence à conserver et qui sera alignée sur la référence.

A partir de ces éléments, on pourra réaliser différentes tâches sur un fichier FASTQ. Tous les éléments ne sont pas nécessaires. Il est par exemple possible de seulement valider (seulement des A, T, C ou G) et raccourcir les lectures pour ne garder que les 30 premiers nucléotides de chaque lecture.

name	start	width	pattern_type	pattern	max_mismatch	readid_prepend
barcode	2	4	constant	ACGT, TATA, …	0	FALSE
recog.seq	5	15	constant	CGGCACC…	1	FALSE
UMI	21	7	nucleotides	ACG	0	TRUE
control.seq	28	5	constant	ACAGA	0	FALSE
readseq	33	18	nucleotides	ATCG	0	FALSE

Table 16 : Table décrivant la structure des lectures PQ-EMOTE contenant l'oligonucléotide Rp8.

2.3.2.2 Alignement des lectures sur un génome de référence

L'alignement des lectures est une étape clé dans une analyse de type RNA-Seq. Le ou les fichiers FASTQ obtenus à l'étape précédente contiennent uniquement les parties correspondant aux transcrits à analyser, et les lectures vont donc pouvoir être alignées sur un génome de référence. Notre librairie se concentre sur l'analyse de données « short-read sequencing », c'est-à-dire que les fragments d'ARN séquencés, et donc les fragments à aligner sur le génome font entre 15 et 300 nucléotides de long. Du fait de l'absence d'exons/introns chez les bactéries, et du fait que les lectures ont été nettoyées des éléments ajoutés lors de la préparation de la librairie et qu'elles sont donc prêtes à être alignées telles quelles sur la référence, nous souhaitons obtenir des alignements parfaits (100 % d'identité sans indel) ou presque parfait (un certain nombre de mésappariements est autorisé mais sans indel).

Les programmes d'alignements produisent en général les alignements dans des fichiers SAM ou BAM. L'étape suivante de quantification des extrémités se fera à partir de ce type de fichiers. L'utilisateur pourra donc utiliser le programme de son choix. Néanmoins, nous avons choisi d'en intégrer un, Bowtie (Langmead et al., 2009), pour lequel une librairie R/BioConductor (Rbowtie Hahne et al.**)** est disponible pour les systèmes d'exploitations Linux, Windows et MacOS.

Ce choix permettra notamment de proposer une fonction plus simple à utiliser pour l'utilisateur car les paramètres adéquats pour Bowtie pour ce type d'analyses seront paramétrés pour lui, et il pourra seulement spécifier le nombre de mésappariements autorisés.

2.3.2.3 Quantification des extrémités 5' à partir des alignements

L'originalité du séquençage des extrémités 5' des ARN réside dans le fait que seule la première position alignée est utilisée, car elle correspond à l'extrémité 5' exacte de la molécule d'ARN. Elle peut donc être définie seulement par sa position sur la séquence de référence sur laquelle elle est alignée.

Par rapport aux analyses habituelles de RNA-Seq qui vont considérer pour chaque intervalle génomique (par exemple les gènes), le nombre de lectures alignées qui la ou le chevauchent, la représentation nécessaire dans notre cas est plus simple. En effet, à partir des alignements, il suffit de raccourcir l'alignement à sa première position alignée (l'extrémité 5' de l'ARN). Ceci implique que tous les alignements ne font plus qu'un seul nucléotide de long et qu'il n'est plus nécessaire de considérer des intervalles, notamment pour déterminer s'ils se chevauchent.

Le résultat de l'étape de quantification consiste donc en une table de comptages par position exacte (par extrémité 5').

Selon la préparation de la librairie, il peut ne pas y avoir d'UMI. Si tel est le cas, les comptages correspondent aux nombres de lectures alignées à la même position de la séquence de référence. Si les lectures contenaient des UMI (ajoutés à l'identifiant de la lecture), alors l'utilisateur pourra choisir les comptages correspondant aux nombre d'UMI différents plutôt qu'au nombre de lectures alignées. Un exemple est donné dans la Table 17 ci-dessous. La première ligne de cette table indique que dans l'échantillon WT_rep1, il y a 67 lectures dont la première position s'aligne à la position 313 sur le brin + de la séquence de référence (chromosome CP010890.1). Pour cette position, seulement 34 UMI différents ont été retrouvés dans les lectures alignées.

sample_id	rname	position	strand	reads	count
WT_rep1	CP010890.1	313	+	67	34
WT_rep2	CP010890.1	313	+	72	44
WT_rep1	CP010890.1	3650	+	107	58
WT_rep2	CP010890.1	3650	+	31	21
WT_rep1	CP010890.1	4128	+	60	30
WT_rep1	CP010890.1	4311	+	31	11

Table 17 : Table de comptages avec le nombre de lectures alignées (reads) et le nombre d'UMI uniques correspondants (count).

Dans le cas où une désaturation est nécessaire, il est possible de conserver les séquences d'UMI dans la table de comptages afin de déterminer leur fréquence d'utilisation. Un exemple est donné dans la Table 18 ci-dessous. La deuxième ligne indique que dans l'échantillon avec le code-barre AAGT, il y a 5 lectures avec le même UMI ACGACAA qui s'alignent en position 313 sur le brin + du chromosome CP010890.1.

sample_id	rname	UMI	position	strand	count
WT_rep1	CP010890.1	AGGGACG	312	+	1
WT_rep1	CP010890.1	ACGACAA	313	+	5
WT_rep1	CP010890.1	ACGACGC	313	+	5
WT_rep1	CP010890.1	ACGCGAA	313	+	2
WT_rep1	CP010890.1	ACGGAGA	313	+	6
WT_rep1	CP010890.1	CAGGACG	313	+	1
WT_rep1	CP010890.1	CGGACGG	313	+	1

Table 18 : ***Table de comptages par séquence d'UMI (count).***

Les fichiers BAM fournis pour la quantification auront pu être générés en utilisant Rbowtie intégré dans la librairie, ou bien ils auront pu être générés par l'utilisateur indépendamment de notre librairie.

La table de comptages est un élément central de la librairie. En effet, c'est à partir de celle-ci que l'utilisateur pourra effectuer les analyses évoquées précédemment comme l'identification des TSS, ou la recherche d'une période dans les profils de dégradation co-traductionnelle. Néanmoins, avant cela, il est souvent nécessaire d'effectuer des étapes filtrage et/ou de normalisation des comptages par échantillon et/ou entre échantillons. Le format des données en entrée et en sortie de ces étapes est le même. Il s'agit d'une table de comptages avec au minimum les colonnes suivantes : identifiant de l'échantillon (*sample_id*), identifiant de la séquence de référence (*rname*), position, brin (*strand*), comptage (*count*).

2.3.2.4 Post-traitements des tables de comptages

Dans cette section, nous détaillons certaines opérations de post-traitement spécifiques aux analyses que nous avons effectuées.

2.3.2.4.1 Désaturation

Les protocoles de préparation de librairies de RNA-Seq incluent souvent une étape de PCR afin d'amplifier la quantité d'ADNc pour le séquençage, et aussi sélectionner les fragments correctement liés aux adaptateurs. En général, si plusieurs lectures contiennent exactement la même séquence, celles-ci sont considérées comme résultant de l'étape de PCR, et une seule lecture est conservée. Pour du séquençage des extrémités ARN, beaucoup d'ARN correspondant au même site de départ de transcription auront la même séquence, notamment à leur extrémité 5'. Dans ce cas, on s'attend, et on cherche même à quantifier le nombre de lectures correspondant à la même extrémité 5'. Pour identifier et éliminer les duplicats de PCR, les séquences UMI sont ajoutées pendant la préparation de la librairie RNA-Seq. Or, plus les séquences UMI sont courtes, et plus le nombre de séquences UMI différentes est petit. Nous avons vu également dans le chapitre 2 que leur fréquence d'utilisation n'est pas uniforme. Par exemple, pour des UMI de 7 nucléotides (A, C ou G) de long, il y a 2 187 séquences UMI différentes, et dans les analyses effectuées dans le chapitre 2, on s'attend à observer une saturation de l'ordre de 5 % à partir de 115 lectures correspondant à la même extrémité 5'.

Dans un premier temps, il s'agit donc de déterminer la distribution des fréquences d'utilisation des UMI à partir des lectures alignées associées à leur UMI. Pour cela, nous utilisons la table de comptages qui contient les UMI (Table 18). Afin de déterminer le nombre d'occurrences de chaque UMI, il ne faut compter chaque occurrence que si l'on est certain qu'elle ne provient pas d'un duplicat de PCR. Pour cela, nous ne conservons parmi les extrémités alignées, que celles pour lesquelles le nombre de séquences d'UMI différentes correspond au nombre de lectures. Ainsi, nous sommes certains qu'il n'y a eu ni duplication, ni réutilisation de la même séquence UMI avec la même extrémité 5'. Après ce filtrage, nous pouvons agréger les nombres d'occurrences des UMI, toutes extrémités 5' confondues. A partir du nombre d'occurrences de chaque UMI, nous en déduisons sa fréquence d'utilisation qui nous servira de probabilité d'apparition dans une lecture.

La méthode de désaturation que nous proposons repose sur le modèle suivant basé sur une distribution binomiale. Pour une séquence correspondant à l'extrémité 5' d'un ARN, un certain nombre de lectures est observé, par exemple 506 (comme dans l'exemple du chapitre 2). Ces 506 lectures sont considérées comme 506 tirages aléatoires avec remise d'un UMI dans une population avec une probabilité de tirage associée à chaque UMI. Parmi les 506 lectures, 3 sont associées au même UMI, l'UMI_C qui a une probabilité de tirage $p(UMI_C) = 0{,}0009$. Ceci se traduit sur une distribution binomiale pour 506 tentatives avec une probabilité de succès de 0,0009 par les valeurs suivantes :

succès (x)	0	1	2	3	4
$p(X = x)$	0,6341	0,2890	0,0657	0,0099	0,0011
$p(X > x)$	0,3659	0,0769	0,0112	0,0012	0,0001
$p(X \geq x)$	1	0,3659	0,0769	0,0112	0,0012

Table 19 : Distribution binomiale pour 0 à 4 succès sur 506 tentatives avec une probabilité de succès égale à 0,0009.

A partir d'un certain seuil α, que l'utilisateur pourra choisir, par exemple 5 %, nous pouvons déterminer que la probabilité que le nombre d'utilisations de l'UMI_C avec 506 tirages soit de 3 ou plus n'est acceptable car $p(X \geq 3) < \alpha$. Par contre, elle est acceptable pour $p(X \geq 2)$.

Ainsi, pour cette extrémité 5' pour l'UMI_C, nous corrigeons sont nombre d'utilisations à 2 (au lieu de 3) car il est très improbable que la 3ème lecture observée avec cet UMI provienne d'une réutilisation de l'UMI_C.

Cette correction est appliquée pour tous les UMI d'une extrémité 5'. Ceci a pour effet de diminuer le comptage associé à l'extrémité 5'. Pour cet exemple, on passe de 506 à 505, mais il aurait pu diminuer davantage si d'autres UMI avaient été corrigés. Le nombre de tentatives ayant diminué, il se peut que certains UMI soient corrigés lors d'une seconde itération si le nombre de succès observés s'avère trop important par rapport à un nombre de tentatives (505) moins grand. Ceci se traduit par l'algorithme suivant :

Entrées :

Soit *t* la table de comptages

Soit *p* la table de fréquences d'utilisation des UMI

Sorties :

t tables des comptages désaturés

Début

Pour chaque extrémités 5' *pos* **dans** *t* :

Tant qu'il y a des changements :

Soit *n* les comptages des UMI à la position *pos*

n_total ← sum(*n*)

Pour chaque UMI *u* **dans** *n* :

Soit *n_max* le nombre maximum tel que pbinom($X \geq n_max$, *n_total*, *p*[*u*]) > α

Si *n_max* < 1 **alors** *n_max* ← *1*

Si *n*[*u*] > *n_max* **alors** *n*[*u*] ← *n_max*

Fin

Algorithme 1 *:* ***Algorithme proposé pour la désaturation des*** *UMI à partir de la table de comptages contenant les séquences UMI (exemple Table 18) et des fréquences d'utilisation des UMI.*

2.3.2.4.2 Filtrage et normalisation des comptages dans des échantillons appariés

Dans le cas d'utilisation n°3, des échantillons pool^{mono} (5'P) et pool^{total} (5'P et 5'PPP) sont appariés pour déterminer la proportion de forme 5'P des extrémités des ARN.

Si pour une extrémité 5', le nombre de comptages observé est de 10 dans le pool^{total}, alors une erreur de 1 sur le comptage dans le pool^{mono} résulte en une erreur de 10 % sur la proportion calculée. Pour limiter le nombre d'erreurs dans les analyses subséquentes dues à une trop faible abondance dans le pool^{total}, l'utilisateur pourra filtrer les extrémités dans les deux échantillons appariés à partir d'un seuil fixé sur les abondances du pool^{total}.

Après ce filtrage éventuel, une normalisation est nécessaire avant de calculer les proportions d'ARN mono-phosphorylés par extrémité 5' pour ajuster les comptages en fonction de la profondeur de séquençage des échantillons appariés. La méthode présentée dans le chapitre 2 utilise un ensemble de contrôle correspondant aux extrémités 5' des deux échantillons qui sont 100 % mono-phosphorylées (ARNr et sites de clivages endoribonucléolytique de la RNase Y). Cet ensemble de contrôle permet d'estimer un facteur de normalisation à appliquer à l'ensemble des comptages des extrémités 5' afin de les rendre comparables entre les deux échantillons appariés pool^{mono} et pool^{total}.

Ces types de filtrage et de normalisation pourraient s'avérer utiles dans d'autres cas. Pour cela, nous proposons de pouvoir les effectuer à partir d'une table contenant des informations (métadonnées) permettant de décrire les échantillons appariés dans les tables de comptages obtenues.

Selon les opérations, cette table de métadonnées pourra contenir les informations suivantes associées à chaque échantillon qui sera référencé par un identifiant d'échantillon :

- **groupe** : les groupes d'échantillons représentant les différentes conditions expérimentales, par exemple sauvage, mutant de délétion Δ1612, Δ0425, Δ2278, …
- **fraction d'ARN** : indique les ensembles d'ARN quantifiés, par exemple pool^{mono} et pool^{total}
- **réplicat** : le numéro de réplicat

Pour le filtrage, nous pouvons utiliser des opérations pivot sur les tables de comptages à partir des métadonnées : l'approche consiste à rassembler sur la même ligne (format large de la table de comptages) les comptages des deux fractions d'ARN correspondant à chaque paire d'échantillon (pool^{mono} et pool^{total}) ayant même groupe et même réplicat, d'effectuer le filtrage des valeurs trop

faibles sur l'échantillon pool^{total}, puis de revenir au format long initial de la table de comptage où pour chaque ligne, il y a un seul comptage pour un groupe, une fraction d'ARN et un réplicat.

Pour la normalisation des comptages entre deux échantillons appariés tels que le pool^{mono} et le pool^{total}, il s'agit d'une opération similaire réalisable avec des opérations de pivot. A partir du format large (comme précédemment) et d'une liste de d'extrémités 5' étant toujours mono-phosphorylées (ensemble de contrôle), il s'agit dans un premier temps de calculer le facteur de normalisation à appliquer à partir de cet ensemble, puis d'effectuer l'ajustement dans l'un des deux échantillons appariés. La procédure est la suivante :

- **Pivoter** de la table de comptages au format large pour avoir les comptages pour les deux fractions, pool^{mono} et pool^{total}, d'ARN sur la même ligne (même groupe, même réplicat),
- **Calculer** le facteur de normalisation (comptage pool^{total} / comptage pool^{mono}) pour chaque élément de l'ensemble de contrôle,
- **Calculer** le facteur de normalisation global (médiane des facteurs de normalisation),
- **Utilisation** du facteur de normalisation global pour normaliser le pool^{mono},
- **Revenir au format long** initial où pour chaque ligne, il y a un seul comptage pour un groupe, une fraction d'ARN et un réplicat

Cette approche utilisant une table de métadonnées est rencontrée dans des librairies existantes, notamment *limma* très utilisée en transcriptomique. Dans cette dernière, initialement développée pour analyser des données de transcriptomes obtenues par microarray mais qui a depuis été mise à jour pour traiter également des données RNA-Seq, les métadonnées sont fournies avec une table *design_matrix* qui, comme ici, permet de fournir des informations sur les échantillons, notamment les conditions expérimentales et les réplicats.

2.3.2.4.3 Normalisation des comptages sur une fenêtre génomique

Le cas d'utilisation n°2 correspond aux analyses effectuées dans le chapitre 1 sur les intermédiaires de dégradation (extrémités 5'P) observés et causés par l'activité 5' exoribonucléasique de la RNase J1 lorsqu'elle suit le dernier ribosome en cours de traduction d'un ARNm. L'analyse globale agrège les profils de comptages de l'ensemble des gènes codants, par exemple autour du codon de départ de la traduction, pour obtenir un profil de méta-comptages. Avant d'agréger les profils de comptages de chacun des gènes, il est nécessaire de les normaliser afin de ne pas attribuer plus de poids aux gènes les plus transcrits (et dégradés de manière co-traductionnelle).

La normalisation à effectuer consiste à diviser les comptages correspondant à une position dans la fenêtre par le nombre total de comptages dans la fenêtre, en d'autres termes à passer des effectifs aux fréquences relatives à la fenêtre considérée.

Dans les 3 applications présentées dans le cas d'utilisation, la même méthode de normalisation est utilisée mais la manière de spécifier la fenêtre diffère :

- par rapport à une caractéristique de l'annotation du génome, par exemple le premier nucléotide du codon de départ (ou de fin) de la traduction : fenêtre d'une certaine largeur autour des positions de la caractéristique,
- pour la FFT et la préférence de phase de traduction, par rapport à l'ensemble des régions codantes des gènes : début et fin des régions.

Pour ces deux manières de spécifier les fenêtres, il s'agit donc de fournir une annotation du génome qui contient la ou les caractéristiques désirées (ici, les positions de début et de fin des CDS). Ceci peut être représenté par une table d'annotation avec au minimum les informations suivantes : l'identifiant de séquence (chromosome), le brin, la position de début, la position de fin, le type de la caractéristique (CDS, TSS ou autre).

Un ensemble de régions peut être déduit à partir d'une caractéristique telle que :

- pour chaque gène, le 1er nucléotide du 1er codon de la traduction pour une largeur de la fenêtre, ou bien
- pour l'ensemble des gènes, déterminer le 1er quartile *q25* de la distribution de leur longueur (ceci pour avoir au moins 75 % des gènes ayant une séquence suffisamment longue), puis, extraire les profils de comptages de la 1ère position des gènes et de longueur *q25* (pour les gènes trop courts, des zéros sont ajoutés au profil).

Pour effectuer la normalisation pour une région donnée, donc un tuple (chromosome, brin, début, fin), il s'agit donc simplement de diviser les comptages concernés de la table de comptages par la somme des valeurs dans cette région.

Une fois que chaque région a été normalisée, leur agrégation consiste à faire la médiane des comptages à chaque position dans les fenêtres.

2.3.3 Réalisation

L'ensemble des fonctions a pu être réalisé uniquement en R. Nous avons cherché un équilibre pour, à la fois, utiliser les fonctionnalités déjà disponibles dans des librairies existantes (et maintenues), tout en utilisant le minimum de dépendances à d'autres librairies.

Une version publique est disponible sur la forge logicielle GitLab du CNRS à l'adresse https://src.koda.cnrs.fr/emote/emote-tk.

La liste des fonctions principales accessibles à l'utilisateur, avec leur courte description est présentée dans la Table 20.

Pre-processing	
Function	**Description**
EMOTE_read_features	Initializes the EMOTE_read_features table with the minimal feature *readseq* corresponding to the region in the read sequence that is expected to align to a reference sequence. Alternatively, a CSV file can be provided to load the read features table from a file.
EMOTE_add_read_features	Adds a read feature to the EMOTE_read_features *tibble* built with EMOTE_read_features or EMOTE_add_read_feature.
EMOTE_parse_reads	Valid reads are written to a dedicated file and if *keep_invalid_reads* is TRUE, invalid reads are written to fastq_file.invalid.fastq.gz. It also writes a feature report which contains features given in the read features parameter and a parsing report which contains the number of reads with valid read features.

EMOTE_demultiplex	Parses each read to retrieve read features in the sequence and test their validity.
Utility	
Function	**Description**
EMOTE_map_fastq	Takes a FASTQ file and a genome FASTA file. An index is created for the reference if not existing, then reads are mapped onto the given reference.
EMOTE_map_stats	Parses the specified BAM file and returns mapping statistics.
Quantification	
Function	**Description**
EMOTE_load_bam_as_tibble	Returns the content of a BAM file in a *tibble*.
EMOTE_summarise_alignments	Processes positions (and UMIs if specified) and returns counts.
EMOTE_quantify_5prime	Processes a BAM file to compute 5'end counts.
EMOTE_quantify_3prime	Processes a BAM file to compute 3'end counts.
Post-processing	
Function	**Description**

EMOTE_filter_paired_samples	Filters out positions from paired samples with a count value lower than the specified threshold in the pool_total sample.
EMOTE_normalize_paired_samples	Normalizes paired samples based on a control set.
EMOTE_normalize_window_counts	Normalizes counts to frequencies in the specified window.
EMOTE_UMI_frequencies	Computes observed UMI frequencies in a count table having UMIs.
EMOTE_desat_count_table	Removes PCR duplicate reads taking into account the observed frequency of each UMI.
Downstream analysis	
Function	**Description**
EMOTE_TSS_model	Computes a p-value for a single position using the beta-binomial distribution (as in TSS-EMOTE).
EMOTE_betabinom_TSS	Computes adjusted p-values for all positions present in a count table. p-values of replicates are combined.
EMOTE_region_profile	Uses an annotation to retrieve positions around the specified feature (*e.g.* start or stop codons) or the region specified with start, end and strand.
EMOTE_metacount_profile	Retrieves regions for around the position of the specified feature (*e.g.* start codon) based on an annotation table and returns the metacounts profile.

EMOTE_FFT	Computes the periodicity of the provided (meta)counts.
EMOTE_frame_preference	Computes the global frame preference of the reads distribution by computing the number of counts falling into each translation frame.
EMOTE_differential_proportion	Statistical comparison of 2 sets of proportion values taking into account replicates.

Table 20 : Liste des fonctions de la librairie EMOTE-tk avec leur courte description

Dans ce qui suit, nous illustrons en Figure 143, 144, 145, 146 et 147, le flux de données et l'articulation des appels aux différentes fonctions de la librairie par rapport aux cas d'utilisation présentés en début de chapitre.

La Figure 143 qui inclut les étapes de pré-traitement, de quantification et de post-traitement. Cette figure rassemble ces étapes et illustre le fait qu'elles sont à la fois connectées, mais potentiellement indépendantes, selon le jeu de données à traiter. Par exemple, il est possible pour l'utilisateur de commencer ses analyses avec notre librairie à partir de fichiers BAM générés par lui-même, c'est-à-dire sans utiliser la fonction EMOTE_map_fastq. De même, pour l'identification des TSS, les fonctions EMOTE_normalize_paired_samples et EMOTE_filter_paired_samples ne sont pas nécessaires et ne sont pas utilisées.

La Figure 143 illustre également que la table de comptages est l'élément central de la librairie. D'une part, certaines fonctions spécifiques permettent de la manipuler dans les post-traitements comme la désaturation ou la normalisation. D'autre part, elle constitue le point de départ à différentes analyses subséquentes plus spécifiques.

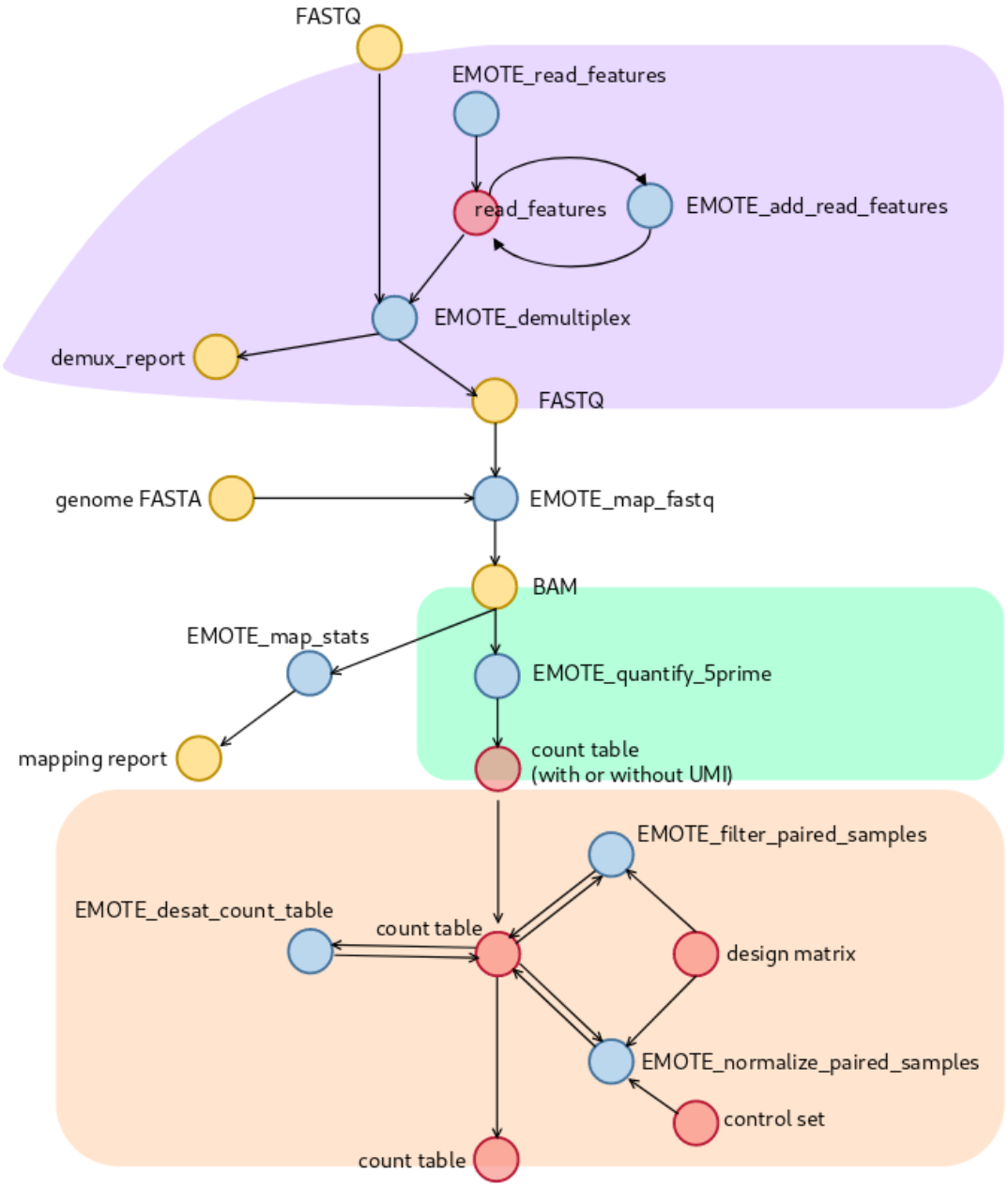


Figure 143 : Flux de données et appels de fonction typiquement effectués pour obtenir les tables de comptages à partir du fichier FASTQ contenant les lectures issues du séquençage 5' RNA-Seq. *Les ronds jaunes correspondent à des fichiers de données. Les ronds rouges correspondent à des structures de données. Les ronds bleus correspondent aux fonctions de la librairie. Le rectangle lilas rassemble les fonctions de la partie pré-traitement. Celui vert menthe-à-l'eau rassemble les fonctions de la partie quantification. Celui orange-brique délavé rassemble les fonctions de la partie post-traitement.*

Pour l'identification des TSS (Figure 144), à partir de la table de comptages obtenue, il s'agit de fournir la table de métadonnées design_matrix décrivant les échantillons appariés et répétitions afin d'effectuer l'identification avec la fonction EMOTE_betabinom_TSS qui renvoie la table de résultats avec les TSS identifiés et leur p-valeur et p-valeur ajustée.

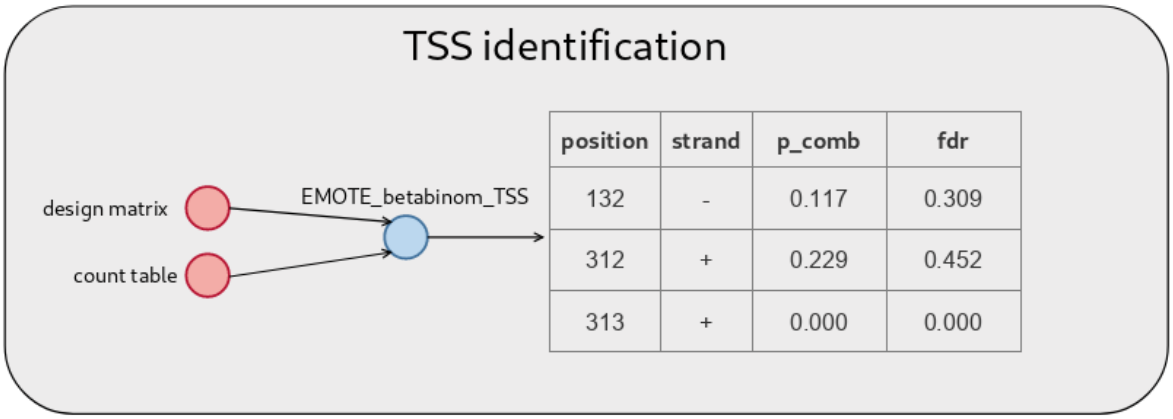


Figure 144 : Flux de données pour l'identification des TSS réalisé à l'aide de la fonction EMOTE_betabinom_TSS à partir de la table de comptage et d'une table de métadonnées*. Le résultat tabulaire renvoyé par la fonction est également représenté. « p_comb » correspondant à la p-valeur combinée des réplicats et « fdr » correspond à la p-valeur combinée ajustée.*

Pour l'étude de la dégradation co-traductionnelle par une exoribonucléase, les pré-traitements permettent d'obtenir les tables de comptages de plusieurs réplicats. L'utilisateur devra appliquer une méthode de normalisation (typiquement CPM) afin de tenir compte de la profondeur de séquençage de chaque réplicat, puis il pourra appeler la fonction EMOTE_metacount_profile en fournissant l'annotation (les positions sur le génome) des caractéristiques à considérer pour construire le profil de méta-comptages autour de la caractéristique d'intérêt (typiquement le premier nucléotide du codon de départ ou de fin de traduction). De même, l'annotation du génome avec les départs et fin de traduction de l'ensemble des gènes codant pour des protéines permettra de construire le profil de méta-comptages pour l'analyse FFT afin de déterminer s'il y a une périodicité dans le méta-profil considéré avec la fonction EMOTE_FFT. Et également, pour l'analyse de préférence de phase de traduction, l'annotation permet de considérer chaque région codante pour agréger les comptages sur chacune des 3 positions dans le codon avec la fonction EMOTE_frame_preference.

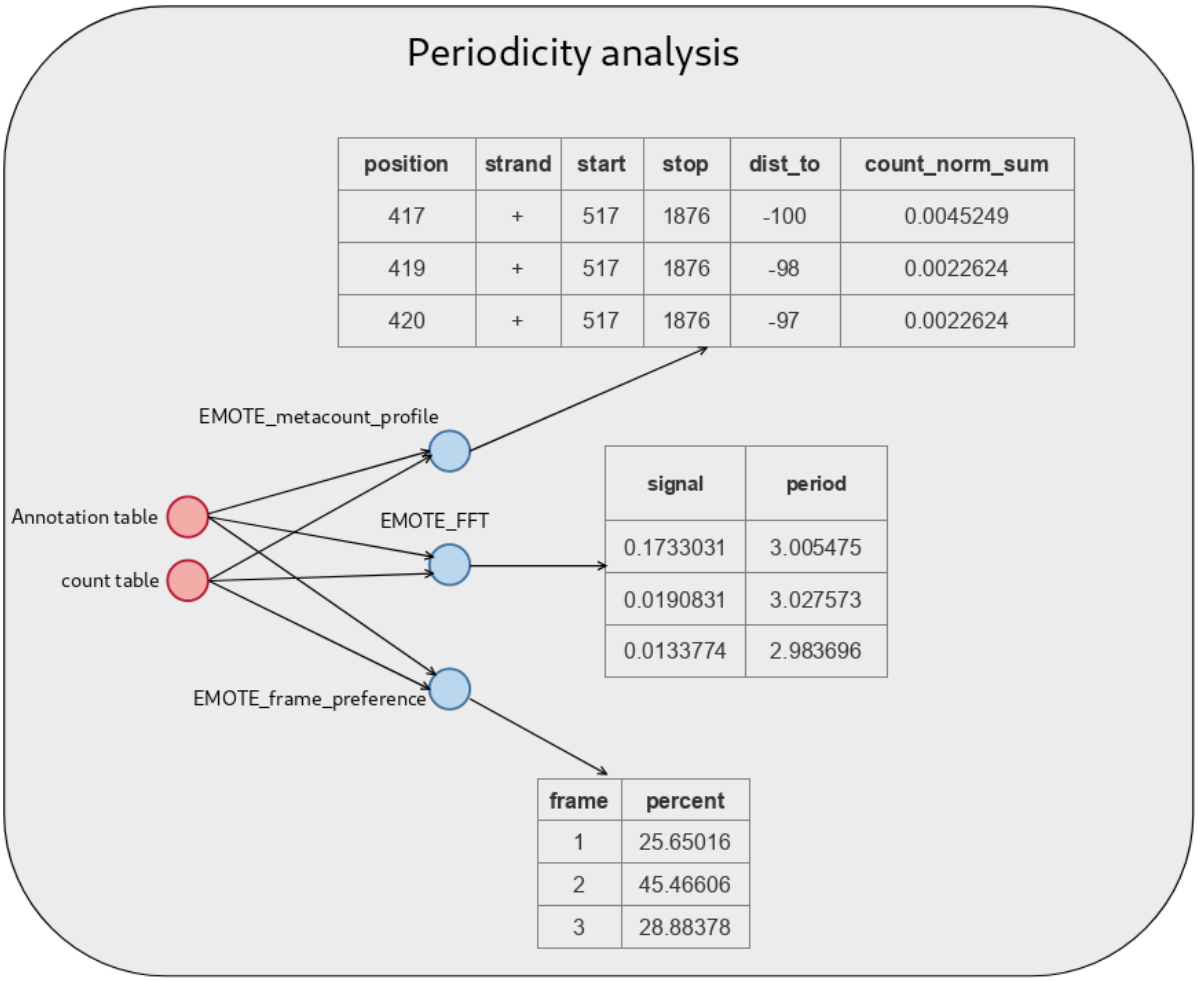
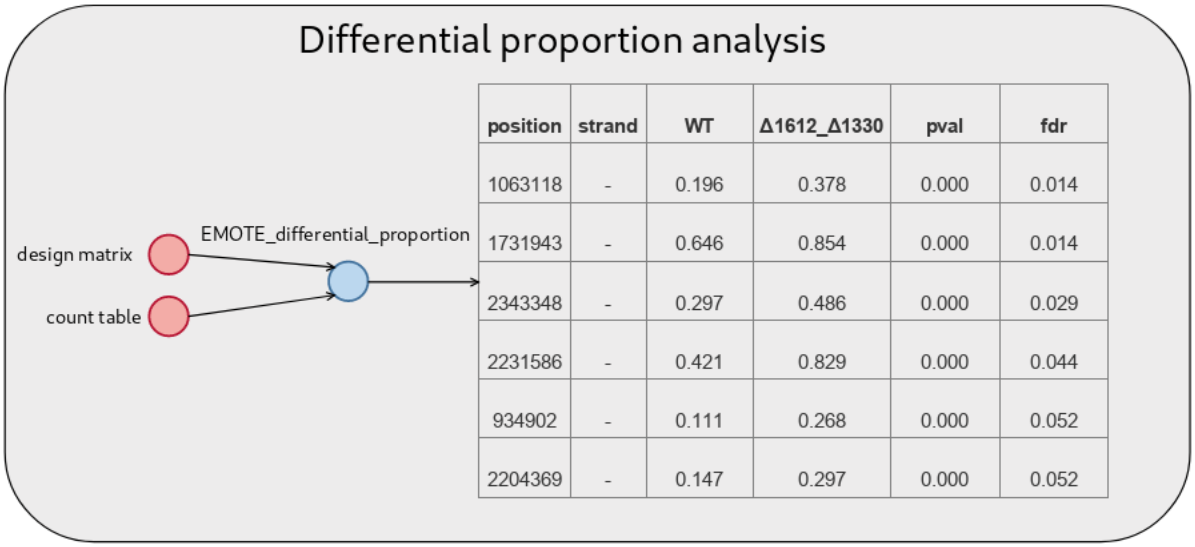


Figure 145 : Flux de données pour l'analyse de la périodicité à l'aide des fonctions EMOTE_metacount_profile, EMOTE_FFT, EMOTE_frame_preference à partir d'une table de comptage et d'une table d'annotation. *Les résultats tabulaires de chacune des fonctions sont également représentés.*

Pour les analyses de données générées par le protocole PQ-EMOTE, la table de comptages aura été post-traitée, éventuellement avec la désaturation pour certaines librairies, pour le filtrage et la normalisation des échantillons appariés. Puis, associée à une table de métadonnées (design_matrix), les extrémités 5' dont la proportion diffère entre deux souches pourront être identifiées avec la fonction EMOTE_differential_proportion (Figure 146*).* Cette fonction utilise la librairie DSS. DSS tient compte des comptages dans les différents réplicats et souches dans le test statistique effectué.

Cette librairie a été conçue pour comparer des proportions de sites méthylés dans des régions et ne tient pas compte du brin. Il a donc été nécessaire de l'englober (*wrapper*) dans la librairie pour proposer à l'utilisateur une utilisation plus logique, et lui éviter de faire les mêmes opérations de manipulation avant et après son utilisation, à chaque fois qu'il souhaite procéder à ce type d'analyse. Pour chaque extrémité 5', les proportions moyennes des formes 5'P sont renvoyées ainsi que la p-valeur et la p-valeur ajustée.



position	strand	WT	Δ1612_Δ1330	pval	fdr
1063118	-	0.196	0.378	0.000	0.014
1731943	-	0.646	0.854	0.000	0.014
2343348	-	0.297	0.486	0.000	0.029
2231586	-	0.421	0.829	0.000	0.044
934902	-	0.111	0.268	0.000	0.052
2204369	-	0.147	0.297	0.000	0.052

Figure 146 : Flux de données pour l'analyse différentielle de proportion à l'aide de la fonction EMOTE_differential_proportion à partir d'une table de comptage et d'une table de métadonnées. *Le résultat tabulaire renvoyé par la fonction est également représenté. Les colonnes « WT » et Δ1612_Δ1330 » correspondent respectivement à la moyenne pondérée de la proportion des réplicats dans la souche sauvage et la souche mutante. La p-valeur ainsi que la p-valeur ajustée sont également affichées.*

Pour le dernier cas d'utilisation, il s'agit de prédire des sites de clivage endoribonucléolytique de la RNase J1 à partir d'une analyse d'expression différentielle entre les comptages dans la souche sauvage comparés à ceux de la souche mutante J1^{AGA}. En pré-traitement, par rapport aux étapes "habituelles", il faut également filtrer les comptages faibles qui devraient correspondre aux intermédiaires de dégradation. Ensuite, le principe est d'utiliser les fonctions disponibles de la librairie EdgeR. Ceci est illustré en Figure 147 dans laquelle, la table de comptages est passée aux nœuds "DGE list formatting" puis 'EdgeR procedure" pour indiquer que ces traitements sont réalisés par des fonctions d'une autre librairie. Cela nous permet d'illustrer l'interopérabilité de notre librairie avec d'autres librairies R/BioConductor.

En effet, pour cet exemple, il s'agit simplement d'appeler le constructeur DGEList de la librairie EdgeR en lui passant la table de comptages (convertie en matrice dans laquelle les lignes représentent les positions et les colonnes les échantillons) en paramètre. A partir de la DGEList obtenue, les fonctions *calcNormFactors*, *estimateCommonDisp* et *estimateTagwiseDisp* peuvent être utilisées pour respectivement mettre à l'échelle la taille des librairies, calculer le paramètre de dispersion globale (entre tous les gènes) et calculer le paramètre de dispersion locale (par gène). Une fois tous ces paramètres estimés, la fonction *exactTest* est utilisée pour tester la différence d'expression de chaque position entre les deux souches en tenant compte des paramètres de dispersion et en se basant sur un modèle de distribution binomiale négative des comptages. Le résultat est présenté comme illustré dans la Figure 147, ou est associé à chacune des positions testées, la variation d'expression en log et la variation de CPM en log entre les deux souches ainsi que la p-valeur associée au test effectué.

Étant donné que chaque position testée fait l'objet d'un test, il est recommandé d'ajuster les p-valeurs obtenues pour contrôler le taux de faux positif (FDR). Cela permet de réduire le risque d'identifier à tort des positions comme statistiquement significatives en raison du grand nombre de tests effectués simultanément.

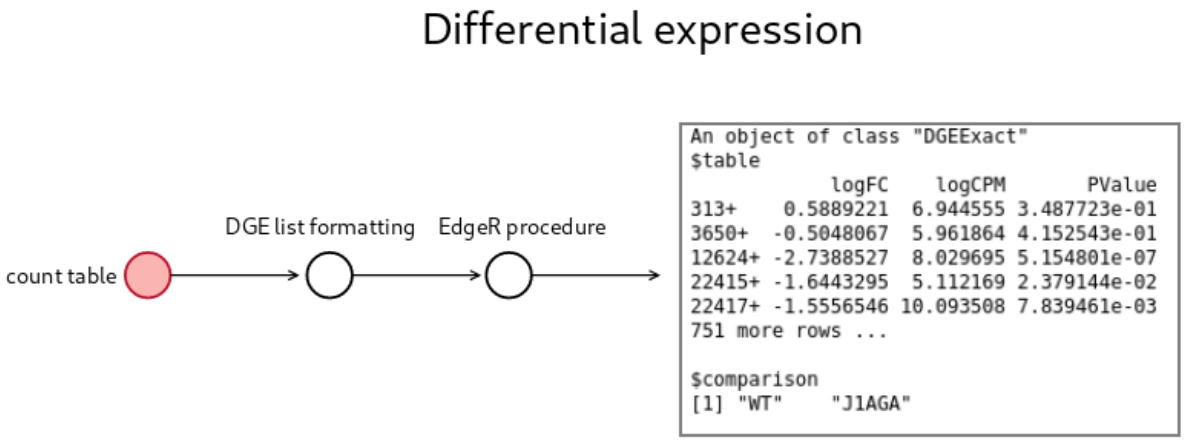


Figure 147 : Flux de données pour une analyse différentielle d'expression de gène à partir d'une table de comptage générée par les fonctions proposées par EMOTE-tk et réalisé à partir du package R EdgeR. *L'étape de formatage de la table de comptage en objet DGEList est représentée sous le nœud « DGE list formatting » et les étapes de filtrage, calcul des facteurs de normalisation, calcul des paramètres de dispersion et test statistique sont représentés sous le nœud « EdgeR procedure ». La table de résultats retournée par cette fonction est également représentée.*

En termes de performances, l'ensemble des traitements sous forme tabulaire sont réalisés avec les fonctionnalités proposées par la librairie *tidyverse* qui utilise les *tibbles* comme structure de données. Lors du développement, des tests et de l'utilisation de EMOTE-tk, nous n'avons pas rencontré de problèmes de performance. Les temps de traitements sur une station de travail ou sur un serveur de calculs n'ont pas été suffisamment longs pour nous inciter à optimiser davantage les traitements effectués, c'est-à-dire que l'ensemble des résultats présentés dans cette thèse peuvent être obtenus en moins de deux jours. De même pour l'occupation mémoire, l'ensemble des traitements peuvent être effectués sur une station de travail disposant de 16Go de mémoire. Ceci est en partie dû au fait que les jeux de données sont relativement petits pour des génomes bactériens dont le génome ne fait que quelques millions de paires de bases.

2.3.4 Discussion

Dans ce chapitre, nous avons présenté les fonctionnalités que nous avons développées pour analyser des données générées par des protocoles RNA-Seq d'extrémités 5'. Nous avons conçu EMOTE-tk avec les principes FAIR (Findable, Accessible, Interoperable, Reusable). Notre objectif était de garantir que l'outil soit accessible et utilisable par la communauté scientifique. À cette fin, nous avons mis en place un dépôt du code source sur une instance GitLab maintenue par le CNRS. De plus, nous avons rédigé des tutoriels détaillés pour guider les utilisateurs dans l'installation, la configuration et l'utilisation de EMOTE-tk, permettant à l'utilisateur d'adapter les exemples à son projet d'étude. Pour améliorer l'interopérabilité, nous avons veillé à ce que les formats de sortie produits par les fonctions de EMOTE-tk soient compatibles avec la plupart des librairies RNA-Seq disponibles dans l'écosystème R/BioConductor. Cela permet aux utilisateurs d'intégrer plus facilement les résultats de EMOTE-tk dans leurs chaînes de traitements existantes, tout en bénéficiant des fonctionnalités disponibles dans d'autres librairies R/BioConductor pour l'analyse et la visualisation des données RNA-Seq. Enfin, nous avons également pris en compte la reproductibilité des analyses en fournissant un environnement Conda permettant à l'utilisateur de configurer le même environnement avec les mêmes versions de librairies dont dépend EMOTE-tk. Cela devrait permettre, dans la plupart des cas, aux utilisateurs de reproduire les résultats obtenus et présentés dans notre travail. La mise à disposition du code des fonctions et leur documentation en ligne est une première étape. La prochaine étape (en cours de réalisation) consiste à proposer une librairie R développée avec la librairie *devtools*. Ceci permettra d'installer la librairie avec la documentation disponible depuis R. L'étape suivante sera de la proposer sur les dépôts de bioconda, ainsi que BioConductor.

Nous n'avons pas présenté dans cette partie les fonctionnalités proposées pour l'analyse de données correspondant aux extrémités 3' des ARN. Leur particularité est que les alignements sur les séquences de référence peuvent être partiels, du fait de modifications post-transcriptionnelles. Ceci nécessite donc d'adapter à la fois les alignements et la quantification des lectures observées qui peuvent être associées à des séquences ajoutées en 3' des ARN. Le développement de ces traitements spécifiques aux extrémités 3' nous a néanmoins permis, à travers un travail collaboratif, d'identifier des traitements communs à ceux effectués sur les extrémités 5', et ainsi à rationaliser et clarifier les parties de code qui étaient communes.

En termes d'amélioration des performances en temps de calcul et d'occupation mémoire, il est possible d'étudier l'utilisation d'une autre librairie pour gérer les tables. En effet, les seuls traitements effectués à l'intérieur de la librairie (sans librairie externe comme pour les fichiers FASTQ avec ShortRead, les alignements avec Rbowtie, et les fichiers BAM avec Rsamtools) concernent la gestion de tables. Dans cette optique, l'utilisation de *data.table* de la librairie *DT*, plutôt que de *tibbles*, serait à considérer. Dans tous les cas, même si la représentation en interne venait à changer, les fonctions proposées resteraient les mêmes et les scripts qui utilisent la librairie n'auraient pas à être modifiés, ou alors très peu.

Pendant ces travaux, nous avons pu adapter les flux de données et d'appels de fonctions décrits dans la partie réalisation, pour son utilisation dans le gestionnaire de *workflows* SnakeMake (Köster & Rahmann, 2012). L'utilisation de SnakeMake permet une gestion efficace des ressources informatiques, notamment en distribuant les processus sur un cluster de calcul avec SLURM (Simple Linux Utility for Resource Management). Cela permet également de paralléliser certains calculs indépendants, comme par exemple pour l'alignement de lectures contenues dans des fichiers FASTQ démultiplexés. La mise en œuvre de cette adaptation a consisté à développer un script R chargeant la librairie EMOTE-tk, puis appelant la fonction avec les paramètres spécifiés en ligne de commande par l'utilisateur. L'avantage de cette approche est de pouvoir réutiliser les chaînes de traitement adaptées à un type d'analyse sur des données différentes. En contrepartie, chaque appel de fonction nécessite de lire et d'écrire les données d'entrée et de sortie depuis des fichiers car chaque appel de fonction est isolé dans un script exécuté de manière isolée.

Le développement de cette librairie sera soumis pour une publication scientifique dont la rédaction est en cours de finalisation.

Matériels et Méthodes

Analyse de gène différentiellement exprimés (DGE)

Alignement des lectures sur un génome de référence

Représentation de la conservation nucléotidique par Weblogos

Comparaison de distribution entre deux échantillons

Ajustement des p-valeurs

Analyse en composante principale (ACP)

Modèle linéaire

Prédiction de structure secondaire d’ARN

3 Matériels et méthodes

3.1 Analyse de gènes différentiellement exprimés (Differential gene expression - DGE)

L'analyse différentielle de l'expression des gènes vise à identifier les gènes dont l'expression varie significativement entre différentes conditions expérimentales.

Les abondances ou les comptages de transcrits pour un gène ne suivent généralement pas une distribution normale. En raison des caractéristiques des données de séquençage, telles que les écarts importants des niveaux d'expression entre les gènes, la forte hétérogénéité des échantillons, et l'incertitude sur la mesure pour les gènes faiblement exprimés qui pourrait avoir un fort impact, il n'est pas possible d'utiliser des modèles basés sur une distribution suivant la loi normale. La distribution binomiale négative est particulièrement adaptée car elle permet de modéliser une variance plus élevée que celle modélisée par la loi normale, rendant ainsi les analyses plus robustes aux valeurs extrêmes (Conesa et al., 2016) .

Un programme couramment utilisé pour l'analyse différentielle de l'expression des gènes à partir de données RNA-Seq est EdgeR (Robinson et al., 2010). Il utilise une distribution binomiale négative et procède à une estimation de la dispersion des données en deux étapes : une estimation globale de la dispersion pour tous les gènes de l'échantillon, suivie d'une estimation locale pour chaque gène. Cette double estimation permet une modélisation de la variabilité biologique et technique. Après avoir estimé les paramètres de dispersion, il utilise un test exact de Fisher modifié pour permettre la prise en compte des paramètres de dispersion estimés, pour déterminer l'expression différentielle des gènes.

Lors de nos travaux nous avons utilisé la version 3.40.0 de librairie EdgeR.

3.2 Alignement des lectures sur un génome de référence

L'alignement des lectures sur un génome de référence est une étape clé dans une analyse RNA-Seq car elle permet d'aligner les lectures obtenues par séquençage à leur(s) emplacement(s) dans le génome de référence. Nous avons déjà abordé l'immense variété de méthodes d'alignement disponibles, parmi lesquelles, Bowtie, un programme très utilisé en raison de sa rapidité, en particulier pour les lectures courtes.

Bowtie est un programme développé par (Langmead et al., 2009) et qui utilise une stratégie de type "seed-and-extend". Cela signifie que dans un premier temps, il divise les lectures en morceaux appelés graines (seeds) de longueur fixe. En utilisant un index BWT (Burrows-Wheeler Transform), Bowtie retrouve dans le génome de référence les régions identiques aux graines. Dans un second temps, il tente d'étendre ces alignements pour couvrir toute la longueur de la lecture tant que le score de similarité de l'alignement dépasse un certain seuil. Cela peut impliquer des mésappariements. Bowtie évalue ensuite la qualité de chaque alignement en fonction du score de l'alignement, calculé à partir du nombre de correspondances (matches) et de mésappariements (mismatches). Les meilleurs alignements sont sélectionnés en fonction de ce critère, en prenant en compte les paramètres définis par l'utilisateur tel que la conservation des lectures multi-alignées.

En 2012, une version améliorée de Bowtie, appelée Bowtie2, a été proposée (Langmead & Salzberg, 2012)**.** Bowtie2 utilise un algorithme différent de celui de Bowtie et permet d'autres fonctionnalités comme la prise en compte d'insertion-délétion ce qui peut ralentir l'exécution. Cependant, nous n'avons pas eu besoin de ces fonctionnalités, et nous avons utilisé Bowtie.

Bowtie bénéficie d'un empaquetage (wrapper) disponible sur R. La version de Bowtie utilisée dans l'ensemble des alignements décrits dans ce manuscrit est la version 1.38.0.

3.3 Représentation de la conservation de séquences nucléotidiques par Weblogos

Afin de détecter d'éventuels motifs par la conservation de nucléotides à des positions spécifiques dans un ensemble de séquences alignées, nous avons utilisé des Weblogos. Ces Weblogos sont basés sur la théorie de l'information et leur construction utilise les notions d'entropie et d'information et permettent d'identifier des motifs biologiquement significatifs qui peuvent être des sites de liaison, des éléments régulateurs ou des sites catalytiques.

Les séquences d'intérêt, doivent d'abord être alignées. Puis, pour chaque position j de l'alignement, il est possible de calculer l'information relative ou bits d'information (I) qui indique à quel point cette position est conservée ou variable. Le calcul est effectué en soustrayant l'entropie observée ($-\sum_i f_{ij} log_2 f_{ij}$) de l'entropie maximale possible ($-\sum_i P_i log_2 P_i$).

P_i (i = 1 à 4) correspond à la fréquence du nucléotide i dans l'ensemble du génome (probabilité théorique) et f_{ij} est la fréquence observée de la base i à la position j sur l'ensemble de séquences.

Pour une séquence d'ARN, où il y a 4 ribonucléotides possibles (A, C, G, U), et où la probabilité de chacun des ribonucléotides est estimée à 0,25, l'entropie maximale est égale à 2. Ceci permet de simplifier la formule du calcul de l'information : ($I(j) = 2 - \sum_i f_{ij} log_2 f_{ij}$)

Cependant, certains organismes présentent des fréquences en nucléotides assez éloignées de 0,25, notamment *S. aureus* riche en AT (33%), et il est préférable d'utiliser la formule non simplifiée qui tient compte de ces biais de fréquences d'apparition.

Les positions du signal qui contiennent de l'information sont celles qui ont une composition très biaisée par rapport à ce qui est attendu de manière aléatoire. Par exemple, si à une position j, une seule base i est présente alors $f_{ij} = 1$ et $\log_2 f_{ij} = 0$, donc $f_{ij} * \log_2 f_{ij} = 0$ et les fréquences observées des autres bases sont nulles. On aura $I(j) = 2$, c'est la quantité d'information maximale.

Les programmes permettant la création de Weblogos affichent à chaque position de l'alignement les ribonucléotides possibles, dont la taille est proportionnelle à leur fréquence à cette position (fréquence * information). Dans cette étude, nous avons utilisé le logiciel WebLogo3 (version 3.7 disponible sur github), pour créer les Weblogos en tenant compte du pourcentage en GC du génome de *S. aureus*.

3.4 Comparaison de distribution entre deux échantillons

Dans cette étude, nous avons comparé à plusieurs reprises les proportions d'extrémités 5'P entre deux ensembles (entre deux souches ou entre les gènes impactés et non impactés).

Les distributions peuvent souvent être asymétriques, ou bien les échantillons à comparer contenir peuvent contenir des valeurs extrêmes (*outliers*). Les tests paramétriques tels que le test de Student ne sont alors pas applicables. Pour comparer ces échantillons, nous utilisons donc un test non paramétrique : le test de Wilcoxon-Mann-Whitney, ou test de rang de Wilcoxon pour deux échantillons, pour comparer deux échantillons indépendants afin de déterminer s'ils proviennent de populations ayant la même distribution.

Pour effectuer le test de Wilcoxon-Mann-Whitney, nous avons utilisé la fonction *wilcoxon.test* du package R *stats* (version 4.2.3).

3.5 Ajustement des p-valeurs

Lors de tests multiples, le risque de rejeter à tort l'hypothèse nulle (risque de première espèce) augmente. Par conséquent, des méthodes d'ajustement sont nécessaires pour maintenir un niveau de faux positifs global acceptable. La méthode de Bonferroni qui consiste à multiplier les p-valeurs par le nombre total de tests effectués est trop conservatrice et des méthodes de contrôle du taux de faux positif (False Discovery Rate - FDR) sont plus adaptées. La méthode de Benjamini-Hochberg (BH), proposée par Yoav Benjamini et Yosef Hochberg en 1995, est couramment utilisée pour ajuster les p-valeurs afin de contrôler le taux de fausses découvertes (FDR). Cette méthode consiste à classer les p-valeurs par ordre croissant. Les p-valeurs les plus faibles sont ajustées de manière plus significative, tandis que les p-valeurs les plus élevées subissent un ajustement moindre. En permettant une proportion acceptée de fausses découvertes, la méthode BH est particulièrement adaptée aux études exploratoires. Elle offre un équilibre entre la réduction des faux positifs et le maintien de la puissance statistique, minimisant ainsi le risque de faux négatifs (risque de deuxième espèce).

Dans notre étude, nous avons utilisé la fonction *p.adjust* du package *stats* (version 4.2.3) pour ajuster les p-valeurs avec la méthode BH.

3.6 Analyse en composantes principales (ACP)

L'ACP est une méthode de réduction du nombre de dimensions/variables d'un jeu de données composé de variables quantitatives. Pour cela, le principe de l'ACP est de maximiser l'information contenue dans les premières dimensions (variance portée par les composantes principales) en changeant de repère par une combinaison linéaire des variables de départ (composantes principales). Les composantes sont calculées (vecteurs propres et valeurs propres) à partir de la matrice de variance-covariance des données (éventuellement après une mise à l'échelle des variables pour leur donner la même importance). Il est alors possible de projetées les données initiales sur les composantes principales, réduisant ainsi la dimensionnalité tout en conservant la structure des données.

Pour effectuer toutes ces étapes, nous utilisons la fonction *prcomp* de la version 4.3.2 du package R *stats*.

3.7 Modèle linéaire

Afin de détecter les gènes dont demi-vie varie de manière statistiquement significative entre les différentes souches tout en tenant compte des réplicats, nous avons utilisé un modèle linéaire.

Ces derniers sont souvent utilisés pour analyser les données biologiques qui présentent une grande dispersion et ne respectent pas l'hypothèse de normalité.

De plus, ils permettent de modéliser l'effet des souches (sauvage *vs* mutante) sur les demi-vies des gènes en tenant compte de la dispersion entre les réplicats (locale) et entre les gènes d'un même échantillon (globale).

Dans l'environnement de programmation R, le package *limma* (Ritchie et al., 2015) fournit un ensemble de fonctions permettant la création et l'ajustement de modèles linéaires.

Nous avons utilisé les fonctions *LmFit* et *eBayes* (*limma* version 3.54.0) pour respectivement construire les modèles linéaires par gène et calculer les paramètres de dispersion globaux et locaux.

3.8 Prédiction de structure secondaire d’ARN

La prédiction de structures secondaires sur des ARN est utilisée pour comprendre la fonction et l'interaction des molécules d'ARN. Ces structures comprennent des éléments tels que les renflements et les tiges-boucles, qui peuvent jouer des rôles dans divers processus biologiques tels que la régulation de leur dégradation.

Les structures secondaires d'ARN sont prédites en utilisant des méthodes, où l'objectif est de trouver la conformation avec l'énergie libre la plus basse et qui correspondrait donc à la structure la plus stable. Les bases de l'ARN peuvent se lier par des liaisons hydrogène, formant des paires de bases Watson-Crick (A-U et G-C) ou des paires non canoniques (G-U).

Les algorithmes utilisent des tables thermodynamiques qui fournissent les énergies libres associées aux différentes paires de bases et motifs structurels. De plus, des algorithmes de programmation dynamique tel que celui de (Zuker & Stiegler, 1981) sont utilisés pour explorer efficacement l’espace des configurations possibles et identifier celle avec l'énergie minimale.

RNAfold est un programme couramment utilisé pour prédire la structure secondaire de l'ARN. Il fait partie de la librairie ViennaRNA dont la première version a été proposée par (Hofacker et al., 1994)**.**

L’ensemble des prédictions de structures secondaire d’ARN présenté dans ce manuscrit a été réalisé à partir du programme RNAfold version 2.5.1.

Discussions et Perspectives

La RNase J1 : L’actrice majeure de la dégradation des ARN chez S. aureus

Etats de phosphorylation des extrémités 5’ des ARN et impact de 4 NUDIX sur la régulation de ces états

EMOTE-toolkit : Une librairie R pour le traitement et l’analyse de données end RNA-Seq

4 Discussions et perspectives

Au cours de ces travaux de thèse, nous nous sommes intéressés à l'étude du métabolisme ARN des Bacillota, et plus précisément à la régulation de la dégradation des ARN par les RNases J chez *S. aureus*. Dans un premier temps, nous avons réalisé des études globales de l'impact de la RNase J1 sur la stabilité des ARN, et sur ses spécificités de substrat. Nous avons ensuite réalisé l'étude d'un mécanisme en particulier lié à la dégradation des ARN, et qui correspond à la déphosphorylation des extrémités 5' des ARN. Pour finir, nous avons exposé les principes de conception et les choix techniques derrière la création de la librairie R EMOTE-tk, développée pour le traitement et l'analyse des données de RNA-Seq des extrémités 5' obtenues dans les deux premières parties de l'étude.

4.1 La RNase J1 : l'actrice majeure de la dégradation des ARN chez *S. aureus*.

Chez *S. aureus*, il est supposé que la RNase J joue un rôle central dans la dégradation des ARN. Cette hypothèse est étayée par plusieurs observations, comme par exemple des études ayant montré que les autres ribonucléases présentes dans cette bactérie, telles que la RNase Y, la PNPase ou encore la RNase III, jouent des rôles moins prédominants et plus spécifiques dans la dégradation des ARN. Les effets pléiotropes sévères associés à la délétion de la RNase J1, qui affectent de nombreux aspects de la physiologie de la bactérie tels que la vitesse de croissance ou la résistance au froid, soutiennent fortement cette hypothèse. Ces observations suggèrent que la RNase J1 n'est pas seulement une ribonucléase parmi d'autres, mais bien une régulatrice majeure de la stabilité et du renouvellement (*turnover*) des ARN chez *S. aureus*. Cependant, à l'heure actuelle, aucune étude *in vivo* n'avait encore été réalisée pour évaluer l'impact de la RNase J sur la stabilité des ARN chez *S. aureus*.

Par ailleurs, la RNase J possède une double activité de dégradation, ce qui lui permettrait, non seulement de contourner les protections des extrémités 5' des ARN grâce à un clivage endoribonucléolytique, mais également de dégrader efficacement les ARN *via* une activité exoribonucléasique. Cependant, bien que son activité endoribonucléasique ait été mise en évidence *in vitro* et *in vivo* sur certains transcrits comme celui de *thrS* (Mathy et al., 2010), aucune étude globale *in vivo* n'avait encore été réalisée pour identifier l'ensemble de ses sites de clivage.

Pour finir, il a été montré qu'un mécanisme de dégradation co-traductionnelle est présent chez *S. aureus* dont l'effet est supposé être dû à la RNase J. Cependant, encore une fois, aucune étude n'avait été entreprise pour le démontrer de manière concluante.

Au cours de cette thèse, nous avons démontré que la dégradation des ARN de 47 % des gènes exprimés dans les conditions étudiées dépend de la RNase J1, avec une préférence pour les gènes fortement exprimés. En outre, nous avons établi que les ARN de 85 % des gènes interagissent avec la RNase J1, ce qui confirme son rôle central dans le processus de dégradation des ARN chez *S. aureus*. Ces résultats indiquent tout de même qu'en l'absence de la RNase J, 53 % (100 - 47) des gènes possèdent des ARN dont la demi-vie reste courte. Cela suggère que bien que la RNase J soit prédominante, d'autres acteurs interviennent également dans la dégradation des ARN.

Nous avons également mis en évidence que la RNase J1 est la principale responsable du phénomène de dégradation co-traductionnelle. À partir de ces résultats, nous avons développé une approche permettant d'identifier les gènes dégradés de manière co-traductionnelle par la RNase J1, révélant que les ARN d'au moins 21 % des gènes exprimés dans les conditions étudiées sont affectés par ce mécanisme. Cette approche ouvre de nouvelles perspectives pour de telles études, néanmoins, sa mise en œuvre reste complexe en raison du besoin de profondeur de séquençage pour obtenir suffisamment d'intermédiaires de dégradation, notamment avec des protocoles axés sur les TSS. Néanmoins, avec un protocole plus adapté ciblant les intermédiaires plutôt que les TSS, cette méthode pourrait s'avérer utile.

Les études globales menées dans ce chapitre intègrent différents types d'approches telles que l'étude de la stabilité des ARN, l'analyse des interactions ARN-RNase J1 et l'exploration de la dégradation co-traductionnelle, révélant divers aspects de la régulation de la dégradation des ARN. Nous avons démontré que certains ARN, dont la demi-vie n'est pas impactée par la délétion de la RNase J1, peuvent néanmoins être dégradés par cette ribonucléase. De même, nous avons observé que des ARN normalement peu ou non dégradés, comme celui du gène *hu*, peuvent interagir avec la RNase J1, suggérant une possible liaison sans dégradation par cette enzyme. Bien que les résultats de ces trois approches concordent, des différences subsistent notamment entre les gènes impactés et les gènes ciblés. Ceci souligne l'importance d'utiliser diverses approches pour caractériser le rôle d'une ribonucléase *in vivo*, étant donné que d'autres facteurs peuvent influencer la régulation des processus biologiques.

Par ailleurs, nos études sur les substrats de la RNase J ont confirmé que cette dernière n'est que très peu sensible aux états de phosphorylation des extrémités 5', et nous n'avons pas été en mesure de déterminer une sensibilité particulière par rapport à leur composition en ribonucléotides ou leur capacité à former des structures secondaires. Cependant, nous avons montré que la RNase J1 cible préférentiellement les gènes qui ne sont pas en première position dans un opéron, ce qui pourrait expliquer que nous n'ayons pas trouvé de signal protégeant ou ciblant les extrémités 5'. De plus, la double activité de la RNase J1 signifie qu'elle pourrait initier la dégradation des ARN par un clivage endoribonucléolytique, puis les dégrader de manière exoribonucléolytique.

Par ailleurs, nos investigations visant à identifier les sites de clivage endoribonucléolytique de la RNase J1 nous ont permis de proposer 125 sites potentiels. La plupart de ces sites se situent juste après une guanosine, comme les sites de clivage de la RNase Y, et sont proches voire à l'intérieur de régions prédites comme structurées. Ces sites sont, comme évoqué à plusieurs reprises, le résultat d'une étude très préliminaire de par le fait que les deux activités de dégradation se font *via* le même site catalytique, et que les sites de clivage potentiels peuvent aussi correspondre à des pauses exoribonucléolytiques. L'analyse de données RNA-Seq d'extrémités 3' des ARN pourrait nous permettre de conclure sur l'origine de ces extrémités 5'. Les fonctionnalités de traitement de ces données ayant déjà été développées et implémentées dans une libraire, l'analyse de ces données pourrait se faire rapidement.

Pour conclure, nos travaux se sont principalement voire exclusivement portés sur l'étude de la RNase J1, notamment car il a été montré que la RNase J2 est enzymatiquement peu active. Nos résultats indiquent que le site actif de la RNase J1 réalise presque à lui seul la totalité de la fonction de dégradation. Toutefois, la conservation de la présence de la RNase J2 suggère un rôle potentiellement important pour la cellule, justifiant des études sur des souches mutantes du site actif de la RNase J2 ou une délétion complète. Étant donné que les chaînes de traitement des données sont déjà en place, ces études pourraient, elles aussi, être menées de manière efficace et rapide.

L'ensemble de ces résultats suggèrent qu'il y a peu de gène non ciblés, non impactés et non dégradés par la RNase J1 dans les conditions étudiées. De plus, n'ayant pas identifié de caractéristiques communes aux ARN que la RNase J1 ne semble pas réguler, cela suggère qu'elle pourrait potentiellement dégrader l'ensemble des ARN de *S. aureus*, à l'exception de quelques cas particuliers nécessitant une étude approfondie cas par cas.

4.2 Etats de phosphorylation des extrémités 5’ des ARN et impact de 4 NUDIX sur la régulation de ces états

Chez les bactéries, les ARN ne possèdent pas de coiffe en 5’ pour les protéger de la dégradation par les ribonucléases ; c'est l’état triphosphorlylé en 5’ des ARN nouvellement synthétisés naturelle des extrémités 5’ qui assure cette protection, notamment contre l'action de la RNase J pour les organismes qui la possède. Des études *in vitro* ont montré chez diverses espèces bactériennes telles que *B. subtilis* (Li de la Sierra-Gallay et al., 2008; Mathy et al., 2007), *Mycobacterium smegmatis* (Taverniti et al., 2011), et *Streptomyces coelicolor* (Pei et al., 2015, p. 201) que la RNase J dégrade de manière préférentielle, voire exclusive, les ARN 5’P, permettant ainsi aux ARN d'être traduits avant d'être dégradés. Cependant, il a été observé que chez *S. aureus*, la RNase J présente une sensibilité nettement moindre aux états de phosphorylation des extrémités 5’ des ARN, avec une préférence d’un facteur 3 pour les ARN 5’P par rapport aux ARN 5’PPP (Hausmann et al., 2017).

Les enzymes capables de déphosphoryler ces extrémités, les ribopyrophosphohydrolases (RppH) jouent donc un rôle important dans la régulation de la dégradation des ARN chez les bactéries. À ce jour, aucune RppH n'a été expérimentalement identifiée chez *S. aureus*. Cependant, (Bonnin & Bouloc, 2015) ont montré que le gène SA1612 présente une certaine similarité (38 %) avec la RppH de *B. subtilis*, ainsi qu'une conservation de la synténie. La délétion de ce gène chez *S. aureus* entraîne seulement quelques différences dans les niveaux d'expression des gènes (seulement 4 gènes sur-exprimé), ce qui est cohérent si l'on considère que la RNase J est capable de dégrader les ARN 5’PPP.

De plus, notre équipe a identifié quatre autres enzymes NUDIX qui pourraient potentiellement présenter une activité RppH. Cependant, l'impact de ces cinq enzymes sur les états de phosphorylation des extrémités 5’ des ARN reste à étudier. Pour mener à bien cette étude, l'équipe avec laquelle nous avons collaboré pendant cette thèse a développé un protocole permettant de quantifier les états de phosphorylation des extrémités 5’ des ARN (PQ-EMOTE). Cette équipe a également construit sept souches mutantes pour 4 des 5 candidats RppH de *S. aureus* (simples et multiples) et a généré des données avec le protocole PQ-EMOTE à partir de celles-ci.

Au cours de cette thèse, nous avons élaboré une chaîne de traitements dédiée au traitement et à l'analyse de ce nouveau type de données produites par le protocole PQ-EMOTE. Parmi ces développements, nous avons mis au point une méthode de désaturation. En effet, nos recherches ont mis en évidence un problème de saturation des UMI, un phénomène peu documenté dans la

littérature et pour lequel aucune méthode de "désaturation" n'avait été proposée. Cette désaturation consiste à éliminer les duplicatas PCR en tenant compte du fait qu'un même identifiant unique de molécules peut être associé à plusieurs copies d'un même ARN. Nous avons évalué les performances de cette méthode, qui permet de réduire significativement le pourcentage d'erreurs dans l'estimation du nombre de transcrits réels, mais dont l'utilisation ne prévaut pas sur l'utilisation d'UMI suffisamment longs.

De plus, nous avons démontré que ce phénomène de saturation est exacerbé par une distribution non uniforme des séquences d'UMI. Bien qu'il serait intéressant de connaitre l'origine de cette non-uniformité, il est surtout important de prendre en compte cette information dans la construction du plan d'expérience, au risque de ne pas pouvoir exploiter les données.

Afin de vérifier la fiabilité des estimations de la proportion d'extrémités 5'P fournies par la chaîne de traitements que nous avons développée, l'équipe collaboratrice est en train d'évaluer cette proportion sur quelques transcrits avec des Northern Blot. Cette comparaison entre les résultats obtenus par les deux approches vise à valider le protocole PQ-EMOTE ainsi que la chaîne de traitements mise en place.

Grâce à cette chaîne de traitements, nous avons déterminé que la proportion médiane d'extrémités 5'P dans la souche sauvage SA564 de *S. aureus* est de 25 %. En revanche, les analyses des états de phosphorylation des extrémités 5'P des ARN dans les souches mutantes pour les candidats RppH n'ont pas révélé de perte d'activité RppH. Nous n'avons donc pas pu identifier l'enzyme ou les enzymes responsables de la déphosphorylation de ces extrémités.

Les résultats que que nous avons obtenus suggèrent donc que la cinquième NUDIX non testée pourrait être responsable de l'activité RppH observée dans la souche quadruple mutante. Par conséquent, en collaboration avec l'équipe, il est prévu d'analyser les états de phosphorylation des extrémités 5' des ARN dans une souche simple mutante pour la NUDIX SA1501, déjà construite, ainsi que dans la souche quintuple mutante actuellement en cours de construction.

Cependant, le protocole a été optimisé parallèlement à la réalisation des analyses et nos développements méthodologiques ont permis l'introduction d'améliorations telles que l'augmentation de la longueur des UMI et le nombre accru de réplicats. Néanmoins, des perspectives d'améliorations sont encore à discuter, comme par exemple le seuil utilisé lors du filtrage sur gel, la profondeur de séquençage ou encore l'étape de séparation en deux pools. Cela signifie qu'il est délicat de se prononcer de manière définitive sur les données de la première librairie.

Il sera donc peut-être nécessaire de générer de nouveau les données de certaines souches, telles que la souche simple mutante pour le gène SA1612, issues d'une librairie peu profonde et peu fiable (avec des UMI courts et peu de réplicats).

Il faut tout de même garder à l'esprit que les NUDIX sont des enzymes difficiles à étudier *in vivo* en raison de leurs nombreuses interactions croisées dues à leurs multiples activités enzymatiques. Par conséquent, si aucun des mutants testés ne montre une diminution globale de la proportion d'extrémités 5'P, il deviendra nécessaire d'envisager des expériences *in vitro* afin de tester l'activité RppH de chacune des cinq NUDIX, en particulier la NUDIX SA1612.

Malgré cette impossibilité à conclure sur l'identité des RppH, nous avons toutefois pu étudier les proportions d'extrémités 5'P des ARN en fonction de leur composition en ribonucléotides. Nous avons ainsi démontré que les ARN commençant par une guanosine sont préférentiellement 5'P. Ces résultats contrastent avec ce qui a été observé pour les RppH de *E. coli* et de *B. subtilis*, où une adénosine est préférée à une guanosine. Nous avons également identifié que les ARN débutant par deux guanosines successives, et plus généralement par une guanosine suivie d'une purine, sont préférentiellement 5'P par rapport à ceux commençant par une guanosine suivie d'une pyrimidine. Ces résultats sont cette fois-ci concordants avec les préférences de *Ec*RppH et *Bs*RppH, bien que cette dernière présente une préférence presque exclusive pour une guanosine en deuxième position, ce qui ne semble pas être le cas chez *S. aureus*.

Il est cependant crucial de considérer que les préférences de *Ec*RppH et *Bs*RppH ont été identifiées *in vitro*, tandis que nos résultats proviennent d'expériences *in vivo*. Il est donc possible que d'autres facteurs interviennent dans la proportion apparente d'extrémités 5'P, notamment leur dégradation.

Pour conclure, bien que ces résultats nécessitent d'être complétés, ils mettent en lumière des différences importantes par rapport aux autres membres du phylum Bacillota, notamment *B. subtilis*, qui est souvent utilisé comme organisme modèle. Ces travaux apportent donc une contribution à la compréhension des mécanismes de dégradation des ARN et soulignent les particularités de *S. aureus* par rapport à d'autres Bacillota.

4.3 EMOTE-tk librairie R pour le traitement et l'analyse de données end RNA-Seq

Les analyses réalisées dans les chapitres précédents ont nécessité le développement de nouvelles méthodes adaptées aux spécificités des données PQ-EMOTE et, plus largement, aux données de séquençage des extrémités des ARN. En particulier, la composition des lectures et la quantification basée uniquement sur le premier nucléotide aligné (extrémité 5') rendent les méthodes existantes inadaptées pour traiter et analyser ces données.

Pour répondre à ce besoin, nous avons développé et intégré un ensemble de méthodes dans une librairie R intitulée EMOTE-tk. Cette librairie inclut plusieurs étapes dont une étape de prétraitement, permettant de vérifier la validité des lectures, de les tronquer et de les répartir dans des fichiers séparés en fonction de leurs conditions expérimentales associées (démultiplexage). Nous avons pensé cette étape pour qu'elle soit adaptable à une grande variété de constructions de lectures, rendant ainsi la librairie utilisable pour d'autres données que celles générées par les protocoles de notre équipe collaboratrice.

Bien que des fonctionnalités d'alignement des lectures sur le génome de référence existent déjà, nous les avons intégrées dans une fonction pour faciliter l'enchaînement du flux de données. La quantification des extrémités des lectures alignées est ensuite réalisée et stockée dans une table de comptages, qui constitue l'élément central de notre librairie EMOTE-tk. À partir de cette table de comptage, il est possible d'effectuer des post-traitements et diverses analyses.

Concernant les post-traitements, nous avons implémenté des méthodes dédiées au filtrage et à la normalisation d'échantillons appariés.

Nous avons également intégré diverses analyses spécifiques aux besoins identifiés lors de ces travaux de thèse. Cela inclut notamment l'analyse différentielle de proportion, utilisant une librairie existante (DSS), ainsi que des analyses de périodicité dans les distributions d'intermédiaires de dégradation.

Notre librairie s'inscrit dans une démarche FAIR (Findable, Accessible, Interoperable, Reusable), assurant ainsi transparence (code ouvert et disponible), documentation au travers d'exemples d'utilisation présentés dans des notebooks, et reproductibilité par la mise à disposition d'un environnement d'exécution (géré par Conda). De plus, nous avons démontré l'interopérabilité de notre librairie en réalisant une analyse différentielle d'expression avec EdgeR à partir de la table de comptages générée par EMOTE-tk.

Pour finir, notre librairie, EMOTE-tk, est conçue pour être évolutive, avec une architecture qui permet l'intégration de nouvelles analyses en fonction des besoins futurs. L'un des aspects clés de cette flexibilité réside dans le format de table de comptages utilisé pour stocker les résultats des analyses. Ce format offre une structure uniforme qui facilite l'intégration de nouvelles méthodes d'analyse ou de nouvelles fonctionnalités au sein de la librairie. Ainsi, EMOTE-tk offre non seulement des fonctionnalités avancées pour l'analyse des données RNA-Seq, mais elle vise également à s’enrichir pour répondre à des potentiels futurs problématiques. En travaillant ensemble pour identifier les besoins émergents et en proposant des solutions adaptées, nous contribuons à faire progresser la recherche dans le domaine de la génomique et à répondre aux défis scientifiques à venir.

4.4 Le mot de la fin

En conclusion, cette thèse met en évidence le rôle central de la RNase J dans la dégradation des ARN chez *S. aureus*, confirmant son statut d'actrice majeure dans ce processus. Nos recherches démontrent que, contrairement aux mécanismes de synthèse des ARN, qui sont très conservés même entre des organismes distants, les mécanismes de dégradation montrent une grande variabilité. En effet, bien qu’ils puissent impliquer des acteurs similaires (exoribonucléase, endoribonucléase et RppH), leur fonctionnement diffère même entre des organismes phylogénétiquement proches tels que *S. aureus* et *B. subtilis*. Cette observation souligne une adaptation spécifique de chaque organisme, reflétant des besoins évolutifs distincts. En explorant davantage les particularités des facteurs impliqués dans la dégradation des ARN et leur interaction avec d'autres régulateurs cellulaires, nous pourrions améliorer notre compréhension des réseaux de régulation génétique et potentiellement identifier de nouvelles cibles pour lutter contre les infections bactériennes résistantes aux antibiotiques. La complexité et la diversité des mécanismes de dégradation des ARN offrent ainsi un vaste champ d'investigation prometteur pour les années à venir.

Références

5 Références

Abbas-Aghababazadeh, F., Li, Q., & Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. *PloS One*, *13*(10), e0206312. https://doi.org/10.1371/journal.pone.0206312

Aboshkiwa, M., Al-Ani, B., Coleman, G., & Rowland, G. (1992). Cloning and physical mapping of the Staphylococcus aureus rplL, rpoB and rpoC genes, encoding ribosomal protein L7/L12 and RNA polymerase subunits β and β'. *Microbiology*, *138*(9), 1875-1880. https://doi.org/10.1099/00221287-138-9-1875

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., & Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Aït-Bara, S., Carpousis, A. J., & Quentin, Y. (2015). RNase E in the γ-Proteobacteria : Conservation of intrinsically disordered noncatalytic region and molecular evolution of microdomains. *Molecular Genetics and Genomics: MGG*, *290*(3), 847-862. https://doi.org/10.1007/s00438-014-0959-5

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)*, *31*(2), 166-169. https://doi.org/10.1093/bioinformatics/btu638

Anderson, K. L., & Dunman, P. M. (2009). Messenger RNA Turnover Processes in Escherichia coli, Bacillus subtilis, and Emerging Studies in Staphylococcus aureus. *International Journal of Microbiology*, *2009*, 525491. https://doi.org/10.1155/2009/525491

Anderson, K. L., Roux, C. M., Olson, M. W., Luong, T. T., Lee, C. Y., Olson, R., & Dunman, P. M. (2010). Characterizing the effects of inorganic acid and alkaline shock on the Staphylococcus aureus transcriptome and messenger RNA turnover. *FEMS Immunology & Medical Microbiology*, *60*(3), 208-250. https://doi.org/10.1111/j.1574-695X.2010.00736.x

Arraiano, C. M., Andrade, J. M., Domingues, S., Guinote, I. B., Malecki, M., Matos, R. G., Moreira, R. N., Pobre, V., Reis, F. P., Saramago, M., Silva, I. J., & Viegas, S. C. (2010). The critical role of RNA processing and degradation in the control of gene expression. *FEMS Microbiology Reviews*, *34*(5), 883-923. https://doi.org/10.1111/j.1574-6976.2010.00242.x

Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., & Scorilas, A. (2021). Third-Generation Sequencing : The Spearhead towards the Radical Transformation of Modern Genomics. *Life (Basel, Switzerland)*, *12*(1), 30. https://doi.org/10.3390/life12010030

Babraham Bioinformatics—FastQC A Quality Control tool for High Throughput Sequence Data. (s. d.). Consulté 12 juillet 2024, à l'adresse https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bandyra, K. J., Said, N., Pfeiffer, V., Górna, M. W., Vogel, J., & Luisi, B. F. (2012). The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. *Molecular Cell*, *47*(6), 943-953. https://doi.org/10.1016/j.molcel.2012.07.015

Bao, L., Hu, J., Zhan, B., Chi, M., Li, Z., Wang, S., Shan, C., Zhao, Z., Guo, Y., Ding, X., Ji, C., Tao, S., Ni, T., Zhang, X., Zhao, G., & Li, J. (2023). Structural insights into RNase J that

plays an essential role in Mycobacterium tuberculosis RNA metabolism. *Nature Communications*, *14*(1), 2280. https://doi.org/10.1038/s41467-023-38045-z

Barrios, H., Valderrama, B., & Morett, E. (1999). Compilation and analysis of σ54-dependent promoter sequences. *Nucleic Acids Research*, *27*(22), 4305-4313. https://doi.org/10.1093/nar/27.22.4305

Baumgardt, K., Gilet, L., Figaro, S., & Condon, C. (2018). The essential nature of YqfG, a YbeY homologue required for 3' maturation of Bacillus subtilis 16S ribosomal RNA is suppressed by deletion of RNase R. *Nucleic Acids Research*, *46*(16), 8605-8615. https://doi.org/10.1093/nar/gky488

Bechhofer, D. H., & Deutscher, M. P. (2019). Bacterial ribonucleases and their roles in RNA metabolism. *Critical reviews in biochemistry and molecular biology*, *54*(3), 242-300. https://doi.org/10.1080/10409238.2019.1651816

Beck, H. J., Fleming, I. M. C., & Janssen, G. R. (2016). 5'-Terminal AUGs in Escherichia coli mRNAs with Shine-Dalgarno Sequences : Identification and Analysis of Their Roles in Non-Canonical Translation Initiation. *PloS One*, *11*(7), e0160144. https://doi.org/10.1371/journal.pone.0160144

Bernstein, J. A., Lin, P.-H., Cohen, S. N., & Lin-Chao, S. (2004). Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(9), 2758-2763. https://doi.org/10.1073/pnas.0308747101

Bessman, M. J., Frick, D. N., & O'Handley, S. F. (1996). The MutT Proteins or "Nudix" Hydrolases, a Family of Versatile, Widely Distributed, "Housecleaning" Enzymes*. *Journal of Biological Chemistry*, *271*(41), 25059-25062. https://doi.org/10.1074/jbc.271.41.25059

Bischler, T., Hsieh, P., Resch, M., Liu, Q., Tan, H. S., Foley, P. L., Hartleib, A., Sharma, C. M., & Belasco, J. G. (2017). Identification of the RNA Pyrophosphohydrolase RppH of

Helicobacter pylori and Global Analysis of Its RNA Targets. *Journal of Biological Chemistry*, *292*(5), 1934-1950. https://doi.org/10.1074/jbc.M116.761171

Blatter, E. E., Ross, W., Tang, H., Gourse, R. L., & Ebright, R. H. (1994). Domain organization of RNA polymerase alpha subunit : C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. *Cell*, *78*(5), 889-896. https://doi.org/10.1016/s0092-8674(94)90682-3

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science (New York, N.Y.)*, *277*(5331), 1453-1462. https://doi.org/10.1126/science.277.5331.1453

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic : A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114-2120. https://doi.org/10.1093/bioinformatics/btu170

Bolstad, B. M., Irizarry, R. A., Astrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, *19*(2), 185-193. https://doi.org/10.1093/bioinformatics/19.2.185

Bonnin, R. A., & Bouloc, P. (2015). RNA Degradation in Staphylococcus aureus : Diversity of Ribonucleases and Their Impact. *International Journal of Genomics*, *2015*, 395753. https://doi.org/10.1155/2015/395753

Bram, R. J., Young, R. A., & Steitz, J. A. (1980). The ribonuclease III site flanking 23S sequences in the 30S ribosomal precursor RNA of E. coli. *Cell*, *19*(2), 393-401. https://doi.org/10.1016/0092-8674(80)90513-9

Brantl, S. (2015). Antisense-RNA mediated control of plasmid replication—pIP501 revisited. *Plasmid*, *78*, 4-16. https://doi.org/10.1016/j.plasmid.2014.07.004

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525-527. https://doi.org/10.1038/nbt.3519

Britton, R. A., Wen, T., Schaefer, L., Pellegrini, O., Uicker, W. C., Mathy, N., Tobin, C., Daou, R., Szyk, J., & Condon, C. (2007). Maturation of the 5' end of Bacillus subtilis 16S rRNA by the essential ribonuclease YkqC/RNase J1. *Molecular Microbiology*, *63*(1), 127-138. https://doi.org/10.1111/j.1365-2958.2006.05499.x

Broglia, L., Lécrivain, A.-L., Renault, T. T., Hahnke, K., Ahmed-Begrich, R., Le Rhun, A., & Charpentier, E. (2020). An RNA-seq based comparative approach reveals the transcriptome-wide interplay between 3′-to-5′ exoRNases and RNase Y. *Nature Communications*, *11*(1), 1587. https://doi.org/10.1038/s41467-020-15387-6

Broglia, L., Le Rhun, A., & Charpentier, E. (2023). Methodologies for bacterial ribonuclease characterization using RNA-seq. *FEMS Microbiology Reviews*, *47*(5), fuad049. https://doi.org/10.1093/femsre/fuad049

Broglia, L., Materne, S., Lécrivain, A.-L., Hahnke, K., Le Rhun, A., & Charpentier, E. (2018). RNase Y-mediated regulation of the streptococcal pyrogenic exotoxin B. *RNA Biology*, *15*(10), 1336-1347. https://doi.org/10.1080/15476286.2018.1532253

Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial transcription initiation. *Nature Reviews. Microbiology*, *2*(1), 57-65. https://doi.org/10.1038/nrmicro787

Browning, D. F., & Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews. Microbiology*, *14*(10), 638-650. https://doi.org/10.1038/nrmicro.2016.103

Bugrysheva, J. V., & Scott, J. R. (2010). The ribonucleases J1 and J2 are essential for growth and have independent roles in mRNA decay in Streptococcus pyogenes. *Molecular Microbiology*, *75*(3), 731-743. https://doi.org/10.1111/j.1365-2958.2009.07012.x

Burrows, M., L, D. J. W. D. I. G. I. T. A., Taylor, R. W., Wheeler, D., & Wheeler, D. (1994). *A Block-sorting Lossless Data Compression Algorithm*.

https://www.semanticscholar.org/paper/A-Block-sorting-Lossless-Data-Compression-Algorithm-Burrows-D.J.WheelerDIGITA/af56e6d4901dcd0f589bf969e604663d40f1be5d

Cahová, H., Winz, M.-L., Höfer, K., Nübel, G., & Jäschke, A. (2015). NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*, *519*(7543), 374-377. https://doi.org/10.1038/nature14020

Campagna, D., Telatin, A., Forcato, C., Vitulo, N., & Valle, G. (2013). PASS-bis : A bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads. *Bioinformatics (Oxford, England)*, *29*(2), 268-270. https://doi.org/10.1093/bioinformatics/bts675

Campo, C. D., Bartholomäus, A., Fedyunin, I., & Ignatova, Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLOS Genetics*, *11*(10), e1005613. https://doi.org/10.1371/journal.pgen.1005613

Cardinale, C. J., Washburn, R. S., Tadigotla, V. R., Brown, L. M., Gottesman, M. E., & Nudler, E. (2008). Termination factor Rho and its cofactors NusA and NusG silence foreign DNA in E. coli. *Science (New York, N.Y.)*, *320*(5878), 935-938. https://doi.org/10.1126/science.1152763

Carpousis, A. J. (2002). The Escherichia coli RNA degradosome : Structure, function and relationship in other ribonucleolytic multienzyme complexes. *Biochemical Society Transactions*, *30*(2), 150-155.

Carpousis, A. J. (2007). The RNA degradosome of Escherichia coli : An mRNA-degrading machine assembled on RNase E. *Annual Review of Microbiology*, *61*, 71-87. https://doi.org/10.1146/annurev.micro.61.080706.093440

Carpousis, A. J., Luisi, B. F., & McDowall, K. J. (2009). Endonucleolytic initiation of mRNA decay in Escherichia coli. *Progress in Molecular Biology and Translational Science*, *85*, 91-135. https://doi.org/10.1016/S0079-6603(08)00803-9

Carrier, M.-C., Morin, C., & Massé, E. (2018). On the Prowl : An In Vivo Method to Identify RNA Partners of a sRNA. *Methods in Enzymology*, *612*, 251-268. https://doi.org/10.1016/bs.mie.2018.09.006

Casas-Pastor, D., Müller, R. R., Jaenicke, S., Brinkrolf, K., Becker, A., Buttner, M. J., Gross, C. A., Mascher, T., Goesmann, A., & Fritz, G. (2021). Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family. *Nucleic Acids Research*, *49*(2), 986-1005. https://doi.org/10.1093/nar/gkaa1229

Cascante-Estepa, N., Gunka, K., & Stülke, J. (2016). Localization of Components of the RNA-Degrading Machine in Bacillus subtilis. *Frontiers in Microbiology*, *7*. https://doi.org/10.3389/fmicb.2016.01492

Castaño, J. G., Tobian, J. A., & Zasloff, M. (1985). Purification and characterization of an endonuclease from Xenopus laevis ovaries which accurately processes the 3' terminus of human pre-tRNA-Met(i) (3' pre-tRNase). *The Journal of Biological Chemistry*, *260*(15), 9002-9008.

Celesnik, H., Deana, A., & Belasco, J. G. (2007). Initiation of RNA decay in Escherichia coli by 5' pyrophosphate removal. *Molecular Cell*, *27*(1), 79-90. https://doi.org/10.1016/j.molcel.2007.05.038

Chao, Y., Li, L., Girodat, D., Förstner, K. U., Said, N., Corcoran, C., Śmiga, M., Papenfort, K., Reinhardt, R., Wieden, H.-J., Luisi, B. F., & Vogel, J. (2017). In Vivo Cleavage Map Illuminates the Central Role of RNase E in Coding and Non-coding RNA Pathways. *Molecular Cell*, *65*(1), 39-51. https://doi.org/10.1016/j.molcel.2016.11.002

Chaudhuri, R. R., Allen, A. G., Owen, P. J., Shalom, G., Stone, K., Harrison, M., Burgis, T. A., Lockyer, M., Garcia-Lara, J., Foster, S. J., Pleasance, S. J., Peters, S. E., Maskell, D. J., & Charles, I. G. (2009). Comprehensive identification of essential Staphylococcus aureus genes using Transposon-Mediated Differential Hybridisation (TMDH). *BMC Genomics*, *10*, 291. https://doi.org/10.1186/1471-2164-10-291

Chen, X., Liu, N., Khajotia, S., Qi, F., & Merritt, J. (2015). RNases J1 and J2 are critical pleiotropic regulators in Streptococcus mutans. *Microbiology (Reading, England)*, *161*(Pt 4), 797-806. https://doi.org/10.1099/mic.0.000039

Chen, Z., Itzek, A., Malke, H., Ferretti, J. J., & Kreth, J. (2013). Multiple roles of RNase Y in Streptococcus pyogenes mRNA processing and degradation. *Journal of Bacteriology*, *195*(11), 2585-2594. https://doi.org/10.1128/JB.00097-13

Cheng, Z.-F., & Deutscher, M. P. (2002). Purification and Characterization of the *Escherichia coli* Exoribonuclease RNase R. *Journal of Biological Chemistry*, *277*(24), 21624-21629. https://doi.org/10.1074/jbc.M202942200

Cheng, Z.-F., & Deutscher, M. P. (2003). Quality control of ribosomal RNA mediated by polynucleotide phosphorylase and RNase R. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(11), 6388-6393. https://doi.org/10.1073/pnas.1231041100

Cheng, Z.-F., & Deutscher, M. P. (2005). An important role for RNase R in mRNA decay. *Molecular Cell*, *17*(2), 313-318. https://doi.org/10.1016/j.molcel.2004.11.048

Chevalier, C., Huntzinger, E., Fechter, P., Boisset, S., Vandenesch, F., Romby, P., & Geissmann, T. (2008). Staphylococcus aureus endoribonuclease III purification and properties. *Methods in Enzymology*, *447*, 309-327. https://doi.org/10.1016/S0076-6879(08)02216-7

Cho, B.-K., Kim, D., Knight, E. M., Zengler, K., & Palsson, B. O. (2014). Genome-scale reconstruction of the sigma factor network in Escherichia coli : Topology and functional states. *BMC Biology*, *12*(1), 4. https://doi.org/10.1186/1741-7007-12-4

Clarke, J. E., Kime, L., Romero A, D., & McDowall, K. J. (2014). Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli. *Nucleic Acids Research*, *42*(18), 11733-11751. https://doi.org/10.1093/nar/gku808

Clouet-d'Orval, B., Rinaldi, D., Quentin, Y., & Carpousis, A. J. (2010). Euryarchaeal β-CASP Proteins with Homology to Bacterial RNase J Have 5′- to 3′-Exoribonuclease Activity. *The*

Journal of Biological Chemistry, *285*(23), 17574-17583. https://doi.org/10.1074/jbc.M109.095117

Commichau, F. M., Rothe, F. M., Herzberg, C., Wagner, E., Hellwig, D., Lehnik-Habrink, M., Hammer, E., Völker, U., & Stülke, J. (2009). Novel activities of glycolytic enzymes in Bacillus subtilis : Interactions with essential proteins involved in mRNA processing. *Molecular & Cellular Proteomics: MCP*, *8*(6), 1350-1360. https://doi.org/10.1074/mcp.M800546-MCP200

Condon, C., Brechemier-Baey, D., Beltchev, B., Grunberg-Manago, M., & Putzer, H. (2001). Identification of the gene encoding the 5S ribosomal RNA maturase in Bacillus subtilis : Mature 5S rRNA is dispensable for ribosome function. *RNA (New York, N.Y.)*, *7*(2), 242-253. https://doi.org/10.1017/s1355838201002163

Condon, C., Pellegrini, O., Gilet, L., Durand, S., & Braun, F. (2021). Walking from E. coli to B. subtilis, one ribonuclease at a time. *Comptes Rendus Biologies*, *344*(4), 357-371. https://doi.org/10.5802/crbiol.70

Condon, C., & Putzer, H. (2002). The phylogenetic distribution of bacterial ribonucleases. *Nucleic Acids Research*, *30*(24), 5339-5346. https://doi.org/10.1093/nar/gkf691

Condon, C., Rourera, J., Brechemier-Baey, D., & Putzer, H. (2002). Ribonuclease M5 has few, if any, mRNA substrates in Bacillus subtilis. *Journal of Bacteriology*, *184*(10), 2845-2849. https://doi.org/10.1128/JB.184.10.2845-2849.2002

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo : A sequence logo generator. *Genome Research*, *14*(6), 1188-1190. https://doi.org/10.1101/gr.849004

Davies, B. W., Köhrer, C., Jacob, A. I., Simmons, L. A., Zhu, J., Aleman, L. M., Rajbhandary, U. L., & Walker, G. C. (2010). Role of Escherichia coli YbeY, a highly conserved protein, in rRNA processing. *Molecular Microbiology*, *78*(2), 506-518. https://doi.org/10.1111/j.1365-2958.2010.07351.x

Deana, A., Celesnik, H., & Belasco, J. G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature*, *451*(7176), 355-358. https://doi.org/10.1038/nature06475

Deikus, G., & Bechhofer, D. H. (2011). 5' End-independent RNase J1 endonuclease cleavage of Bacillus subtilis model RNA. *The Journal of Biological Chemistry*, *286*(40), 34932-34940. https://doi.org/10.1074/jbc.M111.287409

DeLoughery, A., Lalanne, J.-B., Losick, R., & Li, G.-W. (2018). Maturation of polycistronic mRNAs by the endoribonuclease RNase Y and its associated Y-complex in Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), E5585-E5594. https://doi.org/10.1073/pnas.1803283115

Dermit, M., Dodel, M., & K. Mardakheh, F. (2017). Methods for monitoring and measurement of protein translation in time and space. *Molecular BioSystems*, *13*(12), 2477-2488. https://doi.org/10.1039/C7MB00476A

Deutscher, M. P. (2009). Maturation and degradation of ribosomal RNA in bacteria. *Progress in Molecular Biology and Translational Science*, *85*, 369-391. https://doi.org/10.1016/S0079-6603(08)00809-X

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR : Ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15-21. https://doi.org/10.1093/bioinformatics/bts635

Dodt, M., Roehr, J. T., Ahmed, R., & Dieterich, C. (2012). FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*, *1*(3), 895-905. https://doi.org/10.3390/biology1030895

Dominski, Z., Carpousis, A. J., & Clouet-d'Orval, B. (2013). Emergence of the β-CASP ribonucleases : Highly conserved and ubiquitous metallo-enzymes involved in messenger RNA maturation and degradation. *Biochimica Et Biophysica Acta*, *1829*(6-7), 532-551. https://doi.org/10.1016/j.bbagrm.2013.01.010

Dorléans, A., Li de la Sierra-Gallay, I., Piton, J., Zig, L., Gilet, L., Putzer, H., & Condon, C. (2011). Molecular basis for the recognition and cleavage of RNA by the bifunctional 5'-3' exo/endoribonuclease RNase J. *Structure (London, England: 1993)*, *19*(9), 1252-1261. https://doi.org/10.1016/j.str.2011.06.018

Durand, S., Gilet, L., Bessières, P., Nicolas, P., & Condon, C. (2012). Three essential ribonucleases-RNase Y, J1, and III-control the abundance of a majority of Bacillus subtilis mRNAs. *PLoS Genetics*, *8*(3), e1002520. https://doi.org/10.1371/journal.pgen.1002520

Duval, M., Korepanov, A., Fuchsbauer, O., Fechter, P., Haller, A., Fabbretti, A., Choulier, L., Micura, R., Klaholz, B. P., Romby, P., Springer, M., & Marzi, S. (2013). Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biology*, *11*(12), e1001731. https://doi.org/10.1371/journal.pbio.1001731

Ebbert, M. T. W., Wadsworth, M. E., Staley, L. A., Hoyt, K. L., Pickett, B., Miller, J., Duce, J., Alzheimer's Disease Neuroimaging Initiative, Kauwe, J. S. K., & Ridge, P. G. (2016). Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, *17 Suppl 7*(Suppl 7), 239. https://doi.org/10.1186/s12859-016-1097-3

Ebright, R. H., & Busby, S. (1995). The Escherichia coli RNA polymerase alpha subunit : Structure and function. *Current Opinion in Genetics & Development*, *5*(2), 197-203. https://doi.org/10.1016/0959-437x(95)80008-5

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Erlacher, M. D., Lang, K., Wotzel, B., Rieder, R., Micura, R., & Polacek, N. (2006). Efficient ribosomal peptidyl transfer critically relies on the presence of the ribose 2'-OH at A2451 of 23S rRNA. *Journal of the American Chemical Society*, *128*(13), 4453-4459. https://doi.org/10.1021/ja0588454

Even, S. (2005). Ribonucleases J1 and J2 : Two novel endoribonucleases in B.subtilis with functional homology to E.coli RNase E. *Nucleic Acids Research*, *33*(7), 2141-2152. https://doi.org/10.1093/nar/gki505

Farnham, P. J., & Platt, T. (1981). Rho-independent termination : Dyad symmetry in DNA causes RNA polymerase to pause during transcription in vitro. *Nucleic Acids Research*, *9*(3), 563-577. https://doi.org/10.1093/nar/9.3.563

Fazal, F. M., Koslover, D. J., Luisi, B. F., & Block, S. M. (2015). Direct observation of processive exoribonuclease motion using optical tweezers. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(49), 15101-15106. https://doi.org/10.1073/pnas.1514028112

Feklístov, A., Sharon, B. D., Darst, S. A., & Gross, C. A. (2014). Bacterial sigma factors : A historical, structural, and genomic perspective. *Annual Review of Microbiology*, *68*, 357-376. https://doi.org/10.1146/annurev-micro-092412-155737

Feng, H., Conneely, K. N., & Wu, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, *42*(8), e69-e69. https://doi.org/10.1093/nar/gku154

Figaro, S., Durand, S., Gilet, L., Cayet, N., Sachse, M., & Condon, C. (2013). Bacillus subtilis Mutants with Knockouts of the Genes Encoding Ribonucleases RNase Y and RNase J1 Are Viable, with Major Defects in Cell Morphology, Sporulation, and Competence. *Journal of bacteriology*, *195*. https://doi.org/10.1128/JB.00164-13

Foley, P. L., Hsieh, P., Luciano, D. J., & Belasco, J. G. (2015). Specificity and evolutionary conservation of the Escherichia coli RNA pyrophosphohydrolase RppH. *The Journal of Biological Chemistry*, *290*(15), 9478-9486. https://doi.org/10.1074/jbc.M114.634659

Fox, G. E. (2010). Origin and evolution of the ribosome. *Cold Spring Harbor Perspectives in Biology*, *2*(9), a003483. https://doi.org/10.1101/cshperspect.a003483

Frindert, J., Kahloon, M. A., Zhang, Y., Ahmed, Y. L., Sinning, I., & Jäschke, A. (2019). YvcI from Bacillus subtilis has in vitro RNA pyrophosphohydrolase activity. *The Journal of Biological Chemistry*, *294*(52), 19967-19977. https://doi.org/10.1074/jbc.RA119.011485

Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, *19*(1), 531. https://doi.org/10.1186/s12864-018-4933-1

Gao, P., Pinkston, K. L., Bourgogne, A., Murray, B. E., van Hoof, A., & Harvey, B. R. (2017). Functional studies of E. faecalis RNase J2 and its role in virulence and fitness. *PLoS ONE*, *12*(4), e0175212. https://doi.org/10.1371/journal.pone.0175212

Garber, R. L., & Gage, L. P. (1979). Transcription of a cloned Bombyx mori tRNA2Ala gene : Nucleotide sequence of the tRNA precursor and its processing in vitro. *Cell*, *18*(3), 817-828. https://doi.org/10.1016/0092-8674(79)90134-x

Garrey, S. M., Blech, M., Riffell, J. L., Hankins, J. S., Stickney, L. M., Diver, M., Hsu, Y.-H. R., Kunanithy, V., & Mackie, G. A. (2009). Substrate binding and active site residues in RNases E and G : Role of the 5'-sensor. *The Journal of Biological Chemistry*, *284*(46), 31843-31850. https://doi.org/10.1074/jbc.M109.063263

Giraud, C., Hausmann, S., Lemeille, S., Prados, J., Redder, P., & Linder, P. (2015). The C-terminal region of the RNA helicase CshA is required for the interaction with the degradosome and turnover of bulk RNA in the opportunistic pathogen Staphylococcus aureus. *RNA Biology*, *12*(6), 658-674. https://doi.org/10.1080/15476286.2015.1035505

Giuliano, M. G., & Engl, C. (2021). *The Lifecycle of Ribosomal RNA in Bacteria* (p. 27-51). https://doi.org/10.1007/978-3-030-76571-2_2

Godefroy, T. (1970). Kinetics of polymerization and phosphorolysis reactions of Escherichia coli polynucleotide phosphorylase. Evidence for multiple binding of polynucleotide in phosphorolysis. *European Journal of Biochemistry*, *14*(2), 222-231. https://doi.org/10.1111/j.1432-1033.1970.tb00281.x

Górna, M. W., Carpousis, A. J., & Luisi, B. F. (2012). From conformational chaos to robust regulation : The structure and function of the multi-enzyme RNA degradosome. *Quarterly Reviews of Biophysics*, *45*(2), 105-145. https://doi.org/10.1017/S003358351100014X

Gorski, S. A., Vogel, J., & Doudna, J. A. (2017). RNA-based recognition and targeting : Sowing the seeds of specificity. *Nature Reviews. Molecular Cell Biology*, *18*(4), 215-228. https://doi.org/10.1038/nrm.2016.174

Grass, P. (2009). Experimental Design. In *Batch Effects and Noise in Microarray Experiments* (p. 19-31). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470685983.ch3

Green, N. J., Grundy, F. J., & Henkin, T. M. (2010). The T box mechanism : tRNA as a regulatory molecule. *FEBS Letters*, *584*(2), 318-324. https://doi.org/10.1016/j.febslet.2009.11.056

Gruber, T. M., & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annual Review of Microbiology*, *57*, 441-466. https://doi.org/10.1146/annurev.micro.57.030502.090913

Grunau, C., Clark, S. J., & Rosenthal, A. (2001). Bisulfite genomic sequencing : Systematic investigation of critical experimental parameters. *Nucleic Acids Research*, *29*(13), E65-65. https://doi.org/10.1093/nar/29.13.e65

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 Pt 2), 849-857. https://doi.org/10.1016/0092-8674(83)90117-4

Gutiérrez-Preciado, A., Henkin, T. M., Grundy, F. J., Yanofsky, C., & Merino, E. (2009). Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiology and Molecular Biology Reviews: MMBR*, *73*(1), 36-61. https://doi.org/10.1128/MMBR.00026-08

Haldenwang, W. G. (1995). The sigma factors of Bacillus subtilis. *Microbiological Reviews*, *59*(1), 1-30. https://doi.org/10.1128/mr.59.1.1-30.1995

Harbaugh, S., Goodson, M., Chushak, Y., Chávez, J., & Kelley-Loughnane, N. (2019). Riboswitches as Sensor Entities. In G. Thouand (Éd.), *Handbook of Cell Biosensors* (p. 1-30). Springer International Publishing. https://doi.org/10.1007/978-3-319-47405-2_121-1

Hardwick, S. W., Chan, V. S. Y., Broadhurst, R. W., & Luisi, B. F. (2011). An RNA degradosome assembly in Caulobacter crescentus. *Nucleic Acids Research*, *39*(4), 1449-1459. https://doi.org/10.1093/nar/gkq928

Harley, C. B., & Reynolds, R. P. (1987). Analysis of E. coli promoter sequences. *Nucleic Acids Research*, *15*(5), 2343-2361. https://doi.org/10.1093/nar/15.5.2343

Hausmann, S., Guimarães, V. A., Garcin, D., Baumann, N., Linder, P., & Redder, P. (2017). Both exo- and endo-nucleolytic activities of RNase J1 from *Staphylococcus* aureus are manganese dependent and active on triphosphorylated 5′-ends. *RNA Biology*, *14*(10), Article 10. https://doi.org/10.1080/15476286.2017.1300223

Helmann, J. D. (2002). The extracytoplasmic function (ECF) sigma factors. *Advances in Microbial Physiology*, *46*, 47-110. https://doi.org/10.1016/s0065-2911(02)46002-x

Helsen, C., Dubois, V., Verfaillie, A., Young, J., Trekels, M., Vancraenenbroeck, R., De Maeyer, M., & Claessens, F. (2012). Evidence for DNA-binding domain—Ligand-binding domain communications in the androgen receptor. *Molecular and Cellular Biology*, *32*(15), 3033-3043. https://doi.org/10.1128/MCB.00151-12

Helwak, A., & Tollervey, D. (2014). Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nature Protocols*, *9*(3), 711-728. https://doi.org/10.1038/nprot.2014.043

Herskovitz, M. A., & Bechhofer, D. H. (2000). Endoribonuclease RNase III is essential in Bacillus subtilis. *Molecular Microbiology*, *38*(5), 1027-1033. https://doi.org/10.1046/j.1365-2958.2000.02185.x

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chemie / Chemical Monthly*, *125*(2), 167-188. https://doi.org/10.1007/BF00818163

Holmqvist, E., & Vogel, J. (2018). RNA-binding proteins in bacteria. *Nature Reviews. Microbiology*, *16*(10), 601-615. https://doi.org/10.1038/s41579-018-0049-5

Hsiao, Y.-Y., Fang, W.-H., Lee, C.-C., Chen, Y.-P., & Yuan, H. S. (2014). Structural Insights Into DNA Repair by RNase T—An Exonuclease Processing 3′ End of Structured DNA in Repair Pathways. *PLoS Biology*, *12*(3), e1001803. https://doi.org/10.1371/journal.pbio.1001803

Hsieh, P., Richards, J., Liu, Q., & Belasco, J. G. (2013). Specificity of RppH-dependent RNA degradation in Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 8864-8869. https://doi.org/10.1073/pnas.1222670110

Huang, H., Liao, J., & Cohen, S. N. (1998). Poly(A)- and poly(U)-specific RNA 3' tail shortening by E. coli ribonuclease E. *Nature*, *391*(6662), 99-102. https://doi.org/10.1038/34219

Huch, S., Nersisyan, L., Ropat, M., Barret, D., Wang, J., Huerta-Cepas, J., Wei, W., Steinmetz, L. M., Engstrand, L., & Pelechano, V. (2021). *RNA degradation analysis reveals ribosome dynamics in complex microbiome samples* (p. 2021.04.08.439066). bioRxiv. https://doi.org/10.1101/2021.04.08.439066

Hug, H., & Schuler, R. (2003). Measurement of the Number of Molecules of a Single mRNA Species in a Complex mRNA Preparation. *Journal of Theoretical Biology*, *221*(4), 615-624. https://doi.org/10.1006/jtbi.2003.3211

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C., & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 1-6. https://doi.org/10.1038/nmicrobiol.2016.48

Hui, M. P., Foley, P. L., & Belasco, J. G. (2014). Messenger RNA degradation in bacterial cells. *Annual Review of Genetics*, *48*, 537-559. https://doi.org/10.1146/annurev-genet-120213-092340

Hunt, A., Rawlins, J. P., Thomaides, H. B., & Errington, J. (2006). Functional analysis of 11 putative essential genes in Bacillus subtilis. *Microbiology (Reading, England)*, *152*(Pt 10), 2895-2907. https://doi.org/10.1099/mic.0.29152-0

Ibarra, J. A., García-Zacarias, C. M., Lara-Ochoa, C., Carabarin-Lima, A., Tecpanecatl-Xihuitl, J. S., Perez-Rueda, E., Martínez-Laguna, Y., & Puente, J. L. (2013). Further Characterization of Functional Domains of PerA, Role of Amino and Carboxy Terminal Domains in DNA Binding. *PLOS ONE*, *8*(2), e56977. https://doi.org/10.1371/journal.pone.0056977

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (New York, N.Y.)*, *324*(5924), 218-223. https://doi.org/10.1126/science.1168978

Irastortza-Olaziregi, M., & Amster-Choder, O. (2020). Coupled Transcription-Translation in Prokaryotes : An Old Couple With New Surprises. *Frontiers in Microbiology*, *11*, 624830. https://doi.org/10.3389/fmicb.2020.624830

Ishino, Y., Krupovic, M., & Forterre, P. (2018). History of CRISPR-Cas from Encounter with a Mysterious Repeated Sequence to Genome Editing Technology. *Journal of Bacteriology*, *200*(7), e00580-17. https://doi.org/10.1128/JB.00580-17

Jacob, A. I., Köhrer, C., Davies, B. W., RajBhandary, U. L., & Walker, G. C. (2013). Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. *Molecular Cell*, *49*(3), 427-438. https://doi.org/10.1016/j.molcel.2012.11.025

Jäger, S., Fuhrmann, O., Heck, C., Hebermehl, M., Schiltz, E., Rauhut, R., & Klug, G. (2001). An mRNA degrading complex in Rhodobacter capsulatus. *Nucleic Acids Research*, *29*(22), 4581-4588. https://doi.org/10.1093/nar/29.22.4581

Jiang, F., & Doudna, J. A. (2015). The Structural Biology of CRISPR-Cas Systems. *Current opinion in structural biology*, *30*, 100-111. https://doi.org/10.1016/j.sbi.2015.02.002

Jiang, X., & Belasco, J. G. (2004). Catalytic activation of multimeric RNase E and RNase G by 5'-monophosphorylated RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(25), 9211-9216. https://doi.org/10.1073/pnas.0401382101

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118-127. https://doi.org/10.1093/biostatistics/kxj037

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

Kaberdin, V. R., Miczak, A., Jakobsen, J. S., Lin-Chao, S., McDowall, K. J., & von Gabain, A. (1998). The endoribonucleolytic N-terminal half of Escherichia coli RNase E is evolutionarily conserved in Synechocystis sp. And other bacteria but not the C-terminal half, which is sufficient for degradosome assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(20), 11637-11642. https://doi.org/10.1073/pnas.95.20.11637

Kaito, C., Kurokawa, K., Matsumoto, Y., Terao, Y., Kawabata, S., Hamada, S., & Sekimizu, K. (2005). Silkworm pathogenic bacteria infection model for identification of novel virulence genes. *Molecular Microbiology*, *56*(4), 934-944. https://doi.org/10.1111/j.1365-2958.2005.04596.x

Keller, A. N., Yang, X., Wiedermannová, J., Delumeau, O., Krásný, L., & Lewis, P. J. (2014). ε, a new subunit of RNA polymerase found in gram-positive bacteria. *Journal of Bacteriology*, *196*(20), 3622-3632. https://doi.org/10.1128/JB.02020-14

Khemici, V., Poljak, L., Luisi, B. F., & Carpousis, A. J. (2008). The RNase E of Escherichia coli is a membrane-binding protein. *Molecular Microbiology*, *70*(4), 799-813. https://doi.org/10.1111/j.1365-2958.2008.06454.x

Khemici, V., Prados, J., Linder, P., & Redder, P. (2015). Decay-Initiating Endoribonucleolytic Cleavage by RNase Y Is Kept under Tight Control via Sequence Preference and Sub-cellular Localisation. *PLoS Genetics*, *11*(10). https://doi.org/10.1371/journal.pgen.1005577

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907-915. https://doi.org/10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2 : Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36

Kime, L., Jourdan, S. S., & McDowall, K. J. (2008). Identifying and characterizing substrates of the RNase E/G family of enzymes. *Methods in Enzymology*, *447*, 215-241. https://doi.org/10.1016/S0076-6879(08)02212-X

Kleffmann, T., Russenberger, D., von Zychlinski, A., Christopher, W., Sjölander, K., Gruissem, W., & Baginsky, S. (2004). The Arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions. *Current Biology: CB*, *14*(5), 354-362. https://doi.org/10.1016/j.cub.2004.02.039

Koonin, E. V. (1993). A highly conserved sequence motif defining the family of MutT-related proteins from eubacteria, eukaryotes and viruses. *Nucleic Acids Research*, *21*(20), 4847. https://doi.org/10.1093/nar/21.20.4847

Kormanec, J. (2022). Bacterial Regulatory Proteins. *International Journal of Molecular Sciences*, *23*(12). https://doi.org/10.3390/ijms23126854

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520-2522. https://doi.org/10.1093/bioinformatics/bts480

Kothe, U., Wieden, H.-J., Mohr, D., & Rodnina, M. V. (2004). Interaction of helix D of elongation factor Tu with helices 4 and 5 of protein L7/12 on the ribosome. *Journal of Molecular Biology*, *336*(5), 1011-1021. https://doi.org/10.1016/j.jmb.2003.12.080

Kozak, M. (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiological Reviews*, *47*(1), 1-45. https://doi.org/10.1128/mr.47.1.1-45.1983

Kristoffersen, S. M., Haase, C., Weil, M. R., Passalacqua, K. D., Niazi, F., Hutchison, S. K., Desany, B., Kolstø, A.-B., Tourasse, N. J., Read, T. D., & Økstad, O. A. (2012). Global mRNA decay analysis at single nucleotide resolution reveals segmental and positional degradation patterns in a Gram-positive bacterium. *Genome Biology*, *13*(4), R30. https://doi.org/10.1186/gb-2012-13-4-r30

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951-969. https://doi.org/10.1101/pdb.top084970

Laalami, S., Bessières, P., Rocca, A., Zig, L., Nicolas, P., & Putzer, H. (2013). Bacillus subtilis RNase Y activity in vivo analysed by tiling microarrays. *PloS One*, *8*(1), e54062. https://doi.org/10.1371/journal.pone.0054062

Laalami, S., Zig, L., & Putzer, H. (2014). Initiation of mRNA decay in bacteria. *Cellular and Molecular Life Sciences: CMLS*, *71*(10), 1799-1828. https://doi.org/10.1007/s00018-013-1472-4

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359. https://doi.org/10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lasa, I., Toledo-Arana, A., Dobin, A., Villanueva, M., de los Mozos, I. R., Vergara-Irigaray, M., Segura, V., Fagegaltier, D., Penadés, J. R., Valle, J., Solano, C., & Gingeras, T. R. (2011). Genome-wide antisense transcription drives mRNA processing in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(50), 20172-20177. https://doi.org/10.1073/pnas.1113521108

Lasa, I., Toledo-Arana, A., & Gingeras, T. R. (2012). An effort to make sense of antisense transcription in bacteria. *RNA Biology*, *9*(8), 1039-1044. https://doi.org/10.4161/rna.21167

Le Rhun, A., Beer, Y. Y., Reimegård, J., Chylinski, K., & Charpentier, E. (2016). RNA sequencing uncovers antisense RNAs and novel small RNAs in Streptococcus pyogenes. *RNA Biology*, *13*(2), 177-195. https://doi.org/10.1080/15476286.2015.1110674

Lee, D. J., Minchin, S. D., & Busby, S. J. W. (2012). Activating transcription in bacteria. *Annual Review of Microbiology*, *66*, 125-152. https://doi.org/10.1146/annurev-micro-092611-150012

Lee, K., & Cohen, S. N. (2003). A Streptomyces coelicolor functional orthologue of Escherichia coli RNase E shows shuffling of catalytic and PNPase-binding domains. *Molecular Microbiology*, *48*(2), 349-360. https://doi.org/10.1046/j.1365-2958.2003.03435.x

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, *28*(6), 882-883. https://doi.org/10.1093/bioinformatics/bts034

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch

effects in high-throughput data. *Nature reviews. Genetics*, *11*(10), 10.1038/nrg2825. https://doi.org/10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics*, *3*(9), e161. https://doi.org/10.1371/journal.pgen.0030161

Lehnik-Habrink, M., Pförtner, H., Rempeters, L., Pietack, N., Herzberg, C., & Stülke, J. (2010). The RNA degradosome in Bacillus subtilis : Identification of CshA as the major RNA helicase in the multiprotein complex. *Molecular Microbiology*, *77*(4), 958-971. https://doi.org/10.1111/j.1365-2958.2010.07264.x

Leiva, L. E., Elgamal, S., Leidel, S. A., Orellana, O., Ibba, M., & Katz, A. (2022). Oxidative stress strongly restricts the effect of codon choice on the efficiency of protein synthesis in Escherichia coli. *Frontiers in Microbiology*, *13*, 1042675. https://doi.org/10.3389/fmicb.2022.1042675

Leskinen, K., Blasdel, B. G., Lavigne, R., & Skurnik, M. (2016). RNA-Sequencing Reveals the Progression of Phage-Host Interactions between φR1-37 and Yersinia enterocolitica. *Viruses*, *8*(4), 111. https://doi.org/10.3390/v8040111

Li, B., & Dewey, C. N. (2011). RSEM : Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li de la Sierra-Gallay, I., Zig, L., Jamalli, A., & Putzer, H. (2008). Structural insights into the dual activity of RNase J. *Nature Structural & Molecular Biology*, *15*(2), 206-212. https://doi.org/10.1038/nsmb.1376

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Z., & Deutscher, M. P. (1996). Maturation pathways for E. coli tRNA precursors : A random multienzyme process in vivo. *Cell*, *86*(3), 503-512. https://doi.org/10.1016/s0092-8674(00)80123-3

Li, Z., & Deutscher, M. P. (2002). RNase E plays an essential role in the maturation of Escherichia coli tRNA precursors. *RNA*, *8*(1), 97-109.

Li, Z., Pandit, S., & Deutscher, M. P. (1999). RNase G (CafA protein) and RNase E are both required for the 5' maturation of 16S ribosomal RNA. *The EMBO Journal*, *18*(10), 2878-2885. https://doi.org/10.1093/emboj/18.10.2878

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts : An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923-930. https://doi.org/10.1093/bioinformatics/btt656

Linder, P., Lemeille, S., & Redder, P. (2014). Transcriptome-Wide Analyses of 5′-Ends in RNase J Mutants of a Gram-Positive Pathogen Reveal a Role in RNA Maturation, Regulation and Degradation. *PLOS Genetics*, *10*(2), e1004207. https://doi.org/10.1371/journal.pgen.1004207

Lioliou, E., Sharma, C. M., Caldelari, I., Helfer, A.-C., Fechter, P., Vandenesch, F., Vogel, J., & Romby, P. (2012). Global regulatory functions of the Staphylococcus aureus endoribonuclease III in gene expression. *PLoS Genetics*, *8*(6), e1002782. https://doi.org/10.1371/journal.pgen.1002782

Liu, Y., Zhou, J., & White, K. P. (2014). RNA-seq differential expression studies : More sequence or more replication? *Bioinformatics (Oxford, England)*, *30*(3), 301-304. https://doi.org/10.1093/bioinformatics/btt688

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26. https://doi.org/10.1186/1748-7188-6-26

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, Y., Shen, Y., Warren, W., & Walter, R. (2016). *Next Generation Sequencing in Aquatic Models*. https://doi.org/10.5772/61657

Luciano, D. J., Vasilyev, N., Richards, J., Serganov, A., & Belasco, J. G. (2018). Importance of a diphosphorylated intermediate for RppH-dependent RNA degradation. *RNA Biology*, *15*(6), 703-706. https://doi.org/10.1080/15476286.2018.1460995

Luttinger, A., Hahn, J., & Dubnau, D. (1996). Polynucleotide phosphorylase is necessary for competence development in Bacillus subtilis. *Molecular Microbiology*, *19*(2), 343-356. https://doi.org/10.1046/j.1365-2958.1996.380907.x

Lybecker, M., Zimmermann, B., Bilusic, I., Tukhtubaeva, N., & Schroeder, R. (2014). The double-stranded transcriptome of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(8), 3134-3139. https://doi.org/10.1073/pnas.1315974111

Ma, M., Welch, R. D., & Garza, A. G. (2021). The σ54 system directly regulates bacterial natural product genes. *Scientific Reports*, *11*(1), 4771. https://doi.org/10.1038/s41598-021-84057-4

Mackie, G. A. (1998). Ribonuclease E is a 5'-end-dependent endonuclease. *Nature*, *395*(6703), 720-723. https://doi.org/10.1038/27246

Madan Babu, M., & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Research*, *31*(4), 1234-1244. https://doi.org/10.1093/nar/gkg210

Mandal, M., & Breaker, R. R. (2004). Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural & Molecular Biology*, *11*(1), 29-35. https://doi.org/10.1038/nsmb710

Mansour, M., Giudice, E., Xu, X., Akarsu, H., Bordes, P., Guillet, V., Bigot, D.-J., Slama, N., D'urso, G., Chat, S., Redder, P., Falquet, L., Mourey, L., Gillet, R., & Genevaux, P. (2022). Substrate recognition and cryo-EM structure of the ribosome-bound TAC toxin of Mycobacterium tuberculosis. *Nature Communications*, *13*(1), 2641. https://doi.org/10.1038/s41467-022-30373-w

Marincola, G., & Wolz, C. (2017). Downstream element determines RNase Y cleavage of the saePQRS operon in Staphylococcus aureus. *Nucleic Acids Research*, *45*(10), 5980-5994. https://doi.org/10.1093/nar/gkx296

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), Article 1. https://doi.org/10.14806/ej.17.1.200

Martini, M. C., Sun, H., & Shell, S. S. (2021). RNA Sequencing for Transcript 5′-End Mapping in Mycobacteria. In T. Parish & A. Kumar (Éds.), *Mycobacteria Protocols* (p. 513-531). Springer US. https://doi.org/10.1007/978-1-0716-1460-0_22

Mathy, N., Bénard, L., Pellegrini, O., Daou, R., Wen, T., & Condon, C. (2007). 5'-to-3' exoribonuclease activity in bacteria : Role of RNase J1 in rRNA maturation and 5' stability of mRNA. *Cell*, *129*(4), 681-692. https://doi.org/10.1016/j.cell.2007.02.051

Mathy, N., Hébert, A., Mervelet, P., Bénard, L., Dorléans, A., Li de la Sierra-Gallay, I., Noirot, P., Putzer, H., & Condon, C. (2010). Bacillus subtilis ribonucleases J1 and J2 form a complex with altered enzyme behaviour. *Molecular Microbiology*, *75*(2), 489-498. https://doi.org/10.1111/j.1365-2958.2009.07004.x

Mayer, M., Schiffer, S., & Marchfelder, A. (2000). tRNA 3' processing in plants : Nuclear and mitochondrial activities differ. *Biochemistry*, *39*(8), 2096-2105. https://doi.org/10.1021/bi992253e

McDowall, K. J., Lin-Chao, S., & Cohen, S. N. (1994). A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *The Journal of Biological Chemistry*, *269*(14), 10790-10796.

McLennan, A. G. (2013). Substrate ambiguity among the nudix hydrolases : Biologically significant, evolutionary remnant, or both? *Cellular and Molecular Life Sciences: CMLS*, *70*(3), 373-385. https://doi.org/10.1007/s00018-012-1210-3

Mediati, D. G., Wong, J. L., Gao, W., McKellar, S., Pang, C. N. I., Wu, S., Wu, W., Sy, B., Monk, I. R., Biazik, J. M., Wilkins, M. R., Howden, B. P., Stinear, T. P., Granneman, S., & Tree, J. J. (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), 3558. https://doi.org/10.1038/s41467-022-31177-8

Menard, G., Silard, C., Suriray, M., Rouillon, A., & Augagneur, Y. (2022). Thirty Years of sRNA-Mediated Regulation in Staphylococcus aureus : From Initial Discoveries to In Vivo Biological Implications. *International Journal of Molecular Sciences*, *23*(13), 7346. https://doi.org/10.3390/ijms23137346

Milón, P., Maracci, C., Filonava, L., Gualerzi, C. O., & Rodnina, M. V. (2012). Real-time assembly landscape of bacterial 30S translation initiation complex. *Nature Structural & Molecular Biology*, *19*(6), 609-615. https://doi.org/10.1038/nsmb.2285

Minakhin, L., Bhagat, S., Brunning, A., Campbell, E. A., Darst, S. A., Ebright, R. H., & Severinov, K. (2001). Bacterial RNA polymerase subunit omega and eukaryotic RNA polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(3), 892-897. https://doi.org/10.1073/pnas.98.3.892

Mitchell, J. E., Zheng, D., Busby, S. J. W., & Minchin, S. D. (2003). Identification and analysis of ‘extended –10’ promoters in Escherichia coli. *Nucleic Acids Research*, *31*(16), 4689. https://doi.org/10.1093/nar/gkg694

Miyakoshi, M., Chao, Y., & Vogel, J. (2015). Regulatory small RNAs from the 3’ regions of bacterial mRNAs. *Current Opinion in Microbiology*, *24*, 132-139. https://doi.org/10.1016/j.mib.2015.01.013

Mohan, A., Whyte, S., Wang, X., Nashimoto, M., & Levinger, L. (1999). The 3’ end CCA of mature tRNA is an antideterminant for eukaryotic 3’-tRNase. *RNA (New York, N.Y.)*, *5*(2), 245-256. https://doi.org/10.1017/s1355838299981256

Mohr, D., Wintermeyer, W., & Rodnina, M. V. (2002). GTPase activation of elongation factors Tu and G on the ribosome. *Biochemistry*, *41*(41), 12520-12528. https://doi.org/10.1021/bi026301y

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621-628. https://doi.org/10.1038/nmeth.1226

Munro, A. J., Jackson, R. J., & Korner, A. (1964). Studies on the nature of polysomes. *Biochemical Journal*, *92*(2), 289-299.

Murakami, K. S., & Darst, S. A. (2003). Bacterial RNA polymerases : The wholo story. *Current Opinion in Structural Biology*, *13*(1), 31-39. https://doi.org/10.1016/s0959-440x(02)00005-2

Nersisyan, L., Ropat, M., & Pelechano, V. (2020). Improved computational analysis of ribosome dynamics from 5′P degradome data using fivepseq. *NAR Genomics and Bioinformatics*, *2*(4), lqaa099. https://doi.org/10.1093/nargab/lqaa099

Newbury, S. F., Smith, N. H., Robinson, E. C., Hiles, I. D., & Higgins, C. F. (1987). Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell*, *48*(2), 297-310. https://doi.org/10.1016/0092-8674(87)90433-8

Newman, J. A., Hewitt, L., Rodrigues, C., Solovyova, A., Harwood, C. R., & Lewis, R. J. (2011). Unusual, Dual Endo- and Exonuclease Activity in the Degradosome Explained by Crystal Structure Analysis of RNase J1. *Structure*, *19*(9), 1241-1251. https://doi.org/10.1016/j.str.2011.06.017

Obana, N., Nakao, R., Nagayama, K., Nakamura, K., Senpuku, H., & Nomura, N. (2017). Immunoactive Clostridial Membrane Vesicle Production Is Regulated by a Sporulation

Factor. *Infection and Immunity*, *85*(5), 10.1128/iai.00096-17. https://doi.org/10.1128/iai.00096-17

O'Donnell, S. M., & Janssen, G. R. (2002). Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in Escherichia coli. *Journal of Bacteriology*, *184*(23), 6730-6733. https://doi.org/10.1128/JB.184.23.6730-6733.2002

Olson, P. D., Kuechenmeister, L. J., Anderson, K. L., Daily, S., Beenken, K. E., Roux, C. M., Reniere, M. L., Lewis, T. L., Weiss, W. J., Pulse, M., Nguyen, P., Simecka, J. W., Morrison, J. M., Sayood, K., Asojo, O. A., Smeltzer, M. S., Skaar, E. P., & Dunman, P. M. (2011). Small Molecule Inhibitors of Staphylococcus aureus RnpA Alter Cellular mRNA Turnover, Exhibit Antimicrobial Activity, and Attenuate Pathogenesis. *PLOS Pathogens*, *7*(2), e1001287. https://doi.org/10.1371/journal.ppat.1001287

Oussenko, I. A., Sanchez, R., & Bechhofer, D. H. (2002). Bacillus subtilis YhaM, a member of a new family of 3'-to-5' exonucleases in gram-positive bacteria. *Journal of Bacteriology*, *184*(22), 6250-6259. https://doi.org/10.1128/JB.184.22.6250-6259.2002

Paget, M. S. B., & Helmann, J. D. (2003). The sigma70 family of sigma factors. *Genome Biology*, *4*(1), 203. https://doi.org/10.1186/gb-2003-4-1-203

Papenfort, K., & Vanderpool, C. K. (2015). Target activation by regulatory RNAs in bacteria. *FEMS Microbiology Reviews*, *39*(3), 362-378. https://doi.org/10.1093/femsre/fuv016

Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, *32*(10), 1446-1453. https://doi.org/10.1093/bioinformatics/btw026

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417-419. https://doi.org/10.1038/nmeth.4197

Pei, X.-Y., Bralley, P., Jones, G. H., & Luisi, B. F. (2015). Linkage of catalysis and 5' end recognition in ribonuclease RNase J. *Nucleic Acids Research*, *43*(16), 8066-8076. https://doi.org/10.1093/nar/gkv732

Pelechano, V., Wei, W., & Steinmetz, L. M. (2015). Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell*, *161*(6), 1400-1412. https://doi.org/10.1016/j.cell.2015.05.008

Perwez, T., & Kushner, S. R. (2006). RNase Z in Escherichia coli plays a significant role in mRNA decay. *Molecular Microbiology*, *60*(3), 723-737. https://doi.org/10.1111/j.1365-2958.2006.05124.x

Peters, J. M., Vangeloff, A. D., & Landick, R. (2011). Bacterial transcription terminators : The RNA 3'-end chronicles. *Journal of Molecular Biology*, *412*(5), 793-813. https://doi.org/10.1016/j.jmb.2011.03.036

Philippe, N., Salson, M., Commes, T., & Rivals, E. (2013). CRAC : An integrated approach to the analysis of RNA-seq reads. *Genome Biology*, *14*(3), R30. https://doi.org/10.1186/gb-2013-14-3-r30

Phung, D. K., Etienne, C., Batista, M., Langendijk-Genevaux, P., Moalic, Y., Laurent, S., Liuu, S., Morales, V., Jebbar, M., Fichant, G., Bouvier, M., Flament, D., & Clouet-d'Orval, B. (2020). RNA processing machineries in Archaea : The 5'-3' exoribonuclease aRNase J of the β-CASP family is engaged specifically with the helicase ASH-Ski2 and the 3'-5' exoribonucleolytic RNA exosome machinery. *Nucleic Acids Research*, *48*(7), 3832-3847. https://doi.org/10.1093/nar/gkaa052

Piton, J., Larue, V., Thillier, Y., Dorléans, A., Pellegrini, O., Li de la Sierra-Gallay, I., Vasseur, J.-J., Debart, F., Tisné, C., & Condon, C. (2013). Bacillus subtilis RNA deprotection enzyme RppH recognizes guanosine in the second position of its substrates. *Proceedings of the National Academy of Sciences*, *110*(22), 8858-8863.

Płociński, P., Macios, M., Houghton, J., Niemiec, E., Płocińska, R., Brzostek, A., Słomka, M., Dziadek, J., Young, D., & Dziembowski, A. (2019). Proteomic and transcriptomic experiments reveal an essential role of RNA degradosome complexes in shaping the transcriptome of Mycobacterium tuberculosis. *Nucleic Acids Research*, *47*(11), 5892-5905. https://doi.org/10.1093/nar/gkz251

Pobre, V., & Arraiano, C. M. (2015). Next generation sequencing analysis reveals that the ribonucleases RNase II, RNase R and PNPase affect bacterial motility and biofilm formation in E. coli. *BMC Genomics*, *16*(1), 72. https://doi.org/10.1186/s12864-015-1237-6

Prados, J., Linder, P., & Redder, P. (2016). TSS-EMOTE, a refined protocol for a more complete and less biased global mapping of transcription start sites in bacterial pathogens. *BMC Genomics*, *17*(1), 849. https://doi.org/10.1186/s12864-016-3211-3

Purusharth, R. I., Klein, F., Sulthana, S., Jäger, S., Jagannadham, M. V., Evguenieva-Hackenberg, E., Ray, M. K., & Klug, G. (2005). Exoribonuclease R interacts with endoribonuclease E and an RNA helicase in the psychrotrophic bacterium Pseudomonas syringae Lz4W. *The Journal of Biological Chemistry*, *280*(15), 14572-14578. https://doi.org/10.1074/jbc.M413507200

Py, B., Higgins, C. F., Krisch, H. M., & Carpousis, A. J. (1996). A DEAD-box RNA helicase in the Escherichia coli RNA degradosome. *Nature*, *381*(6578), 169-172. https://doi.org/10.1038/381169a0

Raj, R., Nadig, S., Patel, T., & Gopal, B. (2020). Structural and biochemical characteristics of two Staphylococcus epidermidis RNase J paralogs RNase J1 and RNase J2. *The Journal of Biological Chemistry*, *295*(49), 16863-16876. https://doi.org/10.1074/jbc.RA120.014876

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools : A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(Web Server issue), W187-191. https://doi.org/10.1093/nar/gku365

Redder, P. (2015). Using EMOTE to map the exact 5'-ends of processed RNA on a transcriptome-wide scale. *Methods in Molecular Biology (Clifton, N.J.)*, *1259*, 69-85. https://doi.org/10.1007/978-1-4939-2214-7_5

Redder, P. (2018). Molecular and genetic interactions of the RNA degradation machineries in Firmicute bacteria. *Wiley Interdisciplinary Reviews. RNA*, *9*(2). https://doi.org/10.1002/wrna.1460

Redko, Y., Aubert, S., Stachowicz, A., Lenormand, P., Namane, A., Darfeuille, F., Thibonnier, M., & De Reuse, H. (2013). A minimal bacterial RNase J-based degradosome is associated with translating ribosomes. *Nucleic Acids Research*, *41*(1), 288-301. https://doi.org/10.1093/nar/gks945

Redko, Y., Bechhofer, D. H., & Condon, C. (2008). Mini-III, an unusual member of the RNase III family of enzymes, catalyses 23S ribosomal RNA maturation in B. subtilis. *Molecular Microbiology*, *68*(5), 1096-1106. https://doi.org/10.1111/j.1365-2958.2008.06207.x

Redko, Y., Galtier, E., Arnion, H., Darfeuille, F., Sismeiro, O., Coppée, J.-Y., Médigue, C., Weiman, M., Cruveiller, S., & De Reuse, H. (2016). RNase J depletion leads to massive changes in mRNA abundance in Helicobacter pylori. *RNA Biology*, *13*(2), 243-253. https://doi.org/10.1080/15476286.2015.1132141

Redko, Y., Li de la Sierra-Gallay, I., & Condon, C. (2007). When all's zed and done : The structure and function of RNase Z in prokaryotes. *Nature Reviews. Microbiology*, *5*(4), 278-286. https://doi.org/10.1038/nrmicro1622

Richards, J., & Belasco, J. G. (2011). Ribonuclease J : How to lead a double life. *Structure (London, England: 1993)*, *19*(9), 1201-1203. https://doi.org/10.1016/j.str.2011.08.004

Richards, J., Liu, Q., Pellegrini, O., Celesnik, H., Yao, S., Bechhofer, D. H., Condon, C., & Belasco, J. G. (2011). An RNA pyrophosphohydrolase triggers 5′-exonucleolytic degradation of mRNA in Bacillus subtilis. *Molecular cell*, *43*(6), 940-949. https://doi.org/10.1016/j.molcel.2011.07.023

Risso, D., Ngai, J., Speed, T. P., & Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, *32*(9), 896-902. https://doi.org/10.1038/nbt.2931

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Roberts, C., Anderson, K. L., Murphy, E., Projan, S. J., Mounts, W., Hurlburt, B., Smeltzer, M., Overbeek, R., Disz, T., & Dunman, P. M. (2006). Characterizing the Effect of the Staphylococcus aureus Virulence Factor Regulator, SarA, on Log-Phase mRNA Half-Lives. *Journal of Bacteriology*, *188*(7), 2593-2603. https://doi.org/10.1128/jb.188.7.2593-2603.2006

Robertson, H. D., Webster, R. E., & Zinder, N. D. (1968). Purification and properties of ribonuclease III from Escherichia coli. *The Journal of Biological Chemistry*, *243*(1), 82-91.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR : A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. https://doi.org/10.1186/gb-2010-11-3-r25

Rocha, E. P., Danchin, A., & Viari, A. (1999). Translation in Bacillus subtilis : Roles and trends of initiation and termination, insights from a genome analysis. *Nucleic Acids Research*, *27*(17), 3567-3576. https://doi.org/10.1093/nar/27.17.3567

Rodnina, M. V. (2018). Translation in Prokaryotes. *Cold Spring Harbor Perspectives in Biology*, *10*(9), a032664. https://doi.org/10.1101/cshperspect.a032664

Roehr, J. T., Dieterich, C., & Reinert, K. (2017). Flexbar 3.0—SIMD and multicore parallelization. *Bioinformatics (Oxford, England)*, *33*(18), 2941-2942. https://doi.org/10.1093/bioinformatics/btx330

Rojo, F. (2001). Mechanisms of transcriptional repression. *Current Opinion in Microbiology*, *4*(2), 145-151. https://doi.org/10.1016/s1369-5274(00)00180-6

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., & Gourse, R. L. (1993). A third recognition element in bacterial promoters : DNA binding by the alpha subunit of RNA polymerase. *Science (New York, N.Y.)*, *262*(5138), 1407-1413. https://doi.org/10.1126/science.8248780

Roux, C. M., DeMuth, J. P., & Dunman, P. M. (2011). Characterization of components of the Staphylococcus aureus mRNA degradosome holoenzyme-like complex. *Journal of Bacteriology*, *193*(19), 5520-5526. https://doi.org/10.1128/JB.05485-11

Ruff, E. F., Record, M. T., & Artsimovitch, I. (2015). Initial events in bacterial transcription initiation. *Biomolecules*, *5*(2), 1035-1062. https://doi.org/10.3390/biom5021035

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463-5467.

Saunders, K., Bert, A. G., Dredge, B. K., Toubia, J., Gregory, P. A., Pillman, K. A., Goodall, G. J., & Bracken, C. P. (2020). Insufficiently complex unique-molecular identifiers (UMIs) distort small RNA sequencing. *Scientific Reports*, *10*(1), 14593. https://doi.org/10.1038/s41598-020-71323-0

Serganov, A., & Nudler, E. (2013). A decade of riboswitches. *Cell*, *152*(1-2), 17-24. https://doi.org/10.1016/j.cell.2012.12.024

Šetinová, D., Šmídová, K., Pohl, P., Musić, I., & Bobek, J. (2018). RNase III-Binding-mRNAs Revealed Novel Complementary Transcripts in Streptomyces. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02693

Sette, M., van Tilborg, P., Spurio, R., Kaptein, R., Paci, M., Gualerzi, C. O., & Boelens, R. (1997). The structure of the translational initiation factor IF1 from E.coli contains an oligomer-binding motif. *The EMBO Journal*, *16*(6), 1436-1443. https://doi.org/10.1093/emboj/16.6.1436

Shahbabian, K., Jamalli, A., Zig, L., & Putzer, H. (2009). RNase Y, a novel endoribonuclease, initiates riboswitch turnover in Bacillus subtilis. *The EMBO Journal*, *28*(22), 3523-3533. https://doi.org/10.1038/emboj.2009.283

Shepherd, J., & Ibba, M. (2015). Bacterial transfer RNAs. *FEMS Microbiology Reviews*, *39*(3), 280-300. https://doi.org/10.1093/femsre/fuv004

Siegel, A. R., & Wemmer, D. E. (2016). Role of the σ54 Activator Interacting Domain in Bacterial Transcription Initiation. *Journal of molecular biology*, *428*(23), 4669-4685. https://doi.org/10.1016/j.jmb.2016.10.007

Simonetti, A., Marzi, S., Fabbretti, A., Hazemann, I., Jenner, L., Urzhumtsev, A., Gualerzi, C. O., & Klaholz, B. P. (2013). Structure of the protein core of translation initiation factor 2 in apo, GTP-bound and GDP-bound forms. *Acta Crystallographica. Section D, Biological Crystallography*, *69*(Pt 6), 925-933. https://doi.org/10.1107/S0907444913006422

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools : Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491-499. https://doi.org/10.1101/gr.209601.116

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*, Article3. https://doi.org/10.2202/1544-6115.1027

Southern, E., Mir, K., & Shchepinov, M. (1999). Molecular interactions on microarrays. *Nature Genetics*, *21*(1), 5-9. https://doi.org/10.1038/4429

Sprang, M., Andrade-Navarro, M. A., & Fontaine, J.-F. (2022). Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. *BMC Bioinformatics*, *23*(Suppl 6), 279. https://doi.org/10.1186/s12859-022-04775-y

Srouji, J. R., Xu, A., Park, A., Kirsch, J. F., & Brenner, S. E. (2017). The evolution of function within the Nudix homology clan. *Proteins: Structure, Function, and Bioinformatics*, *85*(5), 775-811. https://doi.org/10.1002/prot.25223

Starmer, J., Stomp, A., Vouk, M., & Bitzer, D. (2006). Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors. *PLOS Computational Biology*, *2*(5), e57. https://doi.org/10.1371/journal.pcbi.0020057

Stead, M. B., Marshburn, S., Mohanty, B. K., Mitra, J., Pena Castillo, L., Ray, D., van Bakel, H., Hughes, T. R., & Kushner, S. R. (2011). Analysis of Escherichia coli RNase E and RNase III activity in vivo using tiling microarrays. *Nucleic Acids Research*, *39*(8), 3188-3203. https://doi.org/10.1093/nar/gkq1242

Stent, G. S. (1964). THE OPERON : ON ITS THIRD ANNIVERSARY. MODULATION OF TRANSFER RNA SPECIES CAN PROVIDE A WORKABLE MODEL OF AN OPERATOR-LESS OPERON. *Science (New York, N.Y.)*, *144*(3620), 816-820. https://doi.org/10.1126/science.144.3620.816

Strahl, H., Turlan, C., Khalid, S., Bond, P. J., Kebalo, J.-M., Peyron, P., Poljak, L., Bouvier, M., Hamoen, L., Luisi, B. F., & Carpousis, A. J. (2015). Membrane recognition and dynamics of the RNA degradosome. *PLoS Genetics*, *11*(2), e1004961. https://doi.org/10.1371/journal.pgen.1004961

Sy, B., Wong, J., Granneman, S., Tollervey, D., Gally, D., & Tree, J. J. (2018). High-Resolution, High-Throughput Analysis of Hfq-Binding Sites Using UV Crosslinking and Analysis of cDNA (CRAC). *Methods in Molecular Biology (Clifton, N.J.)*, *1737*, 251-272. https://doi.org/10.1007/978-1-4939-7634-8_15

Symmons, M. F., Jones, G. H., & Luisi, B. F. (2000). A duplicated fold is the structural basis for polynucleotide phosphorylase catalytic activity, processivity, and regulation. *Structure (London, England: 1993)*, *8*(11), 1215-1226. https://doi.org/10.1016/s0969-2126(00)00521-9

Taboada, B., Estrada, K., Ciria, R., & Merino, E. (2018). Operon-mapper : A web server for precise operon identification in bacterial and archaeal genomes. *Bioinformatics*, *34*(23), 4118-4120. https://doi.org/10.1093/bioinformatics/bty496

Taghbalout, A., & Rothfield, L. (2007). RNaseE and the other constituents of the RNA degradosome are components of the bacterial cytoskeleton. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1667-1672. https://doi.org/10.1073/pnas.0610491104

Taverniti, V., Forti, F., Ghisotti, D., & Putzer, H. (2011). Mycobacterium smegmatis RNase J is a 5'-3' exo-/endoribonuclease and both RNase J and RNase E are involved in ribosomal RNA maturation. *Molecular Microbiology*, *82*(5), 1260-1276. https://doi.org/10.1111/j.1365-2958.2011.07888.x

Tedin, K., Moll, I., Grill, S., Resch, A., Graschopf, A., Gualerzi, C. O., & Bläsi, U. (1999). Translation initiation factor 3 antagonizes authentic start codon selection on leaderless mRNAs. *Molecular Microbiology*, *31*(1), 67-77. https://doi.org/10.1046/j.1365-2958.1999.01147.x

Tejada-Arranz, A., Crécy-Lagard, V. de, & Reuse, H. de. (2020). Bacterial RNA Degradosomes : Molecular Machines under Tight Control. *Trends in Biochemical Sciences*, *45*(1), 42-57. https://doi.org/10.1016/j.tibs.2019.10.002

Tejada-Arranz, A., Matos, R. G., Quentin, Y., Bouilloux-Lafont, M., Galtier, E., Briolat, V., Kornobis, E., Douché, T., Matondo, M., Arraiano, C. M., Raynal, B., & De Reuse, H. (2021). RNase R is associated in a functional complex with the RhpA DEAD-box RNA

helicase in Helicobacter pylori. *Nucleic Acids Research*, *49*(9), 5249-5264. https://doi.org/10.1093/nar/gkab283

Thomason, M. K., Bischler, T., Eisenbart, S. K., Förstner, K. U., Zhang, A., Herbig, A., Nieselt, K., Sharma, C. M., & Storz, G. (2015). Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *Journal of Bacteriology*, *197*(1), 18-28. https://doi.org/10.1128/JB.02096-14

Tollerson, R., & Ibba, M. (2020). Translational regulation of environmental adaptation in bacteria. *The Journal of Biological Chemistry*, *295*(30), 10434-10445. https://doi.org/10.1074/jbc.REV120.012742

Tong 2015. (s. d.). *Staphylococcus aureus Infections : Epidemiology, Pathophysiology, Clinical Manifestations, and Management* | *Clinical Microbiology Reviews*. Consulté 12 juillet 2024, à l'adresse https://journals.asm.org/doi/10.1128/cmr.00134-14

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., & Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 180. https://doi.org/10.1186/s13059-020-02090-4

Travis, A. J., Moody, J., Helwak, A., Tollervey, D., & Kudla, G. (2014). Hyb : A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. *Methods (San Diego, Calif.)*, *65*(3), 263-273. https://doi.org/10.1016/j.ymeth.2013.10.015

Tree, J. J., Granneman, S., McAteer, S. P., Tollervey, D., & Gally, D. L. (2014). Identification of Bacteriophage-Encoded Anti-sRNAs in Pathogenic Escherichia coli. *Molecular Cell*, *55*(2), 199-213. https://doi.org/10.1016/j.molcel.2014.05.006

Tucker, B. J., & Breaker, R. R. (2005). Riboswitches as versatile gene control elements. *Current Opinion in Structural Biology*, *15*(3), 342-348. https://doi.org/10.1016/j.sbi.2005.05.003

Udagawa, T., Shimizu, Y., & Ueda, T. (2004). Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria. *The*

Journal of Biological Chemistry, *279*(10), 8539-8546. https://doi.org/10.1074/jbc.M308784200

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database : Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439-D444. https://doi.org/10.1093/nar/gkab1061

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Hieter, P., Vogelstein, B., & Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell*, *88*(2), 243-251. https://doi.org/10.1016/s0092-8674(00)81845-0

Vincent, H. A., & Deutscher, M. P. (2006). Substrate recognition and catalysis by the exoribonuclease RNase R. *The Journal of Biological Chemistry*, *281*(40), 29769-29775. https://doi.org/10.1074/jbc.M606744200

Vogel, J., & Luisi, B. F. (2011). Hfq and its constellation of RNA. *Nature Reviews. Microbiology*, *9*(8), 578-589. https://doi.org/10.1038/nrmicro2615

Wachi, M., Umitsuki, G., Shimizu, M., Takada, A., & Nagai, K. (1999). Escherichia coli cafA gene encodes a novel RNase, designated as RNase G, involved in processing of the 5’ end of 16S rRNA. *Biochemical and Biophysical Research Communications*, *259*(2), 483-488. https://doi.org/10.1006/bbrc.1999.0806

Wade, J. T., & Grainger, D. C. (2014). Pervasive transcription : Illuminating the dark matter of bacterial transcriptomes. *Nature Reviews. Microbiology*, *12*(9), 647-653. https://doi.org/10.1038/nrmicro3316

Wagner, E. G., & Simons, R. W. (1994). Antisense RNA control in bacteria, phages, and plasmids. *Annual Review of Microbiology*, *48*, 713-742. https://doi.org/10.1146/annurev.mi.48.100194.003433

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data : RPKM measure is inconsistent among samples. *Theory in Biosciences = Theorie in Den Biowissenschaften*, *131*(4), 281-285. https://doi.org/10.1007/s12064-012-0162-3

Waters, L. S., & Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*, *136*(4), 615-628. https://doi.org/10.1016/j.cell.2009.01.043

Waters, S. A., McAteer, S. P., Kudla, G., Pang, I., Deshpande, N. P., Amos, T. G., Leong, K. W., Wilkins, M. R., Strugnell, R., Gally, D. L., Tollervey, D., & Tree, J. J. (2017). Small RNA interactome of pathogenic E. coli revealed through crosslinking of RNase E. *The EMBO Journal*, *36*(3), 374-387. https://doi.org/10.15252/embj.201694639

Webb, S., Hector, R. D., Kudla, G., & Granneman, S. (2014). PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biology*, *15*(1), R8. https://doi.org/10.1186/gb-2014-15-1-r8

Weiner, P. (1973). Linear pattern matching algorithms. *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, 1-11. https://doi.org/10.1109/SWAT.1973.13

Weiss, A., Ibarra, J. A., Paoletti, J., Carroll, R. K., & Shaw, L. N. (2014). The δ subunit of RNA polymerase guides promoter selectivity and virulence in Staphylococcus aureus. *Infection and Immunity*, *82*(4), 1424-1435. https://doi.org/10.1128/IAI.01508-14

Wen, T., Oussenko, I. A., Pellegrini, O., Bechhofer, D. H., & Condon, C. (2005). Ribonuclease PH plays a major role in the exonucleolytic maturation of CCA-containing tRNA precursors in Bacillus subtilis. *Nucleic Acids Research*, *33*(11), 3636-3643. https://doi.org/10.1093/nar/gki675

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*(1), 129. https://doi.org/10.1186/s13059-019-1727-y

Wu, H., Wang, C., & Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, *14*(2), 232-243. https://doi.org/10.1093/biostatistics/kxs033

Yamamoto, H., Wittek, D., Gupta, R., Qin, B., Ueda, T., Krause, R., Yamamoto, K., Albrecht, R., Pech, M., & Nierhaus, K. H. (2016). 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria. *Proceedings of the National Academy of Sciences*, *113*(9), E1180-E1189. https://doi.org/10.1073/pnas.1524554113

Yasrebi, H., & Redder, P. (2015). EMOTE-conv : A Computational Pipeline to Convert Exact Mapping of Transcriptome Ends (EMOTE) Data to the Lists of Quantified Genomic Positions Correlated to Related Genomic Information. *Journal of Applied Bioinformatics & Computational Biology*, *04*. https://doi.org/10.4172/2329-9533.1000118

Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, *26*(8), 903-909. https://doi.org/10.1261/rna.074922.120

Zhao, Y., Li, M.-C., Konaté, M. M., Chen, L., Das, B., Karlovich, C., Williams, P. M., Evrard, Y. A., Doroshow, J. H., & McShane, L. M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine*, *19*(1), 269. https://doi.org/10.1186/s12967-021-02936-w

Zhao, Y., Lu, M., Zhang, H., Hu, J., Zhou, C., Xu, Q., Ul Hussain Shah, A. M., Xu, H., Wang, L., & Hua, Y. (2015). Structural insights into catalysis and dimerization enhanced exonuclease activity of RNase J. *Nucleic Acids Research*, *43*(11), 5550-5559. https://doi.org/10.1093/nar/gkv444

Ziemann, M., Kaspi, A., & El-Osta, A. (2016). Evaluation of microRNA alignment techniques. *RNA (New York, N.Y.)*, *22*(8), 1120-1138. https://doi.org/10.1261/rna.055509.115

Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, *9*(1), 133-148. https://doi.org/10.1093/nar/9.1.133

Zuo, Y., & Deutscher, M. P. (2002). The physiological role of RNase T can be explained by its unusual substrate specificity. *The Journal of Biological Chemistry*, *277*(33), 29654-29661. https://doi.org/10.1074/jbc.M204252200

Zuo, Y., Wang, Y., & Malhotra, A. (2005). Crystal structure of Escherichia coli RNase D, an exoribonuclease involved in structured RNA processing. *Structure (London, England: 1993)*, *13*(7), 973-984. https://doi.org/10.1016/j.str.2005.04.015

Watson, J., Baker, T., Bell, S., Gann, A., Levine, M. & Losick, R. (2009). Biologie moléculaire du gène (6ème édition). PEARSON.

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Titre : Régulation de la double activité de dégradation de l'ARN de la RNase J par les modifications épi-transcriptomiques et les structures secondaires

Mots clés : Bioinformatique, traitement des données NGS, Inférence des structures d'ARN, modifications épi-transcriptomiques, dégradation de l'ARN, ARN pyrophosphohydrolases

Résumé : Les ribonucléases permettant de dégrader et/ou maturer les ARN sont vitales pour l'adaptation de la cellule à des changements environnementaux brusques. Elles possèdent deux modes d'action, un mode endonucléolytique clivant les ARN de manière interne et un mode exonucléolytique les dégradant à partir des extrémités 5' ou 3'. Les endoribonucléases reflètent un cas remarquable d'évolution convergente car cette activité est portée par la RNase E, la RNase Y et la RNase J, protéines sans aucune similarité de séquences. Il a aussi été montré qu'en fonction des espèces bactériennes, les substrats et l'activité de ces ribonucléases pouvaient être différents, les rendant parfois essentielles à la survie de la cellule. Mes travaux de thèse portaient sur l'étude du rôle et de la régulation du complexe RNase J chez la bactérie pathogène Staphylococcus aureus. Ce complexe possède une activité duale, endo- et 5' exoribonucléolytique. Cette dernière montre une préférence pour les ARN monophosphorylés en 5' (5'P) par rapport aux ARN triphosphorylés (5'PPP). Lors de mon travail de thèse, j'ai développé des méthodes statistiques et bioinformatiques adaptées au traitement et à l'analyse de données de séquençage des extrémités 5' des ARN, données obtenues par une équipe collaboratrice ayant développé des protocoles expérimentaux permettant d'étudier in vivo et à l'échelle du génome l'activité du complexe RNase J. Ces développements ont été regroupés au sein d'une librairie R libre et gratuite. La première question abordée a été l'identification de la ou des protéines NUDIX ribopyrophosphohydrolases (RppH) responsables de la conversion des ARN 5'PPP en 5'P en vue de leur dégradation par le complexe RNase J. Parmi les cinq RppH prédites in silico chez S. aureus, seule l'étude de la plus similaire à celle de B. subtilis a été publiée concluant à son faible impact sur la stabilité des ARN. Mes travaux ont permis de valider le protocole expérimental par rapport à sa capacité à traduire les abondances en proportions des états de phosphorylation de chaque ARN afin de déterminer statistiquement des différences de proportions. Des mutants de délétion pour les cinq RppH candidates (délétions seules ou combinées) ont été obtenus. Aucun des mutants étudiés n'a perdu totalement l'activité RppH. Le système de régulation de ce mécanisme se révèle donc plus complexe qu'attendu chez S. aureus et la difficulté d'obtention de certains mutants de délétion suggère un mécanisme particulier dans cet organisme.
En parallèle, j'ai exploité des données RNA-Seq déjà disponibles obtenues en bloquant la transcription et en mesurant l'abondance des ARN au cours du temps par analyses transcriptomiques, entre une souche S. aureus sauvage et une souche ΔRNase J. En comparant la stabilité des ARN, j'ai montré que 47% des gènes avaient des ARN dont la dégradation dépend de cette enzyme, avec une préférence pour les gènes fortement exprimés.
J'ai également réalisé l'analyse quantitative des intermédiaires de dégradation par rapport aux cadres ouverts de lecture, obtenus grâce à une autre méthode de quantification in vivo des extrémités 5' des ARN. J'ai mis en évidence, dans la souche sauvage, une distribution périodique de 3 nucléotides correspondant aux codons qui n'est quasiment plus observée dans le mutant ΔRNase J. J'ai ainsi montré que la RNase J peut suivre le dernier ribosome en cours de traduction et qu'elle serait un acteur majeur du mécanisme de dégradation co-traductionnelle.
D'autres analyses de ces mêmes données m'ont permis de prédire 125 sites de clivages endoribonucléolytiques de la RNase J. Ces résultats préliminaires pourront servir de point de départ pour des études complémentaires.
L'ensemble de mes travaux a donc apporté de nouvelles informations sur l'activité in vivo et le rôle du complexe RNase J chez S. aureus et les méthodes que j'ai développées dédiées à l'analyse des données 5' RNA-Seq pourront être utilisées dans le cadre d'autres études de ce type.

Title: Regulation of the dual RNA degradation activity of RNase J by RNA epi-transcriptomic modifications and secondary structures

Key words: Bioinformatics, NGS data processing, RNA structure inference, epi-transcriptomic modifications, RNA degradation, RNA pyrophosphohydrolases

Abstract: Ribonucleases responsible of RNA degradation and/or maturation are vital for the cell's adaptation to sudden environmental changes. They have two modes of action, an endonucleolytic mode cleaving RNAs internally, and an exonucleolytic mode degrading them from their 5' or 3' ends. Endoribonucleases reflect a remarkable case of convergent evolution, as this activity is carried by RNase E, RNase Y and RNase J, proteins with no sequence similarity. It has also been shown that, depending on the bacterial species, the substrates and activity of these ribonucleases can differ, making them sometimes essential for cell survival. My thesis work focused on the role and regulation of the RNase J complex in the pathogenic bacteria Staphylococcus aureus. This complex has dual endo- and 5' exoribonucleolytic activity. The latter shows a preference for 5' monophosphorylated RNAs (5'P) over triphosphorylated RNAs (5'PPP). During my thesis work, I developed statistical and bioinformatics methods adapted to the processing and analysis of RNA 5'-end sequencing data obtained by a collaborating team who had developed experimental protocols to study the activity of the RNase J complex in vivo and on a genome-wide scale. These developments have been integrated into an open-source R package accessible on the CNRS GitLab instance. The first question addressed was the identification of the NUDIX ribopyrophosphohydrolase (RppH) protein(s) responsible for converting 5'PPP RNAs to 5'P for degradation by the RNase J complex. Of the five RppHs predicted in silico in S. aureus, only the one most similar to that of B. subtilis has been published, concluding that it has little impact on RNA stability. My work has validated the experimental protocol in terms of its ability to translate abundances into proportions of the phosphorylation states of each RNA in order to statistically identify differences in proportions. Deletion mutants for the five RppH candidates (single or combined deletions) were obtained. None of the mutants studied totally lost RppH activity. The regulatory system for this mechanism is therefore more complex than expected in S. aureus, and the difficulty to obtain certain deletion mutants suggests a particular mechanism in this organism. In parallel, I exploited previously available RNA-Seq data obtained by blocking transcription and measuring RNA abundance over time by transcriptomic analyses, between a wild-type S. aureus strain and a ΔRNase J strain. By comparing RNA stability, I showed that 47% of genes had RNAs whose degradation depended on this enzyme, with a preference for highly expressed genes. I also carried out a quantitative analysis of degradation intermediates relative to open reading frames, obtained using another method for in vivo quantification of RNA 5' ends. In the wild-type strain, I highlighted a periodic distribution of 3 nucleotides corresponding to codons, which is practically no longer observed in the ΔRNase J mutant. I thus demonstrated that RNase J can follow the last ribosome during translation, and that it is a major player in the co-translational degradation mechanism. Further analysis of the same data enabled me to predict 125 endoribonucleolytic cleavage sites for RNase J. These preliminary results may serve as a starting point for further studies. My work as a whole has therefore provided new information on the in vivo activity and role of the RNase J complex in S. aureus, and the methods I have developed dedicated to the analysis of 5' RNA-Seq data may be used in other studies of this type.